Gender, smoking and blood pressure and the initial presentation of a wide range of cardiovascular diseases: Prospective cohort study in 1.5 million patients using linked electronic health records

Julia Louise George
UCL

PhD in Epidemiology

Declaration

I, Julia Louise George, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Acknowledgments

I was supported in researching and writing this thesis by a National Institute for Health Research (NIHR) Doctoral Fellowship (DRF-2009-02-50).

I am grateful to my supervisors, Harry Hemingway (University College London) and Liam Smeeth (London School of Hygiene and Tropical Medicine) for the time and effort they have put into the supervision of my thesis. I would particularly like to thank Professor Hemingway for supporting my research ambitions from the time of our first meeting. This thesis would not have been possible without the team effort that has gone into creating the CALIBER research platform; I would like to pay particular tribute to the contributions of Emily Herrett and Spiros Denaxas. I received helpful statistical advice for my research from Mai Stafford, Eleni Rapsomaniki, and Owen Nicholas. In the course of developing my research proposal and completing my thesis, I also benefitted from conversations with Adam Timmis, Gene Feder, Sandra Eldridge, Jenny Mindell, Mar Pujades Rodriguez, Sheng Chia Chung, Stephen Bell and the ever perceptive and encouraging Anoop Shah. I was lucky to have a cheerful and responsive proof-reader in Joanne Turner. Thanks also to Natalie Fitzpatrick and my other friends in the Clinical Epidemiology Group for cheering me on during the dark moments that visit anyone trying to complete a PhD. I am grateful to Hannah, Katie and Simon for their unswerving support for my PhD, despite my distraction and absences from home over the past four years.

I would like to dedicate this thesis to my parents, Ann and Rolf George, who demonstrated to me from an early age the value of a life of intellectual endeavour.

Table of Contents

List of Figures

List of Tables

Abstract

Background: Myocardial infarction (MI) and stroke are the predominant endpoints of large-scale epidemiological research on cardiovascular disease (CVD) in healthy populations. This thesis capitalised on opportunities presented by linked electronic health records (EHR) to investigate the association of risk factors with the initial presentation of a wide range of pathologically diverse CVDs, with a focus on gender differences.

Objective: To determine whether gender, smoking and blood pressure have homogeneous associations with the initial presentation of a range of CVDs.

Design: Cohort studies using data from the CALIBER research platform linking four data sources (primary care, disease registry, hospitalisation and mortality records) for 1,758,584 patients free from symptomatic CVD registered with 225 UK general practices between 2001 and 2010.

Main outcome measures: Initial presentation of CVD with stable angina, unstable angina, MI, heart failure, ventricular arrhythmias, coronary death, stroke, abdominal aortic aneurysm (AAA) and peripheral arterial disease (PAD).

Results: 69% of initial presentations of CVD (N=95,267) were neither MI nor stroke. Men had higher rates of all presentations, with age-adjusted gender rate ratios from 1.23 (95% confidence interval 1.19-1.27)) for stroke to 4.27 (3.92-4.65) for AAA. The association of current smoking (compared to non-smoking) varied from an age-adjusted hazard ratio (HR) of 1.01 (0.90-1.13) for arrhythmia to 4.71 (4.15-5.35) for AAA. Gender differences were found only for MI (women: 2.59 (2.37-2.83); men: 2.20 (2.05-2.37)) and AAA (women: 7.00 (5.64-8.70); men: 3.88 (3.33-4.53)). The association of systolic blood pressure was similar across all CVD presentations, excepting AAA, ranging from age-adjusted HR of 1.19 (1.16-1.22) per standard deviation (SD) for heart failure to 1.28 (1.24-1.31) for PAD, with minimal gender differences.

Conclusions: Gender, smoking and blood pressure had heterogeneous associations across a wide range of initial CVD presentations. The implications of these findings for public health, clinical practice and research are discussed.

Introduction

This thesis is concerned with atherosclerotic disease, the underlying processes which manifests in different patients as the different atherosclerotic diseases, such as coronary artery disease, ischaemic stroke, and peripheral arterial disease. Historically, this group of diseases has been approached as if they have the same underlying causal processes and therefore should benefit from similar prevention and treatment strategies. While this commonality of cause may be assumed, the majority of epidemiological studies in this area have focussed on studying an individual disease at a time, with "first myocardial infarction" or "first stroke" being common endpoints. In contrast, the approach of this thesis has been to take this assumption of common underlying disease processes seriously by studying the onset of symptomatic atherosclerotic disease overall as indicated by the first manifestation of *any* symptomatic atherosclerotic disease. At the same time, I have also questioned whether the causal processes for this group of diseases are actually as unified as may be assumed. The overarching question for my thesis is therefore to question whether the association of common risk factors varies with the specific disease which marks the onset of symptomatic atherosclerotic disease. Homogeneity in association of risk factors would indicate commonality of causal processes and vindicate a one-size-fits-all approach to risk factor management. However, heterogeneity in association of risk factors with different diseases would indicate differences in causal processes and raise the question of more bespoke approaches to risk factor management.

1. Terminology

Before further discussion, a few comments about terminology are required. As one of the purposes of this thesis is to question whether the common atherosclerotic diseases share the same causal processes and therefore really are diseases subtypes, I wished to get away from using this term. Throughout this thesis, I have therefore used the phrase "onset of cardiovascular disease with … [*e.g. stable angina*]" to refer to the first manifestation of symptomatic atherosclerotic cardiovascular disease with a specific disease such as stable angina. However, this usage can be overly clunky. I have therefore also used the terminology of "initial presentation" used by Murabito et al in an earlier Framingham paper on initial presentation of coronary heart disease",(1) where onset of coronary heart disease with, say, stable angina, is referred to as patients "presenting with stable angina" or "initial presentation with stable angina". While I risked confusion by using the term "presentation", which can also be used in epidemiology to mean the cluster of symptoms with which an individual patient presents to healthcare services at a particular moment in time, I decided the term was apposite, given that I measured onset of cardiovascular disease (CVD) using records of consultation with health services.

Initial presentation of cardiovascular disease with, say, myocardial infarction (MI) should also be distinguished from first MI, a concept more commonly used epidemiology. First MI (or first other cardiovascular disease) usually refers to first presentation of a specific disease rather than first presentation of any cardiovascular disease. First MI could therefore be preceded by another cardiovascular disease such as stable angina and indeed it is not uncommon for studies to use history of CHD as a co-variable in studies of first MI. *Initial presentation of CVD with MI*, could not, however, be preceded by other symptomatic cardiovascular diseases according to the definition used in this thesis.

Definition of initial versus first presentation: The example of myocardial infarction
Initial presentation of cardiovascular disease with myocardial infarction (MI) – First presentation of any symptomatic cardiovascular disease. Not preceded by any other symptomatic cardiovascular disease
First presentation of myocardial infarction – First presentation of a specific cardiovascular disease i.e. first MI. Could be preceded by another symptomatic cardiovascular disease

For the remainder of this chapter, I develop the rationale for my choice of cardiovascular diseases (endpoints) and risk factors. Then, I briefly describe the way in which electronic health records provide a viable approach to answering my questions. I end the chapter by describing the structure of the remainder of my thesis.

2. Choice of endpoints

2.1. Onset of cardiovascular disease

Much of cardiovascular disease epidemiology is dominated by a focus on the association of risk factors with mortality but this approach conflates the relationship between risk factors which may be associated with onset and subsequent progression of the disease, whether to more acute forms of CVD or mortality, obscuring important aspects of aetiology. For example, the association of smoking(2) or South Asian ethnicity(3) with coronary heart disease mortality is well-documented. However, for both these risk factors, there is evidence that while smokers and South Asians are at increased risk of myocardial infarction (MI), they may be more likely to survive an MI than non-smokers / non-Asians.(4,5) Similarly, there is some evidence that in patients with stable angina the rate of myocardial infarction (MI) or death does not vary by gender,(6) in contrast to the usual male excess in rates of MI or CHD mortality.(7) These examples show how a focus only on mortality can obscure a more complex relationship between disease and several

risk factors, indicating the need to look at the disease as it progresses from onset, through different stages and eventually to mortality. (8) A more detailed understanding of the changing nature of risk would allow an assessment of the relative advantage of interventions at different stages in progression of CVD. The first stage in understanding the aetiology is to understand onset, or as the King in *Alice in Wonderland* says "`Begin at the beginning, … and go on till you come to the end: then stop."(9)

2.2. Specific rather than composite endpoints

The study of aetiology of CVD has been dominated by research focussing on composite endpoints, such as CHD mortality or cerebrovascular disease. Take the example of coronary heart disease. The use of composite endpoints has been partly driven by the concept of CHD as if it were a single disease entity, partly by the need to combine endpoints to generate sufficiently powered studies, and partly by the lack of clinical detail to allow more detailed phenotyping in epidemiological research. However, greater specificity in defining coronary diseases, at least with acute coronary syndrome, has clearly borne fruit. The evolution of the definition of myocardial infraction from a purely fatal disease to non-fatal acute myocardial infarction (MI) to the specific MI subtypes of ST-elevation and non-ST elevation MI(10) has led to an improved understanding of coronary heart diseases with differential risk profiles, (11,12) if not completely differential treatments.(13) Further comparisons between myocardial infarction and unstable angina, although less common, have also proven to be informative. (14,15)

Recent evidence has also suggested a shift in the presentation of CHD over time, with a)a decline in case fatality leaving more people living with CHD as a chronic disease;(16–18) b) a decline in STEMI coupled with a rise in nSTEMI;(19,20) and c) a decline in myocardial infarction of whatever type coupled with a rise in angina pectoris.(21,22) These trends suggest a possible shift toward less acute and more chronic forms of coronary heart disease, which in turn argues for the importance of a more nuanced understanding of the onset of different subtypes of CHD and the progression of the disease through multiple stages as it develops over time.

2.3. Broader than CHD or stroke

Although studies comparing risk factors for myocardial infarction with those for stroke are not uncommon, a broader approach of studying the whole range of cardiovascular disease is still relatively rare,(23,24) despite the commonality of risk factors across the range of arterial beds. This approach was first used in a small number of Framingham studies over twenty years ago,(1,25) but has not been commonly used subsequently. This thesis therefore includes atherosclerotic disease in the cerebral, coronary, abdominal and

peripheral arterial beds, but also cardiac arrhythmias and heart failure as two cardiac diseases which may be significant competing risk to the atherosclerotic CVDs and share many of the same risk factors.

3. Choice of risk factors

For the thesis, I chose to focus on three risk factors: gender, smoking and blood pressure. These three are arguably the most important, most widely recorded, and most widely investigated measures of risk, covering demographic, behavioural and physiological domains.

3.1. Gender

The incidence of specific CVD presentations, when studied individually, varies by gender,(26–29) and there is some evidence that the extent of gender difference in atherosclerosis varies by arterial bed.(30) However, it is not known whether the onset of CVD with specific presentations varies by gender or the extent of such variation; these questions are the subject the first substantive set of analyses for this thesis.

Throughout the thesis, gender-specific results have been reported and the effect modification of gender has been investigated where appropriate. The reasons for this are twofold. First there is some evidence that gender modifies the effect of some common risk factors for at least some cardiovascular diseases, especially smoking(2), and diabetes.(31) It is hoped that the detailed analysis on gender in this thesis will contribute to the debate about gender differences in the possible physiology in cardiovascular disease in women and men. (32) Second, the lack of gender-specific reporting hinders meta-analyses of the impact of gender on treatments and outcomes;(33) a review of gender-based analyses in Cochrane systematic reviews pertaining to CHD found limited studies reporting gender-based results. Twenty per cent of those that did found significant gender differences, suggesting gender differences in treatments and outcomes are uncommon but not unknown.(34) Different areas of CVD research are better than others in reporting gender-specific effects; studies of CHD more commonly report gender-based results than do studies in stroke research.(35) I have therefore considered it important to make a contribution to the better understanding of gender differences to report on gender-based results, even in chapters where the principle focus is on other risk factors.

3.2. Smoking

The second set of substantive analyses investigates the association of smoking with specific presentations which signal the onset of symptomatic CVD. Smoking was chosen because it is one of the leading risk factor for the burden of disease in North America and Western Europe,(36) and remains a common modifiable risk factor in the United

Kingdom, with 23% of men and 19% of women estimated to be current smokers in 2011.(37) There is some evidence that smoking may pose a greater risk for coronary heart disease than for stroke.(38,39) There is also evidence which suggests gender may modify the effect of smoking on CVD.(2)

3.3. Blood pressure

Blood pressure was chosen for investigation because raised blood pressure is one of the larger modifiable risk factors for the burden of disease, particularly cardiovascular disease(40,41) and remains a common modifiable risk factor in United Kingdom, with 31% of men and 28% of women estimated to have hypertension.(37) However, in contrast to smoking, raised blood pressure appears to pose broadly similar risks for mortality from most cardiovascular diseases.(42) There is conflicting evidence of effect modification by gender of the association of raised blood pressure with CVDs, at least for coronary heart disease mortality.(42,43)

4. Linked electronic health records

Linked electronic health records (EHRs) provide a unique opportunity to investigate the questions posed in this thesis because such resources provide the large number of patients required to study multiple disease presentations combined with the clinical detail and event capture needed to specify risk factors and endpoints. The CArdiovascular disease research using LInked Bespoke studies and Electronic health Records (CALIBER) research platform has EHRs for over two million adult patients. CALIBER also links four data sources, including the acute coronary syndrome registry, Myocardial Ischaemia National Audit Project (MINAP), with the clinical details necessary to distinguish specific types of myocardial infarctions and unstable angina. Through incorporation of the primary care record, CALIBER can be used to investigate onset of disease by allowing exclusion of patients with history of disease recorded prospectively (rather than based on recollection). CALIBER, which encompasses multiple sources, including mortality, is also more likely to capture all events, including those which occur out of hospital, then those which rely solely on primary and secondary care records.

5. Aims and objectives

The aim of this thesis is to investigate whether there is heterogeneity of association between specific cardiovascular risk factors and the onset of clinical cardiovascular disease across a wide range of initial disease presentations and whether any such heterogeneity is modified by gender. There are four objectives to address this aim:

1. Investigate whether there is heterogeneity in the association of gender (a non-modifiable risk factor) with the onset of cardiovascular disease across twelve different clinical disease presentations.

2. Investigate whether there is heterogeneity in the association of smoking (a modifiable risk factor - health behaviour) with the onset of cardiovascular disease across twelve different clinical disease presentations.
3. Investigate whether there is heterogeneity in the association of systolic and diastolic blood pressure (a modifiable risk factor – physiological) with the onset of cardiovascular disease across twelve different clinical disease presentations.
4. Investigate whether the associations of smoking and blood pressure with the range of endpoints are modified by gender.

6. Layout of thesis

I begin my thesis by describing electronic health record research in general and the specific research platform used for my research, CALIBER, in detail (Chapter 2). I then describe my methods generally applicable to all of the substantive analyses in my thesis, including definition of risk factors and endpoints, as well as approaches to statistical analysis (Chapter 3). I then describe the derivation of the overall cohort and cohort patients (Chapter 4). In the next three chapters (5-7), I report on the specific methods and findings of my substantive analyses on the association of gender, smoking and blood pressure with initial presentations of cardiovascular disease. Finally, I summarise the overall findings and describe the implications for research, clinical and public health practice (Chapter 8). The appendices include acronyms, codes lists, and variable definitions used, as well as other supporting material.

Source EHR datasets, linkage, creating research data, data quality and governance: The CALIBER research platform

1. Introduction

This chapter describes the elements which were combined to create the CArdiovascular disease research using LIinked Bespoke studies and Electronic health Records (CALIBER) research platform through which the thesis cohort data was generated, as well as the research governance for the thesis.

The use of electronic health records (EHR) in epidemiological research is a growing field, with considerable interest from academic funders, policy leaders and government in the research potential of these resources, in the United Kingdom(44–48) and internationally.(49–52) The United Kingdom has the potential to lead the world in EHR research because of the internationally unique congruence of publically-funded universal healthcare covering primary, secondary and tertiary care; long-established computerisation of general practice; unique personal identifiers in health care; and a range of clinical registries providing depth of clinical detail.

The science of using electronic health records for research, however, is still in its infancy, with on-going debates about different aspects of the methodology. In the last ten years, much of the methodological debate has been about appropriate research and information governance required to protect patient identity/privacy with linked records,(53–56) lack of individual patient consent to use of data,(57–59) and methods of linking data.(60–65) While the debate on these issues continues,(66) it is also now broadening out into how best to make EHR research more transparent, accurate and reproducible,(67–69) and how best to use EHRs for translational research and randomised controlled trials.(70–72) This has occurred at the same time as an explosion in the volume of linked EHR data available and studies using EHR.(73) The fact that STROBE(74) and other reporting guidelines(75) are currently being developed to incorporate issues raised by EHR research is a further indication of the on-going discussion around methodological quality of such research. In this chapter I seek to describe the linked EHR data I use for my research in the context of some of the issues raised by these papers.

First, I describe each of the constituent data sources which have been combined to create the CALIBER research platform, giving a brief description of the context in which the data

originate the type of information recorded, the coding system used, and any known information biases in the dataset. Second, I describe how the linkage was done, the quality of linkage between the datasets – to the extent that it is known - and any information bias that may been introduced in the process of linkage. Third, I summarise the available evidence of the validity of EHR records in general for identifying specific risk factors and endpoints relevant to this thesis. Fourth, I describe the work I undertook, with colleagues, to make the process of creating research-ready data from these raw EHRs more transparent and reproducible. Fifth and finally, I describe the research and information governance used for this study.

2. Data Sources

CALIBER consists of 4 constituent data sources, each of which is described below.

2.1. General Practice Research Database (GPRD)

The GPRD is a primary care database containing anonymised patient records for approximately eight per cent of the UK population (5.2 million patients), owned by the Department of Health from 1994 and curated by the Medicines and Healthcare Regulatory Authority (MHRA).(76) Individual general practices which use the Vision computer system to record data on diagnoses, symptoms, referrals, test results, hospital admissions, prescriptions, procedures, treatments and health behaviours are paid to submit data to GPRD. Patients registered with these practice are informed about the data extraction and offered the opportunity to opt out of having their personal data extracted. GPRD have not published data on how many patients exercised this right. Forty per cent of the 636 general practices participating in GPRD permit additional data to be extracted to allow linkage of individual patient records with other data sources.(77) Data from these practices (n=225), all located in England, are used in the current study. In April 2012, GPRD was incorporated by the Clinical Practice Research Database (CPRD) which will expand data collected to other general practice computer systems and increase the size and coverage of primary care data substantially.

Vision and the other UK general practice computer systems use Read codes, a hierarchical clinical coding system developed for general practice in the United Kingdom (UK). (78) Named after Dr James Read, who developed the codes in the 1980s to allow general practitioners to capture reliably and consistently the range of information required in general practice, such as symptoms and examinations, not just definitive diagnoses.(79) With over 100,000 codes in the Read classification, compared to approximately 20,000 in ICD-10, there

are the potential for multiple codes for individual diseases and endpoints. The hierarchical nature of the coding system means the more specific terms can be related to their more general 'parents'. The Read Codes have now been cross-referenced to the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT), an international hierarchical coding system for use with electronic health records.

GPRD has two key methods of ensuring quality of the data it makes available to the research community. First, they include only patients whose data is of "acceptable research quality", excluding those where gaps or inconsistencies in the patient's record raises concern. The criteria GPRD uses to identify patients of research quality are listed in Table 1. The second method is a data quality audit of all patient records in a practice: recording is assessed against ten practice-based standards, again listed in Table 1.(80) GPRD did not publish the levels or proportion required of a practice to meet their standards. If a practice meets these standards, they are designated "up to standard" (UTS). The first practice designated UTS in CALIBER achieved this status on 19th April 1988. Eighty per cent of the practices in CALIBER achieved UTS status by 1st January 2000, with the other practices achieving UTS in subsequent years. None of the practices in CALIBER have had their UTS status revoked, although this is theoretically possible.

Marked changes in the quality and completeness of data recording in general practices during the period of this study cohort came with radical re-working of the NHS contract with general practitioners. The new GP contract, which came into force in April 2004, introduced the Quality and Outcomes Framework (QoF), which stipulates specific information which must be recorded and levels of preventive treatment to be attained in order to attract payment.(81) Changes to the standards in the QoF have occurred at regular intervals since then.(82) Studies have investigated the effect of the QoF on patient care,(83) on recording of information,(84) and health inequalities, including one finding that recording of CHD quality indicators increased more for men than for women, after the introduction of the QoF.(85)

Table 1: Data quality standards used by GPRD

Patient-level quality standards: Acceptable research quality (ARQ)	Practice-level quality standards: Up to standard (UTS)
An empty or invalid 1st registration date[1]	Percentage of patients of ARQ
Absence of a record for a year of birth	Monthly prescription rate comparable to other practices
A first registration date prior to birth year	Percentage of prescriptions with a medical indication
A transferred out[2] reason with no transferred out date	Death rates comparable to other practices
A transferred out date with no transferred out reason	Cause of death recorded
A transferred out date prior to their first registration date	Outcome of pregnancy recorded
A transferred out date prior to their current registration date	Referral rate comparable to other practices
A current registration date prior to their first registration date	Percentage of referrals with recorded clinical speciality
A current registration date prior to their birth year	
A gender other than Female/Male/Indeterminate	
An age > 115 at end of follow up	
Recorded health care episodes in years prior to birth year	
Registration status of temporary patients	

Patients included in dataset: The population registered with general practice is not synonymous with the general resident population, because not all residents register with a GP. Specific groups are more likely not to register, i.e. asylum seekers, members of the armed forces and their families, convicted prisoners, the homeless and sex workers(86,87) but the number of registered people is not routinely assessed. One recent government report suggests that the number of unregistered people is approximately 1% of the total population.(87) There may be some under-representation of men in a cohort based on GP registration, given that men will constitute a greater part of the unregistered population, but the size of this is small compared to size of the registered population.

[1] Registration here means registration date means date when the first registered with the general practice.

[2] Transferred out refers to when patient leave practice.

GPRD extracts data from specific general practices using the Vision software, also raising the possibility of selection bias if Vision practices are systematically different from practices which use other GP computer systems. GPRD report that the population in their dataset is broadly representative of the population of England and Wales in terms of age and gender profile,(76) though GPRD tends to under-represent certain regions, particularly the north of England. The subset of linkable practices shares the demographic profile of the full GPRD dataset.(77)

Gender differences in consultation rates in primary care: One concern about using primary care EHRs to study gender differences is the differing relationship between women and men with general practice and healthcare in general. Women have been found to consult their general practitioner more overall than men do;(88) however, a substantial portion of this difference can be explained by more consultations by women during the reproductive years,(89) possibly leading to greater completeness in health behaviours and risk factor recording but not necessarily endpoints. Men with symptoms of chest pain recorded on the Rose Angina scale have been found, in one study, to be more likely consult for their chest pain and to be diagnosed with coronary disease than were women.(90) However, other studies have found that when consultations for specific diseases are compared between men and women, there are minimal differences in consultation rates between the two genders, including for established cardiovascular disease.(91,92)

Quality of prescribed medication recording: Prescribing data from GPRD is highly complete and has been used extensively for drug safety research.(93) GPRD only records prescriptions issued and does not measure medications taken, although the requesting of repeat prescriptions does suggest that medication is being taken. A recent study comparing smoking cessation medication in The Health Improvement Network (THIN) (which takes data from the same GP computer system as GPRD) found high concordance between prescribing and dispensing data.(94) Mabotuwana found, albeit in a New Zealand population, a positive predictive value (PPV) of 81.4% (95%CI 78-85) for prescriptions for chronic conditions to be dispensed with seven days, though again this does not necessarily measure medications taken.(95)

Quality of diagnostic codes in GPRD in general: A recent systematic review of studies validating diagnoses in GPRD found those comparing the records in GPRD to non-EHR sources (principally requests to GPs for additional information and/or paper medical

records) had a median positive predictive value (PPV) of 88% across a range of diagnoses.(96) All studies which assessed cardiovascular diagnoses had a median PPV of 85.30 (range: 48-100), while those validating endocrine, nutritional and metabolic diagnoses had a median PPV of 87.70 (range: 53-100). A second systematic review also found considerable variation in the positive predictive value (PPV) of diagnoses; acute conditions were in general less well recorded, although myocardial infarction appeared to be an exception to this finding.(97) However, as both systematic reviews point out, these studies do not identify the negative predictive value of a diagnosis in GPRD, so we do not know the validity of a cardiovascular diagnosis being absent from the GPRD database. More detail on studies of specific cardiovascular disease endpoints is given in Section 4: Accuracy of EHR data.

Quality of risk factor recording: I identified a small number of studies, but no systematic review, investigating the completeness and accuracy of recording of health behaviours, such as smoking, and biomarkers, such as blood pressure or lipid measurement, further described in Section 4: Accuracy of EHR data. Although recent studies are lacking, such studies as exist seem to indicate that risk factors are under-recorded in GPRD.

Conclusions: Although GPRD covers a small proportion of the UK population, the dataset has been found to be representative of that population. Women appear to consult more frequently than men, but this difference may be limited to younger ages. There is some evidence that for the same diagnosis, the rate of consultations is similar between men and women, although one study found that men are more likely to consult for symptoms of coronary disease, i.e. chest pain, which may lead to more complete or earlier diagnosis in men than in women. Diagnostic codes used in GPRD have been found, across the risk factors and cardiovascular disease endpoints relevant to this thesis, to have PPVs over 85%. The recording of risk factors and health behaviours may under-estimate the level of such risk factors in the population.

2.2. Myocardial Ischaemia National Audit Project (MINAP)

As far as I am aware there are only two countries (UK and Sweden) in the world with a national, on-going registry of acute coronary syndrome in which participation is mandated in all hospitals. MINAP is the national registry of patients attending hospitals in England and Wales with suspected acute coronary syndromes (ACS), curated by the National Institute for Cardiovascular Outcomes Research, hosted by University College London. Details of MINAP

have been published elsewhere,(98) but are repeated briefly here. Information on the timing of symptom onset and admission, clinical features and investigations (including ECG results and biomarkers of myocardial necrosis), past medical history, risk factors such as smoking, hospital treatment and discharge diagnosis are collected prospectively at participating hospitals, and submitted to the Central Cardiac Audit Database (CCAD). These records are regularly linked by CCAD to mortality data from the Office of National Statistics and made available in anonymised form to researchers as well as fed back to individual hospitals for quality purposes. MINAP was started in 2000, with a limited number of hospitals submitting data. The Registry achieved near national coverage in 2003, reaching 242 hospitals in 2009, 100% of NHS hospitals in England and Wales. The MINAP data used in the CALIBER research platform runs from January 2003 to August 2009, when the MINAP data was exported for linkage to the trusted third party. Plans for updating the linkage are being considered.

Patients included in dataset: When compared to acute myocardial infarctions (AMI) recorded in hospital episode statistics (HES), ST elevation myocardial infarctions (STEMI) recorded in MINAP have been found in one study to be near complete, but only about half of non ST elevation myocardial infarctions (nSTEMI) were recorded.(98) A consequence of this under-recording is that the MIs of women may be less likely to be recorded, as they are more likely to suffer from an nSTEMI. In a comparison of MINAP with the Patient Episode Database for Wales (PEDW), Lyons and colleagues found that those whose MIs were not recorded were more likely to be older and female.(99,100) Furthermore, they found that the more nSTEMIs a hospital reported, the greater the agreement between MINAP and PEDW. In contrast, a more recent comparison of MINAP with English HES found similar number of women recorded in MINAP as recorded in HES.(101) It is unclear whether using MINAP as a sole source of information on MI is likely to bias estimates of any gender differences.

Quality of data included in dataset: As well as CCAD monitoring 20 fields continually for completeness as the data are submitted, MINAP also conducts an annual data quality exercise, by asking participating hospitals to submit data on 20 fields on a random sample of 20 patients, from the original medical records. The median hospital score from the 2010 audit was 94.8% (IQR 90.0-97.8). I could find no peer-reviewed publication on the agreement between MINAP data and clinical records, whether on the recording of acute coronary syndrome or other risk factor information recorded in MINAP.

Conclusions: MINAP is a useful source of information on MI, particularly as it provides information on MI phenotype, unlike HES. MINAP should be used in conjunction with other sources of information on MI, as it is likely to under-estimate the number of MIs, particularly in women and older people. Although information on risk factors, such as smoking and diabetes are available in MINAP, the recording is, of necessity, retrospective. I therefore made the decision to take risk factor information only from GPRD.

2.3. Hospital Episode Statistics (HES)

HES, owned by the NHS and provided by the NHS Information Centre, consists of data on admissions since the financial year 1989/90, outpatients since 2003/4 and accident and emergency attendance since 2007/8. The hospital admission data includes private patients treated on NHS premises, NHS patients treated by independent providers on behalf of the NHS and patients not resident in England, as well as NHS patients treated in NHS premises. The record for each episode includes up to 20 diagnoses, coded using the International Classification of Disease, Version 10 (ICD-10),(102) and up to 24 procedures, coded using the Office of Population Censuses and Surveys (OPCS) Version 4 codes.(103) HES data available in the CALIBER research platform runs from 1st April 1997 to 31st October 2009.

The funding environment for hospitals changed in 2004 when the NHS switched from block contracts to payment by results (PbR), in which hospitals were paid a fee for each admission, grouped into broad reimbursement bands.(104) One of the impacts of this change in funding environment has been a substantial increase in the number of diagnoses recorded for each admission.

Patients included in dataset: HES is put through a series of data quality checks, principally testing consistency, duplication and completeness.(105) Since 2007/8, HES has published an annual data quality report, which highlight known problems with the data for a given year; of particular relevance is the section on coverage which highlights where individual hospitals have a shortfall between the number of records submitted and the number of admissions reported. The scale of these shortfalls (35,120 in 2009/10)(106) is dwarfed by the overall number of admissions records in HES (14,537,712 admissions in 2009/10),(107) and tends to affect a limited number of trusts in any given year. Thus in any given year the completeness of HES in recording admissions to NHS hospitals is very high.

Admissions to or procedures done privately by independent providers are not included in HES and therefore not included in this study. Relatively little research has been published on the number of patients who might be missed if data from independent providers is not included.(108,109) One report found an increase of 12% in the number of revascularisation procedures in London in from April 2001 to December 2003 if such private procedures in independent facilities are included in the total number of procedures.(109)

Quality of HES data (general): A recent systematic review assessed the quality of discharge coding in HES.(110) Over all the studies, the median diagnostic accuracy was 80.3% (IQR: 63.3–94.1%) ,while the accuracy of procedure coding was 84.2% (IQR: 68.7–88.7%). In those studies conducted since the introduction of Payment by Results, the accuracy of primary diagnosis alone improved to 96.0% (IQR: 89.3–96.2%). The Audit Commission has also conducted an annual audit of discharge diagnoses since PbR was brought in. The 2010 national audit on all hospitals found an average error rate of 11.3% (range 1 to 30%).(111) In the 2011 audit in 60 poorly performing hospitals, they found average error rate of 9%.(112) Details of the accuracy of EHR records for specific endpoints are addressed in Section 4: Accuracy of EHR data.

Conclusions: HES discharge data has nearly complete coverage of admissions and procedures performed in NHS hospitals, though unmeasured private admissions and procedures may be as high as 10% of the overall number for some procedures, particularly in London. The accuracy is also good, and has improved since the introduction of Payment by Results. No evidence of any differences between men and women in accuracy of recording has been reported.

2.4. Mortality data from the Office for National Statistics (ONS)

Data on vital status and cause of death originate from ONS mortality records. By law, any death occurring within England and Wales must be registered with the Registrar of Births and Deaths, normally within 5 days of the death. Such registration must be accompanied by a medical certificate of the cause of death, except in the limited number of cases where death is referred to a coroner. The mortality data for England and Wales is therefore highly complete.(113) UK mortality data do not, however, include people either normally or formerly resident in the UK who die abroad.

Of importance to this thesis is the accuracy of the underlying cause of death on death certificates. I used the example of coronary heart disease (CHD) to research what is known about the accuracy of death certificates. I searched Medline between January 1990 to March 2010 for articles on the accuracy of coding on death certificates for any type of CHD, excluding any articles which assessed deaths before 1985 only. The results are summarised in Appendix B. They fall broadly into four types, assessing:

- coding on death certificates vs autopsy results;
- coding on death certificates vs the results of adjudication of a panel, based on a multiple sources of information;
- agreement between doctors on cause of death; and
- coding on death certificates to diagnoses from contemporaneous hospitalisations.

Only four studies on a UK population could be found, (114–117) three of which were by Goldacre and colleagues who assessed the relationship between hospital discharge diagnosis and cause of death, the third type of death certificate studies. None used ICD-10. The first, in 1993, looked at the cause of death for all hospitalised patients who died with four months of their admission. Of those hospitalised for IHD, 77.8% (8271) had IHD coded as the underlying cause of death, and 90.2% (9596) had circulatory disease mentioned. Overall 85.9% (18,378) of people with admission for circulatory disease had circulatory disease listed as the underlying cause of death. The effect of age on accuracy was assessed and reported as minimal, although no results for IHD are given; gender was not mentioned.(114) The second study analysed any changes in the proportion of death certificates with MI mentioned as a cause of death compared to MI specified as the underlying cause of death between 1979 and 1998. Although the mortality rate for MI fell significantly over the 20 year period, the ratio of underlying to mentioned did not change materially during that time, suggesting that this ratio is a relatively stable measure. This ratio was similar across age groups and between men and women.(115) The third study looked again at hospital discharge data linked to mortality data from 1979 to 1998 in people aged 35-74 following admissions for MI (and stroke). Amongst those who died within 30 days, 85.2% had MI on their death certificate, while 90.2% had CHD (410-414) as an underlying cause of death. As the time period between admission and death lengthened, the proportion dying of the same cause reduced; however, there was little difference in proportion if the death took place in or out of hospital.(116) These three studies, in broad terms, suggest that cause of death coding on UK death certificates is reasonably accurate and that there is little age or gender difference in their accuracy. Mant et al., who looked at overall accuracy of coding of coronary heart

disease deaths,(117) as well as specificity for chronic versus acute CHD, found lower PPV (70%) for codes on death certificates. However, this study included arrhythmias (ICD-9 426-427) and atherosclerosis (440) as well as coronary heart disease (410-414), which may explain the difference with the Goldacre studies.

Most death certificate studies identified looked at the accuracy of coding in ICD-9, no longer used in UK death certification. Only one, by Pajunen and colleagues,(118) looked at coding in ICD-10 using data collected between 1998-2002 for the FINMONICA/FINAMI registries, for both IHD and sudden death (I20-25, I46, R96, R98). They found that for men and women aged 35 to 74, the sensitivity of the death certificate was 92%; for men and women over 75, sensitivity diverges slightly (men 89% to women 90%) but the difference is minimal. Again death certificates appear highly accurate – at least in a Finnish context -- when assessed against a registry of acute coronary syndrome, with minimal gender differences. Of particular importance for this thesis is that FINMONICA/FINAMI is a community registry and therefore includes all of out-of-hospital deaths identified as due to coronary disease.

Of particular concern is the accuracy of death certificate coding when deaths occur out of hospital or suddenly. Goraya et al., using multiple sources of information on cause of death, found that relying on death certificates alone in an American context would like underestimate deaths from IHD that occurred out of hospital by about 5%, with a sensitivity of 91% (85-95) and specificity of 86% (71-95%) for this sub-group of IHD deaths.(119) Several American studies found that death certificates for OOH coronary deaths, when compared to physician-adjudicated or multiple-source cause of death, are relatively poor method to identify sudden cardiac deaths (SCDs) in particular.(119–121) However, in the UK, where a death is sudden and unexpected, deaths are certified by a coroner, usually following a post-mortem. Approximately 40% of deaths from IHD receive a death certificate in this way,(122) while in at least one of the US studies, the post mortem rate was less than 20%, suggesting that the accuracy of death certificates for this sub-group would be higher in the UK than reported in these studies.

Conclusions: The available evidence applicable to the UK suggests that the coding on death certificates for at least coronary heart disease is likely to be accurate, although possibly up to 10% of cases may be missed. However, of greater importance for the current thesis is that there is no evidence to suggest systematic differences between the genders in accuracy of death certificates for these diagnoses.

2.5. Index of multiple deprivation

ONS generated the index of multiple deprivation (IMD) 2007,(123) a combination of scores across seven domains, covering diverse aspects of deprivation such as crime, employment and access to health care, at the level of lower super output area, which covers a population of approximately 1500 people or 400 households. Provided patients have a viable postcode, an IMD score and rank can be attributed to them. IMD score and rank are linked with individual patients, prior to pseudo-anonymisation. IMD has been criticised as a measure of individual deprivation,(124) and it is undoubtedly cruder than individual measures of social class, such as those used in the Whitehall II.(125) However, IMD has been found to be strongly associated with CHD disease and mortality. (126,127) No individual measures of deprivation are currently available within GPRD, so although it would be preferable to have an individual measure of deprivation, IMD is an adequate measure for this PhD.

3. Data Linkage

Unlike the situation for many countries (e.g. US, Australia), the United Kingdom has a unique health identifier for everyone registered with the National Health Service (NHS). The NHS number was introduced with the inception of the NHS in 1948, based originally on the national registration number issued at the start of the Second World War.[3] The computerised NHS Central Register (NHSCR), created in 1991 from local paper records, is responsible for maintaining the system of NHS numbers in England and Wales. A new-style 10-digit NHS number was introduced in 1996, nine digits to uniquely identify individuals with a 10th check digit to confirm the validity of the other nine.(128)NHS numbers are issued to babies born in England and Wales at birth, to anyone who registering with a GP, receiving treatment from a hospital able to issue NHS numbers or on request from the local Primary Care Trust. In the UK, the NHS number is used as the field common to multiple datasets which allows information on individuals in different datasets to be linked. Additional fields such as name or date of birth (patient identifiers) which should remain constant across a range of datasets can also be used to link datasets.

One approach to linking datasets while providing maximal protection for participant privacy, used in creating the CALIBER research platform, is to use a trusted third party (TPP) to make the linkage. Each organisation curating a dataset (the data holders) creates a file containing a dataset-specific pseudo-identifier (identifiers that are specific to patients in each file but

[3] In Scotland, the equivalent number is the Community Health Index, maintained by the Central Register. In Northern Ireland, the equivalent number is the Health and Care Number.

which bear no relation to the real patient identifiable information) for each patient as well as variables with identifiable information which can be used to link data, but no other clinical or other sensitive information. All data holders then send these files the TPP, who then matches the information from all the datasets, either deterministically or probabilistically, and creates a key file with the patient pseudo-identifiers from each dataset. The data holder sends their dataset with clinical information and the patient pseudo-identifier to the researchers. The TTP send the key file, which allows the researcher to link multiple datasets using the linked pseudo-identifiers. Using this approach means that no research organisation or the TPP ever has both patient identifiers and clinical or other sensitive information and no data user has personal information from any dataset other than their own. The process is represented in Figure 1 below.

3.1. CALIBER data linkage process

The CALIBER research platform was created by the linkage of five constituent datasets, which were linked in two stages. For the main linkage to MINAP, Northgate (the TTP used for this linkage) received a list of unique patient identifier fields and demographic characteristics (NHS number, date of birth, sex, and postcode of residence) and their associated source-specific pseudo-identifiers. Records were linked using a pre-defined deterministic linkage algorithm (i.e. records linked only if an exact match is found) and remaining records were linked using a probabilistic method (e.g. records matched with as similar sounding names or common misspellings). Once the linkage was completed, the patient-identifying variables were removed, leaving the key file which links GPRD patient pseudo-identifiers to MINAP hospital record number pseudo-identifiers. MINAP patient pseudo-identifiers are recorded using a one-way cryptographic hashing algorithm and GPRD patient pseudo-identifiers are recorded using an unrelated sequential serial number. Northgate have not provided the CALIBER collaboration with any information on quality of the linkage, i.e. proportion of patients linked either deterministically or probabilistically.

Figure 1: Data linkage using trusted third party to create CALIBER research platform

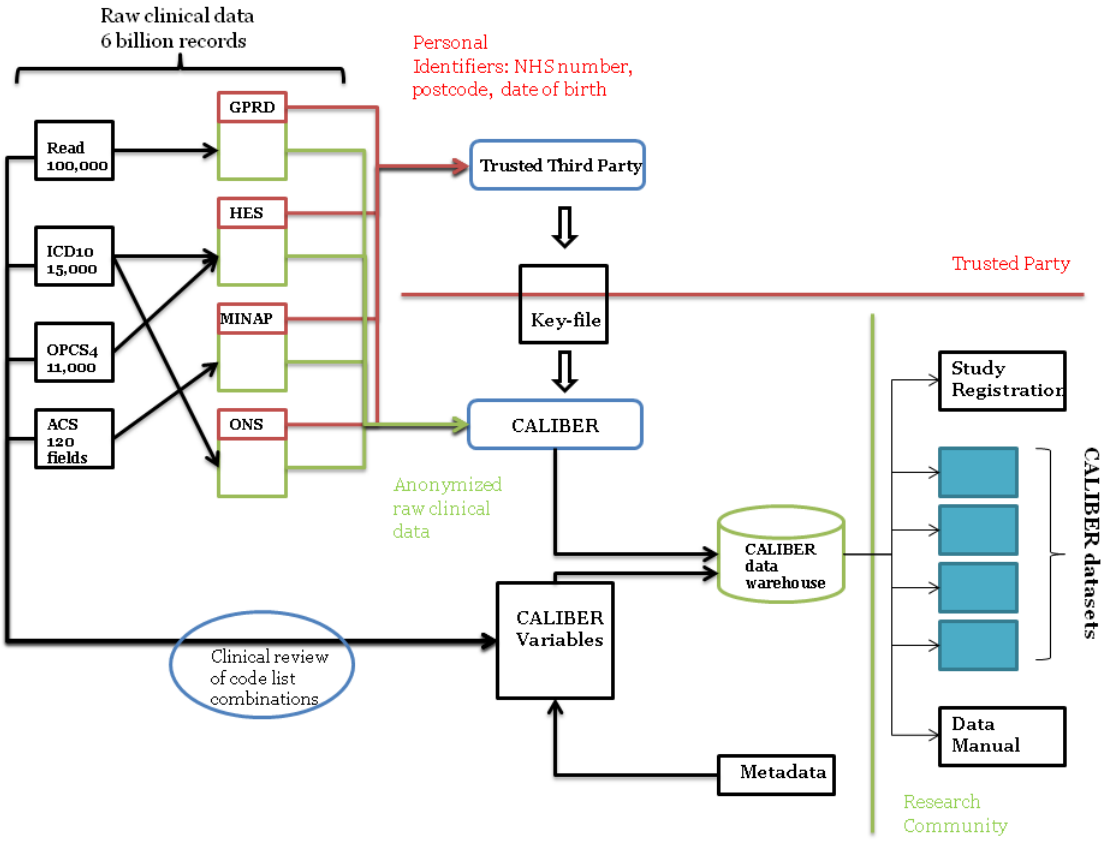


Figure courtesy of Dr Spiros Denaxas, Clinical Epidemiology Group, University College London

Before linkage to MINAP, GPRD was also linked to HES admissions data, mortality data and deprivation data separately, in October 2010, using similar methods to MINAP. GPRD report that for a patient to be put forward for linkage, he or she had to be:

a) Actively registered with a consenting practice at some point during the period for which data are available for the second data source;
b) Been of acceptable research quality in the November 2010 version of GPRD GOLD
c) Had a valid NHS number.(129)

GPRD report that 5.8 million patients were eligible for linkage (but do not report how many were not). They do note that approximately 50% of those without a valid NHS number were not of acceptable research quality so would be excluded for studies using GPRD data in any case, but do not specify the number of patients affected. Of those eligible for linkage, 4.4 million (77%) were matched with HES and 4.2 million were eligible for MINAP linkage.(129) About 99% of the patients eligible for linkage had a valid NHS number and could thus be

linked deterministically.(77) . GPRD do not report a per record linkage statistics which is generally considered the most appropriate measure of linkage performance.

One potential bias with the linked mortality data is that GPRD provides clinical data on a registered population while ONS provides mortality data for resident population. Registered patients who die abroad will therefore not be identified as dead using ONS mortality data. The Foreign Office report that approximately 6,000 British nationals die abroad every year, of whom 55% die of natural causes,(130) which represents approximately 0.5% of the deaths in a given year. No further information could be found on the proportion of these deaths which occur in British residents travelling abroad (who may still be registered with English GP) and those which occur in British nationals resident abroad (who may no longer be registered with an English GP). The number provides the maximal estimation of the extent to which mortality may be underestimated in this cohort for this reason, although the actual number is likely to be much smaller. GPRD provide their own mortality variable derived from registration status, GP own recording of mortality, age and frequency of consultation. This variable can provide another measure to assess the accuracy of ONS mortality status but have not been tested in this thesis.

4. Accuracy of EHR data

Underpinning the work of this thesis is the accuracy of electronic health records (EHRs) in identifying both risk factors and endpoints. In collaboration with other members of the Clinical Epidemiology Group, particularly Marina Daskapolous, I searched for studies on the accuracy of EHRs from primary and secondary care and from death certificates, regardless of country. While sensitivity and specificity are the gold standard measures of accuracy of tests or records because they do not depend on the population prevalence of the condition in a given study, these measures are much less commonly reported than positive predictive value in EHR validation studies. Because our purpose was to undertake a highly sensitive search for studies of EHR accuracy and compare across a single measure of accuracy, we restricted the search to studies which published either positive or negative predictive values. This work benefitted from a number of recent systematic reviews focussing on specific data sources such as GPRD(96,97) and HES(110), as well as reviews focussing on specific endpoints.(131–134) Below, I summarise the findings on the key risk factors and presentations of cardiovascular disease used in this study.

4.1. Risk factors

I identified a small number of studies, but no systematic review, investigating the completeness and accuracy of recording of health behaviours, such as smoking, and biomarkers, such as blood pressure or lipid measurement in GPRD.

One study which compared recorded smoking status in general practice records against both GP questionnaire and levels reported in the General Household Survey found under-recording of current smokers but particularly ex-smokers.(135) However, this study looked at recording prior to the introduction of the Quality and Outcomes Framework in 2004, which, as mentioned above, has been found to increase the level of recording of both smoking status and cholesterol level, although to a lesser extent in women than in men.(85)

Comparing the level of clinically recognised hypertension or dyslipidaemia (defined as recording of clinical diagnosis, biomarkers raised above cut-off or treatment), MacDonald et al. found significant under-recording of these conditions, when compared to the levels reported in the Health Survey for England, particularly in women.(136) Wilkins et al. found evidence in a comparison between GPRD and the Health Survey for England of fewer men identified with hypertension in GPRD than would be expected, in comparison to the Health Survey for England.(88)

Diabetes is the only co-morbid condition used in the analysis for this thesis. A recent systematic review of misclassification and miscoding of diabetes across primary, secondary and registry data concluded that the quantitative measurement of accuracy was poor.(137) The focus of this review, however, was the ability of codes to distinguish between types of diabetes and other related diagnoses such as glucose intolerance. The review did not include any studies in English which reported the positive predictive value (PPV) of EHR codes for diabetes. In two studies of primary care data not included in the review, the PPV of diabetes codes is very high (99-100%).(138–140)

4.2. Presentations of cardiovascular disease

The accuracy of coding of cardiovascular disease endpoints has been assessed in a number of studies across a range of different settings. A total of 25 studies, including 3 systematic reviews, were found which assessed the PPV of codes from either primary or secondary care. None assessed the PPV of linked datasets and none reported the negative predictive value of these codes. A variety of methods were used to estimate the PPV, with the most common

being chart or expert review. Most of the studies used data from the United States, but others were found using data from the UK, Canada, Finland and the Netherlands. Those studies which also published confidence intervals have been included in Figure 2 below. Where the studies specified further characteristics such as fatal or non-fatal presentations or probable or definite diagnoses, this information is also given. A brief summary of the studies by the endpoints used in this thesis has been given below.

Abdominal aortic aneurysm – A 1994 study comparing the Scottish Morbidity Record (the Scottish equivalent of HES) to medical records found a specificity of 98% for AAA but did not report PPV.(141) One recent UK study of coding in HES for abdominal aortic aneurysm (AAA) found a high level of accurate coding (94.9%), although the way in which accuracy was assessed was evaluation of congruence between diagnosis, treatment and method of admission.(142) No studies on the accuracy of coding of AAA from primary care were identified.

Angina pectoris – One US study comparing EHRs and physician-adjudicated cardiovascular endpoints in women only found a low PPV for angina pectoris in hospital EHRs (40%, 95% CI 39-42%)(143), while a European study in men only comparing EHR to disease registry entries found a much higher PPV for unstable angina (78%; 74-83%).(144) A single Scottish study comparing primary care EHRs to paper medical records found a PPV of 100% for angina in primary care data.(138) Another study found the prescription of nitrate had a sensitivity of 73% and specificity of 96% for a GP diagnosis of angina (PPV could not be calculated from the data given.)(145)

Acute myocardial infarction – The PPVs for acute myocardial infarction found in studies of data from secondary care compared to paper medical records have generally been over 85%,(131,144,146) with the exception of a study in women only which found a PPV of 78% (75-80%).(143) A US study which explicitly investigated gender differences in accuracy found lower sensitivity for AMI codes in women compared to men.(147) Compared to the other endpoints, myocardial infarction has a large number of studies assessing the accuracy of coding within GPRD, most of which identified specificity for MI codes of over 90%.(148–153) The PPV in studies which report this measure varied from 85% to 100%.(138,139,148)

Coronary heart disease - A single study from secondary care found a PPV of 91% (88-93%) for the more general codes for coronary or ischaemic heart disease compared to hospital

notes.(144) Studies in primary care found PPVs ranging from of 83% (70-96%) to 100%. (139,140,154)

Heart failure – A systematic review of studies of data from secondary care in North America found the PPV for heart failure codes were generally very high, with most in excess of 90%. Many of the studies in the review included patients seen as outpatients. The one European study of coding accuracy in secondary care identified found a lower PPV of 80%.(144) Studies based on primary care data (GPRD) found PPVs of 82-83%.(155,156)

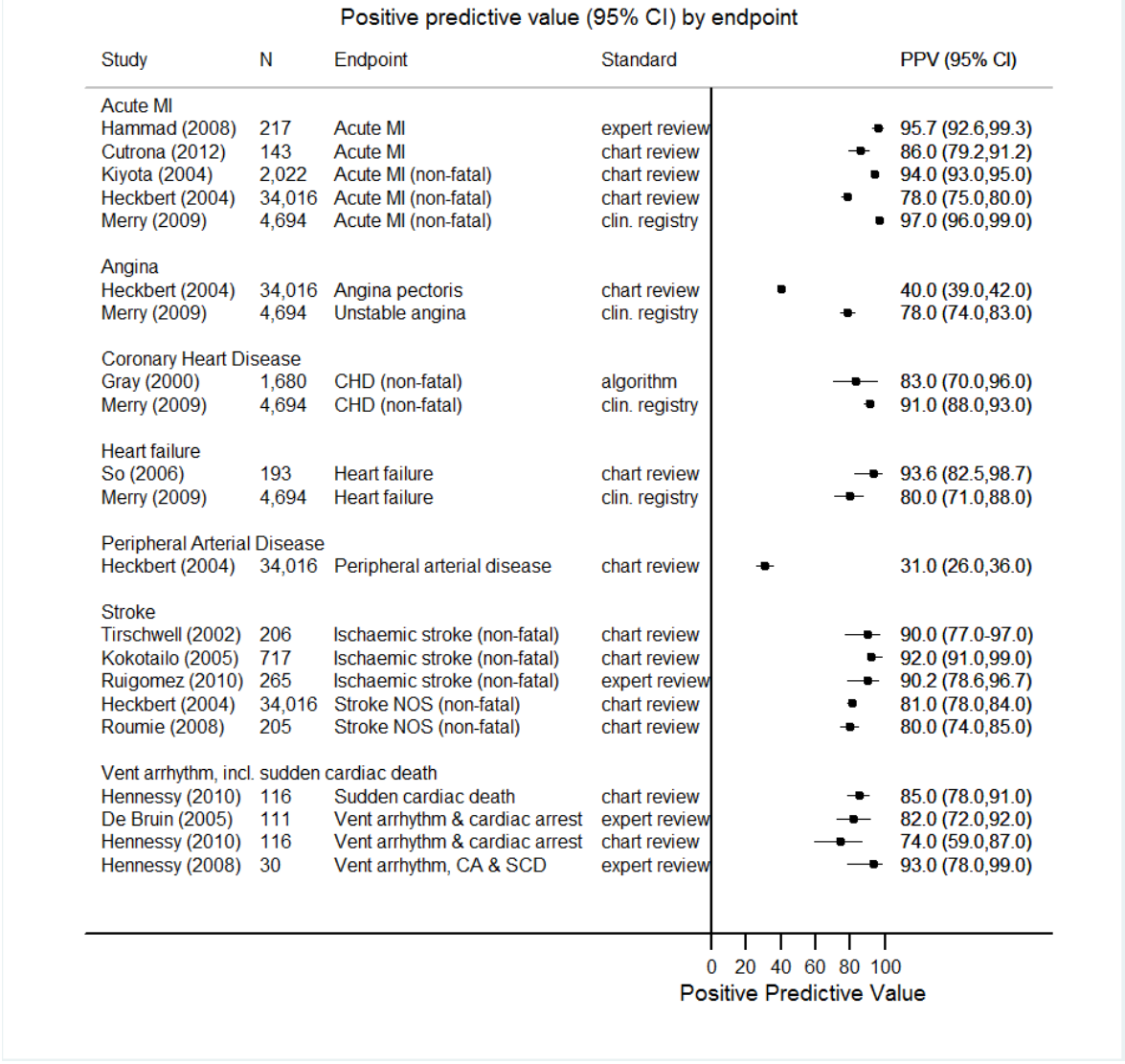
Peripheral arterial disease (PAD) - One study which compared hospital codes and physician adjudication in women only found a low PPV for PAD of 31% (26-36%).(143) One early study of GPRD found a PPV of 75% for PAD.(157)

Stroke, including ischaemic stroke - A systematic review of studies of data principally from secondary care in North America found a wide range of PPV for stroke codes, with the majority of studies finding PPV over 80%; studies of transient ischaemic attack had lower PPV and those of intracranial bleed and subarachnoid haemorrhage, higher.(133) Two older Scottish studies on accuracy of coding in secondary care found PPVs of 75% and 95%.(158,159) One study comparing GPRD stroke codes to paper records found a PPV of 89%,(160) while another study found little agreement between different medical professionals on which codes indicated acute stroke.(67)

Ventricular arrhythmia, cardiac arrest and sudden cardiac death – A systematic review of studies of data from secondary care in North America found PPV between 78-100% for studies using 427.x and 798.x, with the highest PPV (92%) when both were used.(134) Consistent results were found in a European study not included in this review.(161) Two studies of primary care data suggest poor accuracy of ventricular arrhythmias codes in GPRD but did not report confidence intervals.(162,163)

While the PPV varies between endpoints, the majority of studies have found PPV in excess of 80% for all endpoints.

Figure 2: Positive predictive value and 95% confidence intervals for cardiovascular disease presentations recorded in EHR sources



5. From raw data to research-ready data – the CALIBER research platform

A key methodological challenge of epidemiological research using EHRs is to robustly convert raw data to research-ready data. An important aspect of this thesis was not only to create usable data to complete the specific analyses for this PhD thesis, but to do so in such a way that created resources which others could use to make the process of using CALIBER research platform for subsequent research more efficient, transparent and reproducible. This section of the thesis describes the steps to create research-ready data and the CALIBER approach.

In order to understand the process of generating research-ready data from this linked dataset, a basic understanding of the data structures is required. GPRD provides data in several files, which are described in Table 2. All these files are essentially very long lists of unsorted records; for example, the clinical file has 356,446,923 records, at least one and often multiple records for each consultation by every patient registered with all the practices in GPRD over the entire time period of the dataset. The relevant records have to be selected from each relevant file and combined to provide understandable information about individual patients. The same process is used for selecting data from HES and ONS, while MINAP provides data in specified variables. In the CALIBER research platform, each entry specifies the dataset used in creating each variable.

The key stages in creating research-ready data are:

1) Development of code lists used to define exposures, co-variates and endpoints and select relevant data from each data file
2) Definition of such variables, i.e. how the records extracted from data files will be used to specify the value of research variables
3) Algorithms for dealing with duplication and contradiction in the variables created

At each stage of the process, we sought to ensure all our decisions were explicit, documented, reproducible and consistent. Each stage is described in more detail below.

Table 2: Selected constituent data files used for PhD,† including description and number of rows in full CALIBER research platform

Constituent Data files	Description	Records	Used for
GPRD files			
Patient	Year of birth, gender, marital status, registration details and other administrative information	5,372,790	Demographic co-variates
Practice	Region of practice, date practice met data quality standards and last data collection date	226	Modelling data clustered at general practice level
Clinical	Record of each consultation including practice, patient and consultation identifiers, date of data entry, date of consultation, and medcode (GPRD synonym for Read codes)	356,446,923	Specifying cohort, and defining ethnicity, clinical co-variates, and endpoints
Additional	Information, such as measurements (e.g. blood pressure measurement or BMI) or categories (e.g. smoking status – current, ex, non) which are not captured by Read code.	97,244,627	Defining clinical co-variates
Test	Type of test performed, values, units of measure and normal range for laboratory.	227,075,743	Defining clinical co-variates
Therapy	Prescriptions issued included British National Formula (BNF) code, GPRD product code, total quantity prescribed, pack type, number of packs, numeric daily dose, total quantity and number of repeats.	400,859,645	Defining clinical co-variates and specifying endpoints (stable angina)
HES files			
Patient	Patient identifier, birthyear, gender, ethnicity, ethnicity, HES data collection start and end, and matching quality data	2,026,520	Defining demographic co-variate (ethnicity only)
Diagnosis (Admissions)	Patient identifier, admission identifier, date of discharge, ICD10 code, primary diagnosis flag	7,983,022	Specifying cohort, and defining clinical co-variates and endpoints
ONS files			
Mortality	Patient identifier, date of death, under-lying cause of death, cause of death (1-10)	278,088	Specifying cohort and defining endpoints

†GPRD Staff, consultation, referral and immunisation files not used in this PhD.

5.1. Selecting relevant codes to defining risk factors and cardiovascular diseases

In order to select the relevant records from the study population's data, a code list for each variable is needed. At its simplest, a small number of ICD codes are required to select either mortality or hospital diagnosis data; much more extensive lists are required to select records from GPRD. Exchange and discussion about code lists is an important aspect of making EHR research transparent and reproducible. To date, many EHR studies do not publish all the code lists they use, either in the paper or in supplementary material, although this situation is slowly changing. We therefore decided to create auditable STATA do-files used to select code lists for each variable, following the method recommended by Dave and Petersen,(68) as part of our aim to increase transparency and reproducibility in EHR research.

To generate each code list, initial search terms were agreed by at least two clinicians, and the relevant code dictionary searched for matching codes. For ICD-10 codes, OPCS-4 codes and entity codes (used by GPRD to specify the content of subsequent fields in the additional and test files – See Table 2 above), the search was done by hand with reference to any published studies found. For Read codes, STATA do-files were used to search the electronic code dictionary provided by GPRD. Additional codes were identified by hand-searching the NHS Read code browser, asking for suggestions from colleagues who had produced lists for other studies or, in some cases, identifying code lists in published studies or government reports. A preliminary sifting of the identified codes for relevance was completed, and then two clinicians rated the resulting lists for relevance and assigned response categories. Any disagreements were resolved in face-to-face meetings. For the Read code lists, the STATA do-files were amended to include any additions or exclusions, providing an audit trail of the code-selection process and a resource which can be used by other researchers. The STATA do-files were spot checked for accuracy by a researcher who had not written the original do-file. All variable definitions and coding lists used are available through the CALIBER web portal, described in Appendix C.

In creating the variable and coding algorithms, we took an inclusive approach but identified codes which we thought were less definitive as "possible" in our assignment of response categories. We agreed principles on the ways in which we would assign codes to the different response categories, particularly the "history of" and "possible" categories, both to ensure consistency and increase openness about our coding decisions. The principles we used have been given in Appendix C.

The CALIBER variable and coding algorithms were developed through a process of multiple clinical speciality review across two institutions (University College London and the London School of Hygiene and Tropical Medicine). Professor Harry Hemingway (HH) – Cardiovascular Epidemiology, Professor Liam Smeeth (LS) – General Practice, Professor Adam Timmis (AT) – Clinical Cardiology, Dr Anoop Shah (AS) – Clinical Pharmacology, Dr Kate Walters (KW) – General Practice, Dr David Osborne (DO) – Psychiatry, all provided clinical expertise in the development of the CALIBER code lists. Spiros Denaxas (SD) provided technical advice and wrote the programs to extract the relevant records. Emily Herrett (EH) and I wrote the STATA do-files for selecting the Read code lists, further do-files of the mental health portion of the CALIBER manual were developed by Ruzan Udumyan (RU).

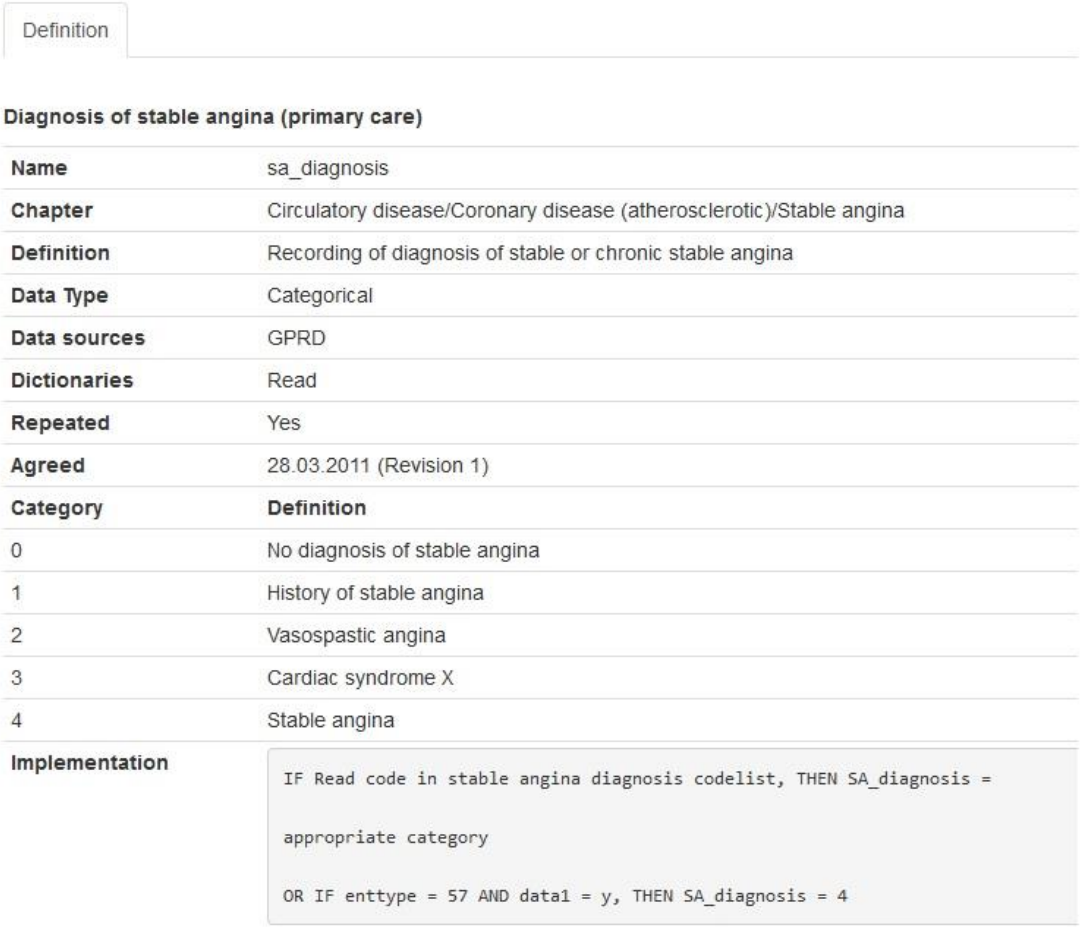
5.2. CALIBER Variable definitions

Once the relevant records can be selected using agreed code lists, the way in which the information in those records will be combined to create research-ready data must be defined. A format for CALIBER variable definitions was agreed which included the following elements:

1) Variable label – plain language name
2) Variable name – variable name used in data files
3) Chapter – Section of CALIBER data portal
4) Variable definition – plain language definition
5) Variable type - continuous, binary, or categorical
6) Data sources – specifies from which constituent CALIBER data source records will be selected
7) Dictionaries – Data dictionary used
8) Agreed: Date on which the variable definition was agreed
9) Units – used for continuous variables only
10) Range – extreme limits were set to exclude biologically implausible values
11) Response categories – for binary and categorical variables only
12) Implementation rule: Logical rule which specifies the way in which the raw data is used to create the variable.

A sample definition for diagnosis of angina in primary care from the CALIBER data portal is shown in Figure 3. Further details on accessing the data portal in 0 Appendix C: CALIBER Data Portal and Developmental Tools.

Figure 3: Diagnosis of stable angina (primary care) from CALIBER data portal



Definition

Diagnosis of stable angina (primary care)

Name	sa_diagnosis
Chapter	Circulatory disease/Coronary disease (atherosclerotic)/Stable angina
Definition	Recording of diagnosis of stable or chronic stable angina
Data Type	Categorical
Data sources	GPRD
Dictionaries	Read
Repeated	Yes
Agreed	28.03.2011 (Revision 1)
Category	**Definition**
0	No diagnosis of stable angina
1	History of stable angina
2	Vasospastic angina
3	Cardiac syndrome X
4	Stable angina
Implementation	`IF Read code in stable angina diagnosis codelist, THEN SA_diagnosis = appropriate category OR IF enttype = 57 AND data1 = y, THEN SA_diagnosis = 4`

The variable definitions were agreed by both clinical and non-clinical researchers. In general we have aimed to give due weight to existing structures in the data and code hierarchies. For example, within GPRD, clinical signs or test results can be recorded either as a Read code (e.g. "O/E blood pressure raised") or as additional data linked to specific entity codes. We therefore created two blood pressure variables, a categorical variable with response categories based on the descriptions in the Read codes and a continuous variable, with the blood pressure data included in the additional file.

As part of the process for defining variables using data from either the GPRD test or additional files (see Table 2 above for description of these files), we checked which codes in the additional file were associated with our Read codes of interest. If the codes appeared to be completely unrelated to the Read code - for example, the Read Code "44PF.00: Total: HDL ratio" was linked to code "265: Sinus x-ray" - we excluded any data linked to those codes. We were surprised by some of these anomalies but as they affected few patients we did not investigate the reasons for this any further. We also excluded records linked to specific entity codes which described laboratory tests for conditions in

which we were not interested, e.g. entity code "363: Lipoprotein electrophoresis" linked to Read codes for low density lipoprotein (LDL). Finally, if there was a contradiction between the Read code used and the name of the linked entity code, we generally did not use any data linked to those records because we had no way of knowing whether the Read code or the entity code was the correct designation, although the two were clearly related. In one exception to this principle, we identified a systematic mismatch between Read Code "44PF.00 - Total: HDL ratio" and entity code "338 - HDL/LDL ratio", which affected >1,900,000 test records. We allowed this mismatched pair in our definition but flagged the problem in the manual. GPRD could not provide an explanation of this apparent systematic mismatch, though informed us that entity codes are allocated by the GP software, not GPRD.[4] For a sample entity code report, see Appendix C. To clarify how some entity codes are used within GPRD, we also ran reports which identified the associated Read codes and number of records. See Appendix C for an example of a reverse entity code report.

Values for observations of clinical signs or test results are only used in the CALIBER variable definitions if they are specific (i.e. equal to a given value) and if the units recorded are those stipulated in our variable definition or were missing. Where they were missing, mean values for those with values and those without were compared in this thesis and those with missing values excluded if they were substantially and significantly different.

For some categorical variables, the response categories from the Read list and the response categories from the associated entity code could theoretically contradict each other. Where there are potential code conflicts between two sources of data, we have identified this in the response categories, so that researchers can determine themselves how they wish to handle these conflicts.

We concentrated on developing definitions for variables which could act as building blocks for more complex variable definitions, which we called base variables (the building blocks) and composite variables (more complex definitions). For example, the specific composite definition of stable angina is:

> Stable angina pectoris identified by diagnosis or prescription of nitrates, Nicorandil, Ivabradine or Ranolazine, and confirmed by test abnormality, or by CABG or PCI without acute myocardial infarction or unstable angina in previous 30 days

[4] Personal communication, A. Gallagher, 26.1.2012.

To implement this definition requires data from eight variables (diagnosis of stable angina, nitrate prescription, atheroma test, ischaemia test, CABG, PCI, MI and UA), six of which are themselves composite variables (e.g. atheroma test composed of invasive angiogram, CT angiogram, MRI angiogram, angiogram modality not specified and angiogram anatomy not otherwise specified combined). While some composite definitions of particular interest to current CALIBER researchers are provided on the CALIBER data portal, we recognized that many projects will wish to develop their own definitions using the base variables as building blocks. As mentioned above, all variable definitions used for this thesis are available on the CALIBER data portal (See Appendix C: CALIBER Data Portal and Developmental Tools for further information).

JG led the development of the first set of CALIBER variable definitions, with draft definitions provided by JG, EH, and RU. The definitions for the initial variables were agreed between HH, LS, EH and JG. SD provided technical advice, checked the association between Read and entity codes, and wrote the implementation rules for the CALIBER research platform, as well as all the computer programs to implement them. AS provided editorial support, additional variable definitions and clinical advice.

5.3. Algorithms for resolving duplicates or conflicts in the data

Steps taken to resolve duplicate entries or conflicts in the data for particular variables were recorded in do-files, which are available on request. However, further work could be undertaken to develop easily understandable algorithms for what steps have been used to resolve inconsistencies in the data.

5.4. The added value of the CALIBER research platform

CALIBER contains over 300 base and composite variables on medical history, diagnosis, investigations, procedures and prescriptions, with the associated code lists created in a transparent and reproducible manner. The variables are recorded in an online data manual, curated by CALIBER staff using established metadata standards(164). Version control is achieved by using a source control repository system. The approach taken is sufficiently flexible that additional variables can be added using the same rigorous approach; this has now been done for base variables required for deriving the Charlton Index for co-morbidities,(165,166) adapted for CALIBER.

6. Information & research governance

Ethics approval was sought by the CALIBER collaboration for linking the constituent datasets to create the CALIBER research platform. In order to use all of the datasets from the research platform, individual researchers have to seek approval from the relevant

authorities to use each constituent dataset. Details of each approval obtained are given below.

6.1. Ethics approval

Ethics approval is not specifically required for the cohort studies undertaken for this PhD because only anonymised data is used in the research. However, ethics approval was obtained for the development of the CALIBER research platform as a whole. The Lewisham Local Research Ethics Committee gave approval for a period of five years for the project entitled "CALIBER dataset", reference number 09/H0810/16, on 8th April 2009.

6.2. Independent Scientific Approval Committee (ISAC) approval

The ISAC of the Medicines and Healthcare products Regulatory Agency (MHRA) requires submission of study protocol as well information on key investigators and sample code lists. Approval of Protocol 10_052R was approved on 8th June 2010. The protocol was revised to include use of HES and ONS datasets, and Protocol 10_052RA was approved on 8th February 2011.

6.3. MINAP Academic Group (MAG) approval

The MAG require for a project protocol and evidence of peer-review, for which they accept ISAC approval. The MAG approval for "Gender differences in the development and prognosis of coronary disease where initial disease manifestation is stable angina, myocardial infarction or unheralded coronary death" was approved on 29.6.2010.

6.4. Study registration

In order to receive CALIBER Scientific Oversight Committee (SOC) approval, all protocols had to be registered with clinicaltrials.gov and the analytic protocol for the study date specified and dated. The clinicaltrials.gov registration for initial presentation of unheralded coronary death (CALIBER-09-05) was obtained on 2.7.210 and revised on 19.12.2011.

6.5. CALIBER SOC approval

CALIBER SOC approval for initial presentation of unheralded coronary death (used for upgrade) was obtained on 8th June 2010 and for initial presentation of multiple endpoints was obtained on 25.11.2011.

6.6. Patient and Public Involvement

On 31.5.2011, I presented the summary of my research proposal to the Clinical Epidemiology Group's Patient and User Involvement Group, which includes carers and patients with coronary disease. The focus of the project on analysing gender differences

across a range of presentations was supported. The conclusions of the research will be fed back to the group in September 2013.

7. Conclusions

In this chapter, I described in detail the constituent datasets of CALIBER, including how the data is recorded, health services factors which may affect the quality and completeness of that EHR data, as well as what evidence is available on the accuracy for my co-variates and endpoints. I further discussed what steps were taken as part of the CALIBER collaboration to make the datasets available in CALIBER more transparent and research using them more reproducible. Further information on the creation of and early studies using the CALIBER collaboration can be found in the protocol paper published in December 2012.(167) Finally, I summarised the information and research governance processes followed, approvals received for this thesis and patient and public involvement in the research project.

Specific methods: Definitions of cohort, risk factors and endpoints and statistical considerations

1. Introduction

In the previous chapter, I described the constituent datasets used in the CALIBER research platform, as well as the methods used by the CALIBER collaboration to develop research-ready data. In this chapter, I have described the study design, population, risk factor and cardiovascular (CVD) endpoint definitions used in my three related studies on the association of gender, smoking and blood pressure with initial presentations of CVD. A key feature of my PhD is the focus on the initial presentation of CVD with any of 12 CVD types (which may have been fatal or non-fatal). Here, I have outlined key issues in modelling these multiple endpoints using competing risks and described the shared statistical methods used for my analyses, as well as treatment of missing data. More details on the statistical methods used for each study are given in the relevant chapters.

2. Study Design

The study design for all three studies (gender, smoking and blood pressure) was a prospective cohort study using patients registered with General Practice Research Database (GPRD) practices and data derived from electronic health records (EHRs). While EHR cohorts are similar to cohorts with bespoke data collection, such as Whitehall II or UK Biobank, it is worthwhile to draw out some differences. Bespoke longitudinal cohort studies record a variety of measurements when the patient enters the study (called baseline measurements), usually completed in the space of a few hours or days. Participants are all enrolled within a relatively short time, such as six months, not least to minimise the cost of employing staff to record the baseline measurements. Participants are then followed up at regular intervals, to update baseline measurements and record events of interest that have happened since the last assessment. In Framingham or Whitehall II, for example, phases of data collection are carried out approximately every two years. With some studies, linkages are made to electronic health records or mortality data, particularly for long term follow-up.(168) EHR cohort studies do not have such control over when or which measurements are recorded, because such measurements are done as part of routine care in general practice or hospital. However, the specification of a baseline time period prior to study entry from which recorded measurements will be used could be considered the equivalent of collection of baseline data at enrolment in bespoke cohorts.

The beginning of the study is 1st January 2001, the date on which the Office for National Statistics (ONS) cause-specific mortality was reliably available in the CALIBER linked

dataset. It would be possible to emulate the brief enrolment period used by bespoke longitudinal cohorts by restricting patients to those who were registered on this date, but one of the strengths of EHR data is that it is possible to include patients as and when they meet the study criteria, as there are no cost implications to doing so. I therefore employed an open or dynamic cohort design, also used in the QRISK EHR cohort studies,(127,169–173) with patients entering the study when they registered with a GPRD practice and met the minimum observation time, age and medical history criteria. The inclusion and exclusion criteria are formally defined below.

3. Cohort population

3.1. Inclusion criteria

GPRD was the only constituent dataset used to define the inclusion criteria for the cohort population, drawn from all patients registered with GPRD practices. I restricted the population to patients who met GPRD's individual data quality standards (called "acceptable research quality", described in greater detail in Chapter 2) and restricted the data I used to define endpoints in GPRD to data collected when the practice met GPRD's practice data quality standards (called "up to standard" (UTS), also described in Chapter 2). I imposed a minimum period of one year of up-to-standard observation prior to study entry to improve the quality of baseline data and increase that likelihood that relevant prior medical history was captured. I included patients aged 30 and above, in line with the Framingham cohort, one of the few existing studies to investigate multiple initial presentation of cardiovascular disease phenotypes.(1) I imposed the high age limit of 100 because of concerns about accuracy of age information at greater ages.[5] The formal inclusion criteria are listed below in the order in which they were implemented by the data manager, over which I had limited control. Patients were included if:

1) They were of acceptable research quality;
2) Their gender was recorded as either male or female;
3) They were registered with the GPRD practice on or after 1st January 2000;
4) They were aged between 30 and 100 the year before study entry;
5) Their current registration period contained at least one year of up-to-standard (UTS) observation time before 1st of January 2001, the beginning of the study.

3.2. Exclusion criteria

A combination of *all* the constituent datasets were used to define the exclusion criteria for the cohort population. Because patients remain registered with a practice until they are

[5] It should be noted that with increases in longevity documented in the latest census, the number of patients over 100 to be expected in an 8% sample of the English population (i.e. patients in GPRD) is not small.

formally removed, it is possible to be registered with a practice and also to be dead. There were a number of patients who met our inclusion criteria but had a date of death prior to study entry (n= 1,678), who were excluded from the study. I also identified a small number of patients (n=222) with two ONS mortality records, presumably indicating a problem with the data linkage for these patients; these patients were similarly excluded. The purpose of these studies was to investigate initial presentation of onset of CVD with different initial presentations, so I excluded any patient with a prior history of CVD in any dataset. I also excluded patients whose first record of a cardiovascular event *after* study entry indicated a historical event, e.g. history of MI or stroke rehabilitation, suggesting I may have missed a prior event. To summarise, patients were excluded from the cohort if:

1) They had an ONS date of death prior to study entry;
2) They had more than one ONS date of death;
3) They had a record in any of CALIBER constituent data sources of symptomatic atherosclerotic disease prior to study entry, i.e. a record of coronary disease, ischaemic cerebrovascular disease, peripheral arterial disease or unspecified atherosclerotic disease. Patients with a record of ventricular arrhythmias, cardiac arrest or heart failure were also excluded from the cohort. Details on the definitions of atherosclerotic and cardiac disease is given below in the description of endpoint variables;
4) They had a code indicating history of disease as the first record after study entry and it was the only code on the day, suggesting that prior history of disease had been missed.

Figure 1 in *Chapter 4 – Description of Cohort Population* shows the derivation of the cohort sample. Table 1 in the same chapter describes the number of people excluded for each specific disease presentation.

3.3. Study entry dates, study exit dates and observation time

With an open cohort, each patient has their own study entry time, although the majority in fact enter the study at the study beginning, i.e. 1st January 2001. Otherwise, patients entered 1 year after a) they registered with an UTS practice, b) their practice achieved UTS status, or c) they reached their 30th birthday. The study end date was 25th March 2010, the last date on which data from any practice was available. Patients were right-censored when they left the practice or their practice provided the last GPRD download, whichever came first.

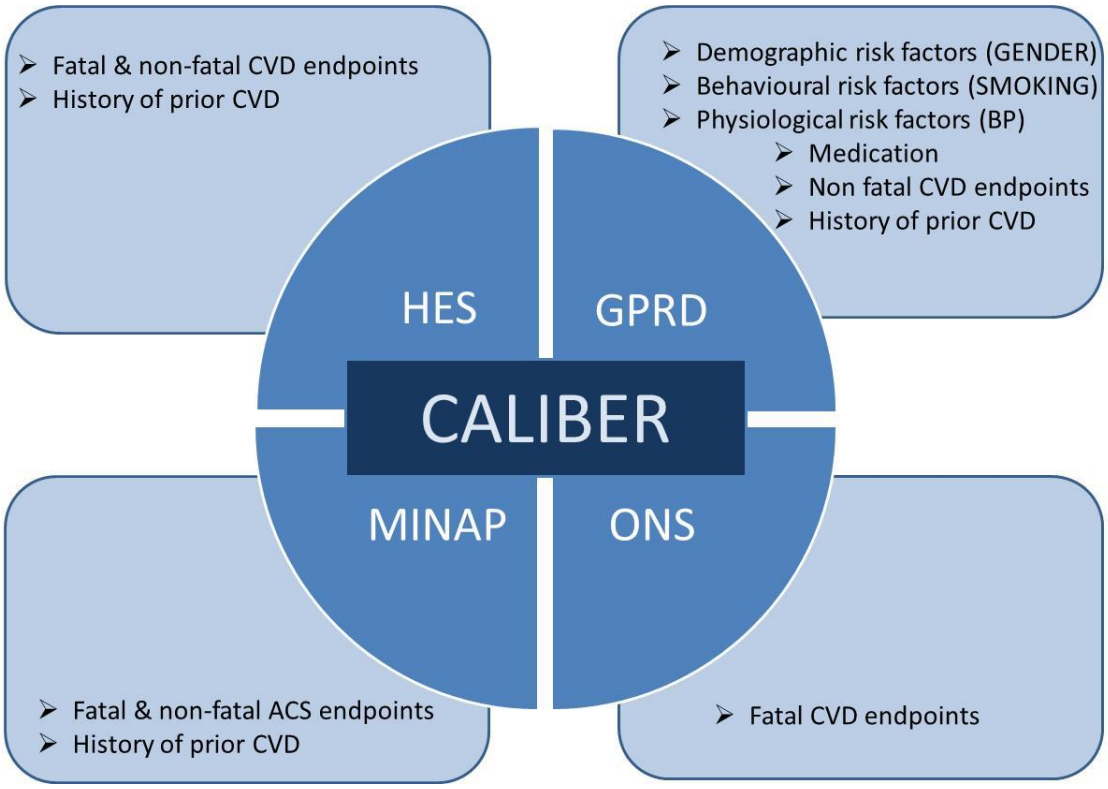
In Section 1of this chapter, I described some of the differences between EHR cohort studies and bespoke longitudinal cohort studies, in particular the range of time during

which data used for defining risk factors and co-variates is taken. As specified in the variable definitions below (Section 56), data used to define some of the co-variates and history of CVD was taken from any point in the participant's record during their current registration period, a portion of which time could be before their practice had achieved UTS status, GPRD's data recording quality standard. The observation time reported for this thesis in Chapter 4 has therefore been divided into overall observation time (from current registration date to cardiovascular endpoint or censoring date), UTS observation time (from current registration or UTS date whichever came later) and time to event (from study entry to cardiovascular endpoint or censoring date). The last equates to the standard observation time reported in cohort studies, but the other two are presented to give some sense of the observation time from which medical history was taken and the observation time during which the practice met data quality standards.

4. Variable Definitions

The variables used to specify risk factors, co-variates and cardiovascular endpoints are defined below. Data from the four constituent datasets was used to create the cohort of patients free from cardiovascular disease and to define the endpoints, while principally GPRD is used to define the risk factors and co-variates. (See Figure 4.)

Figure 4: CALIBER constituent datasets used to define study variables



CVD indicates cardiovascular disease; HES, Hospital Episode Statistics; GPRD, General Practice Research Database; MINAP, Myocardial Ischaemia National Audit Project; ONS, Office for National Statistics.

4.1. Risk factors and co-variates

Each of the risk factors and co-variates are defined below. Age, sex and social deprivation were taken as static values from the GPRD patient file. The other variables have potentially multiple records per patient. For clinical co-variates which change over time, such as smoking status, I searched GPRD for relevant records and used the one recorded on the date closest to the patient's study entry or, in the case of blood pressure, a mean of all values in the two years prior to study entry. (See Figure 6 below for a graphical presentation). Information on which data files were used and how the data was selected are given in variable definitions accessible through the CALIBER web portal (access details provided in Appendix C). Further details on how I arrived at the values for each co-variate are given below.

4.1.1. Definitions

4.1.1.1. Gender

Gender was recorded in the patient file in GPRD. As mentioned above, patients with indeterminate gender were excluded from the study. My findings on the association of gender with the different initial presentations of CVD have been given in Chapter 5.

4.1.1.2. Smoking

Smoking status was recorded in routine general practice, with information obtained by general practitioners, practices nurse, healthcare assistants or administrative staff either in the context of new patient questionnaires or individual consultations. Since 2004, GPs have been incentivised to collect smoking data, under the Quality and Outcomes Framework.(174) This data is provided by GPRD, using a variety of Read codes. For the purposes of this thesis, smoking status was categorised as non-smoker, ex-smoker, or current smoker. Codes which indicated patients were participating in smoking cessation programme were taken to indicate that they were current smokers, rather than ex-smokers. Where a patient's last smoking status prior to study entry was non-smoker, their status was changed to ex-smoker if there was any previous record of being a smoker at any point in the patient's history. See Figure 5 for a graphic representation of this variable definition. My findings on the association of smoking status with the different initial presentations of CVD have been given in Chapter 6.

Figure 5: Definition of smoking status at baseline

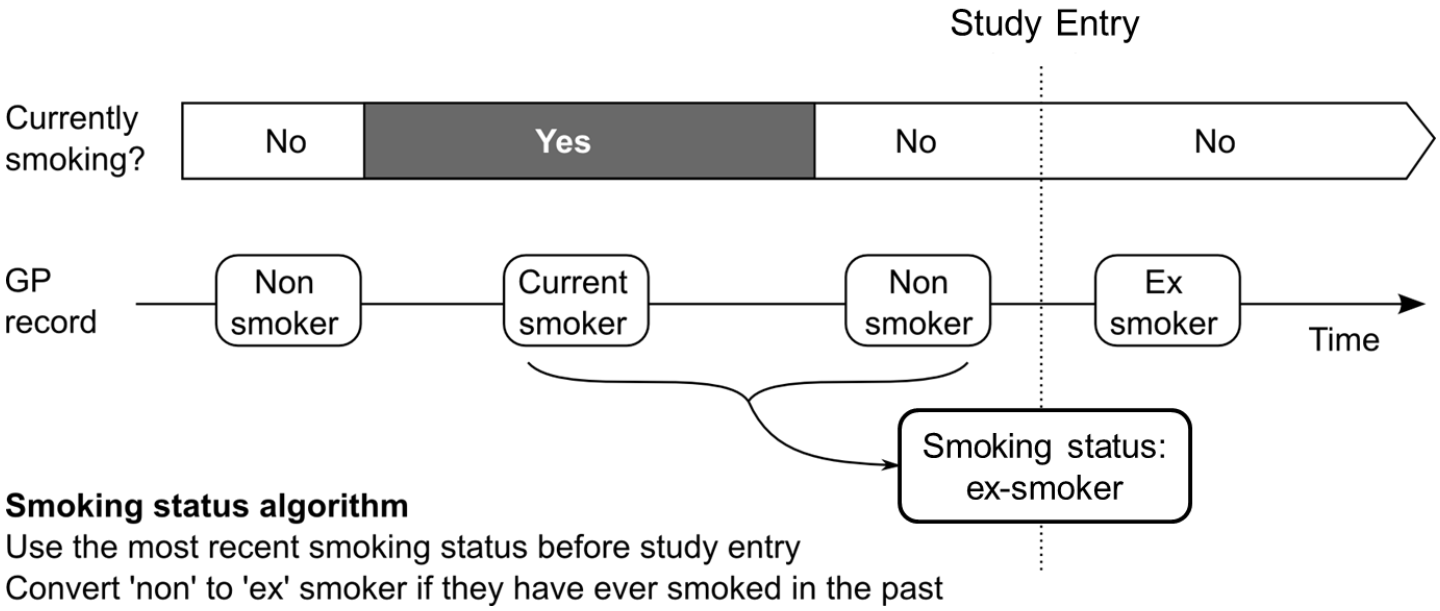


Figure used courtesy of Marina Daskopoulos, Clinical Epidemiology Group, University College London

4.1.1.3. Blood pressure

The blood pressure measurements used in this study were recorded as part of routine care in general practice. The measurements were made by a wide variety of staff, including GPs and practice nurses, using different machines under a variety of conditions, ranging from GP consultations for potentially unrelated illnesses to new patient checks to hypertension clinics. The data was provided by GPRD. No information on the circumstances under which the measurements were taken was available for this study. Systolic blood pressure (SBP) at baseline was defined as the mean SBP over the two years prior to study entry. Diastolic blood pressure (DBP) at baseline was similarly defined. For all analyses which incorporated either measure, SBP and DBP were modelled using change in standard deviation in values. (See Figure 6.) My findings on the association between blood pressure and the initial presentation of CVD have been described in Chapter 7.

4.1.1.4. Age

Age at study entry was derived from the year of birth recorded in GPRD, assuming a date of birth of 1st January of that year. The actual date of birth is not available in pseudo-anonymised data to protect patient identities.

4.1.1.5. Social deprivation

The index of multiple deprivation (IMD) 2007(123) is described in detail in Chapter 2. IMD was used to measure social deprivation. I divided IMD into quintiles with the lowest quintile indicating the greatest deprivation.

4.1.1.6. Ethnic group

Ethnic group is self-reported ethnic group, recorded in either GPRD or Hospital Episode Statistics (HES). Where more than one ethnic group was recorded for a patient, I attempted to resolve any inconsistency within each constituent dataset and then between datasets. Conflicts between records were resolved by taking the lowest level ethnic category that was consistent with the conflicting codes, so, for example if patients were recorded as *White* and *White British*, they were categorised as *White British*. If they were recorded as *White British* and *White European*, they were coded as *White*. Partly because of the code conflicts and partly to simplify analysis, the ethnic groups were then amalgamated into *White, Black, South Asian, Other* and *Unknown* categories.

4.1.1.7. Alcohol consumption

Participant's alcohol consumption was derived from GPRD using data in the clinical and the additional files which indicated:

- Weekly alcohol consumption
- Daily alcohol consumption
- Categorical identification of drinking habits by the GP

Patients were categorised as non-drinker, ex-drinker, occasional drinker, current drinker, and excess drinker, including binge drinker. Where a weekly or daily unit amount was recorded, the patient was allocated to the relevant category based on their gender and the NHS recommended daily/weekly units.

4.1.1.8. Use of anti-hypertension medication

Use of anti-hypertensive medication was derived from at least two successive prescriptions for commonly used agents [beta-blockers, angiotensin-converting-enzyme (ACE) inhibitors, angiotensin receptor blockers (ARBs), thiazides] as well as other less common preparations. The specific medications included in this variable are listed in *Appendix D – Medications*.

4.1.1.9. Lipids

Total cholesterol, high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, and triglycerides in mmol/L were recorded in the GPRD test file. Records with unit in mmol/L or those with none were extracted, but there were significant differences in the distribution of those with units recorded and those with none, so only those with mmol/L specified were used.

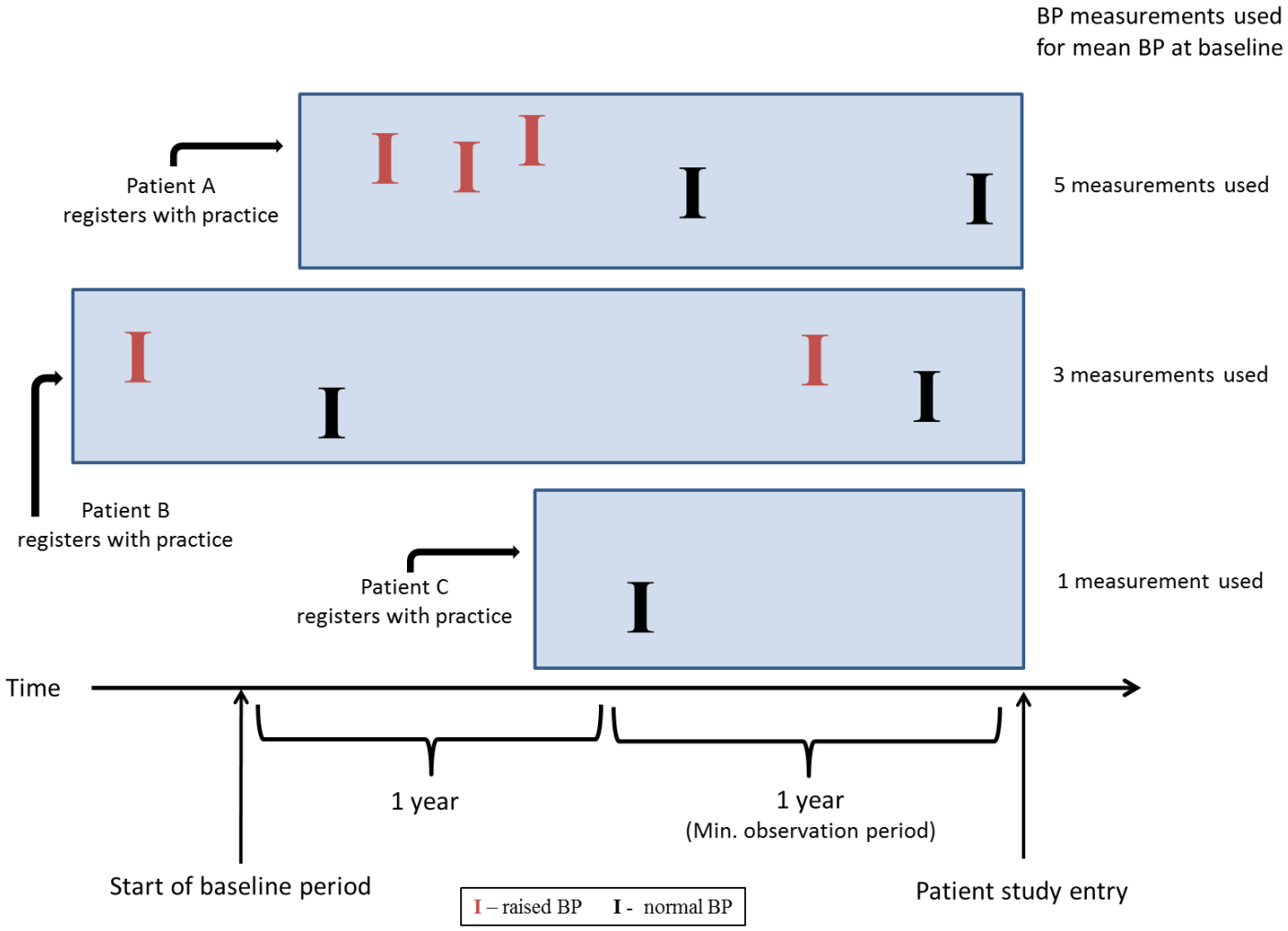
4.1.1.10. Use of statins

Use of statins was derived from a minimum of a single prescription of a statin recorded in GPRD during the two years before study entry. See *Appendix D– Medications* for specific agents.

4.1.1.11. Body mass index

Patients' body mass index (BMI) in kg/m^2 recorded in the GPRD additional file was used for this variable.

Figure 6: Mean blood pressure measurements during baseline period



4.1.1.12. Diabetes mellitus

Patients with diabetes mellitus were identified by any record in GPRD prior to study entry of:

- a diagnosis of Type I, Type II or not further specified;
- monitoring or treatment which specifically identified diabetes;
- complications of diabetes; or
- one or more prescriptions for insulin or oral anti-diabetic medication.

This approach to defining diabetes did not take into account any measurements of blood glucose (although these are available in limited patients in GPRD) and may have under-estimated diabetes. Although the date of onset of diabetes is generally difficult to determine, the approach to analysis used in this thesis did not treat diabetes as a time-varying co-variate, so the timing of onset of diabetes did not pose a problem for these particular set of analyses.

4.1.2. Statistical description of baseline characteristics of cohort

For all these variables, the extent of missingness and the number of available observations for each variable within the dataset is given in *Chapter 4*. For continuous variables, the mean and standard deviation have been presented, while for categorical variables, the number in each response category and proportion have been presented. Where appropriate, comparisons to the Office for National Statistics 2001 Census or Health Survey for England have been made.

4.2. Initial clinical presentations of cardiovascular disease phenotypes

4.2.1. Overall approach

I defined the onset of CVD with specific initial presentations as the recording of a specific diagnosis on first recorded date. This somewhat simplistic approach does not allow for an evolving or developing history of a particular presentation, where a patient may initially be thought to have one presentation (say angina pectoris) which is later clarified to be another presentation (say NSTEMI). However, given that the data used from both HES and the Myocardial Ischaemia National Audit Project (MINAP) are based on discharge rather than admission diagnosis, the extent of this misclassification was likely to be relatively small. A fellow member of the Clinical Epidemiology Group investigated this question, calculating the time elapsed between the first recorded diagnosis and the subsequent one for a sample of patients from CALIBER; she found that the mean number of days between different types of events was over 60 days for all presentations,

suggesting that using the first recorded diagnosis is a robust approach to studying onset of CVD.[6]

Where more than one diagnosis was recorded on one day, I used the most specific and serious presentation. For example, on a given day, a patient admitted with a myocardial infarction might have a record in HES indicating myocardial infarction not otherwise specified, one from GPRD indicating coronary heart disease not otherwise specified and one from MINAP indicating STEMI; the last would be used in the endpoint definition. The ordering of diagnoses, agreed with two clinicians (HH & LS) has been shown in Appendix E.

I used all four constituent data sources (GPRD, MINAP, HES, ONS) to define the following cardiovascular phenotypes. Where codes allowed a distinction to be made between a historical diagnosis and an incident diagnosis, both historical and incident codes were used to exclude patients from the cohort, but only incident codes were used to identify initial presentation of CVD. All presentations were defined to include both fatal and non-fatal presentations, with the exception of stable angina, which does not include fatal presentations, and unheralded coronary death which does not include non-fatal presentations. The definitions have been given below, but are also summarised in Table 3, with the key codes used from each data source. All ICD-10 and Read codes lists used to define variables are available on the CALIBER web portal.

[6] Personal communication, Eleni Rapsomaniki, 21.10.2012

Table 3: Overview of datasets and codes used to define the initial presentation of cardiovascular disease

Endpoint	Primary Care	Disease Registry	Hospital Procedures	Hospital Diagnoses†	Causes of death‡
	GPRD	MINAP	HES	HES	ONS
	Read codes	Registry specific	OPCS-4	ICD-10	ICD-10
Acute Myocardial Infarction	G30X000 Acute ST segment elevation myocardial infarction G307100 Acute non-ST segment elevation myocardial infarction G30..14 Heart attack, G30..15 MI Acute myocardial infarction + 60 other codes as Acute Myocardial Infarction Not Otherwise Specified	MI with or without ST elevation based on initial ECG findings, raised troponins and clinical diagnosis	Not used (there is no code that is specific to primary coronary intervention)	Acute myocardial infarction I21, Current complications of acute myocardial infarction I23	Acute myocardial infarction I21, Current complications of acute myocardial infarction I23
Unstable angina	G311.13/G311100 Unstable angina, G233200 Angina at rest, G311400 Worsening angina + 13 other codes	Discharge diagnosis of unstable angina, no raised ST elevation No raised troponin levels	not used	Unstable or worsening angina I20.0 Acute ischaemic heart disease I24, Coronary thrombosis not resulting in myocardial infarction I24.0, Other forms of ischaemic heart disease I24.8, Acute ischaemic heart disease, unspecified	not used

Endpoint	Primary Care	Disease Registry	Hospital Procedures	Hospital Diagnoses†	Causes of death‡
	GPRD	MINAP	HES	HES	ONS
	Read codes	Registry specific	OPCS-4	ICD-10	ICD-10
				I24.9	
Stable angina	G33..00 Stable Angina, G33z.00 Angina pectoris NOS + 25 other codes for diagnosis of stable angina pectoris 30 codes for evidence of coronary artery disease at angiography (CT,MR, invasive or not specified) 151 Read codes for evidence of myocardial ischaemia (Resting ECG, exercise ECG, stress echo, radioisotope scan) Two or more successive prescriptions for anti-anginals	not used	Coronary Artery Bypass Graft (CABG) K40-K46 or Percutaneous Coronary Intervention (PCI) K49,K50,K75 not within 30 days of an acute coronary syndrome	Stable angina pectoris I20 excluding unstable angina I20.0	not used
Coronary heart disease not otherwise specified	G3…00 Ischaemic Heart Disease + 90 other codes including CHD NOS, chronic ischaemic heart disease, silent myocardial infarction	not used	not used	CHD NOS, chronic ischaemic heart disease, silent myocardial infarction I25 excluding I25.2, old myocardial infarction	not used

Endpoint	Primary Care	Disease Registry	Hospital Procedures	Hospital Diagnoses†	Causes of death‡
	GPRD	MINAP	HES	HES	ONS
	Read codes	Registry specific	OPCS-4	ICD-10	ICD-10
Heart failure *	G58..00 Heart Failure + 92 other Read codes for heart failure diagnosis	not used	not used	I50 Heart failure(including all sub, I11.0 Hypertensive heart disease with (congestive) heart failure, I13.0 Hypertensive heart and renal disease with (congestive) heart failure, I13.2 Hypertensive heart and renal disease with both (congestive) heart failure and renal disease	I50 Heart failure I11.0 Hypertensive heart disease with (congestive) heart failure, I13.0 Hypertensive heart and renal disease with (congestive) heart failure, I13.2 Hypertensive heart and renal disease with both (congestive) heart failure and renal disease
Ventricular arrhythmias, cardiac arrest and sudden cardiac death	G574.00 Ventricular fibrillation and flutter, G757.00 Cardiac arrest + 35 other Read codes for ventricular fibrillation , asystole, cardiac arrest, cardiac resuscitation, electro-mechanical dissociation, G575100 Sudden cardiac death, so described	not used	Implanted cardiac defibrillation device X50, Implantation, revision and renewal of cardiac defibrillator K59	I46 (cardiac arrest) I47.0 (re-entry ventricular arrhythmia) I47.2 (ventricular tachycardia)	I46 (cardiac arrest, includes I46.1 sudden cardiac death) I47.0 (re-entry ventricular arrhythmia) I47.2 (ventricular tachycardia)

Endpoint	Primary Care	Disease Registry	Hospital Procedures	Hospital Diagnoses†	Causes of death‡
	GPRD	MINAP	HES	HES	ONS
	Read codes	Registry specific	OPCS-4	ICD-10	ICD-10
Unheralded coronary death	Any CVD excluded	Any CVD excluded	Any CVD excluded	Any CVD excluded	I20 Angina Pectoris, I21 Acute myocardial infarction, I22 Subsequent myocardial infarction, I23 Certain current complications following acute myocardial infarction, I24 Other acute ischaemic heart diseases and I25 Chronic ischaemic heart disease not preceded by any other CVD presentation
Ischaemic stroke	G64..11 CVA – cerebral artery occlusion, G64..13 Stroke due to cerebral arterial occlusion, G6W..00 Cereb infarct due unspecified occlusion/stenosis of precerebral arteries, G6X..00 Cerebral infarction due to unspecified occlusion or stenosis of cerebral arteries plus 8	not used	not used	I63 Cerebral infarction	I63 Cerebral infarction

Endpoint	Primary Care	Disease Registry	Hospital Procedures	Hospital Diagnoses†	Causes of death‡
	GPRD	MINAP	HES	HES	ONS
	Read codes	Registry specific	OPCS-4	ICD-10	ICD-10
	other codes				
Peripheral arterial disease (PAD)	72 codes for Lower limb peripheral arterial disease diagnosis (including diabetic PAD, gangrene and intermittent claudication Evidence of atherosclerosis of iliac and lower limb arteries based on angiography or Dopplers	not used	L50-L54 Bypass, reconstruction and other open operations on iliac artery L58-L60, L62 Bypass, reconstruction, transluminal operations or other open operations of femoral artery, L65 Revision of reconstruction of artery	I70.2 atherosclerosis of arteries of extremities, I73.9 peripheral vascular disease intermittent claudication Peripheral complications of diabetes including gangrene 0.5 suffix of E10 Insulting dependent diabetes mellitus, E11 Non-insulin-dependent diabetes mellitus, E12 Malnutrition-related diabetes mellitus, E13 Other specified diabetes mellitus, E14 Unspecified diabetes mellitus	I70.2 atherosclerosis of arteries of extremities, I73.9 peripheral vascular disease intermittent claudication, Peripheral complications of diabetes including gangrene 0.5 suffix of E10 Insulting dependent diabetes mellitus, E11 Non-insulin-dependent diabetes mellitus, E12 Malnutrition-related diabetes mellitus, E13 Other specified diabetes mellitus, E14 Unspecified diabetes mellitus

Endpoint	Primary Care	Disease Registry	Hospital Procedures	Hospital Diagnoses†	Causes of death‡
	GPRD	MINAP	HES	HES	ONS
	Read codes	Registry specific	OPCS-4	ICD-10	ICD-10
Abdominal aortic aneurysm (AAA)	G714.00 Abdominal aortic aneurysm without mention of rupture + 11 more codes for AAA diagnosis 13 codes for evidence of AAA on ultrasound or CT scan	not used	L16 Extra anatomic bypass of aorta, L18-L23 Replacement of aneurysmal segment of aorta, bypass of segment of aorta, plastic repair of aorta, L25-L28 Transluminal or endovascular insertion of stent on aneurysmal segment of aorta	I71.3 Abdominal aortic aneurysm, ruptured. 171.4 AAA, without rupture	I71.3 Abdominal aortic aneurysm, ruptured. 171.4 AAA, without rupture

† Primary cause of admission. ‡ Underlying cause of death

4.2.2. Definitions of CVD presentations

4.2.2.1. Stable angina

Stable angina (SA) was defined using three of the four data sources: GPRD, HES and MINAP. Patients with stable angina were identified by a record of:

- a diagnosis of stable angina or angina pectoris not further specified (ICD-10 I20, excluding I20.0);
- a coronary artery bypass graft (CABG) or percutaneous coronary intervention (PCI) not within 30 days of an admission for myocardial infarction or unstable angina;
- two or more subsequent prescriptions for nitrates(145,175,176) or other specific anti-anginal medication (See Appendix D for specific medications); or
- an abnormal result following an exercise ECG, stress echocardiogram, radioisotope scan, or an invasive, computed tomography (CT), magnetic resonance imaging (MRI) or unspecified angiogram.

Patients were also excluded from the cohort if they had a record in MINAP of history of angina.

4.2.2.2. Acute myocardial infarction

Acute myocardial infarction (AMI) includes three subtypes: ST elevation myocardial infarction (STEMI), non-ST elevation myocardial infarction (nSTEMI) and myocardial infarction not otherwise specified (MI NOS). These presentations were defined using three data sources: GPRD, HES & MINAP. Patients with STEMI were identified by a combination of ST-elevation on ECG, raised troponins and clinical findings indicative of MI in MINAP data, as defined by the internationally agreed definition of STEMI.(177) A record of diagnosis of STEMI in GPRD was also included, on the basis that this would reflect information from a discharge letter from an acute admission. Similarly, patients with nSTEMI were identified by no ST-elevation on ECG, raised troponins and clinical findings indicative of MI in MINAP data and diagnosis code alone from GPRD. All other records which did not specify type of myocardial infarction (MI), including AMI codes from HES (ICD-10 I21, I22, & I23), were included in MI not otherwise specified (NOS). Additionally, patients were excluded from the cohort if they had a history of MI in MINAP or a record of old myocardial infarction (I25.2) in HES. Patients with unheralded fatal MI were included in the unheralded coronary death category.

4.2.2.3. Unstable angina

Unstable angina (UA) was defined using 3 data sources: GPRD, HES and MINAP. Patients with unstable angina were identified in MINAP by clinical findings of acute coronary

syndrome but no raised ST elevation on ECG and no raised troponin levels. UA was further identified by diagnosis of unstable angina (ICD10 I20.0) or acute ischaemic heart disease (I24, I 24.0, I24.8 & I24.9) in HES or unstable/worsening angina or acute coronary syndrome in GPRD.

4.2.2.4. Coronary heart disease not otherwise specified (CHD NOS)

Patients with coronary heart disease not otherwise specified (CHD NOS) were those with a variety of non-specific diagnoses such as CHD NOS, chronic ischaemic heart disease and silent myocardial ischaemia (I25 excluding I25.2, old myocardial infarction) in either GPRD or HES.

4.2.2.5. Heart failure

Heart failure (HF) was defined using all four data sources. Patients with heart failure were identified by a record of:

- a diagnosis of heart failure in HES (I50, I26.0, I11.0, I13.0 & I13.2) or GPRD; or
- a result of left ventricular hypertrophy on a resting ECG in GPRD; or
- an underlying cause of death (UCOD) of HF in mortality data as the first presentation of atherosclerotic disease after study entry.

Furthermore, patients were excluded if they had a history of heart failure recorded in MINAP or GPRD.

4.2.2.6. Unheralded coronary death

Unheralded coronary death (UCD) was defined using one data source: ONS. Patients with unheralded coronary death were identified by an UCOD from coronary heart disease (ICD-10 I20-I25) which was not preceded by any other atherosclerotic code.

4.2.2.7. Ventricular arrhythmias, cardiac arrest and sudden cardiac death

Patients with ventricular arrhythmias, cardiac arrest or sudden cardiac death were identified by:

- a diagnosis of ventricular tachycardia (I47.2) or fibrillation (I47.0); or
- asystole, cardiac arrest (I46, I46.0 & I46.9), cardiac resuscitation, electro-mechanical dissociation; or
- implanted cardiac defibrillation device; or
- a diagnosis of sudden cardiac death in HES or GPRD; or
- mortality from cardiac arrest, sudden cardiac death or ventricular arrhythmia.

Patients were also excluded from the cohort if they had a record of history of ventricular tachycardia or fibrillation in GPRD.

4.2.2.8. Ischaemic stroke

Ischaemic stroke was defined using all four data sources: GPRD, HES, MINAP and ONS. Patients with stroke were identified by:

- a record of ischaemic stroke (I63); or
- a record of stroke not otherwise specified (I64, G46.3-G46.7); or
- mortality from any of the above as the first presentation of CVD.

Patients with stroke NOS were included in this category on the basis that the large majority of such unspecified strokes would be ischaemic; however, this category will include some patients with haemorrhagic stroke.

In addition, patients were excluded from the cohort if they had a record for sequelae of stroke (I69) or other cerebrovascular events such as transient ischaemic attack (TIA) (G45.8 & G45.9), a record of other ischaemic cerebrovascular diseases including occlusion and stenosis of cerebral arteries (I65 & I66), other cerebrovascular diseases (I67) and other vascular syndromes of brain in cerebrovascular disease not used elsewhere (G46), abnormal result on a carotid ultrasound or angiogram, abnormal result on cerebral CT; or a procedure on cerebral arteries indicative of atherosclerotic disease.

4.2.2.9. Abdominal aortic aneurysm event

Abdominal aortic aneurysm (AAA) was defined using all four data sources: GPRD, HES, MINAP and ONS. Patients with AAA were identified by:

- a diagnosis of AAA (I71), other than thoracic AAA ; or
- an abnormal results on abdominal ultrasound or CT indicating AAA; or
- a procedure for AAA; or
- mortality from AAA as the first presentation of CVD.

4.2.2.10. Peripheral arterial disease

Peripheral arterial disease (PAD) was defined using all four data sources: GPRD, HES, MINAP and ONS. Patients with PAD were identified by:

- a diagnosis of PAD (I73, I73.1, I73.8, I73.9) including peripheral complications of diabetes, peripheral ischaemia, peripheral vascular disease, gangrene and intermittent claudication; or
- an abnormal results on abdominal ultrasound or CT indicating PAD; or
- a procedure for PAD; or
- Mortality from PAD as the first presentation of CVD.

4.2.3. Statistical description of endpoints

With EHR data, it is possible to compare the number of initial presentations of CVD overall with a specific endpoint to the number of first presentations of that specific endpoint (ignoring any prior presentations with other disease presentations). Such a comparison gives a sense of the importance of distinguishing between initial presentation of CVD overall and first presentation of specific disease presentations and how this might vary between types of CVD. This comparison is presented, separately for men and women, by broad age groups, in Chapter 4.

I have also described the number and proportion of each specific initial presentation of CVD, separately for men and women. I have included in Chapter 4 the number of patients censored and have given the specific reason for censoring. I have given the number of endpoints which were fatal on the same or next calendar day as the endpoint. More detailed timing of endpoints is not available so deaths within 24 hours cannot be calculated. To give a sense of distribution of events by age, I have presented the rate for each initial presentation in 10 year age bands. Finally, I have presented the number of events that are contributed by each constituent data source in CALIBER for each endpoint.

5. Statistical modelling

5.1. Approaches to handling missing data

With any longitudinal cohort study, whether bespoke or EHR, missing data can be an issue both in its potential to create bias and/or the loss of power and precision in effect estimates with complete case analysis. Both types of cohort study have problems with loss to follow up, and missing data among those who are followed up, but the effect of missing data is likely to be different. With the relatively small sample sizes in bespoke cohorts, the loss of precision and power from missing data can become an important issue; with the large (indeed huge) sample sizes available in EHR cohorts, the greater issue is potential information bias, making it important to understand why data may not be recorded in EHR. Rubin first identified the importance of the mechanism that lead to missing data when determining appropriate approaches to dealing with such missingness.(178) He identified the following categories of missingness:

1. Missing completely at random – the missing values are not systematically different from the observed values and the reason for missingness is unrelated to the factor being measured. So, for example, a smoking record is not recorded for a patient because the practice nurse missed work because of a transport strike.
2. Missing at random – missing values are systematically different from the observed values but the difference can be explained by differences in other variables. So, for example, young men are more likely to have a missing smoking record in

GPRD than young women because they do not attend their GP for contraceptive advice and therefore have a smoking history taken prior to prescription of the birth control pill.

3. Missing not at random – missing values are systematically different from the observed values and the reason for missingness is directly related to the factor being observed. So, for example, smokers may be less likely to have their smoking status recorded in GPRD because some smokers may avoid going to the GP in case they are challenged about their smoking.

There are a number of approaches to dealing with missingness, whether it be to use complete case analysis, or to replace the missing values with imputed values. Each approach has specific drawbacks, some of which are outlined in Sterne et al.(179) and which may or may not be appropriate depending on the reason for missingness.[7] Multiple imputation of missing data is becoming the approach of choice for dealing with missingness, but I decided to carry out complete case analysis, for a number of reasons. First, for a limited number of co-variates that capture some of the key risk factors for cardiovascular disease, missingness is low in my covariates. Second, multiple imputation is computationally very intense(179); given that I have to take a pragmatic approach to my modelling to take account of computational capacity, without imputation, dealing with missingness using multiple imputation was not practical. Third, and not least, it is not clear that it was appropriate to use multiple imputation with these EHR in my dataset. A general research programme is currently under way at University College London on the appropriateness of multiple imputation to deal with missingness in health indicators in primary care EHR research. The initial findings, using data from the Health Improvement Network (THIN) database the data from which is sourced from most of the same practices as GPRD, suggest that some data, such as blood pressure and height and weight, are missing at random, while others, particularly reported smoking and drinking habits, are missing not at random, rendering multiple imputation inappropriate for at least these latter variables.(180) In fact, the reasons for missingness across the whole GPRD dataset may be different – Marston et al. was limited to health indicator information recorded within one year of registration, when data are more likely to be recorded as part of general health check rather than because of some indication such as subjective

[7] One approach which is usually given short shrift is replacing missing values with a missing category; however, it is worth noting that this approach could lead to useful information when combined with EHR research. Where patients do not have certain risk factor information recorded, patients who lack this information may have specific risks. In a preliminary analysis for this thesis, I found that no recording of smoking status had a similar association with unheralded coronary death as did smoking. Although this raises intriguing questions, the issue of non-recorded risk factor information was not further pursued in this thesis.

assessment by the GP of excess weight. The data used in this PhD have not been limited to such a narrow time window, adding weight to the arguments against using multiple imputation. Although these potential information biases are also problematic for complete case analysis, I considered it better not to compound the missing data problem with multiple imputation wrongly applied.

5.1.1. Description of missing data and patients excluded for missing data

In Chapter 4, I have identified the extent of missing data for each of my risk factors and co-variates, to the extent it can be identified. So, for example, it is possible to state how many patients do not have a smoking status recorded in their GPRD record, but it is not possible to state how many people who have diabetes do not have a diagnosis of diabetes recorded. I have also described differences between patients in the overall cohort with missing and complete data for IMD, smoking and blood pressure.

5.2. Approaches to modelling competing risks

The main focus of this PhD is to understand the effect of gender, smoking and blood pressure on initial presentation of specific cardiovascular disease phenotypes. As I have shown in Chapters 5-7, very few previous studies have attempted this intuitively important analysis because of small statistical size, lack of assessment of a wide range of phenotypes in a single study, or lack of the temporal resolution necessary to identify phenotypes as initial presentation. In studies potentially large enough to assess multiple endpoints, most studies use composite endpoints which, for example in coronary disease, conflate CHD death with non-fatal MI. Such composite endpoints conflate both multiple phenotypes (fatal MI combined with other fatal coronary disease) and confuse onset and progression. Such approaches also privilege the study of mortality over onset, although this is not always stated. While such an approach might be justified during time periods with high rates of cardiovascular mortality, as mortality declines and morbidity from cardiovascular disease increases,(181,182) such a focus becomes harder to justify. A refocusing on initial presentation of cardiovascular phenotypes and possible approaches to primary prevention across all arterial beds becomes more important.

Different initial presentations can be seen as risks which compete with each other. Only one cardiovascular phenotype can be the initial presentation and the others compete to be an alternate initial presentation. This approach to modelling associations between risk factors and phenotypes is relatively unusual in cardiovascular disease epidemiology – see Chapters 5, 6, and 7 for exemplar studies on association of gender, smoking and blood pressure for specific phenotypes - but is much more common in epidemiological studies

of cancer, where such approaches have been used to identified competing risks of disease reoccurrence, onset of secondary cancers or risks associated with treatment.(183)

So what are the appropriate statistical methods to evaluate whether one specific risk factor has similar or heterogeneous effects on the first of many different endpoints? Consider the following simplified example: We are interested in determining whether women are more likely than men to present with stable angina as the first clinical manifestation of coronary disease. We know that some women may in fact have a myocardial infarction without ever being diagnosed with stable angina. If a woman first develops an MI as her initial coronary disease presentation, she cannot develop stable angina as the first presentation; the MI prevents stable angina from occurring *as an initial presentation*. Below I describe approaches to describing and modelling this situation.

5.2.1. Descriptive methods

The naïve approach to describing time to competing events, called by some the 'naïve' Kaplan Meier approach, would be to treat the MI as a censoring event, as we would, more conventionally, treat withdrawal from the study or the end of study. However, these other administrative reasons for censoring differ in one important regard to another cardiovascular event: theoretically, the patient could still experience stable angina after administrative censoring. We are simply not able to observe this because this other unrelated event stopped our observation of the patient; his or her hazard of stable angina is the same as those patients remaining in the study. However, with the occurrence of a different initial presentation such as MI, the presentation of stable angina as an initial presentation could never occur. His or her hazard of experiencing an initial presentation of angina is not the same as those patients in the study because it is now zero. If we treat a different presentation simply as another reason for censoring, a key assumption of time-to-event analysis is violated: the independence of time to event and time to censoring distributions. Putter et al., amongst others, have shown that treating competing events as censoring events leads to an overestimation of probability of an event at any one point in time (of both the event of interest and the competing risk) such that the probability of either event occurring can be greater than 1, clearly an impossibility.(184,185)

A different approach, called the latent failure time approach, assumes that each person has a hypothetical failure time for each type of competing risk. In our simple example, one woman might have a hypothetical failure time for stable angina of 1.5 years and of 2.3 years for myocardial infarction. We never observe the failure time for myocardial infarction because the failure time for stable angina comes first. The difficulty with this approach is that we, by definition, cannot observe the latent failure times for events

which do not occur; we simply do not know what those times might be. More formally, without making strong additional assumptions, we cannot choose between different joint survival functions which can all fit the observed cause-specific hazards equally well.(184)

A more appropriate approach to describing multiple endpoints is the crude cumulative incidence function, or sub-distribution function, which takes account of other competing risks in calculating the probability of a specific endpoint at a given point in time. To return to our example, we can calculate the probability of an initial presentation of stable angina at a given point in time, say at 5 years, given that an MI has not occurred. The cumulative incidence for a specific endpoint at time *t* is calculated by summing the proportion of individuals at risk who experience that endpoint at every point in time up to time t. Individuals are removed from the risk pool whenever a competing event occurs. This approach allows researchers to look at the probability of occurrence of specific events, taking into account other competing events, but does not in itself allow any analysis of the cause-specific effect of co-variates.

5.2.2. Modelling methods

An approach to modelling competing risks which does allow comparison of cause-specific co-variates is to model co-variates and endpoints jointly in a Cox proportional hazard model. This can be done with record duplication or augmentation, creating a record for each possible failure for every patient in the study, with only one of the duplicate records recording a failure for the failure type that is actually observed. So, in our example, a woman could have a failure from stable angina or she could have a failure from MI; it so happens that she had a failure from stable angina but not MI, but in the data to be analysed she will have two records, one for failure type stable angina and one for failure type MI. This approach, proposed by Lunn and McNeil(186) and recommended by Putter et al. (184) has the virtue that standard approaches and commands for time-to-event analysis, such as Cox proportional hazard, can be used once the dataset has been augmented. This approach allows simultaneous estimation of the cause-specific hazards of different failure types, and the ability to test the equality of the effect of specific covariates. They are more accurate than any hazard calculated in separate individual Cox proportional hazards because they have been estimated jointly. (186)

These are interpreted essentially as any interaction term might be – the effect of the covariate, say smoking, on specific failure types, say angina as compared to MI, with the added caveat that the effect has been estimated in the context of the other potential hazards.(187) Incidentally, this approach can be extended to do multi-state analysis, investigating multiple sequential failure types. Essentially what the cause-specific hazard

is estimating is the "probability of failure specifically resulting from cause *k* in a small interval of time, given that no failure of any kind has occurred thus far."(183) Consequently, the overall hazard for any kind of failure can be derived by adding all the cause-specific hazards. The virtue in this approach to modelling cause-specific risks is that the effect of covariates can be compared in a straight forward way; "the drawback is that the everyday interpretation of these hazards is not intuitive."(184)

It is also possible to estimate the cumulative incidence function for specific endpoints, taking into account the cause-specific hazards of a set of co-variates, essentially combining the hazard ratios from the Cox modelling with the incidence function. Although this approach could provide the most interpretable model while taking account of both competing risks and cause-specific hazards, it is computationally difficult. No STATA program exists to compute these models and is therefore beyond the scope of this PhD.

5.2.3. Computational considerations and specific approach used to model cause-specific hazards

I developed a common Cox proportional hazard model(188) for competing risks, using data augmentation, to estimate the hazard ratios for the co-variates for each presentation in the context of competing risks.(184,186) Data augmentation requires creating a dataset with a stratum for each endpoint which includes all patients, so, for example, with 10 endpoints and 1,000,000 patients, the dataset to be analysed would be 10,000,000 records. Within each endpoint stratum, occurrence of the relevant endpoint is counted as a failure and patients are censored at the point at which a competing event occurs. With datasets of this size, computational capacity and time are serious concerns. Although Lunn and McNeil advocate running the competing risk as a single overall model with interaction terms for all co-variates and strata,(186) I followed an alternative approach suggested by Putter et al.(184) and ran each stratum in sequence, following creation of the appropriate dataset. Even with this approach, the more complex frailty models for a single endpoint took up to 10 hours to converge on a powerful personal computer (Dell computer with core 2, quad 3 GHZ processors, and 8 Gb RAM, local copy of STATA MP4). Using other computing resources, such as the University College London super-computer, Legion, was not an option because of the data permission requirements of the GPRD constituent dataset. Through experimentation with running multiple copies of STATA and increase in computer power (Dell personal computer I7, 8 processor with 3.4 gigahertz (GHZ) and 16 Gb RAM), run times were improved but still remain a substantial issue with survival analysis and datasets of this size.

One drawback to this pragmatic approach to competing risks analysis is that I could not model the random effect of key risk factors between endpoints because each endpoint is modelled separately. In order to confirm the existence and extent of heterogeneity between endpoints, I therefore used techniques from meta-analysis and calculated the I^2 and tau-squared (τ^2).(189)

To further specify the Cox models, I used the Efron method to deal with tied failure times.(190) Patients are clustered at general practice level, so a frailty term was included to take account of this clustering.(191) Where events occurred on the same day as study entry, a factor of 0.5 was added to the event date, so these events were not excluded from the model estimations. I tested the proportional hazards assumption visually using log-log rank graphs. All analyses were performed using STATA version 12 (StataCorp, 4905 Lakeway Drive, College Station, TX 77845, United States).

The cause-specific hazard ratios, with 95% confidence intervals, are presented in the chapters on the association of gender, smoking and blood pressure on specific cardiovascular fatal and non-fatal endpoints. In all chapters, I have first shown the broad cardiovascular endpoints of ischaemic stroke, coronary disease (defined as acute MI and unheralded coronary death), abdominal aortic aneurysm and peripheral arterial disease and then the cardiac presentations of stable angina, unstable angina, acute myocardial infarction, ventricular arrhythmias/ cardiac arrest/sudden cardiac death, heart failure, and unheralded coronary death. I have further presented the cause-specific hazards for ST-elevation myocardial infarction (STEMI), non-STEMI and myocardial infarction not further specified in additional analyses.

6. Summary of descriptive analyses completed

Throughout this chapter, I have noted the descriptive analyses I have presented in Chapter 4 where relevant. Here, for convenience, I summarise those analyses. The results of the modelling of the association of gender, smoking and blood pressure have been presented in *Chapters 5, 6 and 7*. I describe the following information on my cohort:

1. *Derivation of the cohort:* The number (percentage) excluded for each inclusion/exclusion criteria. Exclusions for previous CVD are further broken down by specific CVD presentation.
2. *Observation time:* Mean and standard deviation of observation time in years, separately for women and men in ten year age bands; further separated into overall, UTS observation time and observation time after study entry.
3. *Number of observations for co-variates with potentially multiple records:* For those co-variates where the number of measurements can vary between patients, i.e. excluding

age and sex, I have presented the number of patients, for men and women separately, who had no observations and, amongst those who had at least one observation, the median number of observations and interquartile range.

4. *Baseline characteristics of cohort:* Mean (standard deviation) of all continuous co-variates and number (proportion) of all categorical co-variates, as well as number of patients with no measurements at any point.
5. *Comparison of initial and first presentation for all endpoints*: The number and percentage of endpoints, for initial presentation of CVD overall and for first presentation within a specific CVD, separately for men and women, over all ages and for 3 broad age groups.
6. *Number of endpoint and case fatality:* The number and percentage of endpoints, including number and percentage of patients censored at practice leaving or last practice download. The number of endpoints fatal within 1 calendar day.
7. *Rate of specific initial presentations:* The rate of each specific initial presentation in 10 year age bands.
8. *Data source for endpoints:* The number and percentage of each endpoint to which each constituent dataset contributed.
9. *Patients with missing data for specific co-variates:* Comparison of characteristics of patients with missing data for IMD, smoking and blood pressure compared to those with complete data on these variables.

7. Conclusions

In this chapter, I have described the derivation of the cohort used in my three related studies of the association gender, smoking and blood pressure with the initial presentation of different cardiovascular disease phenotypes. I defined the risk factors, co-variates and endpoints used in my studies. I then justified the need for competing risk analysis for these studies and my decision to use complete case analysis, rather than other approaches to missing data. Finally, I described the statistical methods I used to describe the cohort and endpoints and model the associations between my risk factors and the multiple endpoints, including the concessions I made in my modelling approach on grounds of computational intensity.

Description of Cohort Population

1. Introduction

In the previous two chapters, I described the creation of the CALIBER research platform, the study design, the derivation of the cohort population, the definition of co-variates and the 12 cardiovascular disease (CVD) endpoints used in my three related studies. I also described my statistical approach to describing the cohort and analysing the association between my key risk factors and the endpoints within a competing risk framework. Here I have described the characteristics of the cohort in detail, presenting results of the descriptive analyses common to all three studies.

2. Derivation of the cohort population

The derivation of the overall cohort population, from CALIBER research platform, is shown in Figure 7. The total General Practice Research Database (GRPD) population available within CALIBER was 5,372,790. As described in Chapter 3, I included patients who:

1) were of acceptable research quality,
2) had a definite gender,
3) were registered with their practice on or after 1st January 2000,
4) were aged between 30 and 100 the year before study entry, and
5) had at least 1 year of up-to-standard (UTS) observation after 1st January 2000 in GPRD.

Patients were excluded if they:

6) had an Office of National Statistics (ONS) date of death prior to study entry;
7) had more than one ONS date of death;
8) had a record in any of CALIBER constituent data sources of symptomatic atherosclerotic disease prior to study entry, i.e. a record of coronary disease, ischaemic cerebrovascular disease, peripheral arterial disease, unspecified atherosclerotic disease, ventricular arrhythmias, cardiac arrest or heart failure; or
9) had a first record after study entry of code indicating history of disease.

With electronic health record (EHR) data, unlike with some bespoke cohorts, it is possible to look at all the patients who did not meet the inclusion criteria. I have therefore given the number and proportion of patients in the overall CALIBER research platform who did not met the inclusion criteria as well as showing the number of patients who were excluded by each of the exclusion criteria. (See Figure 1.) After exclusions, 1,758,584 patients aged 30 to 100 with no history of clinically manifest cardiovascular disease (CVD) remained. For each set of analyses in Chapters 5 to 7, patients with relevant risk

factor data were selected to form the cohorts specific to those set of analyses. Further details are given in each of the relevant chapters.

Table 4 shows the number of patients excluded for each prior disease presentation, including those whose first record after study entry indicated history of relevant disease. The majority of the exclusions were for previous cardiac disease, although a large number of patients were also excluded because of previous cerebrovascular disease. The greatest number of patients who were excluded because they had a history of CVD as their first record after study entry were excluded because of a history of stroke. The latter category could include patients with a history of haemorrhagic stroke, who were not explicitly excluded from the study.

Figure 7: Derivation of cohort population, including number of endpoints in cohort

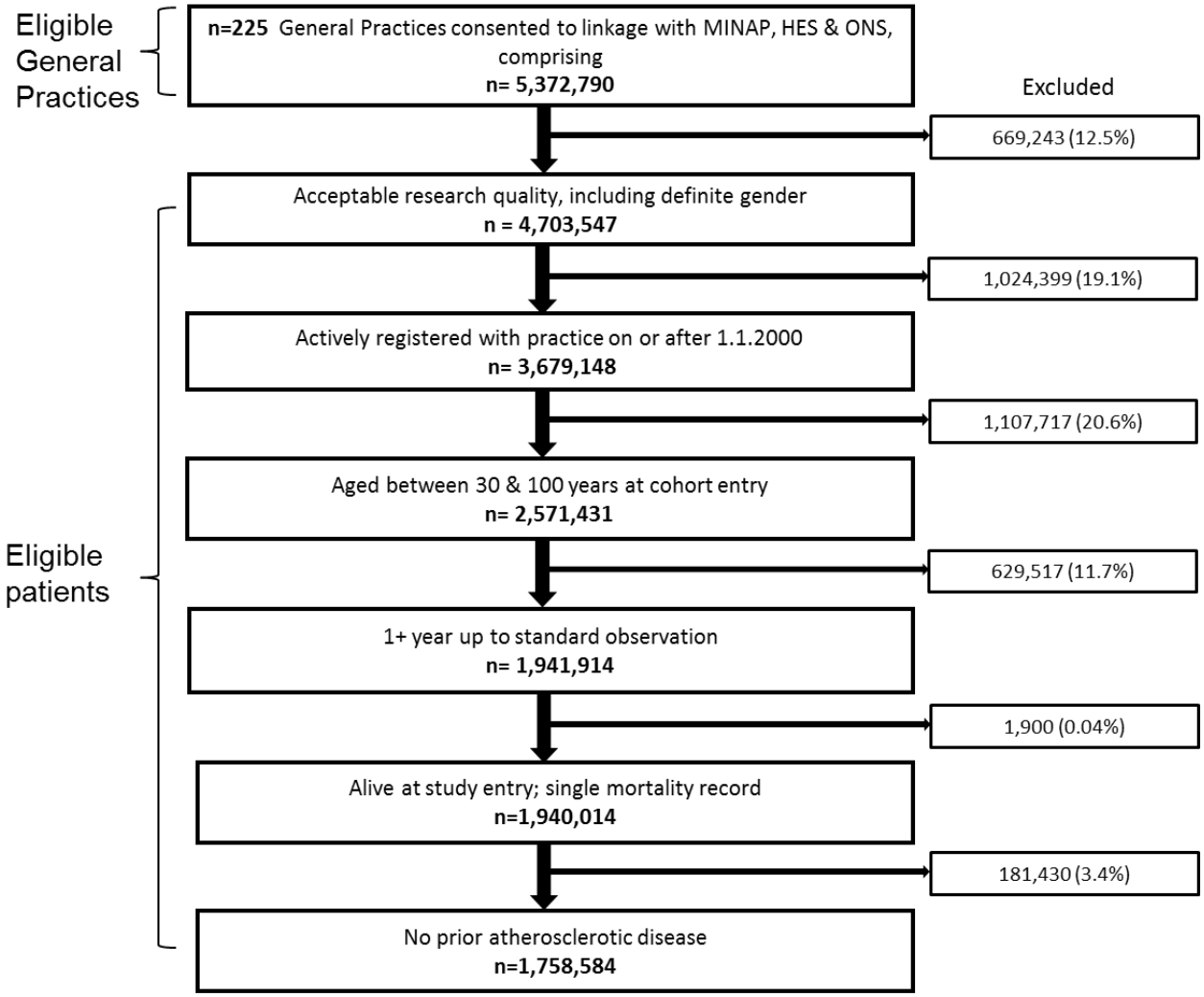


Table 4: Clinically manifest CVD at baseline: Number of patients excluded for specific cardiovascular disease categories

Disease	Prior to study entry	"History of" as 1st record
Stroke, transient ischaemic attack or other ischaemic cerebral disease	37,518	1,573
Abdominal aortic aneurysm	3,765	23
Peripheral arterial disease	11,953	1
Other unspecified atherosclerotic disease	845	47
All cardiac	125,606	99
Stable angina	85,809	--
Unstable angina	548	--
Coronary artery bypass graft and percutaneous coronary intervention	2,732	36
Coronary heart disease not otherwise specified	6,695	4
Acute myocardial infarction	15,533	46
Heart failure	12,261	1
Ventricular arrhythmias, cardiac arrest & sudden cardiac death	2,028	12
Total	179,687	1,743

3. Observation time

As discussed in Chapter 3, the period during which patients in these EHRs could be observed can be divided into three periods. The first is from the very first point at which they register with a practice until they have a specific endpoint or are censored. This is the maximal time during which information is available on these patients. However, many patients registered with a practice before the practice met GPRD-specified data quality standards and became up to standard (UTS). Thus a second possible period of observation is the time patients spent registered with a practice which was UTS. A third possible type of observation time is the more conventional time under observation after study entry which could be called the endpoint follow-up period.

The 1,758,592 patients in the cohort had a total observation time from patient registration to diagnosis/censoring of 27,073,983 person years, with a median total observation time of 12.1 person years (IQR: 5.2-21.4). The total UTS observation time was 16,277,652 person years with a median of 8.9 (4.3-13.0). The total observation time during endpoint follow-up was 9,397,209 person years, median 5.5 (2.1-9.1). Each type of observation time has been shown separately for women and men in 10 year age bands.

Older patients tended to have longer periods of observation time, both before and after study entry, with the longest median endpoint follow-time in the 50-59 age band.

Table 5: Person years of observation time, in years, from patient registration, practice up to standard date and endpoint follow-up

		Patient registration to end of follow-up time			
Age	**Observation time from**	**Women**		**Men**	
		Person yrs	**Median (IQR)**	**Person yrs**	**Median (IQR)**
30-39	Patient registration	3,699,762	8.4 (4.1,15.5)	4,081,184	8.6 (4.1,16.4)
	Practice UTS time	2,718,088	7.5 (3.7,11.4)	2,834,450	7.6 (3.7,11.4)
	Endpoint follow-up	1,495,058	4.0 (1.4,8.2)	1,523,289	3.9 (1.5,8.0)
40-49	Patient registration	2,828,389	13.5 (5.5,22.2)	2,891,166	12.1 (5.,20.9)
	Practice UTS time	1,820,152	9.7 (4.4,13.7)	1,863,302	9.0 (4.1,13.2)
	Endpoint follow-up	1,130,215	7.4 (2.7,9.1)	1,160,918	6.7 (2.5,9.1)
50-59	Patient registration	2,911,600	16.4 (7.5,27.0)	2,831,390	16.2 (7.5,26.0)
	Practice UTS time	1,644,367	10.3 (5.6,14.4)	1,594,646	10.3 (5.4,14.1)
	Endpoint follow-up	993,756	8.3 (3.3,9.1)	954,902	7.6 (3.1,9.1)
60-69	Patient registration	2,081,170	16.5 (7.7,28.9)	1,792,371	16.3 (7.5,28.5)
	Practice UTS time	1,081,902	10.3 (5.4,14.4)	918,960	10.1 (5.,14.1)
	Endpoint follow-up	643,040	7.7 (3.,9.1)	530,827	6.6 (2.7,9.1)
70-79	Patient registration	1,560,571	17.3 (9.1,29.3)	1,062,665	17.1 (9.,30.2)
	Practice UTS time	744,825	10.3 (5.6,14.1)	489,202	10.0 (5.1,13.7)
	Endpoint follow-up	420,442	6.6 (2.7,9.1)	266,247	5.5 (2.3,9.1)
80+	Patient registration	913,550	13.4 (4.9,24.3)	420,165	15.3 (6.7,27.5)
	Practice UTS time	396,110	7.1 (3.4,11.6)	171,646	7.5 (3.6,12.1)
	Endpoint follow-up	197,009	3.2 (1.4,6.2)	81,507	3.2 (1.3,6.1)
Overall	Patient registration	13,995,042	12.3 (5.3,21.5)	13,078,941	11.9 (5.2,21.2)
	Practice UTS time	8,405,444	9.0 (4.3,13.1)	7,872,207	8.7 (4.2,12.9)
	Endpoint follow-up	4,879,520	5.7 (2.1,9.1)	4,517,689	5.2 (2.1,9.1)

4. Number of observations for co-variates

The number of patients with no data recorded for each potential co-variate and the mean number of observations and interquartile range for those with data recorded are presented in Table 4. For the modelling in this thesis, only baseline data was used, but this table indicates that amongst those with any data recorded, there are multiple records for most potential co-variates, particularly the medications reflecting the use of repeat prescriptions. Both the index of multiple deprivation and ethnic group in hospital episode statistics are recorded once for each patient.

Table 6: Number of patients with missing data for baseline characteristics, and the median and interquartile range of observations for patients with data

	No observations		Total Observations	
	Women	**Men**	**Women**	**Men**
Index of Multiple Deprivation	3,878	3,533		
Ethnic Group (HES)	337,531	466,850		
Ethnic Group (GPRD)	621,291	619,428	1 (1-1)	1 (1-1)
Ethnic group (overall)	383,168	471,362		
Smoking status	118,767	186,182	3 (2-6)	3 (1-5)
Systolic & diastolic blood pressure	318,688	478,277	10 (4-20)	8 (3-19)
Total cholesterol	798,875	756,337	3 (1-6)	3 (1-6)
HDL cholesterol	838,259	795,382	2 (1-5)	3 (1-5)
LDL cholesterol	850,272	808,489	2 (1-4)	2 (1-4)
Triglycerides	829,244	785,841	2 (1-5)	3 (1-5)
Diabetes diagnosis	882,721	837,431	2 (1-3)	2 (1-3)
Diabetic medication scripts - Any	882,929	839,975	37 (13-86)	41 (16-88)
- Insulin	897,241	854,494	22 (6-50)	27 (10-56)
- Oral	885,338	843,051	36 (12-85)	39 (15-87)
Body mass index	510,132	567,699	4 (2-8)	3 (1-6)
Alcohol consumption status	571,786	572,134	2 (1-3)	2 (1-3)
Statin scripts	878,899	836,887	23 (10-45)	21 (9-40)
BP lowering scripts - Any	777,792	781,352	30 (5-78)	26 (5-67)
- Diuretics	844,088	832,393	23 (4-55)	21 (5-47)
- β-blockers	843,730	825,012	8 (2-34)	10 (2-33)
- ACEIs & ARBs	859,395	823,298	25 (6-56)	21 (5-51)
- CA-channel blocker	867,730	832,799	16 (4-38)	18 (6-38)

HES indicates Hospital Episode Statistics; GPRD, General Practice Research Database; ACEI, Angiotensin converting enzyme inhibitor; ARB, angiotensin receptor blocker; BP, blood pressure; CA, calcium.

5. Baseline characteristics of cohort

The baseline characteristics of the cohort have been presented, separately for men and women, in Table 7 below. Where comparisons were made with external data sources, data from 2005 were chosen because this year reflects the mid-point of the follow-up time.

Age: The population of the cohort was younger than the English population, with a higher proportion of patients in the younger age bands than the equivalent age bands in the mid-year population estimates for 2005 for England. This information has been included in Table 7 for comparison.(192) The younger age of the research cohort was partly because patients were excluded if they had a history of symptomatic cardiovascular disease, which would exclude more older than younger patients. The cohort was also younger because patients were added to the open cohort when they reached their 30th birthday. If patients who were added to the cohort on attaining their 30th birthday were ignored, the population structure was broadly similar to that of England as a whole.

Ethnic group: Amongst patients with ethnic group recorded, the majority of the patients reported belonging to *White* ethnic group; however, 42.6% of women and 54.9% of men did not have their ethnicity recorded.

Smoking: Substantially more women than men reported being non-smokers (women: 54.0% (95% CI 53.9-54.1%); men: 39.8% (39.7-39.9%); however, men were also much more likely than women not to have their smoking status recorded (men: 21.7% (21.6-21.8); women: 13.2% (13.1-13.3). The proportion of cohort patients who reported being current smokers was slightly lower than the proportion reported in the 2005 Health Survey for England – 27% in men and 24% in women. (193) While the lower rate may reflect missing information, a lower rate of smoking would be expected in a cohort free from symptomatic cardiovascular disease, given the association between smoking and cardiovascular disease.

Blood pressure: Men had higher mean blood pressure and less variability than women (men: 134.4 mmHg (134.4-134.5; women: 128.8 mmHg (128.8-128.9)). However, a substantially higher proportion of men did not have blood pressure recorded (men: 55.7% (55.6-55.8); women: 35.4% (35.3-35.5). Amongst patients with a blood pressure recording, the proportion of patients with raised blood pressure (>140/90) was essentially the same in both genders (women: 20.4% (20.3-20.5); men: 19.1% (19.0-19.2)), but a higher proportion of women than men were on blood pressure-lowering medication (women: 16.0 % (15.9-16.1); men: 10.6% (10.5-10.7)). The higher recording of blood pressure amongst women may be due to younger women having blood pressure measured as part of contraception risk assessment. The proportion of patients with high

blood pressure was substantially lower than that recorded in the Health Survey for England for 2005, which found 34.6% of men and 28.3% of women had high blood pressure.(193) Again, this could be due to the healthy cohort effect or possibly undiagnosed high blood pressure amongst the cohort.

Diabetes: Patients were recorded as having diabetes if they had a diagnosis of diabetes (Type I or II or unspecified) or were prescribed either insulin or oral anti-diabetic medication. There was therefore, by definition, no missing data, although patients who are diabetic may not have been diagnosed by their general practitioner. The proportion of patients who were diabetic, using all diagnostic and prescribing information, was similar in women and men (women: 2.4 (2.4-2.4); men: 2.8 (2.8-2.8)). Again this was somewhat lower than the proportion of people with diabetes reported in the Health Survey for England, which found 4.3% of men and 3.4% of women reported having diabetes.(193)

Lipids: The level of missing data for recording level of the different lipids was exceptionally high, with 88.7% (88.6-88.8) of women and 88.1% (88.0-88.2%) of men having no record at baseline.

Obesity: The mean body mass index (BMI) for women was 26.3 kg/m^2 (26.3-26.3) essentially the same as that for men whose mean BMI was 26.8 kg/m2 (26.8-26.8). These means were also similar to that in the Health Survey for England which found a mean BMI of 26.9 for both men and women.(193) However, there is again a high proportion of missing data amongst the cohort patients, with a higher proportion of data missing in men (women: 56.7% (56.6-56.8); men: 66.2% (66.1-66.3)).

Alcohol consumption: Unusually for this cohort, the proportion of men with no alcohol consumption information recorded (66.9% (66.8-67.0)) was similar to the proportion in women (63.6% (63.5-63.7). Amongst those with this information recorded, about one fifth of both men and women with alcohol consumption recorded were current drinkers, which is higher than the proportion in women in the Health Survey for England but lower than the proportion in men.(193)

Table 7: Baseline characteristics of the cohort, in women and men

	Women			Men		
Socio-demographic factors						
Age in years, n (%)	n	%			%	
		Cohort	England mid-2005*		Cohort	England mid-2005*
30-39	332,390	36.9	22.7	342,067	39.9	24.6
40-49	187,562	20.8	22.1	201,264	23.5	24.6
50-59	155,990	17.3	19.4	153,875	17.9	20.6
60-69	103,675	11.5	15.1	91,221	10.6	15.5
70-79	71,816	8.0	11.9	49,166	5.7	10.6
80+	48,968	5.5	8.9	20,590	2.4	5.1
Total	900,401	--	--	858,183	--	--
Age in years, mean (sd)			48.0 (16)			45.6 (14)
Ethnic group, n (%)						
White	466,852		51.8	348,007		40.6
Black	15,770		1.8	11,434		1.3
South Asian	14,697		1.6	12,453		1.5
Other and mixed	19,914		2.2	14,927		1.7
Not recorded	383,168		42.6	471,362		54.9
Deprivation Quintile, n (%)*						
1st quintile (most deprived)	174,208		19.3	174,207		20.3
2nd quintile	178,099		19.8	171,647		20.0
3rd quintile	181,055		20.1	169,005		19.7
4th quintile	181,475		20.2	169,488		19.7
5th quintile (least deprived)	181,686		20.2	170,303		19.8
Not recorded	3,878		0.4	3,533		0.4
Clinical risk factors						
Smoking Status, n (%)						
Non smoker	486,059		54.0	341,600		39.8
Ex-smoker	113,172		12.6	115,375		13.4
Current Smoker	182,403		20.3	215,026		25.1
Not recorded	118,767		13.2	186,182		21.7
Blood Pressure						
SBP in mmHg, mean (sd)	128.8		19.6	134.4		17.2
DBP in mmHg, mean (sd)	77.7		9.8	81.2		9.8
Raised blood pressure, n (%)	184,120		20.4	164273		19.1
On BP lowering medication, n (%)	143,970		16.0	90,846		10.6
Treated SBP in mmHG, mean (sd)	148.1		18.6	148.6		16.4
Not recorded, n (%)	318,821		35.4	478,413		55.7
Diabetes, n (%)						
Diagnosed with diabetes	17,680		2.0	20,752		2.4
On insulin	3,160		0.4	3,689		0.4
On oral medication	15,063		1.7	15,132		1.8
On any medication	18,223		2.0	18,821		2.2
Diagnosed or on medication	21,957		2.4	23,668		2.8
Lipids in mmol/l, mean (sd)						
Total cholesterol (TC)	5.5		1.1	5.4		1.1
High density lipoprotein	1.5		0.5	1.3		0.4
Low density lipoprotein	3.3		1.0	3.3		1.0
Triglycerides	1.5		1.0	1.9		1.5
On statins at baseline, n (%)	21,502		2.4	21,296		2.5
Total cholesterol on statins	5.4		1.3	5.1		1.3
Not recorded (TC), n (%)	798,875		88.7	756,337		88.1
Obesity						
BMI in kg/m^2, mean (sd)	26.3		5.8	26.8		4.6

	Women		Men	
Not recorded, n (%)	510,132	56.7	567,699	66.2
Alcohol consumption, n (%)				
Non-drinker	68,397	7.6	34,208	4.0
Ex-drinker	5,784	0.6	5,879	0.7
Occasional drinker (<1x a week)	42,916	4.8	23,440	2.7
Current drinker	192,111	21.3	186,400	21.7
Excess drinker, incl. binging	18,540	2.1	34,913	4.1
Not recorded	572,653	63.6	573,343	66.9

[*]Mid-year population estimates from Office for National Statistics. sd indicates standard deviation; TC, total cholesterol; SBP, systolic blood pressure; DBP, diastolic blood pressure; BP, blood pressure; BMI, body mass index.

6. Data source for all endpoints

The constituent data source for each endpoint was identified. GPRD contributes the greatest number of CVD endpoints, followed by HES, ONS and then MINAP, although the relative contribution of each data source varies depending on the specific initial presentation. The number and proportion of patients with each endpoint, from each data source, have been given separately for men and women in Table 8.

Table 8: The proportion of specific endpoint provided by GPRD, HES, MINAP and ONS

	Data Source n (% of category)			
Disease	**GPRD**	**HES**	**MINAP**	**ONS**
Stroke	7,380 (46.3)	7,216 (45.2)	0	1,357 (8.5)
Abdominal aortic aneurysm	1,749 (65.5)	557 (20.9)	0	365 (13.7)
Peripheral arterial disease	8,488 (86.2)	1,281 (13.0)	0	78 (0.8)
Cardiac disease	40,027 (59.9)	19,798 (29.6)	1,136 (1.7)	5,835 (8.7)
Stable angina	20,210 (98.7)	251 (1.2)	24 (0.1)	0
Unstable angina	563 (17.8)	2,542 (80.2)	63 (2.0)	0
Coronary heart disease not otherwise specified	5,660 (71.1)	2,297 (28.9)	0	0
Myocardial infarction	4,314 (32.7)	7,822 (59.3)	1,049 (8.0)	0
ST elevation myocardial infarction	384 (41.5)	0	542 (58.5)	0
Non ST elevation myocardial infarction	847 (62.6)	0	507 (37.4)	0
Myocardial infarction not otherwise specified	3,083 (28.3)	7,822 (71.7)	0	0
Heart failure	8,484 (67.8)	3,455 (27.6)	0	571 (4.6)
Ventricular arrhythmias, cardiac arrest & sudden cardiac death	1,438 (34.0)	2,789 (65.9)	0	7 (0.2)
Unheralded coronary death	0	0	0	5,527(100.0)
Total	57,644 (60.5)	28,852 (30.3)	1,136 (1.2)	7,635 (8.0)

GPRD indicates General Practice Research Database; HES, Hospital Episode Statistics; MINAP, Myocardial Ischaemia National Audit Project; ONS, Office for National Statistics.

7. Comparison of initial presentation of CVD with specific endpoints to first presentation of specific endpoints

Initial presentation of CVD overall with a specific endpoint is defined as the first presentation across any of the possible endpoints; first presentation of a specific endpoint is the first presentation of particular endpoint which may or may not be preceded by another presentation of different CVD endpoint. In Table 9 and Table 10**,** the number of first and initial presentations has been shown for women and for men in each of three broad age groups, as well as the proportion of first presentations which were initial presentation of CVD. The number of first and initial presentations was highest in the 60-79 age band; however, the increase in women was greater because men had high number of initial and first presentations in the youngest age group as well as in the middle age group. For both men and women, the proportion of first presentations which are also initial presentations were lower in cardiac disease phenotypes than in the other CVD phenotypes. In both men and women, the proportion of first presentations which were also initial presentations for each phenotype varies by age. The most common pattern for both men and women was a decline in the proportion of first presentations which were also initial presentations in middle age bands, increasing again in the oldest age group, although with the total presentations the proportion of initial which were also first presentations increased with increasing age in men but not women.

Table 9: Number of initial presentations of any cardiovascular disease endpoints, number of first presentations of cardiovascular disease endpoints, and proportion of initial presentations that also first presentations in women

Women	Age <60 years			Age 60-79 years			Age 80+ years			All ages		
	First	Initial	%	First	Initial	%	First	Initial	%	First	Initial	%
Stroke	1,356	1,269	93.6	4,668	4,110	88.0	4,166	3,682	88.4	10,190	9,061	88.9
Abdominal aortic aneurysm	69	54	78.3	602	481	79.9	247	212	85.8	918	747	81.4
Peripheral arterial disease	1,267	1,151	90.8	2,926	2,510	85.8	1,053	869	82.5	5,246	4,530	86.4
Acute MI & CHD death	1,607	1,311	81.6	4,689	3,463	73.9	2,967	2,159	72.8	9,263	6,933	74.8
Cardiac												
Stable angina	4,917	3,604	73.3	7,965	5,331	66.9	1,695	1,017	60.0	14,577	9,952	68.3
Unstable angina	1,019	559	54.9	1,419	660	46.5	377	184	48.8	2,815	1,403	49.8
Coronary heart disease NOS	2,747	990	36.0	5,414	1,891	34.9	1,018	466	45.8	9,179	3,347	36.5
Acute myocardial infarction	1,341	1,091	81.4	3,249	2,423	74.6	1,536	1,159	75.5	6,126	4,673	76.3
Vent. arrhyth, c arrest & sudden c arrest	682	573	84.0	1,212	873	72.0	265	181	68.3	2,159	1,627	75.4
Heart Failure	768	580	75.5	4,362	3,323	76.2	3,680	3,079	83.7	8,810	6,982	79.3
Coronary death	295	220	74.6	1,742	1,040	59.7	1,732	1,000	57.7	3,769	2,260	60.0
Total	14,461	10,091	69.8	33,559	22,642	67.5	15,769	11,849	75.1	63,789	44,582	69.9

MI indicates myocardial infarction; CHD, coronary heart disease; NOS -not otherwise specified; vent. Arrhyth – ventricular arrhythmia; c. arrest, cardiac arrest.

Table 10: Number of initial presentations of any cardiovascular disease endpoints, number of first presentations of cardiovascular disease endpoints, and proportion of initial presentations that also first presentations in men

	Age <60 years			Age 60-79 years			Age 80+ years			All ages		
Men	**First**	**Initial**	**%**	**First**	**Initial**	**%**	**First**	**Initial**	**%**	**First**	**Initial**	**%**
Stroke	1,989	1,793	90.1	4,388	3,762	85.7	1,532	1,337	87.3	7,909	6,892	87.1
Abdominal aortic aneurysm	359	289	80.5	1,697	1,326	78.1	359	309	86.1	2,415	1,924	79.7
Peripheral arterial disease	2,417	2,087	86.3	3,464	2,773	80.1	554	457	82.5	6,435	5,317	82.6
Acute MI & CHD death	6,184	5,252	84.9	6,770	5,052	74.6	1,720	1,205	70.1	14,674	11,509	78.4
Cardiac												
Stable angina	9,378	4,983	53.1	8,957	4,995	55.8	967	555	57.4	19,302	10,533	54.6
Unstable angina	2,055	1,046	50.9	1,571	628	40.0	199	91	45.7	3,825	1,765	46.1
Coronary heart disease NOS	7,448	2,100	28.2	7,528	2,241	29.8	659	269	40.8	15,635	4,610	29.5
Acute myocardial infarction	5,090	4,307	84.6	4,700	3,569	75.9	888	636	71.6	10,678	8,512	79.7
Vent. arrhyth, c arrest & sudden c arrest	1,626	1,209	74.4	1,934	1,284	66.4	167	114	68.3	3,727	2,607	69.9
Heart Failure	1,679	1,146	68.3	4,254	2,947	69.3	1,804	1,435	79.5	7,737	5,528	71.4
Coronary death	1,228	945	77.0	2,439	1,483	60.8	1,014	569	56.1	4,681	2,997	64.0
Total	33,269	19,905	59.8	40,932	25,008	61.1	8,143	5,772	70.9	82,344	50,685	61.6

MI indicates myocardial infarction; CHD, coronary heart disease; NOS - not otherwise specified; vent. Arrhyth – ventricular arrhythmia; c. arrest, cardiac arrest.

8. Number of endpoints and case fatality

Out of the 1,758,584 cohort patients, after a median of 5.7 years follow up in women and 5.2 in men, 95,267 (5.4%) patients experienced an initial presentation of interest while a further 55,620 died from another cause (3.2%). The remaining 1,663,317 (94.6%) left the practice (26.2%) or were censored at the last practice download (65.3%). Figure 8 shows the number of patients with each specific initial presentation of CVD, while Figure 9 shows the proportion each initial presentation made of all the presentations, in all patients and in women and men separately. Stroke and MI constituted 31% of all initial presentations, with other presentations constituting the majority (69%). While the proportion of presentations which were cardiac were similar between men and women, acute myocardial infarction formed a greater proportion of initial presentation in men, while heart failure was more common initial presentation in women. Women were also more likely to have stroke as an initial presentation compared to men.

Twelve per cent of the cardiovascular events were fatal within 1 calendar day of the event date (n=11,403), with a considerable range in the proportion of fatal events, depending on the diagnosis. As expected, unheralded coronary death is 100% fatal within 1 day, while stable angina is virtually 0%. Fatalities in the other presentations range from 0.6% for unstable angina to 23.1% for abdominal aortic aneurysm. Table 11 below shows the number and proportion of endpoints, as well as the number which were fatal within 1 calendar day. Of note is the higher case fatality in women from almost all initial presentations, with stroke and abdominal aortic aneurysm showing the greatest differences between genders.

Figure 8: Number of initial presentations in patients with no clinically manifest CVD at baseline

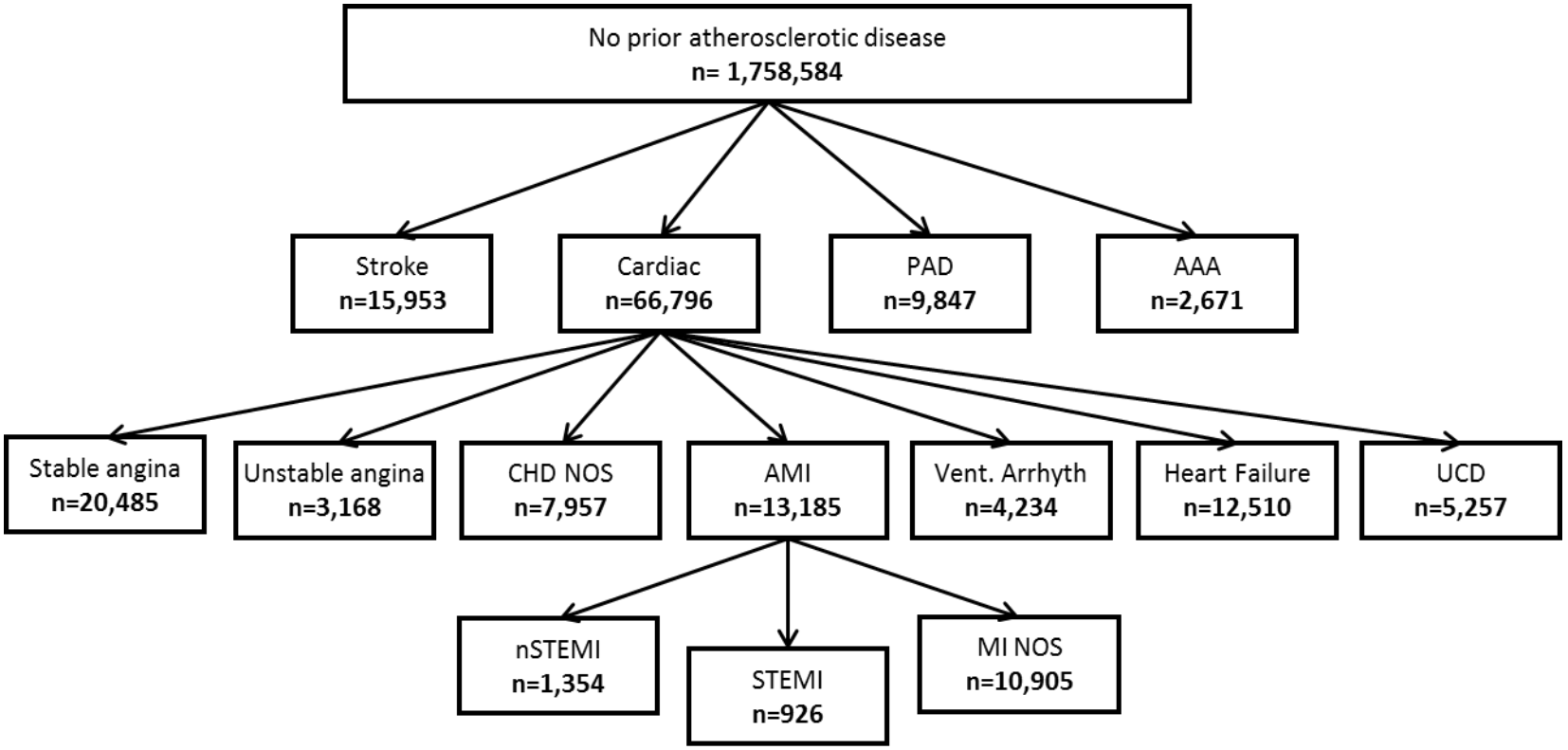
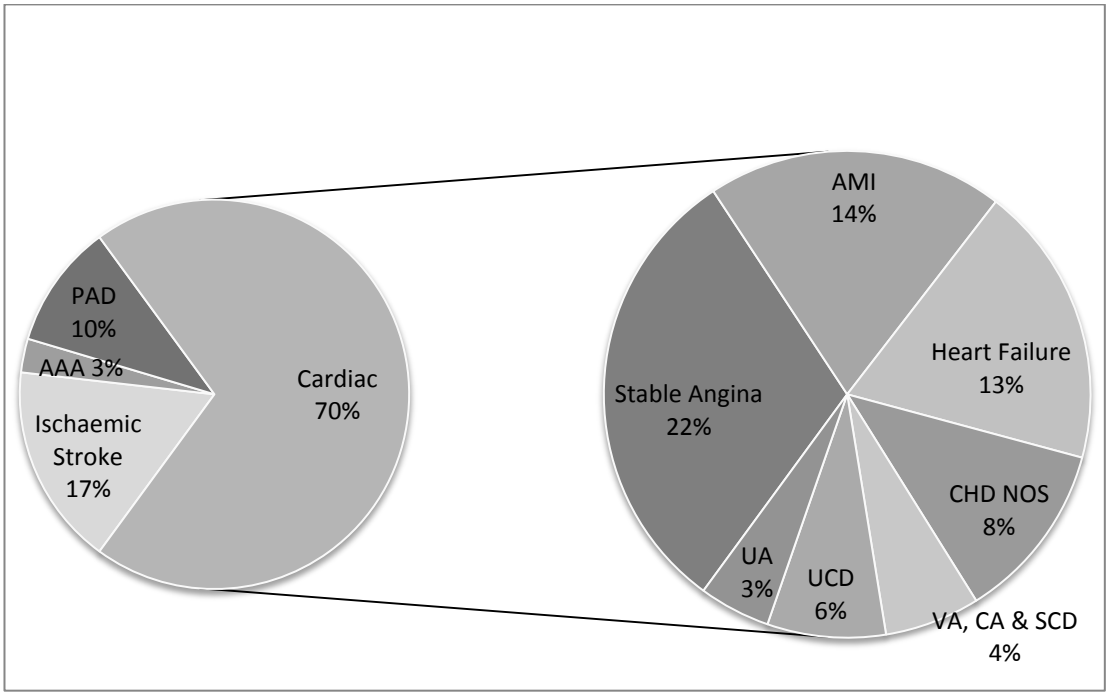
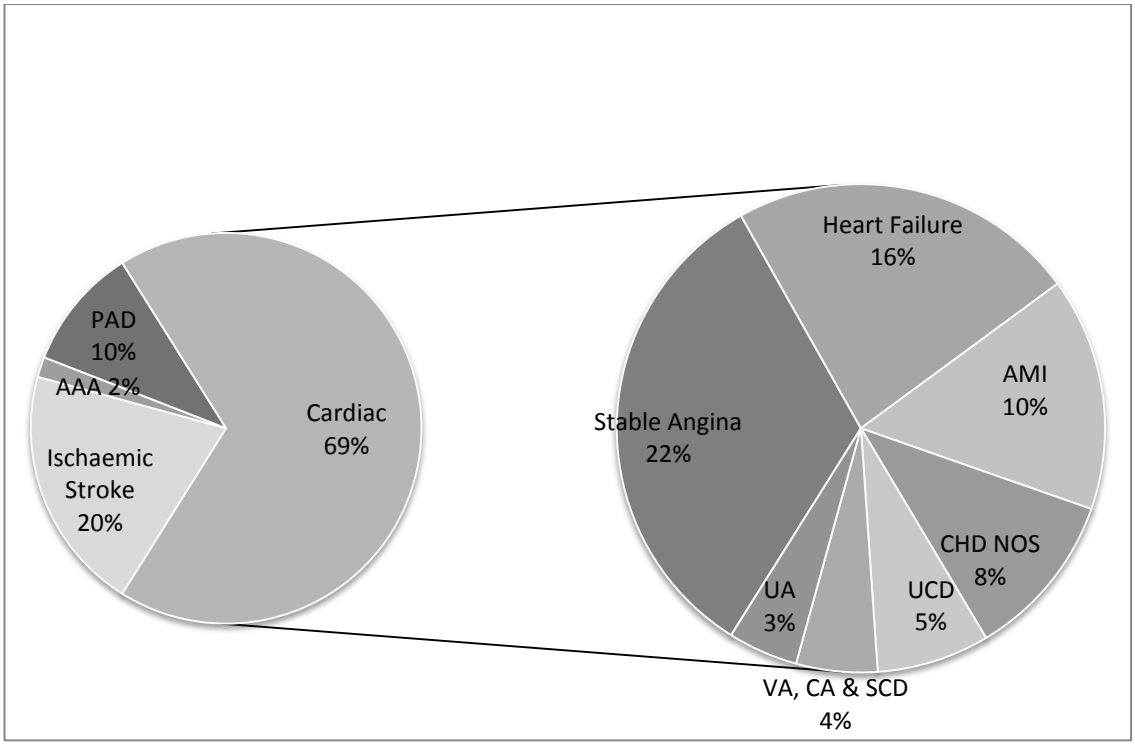
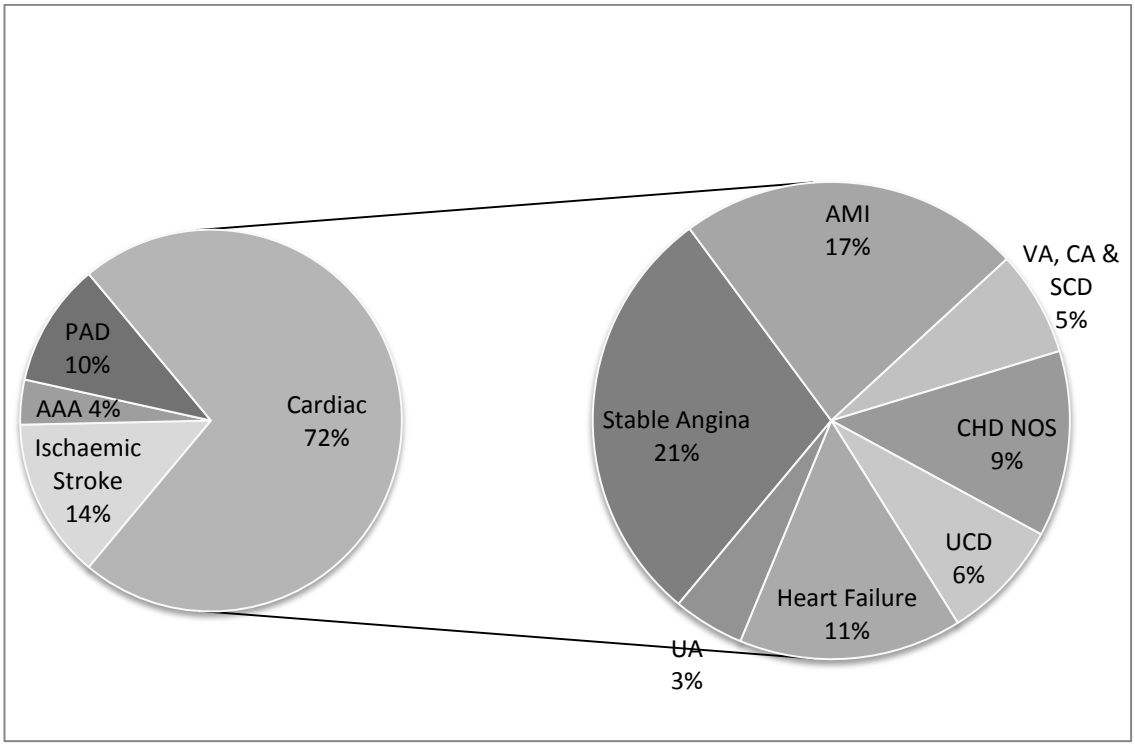


Figure 9: Proportion of initial presentation of CVD for each presentation, overall and in men and women







PAD indicates peripheral arterial disease; AAA, abdominal aortic aneurysm; AMI, acute myocardial infarction; CHD NOS, coronary heart disease not otherwise specified; LVA, CA & SCD, ventricular arrhythmias, cardiac arrest and sudden cardiac death; UA, unstable angina.

Table 11: Number of women and men with each endpoint and proportion fatal within one calendar day

	Endpoints n (% CVD endpoints)		Fatal within 1 calendar day n (% of category)	
Disease	**Women**	**Men**	**Women**	**Men**
Stroke	9,061 (20.3)	6,892 (13.6)	1,293 (14.3)	583 (8.5)
Abdominal aortic aneurysm	747 (1.7)	1,924 (3.8)	209 (28.0)	409 (21.3)
Peripheral arterial disease	4,530 (10.2)	5,317 (10.5)	85 (1.9)	39 (0.7)
Cardiac disease	30,244 (67.8)	36,552 (72.1)	4,101 (13.6)	4,684 (12.8)
Stable angina	9,952 (22.3)	10,533 (20.8)	6 (0.1)	11 (0.1)
Unstable angina	1,403 (3.1)	1,765 (3.5)	5 (0.4)	13 (0.7)
Coronary heart disease not otherwise specified	3,347 (7.5)	4,610 (9.1)	278 (8.3)	313 (6.8)
Myocardial infarction	4,673 (10.5)	8,512 (16.8)	436 (9.3)	481 (5.7)
ST elevation myocardial infarction	282 (0.6)	644 (1.3)	12 (4.3)	8 (1.2)
Non ST elevation myocardial infarction	505 (1.1)	849 (1.7)	3 (0.6)	3 (0.4)
Myocardial infarction not otherwise specified	3,886 (8.7)	7,019 (13.8)	421 (10.8)	470 (6.7)
Heart failure	6,982 (15.7)	5,528 (10.9)	815 (11.7)	517 (9.4)
Ventricular arrhythmias, cardiac arrest & sudden cardiac death	1,627 (3.6)	2,607 (5.1)	301 (18.5)	352 (13.5)
Unheralded coronary death	2,260 (5.1)	2,997 (5.9)	2,260 (100.0)	2,997 (100.0)
Total	44,582 (100)	50,685 (100)	5,688 (12.8)	5,715 (11.3)

9. Rate of initial presentation of CVD with specific endpoints

The rate of initial presentation per 100,000 person years in 10 year age bands was calculated. The rates for most presentations increased up to age 70-79 years and declined thereafter. Stable angina has the greatest increase amongst these presentations up to this age, but both heart failure and stroke as initial presentations continued to rise, indeed rise steeply, by age 80 and did not decline. The rates per 100,000 person years in ten year age bands and overall, are shown in Table 12 below.

10. Patients with missing data

One of the concerns with electronic health record data is the extent of data, particularly risk factor data, which is unrecorded for patients. Although unrecorded data is not missing in the sense that a survey response might be missing – patients have either not attended general practice or not had their risk factors assessed – this data will be referred to as missing henceforth. Below I have compared in greater detail the characteristics of patients with missing data for four key risk factors which have been incorporated in later analyses.

10.1. Index of multiple deprivation not recorded

Very few patients (n= 7,411, 0.4%) had the index of multiple deprivation (IMD) missing, because IMD is missing only in those patients who did not have a valid postcode recorded when the data was linked. Lacking a valid postcode could be due to clerical error, but could also be due to patients being homeless with no fixed address. The mean age of those with IMD missing (46.1 years (45.8-46.4) was marginally younger than those with IMD recorded (47.8 years (47.8- 47.9). Both men and women were just as likely to have their IMD status recorded (women: 52.3% (51.2-53.5%); men: 48.8% (48.7-48.9%)) as not (women: 51.2% (51.1-51.3%); men: 47.8% (46.5-48.8%)). Other differences in terms of smoking status or smoking status recorded, diabetic or not, blood pressure recorded or not, were also negligible.

10.2. Smoking status not recorded

Overall 304,949 (17.3%) patients in the overall cohort had no smoking status recorded in GPRD. They were slightly older than those with smoking status recorded; mean age for those with no smoking status was 49.3 years (95%CI 49.2- 49.3) compared to 47.6 (47.5 - 47.6) in those with data on this risk factor. The greatest difference was in patients aged 80 and over where 29.3% (29.0-29.6) of patients had no smoking status recorded compared to 16.3% (16.2-16.4%) in the 30-39 year olds. Men were much less likely to have their smoking status recorded (61.1% (60.9-61.2) than women (38.9% (38.8-39.1). The proportion of patients with missing smoking status increased somewhat with increasing social deprivation, so 18.4% (18.2-18.5) in the least deprived quintile were missing data

compared to 22.1% (21.9-22.2) in the most deprived quintile. Patients who were diabetic were much more likely to have their smoking status recorded 92.3% (92.1-92.5) than not 7.7% (7.5-7.9). There was also wide variation between practices in the proportion of patients with smoking status missing, ranging from 2.96% to 58.36%.

10.3. Blood pressure not recorded

Of patients with no recorded smoking status, 245,228 (80.4%) also had blood pressure (BP) missing. However, considerably more patients had no recent BP measurement than were missing smoking status; in total 797,234 patients, 45.3% of the overall cohort. Unlike smoking, those with BP missing were younger than those with BP recorded: 46.6 (46.6- 46.7) compared to 48.9 years (48.8-48.9). The proportion of patients with missing BP data was greatest in patients under 50. BP was missing in 60.0% (59.9-60.1) of men compared to 40.0% (39.9-40.1) of women. The proportion of patients in each IMD category was similar in patients with and without missing BP measurements. Ten per cent of patients with diabetes had no BP measurement at baseline. There was variation between practices in the proportion of patients with a BP measurements but this did not vary as widely as the recording of smoking status.

10.4. IMD, smoking and blood pressure not recorded

Overall, 860,692 (49.9%) patients in the overall cohort did not have IMD, smoking and blood pressure recorded at baseline. The mean age of those with missing data was 47.1 years (47.1- 47.2), slightly younger than those with these data recorded (48.5 years (48.5-48.6). The age group with the highest proportion of missing data was 40-49 with 53.2% (53.1-53.4%) missing compared to 48.9% (48.8- 49.1) overall. The greatest difference was in the proportion of men with missing data (58.6% (58.4-58.6) compared to women with missing data (41.5% (41.1-41.6). Those with diabetes were much more likely to have data recorded than not (85.1% (85.1- 85.1) vs 14.9% (14.9- 14.9)), though it is worth noting that over 10% of diabetic patients had no smoking status and no recent blood pressure recorded.

Table 12: Rate of different initial presentations per 100,000 person years of observations in 10 year age bands

Endpoints	30-39	40-49	50-59	60-69	70-79	80-89	90+	All ages
Stroke	12.1 (10.9-13.4)	35.7 (33.3-38.2)	96.2 (91.9-100.7)	262.2 (253.1-271.6)	695.5 (676.0-715.4)	1653.2 (1603.1-1704.9)	2869.1 (2693.4-3056.4)	169.5 (166.9-172.2)
Abdominal aortic aneurysm	0.8 (0.6-1.2)	2.2 (1.7-2.9)	13.7 (12.1-15.4)	62.8 (58.4-67.5)	155.2 (146.1-164.8)	193.1 (176.5-211.3)	140.3 (105.4-186.8)	28.4 (27.3-29.5)
Peripheral arterial disease	9.5 (8.5-10.7)	36.7 (34.3-39.3)	108.0 (103.5-112.7)	226.8 (218.3-235.5)	379.9 (365.6-394.8)	477.1 (450.5-505.2)	462.8 (395.4-541.7)	104.6 (102.6-106.7)
Cardiac disease	61.4 (58.7-64.3)	267.0 (260.4-273.8)	607.7 (596.8-618.7)	1101.7 (1082.9-1120.9)	1641.4 (1611.4-1671.9)	2094.8 (2038.3-2152.9)	2400.4 (2240.1-2572.2)	531.8 (527.2-536.5)
Stable angina	24.7 (23.0-26.5)	113.2 (109.0-117.7)	268.6 (261.4-275.9)	494.7 (482.1-507.5)	654.6 (635.8-674.0)	574.4 (545.2-605.2)	483.7 (414.6-564.2)	217.7 (214.7-220.7)
Unstable angina	7.5 (6.6-8.5)	25.2 (23.2-27.3)	41.0 (38.3-44.0)	64.5 (60.1-69.3)	76.9 (70.6-83.7)	97.8 (86.2-111.0)	104.5 (75.0-145.5)	33.7 (32.5-34.9)
Coronary heart disease not otherwise specified	6.9 (6.0-7.9)	35.8 (33.4-38.3)	105.6 (101.1-110.2)	201.6 (193.7-209.9)	255.7 (244.0-267.9)	266.0 (246.4-287.2)	244.8 (197.2-304.0)	84.6 (82.7-86.4)
Myocardial infarction	18.5 (17.0-20.1)	77.8 (74.3-81.5)	156.4 (150.9-162.0)	258.4 (249.4-267.7)	428.9 (413.7-444.6)	633.9 (603.2-666.2)	713.6 (628.6-810.0)	140.1 (137.7-142.5)
ST elevation myocardial infarction	2.4 (1.9-3.0)	8.0 (6.9-9.3)	12.8 (11.3-14.4)	16.7 (14.6-19.3)	22.2 (19.0-26.0)	25.3 (19.7-32.4)	29.9 (16.1-55.5)	9.8 (9.2-10.5)
Non ST elevation myocardial infarction	2.0 (1.5-2.6)	8.8 (7.6-10.1)	15.8 (14.1-17.6)	27.7 (24.9-30.9)	41.6 (37.0-46.7)	64.0 (54.7-74.8)	47.8 (29.3-78.0)	14.4 (13.6-15.2)
Myocardial infarction not otherwise specified	14.2 (12.9-15.6)	61.1 (58.0-64.3)	127.9 (122.9-133.0)	213.9 (205.7-222.5)	365.1 (351.1-379.7)	544.7 (516.3-574.7)	635.9 (556.0-727.3)	115.9 (113.7-118.1)
Heart failure	5.7 (4.9-6.7)	19.6 (17.9-21.5)	56.5 (53.3-59.9)	181.1 (173.6-189.0)	601.3 (583.3-619.9)	1503.7 (1456.0-1553.0)	2457.1 (2294.9-2630.9)	132.9 (130.6-135.3)
V. arrhythmias, cardiac arrest & sudden cardiac death	10.1 (9.0-11.3)	20.4 (18.6-22.3)	51.7 (48.6-55.0)	106.1 (100.3-112.1)	132.1 (123.8-140.9)	108.0 (95.7-121.8)	89.6 (62.6-128.1)	45.0 (43.7-46.4)
Unheralded coronary death	3.9 (3.2-4.6)	15.0 (13.5-16.6)	36.1 (33.5-38.9)	82.6 (77.6-88.0)	225.3 (214.4-236.8)	522.7 (494.9-552.1)	853.9 (760.4-958.8)	55.9 (54.4-57.4)

11. Conclusions

A total of 1,758,584 met the basic inclusion and exclusion criteria for the overall cohort without reference to the completeness of co-variate data. As a whole the cohort was relatively young, as would be expected in a healthy cohort, and included a slightly higher proportion of women than men. Differences in the measured values of smoking and blood pressure between the current cohort and national averages recorded in the Health Survey for England could be due to missing data but could also be due to the healthy population in the cohort. The level of lipid, alcohol consumption and BMI was disappointly low, so the decision was taken not to include these variables in the main modelling in subsequent chapters. Further work within the Clinical Epidemiology Team to develop robust approaches to multiple imputation to allow the association of these risk factors with initial presentation, as well as make suitable adjustments in modelling, are currently underway.

The comparison between the onset of CVD with specific diseases and the first presentation of a specific disease revealed both gender and age differences. This distinction is important for understanding gender differences in the disease pathways: the initial presentation forms a lower proportion in men compared to women of first presentations across several diseases. This difference is particularly notable with stable angina, where 54.6% of first presentations in men of all ages were also the initial presentation of CVD, while in women the proportion is 68.3%, with more extreme differences in the youngest age group. These differences point to the need to study the patient pathway from onset with any presentation through to subsequent events with greater precision than has been the case in previous research. The importance of a broader approach is reinforced by the proportion of initial presentations which are neither stroke nor acute MI – 69% overall (70% in women and 69% in men). These findings show the importance of including the range of diseases with which CVD can initially present and indicates potentially wider opportunities for secondary prevention than currently exploited.

Given the higher case fatality from stroke and AAA in women than in men, those secondary prevention opportunities may be more limited in women. Further research to determine whether women are receiving similar levels of primary prevention compared to men and whether any such differences might explain these differences in case fatality should be undertaken.

Although the rate of initial presentations increases with age across most CVD subtypes up to age 70-79, the relatively shallow increase below age 70 is perhaps surprising, but may be a reflection of the rate of increase in *initial presentation*. Patients who have a second or third event in later middle age will not be included in these rates. The rapid increase in rate of both heart failure and stroke, starting at 70 and above, is likely a reflection of increasing blood pressure, and the long-term damage caused by that increase, with age. The association of blood pressure with initial presentation as modified by age is investigated in greater detail in Chapter 7.

The examination of patients with data missing on key variables (IMD, smoking status and blood pressure) which are generally well recorded in the dataset showed some patterning by age, but principally by gender. Men were significantly more likely to have smoking and blood pressure data missing than women so that 39.5% of patients with complete data on these variables were male compared to 60.5% female. Further discussion of the specific cohort used in Chapters 5-7 can be found in those chapters.

Gender and the initial presentation of a wide range of cardiovascular diseases: Influence of age, smoking and diabetes

1. Abstract

Background: Women are known to have lower rates of some cardiovascular diseases (CVD), e.g. myocardial infarction (MI) and abdominal aortic aneurysm (AAA), compared to men. However, it is unknown the extent to which gender is differentially associated with the onset of CVD across a wide range of presentations. Based in part on previous research,(2,31) I hypothesised that there would be heterogeneity in the gender ratios between different initial presentations, and that these gender ratios would be stronger at younger ages, and weaker among smokers and people with diabetes.

Objectives: To determine the heterogeneity of association of gender with the onset of CVD across a wide range of cardiovascular disease presentations and to determine whether those associations were modified by age, smoking or diabetes.

Design: Cohort study using data from the CALIBER research platform linking four data sources (primary care, disease registry, hospitalisation and mortality records) for 1,758,584 patients free from symptomatic CVD.

Main outcome measures: Initial presentation of CVD with stable angina, unstable angina, MI, heart failure, ventricular arrhythmias, unheralded coronary death (UCD), stroke, AAA and peripheral arterial disease (PAD).

Results: In both men and women, CVD most commonly presented with stable angina, followed in women by stroke and in men by MI. There was marked heterogeneity in gender effect, with an almost fourfold difference in gender rate ratios. The most extreme were for AAA (4.27 (95%CI 3.92-4.65) and MI (2.51 (2.42-2.60)) with only modest effects for stroke, angina, and heart failure. These rate ratios for MI and UCD were 4 and 6 times greater in the under-40s compared to those 80 and over. Other presentations showed quite different effect modifications with age. Amongst diabetics, the male excess for onset of CVD with all presentations except AAA was reduced, most substantially for the acute coronary presentations. The only clear effect modification for smoking was a reduction in the male excess amongst smokers for initial presentation with AAA.

Conclusion: Onset of symptomatic CVD with the most common presentations (heart failure, stroke, stable angina) show only a modest male preponderance, while other presentations showed either a high (PAD, MI, UCD) or extremely high male preponderance (AAA). The extent of modification of the gender rate ratios by age, smoking status and diabetes varied considerably by initial CVD presentation.

2. Introduction

The previous chapters described the creation of the CALIBER research platform, the general methods used for the thesis and general results applicable to the whole thesis. In this chapter, I examine the association of gender with the initial presentation of cardiovascular disease.

Women are known to have much lower rates of some cardiovascular diseases (CVD), such as myocardial infarction (MI) and abdominal aortic aneurysm (AAA), compared to men. For other CVDs, the incidence ratios are not clear. (194) There are a lack of large scale, prospective studies which examine the way in which gender patterns the initial presentation (non-fatal and fatal) of a wide range of cardiovascular diseases. In studies of first presentation of individual CVD presentations, men have been found to have higher incidence rates than women across a range of individual cardiovascular disease presentations, including stable angina,(195) unstable angina,(144) myocardial infarction,(144,196) ischaemic stroke(26,197), abdominal aortic aneurysm (198,199), heart failure(200,201), sudden cardiac death(202,203) and peripheral arterial disease.(204) However, these studies only investigate the onset or first presentation *within* a single disease presentation and disregard the possibly earlier onset of other CVD presentations. This approach does not give due regard to more generalised nature of atherosclerotic disease, with common risk factors potentially affecting the cerebral, coronary, abdominal and peripheral arterial beds. *The principal aim of this chapter's analyses is to investigate the association of gender* with the onset of CVD, looking at the first presentation of any disease *across* the entire range of CVD presentations.

Although the majority of cardiovascular risk factors have been found to have similar risks for men and women, age, smoking and diabetes appear to modify the effect of gender on at least some cardiovascular disease presentations. Large-scale cohort studies or meta-analyses support these being robust effects. I briefly summarise the evidence for the interaction between these risk factors and gender and the implications for my analyses below.

The interaction between gender and age in the association with specific cardiovascular disease presentations has been documented in a number of studies mainly focussing on single disease presentations (fatal and non-fatal). Taking coronary heart disease (CHD) as an example, women develop coronary heart disease about 10 years later than men, (205) with an up to five-fold greater risk of CHD mortality in men compared to women, at least in middle-age. (206) The difference in risk declines with age, reaching near parity in

people in their 70s or 80s.(207) Although some have attributed the closing of the gap in risk to increase in women's risk after the menopause, women appear to have a steady increase in risk with age, while men have an excess risk at a younger age; it is the decline in risk in men which leads to the reduction in risk ratio at older ages.(208) Stroke appears to follow a similar pattern of male excess declining with age.(26) While other CVD presentations may not have similar interactions between age and gender, there is evidence that they all have gender differences in the mean age of onset. The extent of that difference varies from as little as two years for heart failure,(209) three for peripheral arterial disease,(28) four for stroke,(26) five for stable angina(210) and ruptured abdominal aortic aneurysm,(211) six for sudden cardiac death,(212) to ten years for acute myocardial infarction (205) and CHD mortality.(213) Given these differences, age is likely to modify the association of gender with the range of specific initial presentations of cardiovascular disease. The hypothesis that there is no effect modification by age will be tested in the analysis.

Diabetes has also been shown, in large robust meta-analyses, to modify the effect of gender on coronary heart disease mortality, with a greater increase in the risk of CHD mortality in women with diabetes compared to the increase men,(31,214,215) to the extent that some authors have suggested that the male excess in risk of CHD is negated by diabetes.(216) Although meta-analyses are lacking, diabetes has been shown to have a greater effect on women's risk of ischaemic stroke compared to men's,(217) but no effect on the relative risk of sudden cardiac death.(218) Similarly, smoking has been found to pose a greater risk for women than for men for CHD(2), stroke,(39) and possibly PAD.(204) However, it is not clear whether the effect modification of gender by diabetes and smoking extends to a) initial presentation of cardiovascular disease rather than CVD mortality, and b) across the range of different possible CVD presentations. The hypothesis that there is no effect modification of gender by either risk factor will be tested in the analyses.

My cohort study (2001-2010) takes place at a time of continuing reductions in the rate of cardiovascular mortality in both men and women,(219) but there is some evidence that the rate of decline in women, particularly younger women, has been reversed.(220) To take an extreme example, if the rate of a specific initial presentation declined in men but remained constant in women over a given time period, the rate ratio would reduce over this time period. I have therefore also investigated whether the rate ratios did indeed decreased between three equal time periods, 2001-03, 2004-06 and 2007-09.

This chapter therefore has four key objectives:

1. To summarise the existing literature on the association of gender with the onset of CVD across a wide range of disease presentations.
2. To determine whether the risks comparing men to women for the onset of CVD vary across a wide range of cardiovascular disease presentations.
3. To test whether any gender differences found are modified by age, diabetes and smoking of patients.
4. To determine whether there is a declining trend in any gender differences for the range of disease presentations over the time period of the cohort.

3. Literature review

3.1. Search strategy

The aim of my literature search for this chapter was to find studies which a) reported the gender-specific rate of onset of CVD with specific cardiovascular presentations in populations without clinically manifest cardiovascular disease and b) which included at least two endpoints. I used a specific search strategy focussing on the rate of initial CVD presentation. The search terms I used are reproduced in Table 22 below. I limited both searches to papers in English with the full text available, published since 1992 in PubMed. I supplemented the studies found through the formal search strategy by searching reference lists of relevant papers, using forward citation searches of earlier relevant papers and asking other researchers for the details of any papers of which they were aware. Relevant studies published before 1992 which were identified have also been included.

Table 13: Search strategy used to identify studies on gender specific rate of onset of CVD with specific cardiovascular presentations

Concepts	Boolean operator	Terms used
Specific search		
Rate		(Incidence OR incident) rate
Endpoints	AND	"myocardial infarction" OR STEMI OR nSTEMI OR "acute coronary syndrome" OR “angina” OR “coronary insufficiency” or “acute coronary syndrome” OR “heart attack”
		“abdominal aortic aneurysm” OR “peripheral arterial disease” OR “peripheral vascular disease” OR PAD OR “intermittent claudication” OR “peripheral ischaemia”
		Stroke OR “cerebrovascular disease” OR “cerebrovascular accident”
		(ventricu* AND arrhythm*) OR “cardiac arrest” OR “sudden cardiac death”
		“heart failure”
Gender	AND	“sex factors”[MESH] OR (sex OR gender) OR (women AND men) OR (male AND female)

3.2. Findings of literature search

Five studies which described the rate of onset of CVD across several endpoints were found. Only one small study, limited by cohort size (n=931) but not the number of endpoints investigated (n=6), could be found which reported rates of onset of CVD with different cardiovascular diseases in both men and women,(221) although Murabito et al. also reported the proportion of different presentations amongst participants who developed CHD in the Framingham cohort.(1) Another four studies reported rates in men only.(222–225) The five studies of the overall onset of CVD are all bespoke cohort studies.

I also identified a number of other studies which estimated gender-specific incidence rates for several cardiovascular diseases measured concomitantly, where prior manifestations with other cardiovascular disease types were ignored.(84,144,195–197,226,227) These studies provide comparative rates, within the same population, of the cardiovascular diseases of interest to this thesis, even if they do not report the onset of CVD with specific presentations. These additional studies of onset of specific cardiovascular diseases include bespoke cohort studies,(144,195–197,227) electronic health record studies,(84) and combinations of the two.(226) I summarised the findings of these studies in Table 14.

From this review, it is clear that the evidence of gender differences in the rate of onset of CVD overall with different specific diseases is limited. The single study found which included both men and women found a higher rate of acute myocardial infarction (AMI) in men than in women, with similar rates of stable angina, heart failure, stroke and peripheral arterial disease. However, because of the small cohort size, the confidence intervals are wide, so differences may have been missed. The studies of first presentations of specific cardiovascular diseases (regardless of previous differing CVD disease presentations) appear to show differing rate ratios (women compared to men) across a limited range of first presentations, with greatest male excess in rates of AMI.(84,144,195,227–229) Single studies also found a male excess in rates of unstable angina(144) and heart failure.(196)

Table 14 : Age-adjusted rate per 1,000 person years for a range of specific cardiovascular diseases in men and women

Study Details					Cardiovascular Disease Presentations Age-adjusted rate per 1,000 person years (95% confidence intervals)								
First Author & publ. year	**Cohort size**	**Sex**	**Age at entry**	**CVD Events**	**Stable Angina**	**Unstable Angina**	**Acute MI**	**Unher'd Coronary Death**	**Vent Arrhyth .**	**Heart Failure**	**Stroke**	**AAA**	**PAD**
Studies of overall onset of cardiovascular disease with specific presentations													
Baena-Diez 2010(221)[a]	931	W	35-84	71	2.2 (1.0-3.8)		1.2 (0.3-2.0)			2.7 (1.4-4.0)	3.3 (1.9-4.6)		1.8 (0.7-2.9)
		M		63	2.4 (0.9-3.9)		6.1 (3.6-8.5)			2.2 (0.9-3.5)	2.0 (0.6-3.4)		1.4 (0.4-2.4)
Canoui-Poitrine 2009(222)	9,758	M	50-59	292	2.4 (2.0-2.8)		5.0 (4.2-5.7)						
Glynn 2005 (223)	18,662	M	40-84	2,250			4.1				2.7		
Ducimetiere 2001(224)	7,359	M[b]	50-59	200	2.6 (2.1-3.1)		2.9 (2.4-3.5)						
	2,399	M[c]		121	5.4 (4.1-6.7)		5.2 (3.9-6.6)						
Dagenais 1990(225)	4,576	M	35-64	603	6.7		4.7	2.2					

Internally age-standardised rate per 1,000 person years unless otherwise specified. A merged cell indicates composite endpoint. CVD indicates cardiovascular disease; MI, myocardial infarction; unher'd, unheralded, vent arrhyth, ventricular arrhythmias, cardiac arrest and sudden cardiac death; AAA, Abdominal aortic aneurysm; PAD, peripheral arterial disease; M, Men; W, Women; confidence intervals omitted where confidence intervals not specified in paper. [a] *standardised to European population* [b] *France* [c] *Ireland. Where no confidence intervals are given, none were reported in the study.*

Study Details					Cardiovascular Disease Presentations Age-adjusted rate per 1,000 person years (95% confidence intervals)								
First Author & publ. year	**Cohort size**	**Sex**	**Age at entry**	**CVD Events**	**Stable Angina**	**Unstab Angina**	**Acute MI**	**Unher'd Coronary Death**	**Vent Arrhyth.**	**Heart Failure**	**Stroke**	**AAA**	**PAD**
Studies of first presentation of specific cardiovascular disease presentations													
Bhattarai 2012(84)	300,020	W	30+	12,495	0.6[d]		0.6						
		M			1.0[d]		1.5						
Merry 2009(144)	21,148	W	20-69	277		1.0	0.9			0.5			
		M		692		2.6	3.5			0.9			
Ishikawa 2008(197)[e]	12,388	W	20-69	153			0.3				1.4		
		M		229			0.8				2.3		
Silventoinen 2008(226)	1,145,758	M	16-25	65,611			0.2				0.3		
Arnold 2005(196)[f]	5,888	W	65+	1,136			9.4 (8.2-10.6)			18.4 (16.8-20.0)	13.7 (12.3-15.1)		
		M		1,094			19.3 (17.2-21.4)			29.2 (26.7-31.7)	14.7 (12.9-16.5)		
Rothwell 2005(227)[g]	91,106	W	35-85	348		0.3	0.3		0.3		1.5	0.1	0.2
		M		437		0.3	0.7		0.3		1.4	0.2	0.2
Haan 1996(195)[h]	3032	W	65+	446	7.1		4.8			10.6	10.1		
		M		447	13.4		6.3			9.1	10.0		

Internally age-standardised rate per 1,000 person years unless otherwise specified. A merged cell indicates composite endpoint. CVD indicates cardiovascular disease; MI, myocardial infarction; unher'd, unheralded, vent arrhyth, ventricular arrhythmias, cardiac arrest and sudden cardiac death; AAA, abdominal aortic aneurysm; PAD, peripheral arterial disease; M, Men; W, Women; confidence intervals omitted where confidence intervals not specified in paper. d incudes unstable angina and acute coronary syndrome; e rate in 2009 standardised to Japanese population of 1985; f overall crude rates after 10 years of follow-up; g crude rates, AMI = rate for STEMI; PAD = critical limb ischaemia; Vent arrhyth = sudden cardiac death only; h Age-standardised rate calculated from age-specific rates published in paper. Where no confidence intervals are given, none were reported in the study.

4. Methods

The data sources, population, risk factor and endpoint definitions are described in detail in Chapters 2 and 3, but are briefly summarised here for ease of reference. Any methodological details specific to the set of analyses in this chapter are also noted here.

4.1. Data sources

The data sources are described more fully in Chapter 2. In brief, the Cardiovascular disease research using Linked Bespoke studies and Electronic Records (CALIBER) e-health research platform links NHS primary care data from the General Practice Research Database (GPRD),(76) to data for acute coronary syndrome (ACS) admissions from the Myocardial Ischaemia National Audit Project registry (MINAP),(98) NHS hospital admissions data from Hospital Episodes Statistics (HES),(230) and mortality and social deprivation data from the Office of National Statistics (ONS).(113,123) Records were primarily linked using a pre-defined deterministic linkage algorithm based on NHS number, with a small minority linked using a probabilistic method using DOB and postcode.(77) A web-based portal documenting the creation of all CALIBER data items, from these multiple data sources, is available at www.caliberresearch.org and further details on the creation of the CALIBER platform have been published elsewhere.(231)

4.2. Population

The identification of the general cohort is described in detail in Chapter 3. All patients in the overall cohort had a valid age and gender recorded. For the main analysis in this chapter, I used the overall thesis cohort (overall n=1,758,584; women=900,401; men=858,183). For the analysis of effect modification by smoking status, the cohort was restricted to those with smoking status recorded at baseline, 82.7% of the overall cohort (women=781,634, men=672,001). More detail on patients excluded because of missing smoking data is given in Chapter 4. Because patients were defined as diabetic if there was a positive record in GPRD indicating diabetes and defined as non-diabetic if there was no such record, there were, by definition, no patients with missing diabetic status. Therefore the overall cohort was used for the analysis of effect modification by diabetes.

4.3. Risk Factors:

Gender

Gender was defined as the gender recorded in the GPRD patient data file.

Other risk factors

Age at entry was defined as the age in years in January of the first year of endpoint follow-up; more precise estimates of age were not possible as only year of birth is supplied with

the data to protect patient identities. Smoking status was defined as the GPRD record of smoking status with the last possible date before endpoint follow-up and categorised as *non-smoker*, *ex-smoker*, or *current smoker*. If non-smokers had a previous record indicating smoking in their entire GPRD history, they were counted as an ex-smoker. Diabetes mellitus was defined as a diagnosis of Type 1, Type 2 or unspecified diabetes or at least one prescription for insulin or oral hypoglycaemic agent in the two years before the start of endpoint follow-up.

Other risk factors were not used in the analyses in this chapter, but baseline values are presented to aid with interpretation and understanding of the analyses in this chapter. Except for ethnic group, all these risk factors were taken from GPRD data. Social deprivation was measured by the index of multiple deprivation (IMD) 2007,(123) dividing IMD into quintiles. Ethnic group was categorised as *White*, *Black*, *South Asian* or *Other*, using self-reported ethnic group recorded in GPRD and HES, with unresolvable code conflicts between the two data sources recorded as missing. Systolic blood pressure (SBP) was defined as the mean SBP recorded in the two years prior to the start of endpoint follow-up. For all remaining risk factors, I used the most recent record from GPRD in the two years before the start of endpoint follow-up. Total cholesterol level was taken from laboratory results from plasma or serum samples, recorded mmol/L units. Lipid-lowering medication at baseline was derived from one or more prescriptions for a statin. Body mass index (BMI) was defined as weight in kilograms over height in metres squared. The number of different blood-pressure-lowering medication classes prescribed was based on prescriptions categorised into thiazides, potassium-sparing diuretics, beta-blockers, ace inhibitors and other less common medications. A complete list of drug classes and specific preparations is available in Appendix D.

4.4. Endpoints: Initial symptomatic presentations of cardiovascular disease

The definition of my primary endpoints are described in more detail in Chapter 3, with additional information on the contribution of the different data sources to each endpoint described in Chapter 4. To summarise briefly, my primary endpoints were fatal and non-fatal presentations of a range of cardiovascular diseases encompassing coronary heart disease (stable angina, unstable angina, ST-elevation myocardial infarction (STEMI), non-ST-elevation myocardial infarction (NSTEMI), myocardial infarction not otherwise specified (MI NOS), and coronary death unheralded by prior symptomatic disease), stroke, peripheral arterial disease (PAD), abdominal aortic aneurysm (AAA), heart failure, and ventricular arrhythmias including cardiac arrest and sudden cardiac death (SCD). Given the limited number of myocardial infarctions specified as STEMI or NSTEMI, the

main analyses used a composite endpoint of all acute myocardial infarctions (AMI) as an endpoint. Secondary analyses with specific myocardial infarction types have also been presented. For ease of comparison with previous studies, a composite endpoint of AMI and unheralded coronary death has also been included. Diagnoses were identified using codes from the International Classification of Diseases 10th Revision (ICD 10)(232) for the hospital data (HES) and mortality data (ONS), from Read Codes(78) for primary care data (GPRD) and bespoke variables in the ACS registry (MINAP).

4.5. Statistical analysis

My primary focus in this chapter was to investigate the heterogeneity of the association of gender with a range of initial presentations of CVD. I first described the gender-specific rates, using person-time at risk as the denominator, standardised to the internal age structure for the whole cohort, using single years of age at entry, for all initial presentations. For each presentation, person years were defined as the time from each patient's study entry until the onset of CVD with one of the initial presentations or until the patient left the practice or the date of the last practice data download. Confidence intervals for the rates were calculated using the method described by Breslow and Day.(233)

I then estimated the rate ratios per 1,000 person years for each presentation, comparing men to women, using Poisson regression. I adjusted for age at entry only. To investigate whether any of these associations were modified by age, I estimated the gender rate ratios for each presentation stratified by 10 year age bands (30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90+). The same stratified approach was used to investigate effect modification by smoking status (non-smoker, ex-smoker and current smoker) and diabetes (non-diabetic, diabetic).

To investigate whether the gender rate ratios declined over time, I divided the cohort into three time periods (2001-03, 2004-06, 2007-09) and calculated the rate ratios for each time period. Patient-time was allocated to the relevant decade, so patients could contribute time to all three periods, if they entered the study in the first and did not have a presentation or leave the cohort until the last.

Continuous variables are shown as the mean (standard deviation) and categorical variables shown as frequencies (percentage). All analyses were performed using STATA version 12 (StataCorp, 4905 Lakeway Drive, College Station, TX 77845, United States).

5. Results

For the main analyses in this study, there was a cohort of 1,758,584 people, with a total of 9,397,192 person years of observation time (women: 4,879,503 person years; men: 4,517,689 person years) and a median of 5.46 person years (Interquartile range (IQR) 2.08-9.09).

5.1. Baseline characteristics

The mean age in women was 49.0 years (standard deviation (sd) 16.0 years) and 46.6 (sd 14.0) in men. Further baseline characteristics in five age categories and overall are given in Table 15 below. The proportions/means were calculated for patients with complete data for each co-variate; where this was less than the complete cohort used in this chapter, the number of patients with data for a given co-variate is shown at the base of the Table.

The cohort was young, as would be expected in a healthy cohort, with women on average five years older than men. At all ages, men were more likely to be either smokers or ex-smokers and diabetic. Below the age of 60, men had higher blood pressure than women, with blood pressure increasing in women until they had higher mean blood pressure at aged 70 and over. Even at higher ages, women were less likely to be on blood pressure medication. Otherwise, women and men had broadly similar characteristics, including mean BMI and mean total cholesterol.

Table 15- Baseline characteristics of patient cohort, in different age groups at baseline and overall

	<50 years		50-59 years		60-69 years		70-79 years		80+ years		All ages	
	Women	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**
Number	519,952	543,331	155,990	153,875	103,675	91,221	49,166	71,816	48,968	20,590	900,401	858,183
Person-time in years, median (IQR)	4.9 (1.8-9.1)	4.6 (1.8-9.1)	8.3 (3.3-9.1)	7.6 (3.1-9.1)	7.7 (3.0-9.1)	6.6 (2.7-9.1)	6.6 (2.7-9.1)	5.5 (2.3-9.1)	3.2 (1.4-6.6)	3.2 (1.3-6.4)	5.7 (2.1-9.1)	5.2 (2.1-9.1)
Ethnic Group, n (%)												
White	256,576 (86.7)	186,488 (86.0)	76,690 (93.4)	68,100 (94.0)	58,249 (94.4)	50,007 (95.0)	45,174 (96.6)	30,747 (96.1)	30,163 (98.3)	12,665 (97.8)	466,852 (90.3)	348,007 (90.0)
Black	12,778 (4.3)	9,077 (4.2)	1,403 (1.7)	1,098 (1.5)	1,069 (1.7)	797 (1.5)	437 (0.9)	391 (1.2)	83 (0.3)	71 (0.5)	15,770 (3.0)	11,434 (3.0)
S. Asian	11,188 (3.8)	9,506 (4.4)	1,728 (2.1)	1,462 (2.0)	1,109 (1.8)	939 (1.8)	537 (1.1)	456 (1.4)	135 (0.4)	90 (0.7)	14,697 (2.8)	12,453 (3.2)
Deprivation quintile, n (%)												
Least	105,181 (20.3)	105,636 (19.5)	33,332 (21.4)	32,784 (21.4)	20,907 (20.2)	18,494 (20.3)	13,456 (18.8)	9,614 (19.6)	8,810 (18.1)	3,775 (18.4)	181,686 (20.3)	170,303 (19.9)
Most	108,269 (20.9)	118,994 (22)	25,526 (16.4)	26,646 (17.4)	17,549 (17.0)	15,599 (17.2)	13,498 (18.9)	9,005 (18.4)	9,366 (19.2)	3,963 (19.3)	174,208 (19.4)	174,207 (20.4)
Smoking status, n (%)												
Non-smoker	277,628 (60.1)	217,280 (51.5)	83,425 (61.1)	59,223 (48.9)	58,442 (64.7)	36,466 (48.9)	40,272 (68.6)	20,511 (51.8)	26,292 (76.7)	8,120 (54.5)	486,059 (62.2)	341,600 (50.8)
Ex-smoker	62,358 (13.5)	55,081 (13.1)	20,562 (15.1)	24,235 (20.0)	15,139 (16.8)	19,681 (26.4)	10,022 (17.1)	11,780 (29.8)	5,091 (14.8)	4,598 (30.9)	113,172 (14.5)	115,375 (17.2)
Current smoker	121,697 (26.4)	149,524 (35.4)	32,614 (23.9)	37,555 (31.0)	16,786 (18.6)	18,470 (24.8)	8,395 (14.3)	7,304 (18.4)	2,911 (8.5)	2,173 (14.6)	182,403 (23.3)	215,026 (32.0)

	<50 years		50-59 years		60-69 years		70-79 years		80+ years		All ages	
	Women	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**
Blood pressure, mean (sd)												
SBP in mmHg	119.3 (13.7)	128.5 (14.4)	134.0 (17.2)	137.4 (16.7)	142.3 (17.3)	143.0 (16.8)	149.8 (17.4)	147.7 (17)	151.2 (19.6)	148.3 (18.6)	128.4 (19.4)	134.2 (17.1)
BP medication classes, n (%)												
None	501,646 (96.5)	527,896 (97.2)	132,263 (84.8)	134,837 (87.6)	76,469 (73.8)	70,484 (77.3)	45,568 (63.5)	33,833 (68.8)	29,748 (60.7)	14,121 (68.6)	785,694 (87.3)	781,171 (91.0)
1	15,807 (3.0)	12,553 (2.3)	17,805 (11.4)	13,922 (9)	18,599 (17.9)	14,461 (15.9)	17,173 (23.9)	10,579 (21.5)	13,222 (27.0)	4,653 (22.6)	82,606 (9.2)	56,168 (6.5)
2	2,089 (0.4)	2,204 (0.4)	4,925 (3.2)	3,948 (2.6)	6,830 (6.6)	4,691 (5.1)	7,015 (9.8)	3,674 (7.5)	4,688 (9.6)	1,393 (6.8)	25,547 (2.8)	15,910 (1.9)
3	410 (0.1)	678 (0.1)	997 (0.6)	1,168 (0.8)	1,777 (1.7)	1,585 (1.7)	2,060 (2.9)	1,080 (2.2)	1,310 (2.7)	423 (2.1)	6,554 (0.7)	4,934 (0.6)
Other risk factors												
Total cholesterol in mmol/L, mean (sd)	5.1 (1.0)	5.4 (1.2)	5.7 (1.1)	5.5 (1.1)	5.8 (1.2)	5.4 (1.1)	5.8 (1.2)	5.2 (1)	5.6 (1.2)	5.0 (1.0)	5.5 (1.1)	5.4 (1.1)
On statins, n (%)	2,307 (0.5)	5,323 (1.0)	4,616 (3.4)	6,207 (4.0)	7,132 (7.9)	6,185 (6.8)	4,803 (8.2)	3,045 (6.2)	1,454 (4.2)	536 (2.6)	20,312 (2.6)	21,296 (2.5)
Diabetes mellitus, n (%)	6,373 (1.2)	7,125 (1.3)	3,803 (2.4)	5,509 (3.6)	4,770 (4.6)	5,683 (6.2)	4,294 (6.0)	3,924 (8.0)	2,717 (5.5)	1,427 (6.9)	21,957 (2.4)	23,668 (2.8)
BMI in kg/m^2, mean (sd)	25.9 (5.9)	26.6 (4.7)	27.2 (5.8)	27.6 (4.7)	27.4 (5.6)	27.3 (4.4)	26.7 (5.3)	26.5 (4.1)	24.7 (4.8)	25.1 (3.9)	26.3 (5.8)	26.8 (4.6)

Overall N= 1,758,584. N for variables with missing data: Ethnic group n= 904,054; IMD n= 1,751,173; smoking n=1,453,635; SBP n= 961,350; total cholesterol n= 203,372; BMI n= 680,753. IQR indicates inter-quartile range; SBP, systolic blood pressure; BP, blood pressure; BMI, body mass index.

5.2. Rates of initial presentations of cardiovascular disease

There were a total of 77,957 diagnoses, 41,401 in men and 36,556 in women. The number of each presentation is shown in Table 16 below. The age-standardised rates of initial presentation across the range of endpoints are shown in Table 16 below. Across all presentations, men had higher age-standardised rates than women. Amongst the cardiovascular presentations in the head, heart, abdomen and peripheral arterial beds, stroke was the most common presentation in women, followed by AMI combined with unheralded coronary death, while in men the rate of AMI was higher than that of stroke. Amongst the cardiac presentations, stable angina was the most common presentation in both men and women, while the second most common presentation in women was heart failure and AMI in men. Unstable angina was the least common initial cardiac presentation for both men and women. The greatest gender differences were in the rate of presentation with AMI (and consequently AMI with unheralded coronary death), with men having more than double the rate of women. There is almost no difference in women between the rates of initial presentation with stroke and with stable angina, while in men the rate of presentation with AMI and stable angina are both significantly higher than initial presentation with stroke. The least common presentation in women was AAA, at approximately 10% of the rate of stable angina, the most common presentation. In men, the least common presentation was unstable angina, approximately 15% of the rate of stable angina.

Table 16: Age-standardised rates per 1,000 person years with 99% confidence intervals across wide range of initial presentations of cardiovascular disease

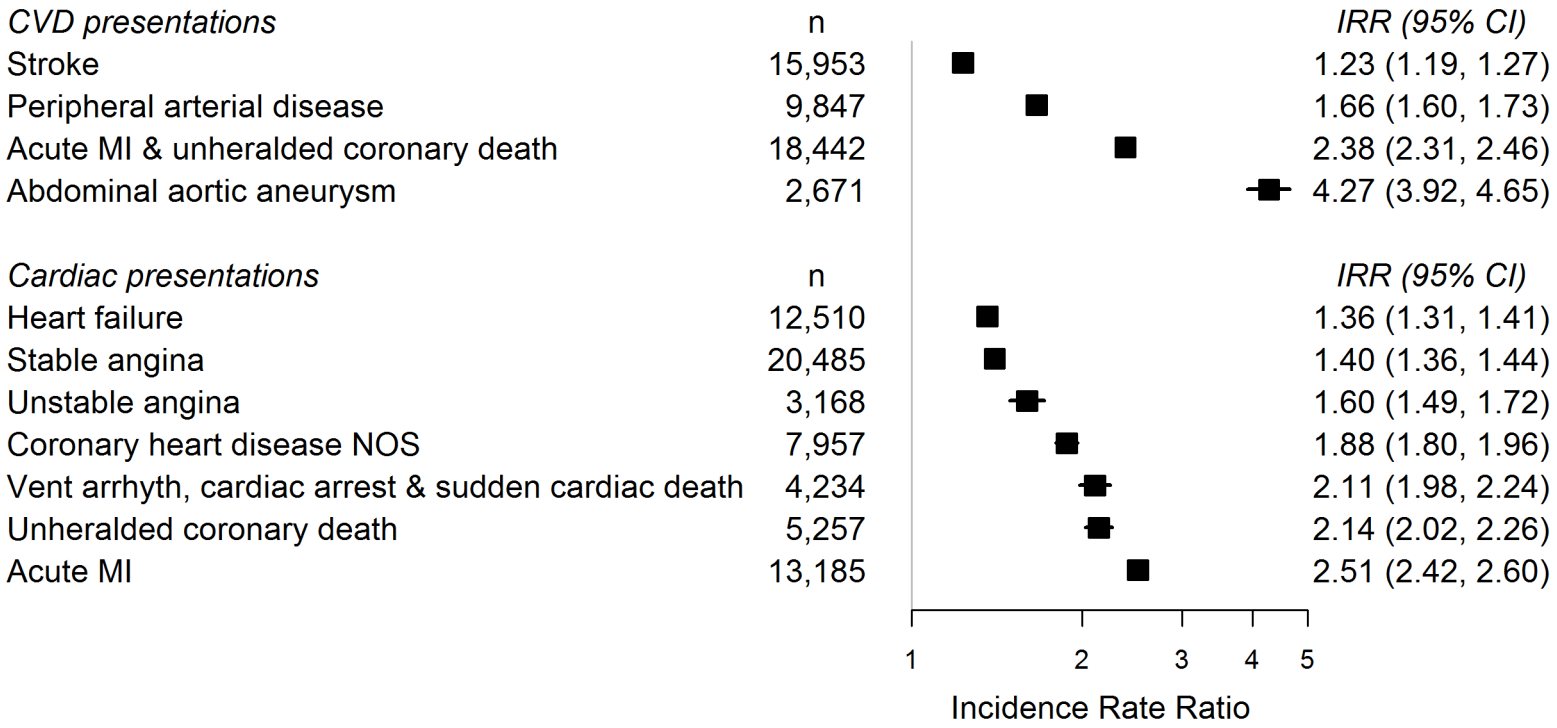
Initial Presentation	Women		Men	
	n	Rate (99% CI)	n	Rate (99% CI)
CVD presentations				
Stroke	6,988	1.71 (1.69-1.73)	5,489	1.93 (1.90-1.96)
Acute myocardial infarct. & unheralded coronary death	5,458	1.27 (1.25-1.29)	9,102	2.80 (2.77-2.84)
Peripheral arterial disease	3,855	0.81 (0.79-0.83)	4,455	1.25 (1.23-1.28)
Abdominal aortic aneurysm	606	0.13 (0.13-0.14)	1,592	0.52 (0.50-0.53)
Cardiac presentations				
Stable angina	8,695	1.77 (1.74-1.79)	8,846	2.35 (2.31-2.38)
Heart failure	5,485	1.31 (1.29-1.33)	4,460	1.69 (1.66-1.72)
Acute myocardial infarct. (all)	3,853	0.85 (0.83-0.86)	6,835	1.96 (1.93-1.99)
ST-elevation myocardial infarct.	219	0.05 (0.05-0.05)	503	0.14 (0.13-0.15)
Non ST-elevation myocardial infarct.	428	0.09 (0.09-0.10)	684	0.19 (0.18-0.20)
Myocardial infarction NOS	3,206	0.70 (0.69-0.72)	5,648	1.63 (1.60-1.66)
Coronary heart disease NOS	2,888	0.59 (0.58-0.61)	3,832	1.03 (1.01-1.05)
Unheralded coronary death	1,605	0.42 (0.41-0.44)	2,267	0.84 (0.82-0.86)
Ventricular arrhythmias, cardiac arrest & sudden cardiac death	1,357	0.30 (0.28-0.31)	2,146	0.58 (0.57-0.60)
Unstable angina	1,224	0.25 (0.24-0.26)	1,479	0.39 (0.38-0.40)

Gender-specific rates per 1,000 person years at risk, standardised to internal age structure for men and women combined, using single years of age at entry. Rates ordered from highest to lowest in within CVD presentations and within specific cardiac presentations using the rates in women. CVD indicates cardiovascular disease; NOS, not otherwise specified.

5.3. Rate ratios of initial presentations of cardiovascular disease

The gender rate ratios (RR) for each CVD presentation, comparing women to men, are presented in Figure 10 below. All RRs show a significant male excess across the range of CVD presentations. The highest RR was for AAA, while the lowest was for stroke, with considerably heterogeneity in effect of gender, with a rate ratio for AAA almost 4 times the rate ratio for stroke. Within the cardiac presentations, acute MI had almost twice the RR as the chronic presentations of heart failure and stable angina.

Figure 10: Age-adjusted gender rate ratios (men compared to women) for initial presentation of a range of cardiovascular disease presentations

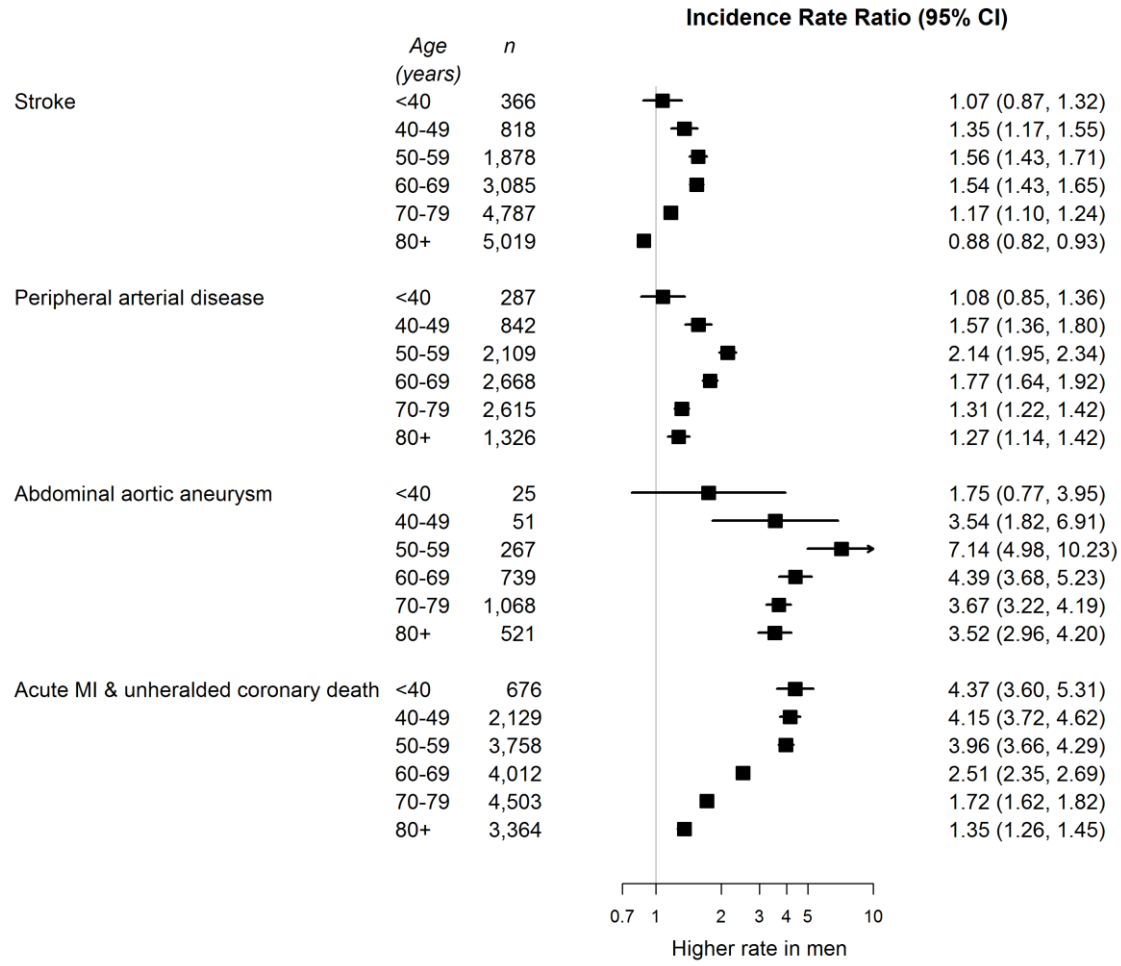


N=1,758,584. CI indicates confidence interval; MI, myocardial infarction; NOS, not otherwise specified; CVD, cardiovascular disease; RR, gender rate ratios; Vent arrhyth, ventricular arrhythmias.

5.4. Effect modification of gender rate ratios by age

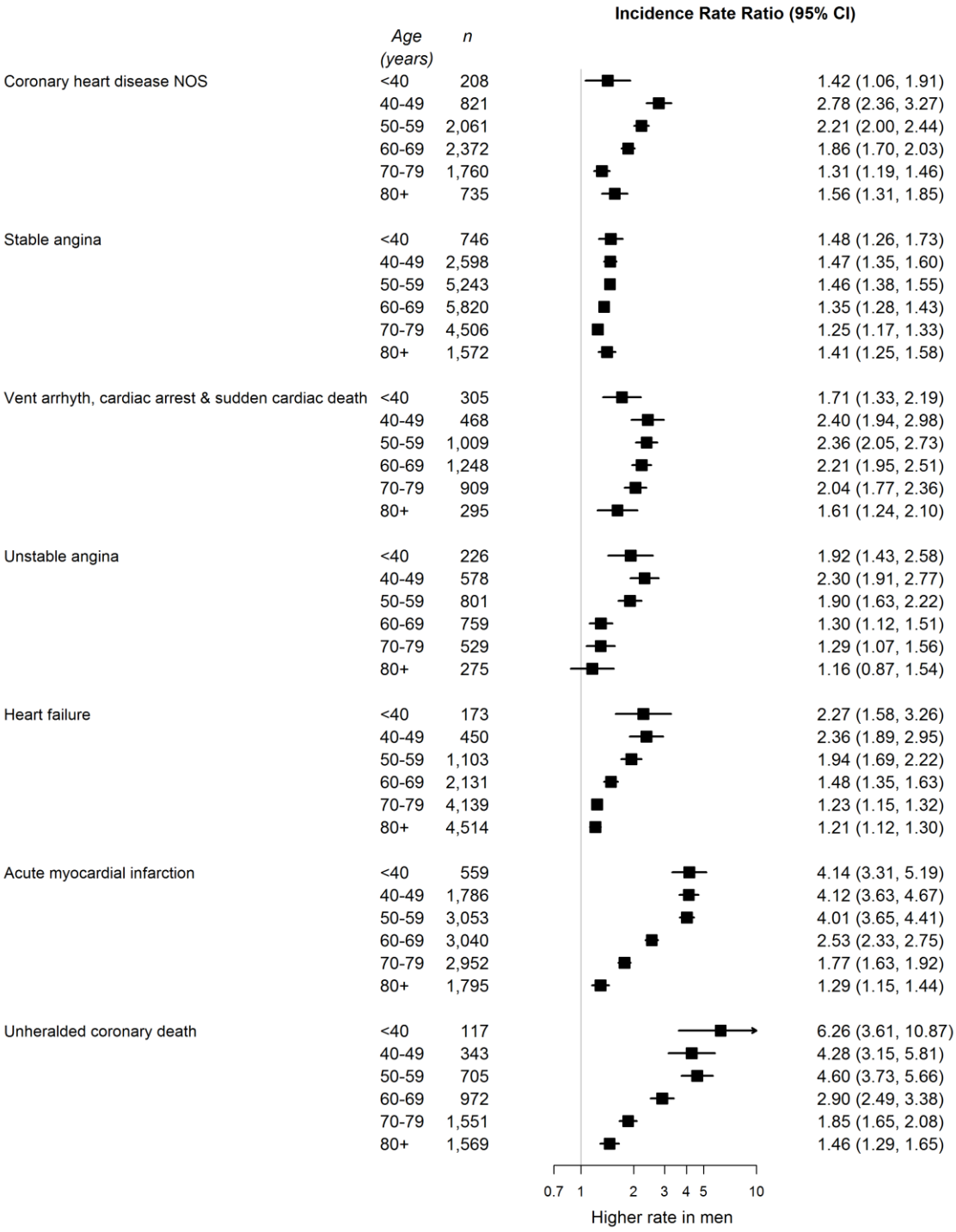
The age-stratified gender rate ratios for CVD presentations are shown in Figure 11 below, while those for the cardiac presentations are shown in Figure 12. Additional results for the myocardial infarction types are shown in Appendix F. The extent of effect modification by age varied by initial presentation. There was no effect modification by age for stable angina. There was a U-shaped relationship between age and gender for stroke, PAD, the ventricular arrhythmias and possibly for CHD NOS, with minimal gender differences in the rate of initial presentation in the youngest age group, increasing male excess through middle age and declining gender differences in the older age groups. For the remainder of the cardiac presentations (heart failure, AMI including all subtypes, unheralded coronary death and to a lesser extent unstable angina), the male excess was greatest in the youngest age groups and declined from 60 and over, if not at younger ages. The relationship between age, gender and initial presentation with AAA did not fit clearly into any of these patterns but bore similarities to the U-shaped pattern, with minimal gender differences in the youngest age groups and the male excess increasing with increasing age; however, AAA showed no decline in the male excess in the oldest age groups. However, the number of AAA presentations, especially at the younger ages, was very small.

Figure 11: Gender rate ratios (men compared to women) for initial presentation of cardiovascular disease, stratified by age group



N=1,758,584; CI indicates confidence interval; MI, myocardial infarction.

Figure 12: Gender rate ratios (men compared to women) for initial presentation of specific cardiac presentations, stratified by age group

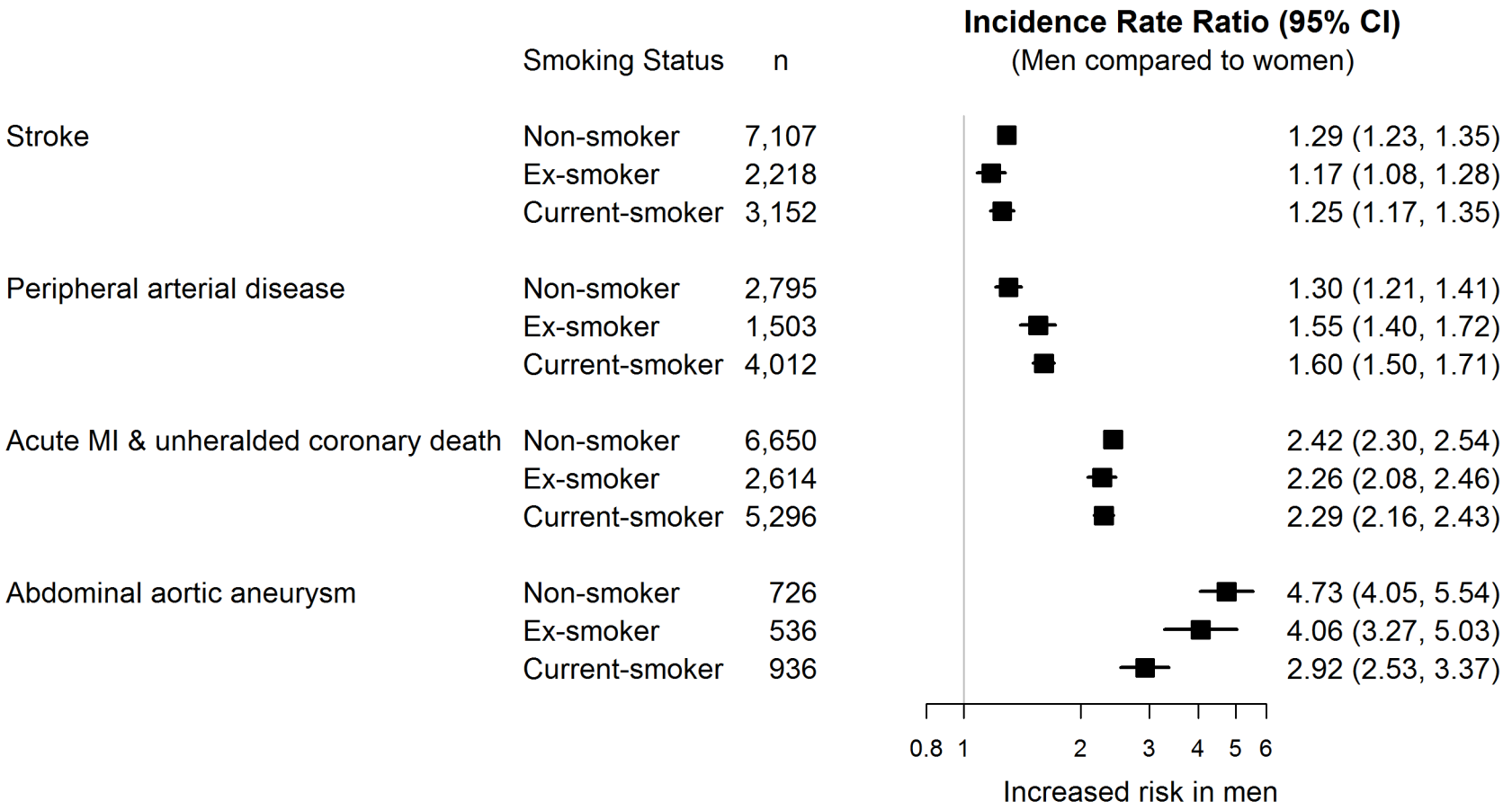


CI indicates confidence interval; NOS, not otherwise specified; Vent arrhyth, ventricular arrhythmias.

5.5. Effect modification of gender rate ratios by smoking

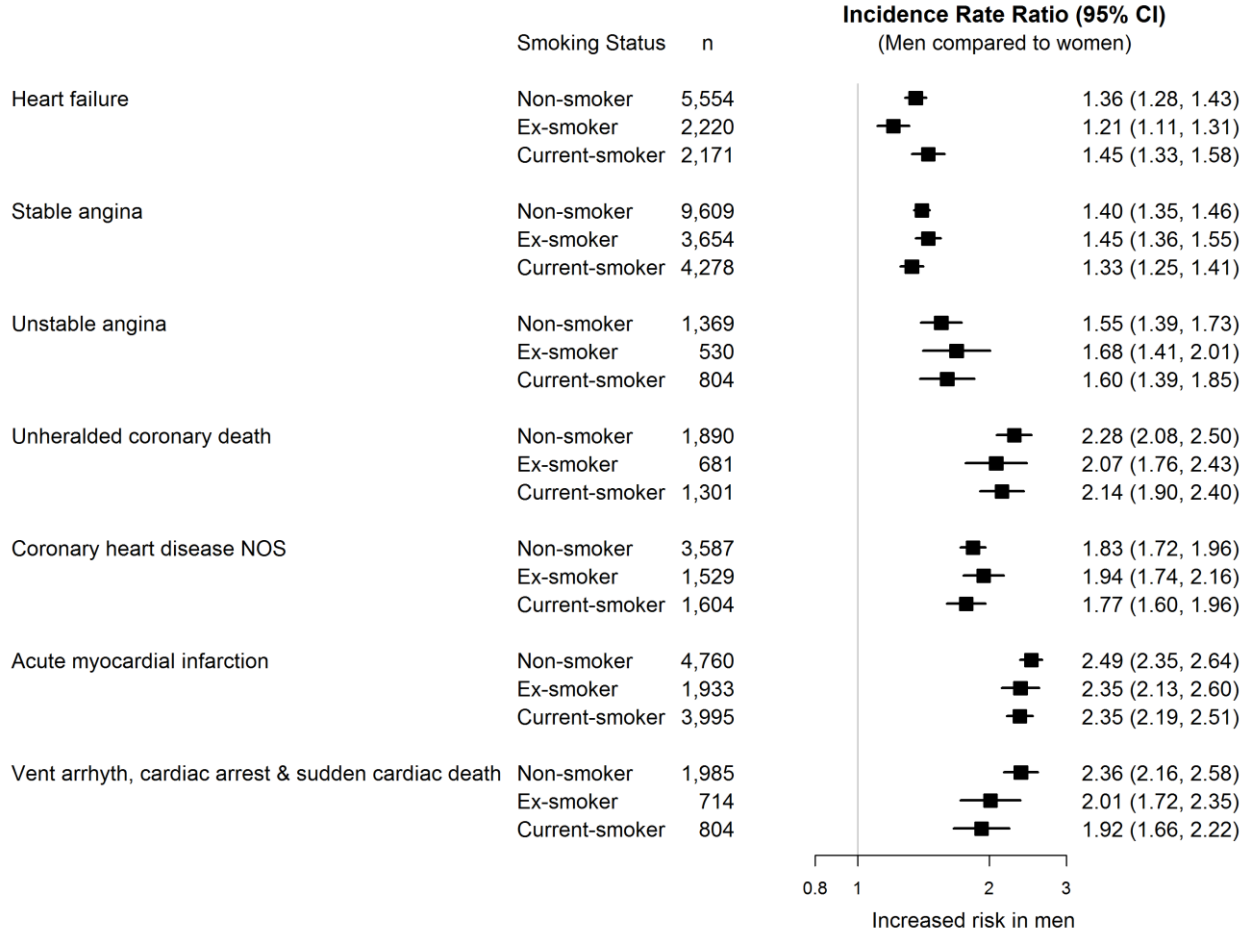
In a subgroup of patients with smoking status recorded (n=1,453,635, 53.8% women), the effect modification of smoking on age-adjusted gender rate ratios for initial CVD presentations was calculated. These rate ratios, stratified by smoking status, are shown in Figure 13 and Figure 14 below. The association between gender and initial presentation of CVDs did not show a consistent pattern across the range of cardiovascular diseases. Amongst current smokers, the male excess was considerably reduced for AAA, and, to a lesser extent, for the ventricular arrhythmias including cardiac arrest and sudden cardiac death. Amongst ex-smokers, there was a reduction in the male excess in heart failure, and the ventricular arrhythmias, although these effects are weak. Amongst non-smokers, the male excess was reduced in peripheral arterial disease, again a relatively weak effect. Amongst the MI subtypes, the only presentation where an effect modification was evident was for nSTEMI, with a reduction in the male excess for both current and ex-smokers. (See Appendix F for results for the myocardial infarction types.) For the remainder of the presentations, there appeared to be minimal effect modification of the effect of gender by smoking.

Figure 13: Gender rate ratios (men compared to women) for initial presentation of cardiovascular disease, stratified by smoking status



N=1,453,635. CI indicates confidence interval; MI, myocardial infarction.

Figure 14: Gender rate ratios (men compared to women) for initial presentation of specific cardiac presentations, stratified by smoking status

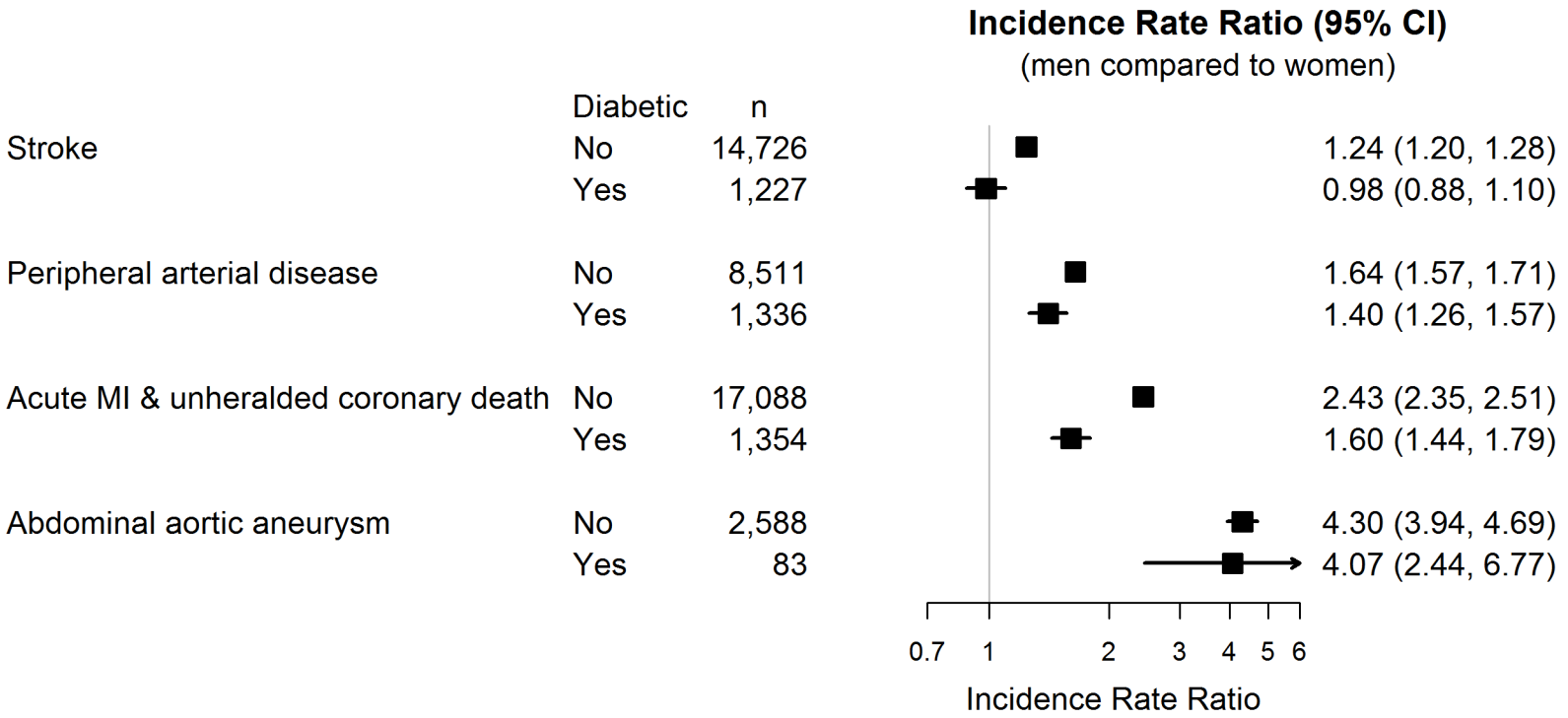


N=1,453,635. CI indicates confidence interval; NOS, not otherwise specified; Vent arrhyth, ventricular arrhythmias.

5.6. Effect modification of gender rate ratios by diabetes

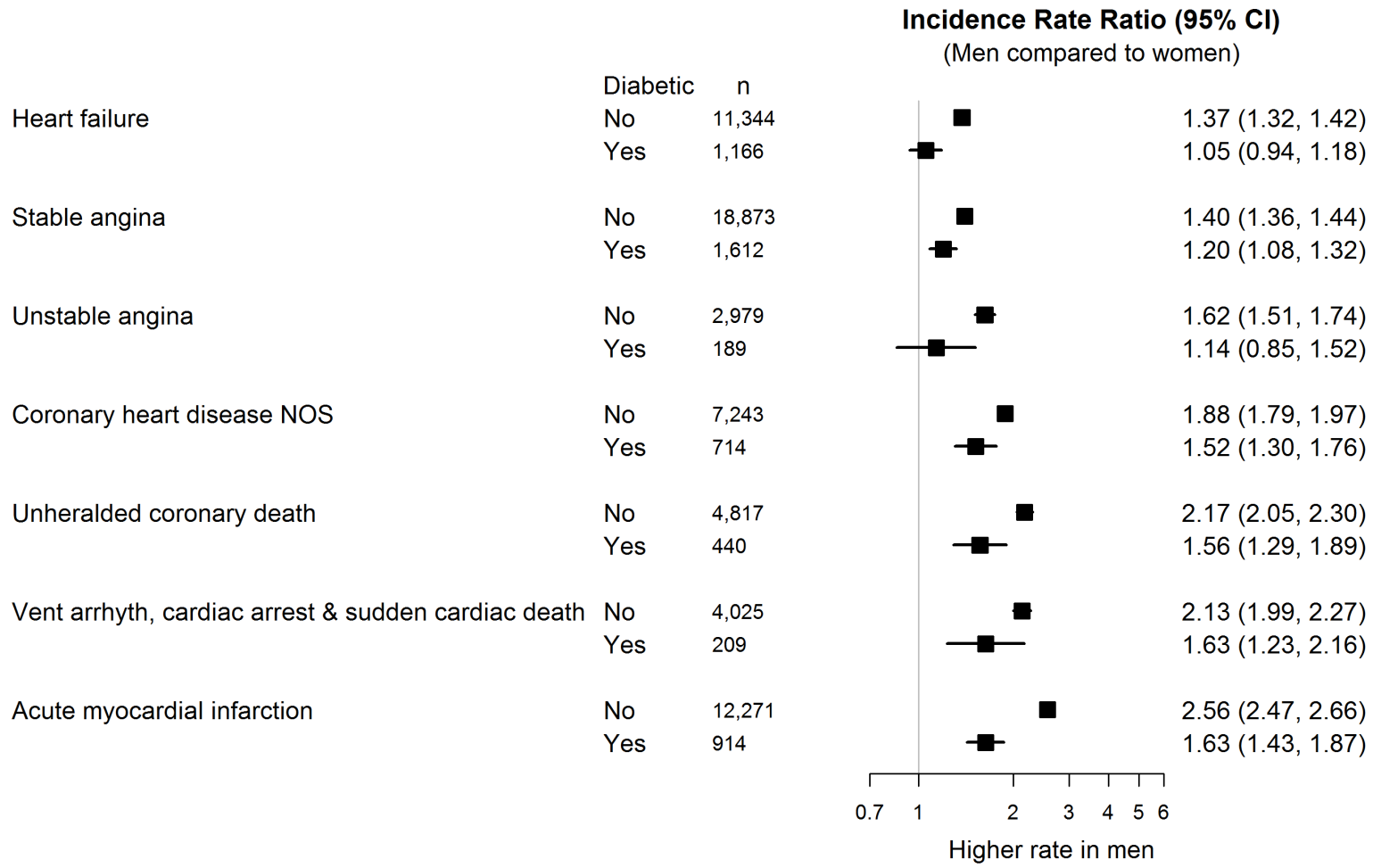
The effect modification of diabetic status on age-adjusted gender rate ratios for initial CVD presentations was calculated. Except for AAA, the male excess was reduced for all presentations. The strongest effect modification was for AMI and the composite AMI and unheralded coronary death, with diabetic patients having a two thirds reduction in the male excess compared to non-diabetic patients. The other acute coronary presentations (unstable angina and unheralded coronary death) also showed similar sizeable reductions in gender differences. For stroke and heart failure, although the effect modification was smaller, the gender difference was eliminated amongst diabetics. (See Figure 15 and Figure 16 below. Additional results for MI subtypes in Appendix F.)

Figure 15: Gender rate ratios (men compared to women) for initial presentation of cardiovascular disease, stratified by diabetic status



N=1,758,584. CI indicates confidence interval; MI, myocardial infarction.

Figure 16: Gender rate ratios (men compared to women) for initial presentation of cardiac presentations, stratified by diabetic status



N=1,758,584. CI indicates confidence interval; NOS, not otherwise specified; Vent arrhyth, ventricular arrhythmias.

5.7. Time trends

For AMI plus unheralded coronary death, stroke, AAA and PAD, there were no significant trends in the rate ratios (men compared to women) over the three time periods between 2001 and 2009. (See Table 17 below.) However, amongst the cardiac presentations, the RR for stable angina and CHD NOS increased over the time period, with non-overlapping 99% confidence intervals. There was a smaller increase in CVD onset with heart failure, while there was a slight decrease in presentation of CVD with AMI, although with overlapping 99% confidence intervals. To investigate the possible reasons for this, the gender-specific age-adjusted rates for all presentations were calculated for each of the three time periods and presented in Figure 17 below.

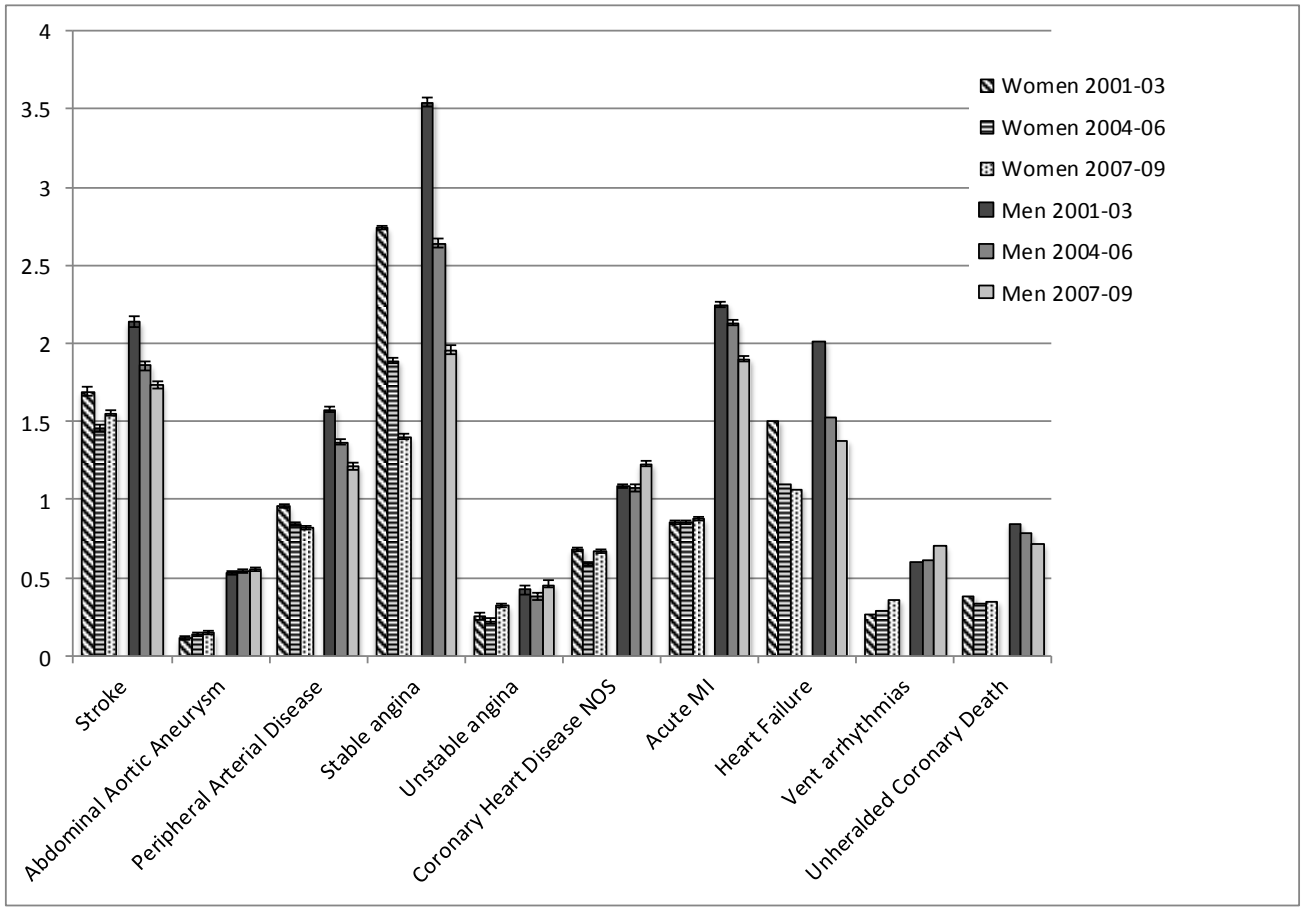
For AMI and unheralded coronary death, men showed a clear decrease in incident rates between 2001-03 and 2007-09 while the rates for women remained largely unchanged. For both stroke and heart failure, both men and women showed a reduction between 2001-03 and 2004-06 which then continued in men but halted or reversed in women. Men appeared to have a greater reduction in the rate of stable angina than women (though the difference is small). Both men and women showed similar small increases in the rates of unstable angina and the ventricular arrhythmias. The rates of CHD NOS did not exhibit any clear trend, while the rates for AAA remained unchanged in both men and women throughout this period.

Table 17: Age-adjusted rate ratio (men compared to women) for range of initial presentations of cardiovascular disease for 2001-3, 2004-6, and 2007-9

	Rate ratio of men compared to women (99% confidence interval)		
Presentation	**2001-3**	**2004-6**	**2007-9**
CVD presentations			
Acute MI & unheralded coronary death	2.57 (2.42, 2.73)	2.55 (2.41, 2.71)	2.41 (2.27, 2.55)
Stroke	1.31 (1.23, 1.39)	1.33 (1.25, 1.42)	1.27 (1.19, 1.35)
Abdominal aortic aneurysm	5.09 (4.25, 6.11)	4.29 (3.65, 5.05)	4.38 (3.75, 5.12)
Peripheral arterial disease	1.74 (1.61, 1.87)	1.73 (1.60, 1.87)	1.69 (1.56, 1.82)
Other deaths	1.36 (1.31, 1.41)	1.34 (1.29, 1.38)	1.28 (1.24, 1.32)
Specific cardiac presentations			
Stable angina	1.35 (1.29, 1.41)	1.46 (1.39, 1.54)	1.55 (1.46, 1.64)
Unstable angina	1.76 (1.53, 2.03)	1.68 (1.47, 1.92)	1.56 (1.38, 1.77)
Coronary heart disease NOS	1.68 (1.54, 1.84)	1.96 (1.79, 2.14)	2.04 (1.88, 2.21)
Acute myocardial infarction	2.73 (2.54, 2.93)	2.62 (2.45, 2.81)	2.43 (2.27, 2.60)
Heart failure	1.39 (1.30, 1.49)	1.42 (1.32, 1.52)	1.48 (1.37, 1.59)
Vent arrhyth, cardiac arrest & sudden cardiac death	2.39 (2.09, 2.73)	2.20 (1.95, 2.48)	2.15 (1.93, 2.40)
Unheralded coronary death	2.25 (2.01, 2.51)	2.40 (2.15, 2.69)	2.40 (2.14, 2.69)

CVD indicates cardiovascular disease; MI, myocardial infarction; NOS, not otherwise specified; Vent arrhyth, ventricular arrhythmia.

Figure 17: Age-standardised rates per 1,000 person years for initial CVD presentations in men and women for three time periods



NOS indicates not otherwise specified; MI, myocardial infarctionvent, ventricular.

6. Conclusions

The incidence rate of atherosclerotic cardiovascular disease in men exceeded that of women for every initial presentation. However, the extent of the male excess varied considerably. In each section below, I summarise my findings for each objective. I then describe the strengths and limitations of this set of analyses. The implications for clinical care, public health practice, and research of my findings from these and my other analyses are discussed in Chapter 8 – Conclusions and Recommendations.

6.1. Summary of existing literature (Objective 1)

Five studies which described the gender-specific rate of onset of CVD across several disease presentations were found in the literature search, only one of these included both men and women. This study identified a clear male excess for AMI but other no significant gender differences in the rate of initial presentation with other presentations;(221) however, given the small numbers and wide confidence intervals, the accuracy of the effect estimates and hence any gender ratio must be questioned. Studies describing the rate of first presentation of specific diseases, which ignored any earlier presentation with other cardiovascular diseases presentations, found a clear male excess in rates of AMI, heart failure and unstable angina.

6.2. Differences between men and women in rates of initial presentations of cardiovascular disease (Objective 2)

6.2.1. Gender-specific rates of initial presentations

Women had consistently lower age-standardised incidence rates per 1,000 person years of all the specific CVD presentations, but there was considerable heterogeneity in the rates across presentations both within and between genders. The greatest gender difference was in the rate of AMI where men had almost double the rate in women; stable angina was the most common cardiac presentation in men followed by AMI while stable angina and stroke the most common presentations in women. In both men and women, the rate of the most common presentation was approximately 10 times the rate in the least common.

It is not clear how meaningful it is to compare these results to the effect sizes found to the existing literature because existing studies have investigated older patients, collected data in earlier time periods than the current study, used different definitions of endpoints, or, in the limited initial presentation studies, employed fewer endpoints, all of which would affect the size of the rates. The male-only studies of initial presentation of specific CVD presentations found consistently higher rates than the present analyses but used no more than three endpoints.(222–225) If additional presentations had been included, the rates

of initial presentation would have been reduced. The one study of initial presentation which included both men and women, as well as covering a similar time period, number of presentations and age range as the present analyses, found a substantially higher rate of acute MI in men than the current analyses; all other rates were potentially consistent with the current analyses given the wide confidence intervals.(221) Out of all the first presentation studies which include multiple endpoints, the OXVASC study by Rothwell et al.(227) is the most similar to the current analyses, covering a similar age group, time period and wide range of specific CVD presentations. They found substantially lower rates of AMI in both genders and lower rate of stroke and AAA in men. The reasons for these discrepancies are not clear, although may be related to different cohort populations – at least amongst the patients with an ethnic group recorded, there was a higher proportion of South Asian and Black patients in the current analyses compared to the OXVASC cohort.

6.2.2. Gender rate ratios for initial presentation of cardiovascular disease

The gender rate ratios (RRs) varied considerably between the different CVD initial presentations. Onset of symptomatic CVD with the most common presentations (heart failure, stroke, stable angina) show only a modest male preponderance, while other presentations showed either a high (PAD, MI, UCD) or extremely high male preponderance (AAA). The highest RR was for abdominal aortic aneurysm, while the lowest RR was for stroke. The RR for the coronary disease presentations increased in a linear fashion from the chronic (stable angina) to the most acute (AMI & UCD), with the RR for heart failure similar to that of stable angina and the RR of the arrhythmias similar to that of AMI.

No other study has published gender rate ratios with confidence intervals for the onset of CVD with the current wide range of initial presentations. These results show in a single study the considerable variation in the association of gender with the different presentations which signal the onset of symptomatic CVD.

6.3. Modification by gender rate ratios by age, smoking or diabetic status (Objective 3)

I found clear evidence of modification of the effect of gender by age, diabetic status, and to a lesser extent smoking. The effect of age varied by presentation with the most common pattern being one of lower gender differences at younger ages, increasing in middle age and then declining in older age. However, for the more acute coronary presentations, particularly AMI, gender differences were high at the youngest age and then declined

thereafter. Studies of gender differences in incidence of AMI or composite CHD have found age-related patterns similar to the current analyses,(207) but lack information on the age-related patterns of onset of CVD across the wide range of presentations used here. These findings are needed to combat the conventional wisdom, perpetuated even in a paper calling for improved research on gender differences in cardiovascular disease, that women develop CVD between 7 and 10 years later than men.(234) General practitioners may discount symptoms and signs of the onset of CVD if women are perceived to be too young to be developing the disease.

Diabetes reduced substantially the male excess for most initial presentations of CVD, with the exception of AAA. Diabetes appears to have either no effect(235) or possibly a protective effect on AAA(236), so it is perhaps not surprising that there is no effect modification on the association of gender with AAA. The strongest effect modification was on presentation with AMI, which is in keeping with findings from studies including AMI as an endpoint.(14,207,237) The current analyses extend that finding to other acute coronary presentations (unstable angina, unheralded coronary death and possibly the ventricular arrhythmias). Additionally, the current analyses indicate that diabetes essentially eliminates the already small gender differences in presentation with stroke and heart failure, in keeping with previous studies on heart failure(238) and stroke(239,240).

Smoking status modified the effect of gender in relatively few presentations compared to age and diabetes. The strongest effect modification was for AAA, where the male excess was considerably reduced amongst current smokers and to a lesser extent amongst ex-smokers. A similar, but weaker, pattern was seen in initial presentation with ventricular arrhythmias. The gender difference in PAD was reduced amongst non-smokers. The links with AAA of both smoking and male gender are well-established(235), but no studies which identified an effect modification of smoking on the effect of gender have been identified. This paucity of research on women and AAA has been commented on previously.(241) This observed reduction in the gender risk ratio amongst both current smokers and ex-smokers supports the idea of risk stratification being a useful adjunct to any screening programmes for AAA(242) which is currently restricted in men in the UK.(243) The reduction in male risk for PAD for non-smokers seems at odds with the limited studies which suggest either a greater sensitivity to smoking in women compared to men or no gender difference,(15,59) but this may be because previous studies have included asymptomatic PAD, unlike the current set of analyses. These differences to existing literature may be a consequence of focussing on onset of cardiovascular disease

rather than any disease presentation, but could also be due to the lack of large sufficiently powered studies incorporating both men and women.

6.4. Trends in gender rate ratios (Objective 4)

I found clear evidence of a reduction over time in gender rate ratios for the majority of the cardiac presentations, but not for the other CVD presentations. In a post-hoc analysis of age-standardised rates for three time periods, the rate ratio of AMI and UCD decreased in men but remained unchanged in women. HF rates also declined across all three time periods in men but plateaued in women from 2004-2006. The rate of arrhythmias appears to be increasing for both men and women, while the rate of stable angina declined in both men and women, with a possible slowing of reduction in women in 2007-09. These findings suggest that the recent gains made in reducing coronary heart disease are not shared equally between men and women, particularly for the more acute presentations.

These findings are consistent with at least one other study on trends with multiple presentations in a single study. An Australian study of trends in hospitalisations between 2000 to 2007 found a reduction in the rate of incident coronary heart disease hospitalisations in men but not in women and similar decline in both women and men in peripheral arterial disease. Unlike my findings, a greater reduction in the rates of admissions for cerebrovascular presentations was found in women.(24) A study using primary care data between 2000 to 2009 found an increase in the rate ratio (women to men) for stable angina and no change for AMI, but these differences could be due to single source of data and different case definitions.(84) Studies of trends with single endpoints found similar patterns as those in the current study for AMI(245–247) , possibly for heart failure(248), stroke(249) and AAA (250,251) and different trends for peripheral arterial disease.(252) No studies on trends of stable angina or the arrhythmias covering the same or similar period could be found.

6.5. Strengths

One of the main strengths of this study, in common with other analyses, is the size of the cohort which is sufficiently large to allow the detailed examination of effect modification of the association of gender by age, smoking status and diabetes. In addition to replicating findings from studies which focus exclusively on individual endpoints like AMI, this study is able to extend findings on effect modification to less frequently studies presentation such as unstable angina and heart failure. The size of the cohort also allows detailed study of gender differences in relatively rare presentations such as AAA, leading to potentially

new insights on the effect modification of smoking on the association of gender with this presentation.

6.6. Limitations

One of the limitations of the current analyses is that no account was taken of other possible co-variates, which could affect the strength of the association of gender with the range of initial presentations. In addition to the common other risk factors such as lipids or BMI, no adjustment was made for prescription of either oral contraceptives or hormone replacement therapy (HRT) in women, both of which can increase the risk of at least some of these cardiovascular disease presentations.(253,254) This omission is potentially less serious for oral contraceptives: While HRT has been shown to have a significant impact on cardiovascular events, given the relatively low increase in risk and the relatively low rate of events in young women taking oral contraceptives, the effect of not adjusting for oral contraceptive use is likely to be minimal. Another limitation is the use of diagnoses to identify onset of symptomatic CVD, where cardiovascular disease can be under-diagnosed in women. For example, angina has been found to be under-diagnosed in women who have chest pain similar to men,(255,256) as well as in women with differing angiographic results but similar prognosis as men.(257,258)

As mentioned above, the implications of these findings for clinical practice, public health and research are further explored in Chapter 8 - Conclusions and Recommendations.

7. Summary of Findings

- Evidence of gender differences in the rates of initial presentation of CVD across a wide range of presentations has been lacking.
- Common CVDs (heart failure, stroke, stable angina) show only a modest male preponderance.
- There was marked heterogeneity in gender effect, with an almost fourfold difference in the male excess in the gender rate ratios, highest for AAA and lowest for stroke.
- There is clear effect modification with age and diabetes across a wide range of presentations but the modification of effect of gender by smoking status was limited mainly to AAA.

Association of smoking with initial presentation of cardiovascular disease phenotypes

1. Abstract

Background: It is unknown whether smoking status is differentially associated with the onset of cardiovascular disease (CVD) across a wide range of presentations, covering the cerebral, coronary, abdominal and peripheral arterial circulations.

Objectives: To determine whether current smokers and ex-smokers differ in their associations with a wide range of initial CVD presentations compared to non-smokers, and whether those associations are modified by gender.

Design: Cohort study using data from the CALIBER research platform linking four data sources (primary care, disease registry, hospitalisation and mortality records) for 1,453,635 patients free from symptomatic CVD with smoking status recorded.

Main outcome measures: Initial presentation of CVD with stable angina, unstable angina, myocardial infarction (MI), heart failure, ventricular arrhythmias, coronary death, stroke, abdominal aortic aneurysm (AAA) and peripheral arterial disease (PAD).

Results: The association of being a current smoker compared to being a non-smoker varied considerably across all CVD presentations, ranging from no association with arrhythmia (age-adjusted hazard ratio (HR) 1.01 (0.90-1.13) to a strong association with AAA (HR 4.63 (4.08-5.25)). For all presentations, the association with being an ex-smoker was considerably smaller and the heterogeneity of effect was considerably reduced, but particularly amongst the cardiac presentations. For most presentations, women who were current smokers had similar risks to men, with the exception of modestly increased association with acute MI (women: HR 2.59 (2.37-2.83); men: 2.20 (2.05-2.37)) and markedly increased association with AAA (women: HR 7.12 (5.74-8.81); men: 3.77 (3.24-4.39)). Similar effect estimates were found with multivariable models, and for models with an extended cohort of patients with missing co-variate data.

Conclusions: Strongly heterogeneous effects of current smoking on initial CVD presentations were found. However, the heterogeneity of effect was much less pronounced in ex-smokers. Gender differences were limited to AAA and acute MI presentations in current smokers compared to non-smokers.

2. Introduction

The previous chapter described the first substantive set of analyses for this thesis examining the association of gender with the range of initial presentations of cardiovascular disease. In this chapter, I investigate the association of smoking with these presentations.

The effect of smoking on individual cardiovascular diseases is well-documented. Current smokers have up to a threefold risk for acute myocardial infarction (AMI) compared to non-smokers, with substantial attenuation of the risk in ex-smokers particularly in women.(205,259) Being a current smoker increases the risk of a stroke by 50%, with ex-smokers having an attenuated but still increased risk, compared to non-smokers.(39) The link between smoking and peripheral arterial disease (PAD), including abdominal aortic aneurysm (AAA), is similarly well-established, with levels of risk similar to AMI.(242,260) Although smoking increases the risk for ventricular arrhythmias, cardiac arrest and sudden cardiac death, the level of risk is less clear.(261–264) While the link between smoking and these common cardiovascular diseases (CVD) studied largely in isolation from each other have been robustly demonstrated, existing research disregards the possibly earlier onset of CVD with other disease presentations. This approach does not give due regard to more generalised nature of atherosclerotic disease, with common risk factors potentially affecting the cerebral, coronary, abdominal and peripheral arterial beds. As with the previous chapter on gender, t*he principal aim of this chapter's analyses is to investigate the association of smoking* with the onset of CVD, looking at the first presentation of any disease *across* the entire range of CVD presentations.

If the onset of CVD with common cardiovascular diseases differ in their association with smoking, this might lead to important insights into aetiology, including the design and interpretation of studies which evaluate the interplay between cardiovascular diseases and genetic variants.(265) Smoking has been shown to have both acute effects such as increasing blood pressure, heart rate, levels of carboxyhaemoglobin, and pro-thrombotic states, as well as chronic effects such as increasing the level of low-density lipoprotein cholesterol, fibrinogen, and platelet aggregation, so differing effects on distinct cardiovascular presentations can be expected.(266,267) Chronic effects might be expected to persist, even after quitting smoking, so I will investigate the effect of being an ex-smoker both by comparing ex-smokers to non-smokers in the main analyses and by further analyses of the sub-group of patients for whom more detailed information on quitting is available. Additionally, there is evidence to suggest that smoking may pose a higher risk of several cardiovascular diseases in women compared to men, at least for

composite endpoints,(2,39,268) raising the question of whether such gender differences exist with specific initial presentations of CVD.

My objectives for the analyses in this chapter are to:

1. Summarise the existing literature on the association of current and ex-smoking compared to non-smoking with initial presentation of CVD across a wide range of presentations;
2. Investigate whether, compared to non-smokers, being a current smoker or an ex-smoker differ in the associations with these presentations;
3. Determine whether any such effects are modified by gender; and
4. Determine the effect of smoking cessation compared to continued smoking on initial cardiovascular disease presentations.

3. Literature Review

3.1. Search Strategy

The aim of my literature search was to find cohort studies on initial presentation of CVD with specific presentations in populations without clinically manifest cardiovascular disease with at least two endpoints which reported gender-specific associations with smoking on those endpoints. I used both a specific strategy focussing on the association of smoking with initial CVD presentation for different specific presentations and a sensitive search focussing on observational studies on the association of smoking with cardiovascular disease in general, when my specific search found few papers. The search terms I used are reproduced in Table 18. I limited both searches to papers in English with the full text available, published since 1992 cited in the Medline database. For the specific search, I searched on pairs of specific endpoint groups in turn to make the process of sifting through papers manageable. I supplemented the studies found through the formal search strategy by searching reference lists of relevant papers, using forward citation searches of earlier relevant papers and asking other researchers for the details of any papers of which they were aware. Given the paucity of studies found, relevant studies published before 1992 which were identified through searching references were also included. Studies which have focussed exclusively on mortality endpoints, such as the recent Million women Study, (269) or specific endpoints, such as the recent systematic review and meta-analysis on the risk of smoking for CHD,(2) have not been included.

3.2. Findings from literature review

Five studies which investigated the association between smoking and initial presentation of CVD with at least two endpoints were found, only one of which included women.(25,222,223,225,264,270) These studies have been summarised in Table 19;

additional studies where the endpoints were first presentation of a specific cardiovascular disease CVD, rather than initial presentation of CVD across any disease type, and which reported multiple endpoints in the same patient population have also been included. (15,271,272)

Table 18: Search strategy used to identify studies on association of smoking with initial presentation of CVD

Concepts	Boolean operator	Terms used
Specific search		
Smoking		smoking OR cigarettes OR tobacco OR "Smoking/adverse effects"[MeSH]
Endpoints	AND	"myocardial infarction" OR STEMI OR nSTEMI OR "acute coronary syndrome" OR “angina” OR “coronary insufficiency” or “acute coronary syndrome” OR “heart attack”
	AND	“abdominal aortic aneurysm” OR “peripheral arterial disease” OR “peripheral vascular disease” OR PAD OR “intermittent claudication” OR “peripheral ischaemia”
	AND	Stroke OR “cerebrovascular disease” OR “cerebrovascular accident”
	AND	(ventricu* AND arrhythm*) OR “cardiac arrest” OR “sudden cardiac death”
	AND	“heart failure”
Gender	AND	“sex factor”[MESH] OR (gender OR sex) OR (women AND men) OR (male AND female)
Initial presentation		((first OR initial) AND (presentation OR manifestation)) OR (onset)
Sensitive search		
Smoking		smoking OR cigarettes OR tobacco OR "Smoking/adverse effects"[MeSH]
Endpoints	AND	“coronary disease” OR “CHD” OR “IHD” OR ((“ischemic” OR “ischaemic”) AND “heart disease”) OR “cerebrovascular disease” OR “peripheral arterial disease” OR “cardiovascular disease” OR “CVD”
Gender	AND	“sex factor”[MESH] OR (gender OR sex) OR (women OR men) OR (male OR female)
Cohort	AND	“observational study” OR “cohort studies”[MESH]

None of the studies were conducted solely with electronic health records (EHRs), like the current analyses, although two used EHRs for endpoint identification.(15,270) Constituent studies of the one systematic review and meta-analysis.(272) were not included in the table. No studies were found which included amongst multiple endpoints ventricular arrhythmias, cardiac arrest, sudden cardiac death, heart failure, AAA, or PAD. In general the studies found current smoking to be a greater hazard for coronary heart disease presentations than for stroke, with a dose-response found in those studies which measured the level of smoking.(25) The one study of first presentation with CHD or stroke which included both genders found women to be at increased risk of CHD but not stroke compared to men only amongst participants smoking the highest number of cigarettes per day.(272)

There were several limitations to the initial presentation studies. First, they included a narrow range (typically 2 or 3) of endpoints rather than examining a wide range of common cardiovascular disease presentations. Second, most were too small (typically analysing between 500-1,000 events) to assess reliably differences amongst multiple endpoints. Third, only one included women.(25) The 'first presentation' studies lacked temporal resolution, having limited ability to distinguish the association between smoking and initial or subsequent presentations, an important distinction in understanding how cardiovascular disease develops.

Table 19: Cohort studies comparing risk of smoking between difference cardiovascular disease presentations

Study Details						Exposure	Cardiovascular Presentation: Hazard Ratio of smokers compared to non-smokers (95% CI)				
First Author & year of publication	Data years & country	Cohort size	Events	Age Group	Gender	Smoking status	Stable Angina	Unstable Angina	Acute MI	Unheralded Coronary Death	Stroke
Studies of overall onset of cardiovascular disease with specific presentations											
Canoui-Poitrine 2009(222)	1991-1998 France	9,758	292	50-59	M	Current	1.3 (0.8-2.0)	2.1 (1.4-3.0)			
Lawlor 2008(270)	1992-2001 S. Korea	648,346		30-64	M	Ex			1.3 (1.2-1.5)		1.0 (0.9-1.1)
						<10			2.0 (1.7-2.3)		1.4 (1.3-1.6)
						10-19			2.3 (2.0-2.6)		1.6 (1.5-1.8)
						20+			2.9 (2.6-3.3)		1.7 (1.6-1.9)
Glynn 2005(223)	1982-2003 USA	22,071	1,348	40-84	M	Ex			1.1 (1.0-1.3)		1.1 (0.9-1.3)
						Current			1.8 (1.6-2.2)		1.83 (1.5-2.2)
Dagenais 1990(225)[a]	1973-1986 Canada	4,576	603	35-64	M	Ex	1.4 (0.9-2.2		0.5 (0.3-1.0)	1.1 (04-3.3)	
						Cigars/pipes	1.2 (0.6-2.4)		1.2 (0.6-2.5)	2.4 (0.7-7.8)	
						1-20	1.3 (0.8-2.0)		1.4 (0.8-2.2)	2.3 (0.9-6.0)	
						>20	1.7 (1.1-2.6)		1.6 (1.0-2.5)	3.4 (1.3-8.5)	
Stokes 1987(25)[b]	1948-1978 USA	5,209	--	<65	W	Current	1.09		1.04	1.44	
					M		0.86		1.57	1.36	

Study Details						Exposure	Cardiovascular Presentation: Hazard Ratio of smokers compared to non-smokers (95% CI)				
First Author & year of publication	**Data years & country**	**Cohort size**	**Events**	**Age Group**	**Gender**	**Smoking status**	**Stable Angina**	**Unstable Angina**	**Acute MI**	**Unheralded Coronary Death**	**Stroke**
Studies of first presentation of specific cardiovascular disease presentations											
Merry 2011(15)	1987-2003 Netherlands	19,096	694	20-59	M	Ex		1.4 (1.0-1.9)	1.3 (0.9-1.8)		
						Current		1.6 (1.2-2.3)	3.1 (2.3-4.1)		
Kondo 2011(271)	2001-2008 Japan	25,464	110	20-61	M	Ex-smoker			0.8 (0.2-4.5)		1.0 (0.4-2.4)
						Light			6.8 (1.7-33.3)		0.7 (0.1-2.6)
						Moderate			3.9 (1.3-17.0		2.5 (1.3-5.3)
						Heavy			5.8 (1.8-25.9)		2.2 (1.0-5.2)
Woodward 2005(272)	1966-1999 Asia Pacific Region	562,338	12,462	>=20	M	Current			1.6 (1.4-1.7)		1.3 (1.2-1.4)
					W				1.7 (1.5-2.0)		1.4 (1.3-1.6)

Hazard ratios in original study unless otherwise specified. MI indicates myocardial infarction; [a] Relative risk; [b] Cochran Mantel Haenzel test to compare age-adjusted prevalences for smoking.

4. Methods

The data sources, population, risk factor and endpoint definitions, as well as composition of the cohort, are described in detail in Chapters 3 and 4. These are, however, briefly summarised here for ease of reference. Any methodological details specific to the set of analyses in this chapter are also noted.

4.1. Data sources

The Cardiovascular disease research using Linked Bespoke studies and Electronic Records (CALIBER) e-health research platform links primary care data from General Practice Research Database (GPRD),(76) to admissions for acute coronary syndrome (ACS) from the Myocardial Ischaemia National Audit Project registry (MINAP),(98) hospital admissions data from Hospital Episodes Statistics (HES),(230) and mortality and deprivation data from the Office of National Statistics (ONS).(113,123) Records were primarily linked using a pre-defined deterministic linkage algorithm based on NHS number, with a small minority linked using a probabilistic method using DOB and postcode.(77) A web-based portal documenting the creation of all CALIBER data items, from these multiple data sources, is available at www.caliberresearch.org and further details on the creation of the CALIBER research platform have been published elsewhere.(231)

4.2. Population

The identification of the general cohort is described in detail in Chapter 3. All patients in the overall cohort had a valid age and gender recorded. For the analysis in this chapter, the cohort was restricted to those patients who had smoking status recorded prior to cohort entry (82.7% of the overall cohort for the PhD). In Chapter 4, I have described in greater detail the ways in which patients with missing data differed to those with more complete data.

4.3. Smoking

Smoking status was defined as the GPRD record of smoking status with the last possible date before endpoint follow-up and categorised as *non-smoker*, *ex-smoker*, or *current smoker*. If non-smokers had a previous record indicating smoking in their entire GPRD history, they were counted as an ex-smoker. The formal variable definition is given in the CALIBER research platform accessible through the web-portal. Access to the web portal is described in Appendix C.

4.4. Other risk factors

More detailed information on the definition of these risk factors is given in Chapter 3, with a brief summary given here. Gender was defined as the gender recorded in the GPRD patient data file. Age at entry was defined as the age in years in January of the first year of endpoint follow-up; more precise estimates of age were not possible as only year of birth is supplied with the data to protect patient identities. Social deprivation was measured by the index of multiple deprivation (IMD) 2007,(123) dividing IMD into quintiles. Ethnic group was categorised as *White*, *Black*, *South Asian* or *Other*, using self-reported ethnic group recorded in GPRD and HES, with unresolvable code conflicts between the two data sources recorded as missing. Systolic blood pressure (SBP) was defined as the mean SBP recorded in the two years prior to the start of endpoint follow-up. Diabetes mellitus was defined as a diagnosis of Type 1, Type 2 or unspecified diabetes or at least one prescription for insulin or oral hypoglycaemic agent before endpoint follow-up. For all remaining risk factors, I used the most recent record from GPRD in the two years before the start of endpoint follow-up. Total cholesterol level was taken from laboratory results from plasma or serum samples, recorded mmol/L units. Body mass index (BMI) was defined as weight in kilograms over height in metres squared. Lipid-lowering medication at baseline was derived from one or more prescriptions for a statin. The number of different blood-pressure-lowering medication classes prescribed was based on prescriptions for thiazides, potassium-sparing diuretics, beta-blockers, ace inhibitors and other less common medications. A complete list of drug classes and specific preparations is available in Appendix D.

4.5. Endpoints: Initial clinical presentations of cardiovascular disease

The definition of my primary endpoints has been described in more detail in Chapter 3, with additional information on the contribution of the different data sources to each endpoint described in Chapter 4. To summarise briefly, my primary endpoints were fatal and non-fatal presentations of a range of cardiovascular diseases encompassing coronary heart disease [(CHD) including stable angina, unstable angina, ST-elevation myocardial infarction (STEMI), non-ST-elevation myocardial infarction (NSTEMI), myocardial infarction not otherwise specified (MI NOS), and coronary death unheralded by prior symptomatic disease], stroke, peripheral arterial disease (PAD), abdominal aortic aneurysm (AAA), heart failure (HF) and ventricular arrhythmias including cardiac arrest and sudden cardiac death (SCD). Given the limited number of myocardial infarctions specified as STEMI or NSTEMI, the main analyses used a composite endpoint of all acute myocardial infarctions (AMI) as an endpoint. Secondary analyses with specific myocardial infarction types have also been presented. For ease of comparison with previous studies,

a composite endpoint of AMI and unheralded coronary death has also been included. Diagnoses were identified using codes from the International Classification of Diseases 10th Revision (ICD 10)(232) for the hospital data (HES) and mortality data (ONS), from Read Codes(78) for primary care data (GPRD) and bespoke variables in the ACS registry (MINAP).

4.6. Statistical analysis

To confirm the validity of smoking status as recorded in GPRD, I calculated the Cox proportional hazard for current and ex-smokers compared to non-smokers, adjusted for age and sex, for lung cancer diagnosis, admission or death (ICD-10 C34).

I developed a stratified Cox proportional hazard model for competing risks, using data augmentation, to estimate the hazard ratios for smoking for each presentation overall and separately for men and women.(184) More detail on this method of data analysis has been given in Chapter3. I produced three models, the first adjusted for gender, age and age-squared, the second also adjusted for deprivation, blood pressure, use of blood-pressure-lowering medication, use of statins, and diabetes, the third also including a shared frailty term to take account of patients clustered at general practice level.(191) I used the Efron method to deal with tied failure times.(190) I tested the proportional hazards assumption visually using log-log plots. The heterogeneity of the HR was tested using the I-squared (I^2) test for heterogeneity and size of the heterogeneity estimated using Tau-squared (T^2).(189)

Multiple imputation for missing data was not undertaken, as there is some evidence that, at least for some risk factors, data recorded in these kind of primary care data are not missing at random.(180,273) Given the level of missingness in ethnic group, lipids and BMI, these variables were not included in the analyses. Gender, age, social deprivation, systolic blood pressure, use of blood pressure lowering medication, use of statins, and diabetes, in addition to smoking status, were used. All analyses were conducted on complete cases, except where specified. Continuous variables are shown as the mean (standard deviation) and categorical variables shown as frequencies (percentage), separately for men and women.

I have first presented the association between being a current smoker or ex-smoker compared to non-smokers with presentations in the cerebral, coronary, abdominal and peripheral arterial circulations (stroke, AMI plus UCD presentations, AAA, and PAD) and then with the cardiac presentations (stable angina, unstable angina, CHD NOS, AMI, ventricular arrhythmias/cardiac arrest/sudden cardiac death combined, heart failure and

UCD). Finally, the association between smoking status and specific myocardial infarction endpoints (STEMI, NSTEMI, and MI NOS) have been shown in Appendix G.

In a sub-group analysis, I repeated this modelling disaggregating AMI subtypes with participants still in the cohort from January 2003, when MINAP data became available, to check whether any association found in the main analysis was an artefact of specifying the AMI subtypes prior to the inclusion of MINAP data. In a second subgroup analysis to test the effect of quitting on my endpoints, I selected participants who had a) at least two further years of observation at baseline and b) a record of smoking status within those two years. I then modelled the association with my endpoints for participants who quit smoking against those who continued to smoke and against non-smokers. In a post hoc analysis, I also modelled the association of smoking, adjusting for age and sex only, for all participants who met the study entry criteria including those in whom other covariate information was missing.

In a post hoc analysis, I modelled the association of smoking, adjusting for age and sex only, for all participants who had smoking status recorded but for whom other covariate information was missing.

All analyses were performed using STATA version 12 (StataCorp, 4905 Lakeway Drive, College Station, TX 77845, United States).

5. Results

From the overall healthy cohort of 1,758,584, 304,949 (17.3%) patients did not have smoking recorded. Out of the 1,453,635 remaining patients with smoking status recorded, a further 555,743 (38.2%) did not have deprivation or blood pressure at baseline recorded. Therefore a total of 897,892 patients (60.5% women) were included in the main set of analyses for this chapter. Those excluded from the cohort because of missing data were more likely to be male, younger and a current smoker than those included in the complete case analysis. More detailed information on difference between patients with missing data and those with complete data is given in Chapter 4. Women had a follow-up time from study entry of 2,735,785 person years and men, 1,652,664 person years, with a median of 4.82 person years (interquartile range (IQR): 1.88-9.09) for women and 4.16 (IQR: 1.69-8.09) for men.

Risk of lung cancer was much higher in current smokers at baseline compared to non-smokers (Hazard Ratio (HR) = 9.50 (95% CI 8.83-10.22; p=0.000)) and weaker in ex-smokers compared to non-smokers (HR=3.93 (CI 3.63-4.26; p=0.000).

5.1. Baseline characteristics

As shown in Table 18**,** men are more likely than women to be current or ex-smokers and have a higher mean age than women, except amongst non-smokers. Men are also more likely to have a history of diabetes recorded and to be on statins, regardless of smoking status. Women had lower mean systolic blood pressure than men, but were not necessarily more likely to be on blood pressure lowering medication. The proportion of current smokers increased with increasing deprivation, while the proportion of non-smokers decreased; there was no significant difference in the proportion of ex-smokers between deprivation groups.

Table 20: Baseline characteristics of patients by smoking status and gender

	Non-smoker		Ex-smoker		Current smoker	
	Women	**Men**	**Women**	**Men**	**Women**	**Men**
Overall, n (%)	335,681 (61.8)	174,307 (49.1)	87,230 (16.1)	76,456 (21.5)	119,986 (22.1)	104,232 (29.4)
Age in years, mean (sd)	49.3 (16.4)	48.7 (14.7)	49.5 (16.1)	53.5 (15.2)	44.9 (13.3)	45.6 (12.8)
Ethnic group: White	173,812 (85.4)	75,316 (84.8)	53,104 (95.7)	39,414 (92.3)	72,012 (95.3)	50,479 (89.3)
Black	9,411 (4.6)	4,145 (4.7)	639 (1.2)	919 (2.2)	1,194 (1.6)	1,674 (3.0)
South Asian	9,807 (4.8)	4,843 (5.5)	351 (0.6)	936 (2.2)	531 (0.7)	1,854 (3.3)
Deprivation (IMD): Least deprived quintile	76,422 (22.8)	40,009 (23.0)	18,598 (21.3)	15,943 (20.9)	16,091 (13.4)	14,282 (13.7)
Most deprived quintile	55,552 (16.5)	29,474 (16.9)	14,815 (17.0)	13,069 (17.1)	34,200 (28.5)	30,560 (29.3)
Systolic BP in mmHG, mean (sd)	129 (19.7)	134 (16.9)	129 (19.3)	137 (17.3)	126 (18.3)	132 (17.0)
Classes of BP lowering medication: 0	274,301 (81.7)	142,446 (81.7)	71,058 (81.5)	57,998 (75.9)	105,312 (87.8)	91,581 (87.9)
1	43,107 (12.8)	22,560 (12.9)	11,320 (13.0)	12,943 (16.9)	10,959 (9.1)	9,396 (9.0)
2	14,333 (4.3)	7,067 (4.1)	3,841 (4.4)	4,149 (5.4)	2,974 (2.5)	2,496 (2.4)
3+	3,940 (1.2)	2,234 (1.3)	1,011 (1.2)	1,366 (1.8)	741 (0.6)	759 (0.7)
Total cholesterol in mmol/L, mean (sd)	5.5 (1.1)	5.3 (1.1)	5.6 (1.1)	5.4 (1.1)	5.6 (1.2)	5.4 (1.2)
Statin prescription	11,609 (3.5)	8,296 (4.8)	4,060 (4.7)	6,151 (8.0)	3,353 (2.8)	4,231 (4.1)
Diabetes	11,982 (3.6)	9,407 (5.4)	3,393 (3.9)	6,182 (8.1)	3,254 (2.7)	4,608 (4.4)
BMI in kg/m2, mean (sd)	26 (5.7)	27 (4.5)	27 (5.8)	28 (4.5)	26 (5.7)	26 (4.6)

Values are n (%) unless otherwise specified. IMD - index of multiple deprivation, sd – standard deviation, BMI – body mass index, BP – blood pressure. Data presented for subset of patients for ethnic group (n= 522,666 (58.2%)), total cholesterol (n= 176,662 (19.7%)), and BMI (n= 608,999 (67.8%)).

5.2. Events

Amongst patients with complete data for this chapter's analyses, there were a total of 52,581 CVD endpoints with a further 26,590 deaths from other causes. The number of specific CVD presentations for women and men are shown in Table 21 below.

Table 21: Number of specific initial presentations for women and men

	Women		Men	
Presentation*	**n**	**% CVD**	**n**	**% CVD**
Stable angina	6,495	24.4	5,686	21.3
Stroke	5,039	18.9	3,448	12.9
Heart Failure	4,026	15.1	2,933	11.0
Peripheral arterial disease	2,810	10.5	2,828	10.6
Acute myocardial infarction (all)	2,747	10.3	3,943	15.2
Myocardial infarction not otherwise specified	2,258	8.5	3,160	11.9
Coronary heart disease not otherwise specified	2,141	8.0	2,520	9.5
Unheralded Coronary Death	1,114	4.2	2,933	11.0
Vent arrhythmias, cardiac arrest & sudden cardiac death	971	3.6	1,280	4.8
Unstable angina	875	3.3	910	3.4
Abdominal aortic aneurysm	435	1.6	1,007	3.8
Non-ST elevation myocardial infarction	329	1.2	450	1.7
ST elevation myocardial infarction	160	0.6	333	1.2
All CVD presentations	26,653	--	25,928	--
Deaths from other causes	15,369	--	11,221	--

* *Presentations ordered in descending order of frequency in women*

5.3. Testing model assumptions – proportional hazards of current and ex smoking compared to non-smoking

The proportion hazard assumption was tested using the log-log graph method for both current smokers and for ex-smokers compared to non-smokers. These graphs are shown in Appendix G. The proportional hazard assumption for both smokers and ex-smokers compared to non-smokers appears to be violated for the following endpoints only: AAA, AMI, stroke and unstable angina. However, given the size of the current cohort any test of proportional hazard is likely to be violated and the hazard lines are broadly parallel for all endpoints, I decided that using the Cox proportional hazard model would be a reasonable approach to modelling the association with smoking.

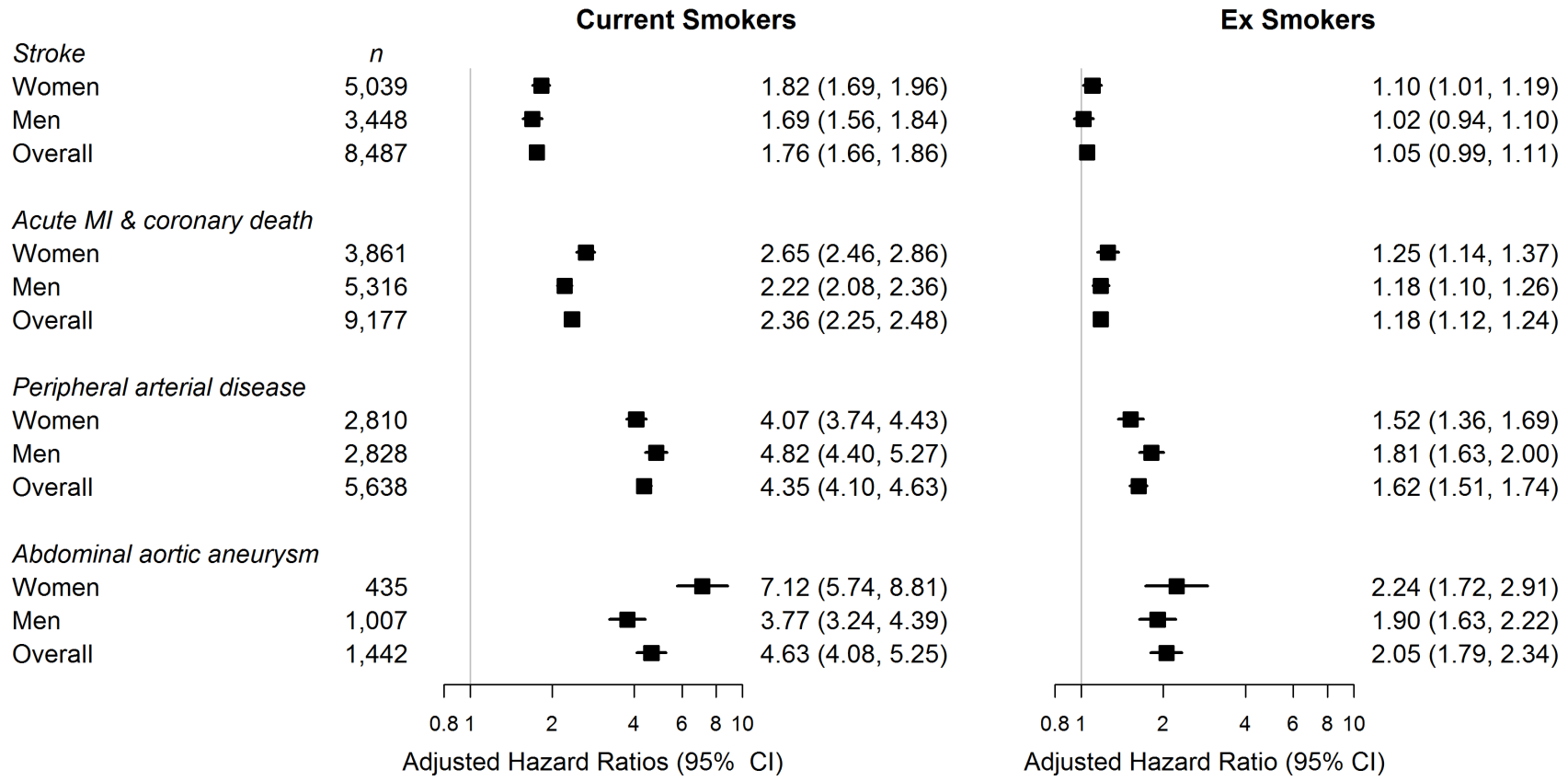
5.4. Association of smoking status with initial presentation of stroke, acute MI and unheralded coronary death, peripheral artery disease and abdominal aortic aneurysm

The association of current smoking differed markedly across the cerebral, coronary, peripheral and abdominal endpoints, as shown in Figure 18. The effects were weakest for AMI and unheralded coronary death (UCD), moderate for stroke and PAD and strongest for AAA. For all patients, the association of smoking with different endpoints was decisively heterogeneous (I^2= 99.5%), with a variance between HRs as measured by T^2 of 0.1951. The inclusion of additional variables in the multivariable Cox proportional hazard model had almost no effect on the size of the HR or measures of heterogeneity. (See Figure 19.) Inclusion of a random effect variable for GP practice in the model also had minimal impact on the effect estimates. (See Figure 20.)

Compared to men, women had a modestly increased hazard associated with being a current smoker, for AMI and UCD, but a markedly increased hazard for AAA. There was no gender difference in hazard ratio for stroke and minimal for PAD. In both men and women, the association of smoking with these endpoints was heterogeneous, with similar variance in association between endpoints (men I^2= 99.1%, T^2= 0.2210; women I^2= 98.9%, T^2= 0.2006).

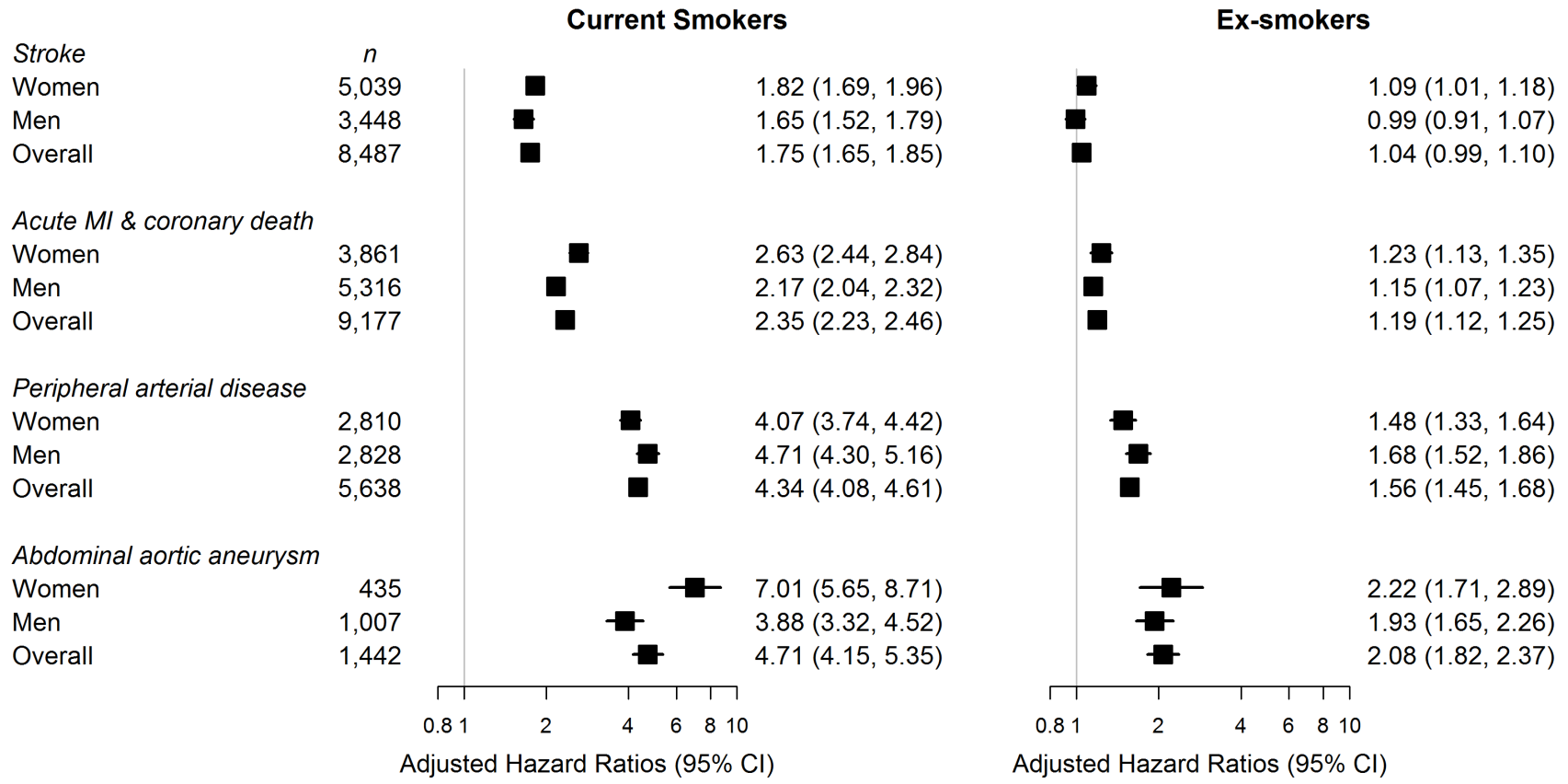
Although being an ex-smoker was associated with increased hazard for most of the cardiovascular presentations compared to non-smokers, the effect size was considerably smaller than for current smokers. For AMI and UCD, PAD and AAA, the hazard was less than half of the hazard for being a current smoker. (Figure 18.) Also, the variance in the association of being an ex-smoker with these presentations was negligible, although still significant (overall: I^2= 97.9%, T^2= 0.0593; men: I^2= 97.1%, T^2= 0.0771; women: I^2= 92.8%, T^2= 0.0387). Again, the inclusion of additional variables or the frailty term had almost no effect on the strength of the association. (Figure 19 and Figure 20 below.) The gender difference in HR for AAA present for smokers was absent for ex-smokers.

Figure 18: Age-adjusted hazard ratios for initial presentations of cardiovascular disease associated with being current smoker or ex-smoker compared to non-smokers, overall and in women and men



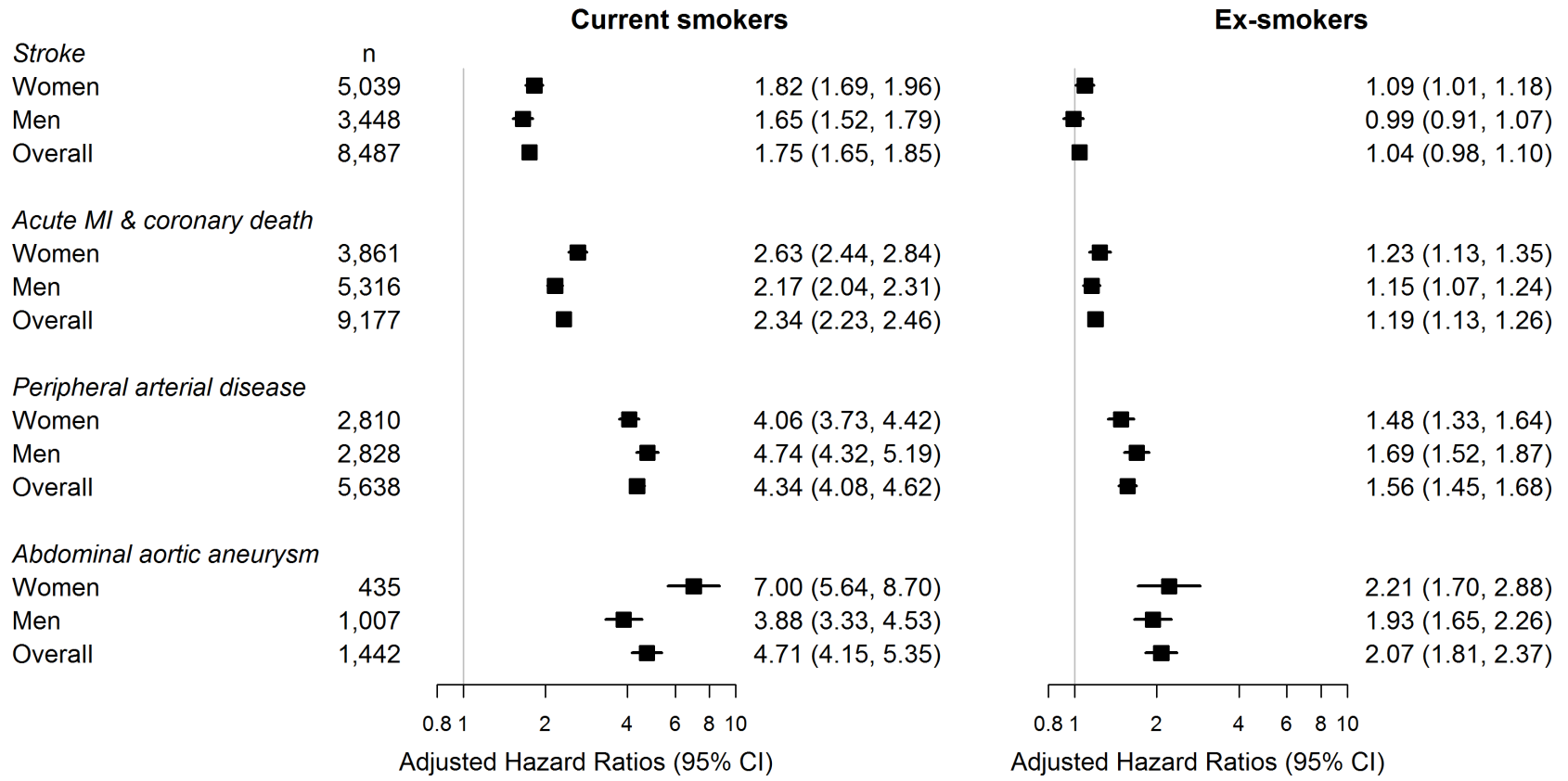
Hazard ratios for current smokers or ex-smokers compared to non-smokers, adjusted for age at baseline for men and women, and additionally for sex for all patients, in complete cases. N= 897,892. MI indicates myocardial infarction; CI, confidence interval; HR, hazard ratio.

Figure 19: Multivariable adjusted hazard ratios for initial presentations of cardiovascular disease associated with being current smoker or ex-smoker compared to non-smokers, overall and in women and men



Hazard ratios adjusted for age at baseline, sex, deprivation (index of multiple deprivation quintile), smoking status (current smoker, ex-smoker, non-smoker), diabetes mellitus (yes/no), mean systolic blood pressure at baseline, blood pressure medication (yes/no) and statin use at baseline (yes/no), in complete cases. N= 897,892. MI indicates myocardial infarction; CI, confidence interval; HR, hazard ratio.

Figure 20: Multivariable adjusted hazard ratios (with random effects for GP practice) for initial presentations of cardiovascular disease associated with being current smoker or ex-smoker compared to non-smokers, overall and in women and men



Hazard ratios adjusted for age at baseline, sex, deprivation (index of multiple deprivation quintile), smoking status (current smoker, ex-smoker, non-smoker), diabetes mellitus (yes/no), mean systolic blood pressure at baseline, blood pressure medication (yes/no) and statin use at baseline (yes/no), with frailty term to take account of clustering at practice level, in complete cases. N= 897,892. MI indicates myocardial infarction; CI, confidence interval; HR, hazard ratio.

5.5. Association of smoking with cardiac disease

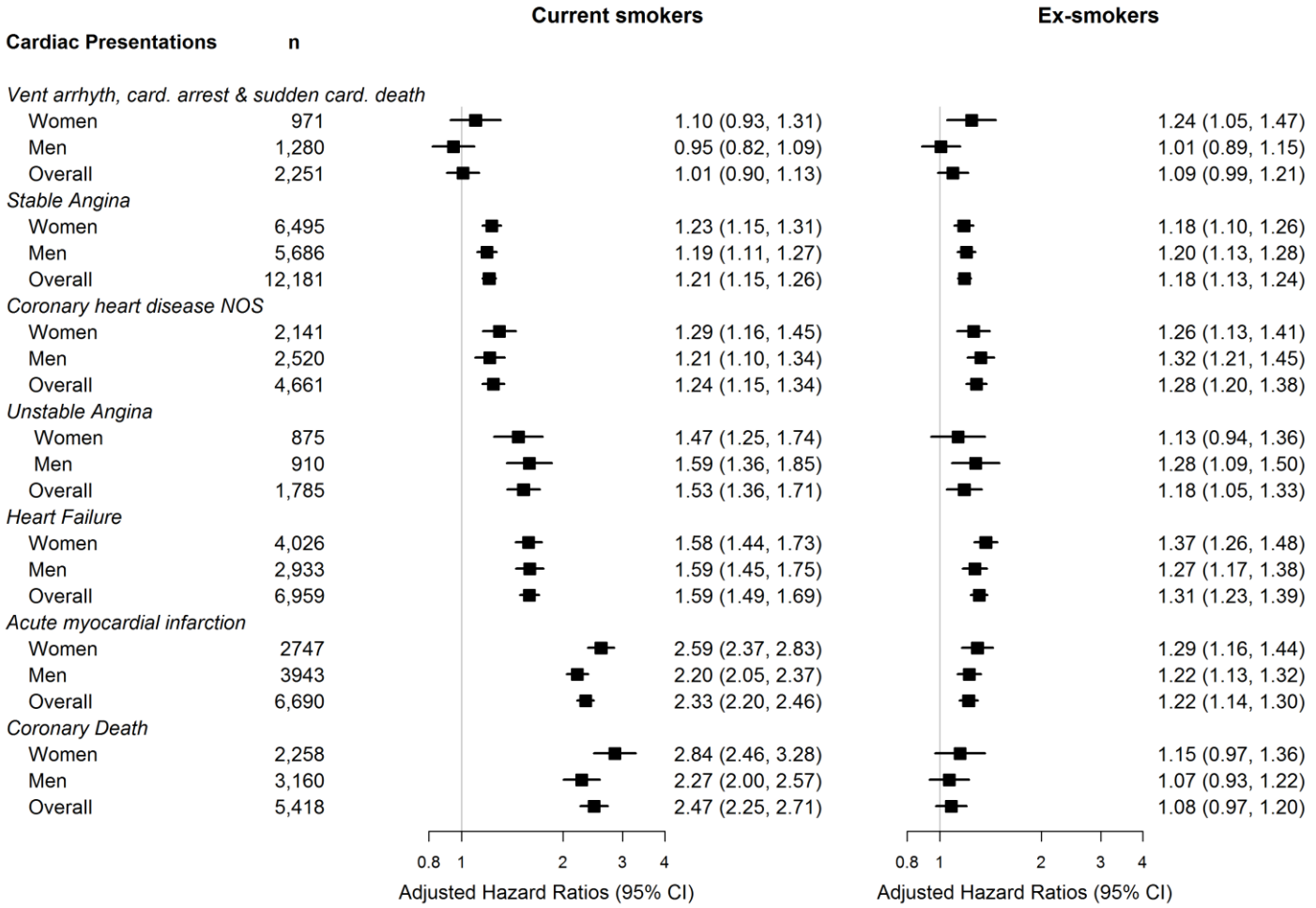
As with cardiovascular endpoints across all four arterial beds, I found considerable variation in the association with being a current smoker at baseline and the specific cardiac endpoints. There were much stronger associations with myocardial infarction and unheralded coronary death than with the other more chronic presentations such as stable angina, unstable angina, coronary heart disease not otherwise specified (CHD NOS) and heart failure and no increased hazard of ventricular arrhythmias, cardiac arrest and SCD. (Figure 21) This pattern was unchanged when additional variables were added in the multivariable model (Figure 22) or the frailty model (Figure 23). The variance in association of current smoking with these endpoints was lower than the variance in association found across all the arterial beds (Overall: I^2= 98.9%, T^2= 0.1101)

Smoking posed a similar hazard for men and women for all cardiac endpoints except acute MI and possibly unheralded coronary death, with women were at greater risk. Again, the variance in association between endpoints was similar in men and women (men: I^2= 97.4%, T^2= 0.0951; women: I^2= 97.8%, T^2= 0.1339).

The association of being an ex-smoker with the cardiac endpoints was smaller across all endpoints than the association of being a current smoker. (See Figure 21.) However, the extent of the difference was much greater with the more acute presentations, with the hazard of acute MI and unheralded coronary death less than half that of current smokers. The variance in the association between the cardiac endpoints is negligible (Overall: I^2= 68.3%, T^2= 0.0028). The addition of other variables or a frailty term into the model made little difference to the effect estimates (Figure 22 and Figure 23). There were no gender differences in the association of being an ex-smoker with these endpoints.

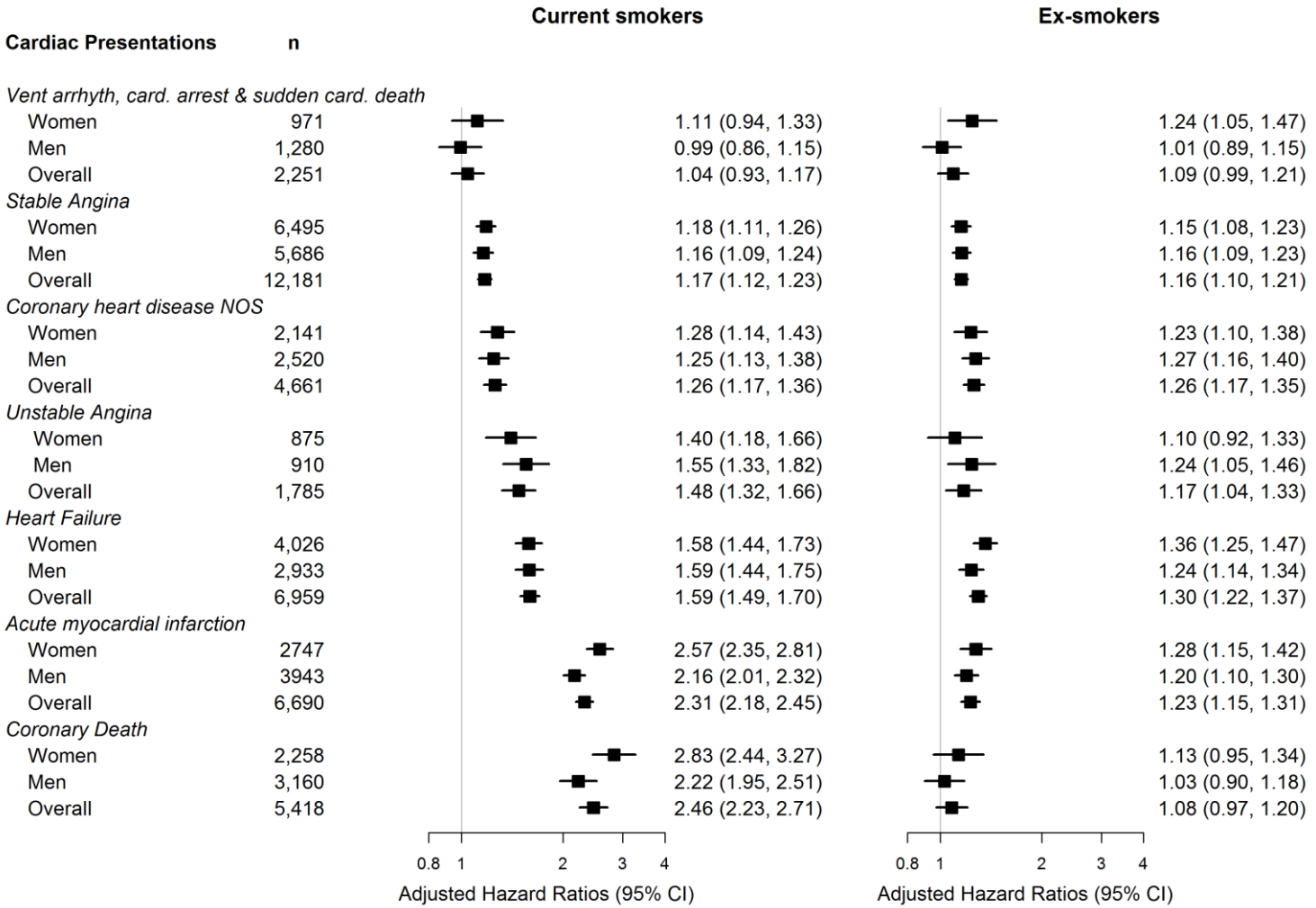
I also investigated the association of smoking with initial presentation of AMI subtypes, with the age-adjusted estimates, multivariable-adjusted estimates, and multivariable-adjusted with random effect for GP all shown in Appendix G. Overall, current smoking had the strongest association with STEMI, followed by MI NOS, followed by NSTEMI, but with overlapping confidence intervals. These analyses were repeated on a subset of the cohort with initial presentations after January 2003 (n= 847,415), with similar effect sizes and gender differences. Smoking posed the greatest hazard for STEMI (HR in women: 3.09 (CI 2.15-4.44; HR in men: 2.63 (2.04-3.38)), followed by MI not otherwise specified (women: 2.46 (2.23-2.72); men: 2.09 (1.92-2.27)) and nSTEMI (women: 2.57 (1.97-3.34); men: 1.83 (1.47-2.29)).

Figure 21: Age-adjusted hazard ratios for initial presentations of cardiac disease associated with being current smoker or ex-smoker compared to non-smokers, overall and in women and men



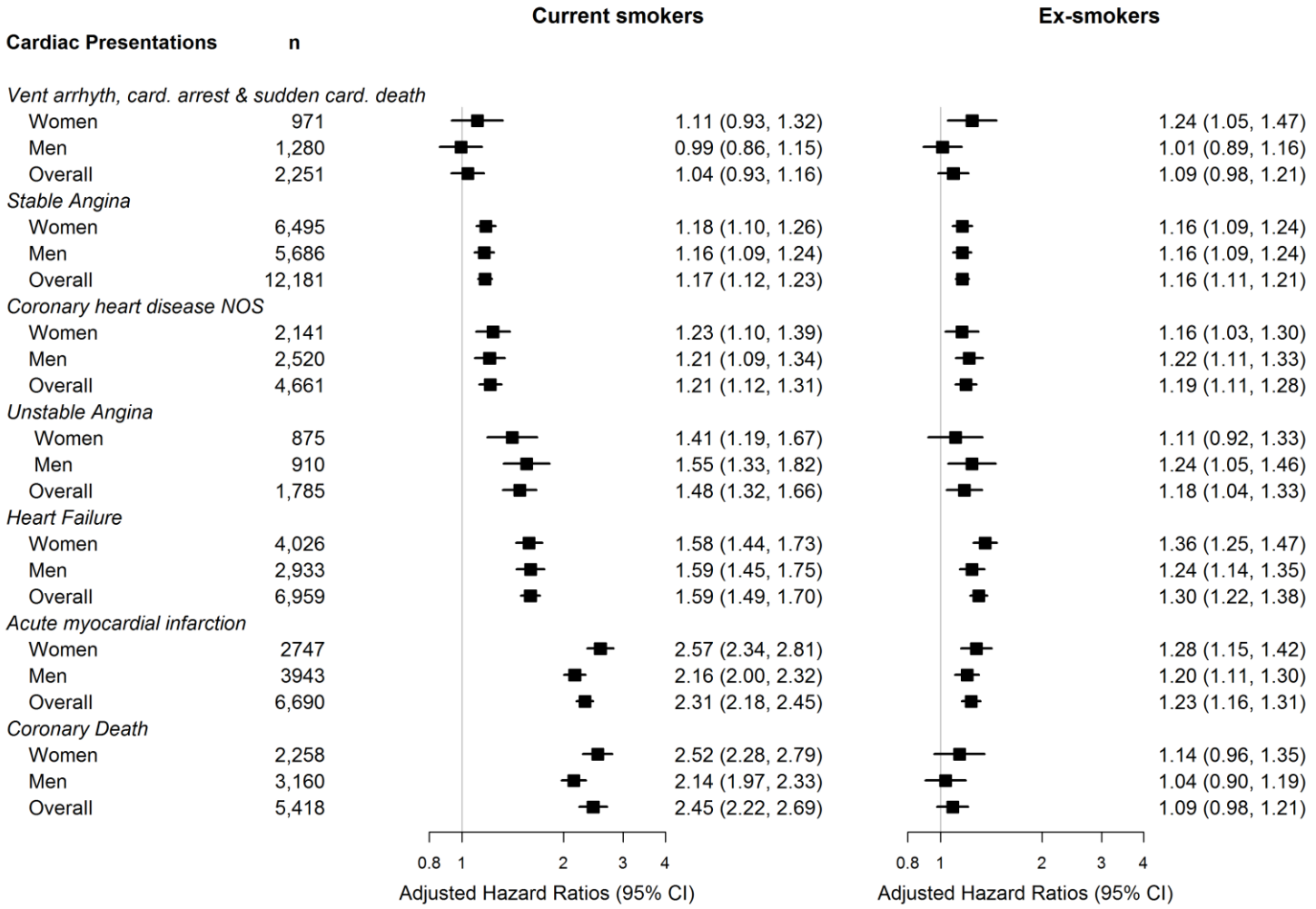
Hazard ratios adjusted for age at baseline for men and women, and additionally for sex for all patients, in complete cases. N= 897,892. MI indicates myocardial infarction; CI, confidence interval; NOS, not otherwise specified.

Figure 22: Multivariable adjusted hazard ratios for initial presentations of cardiac disease associated with being current smoker or ex-smoker compared to non-smokers, overall and in women and men



Hazard ratios adjusted for age at baseline, sex, deprivation (index of multiple deprivation quintile), smoking status (current smoker, ex-smoker, non-smoker), diabetes mellitus (yes/no), mean systolic blood pressure at baseline, blood pressure medication (yes/no) and statin use at baseline (yes/no), in complete cases. N= 897,892. MI indicates myocardial infarction; CI, confidence interval; NOS, not otherwise specified.

Figure 23: Multivariable adjusted hazard ratios (with random effects for GP practice) for initial presentations of cardiac disease associated with being current smoker or ex-smoker compared to non-smokers, overall and in women and men

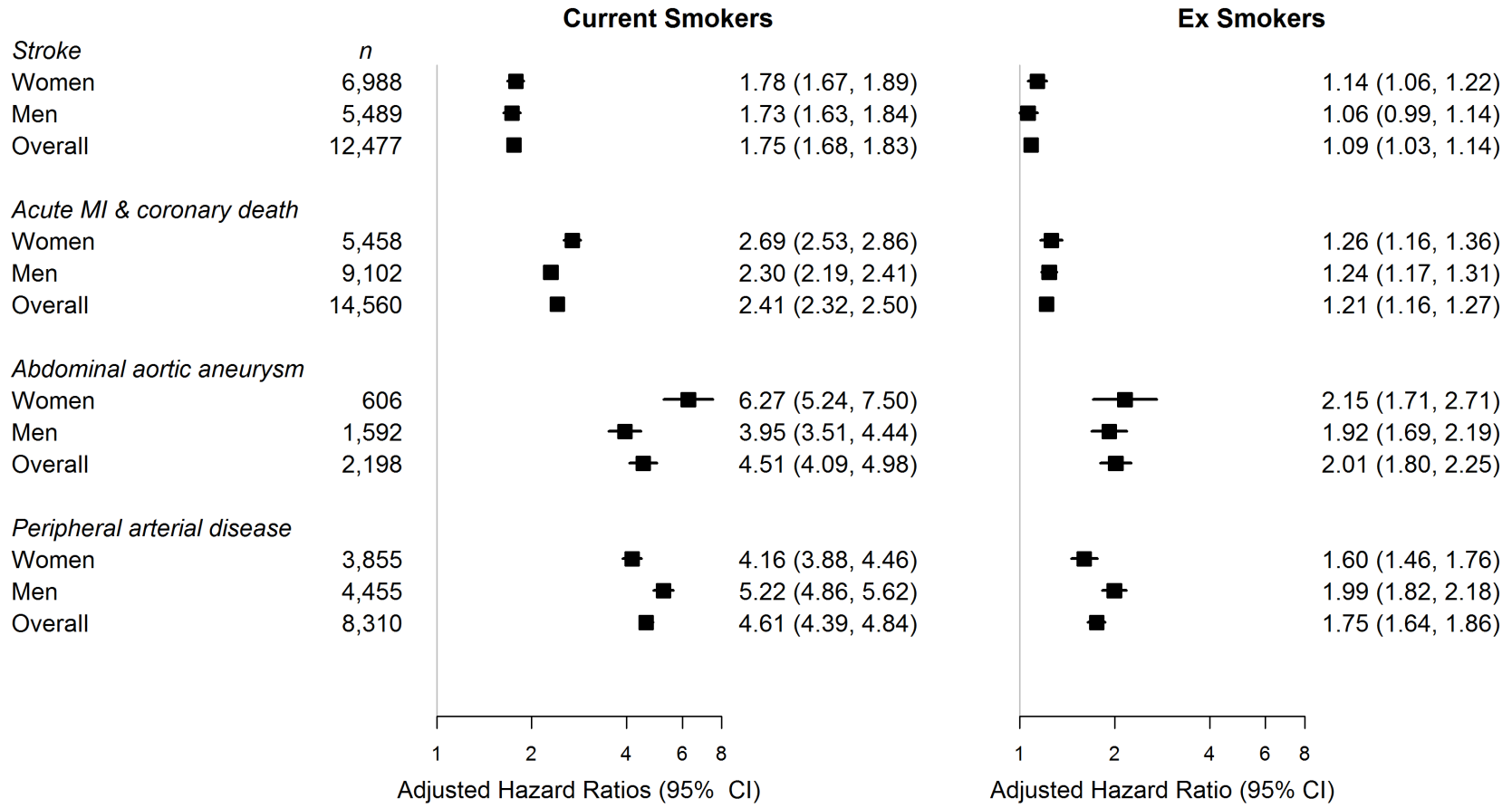


Hazard ratios adjusted for age at baseline, sex, deprivation (index of multiple deprivation quintile), smoking status (current smoker, ex-smoker, non-smoker), diabetes mellitus (yes/no), mean systolic blood pressure at baseline, blood pressure medication (yes/no) and statin use at baseline (yes/no), with frailty term to take account of clustering at practice level, in complete cases. N= 897,892. MI indicates myocardial infarction; CI, confidence interval; NOS, not otherwise specified.

While the likelihood ratio tests comparing all three models showed that the fully adjusted model with shared frailty was the best fit to the data for all endpoints, the change in effect size for the hazard of smoking from the model simply adjusted for age was minimal. I therefore reran the analyses for all cohort participants, including those with missing data, to determine whether similar effect sizes would be seen with all patients, given the potential bias in analysing only patients with complete data. The pattern of association and size of effect with cardiac endpoints remained virtually unchanged in the expanded cohort. There was no association with ventricular arrhythmias, small associations with the more chronic presentations and stronger association with the acute presentations of MI and unheralded coronary death. Women had a stronger association than men for the latter two endpoints. These findings have been shown in Figure 24 and Figure 25 below.

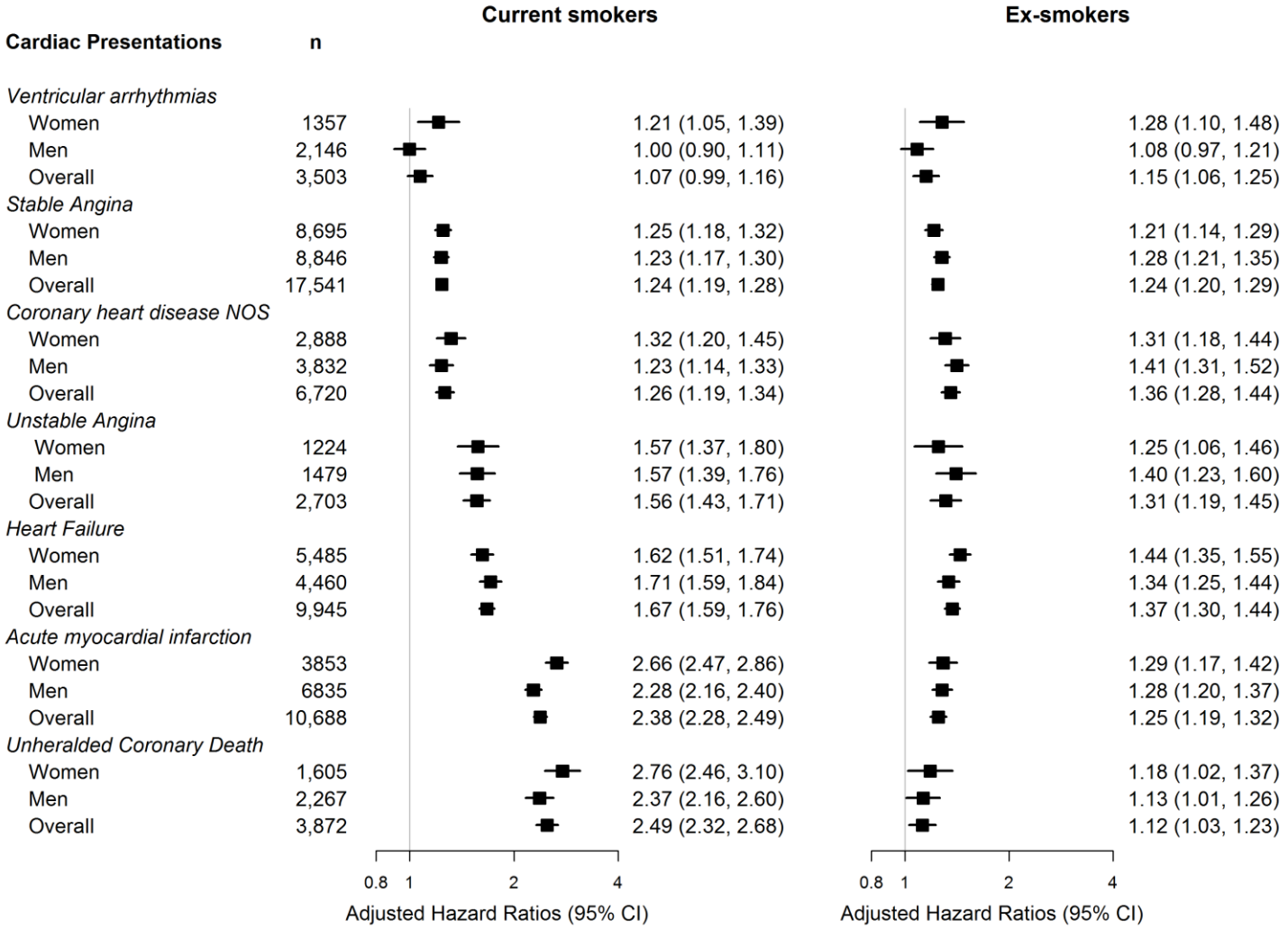
In my subgroup analysis of the effect of quitting smoking, I had 232,416 participants with second smoking status record and a further two years of observation time after the start of endpoint follow-up. Of these, 45,320 women and 27,267 men continued to smoke and 2,814 women (5.84% of female smokers) and 1,691 men (5.84% of male smokers) quit. When modelling the effect of quitting, we found reduced risks with quitting for most endpoints, but with wide confidence intervals, particularly in women, due to the limited number of patients. The largest differences between quitters and continuers were for AMI, unstable angina and unheralded coronary death for both men and women. The hazard ratios for the cardiac presentations for patients who continued to smoke and patients who quit compared to non-smokers are shown in Figure 26 and Figure 27.

Figure 24: Age-adjusted hazard ratios for initial presentations of cardiovascular disease associated with being current smoker or ex-smoker compared to non-smokers, overall and in women and men, in all patients with smoking record



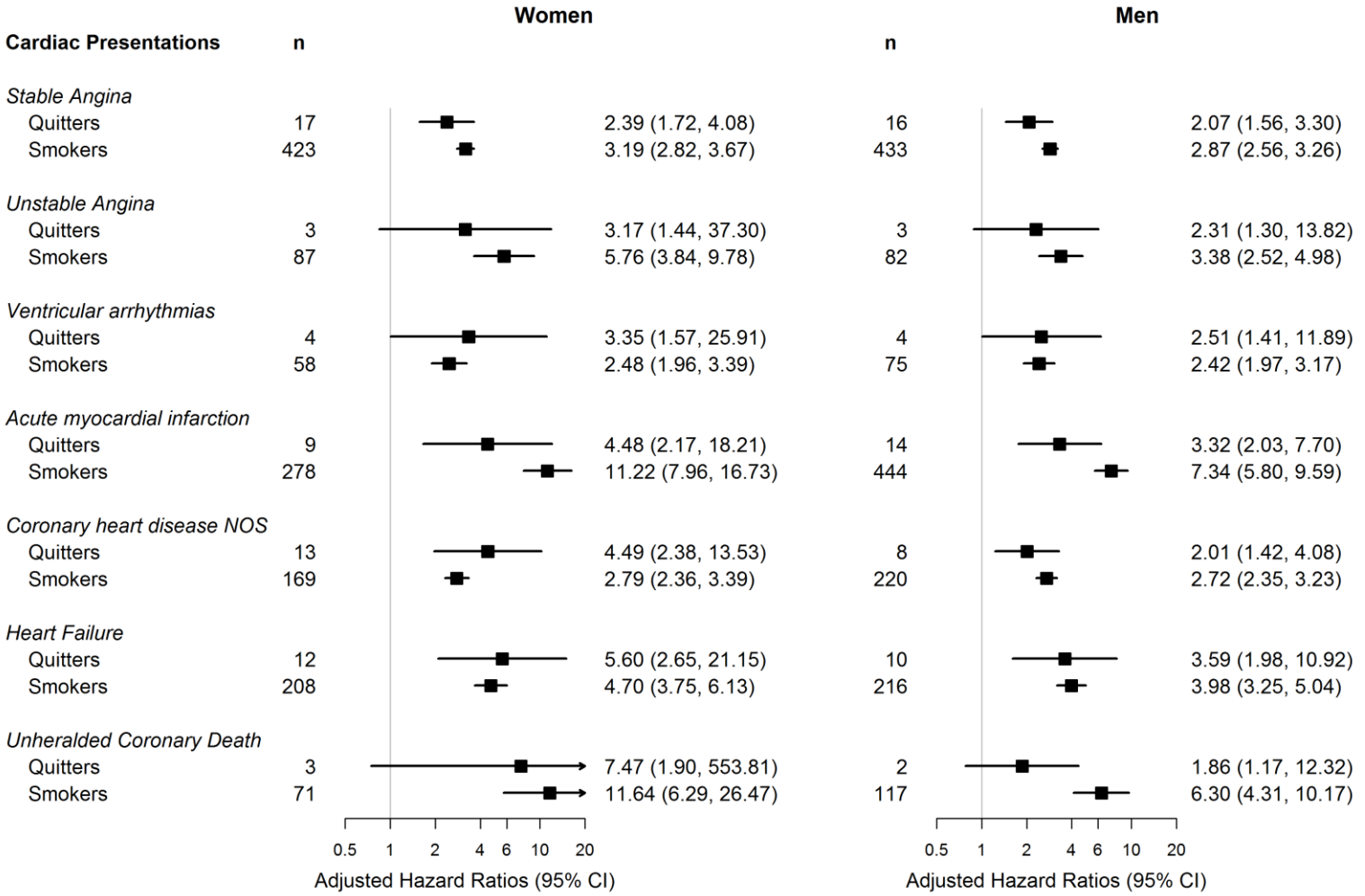
Hazard ratios for current smokers or ex-smokers compared to non-smokers, adjusted for age at baseline for men and women, and additionally for sex for all patients, in all patients with smoking record. N=1,453,635. MI indicates myocardial infarction; CI, confidence interval.

Figure 25: Age-adjusted hazard ratios for initial presentations of cardiac disease associated with being current smoker or ex-smoker compared to non-smokers, overall and in women and men, in all patients with smoking record



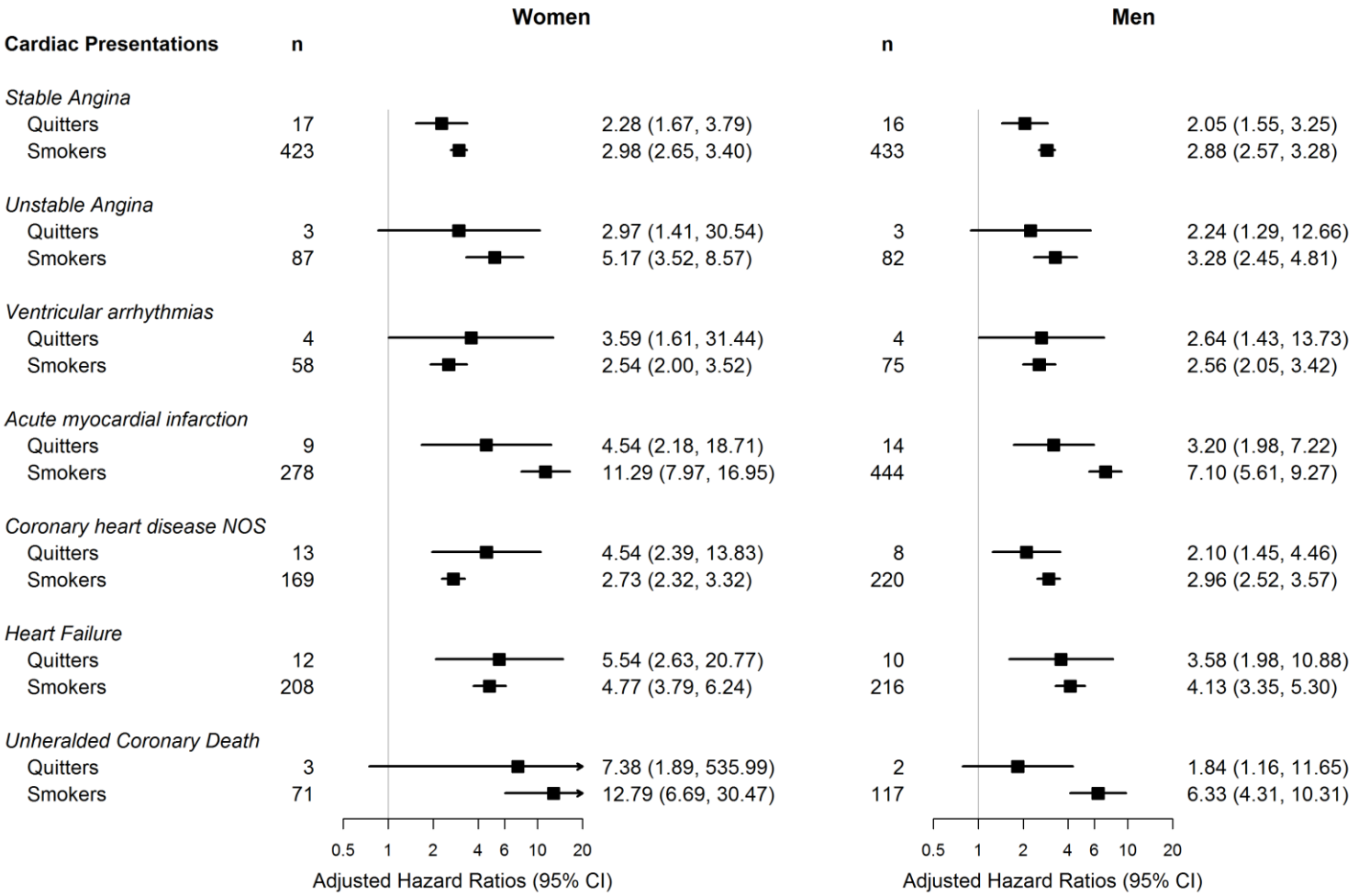
Hazard ratios adjusted for age at baseline for men and women, and additionally for sex for all patients, in complete cases. N= N=1,453,635. MI indicates myocardial infarction; CI, confidence interval; NOS, not otherwise specified.

Figure 26: Age-adjusted hazard ratios for initial presentations of cardiac disease associated with continuing to smoke or quitting compared to non-smokers, overall and in women and men



Hazard ratios adjusted for age at baseline for men and women, and additionally for sex for all patients, in patients with two years of observation after study entry and 2nd smoking status recorded. N=. MI indicates myocardial infarction; CI, confidence interval; NOS, not otherwise specified.

Figure 27: Multivariable-adjusted hazard ratios for initial presentations of cardiac disease associated with continuing to smoker or quitting compared to non-smokers, overall and in women and men



Hazard ratios adjusted for age at baseline, sex, deprivation (index of multiple deprivation quintiles), smoking status (current smoker, ex-smoker, non-smoker), diabetes mellitus (yes/no), mean systolic blood pressure at baseline, blood pressure medication (yes/no) and statin use at baseline (yes/no), in patients with two years of observation after study entry and 2nd smoking status recorded. N=. MI indicates myocardial infarction; CI, confidence interval; NOS, not otherwise specified.

6. Discussion

I found in a cohort of almost 900,000 people with more than 56,000 endpoints of 12 different types that current smoking has heterogeneous effects on the initial presentation of cardiovascular diseases. There have been no other prospective cohorts in which such a range of initial presentations of cardiovascular disease has been assessed. A novel feature of my analysis was to study initial presentations with non-fatal or fatal endpoints. Most cohort studies do not do this; an 'incident' acute myocardial infraction includes patients who have previously manifested another symptomatic cardiovascular disease. This matters because the causes of onset and causes of progression may differ. The considerable heterogeneity in association with different initial presentations empirically demonstrates the importance of making this distinction.

In each section below, I summarise my findings for each objective. I then describe the strengths and limitations of this set of analyses. The implications for clinical care, public health practice, and research of my findings from these and my other analyses are discussed in Chapter 8 – Conclusions and Recommendations.

6.1. Summary of existing literature (Objective 1)

There are a dearth of studies investigating the association between smoking and the onset of CVD across multiple presentations, with only five studies identified in which at least two endpoints were compared, only one of which included women. None of the studies identified included ventricular arrhythmias, cardiac arrest, sudden cardiac death, heart failure, AAA, or PAD as endpoints. These studies found smoking to be more strongly associated with coronary heart disease than stroke and women to be at greater risk than men of CHD, but not stroke, only amongst the heaviest smokers.

6.2. Association of smoking with onset of symptomatic cardiovascular disease across a wide range of disease presentations (Objective 2)

In general, the associations of smoking with specific presentations which I found fit into a pattern of both acute and chronic effects of smoking, as would be expected from the physiological effects of smoking. However, for the first time the size of these effects on clinical presentations of cardiovascular disease were estimated within a single study, allowing direct comparison between presentations. I found the strongest association in both men and women between smoking and PAD and AAA, as well as the acute coronary presentations of AMI and UCD, suggesting an acute triggering effect of smoking such as thrombus formation. I found much lower associations between smoking and the more chronic forms of coronary disease such as stable angina and heart failure, indicating a more limited chronic effect possibly through effects on lipid levels and platelet formation.

I found no additional risk for ventricular arrhythmias (including cardiac arrest and sudden cardiac death). The effect size for ex-smokers for the acute presentations was similar to that for chronic presentations, and the heterogeneity between effect sizes also became negligible, again supporting the more limited chronic (and apparently irreversible) effect of having previously been a smoker, when compared to the much larger effects of being a current smoker. These findings are robust, remaining essentially unchanged with adjustment for additional risk factors and clustering at GP practice level, as well as being replicated in a much larger cohort of patients for whom gender, age and smoking status only were recorded.

My findings are consistent with previous literature for initial presentations of CVD.(222,225,270,274) Small, retrospective case-control studies comparing only two or three morbidities have suggested that smoking is more strongly associated STEMI than nSTEMI;(275) with acute myocardial infarction (AMI) than unstable angina;(276) with AMI than stable angina;(14,277–279)and with unstable versus stable angina,(280) all findings consistent with the dual impact of chronic and acute effects. The heterogeneity of association is also consistent with findings from a recent large study of the association of smoking with a range of different causes of mortality in healthy women.(269)

6.3. Gender differences in the association of smoking with initial CVD presentations (Objective 3)

I found limited evidence of gender differences in the effect of smoking for most presentations with the association of being a current smoker being somewhat stronger in women for an initial presentation of AMI and considerably stronger in women for AAA. While the AMI finding is consistent with existing studies(2,272), my findings do not apply generally to coronary heart disease; by using composite endpoints, previous studies may disguise the differences between more specific endpoints and may exaggerate the extent of gender differences in effect of smoking. The strong association between AAA and smoking, particularly in women, is consistent with the finding of both a recent individual participant meta-analysis of rupture rates of screening-identified AAA(236) and a recent large study of the association between smoking and cause-specific mortality in women, (269) as well as a study of AAA clinically significant events in a women-only cohort.(241)

6.4. Association of smoking cessation compared to continuing to smoke on initial CVD presentations (Objective 4)

In the subgroup of patients for whom quitting or continuing to smoke could be ascertained, quitters had lower hazard of onset of symptomatic CVD across most endpoints, particularly the acute coronary presentations compared to those patients who continued to smoke. However, the number of quitters was only 5% of this subgroup, so

the confidence intervals around these estimates are large. Although the lack of precision is disappointing, these findings are in keeping with existing literature on the impact of quitting on risk of cardiovascular disease.(272,281,282)

6.5. Strengths

The strengths of these analyses, like much of my thesis, lie in the size of the cohort which enables me to investigate individual cardiovascular phenotypes, as well as the prospective nature of the data allowing me to identify initial presentations. For this chapter's analyses, I was also able to assess my measurement of current smoking and ex-smoking with lung cancer, providing face validity to my exposure measurement.

6.6. Limitations

The first limitation of the current study is that level of missing data, leading to greater exclusion of men from the study than women. However, my analyses with all patients who had a smoking status recorded regardless of other missing data also found similar associations to my complete case analyses, indicating that the complete case analysis did not introduce serious bias. Second, as with all electronic health record research, the data was not collected according to protocol so GPs may differ in their coding practices as well as diagnostic approaches. Although my models incorporating shared frailty at practice level were a better fit to the data, the effect estimates did not differ materially suggesting that any differences in coding which may exist do not substantially affect my findings. The third limitation is the lack of greater detail on smoking status, either in terms of amount smoked or duration of smoking. The fourth limitation is the potential for misclassification of my endpoints; of concern here is CHD NOS which is likely to be a mixture of patients with coronary artery disease, stable angina and acute coronary syndromes. Although I could not accurately allocate these patients to a more specific category, the inclusion of CHD NOS as an endpoint allows an estimation of the size of risk attributable to this rather nebulous endpoint. I found that the strength of association of smoking with CHD NOS to be between that of stable and unstable angina; while this known misclassification may affect the size of the effect estimates, it is unlikely to affect the relative effect size between these endpoints. Fifth, it is possible that I missed some presentations, particularly deaths abroad as these are not included in ONS data, but the number of these is likely to be small.

The implications of these finding for research, clinical and public health practice are discussed in Chapter 8.

7. Summary of Findings

- Little is known from previous studies about heterogeneity in the association of smoking status with the onset of symptomatic CVD across a wide range of presentations
- The association of being a current smoker (compared to being a non-smoker) with the initial presentations of CVD varied considerably across all presentations, ranging from no association with the ventricular arrhythmias to a strong association with AAA.
- The association of being an ex-smoker (compared to being a non-smoker) was considerably smaller and less variable across the range of CVD presentations than those found for current smokers.
- Women who smoked had moderately increased risk of AMI compared to men, but markedly increased risk of AAA. No other significant gender differences were identified.
- Weak evidence was found for the benefit of quitting compared to continuing to smoke, particularly for the more acute coronary presentations (unstable angina, AMI, and unheralded coronary death).

Differential effects of systolic and diastolic blood pressure on initial presentations of cardiovascular diseases in women and men

1. Abstract

Background: It is unknown whether blood pressure is differentially associated with the onset of cardiovascular disease (CVD) across a wide range of presentations, covering the cerebral, coronary, abdominal and peripheral arterial circulations.

Objectives: To determine whether systolic blood pressure (SBP) and diastolic blood pressure (DBP) differ in their associations with a wide range of initial CVD presentations and whether those associations are modified by gender or age.

Design: Cohort study using data from the CALIBER research platform linking four data sources (primary care, disease registry, hospitalisation and mortality records) for 961,350 patients free from symptomatic CVD, with a BP measurement recorded in the two years before endpoint follow-up.

Main outcome measures: Initial presentation of CVD with stable angina, unstable angina, myocardial infarction (MI), heart failure, ventricular arrhythmias, coronary death, stroke, abdominal aortic aneurysm (AAA) and peripheral arterial disease (PAD).

Results: There were a total of 52,581 CVD events in the 4,742,025 years of follow-up. The association of SBP with most CVD endpoints was modest, ranging from HR 1.19 (1.16-1.22) per standard deviation (SD) for heart failure to 1.28 (1.24-1.31) for PAD. The exception was AAA where no association with raised SBP was found (1.00, 0.95-1.06). The association of DBP with all endpoints was slightly lower across all endpoints than that found for SBP, with the exception of stronger association with AAA (1.26, 1.19-1.34) and none with PAD (1.02, 0.99-1.05). The increased hazard of most initial presentations was seen at the lowest blood pressure categories. SBP and DBP had a stronger association with stroke in men than in women (men HR 1.27 (1.22-1.32); women 1.17 (1.14-1.20)) while women were at greater risk of arrhythmia than men with higher SBP and DBP (women 1.35 (1.26-1.44); men 1.12 (1.06-1.20)). The association of SBP and DBP decreased with age up to 70-79 years across all endpoints, with the strongest effect modification by age for onset of CVD with heart failure and unheralded coronary death at younger ages.

Conclusions: For most initial presentations, the effects of SBP and DBP were broadly homogeneous and modest, with exception of no association of SBP with AAA and of DBP with PAD. Age was an important effect modifier across all presentations but gender was not.

2. Introduction

The previous two chapters covered the association of gender and smoking with the initial presentations of cardiovascular disease. In this set of analyses, I investigate whether systolic blood pressure (SBP) and diastolic blood pressure (DBP) are differentially associated with the onset of CVD across a wide range of presentations, covering the cerebral, coronary, abdominal and peripheral arterial circulations.

The role of raised blood pressure (BP) as a risk factor for cardiovascular death is well-documented. The Prospective Studies Collaboration found increased risk of mortality from stroke, ischaemic heart disease (IHD) and other vascular causes, even at lower levels of SBP and DBP, across all ages.(283) These findings are supported by other smaller studies of both blood pressure(284,285) and hypertension(286,287) in IHD and stroke deaths. Most studies have also found broadly homogeneous associations between blood pressure and the different causes of death.(283–285) However, by focussing on the association between baseline blood pressure and mortality, these studies conflate the association between development of the different cardiovascular disease (CVD) phenotypes, disease progression and mortality from cardiovascular causes.

A better understanding of the association between blood pressure and the way in which cardiovascular disease first presents, across a range of different cardiovascular disease phenotypes, will help to illuminate the aetiological mechanisms underlying that association. In particular, potentially differing associations with more finely-grained diseases than the composite cardiovascular disease (CVD) endpoints commonly used in mortality studies could indicate differing mechanisms for onset and disease progression. There is considerable debate about which elements of blood pressure best capture the increases in risk associated with cardiovascular disease,(283,288–290) so, in addition to investigating the independent association of SBP and DBP the range of different initial presentations, I will investigate any differences between these two measures of blood pressure for each presentation.

Not only do the levels of blood pressure change with age,(291,292) but the level of risk associated with blood pressure appears to increase with age, at least for some presentations. For example, SBP has been found to be the chief contributor to age-related excess in coronary heart disease (CHD) mortality.(207,293) There is also conflicting evidence whether the risk associated with SBP, DBP and other blood pressure measures, such as pulse pressure, change relative to each other with age.(284,294) I will therefore investigate whether the association between the initial presentation of a range of

cardiovascular diseases and SBP and DBP is modified by age, both independently and compared to each other.

Many studies have found no gender differences in the risk of cardiovascular disease attributable to blood pressure(283,295) or in the response to treatment(296), while others have identified increases in blood pressure, usually SBP, as posing a greater risk to women.(1,297,298) The modification of any association between SBP and DBP and the initial presentations of cardiovascular diseases by gender will therefore be investigated.

The objectives of this chapter are to:

1. Summarise the existing literature on the association of blood pressure with the initial presentation of CVD across a wide range of presentations;
2. Investigate whether SBP and DBP differ in their individual associations with the onset of CVD across a wide range of disease presentations;
3. Investigate whether the association with specific initial presentations differs between SBP and DBP;
4. Investigate whether these associations are modified by either gender or age.

3. Literature review

3.1. Search Strategy

The aim of my literature review was to find cohort studies on the initial presentation of symptomatic CVD with specific diseases in populations without clinically manifest cardiovascular disease. To be included, any study had to include at least two endpoints and report the gender-specific association of systolic or diastolic blood pressure or hypertension with those endpoints. Studies which investigated only men or women were included. I used both a specific strategy focussing on the association of blood pressure with initial CVD presentation for different specific presentations and a sensitive search focussing on observational studies on the association of blood pressure with cardiovascular disease in general, when my specific search found few papers. The search terms I used are reproduced in Table 22 below. I limited both searches to papers in English with the full text available, published since 1992, using PubMed as my principal index. For the specific search, I searched on pairs of specific endpoint groups in turn to make the process of sifting through papers manageable. I supplemented the studies found through the formal search strategy by searching reference lists of relevant papers, using forward citation searches of earlier relevant papers and asking other researchers for the details of any papers of which they were aware. Where I found relevant studies published before 1992, I included those in the Table.

Table 22: Search strategy used to identify studies on association of blood pressure with initial presentation of CVD

Concepts	Boolean operator	Terms used
Specific search		
Blood pressure		"diastolic blood pressure" OR "systolic blood pressure" OR "hypertens*"
Endpoints	AND	"myocardial infarction" OR STEMI OR nSTEMI OR "acute coronary syndrome" OR "angina" OR "coronary insufficiency" or "acute coronary syndrome" OR "heart attack"
	AND	"abdominal aortic aneurysm" OR "peripheral arterial disease" OR "peripheral vascular disease" OR PAD OR "intermittent claudication" OR "peripheral ischaemia"
	AND	Stroke OR "cerebrovascular disease" OR "cerebrovascular accident"
	AND	(ventricu* AND arrhythm*) OR "cardiac arrest" OR "sudden cardiac death" OR "sudden death"
	AND	"heart failure"
Gender	AND	"sex factor"[MESH] OR (gender OR sex) OR (women AND men) OR (male AND female)
Initial presentation		((first OR initial) AND (presentation OR manifestation)) OR (onset)
Sensitive search		
Blood pressure		"diastolic blood pressure" OR "systolic blood pressure" OR "hypertens*"
Endpoints	AND	"coronary disease" OR "CHD" OR "IHD" OR (("ischemic" OR "ischaemic") AND "heart disease") OR "cerebrovascular disease" OR "peripheral arterial disease" OR "cardiovascular disease" OR "CVD"
Gender	AND	"sex factor"[MESH] OR (gender OR sex) OR (women OR men) OR (male OR female)
Cohort	AND	"observational study" OR "cohort studies"[MESH]

3.2. Findings of Literature Review

I found seven studies(222,223,264,299–301) which reported the association between a measure of blood pressure and initial presentation of CVD in at least two endpoints, only two of which included both women and men in the study,(300,301) neither of which reported sex-specific results. The most common endpoints compared are either stable angina versus acute myocardial infarction (AMI)(222,225,299) or AMI versus stroke,(223,300) although one study compared AMI to heart failure(301) and another AMI to sudden cardiac death.(264) No studies which included ventricular arrhythmias or abdominal aortic aneurysm were found. Methods used by these studies included bespoke

cohorts, use of electronic health records (EHRs), and systemic review with individual participant meta-analysis. The studies are summarised in Table 23 below.

Some studies found possibly heterogeneous associations with different presentations,(225,286,299) while most did not. Given the limited number of papers I found on initial presentation, I have included in the table papers which report results in combined men and women, as well as papers on first presentation or mortality endpoints where these met my criteria of publishing sex-specific associations of blood pressure on more than one endpoint.

These studies have a number of limitations; specifically, they:

a) are restricted to a few endpoints;(222,223,225,264,299,300,302)
b) have investigated combined systolic and diastolic hypertension only;(286,287,299,303)
c) are too small to estimate modification by age or gender(300,302) or are limited to one sex (222,223,225,226,304)

The additional studies investigated first presentation or mortality within a specific cardiovascular disease subtype (e.g. first stroke even if preceded by an MI) rather than first of any cardiovascular presentation.(226,283–287,303,304)

Table 23: Summary of studies investigating the strength of association of systolic blood pressure, diastolic blood pressure or hypertension with specific cardiovascular disease presentations (initial, first within presentation, and mortality)

Study Details			Adjusted for						BP measurement	Groups		Cardiovascular Disease Presentations HR or RR (95%CI) per 1 standard deviation unless otherwise noted						
First Author & publ. year	Cohort size	CVD Events	Age	Smoking	Lipids	Obesity	Diabetes	Alcohol		Gender	Age at entry	Stable Angina	Acute MI	Unheralded Coronary Death	Sudden Cardiac Death	Heart Failure	Stroke	Peripheral Arterial Disease
Studies of overall onset of cardiovascular disease with specific presentations																		
Canoui-Poitrine 2009(222)	9,758	292	●	●	●	●	●	●	SBP	M	50-59	1.3 (1.1-1.5)	1.3 (1.2-1.5)[a]					
									DBP			1.0 (0.9-1.3)	1.3 (1.2-1.5)[a]					
Conen, 2007(299)	37,787	930	●	●	●	●	●	●	HT	W	45+	1.1 (0.9-1.3)[b]	1.0 (0.8-1.4)				1.6 (1.2-2.1)	
Glynn 2005(223)[c]	18,662	2,250	●	●	●	●	●	●	HT	M	40-84		1.7 (1.6-2.0)				1.9 (1.7-2.2)	
Mattace-Raso, 2004(300)	4,234	342		●	●	●	●		SBP	60% W	55+		1.2 (1.1-1.5)				1.6 (1.4-1.9)	
									DBP				1.1 (0.9, 1.3)				1.3 (1.1, 1.5)	
Vaccarino, 2000(302)	2,153	552	●	●		●	●		SBP	61% W	65+		1.1 (1.0-1.2)			1.1 (1.0-1.2)		
									DBP				1.0 (0.9-1.2)			1.0 (0.9-1.2)		

Study Details			Adjusted for						BP measurement	Groups		Cardiovascular Disease Presentations HR or RR (95%CI) per 1 standard deviation unless otherwise noted						
First Author & publ. year	Cohort size	CVD Events	Age	Smoking	Lipids	Obesity	Diabetes	Alcohol		Gender	Age at entry	Stable Angina	Acute MI	Unheralded Coronary Death	Sudden Cardiac Death	Heart Failure	Stroke	Peripheral Arterial Disease
Dagenais 1990(225)[f]	4,576	603	●	●	●				SBP 125-132	M	35-64	0.9(0.6-1.3)[a]	1.0 (0.6-1.7)[g]	0.8 (0.3-2.1)				
									SBP 133-140			1.0(0.7-1.5)[a]	0.8 (0.5-1.4)[g]	2.1 (0.9-4.7)				
									SBP 141- 152			1.3(0.9-1.9)[a]	1.3 (0.9-2.1)[g]	2.4 (1.1-5.2)				
									SBP >152			1.4(1.0-2.1)[a]	1.6 (1.0-2.5)[g]	3.1 (1.5-6.5)				
									DBP 78-82			1.0 (0.7-1.4)[a]	1.2 (0.7-1.9)[g]	1.1 (0.6-2.3)				
									DBP 83-87			1.1 (0.7-1.6)[a]	1.1 (0.7-1.8)[g]	1.1 (0.5-2.3)				
									DBP 88-92			1.2 (0.8-1.7)[a]	1.3 (0.8-2.1)[g]	1.7 (0.9-3.2)				
									DBP >92			1.6 (1.1-2.2)[a]	2.3 (1.5-3.5)[g]	2.5 (1.3-4.5)				

Study Details			Adjusted for						BP measurement	Groups		Cardiovascular Disease Presentations HR or RR (95%CI) per 1 standard deviation unless otherwise noted						
First Author & publ. year	Cohort size	CVD Events	Age	Smoking	Lipids	Obesity	Diabetes	Alcohol		Gender	Age at entry	Stable Angina	Acute MI	Unheralded Coronary Death	Sudden Cardiac Death	Heart Failure	Stroke	Peripheral Arterial Disease
First presentation																		
Gu, 2009(286)	158,666	1,371	●	●		●	●	●	HT[h]	53% W	40+		1.8 (1.6-2.1)				3.1 (2.8-3.3)	
Silven-toinen 2008(226)	1,145758	65,611	●						SBP	M	16-25		1.2 (1.2-1.2)				1.1 (1.0-1.1)	
									DBP				1.2 (1.2-1.2)				1.1 (1.1-1.2)	
O'Donnell 1997(304)	18,682	1,562		●		●	●	●	HT	M	40-84		1.8 (1.5-2.1)				2.2 (1.8-2.7)	
Mortality																		
Sauvaget, 2010(285)	167,331	4,007	●	●		●		●	SBP[h]	W	35-90		1.7 (1.3-2.2)				1.9 (1.4-2.7)	
										M			1.5 (1.3-1.8)				2.2 (1.6-2.9)	
									DBP[h]	W			1.0 (0.8-1.2)				1.2 (1.0-1.5)	
										M			1.1 (0.9-1.2)				1.3 (1.0-1.6)	
Gu, 2009(286)[h]	158,666	1,310	●	●		●	●	●	HT[h]	53% W	40+		1.9 (1.6-2.3)				3.0 (2.6-3.3)	
Jouven, 1999(264)	7,746	310	●	●	●	●	●		SBP	M	43-52		1.5 (1.2-1.8)[d]		1.2 (1.0-1.5)			

Study Details			Adjusted for						BP measurement	Groups		Cardiovascular Disease Presentations HR or RR (95%CI) per 1 standard deviation unless otherwise noted						
First Author & publ. year	Cohort size	CVD Events	Age	Smoking	Lipids	Obesity	Diabetes	Alcohol		Gender	Age at entry	Stable Angina	Acute MI	Unheralded Coronary Death	Sudden Cardiac Death	Heart Failure	Stroke	Peripheral Arterial Disease
Lawes 2003(284)	94,147	3,491		●	●				SBP	W	<50		2.4 (1.9-3.1)				1.9 (1.6-2.4)	
										M			1.5 (1.3-1.7)				2.0 (1.7-2.3)	
									DBP	W	<50		2.4 (1.7-3.5)				1.8 (1.4-2.4)	
										M			1.5 (1.3-1.8)				2.5 (2.1-3.0)	
									SBP	W	50-59		1.5 (1.3-1.7)				1.7 (1.4-1.9)	
										M			1.4 (1.2-1.6)				2.0 (1.8-2.2)	
									DBP	W	50-59		1.6 (1.4-1.9)				1.6 (1.4-2.0)	
										M			1.4 (1.2-1.6)				1.9 (1.7-2.2)	
									SBP	W	60-69		1.5 (1.4-1.7)				1.5 (1.3-1.7)	
										M			1.3 (1.2-1.4)				1.5 (1.3-1.7)	
									DBP	W	60-69		1.5 (1.3-1.7)				1.5 (1.3-1.7)	
										M			1.1 (1.0-1.2)				1.4 (1.3-1.6)	
									SBP	W	70+		1.2 (1.1-1.5)				1.3 (1.1-1.4)	
										M			1.1 (1.0-1.3)				1.4 (1.2-1.6)	
									DBP	W	70+		1.2 (1.0-1.4)				1.2 (1.0-1.3)	

Study Details			Adjusted for						BP measurement	Groups		Cardiovascular Disease Presentations HR or RR (95%CI) per 1 standard deviation unless otherwise noted						
First Author & publ. year	Cohort size	CVD Events	Age	Smoking	Lipids	Obesity	Diabetes	Alcohol		Gender	Age at entry	Stable Angina	Acute MI	Unheralded Coronary Death	Sudden Cardiac Death	Heart Failure	Stroke	Peripheral Arterial Disease
										M		1.1 (0.9-1.3)					1.1 (0.9-1.3)	
Lida, 2003(303)	9,633	446	●						HT[h]	W	30+	1.2 (0.6-2.6)					3.5 (1.3-9.3)	
										M		7.7 (1.1-55.2)					2.3 (0.6-8.6)	
Lewington, 2002(283)	1,000,000	56,137							SBP[j]	n/s	40-49	0.5 (0.5-0.5)					0.4 (0.3-0.4)	0.4 (0.4-0.5)
											50-59	0.5 (0.5-0.5)					0.4 (0.4-0.4)	0.5 (0.5-0.5)
											60-69	0.5 (0.5-0.6)					0.4 (0.4-0.5)	0.5 (0.5-0.6)
											70-79	0.6 (0.6-0.6)					0.5 (0.5-0.5)	0.6 (0.6-0.7)
											80-89	0.7 (0.6-0.7)					0.7 (0.6-0.7)	0.7 (0.7-0.8)
									DBP[j]	n/s	40-49	0.5 (0.4-0.5)					0.4 (0.3-0.4)	0.4 (0.4-0.5)
											50-59	0.5 (0.5-0.6)					0.3 (0.3-0.4)	0.5 (0.4-0.5)
											60-69	0.6 (0.5-0.6)					0.4 (0.4-0.4)	0.5 (0.5-0.5)
											70-79	0.6 (0.6-0.6)					0.5 (0.5-0.5)	0.6 (0.6-0.7)

Study Details			Adjusted for						BP measurement	Groups		Cardiovascular Disease Presentations HR or RR (95%CI) per 1 standard deviation unless otherwise noted						
First Author & publ. year	Cohort size	CVD Events	Age	Smoking	Lipids	Obesity	Diabetes	Alcohol		Gender	Age at entry	Stable Angina	Acute MI	Unheralded Coronary Death	Sudden Cardiac Death	Heart Failure	Stroke	Peripheral Arterial Disease
											80-89		0.7 (0.7-0.7)				0.6 (0.6-0.7)	0.7 (0.6-0.8)
Antikainen, 1998(287)	21,493	2,295	●	●	●	●			HT[h]	W	45-64		2.3 (1.6-3.4)				4.4 (2.3-8.4)	
										M			*1.9 (1.6-2.3)*				3.3 (1.9-5.8)	

HR/RR in merged cells indicates combined endpoint; SBP and DBP per standard deviation unless otherwise specified. Studies adjusted for additional co-variates included social factors (Conen, Dagenais, Gu, Sauvegnet), family history of CVD (Canoui-Poitrine, Jouven, O'Donnell), BP medication (Canoui-Poitrine, Gu).HR indicates hazard ratio; RR, relative risk; CI, confidence interval; CVD, cardiovascular disease; BP, blood pressure; MI, myocardial infarction; SBP, systolic blood pressure; DBP, diastolic blood pressure; HT, hypertension; M, men; W, women; ns – not specified in paper; [a] Includes unstable angina; [b] SA here coronary revascularisation only; [c] Venous thromboembolism included as a competing risk; [d] AMI here fatal AMI only; [f] HR per quintile - reference category <125 mmHg SBP and <78 mmHg DBP; [g] AMI here non-fatal AMI only; [h] Stage 1 hypertension (140-159 mmHg SBP and 90-99 mmHg DBP, other hypertension categories also given in paper; [j] HR for reduction of 20 mm Hg SBP and of 10 mm Hg DBP.

4. Methods

The data sources, population, risk factor and endpoint definitions, as well as composition of the cohort, are described in detail in Chapters 3 and 4. These are, however, briefly summarised here for ease of reference. Any methodological details specific to the set of analyses in this chapter are also noted.

4.1. Data sources

The data sources are described more fully in Chapter 3. In brief, the Cardiovascular disease research using Linked Bespoke studies and Electronic Records (CALIBER) e-health research platform links NHS primary care data from the General Practice Research Database (GPRD),(76) to data for acute coronary syndrome (ACS) admissions from the Myocardial Ischaemia National Audit Project registry (MINAP),(98) NHS hospital admissions data from Hospital Episodes Statistics (HES),(230) and mortality and social deprivation data from the Office of National Statistics (ONS).(113,123) Records were primarily linked using a pre-defined deterministic linkage algorithm based on NHS number, with a small minority linked using a probabilistic method using DOB and postcode.(77) A web-based portal documenting the creation of all CALIBER data items, from these multiple data sources, is available at www.caliberresearch.org and further details on the creation of the CALIBER research platform have been publish elsewhere.(231)

4.2. Population

The identification of the general cohort is described in detail in Chapter 3. For the analysis in this chapter, the cohort was restricted to those patients who had at least one blood pressure measurement in the two years prior to endpoint follow-up. In the overall cohort for this PhD, I found 45.3% of adult patients had no such blood pressure measurement. Further information on patients excluded due to missing blood pressure data are given in Chapter 4.

4.3. Blood pressure measurements

Blood pressure used in the present study was measured in general practice as part of routine clinical care. Since April 2004, UK general practitioners have received performance-related pay, worth approximately 25% of a practice's income,(81) for measuring and recording a number of health indicators, including annual blood pressure measurements of anyone over 45 or with hypertension, coronary heart disease, chronic kidney disease or diabetes. Consequently, recording of key indicators increased substantially,(85) with over 80% of adults on average having blood pressure recorded.(305) Among the patients in this thesis cohort with at least one blood pressure

measurement, patients had a median of 3 blood pressure measurements (interquartile range: 1-10). I calculated mean systolic blood pressure (SBP) and diastolic blood pressure (DBP) recording from the two years prior to endpoint follow-up as my baseline measurements.

4.4. Other risk factors

More detailed information on the definition of other risk factors is given in Chapter 3. To recapitulate briefly, I defined the other risk factors as follows. Except for ethnic group, all these risk factors were taken from GPRD data. Gender was defined as the gender recorded in the GPRD patient data file. Age at entry was defined as the age in years in January of the year of study entry. Social deprivation was measured by the index of multiple deprivation (IMD) 2007,(123) dividing IMD into quintiles. Ethnic group was categorised as *White*, *Black*, *South Asian* or *Other*, using self-reported ethnic group recorded in GPRD and HES, with unresolvable code conflicts between the two data sources recorded as missing. Smoking status was defined as the GPRD record of smoking status with the last possible date before study entry and categorised as *non-smoker*, *ex-smoker*, or *current smoker*. If non-smokers had a previous record indicating smoking in their history, they were counted as an ex-smoker. For all remaining risk factors, I used the most recent record from GPRD in the two years before study entry. Total cholesterol level was taken from laboratory results from plasma or serum samples, recorded mmol/L units. Lipid-lowering medication at baseline was derived from one or more prescriptions for a statin. Diabetes mellitus was defined as a diagnosis of Type 1, Type 2 or unspecified diabetes or at least one prescription for insulin or oral hypoglycaemic agent. Body mass index (BMI) was defined as weight in kilograms over height in metres squared. The number of different blood pressure-lowering medication classes prescribed was based on prescriptions for thiazides, potassium-sparing diuretics, beta-blockers, ace inhibitors and other less common medications. A complete list of drug classes and specific preparations is available in Appendix D.

4.5. Outcomes: Initial symptomatic presentations of cardiovascular disease

The definition of my primary endpoints are described in more detail in Chapter 3, with additional information on the contribution of the different data sources to each endpoint described in Chapter 4. To summarise briefly, my primary endpoints were fatal and non-fatal presentations of a range of cardiovascular disease phenotypes encompassing coronary heart disease (stable angina, unstable angina, coronary heart disease not otherwise specified (CHD NOS), ST-elevation myocardial infarction (STEMI), non-ST-elevation myocardial infarction (NSTEMI), myocardial infarction not otherwise specified (MI NOS), and coronary death unheralded by prior symptomatic disease (UCD)), stroke,

peripheral arterial disease, abdominal aortic aneurysm (AAA), heart failure and ventricular arrhythmias including cardiac arrest and sudden cardiac death (SCD). Given the limited number of myocardial infarctions specified as STEMI or NSTEMI, the main analyses used a composite endpoint of all acute myocardial infarctions (AMI) as an endpoint. Secondary analyses with specific myocardial infarction types have also been presented. For ease of comparison with previous studies, a composite endpoint of AMI and unheralded coronary death has also been included. Diagnoses were identified using codes from the International Classification of Diseases 10th Revision (ICD 10)(232) for the hospital data (HES) and mortality data (ONS), from Read Codes(78) for primary care data (GPRD) and bespoke variables in the ACS registry (MINAP).

4.6. Statistical analysis

The principal analysis assessed the relationship between SBP and DBP separately, as linear continuous variables, with the initial presentation of all of the cardiovascular presentations. I developed a stratified Cox proportional hazard model for competing risks, using data augmentation,(184) to estimate the hazard ratios and 95% confidence intervals for one standard deviation difference in SBP and DBP for each presentation. The SBP and DBP measurements were scaled using the standard deviation estimated from the entire cohort, with the same standard deviation used in all analyses. Although there are cogent arguments against using standardised regression coefficients as the measure of effect, not least because such a practice does not allow comparison between studies,(306) I took this approach here to allow direct comparison between SBP and DBP within the current study. Without this standardisation, these two measures of blood pressure could not be compared to each because they are measured on different scales.(284,300) The cohort-wide standard deviation for SBP at baseline in the cohort was 18.9 mmHg, and for DBP, 10.0 mmHg, similar to 20 mmHg SBP and 10 mmHg used to report BP in the Prospective Studies Collaboration study of blood pressure.(283)
I produced three models for each measure of blood pressure, the first adjusted for gender, age and age-squared, the second also adjusted for deprivation, smoking, use of statins, and diabetes, and the third also including a shared frailty term to take account of patients clustered at general practice level. Given the level of missingness in ethnic group, lipids and BMI, these variables were not included in the principal analyses. I explicitly tested for interaction between gender and SBP and DBP and age and SBP and DBP. Subsequently, models stratified for gender and for age were estimated to test whether these modified the association of SBP and DBP with the range of endpoints. All analyses were conducted on complete cases. I used the Efron method to deal with tied failure times (except for the model for stroke in women where Breslow method had to be used for the model to converge). (190)

Multiple imputation for missing data was not undertaken, as there is some evidence that, at least for some risk factors, data recorded in these kind of primary care data are not missing at random.(180,273)The existence of heterogeneity in the hazard ratios (HRs) was formally tested using the I-squared (I^2) test for heterogeneity, while the extent of the heterogeneity across CVD phenotypes was estimated using Tau-squared (T^2).(189)

Given the digit preference for recording blood pressure measurements in multiples of ten evident from the data (See Figure 28 below), I also investigated whether modelling the association of BP with my endpoints using categories of blood pressure centred around the preferred digits produced similar results to those using continuous BP variables. For SBP, I used the following categories <124, 125-134, 135-144, 145-154, 155-164, 165+ mmHG and for DBP, <75, 75-84, 85-94, 95-104, 105+ mmHg). I tested the proportional hazards assumption for SBP and DBP visually using log-log plots on these categorical SBP and DBP variables. I also tested whether the standard deviation of SBP and DBP was substantially influenced by the digit preference evident in the data. If all values that were multiples of 10 were dropped from the dataset, the standard deviation was similar to that derived from all measurements, leading to the conclusion that the standard deviation was not influenced by GPs' digit preference.

Continuous variables are shown as the mean (standard deviation) and categorical variables shown as frequencies (percentage). All analyses were performed using STATA version 12 (StataCorp, 4905 Lakeway Drive, College Station, TX 77845, United States).

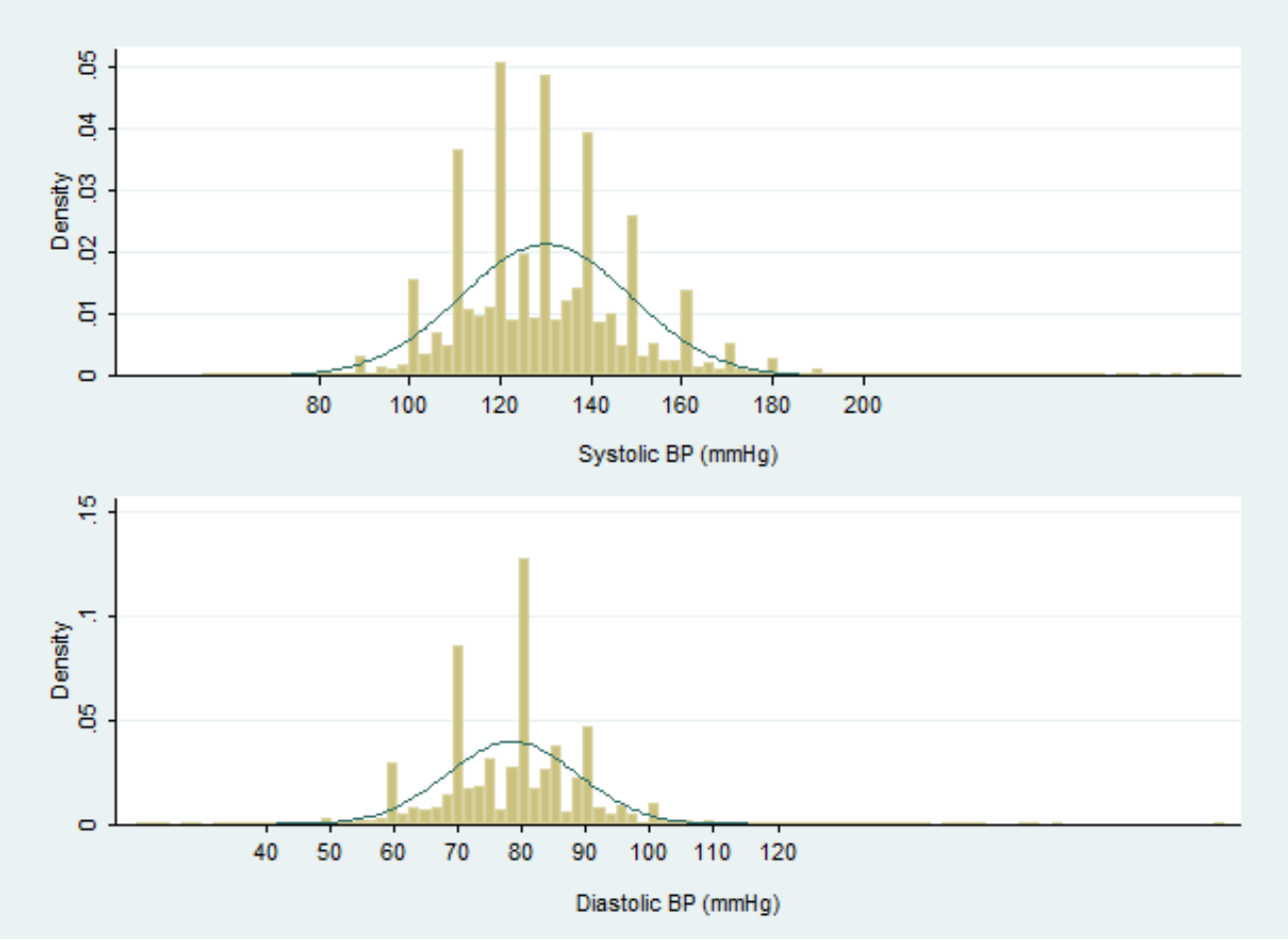
I have first presented the association between SBP and DBP with presentations in the cerebral, coronary, abdominal and peripheral arterial circulations (Stroke, AMI plus UCD presentations, AAA, and PAD) and then with the specific cardiac presentations (stable angina, unstable angina, CHD NOS, AMI, ventricular arrhythmias/cardiac arrest/sudden cardiac death combined, heart failure and UCD). Finally, the specific myocardial infarction subtypes (STEMI, NSTEMI, and MI NOS) have been presented in Appendix H.

5. Results

797,234 patients, from the overall healthy cohort of 1,758,584, did not have a BP measurement recorded at baseline. Therefore a total of 961,350 patients (581,580 women and 379,770 men) met the inclusion criteria for this set of analyses. (Further information on difference between patients with missing data and those with complete data is given in Chapter 4.) Women had a follow-up time from study entry of 2,952,608

person years and men, 1,792,417 person years, with a median of 4.94 person years (1.92-9.09) for women and 4.25 (1.73 -8.20) for men.

Figure 28: Distribution of systolic and diastolic blood pressure before study entry showing digit preference in recording



BP indicates blood pressure.

5.1. Baseline characteristics

The baseline characteristics of the cohort y are shown in Table 24 and Table 25 below, stratified by 10 mmHg bands of SBP and DBP. As expected, the mean age increased with each increase in BP category, for both SBP and DBP. Women formed a greater proportion of the lower blood pressure categories, as did those from Black and South Asian ethnic groups, though it should be noted that ethnic group is missing for 32% of the cohort. The proportion of ex-smokers increased with SBP and DBP levels, as did the proportion of patients on statins. The proportion of patients with a diabetes diagnosis increased with SBP but not DBP. The mean BMI did not vary appreciably across the levels of SBP but did increase with levels of DBP.

Table 24: Characteristics of cohort stratified by systolic blood pressure levels

Systolic Blood pressure		<120	120-129	130-139	140-149	150-159	≥160
Number		258,907	204,238	183,364	147,367	88,442	79,032
Age at cohort entry in years, mean (95% CI)		39.6 (10.1)	43.7 (12.5)	49.0 (14.3)	55.4 (14.8)	61.0 (14.1)	66.4 (13.5)
Female, %		197,612 (76.3)	121,861 (59.7)	93,738 (51.1)	73,804 (50.1)	47,793 (54)	46,772 (59.2)
Ethnic group:	White	128,634 (83.6)	100,286 (87.8)	91,704 (89.6)	78,468 (92.2)	49,703 (94)	46,921 (95.2)
	Black	6,523 (4.2)	4,392 (3.8)	3,617 (3.5)	2,460 (2.9)	1,275 (2.4)	1,026 (2.1)
	South Asian	8,394 (5.5)	4,238 (3.7)	3,160 (3.1)	1,910 (2.2)	932 (1.8)	587 (1.2)
Deprivation:	Least deprived quintile	53,894 (20.9)	41,440 (20.4)	36,659 (20.1)	29,494 (20.1)	17,403 (19.7)	14,337 (18.2)
	Most deprived quintile	53,001 (20.6)	40,732 (20.0)	35,519 (19.5)	27,759 (18.9)	16,827 (19.1)	15,590 (19.8)
Smoking status:	Ex-smoker	39,112 (15.9)	33,117 (17.1)	31,993 (18.5)	27,936 (20.4)	17,360 (21.3)	14,881 (21)
	Current smoker	66,939 (27.2)	52,732 (27.3)	43,586 (25.2)	31,125 (22.7)	16,688 (20.5)	14,093 (19.9)
Total cholesterol in mmol/L, mean (sd)		5.2 (1.1)	5.3 (1.2)	5.4 (1.1)	5.5 (1.1)	5.6 (1.1)	5.7 (1.1)
On statins		2,869 (1.1)	5,461 (2.7)	9,029 (4.9)	10,111 (6.9)	6,654 (7.5)	5,240 (6.6)
Diabetes mellitus		4,209 (1.6)	6,237 (3.1)	9,017 (4.9)	9,262 (6.3)	6,508 (7.4)	5,844 (7.4)
BMI in kg/m2, mean (sd)		24.5 (4.5)	26.3 (5)	27.4 (5.3)	28.1 (5.5)	28.4 (5.6)	28.4 (5.8)
Classes of BP lowering medication:	0	251,640 (97.2)	191,017 (93.5)	155,880 (85)	106,247 (72.1)	50,933 (57.6)	36,779 (46.5)
	1	6,218 (2.4)	9,968 (4.9)	19,348 (10.6)	28,433 (19.3)	26,429 (29.9)	29,825 (37.7)
	2	831 (0.3)	2,560 (1.3)	6,389 (3.5)	9,927 (6.7)	8,565 (9.7)	9,554 (12.1)
	3+	218 (0.1)	693 (0.3)	1,747 (1)	2,760 (1.9)	2,515 (2.8)	2,874 (3.6)

Values are n (%) unless otherwise specified. CI indicates confidence interval; sd, standard deviation; BMI, body mass index; BP, blood pressure. Data presented for subset of patients for ethnic group (n=557,615 (58.0%)), deprivation status (n=957,382 (99.6%)), smoking status (n=901,629 (93.8%)), total cholesterol (n=185,223 (19.3%)), BMI (n=627,175 (65.2%)).

Table 25: Characteristics of cohort stratified by diastolic blood pressure levels

Diastolic Blood pressure		<70	70-79	80-89	90-99	≥100
Number of patients		133,837	315,352	362,809	126,425	22,927
Age at cohort entry in years, mean (sd)		41.2 (13.5)	46.1 (15.2)	51.9 (15.4)	54.5 (14.2)	53.6 (13.6)
Female, %		101,897 (76.1)	207,461 (65.8)	200,817 (55.4)	61,929 (49)	9,476 (41.3)
Ethnic group:	White	68,863 (84.6)	160,954 (87.8)	188,143 (90.6)	66,284 (91.3)	11,478 (89.7)
	Black	3,276 (4)	6,533 (3.6)	6,320 (3)	2,527 (3.5)	637 (5)
	South Asian	4,104 (5)	6,968 (3.8)	6,097 (2.9)	1,757 (2.4)	295 (2.3)
Deprivation:	Least deprived quintile	28,438 (21.3)	64,008 (20.4)	72,663 (20.1)	24,088 (19.1)	4,030 (17.6)
	Most deprived quintile	27,458 (20.6)	62,230 (19.8)	69,541 (19.2)	25,054 (19.9)	5,145 (22.5)
Smoking status:	Ex-smoker	21,402 (16.9)	52,966 (17.7)	63,338 (18.7)	22,863 (19.6)	3,830 (18.5)
	Current smoker	34,450 (27.1)	77,150 (25.8)	81,066 (23.9)	26,986 (23.1)	5,511 (26.7)
Total cholesterol in mmol/L, mean (sd)		5.1 (1.1)	5.3 (1.1)	5.5 (1.1)	5.6 (1.1)	5.7 (1.1)
On statins		2,403 (1.8)	11,373 (3.6)	17,529 (4.8)	6,896 (5.5)	1,163 (5.1)
Diabetes mellitus		3,217 (2.4)	13,681 (4.3)	17,576 (4.8)	5,709 (4.5)	894 (3.9)
BMI in kg/m^2, mean (sd)		24.0 (4.1)	25.7 (4.8)	27.4 (5.3)	29.1 (5.8)	30.1 (6.2)
Classes of BP lowering medication:	0	127,562 (95.3)	282,991 (89.7)	287,042 (79.1)	81,377 (64.4)	13,524 (59)
	1	4,538 (3.4)	22,089 (7)	52,813 (14.6)	33,481 (26.5)	7,300 (31.8)
	2	1,208 (0.9)	7,627 (2.4)	18,148 (5)	9,224 (7.3)	1,619 (7.1)
	3+	529 (0.4)	2,645 (0.8)	4,806 (1.3)	2,343 (1.9)	484 (2.1)

Values are n (%) unless otherwise specified. CI indicates confidence interval; sd, standard deviation; BMI, body mass index; BP, blood pressure. Data presented for subset of patients for ethnic group (n=557,615 (58.0%)), deprivation status (n=957,382 (99.6%)), smoking status (n=901,629 (93.8%)), total cholesterol (n=185,223 (19.3%)), BMI (n=627,175 (65.2%)).

5.2. Events

Amongst patients with complete data for this chapter's analyses, there were a total of 52,581 CVD endpoints with a further 26,590 deaths from other causes. The number of specific CVD presentations for women and men are shown in Table 21 above, in descending order of frequency of CVD presentation in women.

Table 26: Number of specific initial presentations for women and men

	Women		Men	
Presentation	**n**	**% CVD**	**n**	**% CVD**
Stable angina	6,495	24.4	5,686	21.3
Stroke	5,039	18.9	3,448	12.9
Heart Failure	4,026	15.1	2,933	11.0
Peripheral arterial disease	2,810	10.5	2,828	10.6
Acute myocardial infarction (all)	2,747	10.3	3,943	15.2
Myocardial infarction NOS	2,258	8.5	3,160	11.9
Non-ST elevation myocardial infarction	329	1.2	450	1.7
ST elevation myocardial infarction	160	0.6	333	1.2
Coronary heart disease NOS	2,141	8.0	2,520	9.5
Unheralded coronary death	1,114	4.2	2,933	11.0
Vent arrhythmias, cardiac arrest & sudden cardiac death	971	3.6	1,280	4.8
Unstable angina	875	3.3	910	3.4
Abdominal aortic aneurysm	435	1.6	1,007	3.8
All cardiovascular disease presentations	26,653	--	25,928	--
Deaths from other causes	15,369	--	11,221	--

Presentations listed in order of frequency in women. CVD indicates cardiovascular disease; NOS, not otherwise specified.

5.3. Testing Model Assumptions

The proportion hazard assumption, tested using the log-log graphs using the categorical variables for SBP and DBP, was broadly met for most endpoints although there was some overlap between the hazard lines. The presentations where the hazards were less proportional were AAA, unstable angina, and ventricular arrhythmias for both SBP and DBP and for unheralded coronary death for SBP only. However, given the size of the current cohort any test of proportional hazard is likely to be violated. I therefore decided that using the Cox proportional hazard model would be a reasonable approach to modelling the association between initial presentations and blood pressure. These graphs are shown in Appendix H.

5.4. Association of blood pressure with initial presentation of stroke, acute MI and unheralded coronary death, peripheral artery disease and abdominal aortic aneurysm

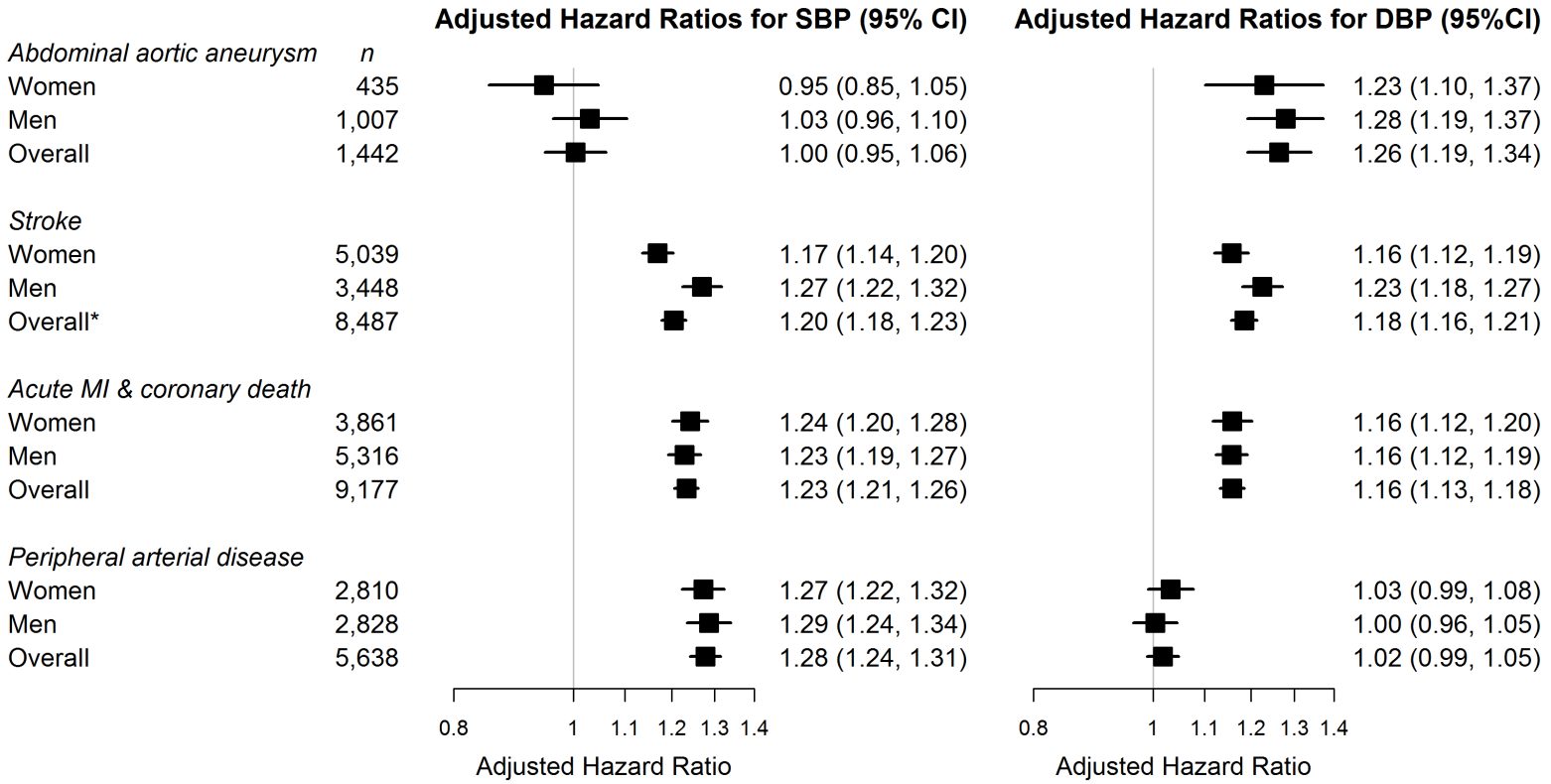
5.4.1. Overall

For all patients in the age-adjusted model, one SD increase in SBP was associated with a moderate increased risk (25%) of an initial presentation of CVD with stroke, acute MI/ coronary death and PAD, but not AAA. Ignoring AAA, the lowest HR for an increase in SBP was associated with stroke and highest with PAD. Heterogeneity between the HRs was minimal, with a variance between the HRs, as measured by T^2, of 0.0107. In contrast, one SD increase in DBP was associated with a moderate risk of an initial presentation with stroke, acute MI/ coronary death and AAA, but not PAD. Ignoring PAD, the lowest HR for an increase in DBP was acute MI/coronary death and the highest with AAA. Again, heterogeneity between the HRs for DBP presentations was limited (T^2= 0.0060). These results are shown in Figure 29.

The inclusion of additional variables in the multivariable Cox proportional hazard model had minimal effect on the size of the HR except for HR for DBP and PAD, where the HR and 95% CI increased above 1 (Figure 30). Adding a random effect variable for GP practice into the model also had minimal impact on the effect estimates (Figure 31).

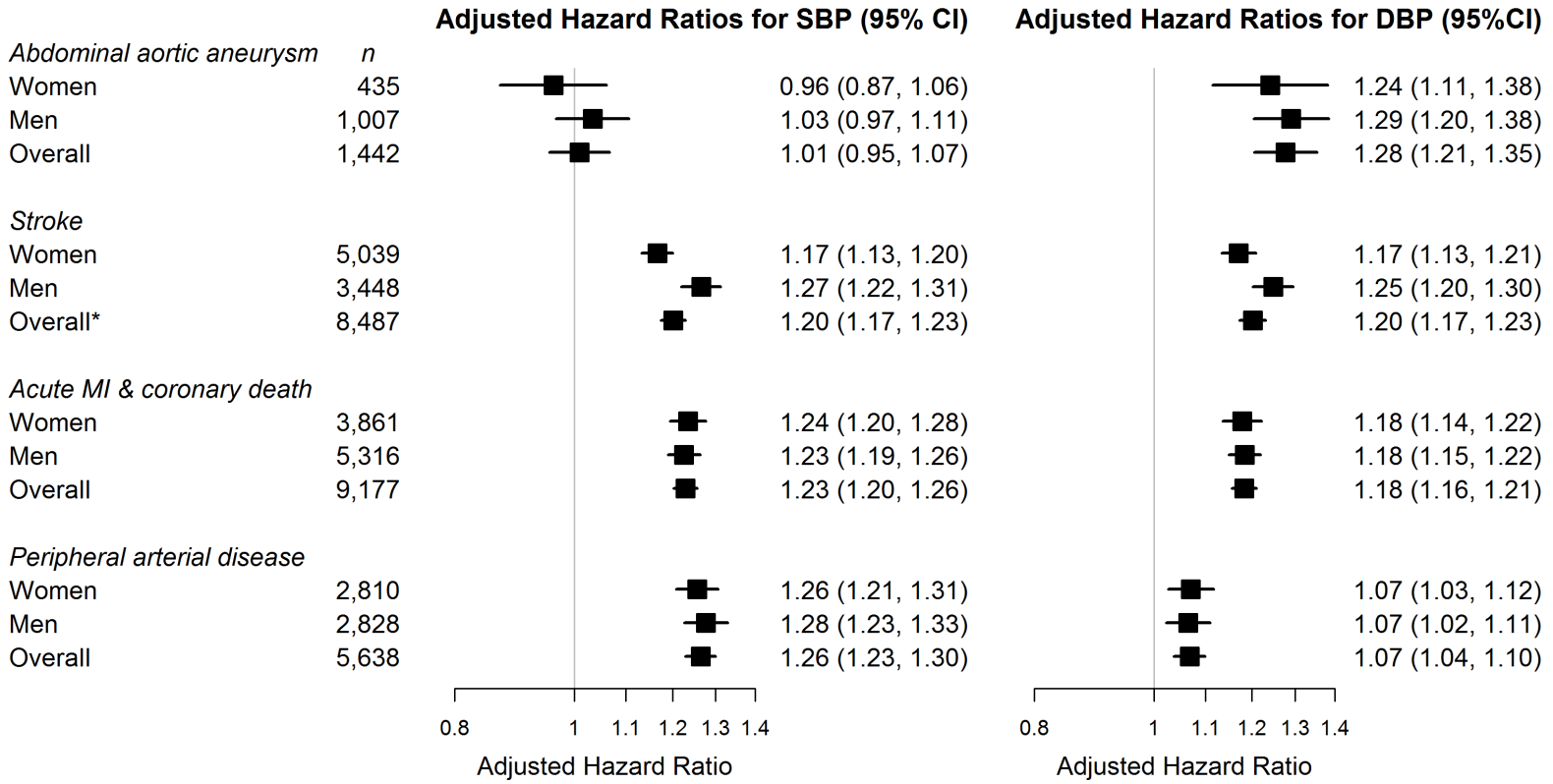
Given the sizeable differences in effect estimates of the association with peripheral arterial disease between SBP and DBP, in a post-hoc analysis I examined whether this finding could be the result of outlying values by estimating a model with measurements below the 5th centile and above the 95th centile. Although the effect size was slightly reduced, the effect estimate for DBP remained substantially lower than that for SBP.

Figure 29: Age-adjusted hazard ratios for initial presentations of cardiovascular disease associated with 1 standard deviation increase in systolic and diastolic blood pressure overall and in women and men



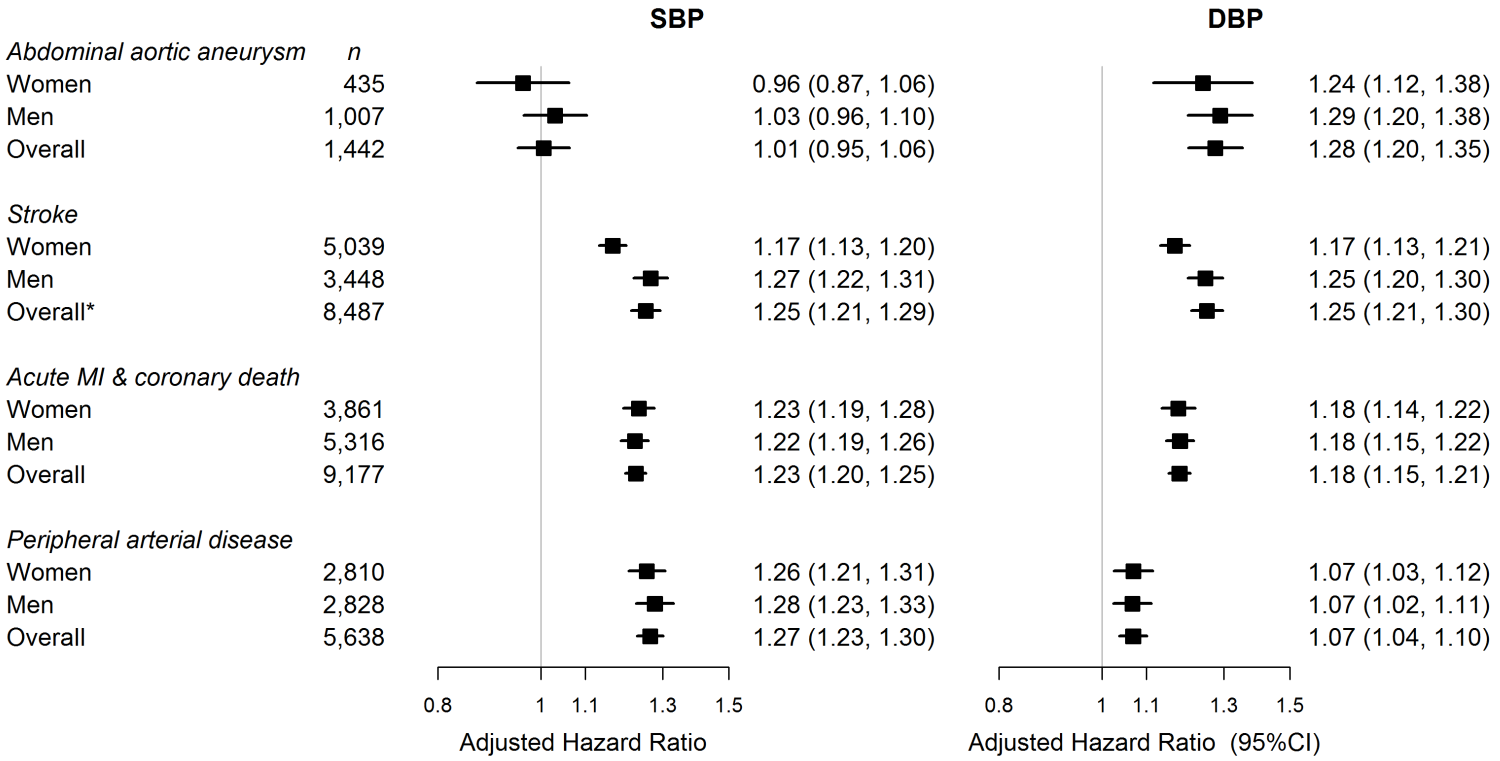
*Hazard ratios for mean systolic or diastolic blood pressure at baseline (per 1 standard deviation (sd)), adjusted for age at baseline in men and women, and additionally for sex in all patients, in complete cases. N= 897,892. For SBP, 1 sd = 18.9 mmHg; for DBP, 1 sd=10 mmHg. Endpoints ordered by size of HR in all patients. MI indicates myocardial infarction; CI, confidence interval; DBP, diastolic blood pressure; HR, hazard ratio; SBP, systolic blood pressure. *Sex-BP interaction not included in model used to estimate HR for figure.*

Figure 30: Multivariate adjusted hazard ratios for initial presentations of cardiovascular disease associated with 1 standard deviation increase in systolic and diastolic blood pressure overall and in women and men



*Hazard ratios for mean systolic or diastolic blood pressure at baseline (per 1 standard deviation (sd)), adjusted for age at baseline, sex, deprivation (index of multiple deprivation quintile), smoking status (current smoker, ex-smoker, non-smoker), diabetes mellitus (yes/no) and statin use at baseline (yes/no) , in complete cases. N= 897,892. For SBP, 1 sd = 18.9 mmHg; for DBP, 1 sd=10 mmHg. MI indicates myocardial infarction; CI, confidence interval; DBP, diastolic blood pressure; HR, hazard ratio; SBP, systolic blood pressure. *Sex-BP interaction not included in model used to estimate HR for figure.*

Figure 31: Multivariate adjusted hazard ratios (with random effects for GP practice) for initial presentations of cardiovascular disease associated with 1 standard deviation increase in systolic and diastolic blood pressure overall and in women and men



*Hazard ratios for mean systolic or diastolic blood pressure at baseline (per 1 standard deviation (sd)), adjusted for age at baseline, sex, deprivation (index of multiple deprivation quintile), smoking status (current smoker, ex-smoker, non-smoker), diabetes mellitus (yes/no) and statin use at baseline (yes/no), , with frailty term to take account of clustering at practice level, in complete cases. N= 897,892. For SBP, 1 sd = 18.9 mmHg; for DBP, 1 sd=10 mmHg. MI indicates myocardial infarction; CI, confidence interval; DBP, diastolic blood pressure; HR, hazard ratio; SBP, systolic blood pressure. *Sex-BP interaction not included in model used to estimate HR for figure.*

5.4.2. Modification of effect of SBP and DBP by gender

As can be seen from Figure 29 to Figure 31 above, there were no gender differences in the association of SBP with the initial presentations of acute MI/coronary death and PAD, but men were at somewhat greater risk of stroke from increases in SBP than women. Increases in SBP appear to have a protective effect in women for AAA, but not in men; however, given the limited number of AAA events in women these results should be treated with caution. In both men and women, the association of SBP with these endpoints was heterogeneous, but with minimal variance in the association between SBP and the endpoints (men I^2= 97.7%, T^2= 0.0105; women I^2=98.0%, T^2=0.0111). Men were also at greater risk of stroke from increases in DBP compared to women, but the gender differences were smaller. There were no other gender differences in the association of DBP with the other endpoints. The extent of heterogeneity in the association of DBP with the endpoints was minimal in both men and women (men I^2= 96.7%, T^2= 0.0073; women I^2=94.4%, T^2=0.0045).

I explicitly tested for interaction between blood pressure and gender, across these endpoints. The hazard ratios for the interaction terms, with 95% confidence interval and p-value are shown in Appendix H, along with the results of the likelihood ratio test comparing the model with the interaction term to the model without. Stroke was the only endpoint where the interaction term was significant showing a smaller effect of increase in both systolic and diastolic blood pressure on the hazard of stroke in women compared to men. The inclusion of the interaction term marginally improved the fit of the model.

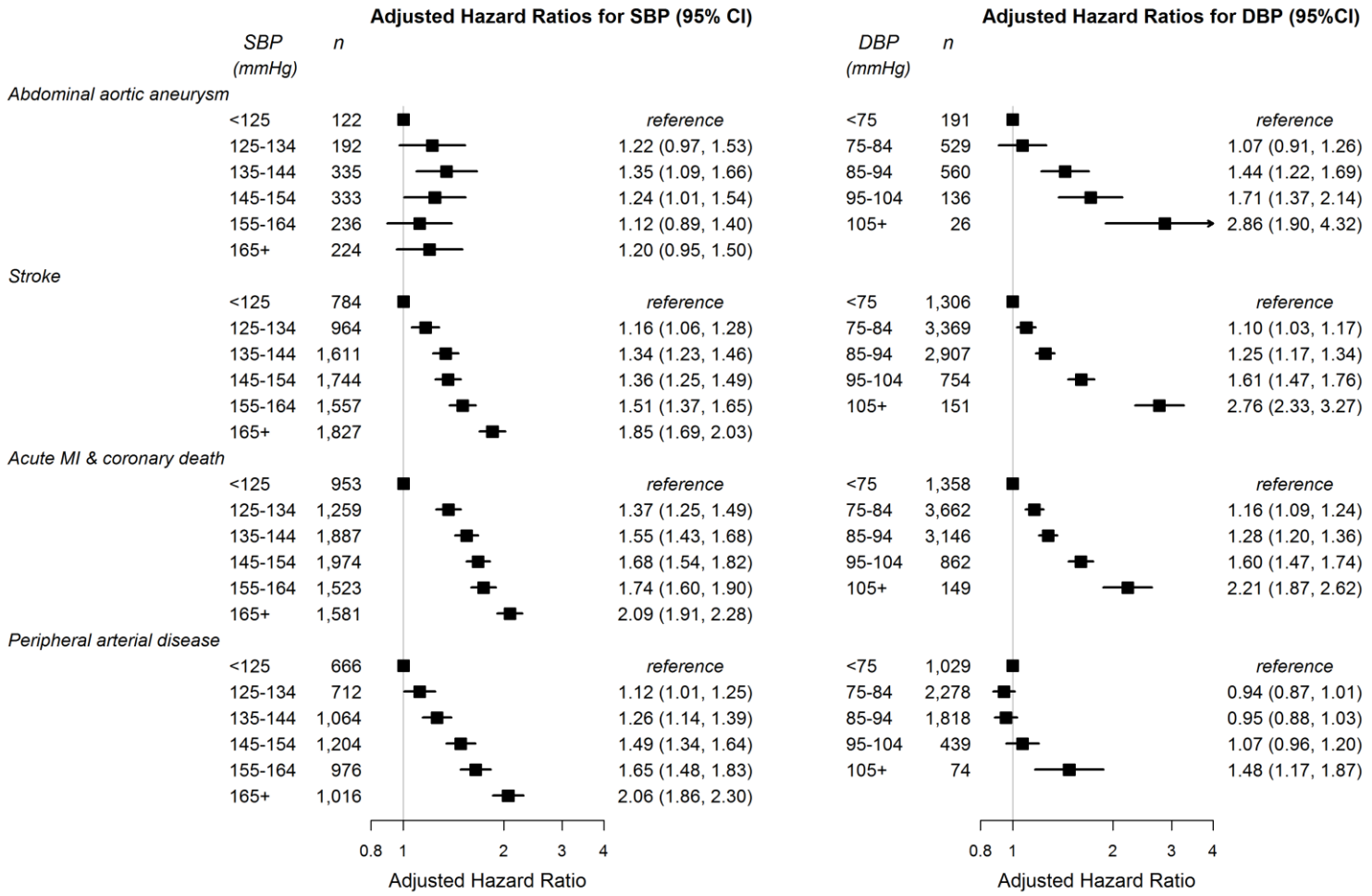
5.4.3. Categorical SBP and DBP variables

Similar results were found when the associations of SBP and DBP with these endpoints were modelled using categorical blood pressure measures, with categories centred on the ten-digit numbers, rather than continuous measures of SBP and DBP. There was an increasing trend with each increase in BP category across the endpoints where the continuous variable showed an increased hazard (i.e. for stroke and AMI/coronary death with SBP and DBP, for PAD with SBP and for AAA with DBP) and no increasing trend where there was no positive association with the continuous BP variable and endpoint. The hazard between the lowest and the highest categories of SBP doubled for stroke, acute MI/coronary death and PAD. The hazard of AAA and stroke tripled from lowest to highest category of DBP and doubled for acute MI/coronary death. Additionally, there was an increased hazard for all endpoints for the highest categories of both SBP and DBP (Figure 32). As with the continuous SBP and DBP variables, there was little change to the estimates with the inclusion of additional variables into the model (Figure 33).

The gender difference identified with continuous blood pressure measure was replicated in these analyses, but the excess hazard for stroke in men was restricted to the highest BP categories (165+ mmHG SBP and 105+ mmHg DBP).

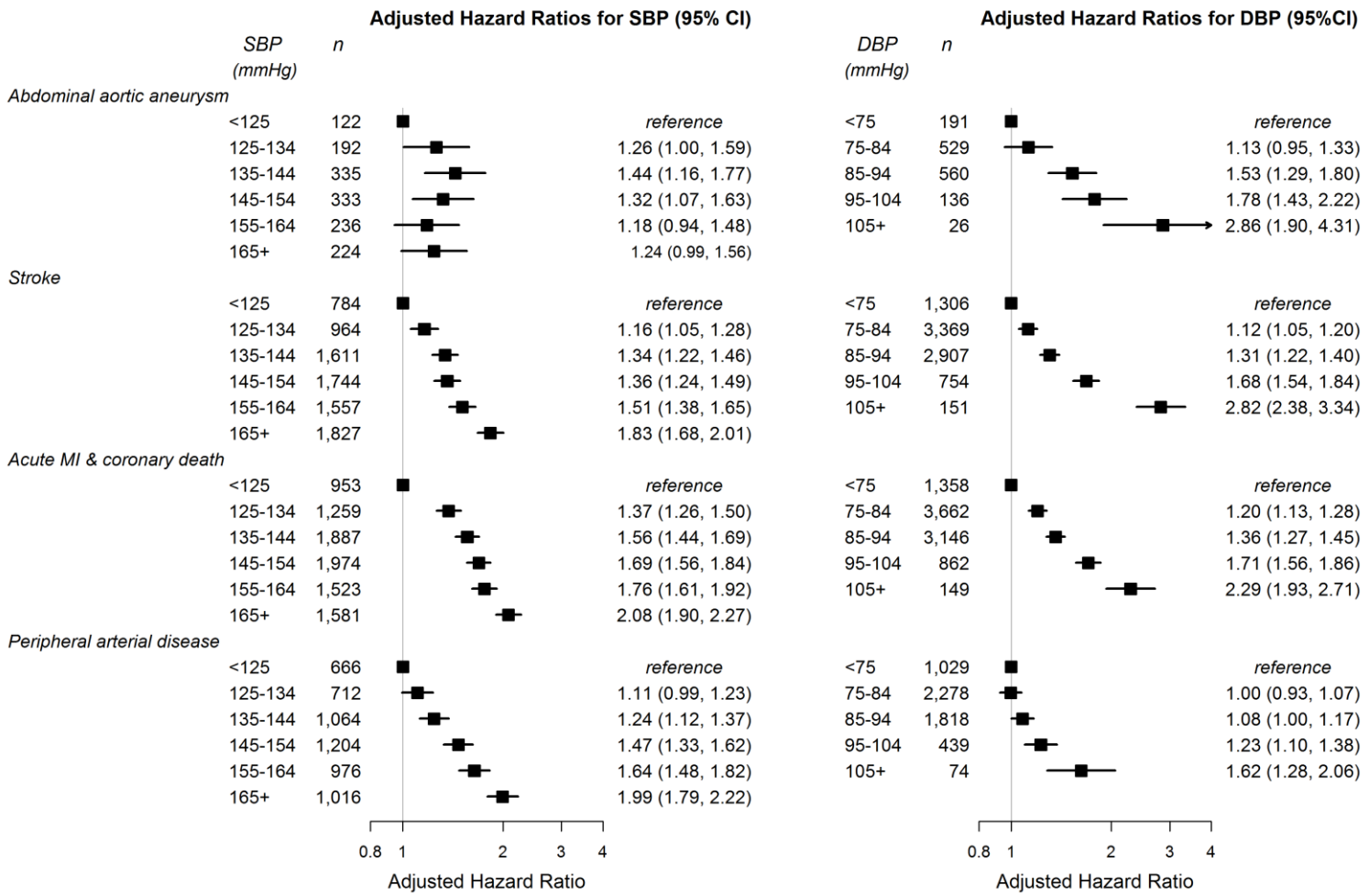
Given the linear trend evident in these categorical analyses, no attempt was made to improve the fit to the data with inclusion of quadratic terms for SBP and DBP.

Figure 32: Age-sex adjusted hazard ratios for initial presentations of cardiovascular disease by baseline systolic and diastolic blood pressure categories



Hazard ratios for blood pressure category compared to lowest category (SBP: <125 mmHg; DBP: <75 mmHg), adjusted for age at baseline and gender, in complete cases. N= 897,892. MI indicates myocardial infarction; CI, confidence interval; DBP, diastolic blood pressure; SBP, systolic blood pressure.

Figure 33: Multivariable adjusted hazard ratios for initial presentations of cardiovascular disease for baseline systolic and diastolic blood pressure categories



Hazard ratios for blood pressure category compared to lowest category (SBP: <125 mmHg; DBP: <75 mmHg), adjusted for age at baseline, gender, deprivation (index of multiple deprivation quintile), smoking status (current smoker, ex-smoker, non-smoker), diabetes mellitus (yes/no) and statin use at baseline (yes/no) , in complete cases. N= 897,892. MI indicates myocardial infarction; CI, confidence interval; DBP, diastolic blood pressure; SBP, systolic blood pressure.

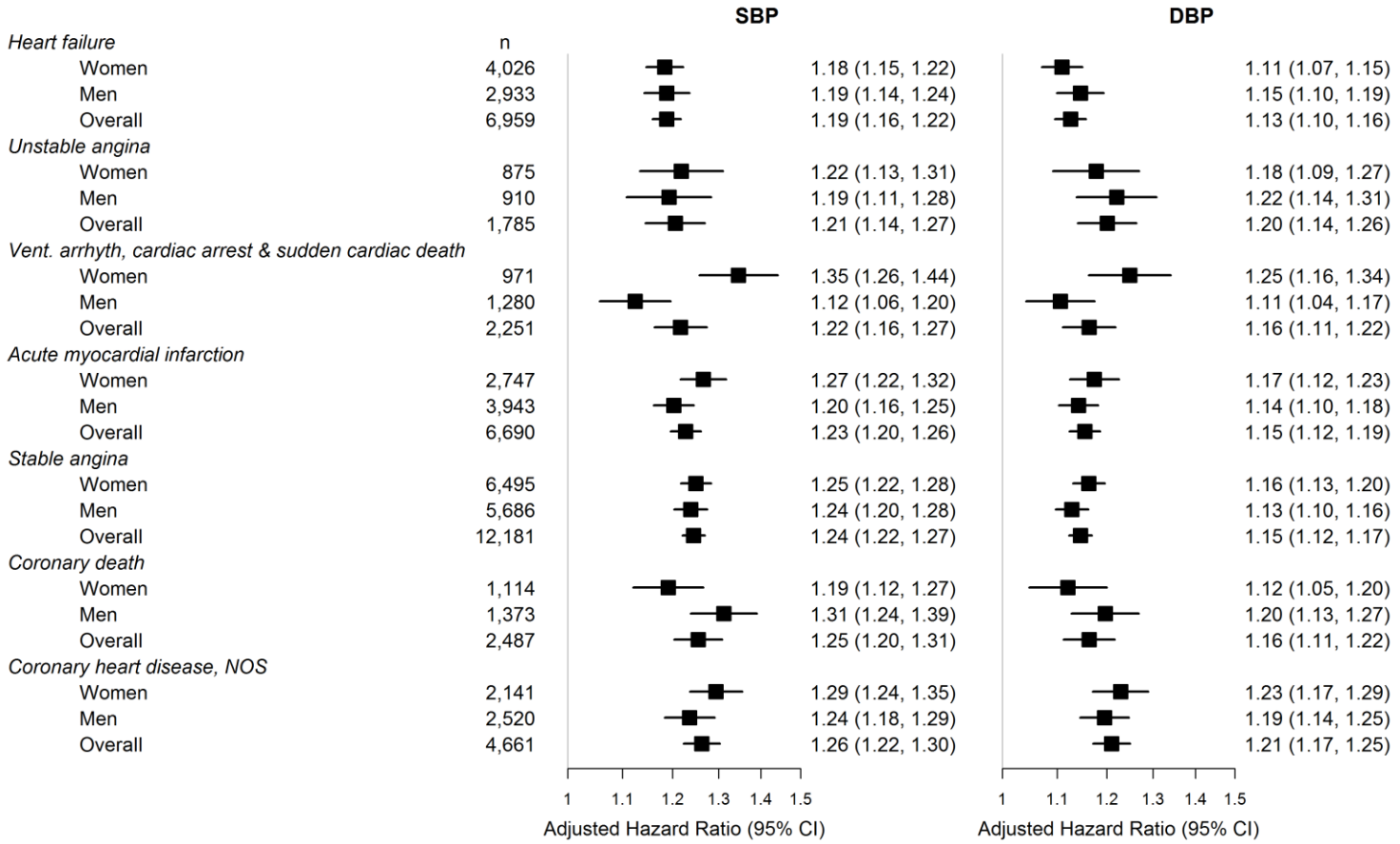
5.5. Association of blood pressure with cardiac diseases

5.5.1. Overall

As with cardiovascular endpoints across all four arterial beds, there was a modest increased risk (20-25%) with each standard deviation increase in SBP in the age-adjusted model. For most presentations, there was a slightly smaller increased hazard for DBP (~15%) with the exception of unstable angina and CHD NOS where the hazard was the same between SBP and DBP. However, all these differences were small (Figure 34). There was also little heterogeneity in the association between blood pressure (both systolic and diastolic) and the different cardiac endpoints for patients overall, reflected in the low Tau-squared for both SBP and DBP (SBP: I^2= 98.9%, T^2= 0.0138; DBP: I^2= 97.9%, T^2= 0.0079).

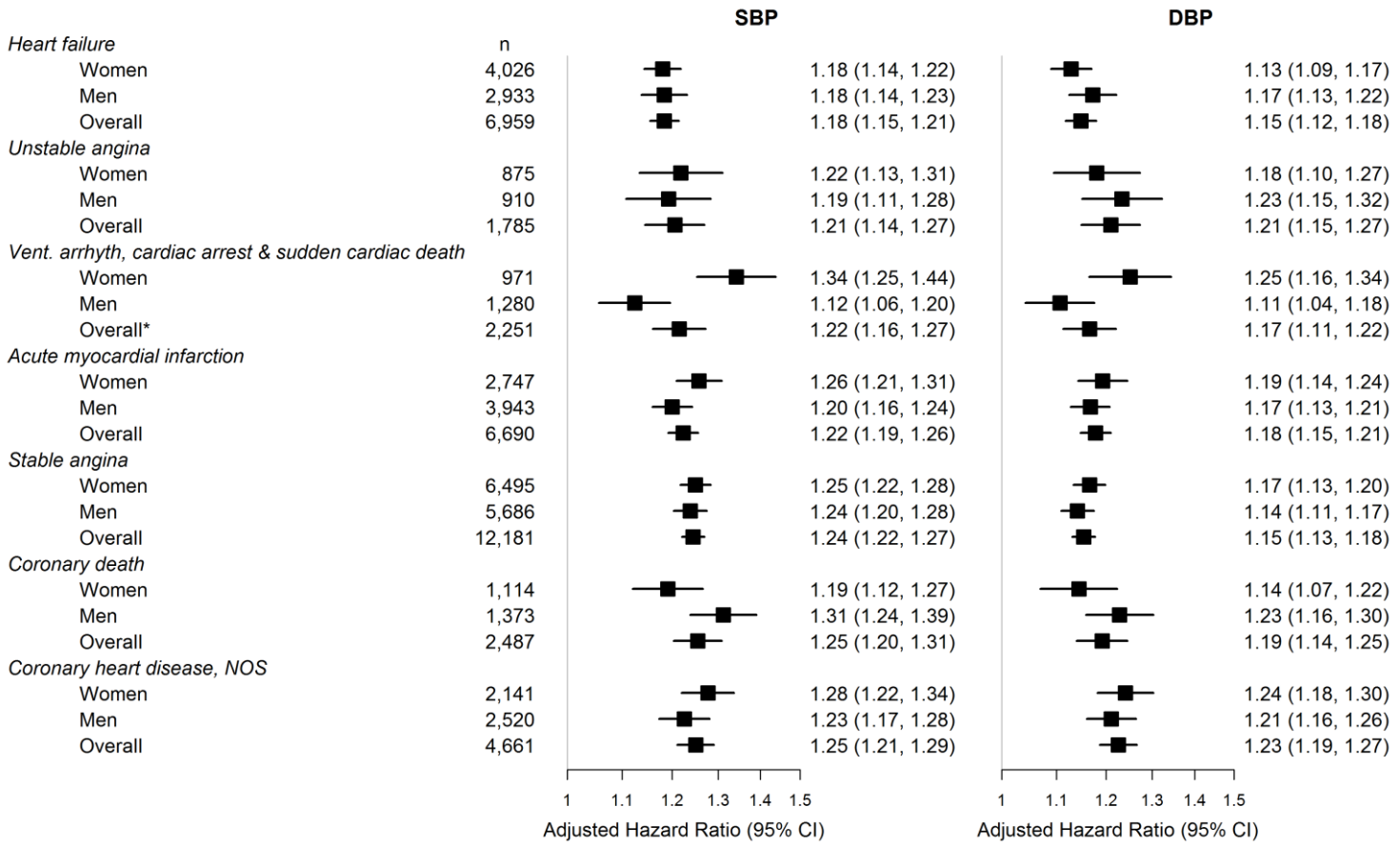
As with stroke, AAA, PAD and the composite CHD endpoint, the multivariable adjusted model slightly attenuated the effect estimates across all endpoints (Figure 35). Inclusion of a random effect variable for GP practice in the model had minimal impact on the effect estimates (Figure 36below).

Figure 34: Age-adjusted hazard ratios for initial presentations of cardiac disease associated with 1 standard deviation increase in systolic and diastolic blood pressure overall and in women and men



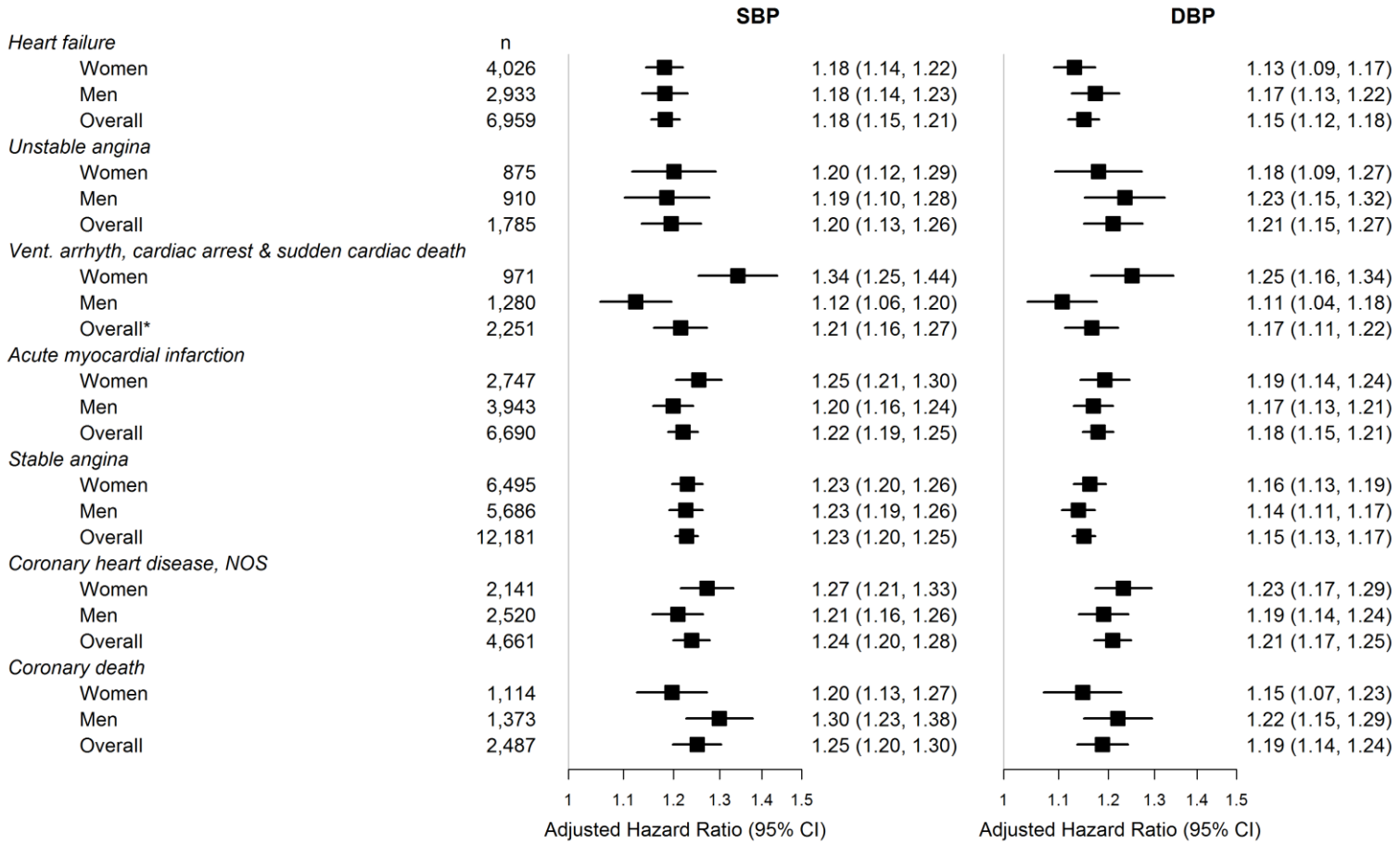
Hazard ratios for mean systolic or diastolic blood pressure at baseline (per 1 standard deviation (sd)), adjusted for age at baseline and gender, in complete cases. N= 897,892. For SBP, 1 sd = 18.9 mmHg; for DBP, 1 sd=10 mmHg. CI indicates confidence interval; DBP, diastolic blood pressure; NOS, not otherwise specified; SBP, systolic blood pressure; Vent arrhyth, ventricular arrhythmias.

Figure 35: Multivariable adjusted hazard ratios for initial presentations of cardiac disease associated with 1 standard deviation increase in systolic and diastolic blood pressure overall and in women and men



*Hazard ratios for mean systolic or diastolic blood pressure at baseline (per 1 standard deviation (sd)), adjusted for age at baseline, sex, deprivation (index of multiple deprivation quintile), smoking status (current smoker, ex-smoker, non-smoker), diabetes mellitus (yes/no) and statin use at baseline (yes/no), in complete cases. N= 897,892. For SBP, 1 sd = 18.9 mmHg; for DBP, 1 sd=10 mmHg. MI indicates myocardial infarction; CI, confidence interval; DBP, diastolic blood pressure; SBP, systolic blood pressure. *Sex-BP interaction not included in model used to estimate HR for figure.*

Figure 36: Multivariable adjusted hazard ratios (with random effects for GP practice) for initial presentations of cardiac disease associated with 1 standard deviation increase in systolic and diastolic blood pressure overall and in women and men



*Hazard ratios for mean systolic or diastolic blood pressure at baseline (per 1 standard deviation (sd)), adjusted for age at baseline, sex, deprivation (index of multiple deprivation quintile), smoking status (current smoker, ex-smoker, non-smoker), diabetes mellitus (yes/no) and statin use at baseline (yes/no) with frailty term for general practice, in complete cases. N= 897,892. For SBP, 1 sd = 18.9 mmHg; for DBP, 1 sd=10 mmHg. CI indicates confidence interval; DBP, diastolic blood pressure; SBP, systolic blood pressure; NOS, not otherwise specified. *Sex-BP interaction not included in model used to estimate HR for figure.*

5.5.2. Modification of effect of SBP and DBP by gender

The effect of SBP and DBP on the ventricular arrhythmias was clearly modified by gender, with a substantially increased hazard for women compared to men from increases in both SBP and DBP. Other gender differences were less clear, although men appeared to have a greater hazard than women of unheralded coronary death from increases in SBP and DBP and of unstable angina from increases in DBP only. Women appeared to have a greater hazard than men of acute MI and CHD NOS from increases in SBP only.

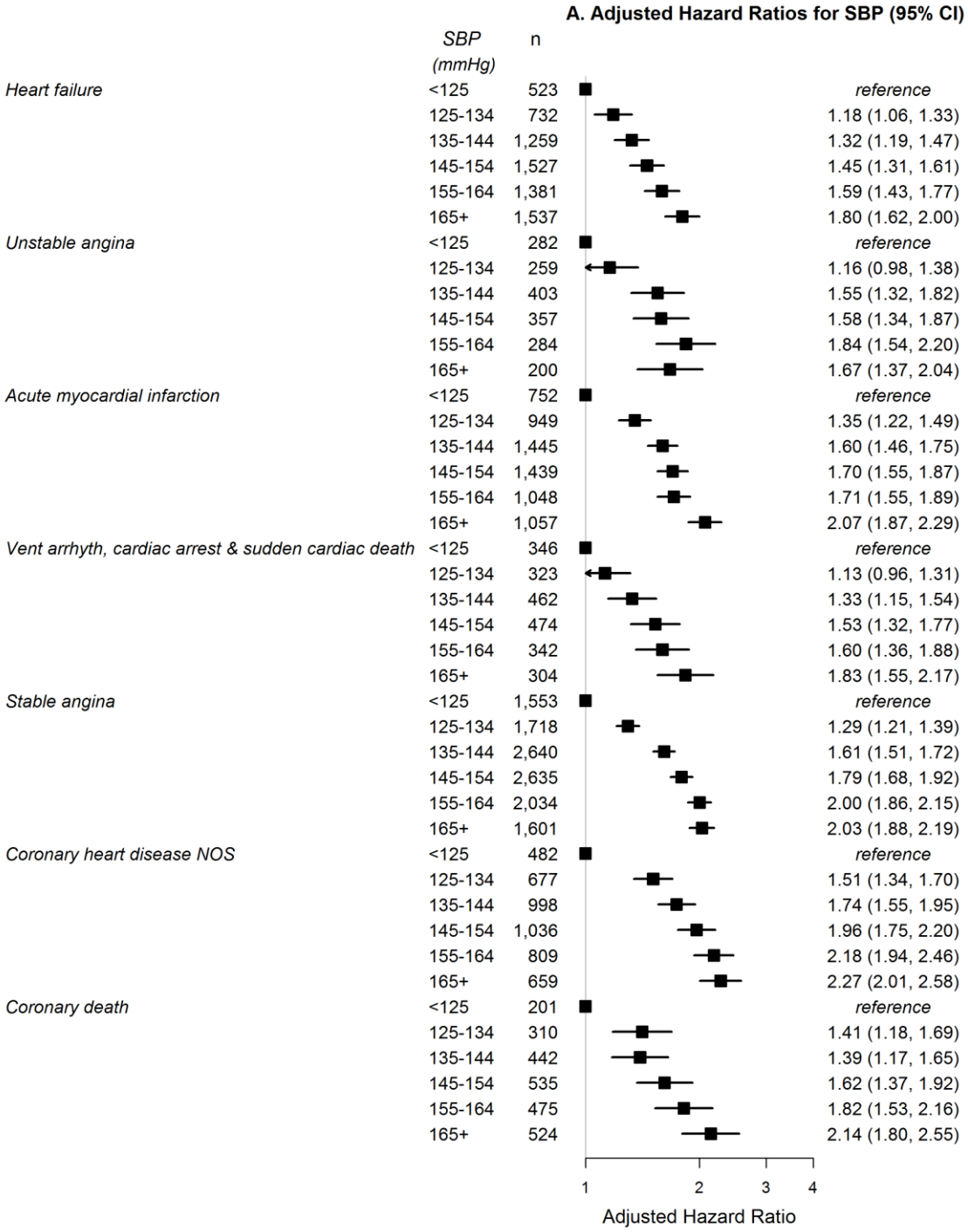
In my modelling of association of SBP and DBP with the specific cardiac endpoints, the interaction term with gender and BP was only significant for the arrhythmias, showing increased risk for women compared to men with each standard deviation change in SBP, and to a lesser extent for DBP. See Appendix H for the HR for the interaction term and the likelihood ratio (LR) test values and p values for the model fit comparing the model with the interaction term and without.

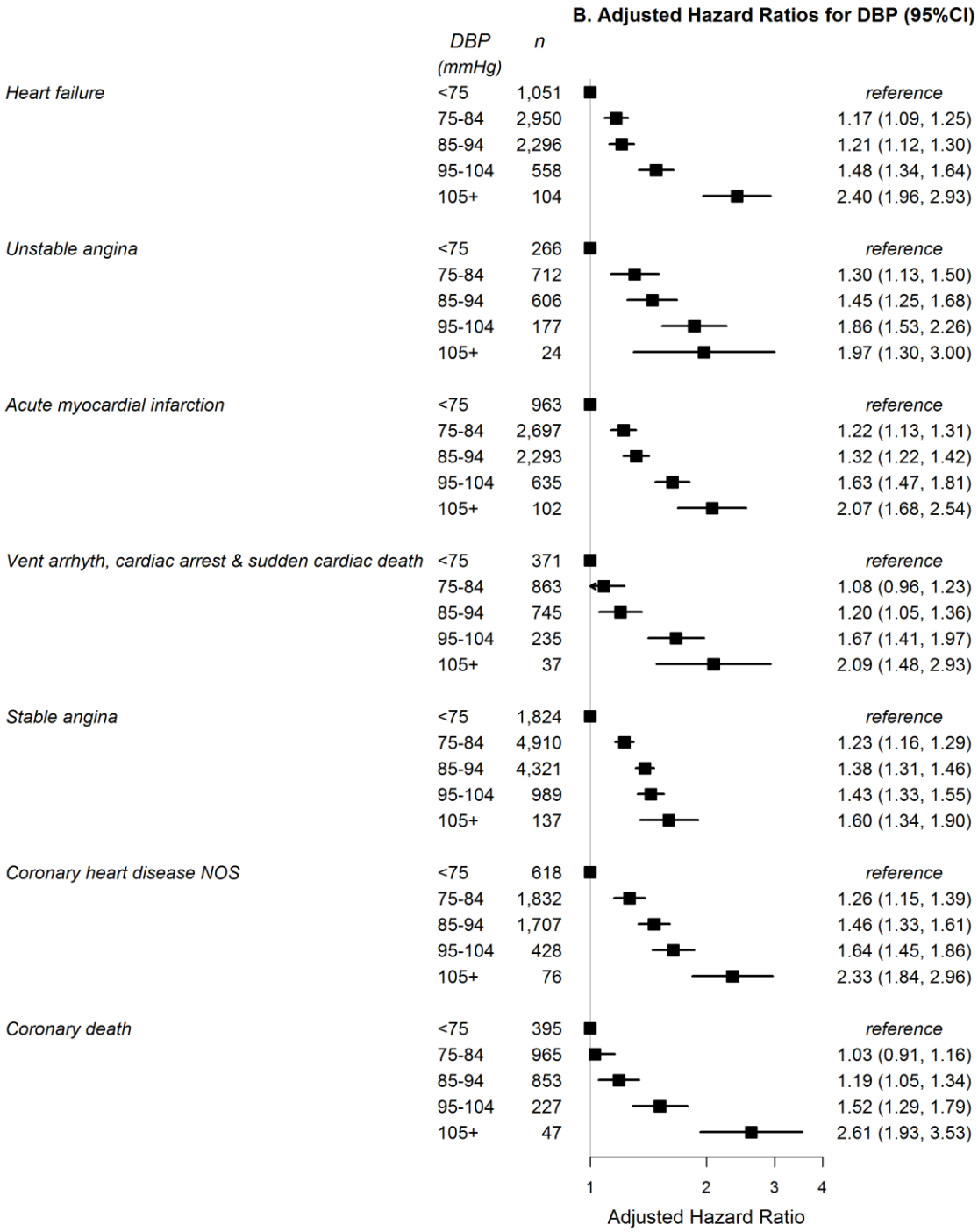
5.5.3. Categorical SBP and DBP variables

The hazard of SBP increased twofold from a SBP below <125 mmHg to a SBP 165+ mmHg across most presentations, with a broadly linear progression for each 10 mmHg increase. The increase in hazard from lowest to highest SBP was smallest for unstable angina (67%) and highest for CHD NOS (227%). With the stable angina and CHD NOS endpoints, the rate of increased appeared to slow above 154 mmHg. For DBP, the hazard also doubled from the lowest (<75 mmHg) to the highest (105+) category, again with a broadly linear progression for each 10 mmHg increase. The smallest increase was for stable angina (60%) and the largest was for heart failure (240%). (See Figure 37.)

There was little change in these effects with the addition of other variables to the model. (Figure 38) These results are similar to those found when the association of SBP and DBP with the cardiac endpoints was modelled using continuous variables. As with the continuous SBP variables, SBP had greater association with arrhythmias in women than in men, although only at SBP of 145 mmHg or greater. There were no gender differences in the association of DBP with cardiac endpoints.

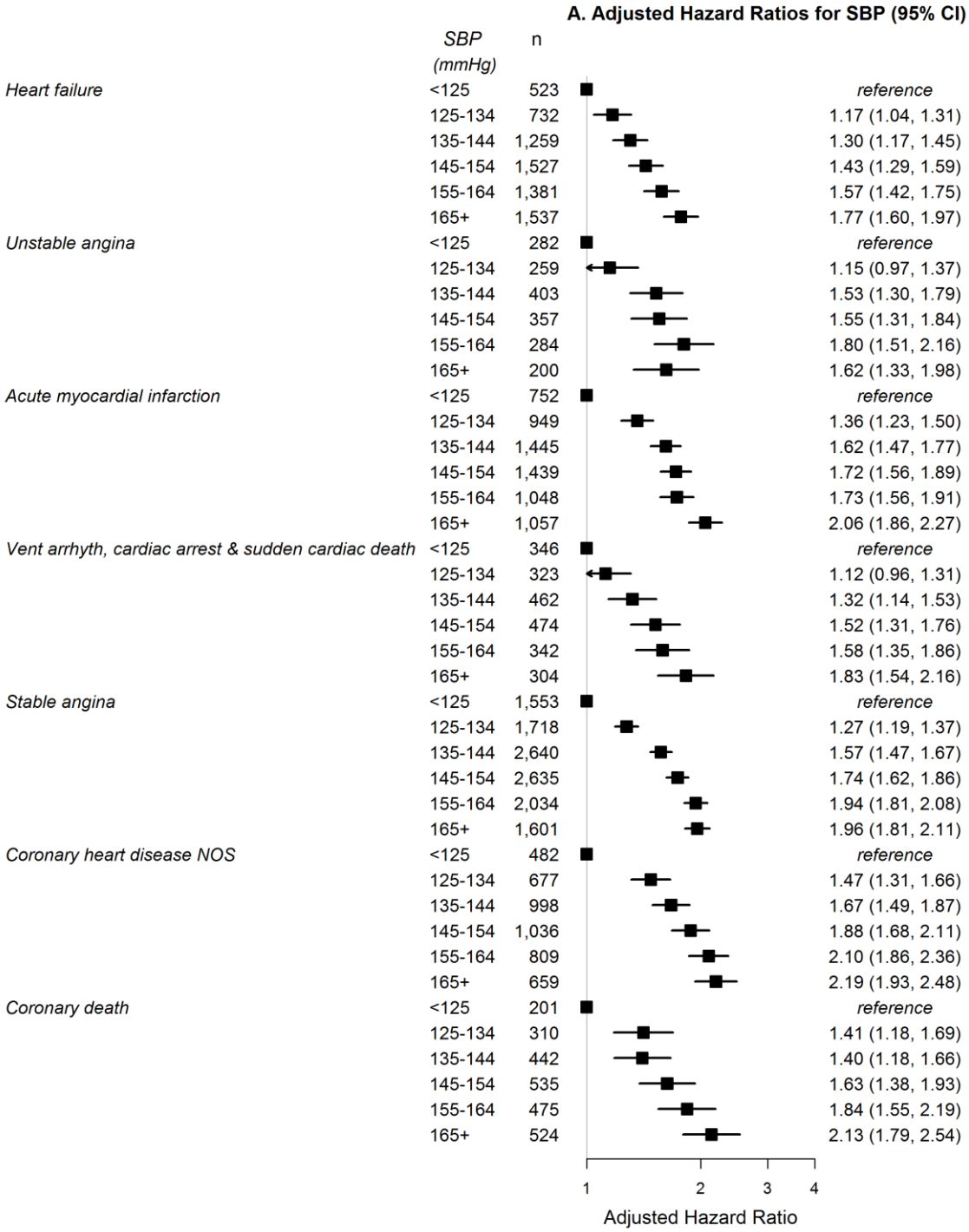
Figure 37: Age-sex adjusted hazard ratios for initial presentations of cardiac disease associated with categorical increase in baseline systolic blood pressure (A) and diastolic blood pressure (B)

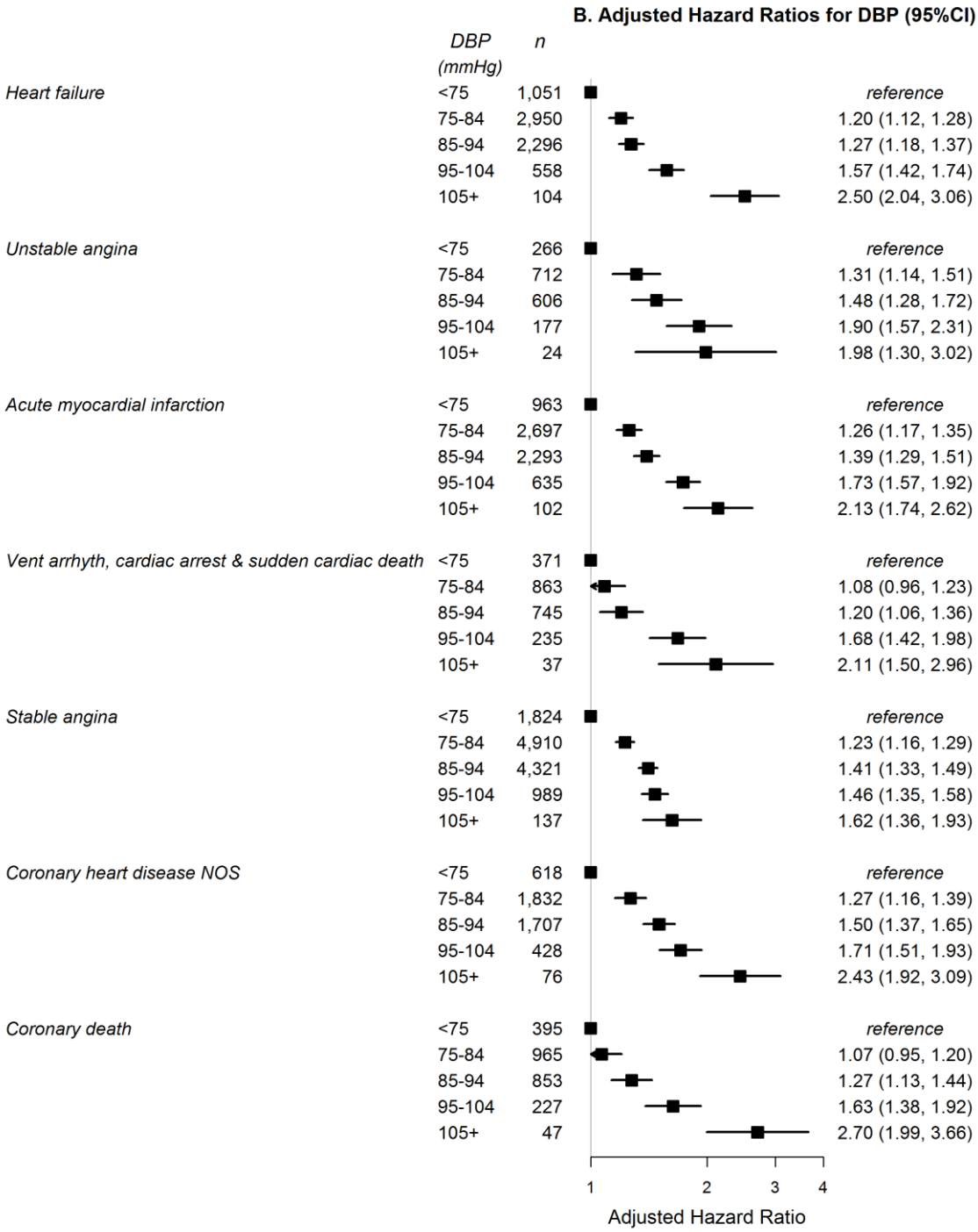




Hazard ratios for blood pressure category compared to lowest category (SBP: <125 mmHg; DBP: <75 mmHg), adjusted for age at baseline and gender, in complete cases. N= 897,892. CI indicates confidence interval; DBP, diastolic blood pressure; NOS, not otherwise specified; SBP, systolic blood pressure; Vent arrhyth, ventricular arrhythmias.

Figure 38: Multivariable adjusted hazard ratios for initial presentations of cardiac disease associated with categorical increased in baseline systolic blood pressure (A) and diastolic blood pressure (B)





Hazard ratios for blood pressure category compared to lowest category (SBP: <125 mmHg; DBP: <75 mmHg), adjusted for age at baseline, sex, deprivation (index of multiple deprivation quintile), smoking status (current smoker, ex-smoker, non-smoker), diabetes mellitus (yes/no) and statin use at baseline (yes/no) , in complete cases. N= 897,892. CI indicates confidence interval; DBP, diastolic blood pressure; NOS, not otherwise specified; SBP, systolic blood pressure; Vent arrhyth, ventricular arrhythmias.

5.6. Association of blood pressure with specific AMI subtypes

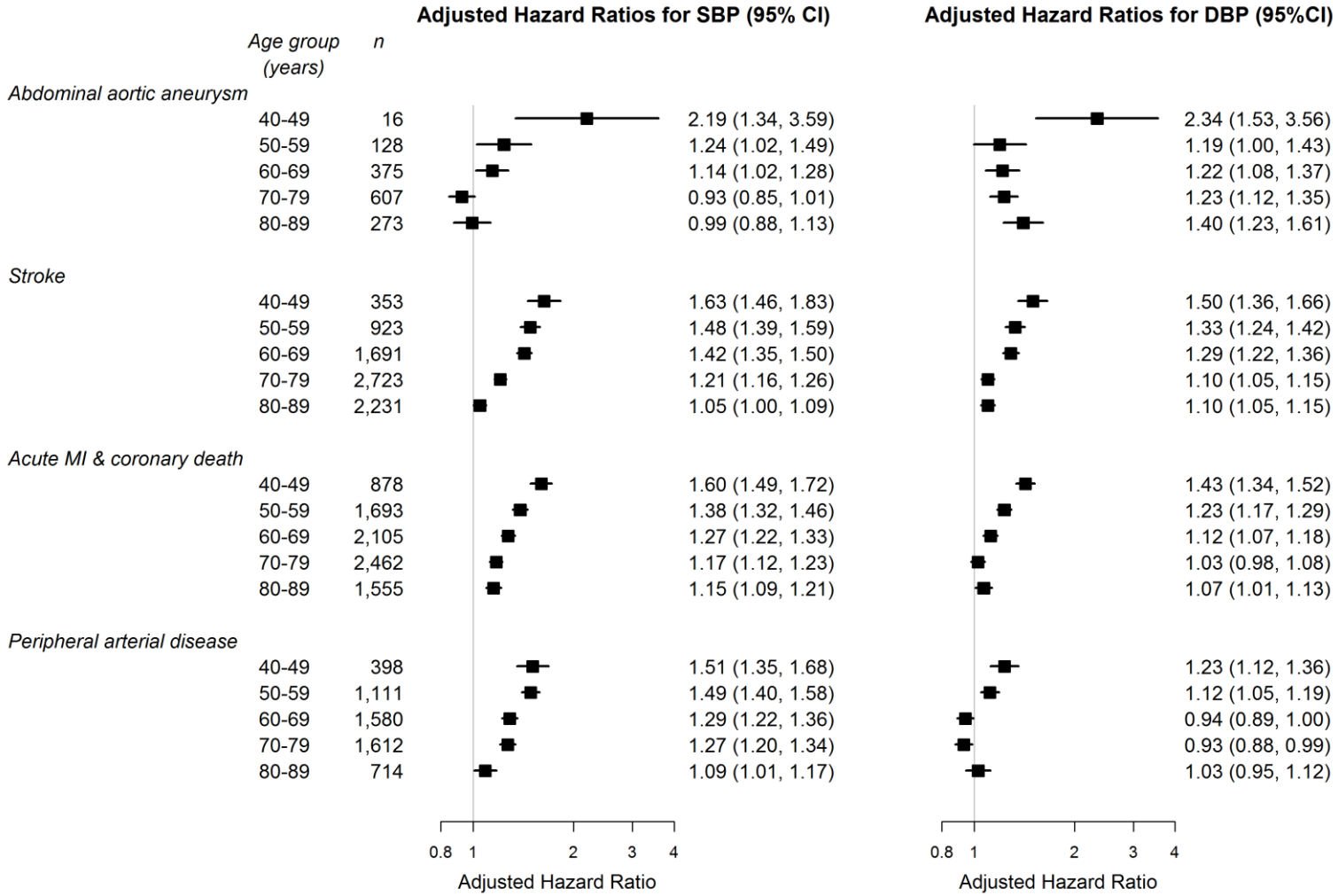
When acute myocardial infarction is divided into the constituent endpoints of STEMI, nSTEMI and MI NOS, the pattern of homogeneity of effect among endpoints, no gender interaction and similar effect size between diastolic and systolic seen with acute MI overall remains unchanged. (See Appendix H for figures.) No interaction effects between gender and either systolic or diastolic blood pressure were significant. (See Appendix H for table with interaction effects)

5.7. Modification of Association by Age

The strength of association with increases in both SBP and DBP and the initial CVD presentations of stroke, AMI and UCD, and peripheral arterial disease declined with each additional decade of age up to age 70-79; thereafter there was a slight increase in the strength of association, particularly with DBP. The modification by age of the association between SBP and DBP with AAA was less clear, but is potentially also consistent with a linear decrease up to age 70-79. (See Figure 41 below.) Age modified the effect of SBP and DBP on stroke to the greatest extent and on peripheral arterial disease to the least extent, but these differences were small. The impact of age on the association of SBP and DBP with these endpoints remained unchanged in the multivariable model, although the effect estimates were attenuated, particularly for peripheral arterial disease. (See Figure 42 below.)

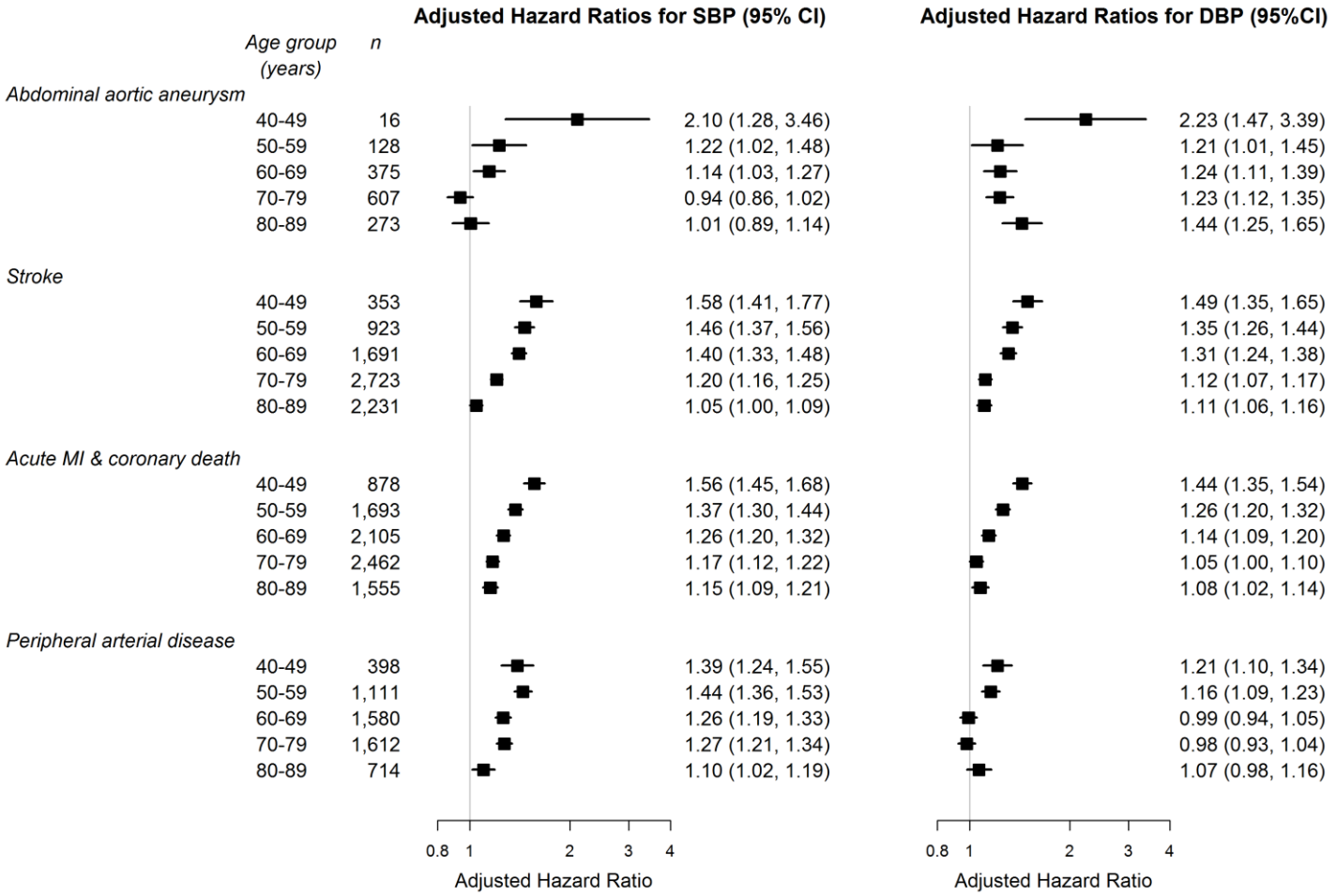
Among most of the cardiac presentations, there was again a linear decreasing association between SBP and DBP and all the endpoints with each decade of age up to age 70-79, with slight increase in the association above age 80, particularly for DBP. The exceptions to this pattern were unstable angina and unheralded coronary death where the association continued to decline above age 70-79. There was greater heterogeneity in the modification of SBP and DBP by age among the cardiac presentations than among the other CVD presentations, with more marked effects of an increase in blood pressure on unheralded coronary death and heart failure in younger age groups, compared to the other presentations. Estimates were slightly attenuated in the multivariable model, but otherwise the patterns of associations remained unchanged.

Figure 39: Sex adjusted hazard ratios for initial presentations of cardiovascular disease associated with 1 standard deviation increase in systolic and diastolic blood pressure, stratified by age group



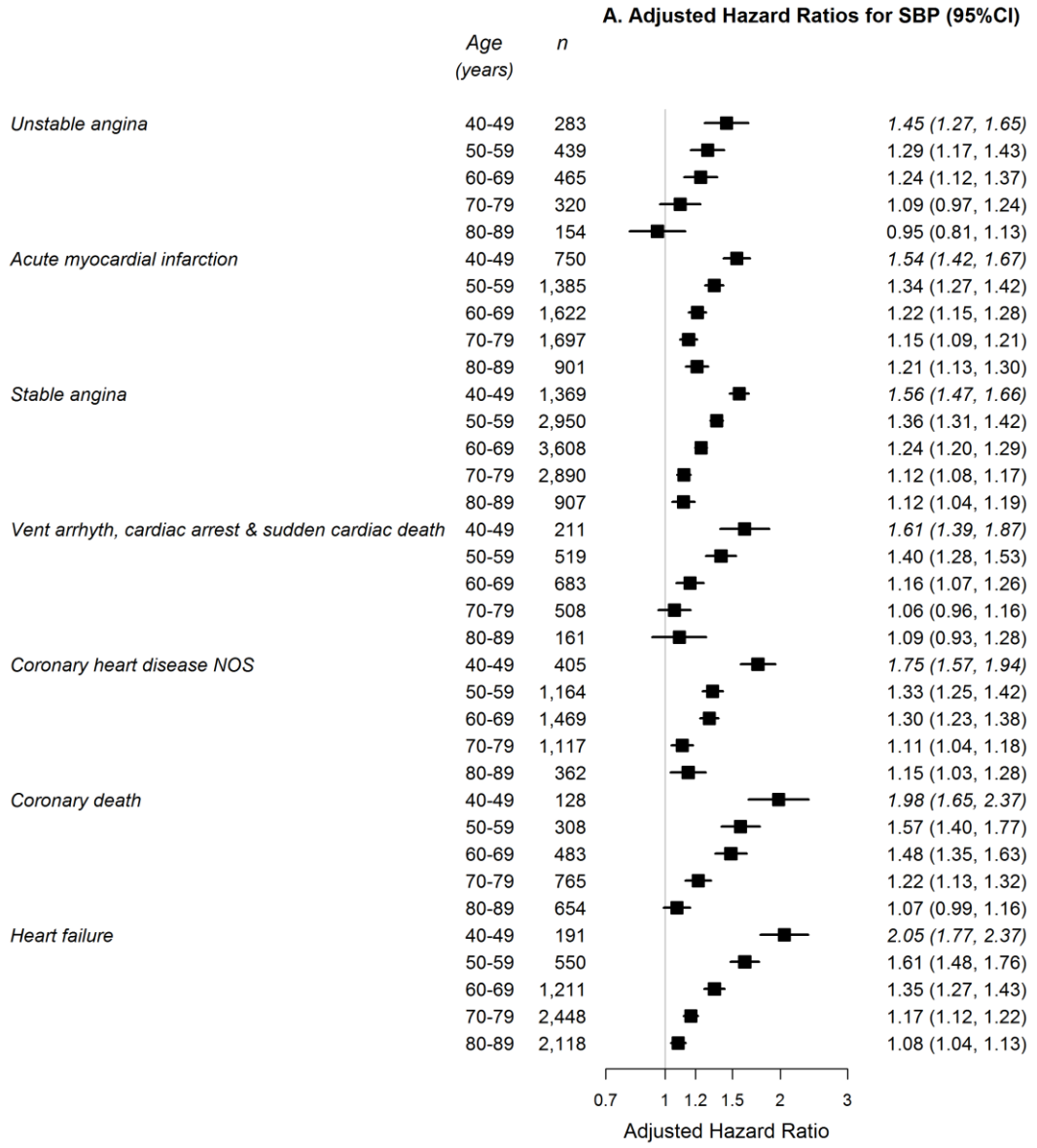
Hazard ratios for mean systolic or diastolic blood pressure at baseline (per 1 standard deviation (sd)), adjusted for age at baseline and gender, in complete cases, stratified by age group. HRs for <40 and 90+ years not shown. N= 897,892. For SBP, 1 sd = 18.9 mmHg; for DBP, 1 sd=10 mmHg. MI indicates myocardial infarction; CI, confidence interval; DBP, diastolic blood pressure; SBP, systolic blood pressure.

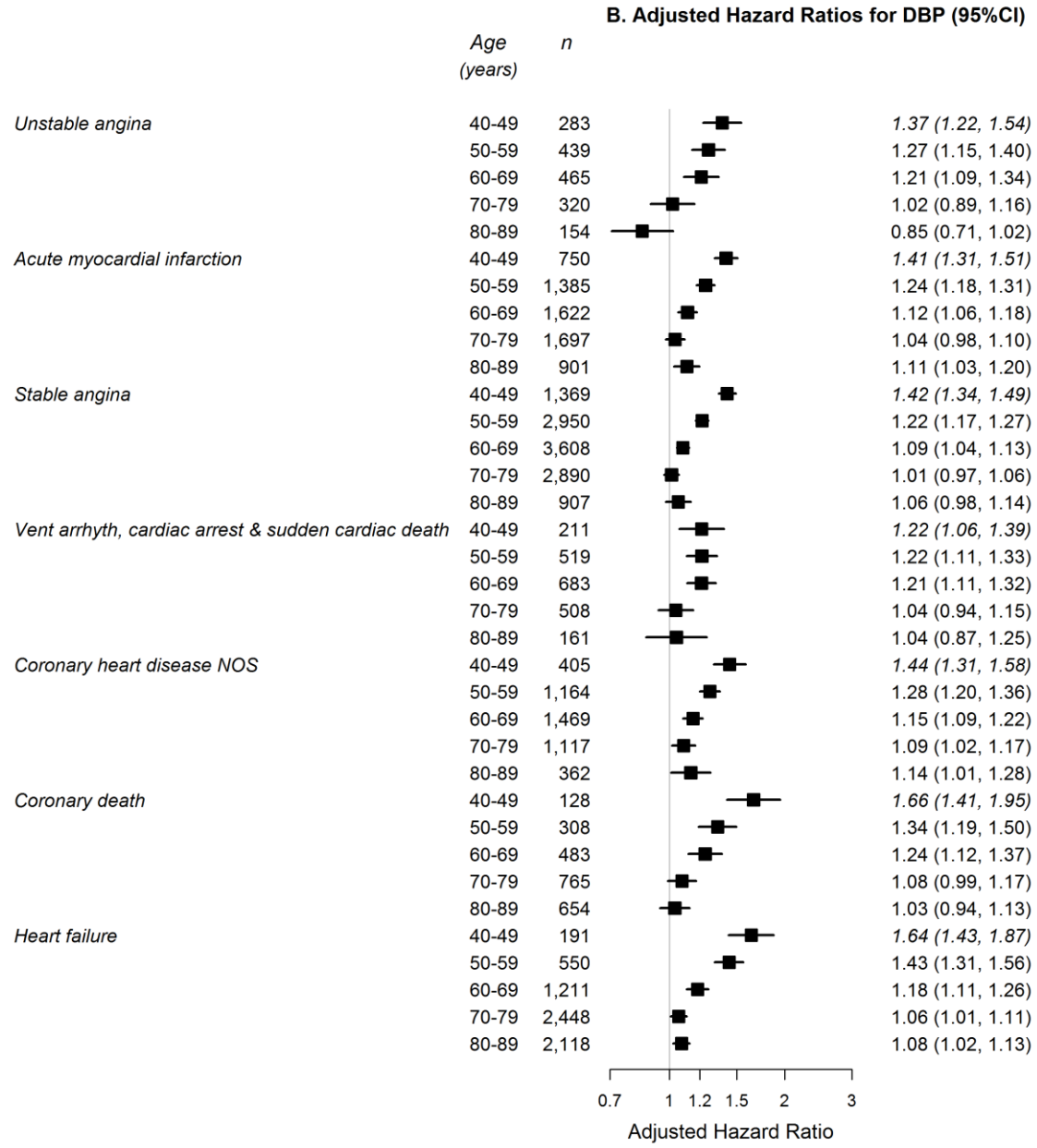
Figure 40: Multivariable adjusted hazard ratios for initial presentations of cardiovascular disease associated with 1 standard deviation increase in systolic and diastolic blood pressure, stratified by age group



Hazard ratios for mean systolic or diastolic blood pressure at baseline (per 1 standard deviation (sd)), adjusted for gender, deprivation (index of multiple deprivation quintile), smoking status (current smoker, ex-smoker, non-smoker), diabetes mellitus (yes/no) and statin use at baseline (yes/no), in complete cases, stratified by age group. HRs for <40 and 90+ years not shown. N= 897,892. For SBP, 1 sd = 18.9 mmHg; for DBP, 1 sd=10 mmHg. MI indicates myocardial infarction; CI, confidence interval; DBP, diastolic blood pressure; SBP, systolic blood pressure.

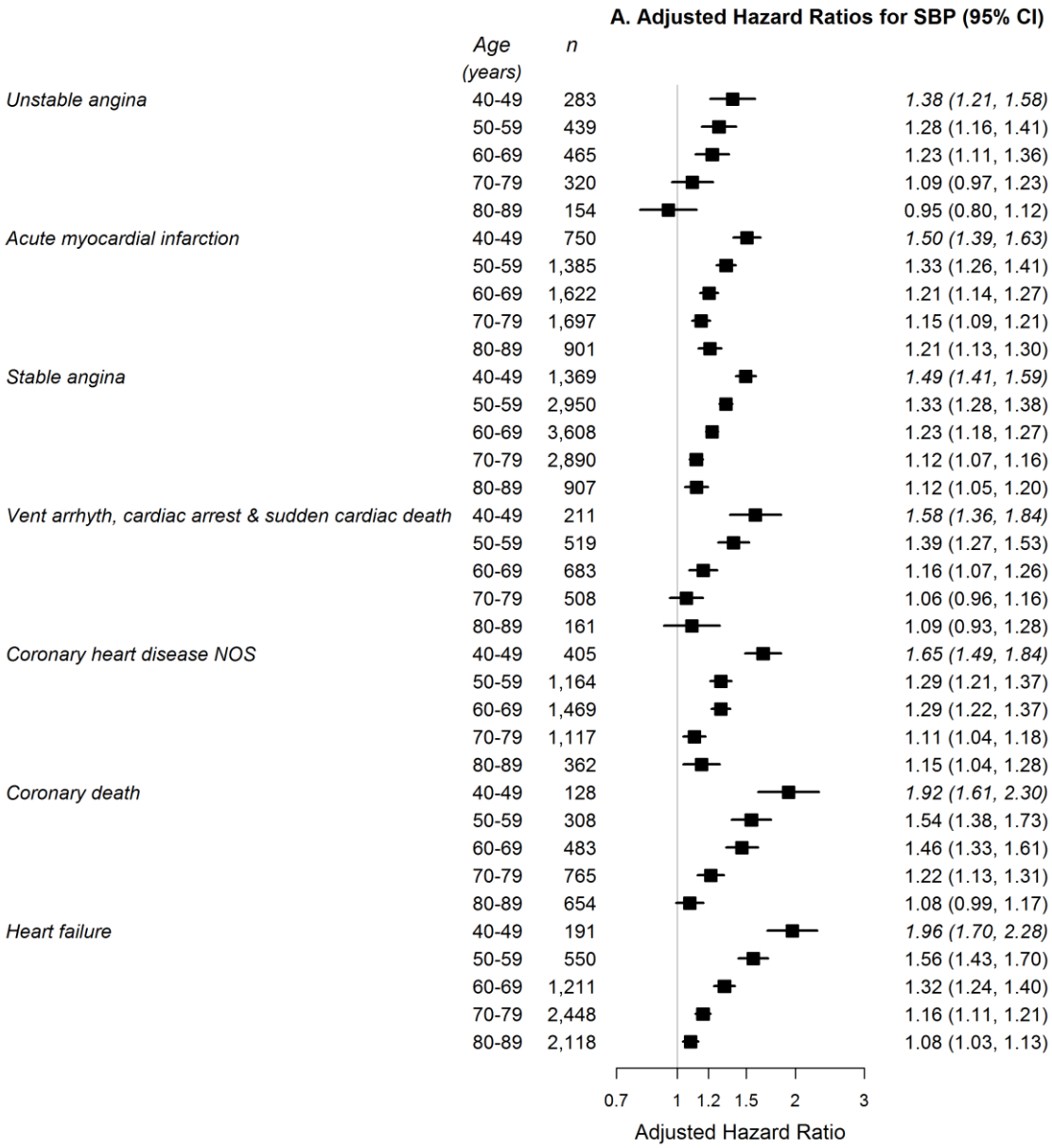
Figure 41: Hazard ratios, adjusted by gender, for initial presentations of cardiac phenotypes associated with 1 standard deviation increase in systolic blood pressure (A) and diastolic blood pressure (B), stratified by age group

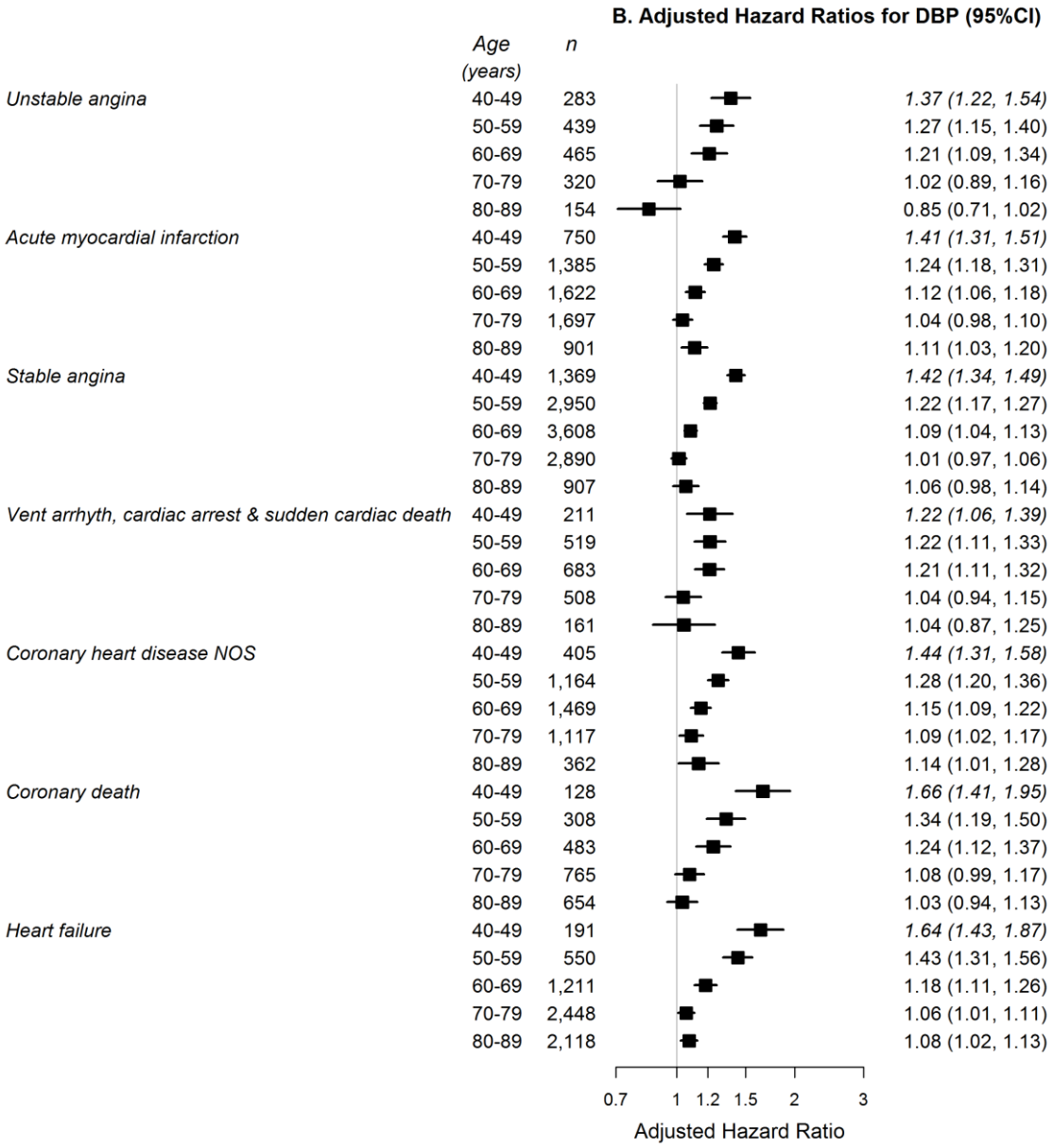




Hazard ratios for mean systolic or diastolic blood pressure at baseline (per 1 standard deviation (sd)),adjusted for age at baseline and gender, in complete cases, stratified by age group. HRs for <40 and 90+ years not shown. . N= 897,892. For SBP, 1 sd = 18.9 mmHg; for DBP, 1 sd=10 mmHg. CI indicates confidence interval; DBP, diastolic blood pressure; NOS, not otherwise specified; SBP, systolic blood pressure; Vent arrhyth, ventricular arrhythmias.

Figure 42: Multivariable adjusted hazard ratios for initial presentations of cardiac phenotypes associated with 1 standard deviation increase in systolic blood pressure (A) and diastolic blood pressure (B), stratified by age group





Hazard ratios for mean systolic or diastolic blood pressure at baseline (per 1 standard deviation (sd)), adjusted for age at baseline, sex, deprivation (index of multiple deprivation quintile), smoking status (current smoker, ex-smoker, non-smoker), diabetes mellitus (yes/no) and statin use at baseline (yes/no), in complete cases, stratified by age group. HRs for <40 and 90+ years not shown. N= 897,892. For SBP, 1 sd = 18.9 mmHg; for DBP, 1 sd=10 mmHg. CI indicates confidence interval; DBP, diastolic blood pressure; NOS, not otherwise specified; SBP, systolic blood pressure; Vent arrhyth, ventricular arrhythmias.

6. Discussion

I found in a cohort of almost 900,000 with more than 56,000 endpoints of 12 different types that SBP and DBP have broadly homogenous effects on most initial presentations of cardiovascular disease, with the exception of AAA and PAD. The increased hazard associated with increases in SBP and DBP were evident even at the lowest blood pressure categories. While the effect of one standard deviation increase in SBP and DBP was moderate, the increase in hazard between patients with the lowest BP compared to those

with highest were doubled for SBP and, for some presentations, tripled for DBP. There have been no other prospective cohorts in which such a range of initial presentations (both fatal and non-fatal) of cardiovascular disease have been assessed. Most cohorts do not take this approach; an 'incident' myocardial infarction, for example includes patients who have previously manifested other symptomatic cardiovascular disease. Understanding differences in the causes of onset versus the causes of progression of cardiovascular disease cannot be unpicked unless the onset of symptomatic cardiovascular disease is recognised across a range of potential presentations. The homogeneity in association with a wide range of initial presentations suggests a common physiological mechanism for the role of blood pressure in the onset of cardiovascular disease, regardless of initial presentation, with the exception of AAA and PAD.

In each section below, I summarise my findings for each objective. I then describe the strengths and limitations of this set of analyses. The implications for clinical care, public health practice, and research of my findings from these and my other analyses are discussed in Chapter 8 – Conclusions and Recommendations.

6.1. Summary of existing literature (Objective 1)

The limited evidence on the association between blood pressure and the initial presentation of CVD suggest a possible increased risk of unheralded coronary death and stroke compared to acute MI, with no difference in risk between cardiac diseases. However, the available literature is limited by inclusion of few endpoints, small number of participants, the use of hypertension as the only measure of blood pressure and a lack of sex-specific estimates of risk in both men and women. One limitation of the search strategy was the omission of epidemiological studies using data from clinical trials; inclusion of such studies in the initial search might have yielded additional studies for inclusion in the review.

6.2. Association of SBP and DBP with onset of symptomatic cardiovascular disease across a wide range of disease presentations (Objective 2)

SBP and DBP are both positively associated with the initial presentation of CVD across all endpoints with the exception of SBP and AAA, where no effect of increasing SBP on onset of AAA was shown. With the exception of AAA, there is little heterogeneity in the size of the effect estimates, which range from an age-adjusted HR for one sd increase in SBP of 1.19 (95% CI 1.16-1.22) for heart failure to 1.28 (1.24-1.31) for peripheral arterial disease. This relatively low effect estimate of the association of increases in BP with heart failure is surprising, given the generally strong association of two to three times the risk of heart failure with hypertension commonly found in the general literature.(307) The

increased hazard associated raised SBP and DBP was evident with even the lowest blood pressure categories. While the increase in hazard is modest for one standard deviation, the differences in hazard between patients with the lowest and the highest categories of blood pressure are substantial: twofold for SBP and triple for DBP for some presentations.

In addition to assessing the effect of blood pressure across such a wide range of presentations, examining the independent effects of SBP and DBP adds to the debate about which measures of blood pressure are the most accurate measure of risk. My findings indicate that for most presentations there is no difference between SBP and DBP, although SBP poses a slightly higher risk.

The size of the effect estimates are consistent with some of previous studies which specifically examined initial presentation of CVD using more chronic CVD endpoints, such as stable angina and heart failure(222,301) but not all.(225) Previous studies which were restricted to acute forms of CVD (MI and stroke) found stronger associations with stroke than found in the current analyses, for both initial presentation(223,299,300) and first presentation within phenotype(286,304). This discrepancy in findings indicates that not including a wider range of endpoints in any analysis of initial presentation may over-estimate the effect of BP on onset of CVD with stroke but also suggests the importance of blood pressure in the development of stroke after onset of CVD with other presentations. When they become available, the findings of studies using the REACH registry, which is investigating subsequent events in patients with atherosclerotic disease, may provide further information on this point.(308)

6.3. Differences in association of SBP and DBP with initial presentations of CVD (Objective 3)

As mentioned above, the effect estimates for SBP are slightly higher than the estimates for DBP for all endpoints. The exceptions are a) PAD where the estimate for DBP was considerably lower than that for SBP and b) AAA where the effect estimate for DBP was significantly higher than that for SBP. Unlike the current analyses, previous studies which have investigated the differences in different blood pressure measures have tended to focus on one or two individual presentations,(300) global CVD risk,(289) or mortality.(283)

The lack of effect of increases in SBP on AAA is potentially inconsistent with previous studies which consistently show an effect of hypertension of AAA, (235,242) but most previous studies have used hypertension as the only blood pressure measure rather than

looking at the individual elements. One possible explanation for the lack of association of SBP with AAA is that patients also have dilation of thoracic aorta, causing lower resistance to the heart pumping and therefore lower SBP.

6.4. Modification by gender and age (Objective 4)

Men and women do not differ in the association of SBP and DBP with initial presentation of CVD endpoints, with the exception of stroke and the ventricular arrhythmias. Men have a greater hazard of SBP and DBP for initial presentation of stroke compared to women; the analysis of categorical measures of SBP and DBP indicates this difference is principally due to an effect at the highest SBP / DBP categories. Women had a markedly higher hazard from increases in SBP/ DBP than men for ventricular arrhythmia, cardiac arrest and sudden cardiac death.

No studies with multiple endpoints of initial presentation of CVD or first presentation within phenotype produced gender-specific effect estimates for SBP or DBP for comparison. The existing literature on risks associated with increases in blood pressure which have looked at gender-specific effects have not found consistent associations. Sauvaget found a higher risk from SBP of mortality from stroke in men compared to women in a South Asian population,(285) but Lida found a higher risk from SBP in women in a Japanese population.(303) Hypertension has been found to be a major risk factor for sudden cardiac death in a woman-only study(309) but has been found to pose a lower risk to women than men in an earlier Framingham study.(310) Results from a study on new onset hypertension in participants in the Framingham cohort found a higher proportion of men presenting with stable angina, PAD (intermittent claudication) and AMI, while a higher proportion of women presented with stroke when compared to controls who experienced non-CVD death.(311)

A clear reduction in the hazard posed by increases in BP with age was found across all endpoints, with the strongest modification of the effect of age in unheralded coronary death and heart failure. Indeed, increases in blood pressure doubled the association with heart failure in patients below the age of 50 compared to the oldest age group. This age effect may reflect different aetiological causes of heart failure at different ages, with poor left ventricular function, i.e. subclinical heart failure, causing low SBP in the older age groups, while the presentations in the youngest groups caused exclusively by hypertension.

These findings are consistent with mortality studies of the effect of increases in blood pressure which show a decline in effect with age for acute myocardial infarction, stroke and vascular disease.(283,284) However, the size of effect estimates in the current analyses are somewhat smaller, which is consistent with increased blood pressure not only being associated with onset but also conferring additional risk for progression to subsequent CVD events or CVD mortality. Unlike the current analyses, Franklin et al. found DBP to be more strongly associated with CVD in patients under 50 while SBP and pulse pressure assumed greater importance in older people.(294)

6.5. Strengths

The general strengths of data from the CALIBER research platform have been rehearsed elsewhere in this thesis, but include the large cohort size, the ability to identify with some degree of certainty the initial presentation of CVD, and the detailed risk factor and prescribing data available. With specific reference to this set of analyses, one key strength is the availability of continuous measures of both systolic and diastolic blood pressure, allowing comparison of the effects of each measure.

6.6. Limitations

A key limitation of cohort studies using EHR is missing data and the biases introduced by losing patients without complete data. A detailed discussion of this issue can be found in Chapters 4, but of concern for this set of analysis is that younger patients were more likely to be excluded from the cohort because of missing blood pressure data. My finding of stronger association of raised blood pressure at younger ages may be due to recording bias, with GPs recording blood pressure in younger people about whom they had concerns for other reasons such as obesity. A further limitation of EHRs is the quality of recording of blood pressure. Figure 28 above showed that GPs have a clear preference for recording data in multiples of 10 (and to a lesser extent 5). However, additional modelling of the association of blood pressure using categorical variables for SBP and DBP found the same effects as models using blood pressure as a continuous variable. There is concern within the epidemiological literature about the use of standardised regression coefficients, such as used in the present analyses, because such standardisation has the potential to confound aspects of study design which drive the variance of the standardised measure with effect of interest, as well as militating against valid comparisons between studies.(306) However, the first concern does not apply in the current situation where the same standard deviation is used across the whole study, and the same covariates are used in models for SBP and DBP. (312) The value of being able to make a direct comparison between SBP and DBP in the current study was thought to outweigh the potential for lack of comparability with future studies.(313)

The current analyses did not use repeated measures of blood pressure, so there was the potential for regression dilution.(314) However, with a mean follow-up time of approximately 5 years, the extent of the dilution would be considerably less than in some of the longer term cohort studies. Only linear effects for both SBP and DBP were modelled; a possible j-shaped association, particularly with DBP, which has been found in the previous literature,(315) was not assessed. Although the relationship looks broadly linear in the categorical analysis, the lowest DBP category is higher than in some other studies. The lower effect estimate for DBP across most endpoints could be due to increased risk at the lowest levels of DBP. The proportional hazard assumptions appeared to be violated with the specific presentation, AAA, which appears to differ from the other presentations in lack of effect of SBP. It is possible that alternative approaches to modelling the association with SBP and AAA would better capture the relationship. This may also hold true for the ventricular arrhythmias, the only presentation where there were clear gender differences in effect.

Although data on blood pressure-lowering medication was available in this dataset, it was not included in the model estimating the effect of increases in blood pressure on initial presentation of CVD because medication would be in the causal pathway between exposure and outcome. However, patients on medication whose blood pressure is not controlled may have additional risk not simply related to their blood pressure level. This complexity would not be adequately captured in the model used. Also, there is good evidence that calcium-channel blockers have a greater preventative effect on stroke than other blood-pressure lowering medications, so that medication has an effect on likelihood of different initial presentations.(316)

7. Summary of Findings

- Little is known from previous studies about heterogeneity in the association of SBP and DBP with the initial presentation of CVD across a range of presentations.
- One standard deviation increase in SBP and DBP was associated with increased hazard of initial presentation across all presentations, with the exception of AAA (SBP only) and PAD (DBP only).
- Some heterogeneity in the strength of association of SBP and DBP with a range of CVD endpoints was found, but was limited compared to heterogeneity of effects found with smoking.

- The increased hazard associated with higher SBP and DBP were evident even with the lowest blood pressure categories.
- The association of increases in BP with initial presentation is modified by gender only for stroke, with increased hazard in men, and for ventricular arrhythmias, with increased hazard in women.
- The association of SBP and DBP with most endpoints decreases with age up to 70-79 years. This effect modification is particularly notable for heart failure where one standard deviation increase in SBP had double the hazard for patients under 50 as for those over 80.

Overall Conclusions and Implications for Public Health, Clinical Practice and Research

1. Introduction

In this thesis, I have described briefly the field of electronic health record (EHR) research and the specific research platform, CALIBER, used for my research (Chapter 2). I then went on to describe my general methods (Chapter 3) and report on results applicable to all the analyses in this thesis (Chapter 4). In chapters 5 to 7, I discussed the results of my investigation of the association between gender, smoking and blood pressure on the onset of cardiovascular diseases (CVD) across a range of initial presentations. Here, I pull together the threads of the thesis and discuss some of the wider implications of my findings. I begin with a summary of my methodological approach and main findings, discuss the overall strengths and limitations of the cohort and my analyses, followed by my views on the implications of this research for clinical and public health practice and recommendations for research.

2. Overall approach and main findings

The cornerstone of my thesis has been the development, with colleagues, of the CALIBER research platform.(167) CALIBER incorporates linked electronic health records data from four data sources, a common data model and transparency through analytic protocols registered in the public domain. CALIBER was used to implement the current research study but provides the tools for building programmes of epidemiological research such those currently being undertaken by the Clinical Epidemiology Group at University College London (UCL), the London School of Hygiene and others. Using CALIBER, I developed a large cohort of 1,758,584 patients free from symptomatic CVD with which to investigate the association of gender, smoking and blood pressure with the onset of cardiovascular disease with a wide range of initial presentations.

My principal findings were:

- 69% of initial presentations (N=95,267) of CVD were neither acute myocardial infarction (AMI) nor stroke
- Men had higher rates of all presentations, with age-adjusted gender rate ratios from 1.23 (95% confidence interval 1.19-1.27)) for stroke to 4.27 (3.92-4.65) for AAA.
- The association of current smoking (compared to non-smoking) varied from an age-adjusted hazard ratio (HR) of 1.01 (0.90-1.13) for arrhythmia to 4.71 (4.15-5.35) for AAA.
- Gender differences were found only for MI (women: 2.59 (2.37-2.83); men: 2.20 (2.05-2.37)) and AAA (women: 7.00 (5.64-8.70); men: 3.88 (3.33-4.53)).

- The association of systolic blood pressure was similar across all CVD presentations, excepting AAA, ranging from age-adjusted HR of 1.19 (1.16-1.22) per standard deviation (SD) for heart failure to 1.28 (1.24-1.31) for PAD, with minimal gender differences.

3. Overall strengths

This thesis provides a demonstration of the value of linked electronic health records for programmes of cohort research. By linking the longitudinal record from primary care data to the episodic records provided by secondary care and clinical registry data with final outcome data from mortality data, a detailed picture of the patient journey over time can be seen. The current work has focussed on initial presentation of cardiovascular disease, using the primary record to exclude patients with prior history, provide detailed baseline risk factor information and, along with the secondary care, clinical registry and mortality data, to provide endpoint data.

The results from the thesis combine replicating results from previous studies, which provides reassurance about the accuracy of the findings overall, with novel results which extend knowledge. The association I found in the linked EHR between smoking and lung cancer is consistent with results of a meta-analysis of the studies based on investigator led bespoke studies used in the International Agency for Research on Cancer (IARC) review of smoking and lung cancer,(317) but I also showed the relationship between smoking and onset of CVD across a wide range of CVD endpoints for the first time. The association I found between being an ex-smoker and multiple CVD endpoints is consistent with findings of the Million Women study,(269) but shows the relationship with the overall onset of CVD rather than with the first occurrence of CVD endpoints, regardless of previous disease. The pattern of the relationship between raised blood pressure and cardiovascular endpoints, as modified by age, replicates the findings of the Prospective Studies Collaboration(283) but extends both the range of endpoints investigated and focusses on onset of CVD rather than CVD mortality.

Other novel insights provided by the approach of focussing on initial presentation of CVD across a broad range of specific presentations include the frequency with which CVD first presents with diseases *other* than MI and stroke – almost 70% of endpoints are other presentations. Most previous investigator-led studies have relied on so-called hard endpoints such as MI and stroke. The presentations of stable angina and peripheral arterial disease which I found to be more common have better initial prognosis and more chronic presentations than either MI or stroke, allowing greater opportunities for intervention. I also

identified the heterogeneity of association of gender and smoking with multiple CVD endpoints in contrast to the relative homogeneity of association of blood pressure with these endpoints.

4. Overall limitations

There are four key limitations to the dataset and the research presented in the thesis which are discussed here. Future research, either currently planned or potential, to address these limitations is described in 6.1 Implications for research methods.

4.1. Selection of cohort patients

The patients included in the overall cohort were those who met the age and observation time criteria and had no prior history of cardiovascular disease. However, for the specific analyses on gender, smoking and blood pressure, patients were excluded if risk factor data were missing. Approximately a fifth of patients who met the general inclusion and exclusion criteria were excluded for the analyses in Chapter 6 and half of all possible patients for Chapter 7. (See Chapter 4 for more detailed analysis of missing data). The majority of those excluded were younger men, who are less likely to attend general practice even if registered. This potential selection bias does raise concerns, but which are common to cohort studies using this sort of data, particularly those investigating aetiological questions. One approach to addressing this concern has been to investigate the possibilities for imputation, which are further discussed in 6.1 below. Another approach to dealing with missingness, which I favour, is to investigate the informativeness of missing data, also discussed below.

One question worth asking is whether this selection bias I have identified is in fact worse or simply more explicit than other studies. With electronic health records research, it is easier to identify and characterise those not included in a cohort compared to investigator-led cohorts. Such studies may not be able to describe potentially eligible participants who were excluded because they did not volunteer to participate, which could be seen as the equivalent of registered population excluded because of missing data. Only individual participant meta-analyses are currently able to compete with EHR studies on the size of cohort, but these are a collection of multiple different individual studies, all with different approaches to patient recruitment and retention. For example, in the Prospective Studies Collaboration study on the association of blood pressure with cardiovascular mortality,(42) it is not clear how many potential participants have been excluded due to missing data because the overall figure across the 61 studies has not been reported. Further discussion on methods of comparison

between EHR studies, investigator-led cohort studies and individual participant meta-analyses would be helpful so that biases introduced by each could be fairly adjudicated.

4.2. Determination of endpoints

In conventional bespoke cohorts, the determination of endpoints is commonly accomplished by an adjudication panel, often using multiple sources of information to arrive at a conclusion about each specific endpoint. There is a large literature describing these processes, if not evaluating them, for a wide variety of studies and endpoints. For electronic health record studies the determination of endpoints is conventionally an assessment of the sensitivity, specificity or positive and negative predictive value (PPV) of specific codes or coding algorithms, determined by comparing codes against a gold standard such as medical records. Those validation studies which relate to the endpoints in this thesis have been described in Chapter 2. It is not clear whether a similar approach is even theoretically possible with linked datasets, such as CALIBER, because there is no obvious gold standard. One approach to addressing this issue has been to assess the extent of overlap between multiple EHR data sources, the most recent of which examined the overlap in myocardial infarction cases between the constituent datasets within CALIBER.(101) This study found that each individual data source missed a substantial number of cases (between 25-50%), substantiating the importance of using multiple sources of information, as has been done in this thesis.

In order to provide sound determination of endpoints, the CALIBER research collaboration implemented a rigorous, documented and auditable approach to:

a) Defining reusable algorithms to identify relevant codes – agreed by at least two clinicians, including a practising GP, and recorded as STATA do-files which can be rerun and updated
b) Defining base variables – specifying the algorithm used to create data variables from a single data source and identifying multiple response categories which can be combined or excluded depending on the purpose of individual researchers
c) Defining risk factors and endpoints – combining base variables from multiple data sources in specified and documented ways to arrive at endpoint and risk factor definitions
d) Recording metadata – all variables include meta-data which includes a plain English definition, specification of minimum and maximum values if appropriate, response categories, data sources, and version control information.(167)

The determination of some endpoints in CALIBER, specifically the acute coronary ones, is likely to be more accurate than others because of the linkage with the acute coronary syndrome registry, MINAP. At present, there are no other equivalent registries for the other endpoints although there are some promising developments. The only national clinically collected data source for stroke is a bi-annual sentinel survey of two months of stroke admissions.(318) There is a relatively new registry for heart failure, curated by the National Institute for Cardiovascular Outcomes Research, who curate MINAP, but the proportion of cases in this four-year-old audit remains relatively low with 58% of heart failure admissions submitted to the audit.(319) The National Cardiac Arrest Registry (NCAR) which aims to capture all in-hospital cardiac arrests involving a resuscitation team,(320) has not yet begun to collect data. From the evidence on comparison of data sources for acute myocardial infarction, (101) addition of any of these data sources could increase the number of endpoints over inclusion of HES and could improve the quality of clinical information about stroke, heart failure or cardiac arrest. The possibility of linkage with these sources of data should be considered for future development of CALIBER.

4.3. Missing data for adjustment

The section above highlights the issue of patients excluded from the cohort because of missing information. Data on other key risk factors such as lipid levels, ethnic group, alcohol consumption and body mass index were also missing to such an extent (Chapter 4) that they were not included in the adjusted models. Future work on electronic health record data needs to address this issue of missingness. Possible approaches to addressing this methodological issue are described in *6.1 Implications for research methods* below.

4.4. Measurement of risk factors

The measurement of the risk factors used in this research were completed in real-world settings and are likely to have greater errors of measurement than those collected under research conditions, particularly the blood pressure measurements. I have already shown GP's preference for recording blood pressure in multiples of 5 or 10, providing evidence of systematic under or over-estimation of actual blood pressure (Chapter 7). The possibility of measurement by indication is likely; patients with greater risk factor profile or higher baseline blood pressure maybe more likely to have their blood pressure recorded in subsequent periods. Additionally, there is the possibility of changes in risk factors over time, with smokers quitting or ex-smokers resuming and blood pressure increasing with age, raising the possibility of regression dilution.

These short-comings in the quality of the data could explain, for instance, differences in the size of association of blood pressure found in the current analyses compared with those found in the Prospective Studies Collaboration (PSC) study on blood pressure.(283) I did not adjust for regression dilution, while the PSC did. However, the maximum period of follow-up for the current cohort is 10 years, with a median of 5, which Clarke, Shipley, Lewington et al. estimate would under-estimate the risk by approximately one third in the first decade,(314) bringing my estimates more in line with those of the PSC. The studies which contributed data to the PSC measured blood pressure under research conditions, again another explanation of possible different effects. However, the obvious small errors in measurements in the current data due to digit preference did not appear to have a substantial impact of the findings from the current analyses, with similar results from both continuous and categorical measurements of blood pressure (Chapter 7). Differences in findings with Prospective Studies Collaboration could also be due to the endpoints investigated – onset of CVD versus CVD mortality.

4.5. Use of relative measure of risk

Throughout this thesis, relative measures of risk have been used to investigate the hetereogeneity of association between the risk factors and the endpoints:- rate ratios for the association of gender and hazard ratios for the association of smoking and blood pressure on the initial presentations of CVD. The appropriateness of relative measures of risk has been extensively debated within the epidemiological literature, particularly in the assessment of interactions or effect modification.(321) Weed and Trock(322) have suggested that content of these debates can be divided into three sorts of questions:

1. What is the definition of effect modification? These ontological questions have been addressed for example in Rothman, Greenland and Lash,(323) Siemiatycki and Thomas,(324) and Weinburg.(325)
2. How should we measure or infer the existence of effect modification? These epistemiologial or methodological questions have been addressed for example in Walter and Holdford,(326) and Pearce(327); and
3. What action should we take as a consequence of such identification? These ethical or public health/clinical questions have been addressed for example in Weed and Trock(322) and Spiegelhalter.(328)

Rothman, Greenland and Walker proposed a similar taxonomy of questions, dividing implications for action into those at population level and individual level. (329)

Conversion of relative measures of risk derived from time-to-event analyses into measures of risk difference in order to assess effect modification or determine appropriate individual or population-level action are not as straight-forward as with other regression analyses, although approaches have been suggested.(330,331) Time-to-event analyses with competing risks provide an additional complication in seeking to assess differences in risk but differences in cumulative incidence which take account of cause-specific hazard ratios have been proposed.(332) Such a measure has the virtue providing a measure of the absolute risk difference between groups which takes into account both frequency of events in such groups and relative hazards.(183) In order to assess the impact of potential risk reduction measures, for populations, such absolute measures of risk are needed in order to estimate the effect on incidence of specific presentations in the presence of competing risks.(333) For individuals to determine their own risk reduction with changes in behaviour or treatment, the same absolute measures are required. The current analyses are therefore not sufficient to determine individual or public health action but have the potential to indicate areas where, with further modelling, fruitful approaches may be determined. For example, in a development of the approach used in this thesis, differences in absolute (lifetime) risks for different CVD disease presentations between men and women identified similar male excess for myocardial infarction and abdominal aortic aneurysm as indicated by the current relative measures of risk.(334) Further work to develop calculations of absolute measures of risk difference to assess the impact of smoking and blood pressure are currently underway.

5. Implications of findings for public health & clinical practice

5.1. Development and use of risk scores

The choice of endpoints included in the risk scores in common use in the UK or recommended by National Institute for Health and Care Excellence (NICE) guidance (Table 27) varies widely.(335) Few risk scores include peripheral arterial disease or heart failure in the endpoints assessed.(336) This situation prevails despite NICE guidance recommending inclusion of all symptomatic atherosclerotic vascular disease, including revascularisation and PAD, as endpoints in the risk equations used for formal risk assessment. My research demonstrates the extent to which the range of endpoints selected for inclusion in the calculation of risk scores matters. I have shown that the onset of symptomatic CVD is with presentations other than MI and stroke in approximately 70% of cases. Many of these other presentations are more chronic and potentially treatable diseases than either of these so-called hard endpoints (Chapter 4). From the perspective of specific risk factors, smoking is

used in all widely used risk prediction models for CVD onset. However, I found that the most common initial presentation, stable angina, is only weakly associated with smoking (Chapter 6). Clinicians may be underestimating the risk of chronic presentations of CVD and therefore missing opportunities for targeting primary prevention therapy for patients at risk of developing CVD.

The latest Framingham risk score(41) does include the widest range of CVD endpoints, but given findings on the importance of ethnicity(337) and social deprivation(338) in the UK context, there is an urgent need to create – and use - a UK risk score for onset of CVD which includes both a broader range of endpoints and risk factors than any of the risk equations currently in use. The importance of this recommendation is reinforced by evidence about the seriousness of onset of CVD with any presentation. For example, the rate of progression from less acute presentation of CVD with PAD to more acute presentations of CVD or death is high: 20-40% of patients with PAD have subsequent acute MI or stroke and 50% die within 6 years.(339,340)

The finding of heterogeneity in the gender rate-ratios suggests that disease-specific gender coefficients, as well as gender-specific baseline survival, would lead to a risk prediction model with better discrimination than one which does not distinguish between a range of endpoints. However, the interaction between smoking and gender in the risk of onset with acute MI and AAA also suggests that gender-specific risk scores may better capture the risk of onset, if indeed there is modification of the effect of gender by smoking. Simply including a coefficient term for gender, even ones which are disease-specific may not adequately capture differences in risk for men or women across the range of endpoints. Further work is indicated to assess the discrimination of risk prediction models which use disease-specific sex coefficients or separate sex-specific risk prediction models. The latest Framingham risk score for use in primary care, which uses the sex-specific approach, suggests adding a calibration factor to the overall CVD risk score to predict disease-specific risks, (41) making it an ideal benchmark against which to measure the performance of the approaches mooted above.
The range in the strength of association between the risk factors studied in this thesis and the specific diseases with which CVD initially presents also raises the question of whether primary prevention treatment might be tailored to the individual elements which contribute to a risk score, where that specific risk element has a greater association with particular presentations. For example, a systematic review of blood pressure lowering medication found that although different regimens reduced overall cardiovascular risk, their effect on

cause-specific risk varied.(341) The endpoints studied in this review did not include PAD, but on the basis of my findings in Chapter 7, a blood-pressure-lowering regimen which reduced the risk of PAD the most, for example, might be preferred to one which acted more effectively on prevention of other endpoints.

Risk scores could also potentially be personalised to take account of individual priorities. Patients may have specific CVD diseases which they are more motivated to avoid; for example, my father, a life-long academic, is most fearful of suffering a stroke and losing his intellectual capacities than suffering other CVD presentations. Other patients may have seen family members die from specific diseases and wish to avoid their fate. A range of risk prediction scores which cover specific endpoints could be used to tailor specific prevention strategies to patient priorities, which in turn could enhance compliance.

Table 27: Cardiovascular Risk Scores commonly used in England and Wales

Endpoints included	FRS - Original Total CVD (1991) (342)	FRS - CHD Total CHD (1998) (259)	FRS - ATP-III Hard CHD (2001) (343)	FRS - Updated Total CVD (2008) (41)	QRisk – 1 CVD (2007) (169)	QRisk – 2 Total CVD (2008) (344)	Score Fatal CVD (2003) (345)	Assign CVD (2007) (338)
Endpoints included								
Ischaemic stroke	X			X	X	X	X	X
Haemorrhagic stroke	X			X	X	X	X	X
Transient ischaemic attack	X			X	X	X	X	X
Angina Pectoris	X	X		X		X	X	X
Unstable angina	X	X		X		X	X	X
Myocardial infarction	X	X	X	X	X	X	X	X
Sudden CHD death	X	X	X	X	X	X	X	X
Revascularisation procedure								X
Congestive heart failure	X			X		X	X	
Peripheral arterial disease	X			X		X	X	
Abdominal aortic aneurysm							X	
Proportion of women in cohort	53.5%	53.4%	53.4% (assumed)	53.3%	50.4%	50.4%	42.9%	49.2%

Adapted from information in Systematic review of cardiovascular disease risk assessment tools(346)

5.2. Gender and cardiovascular disease

The association of gender with onset of a wide range of cardiovascular presentations clearly varies. As has been previously documented, acute MI and coronary death are more common in men at a younger age than in women (Chapter 5). However, similar effect modification with age did not hold for other disease presentations. These findings point to the importance both of more finely specified endpoints in understanding cardiovascular risk, but also the importance of adequately powering studies so that effects in men and women can be studied separately. Furthermore, it raises concerns about the generalisation of earlier CHD risk in men being extended to all CVD, as can be found, for example on popular website such as the British Heart Foundation (http://www.bhf.org.uk/heart-health/conditions/cardiovascular-disease.aspx). If such conflation of risk is a view also held by primary or secondary care doctors, it could lead to the under-estimation of risk in women for presentations other than coronary heart disease at younger ages.

Although the gains for public health made in reducing the rate of onset of both AMI and unheralded coronary death (UCD) in men, the apparent lack of similar progress in women is of grave concern, particularly given that coronary heart disease remains the most common cause of death in women and therefore reducing onset must remain a priority. Although the reasons for the lack of change in the rate ratio over the 10 years of this study are still unclear, clinicians should ensure that women with similar risk factor profile to men are offered the same primary prevention interventions.

5.3. Smoking

5.3.1. Better smoking data should be recorded

The lack of detail recorded in general practice on patients who smoke, i.e. number of cigarettes per day, pack years or age of taking up smoking, is of grave concern. Being a current smoker has a strong association with onset of CVD across a wide range of presentations, but little detail is recorded which would enable a better understanding and management of this risk. GPs would not consider using a blood pressure machine which simply records hypertension or not, so one must ask why are they are prepared to record a binary measure for smoking, when smoking is a risk factor which poses such an acute risk for some CVD presentations and is at least as serious a risk factor as blood pressure for overall risk? Under the Quality and Outcomes Framework, which provides financial incentives for general practitioners to record risk factor information about their patients, more detailed information than simply smoking status could be required. For example,

GPs are required to measure HbA1c for diabetic patients and cholesterol levels for patients with CHD. Recording the number of cigarettes smoked per day or pack years for smokers would be both feasible and appropriate.

5.3.2. Importance of quitting

My findings on the variation in the strength of association between smoking status and different endpoints support both an acute and a chronic effect of smoking on the onset of CVD. The large heterogeneity between different endpoints for current smoking and the much more limited heterogeneity and weaker associations for ex-smokers demonstrates the large gain to be made from quitting but also that some damage caused by having smoked cannot be undone. These conclusions are similar to those reached in the Million Women study based on a narrow set of endpoints.(269)

The large risk of smoking with acute MI and unheralded coronary death, which is physiologically consistent with smoking playing a role in platelet aggregation and thrombus formation, poses the question of whether primary prevention should be tailored to reduce thrombus formation in smokers. Clearly quitting smoking would have the largest impact on risk of such an event, but a harm-reduction strategy, such as the prescription of antiplatelet drugs for smokers who cannot or will not quit should be explored. Indeed, the UK National Institute for Health and Care Excellence has recently published guidance on harm reduction with tobacco use.(347)

5.3.3. Risks of smoking in women

I found smoking was much more strongly associated with onset of CVD presenting with AAA in women than in men. Currently in the UK, women are not screened for AAA for cost-effectiveness reasons, because AAA is a rare event in women.(243) My findings raise the question of whether a two-stage screening process for AAA should be explored, starting with assessment of current and past smoking behaviour and then proceeding with abdominal ultrasound screening in smokers or ex-smokers, but not in non-smokers. If such an approach were shown to be effective, the case finding in the second stage of screening, and therefore the cost-effectiveness of the AAA screening programme as a whole, could be improved. The conclusion of an individual-participant meta-analysis of screen-detected AAA suggested the importance of earlier intervention in women and smokers,(236) both conclusions of which would be supported by my current findings.

Evidence indicates increasing prevalence of AAA in women(242) with these increases potentially accelerating as the effect of the wide uptake of smoking in women in the 1950s, 60s, and 70s plays out. Coupled with my findings on the greater relative risk of

smoking in women than in men, these changes in epidemiology should be considered in the next review of the National Screening Committee's (NSC) policy on AAA screening. Given the current Cochrane review of AAA screening which informs the NSC policy was last updated in 2007 and includes only one study which included women,(243) there is also a need to undertake further research on investigate the potentially changing risk of AAA in women.

5.4. Blood pressure

5.4.1. Increased risk of raised blood pressure at all levels

The most important finding from my analyses of the association of blood pressure with the onset of cardiovascular disease is that the risk of onset of CVD across virtually all presentations is increased even at the lowest levels of systolic and diastolic blood pressure. This finding supports the view that blood pressure should be addressed at a population level. Individual treatment with medication across such a large population is neither feasible nor desirable. Although NICE has clinical standards on hypertension, reducing blood pressure at a population level does not feature in the list of topics for which NICE are developing public health standards. Interventions to reduce blood pressure tend to be scattered throughout guidance on other topics such as increasing physical activity and reducing alcohol consumption. Action on reducing salt intake does not yet have the consensus and impetus behind it that tobacco control policies do, but has the potential to have significant effects on blood pressure levels at population level. The latest Health Survey for England identified that mean SBP and DBP fell "marginally but significantly" between 2003 and 2011, which the survey suggests equates to a reduction in ischaemic heart disease deaths of 9% in men and 12% in women and a greater reduction in stroke deaths (men:13%, women 18%).(37)

5.4.2. AAA and blood pressure

I found that SBP was not associated with initial presentation of CVD with AAA, while DBP was. Some studies have previously identified raised DBP as a risk factor for AAA.(348,349)but much of the literature focuses on hypertension, rather than SBP or DBP as individual risk factors. One systematic review and meta-analysis of screen-detected AAA found a modest effect of hypertension,(350) while other studies have found no effect(351) or only in women(352). None of the previous studies have the scale of the current study, although previous studies have the benefit of being able to distinguish between screen-detected AAA and AAA events, unlike the current study. Greater clinical attention should be paid to raised DBP, particularly in current or ex-smokers, who may be at greater risk of AAA.

5.4.3. Blood pressure in people under 50 years

The strongest association of increases in SBP and DBP were seen with heart failure and unheralded coronary death in people under 50 (Chapter 7), although it must be emphasized that this is a relative measure. These findings support the recent revisions to National Institute of Clinical Excellence guidance on hypertension which recommended further investigation of hypertension in people aged under 40 whose 10 year risk might be under-estimated by standard risk scores.(353) Further research would be required to determine whether the most appropriate public health target group for blood pressure reduction would be the few younger people at greater relative risk or the more numerous older people at lower relative risk. Measures of absolute risk would, no doubt, favour interventions with older people given the frequency of presentations in this group, but measurement of the years of life lost or reduction in disease-free years should also be considered, given the estimated association of higher blood pressure with presentations which have such poor prognosis. Broad public health measures to reduce blood pressure, such as reducing the population average intake of salt(354), have the potential to address a broad range of age groups

6. Implications for research methods and areas for research

6.1. Implications for research methods

The success of this thesis in identifying novel findings on the initial presentation of CVD raises the possibility of linked electronic health records replacing investigator-led registries. To take an example from prognosis research, the REACH registry is the largest current international cohort seeking to understand the nature and frequency of subsequent CVD events in patients with stable atherosclerotic disease. (23) It aims to recruit a total of 60,000 patients, and is in the early stages of endpoint follow-up. My cohort from the current analyses has data on over 80,000 patients with CVD endpoints which they survived for at least one day with up to 10 years of follow-up time. This simple comparison shows the potential for research similar to that of the REACH registry using EHR records at a fraction of the cost and with the potential for quicker publication of results.

Missingness continues to be an issue with EH-based research. As discussed in Chapter 3, researchers are investigating approaches to imputation with longitudinal primary care data similar to GPRD, which may lead to a better understanding of the suitability of imputation in these datasets, where missingness is likely not to be completely at random.(180) The breadth of additional variables available to use for imputation and the

large number of patients may make these approaches feasible, although imputation with such large datasets is computationally intensive. Repeating the analyses in this thesis with a multiple imputation dataset is planned and the results of the comparison with the complete case analysis presented here will be informative.

Another approach to handling missingness, particularly where data is not missing at random, is to treat missing data as informative. For example, in a preliminary analysis not presented in this thesis, I found that having no smoking status recorded was as strongly associated with unheralded coronary death as was being a current smoker, at least in men. Further work should be undertaken to investigate whether risk scores or stratification based on missing information, although not intuitive for clinicians, could be valid approach to dealing with missing data in electronic health records.

To date, CALIBER has not included more complex algorithms to verify endpoints, as has been done in other EHR research. Some have sought confirmation of diagnoses, such as myocardial infarction, by requiring evidence of hospital admission, raised enzymes or subsequent appropriate treatment.(150,355,356) However, evidence from Herrett et al. indicates that such complex approaches may not be necessary, with the PPV of 92.2% (95% confidence interval 91.6-92.8%) for a record of acute AMI in GPRD and 91.5% (90.8-92.1%) for a record in HES.(101) Planned further research by UK Biobank will identify the PPV of different algorithms and combinations of linked electronic health record sources against manual review of the full, locally held, case notes for stroke, diabetes and coronary disease as part of the endpoint ascertainment, validation and phenotyping project. It is possible that the results of the UK Biobank and other related initiatives may lead to refinement in the algorithms used to define variables in CALIBER. Research to create codes from free text fields, which often include confirmatory information,(357,358) as well as a pilot project to make data from ECG readings machine interpretable for epidemiological research, will, in time, improve the specificity of endpoints from EHR data.

Heterogeneity in smoking effects has been observed in observational and clinical trials with composite endpoints. My study suggests that the specific composition of those composite endpoints may itself to be the source of the heterogeneity, rather than necessarily differences in actual effect of the exposure, indicating the need for greater specificity of CVD endpoints in studies of smoking. As efforts to understand the interplay between genes and environmental factors such as behaviour increase, my findings underscore the importance of specifying the phenotypes under study. This was

demonstrated by recent study identifying variant related to smoking and well-defined endpoint of PAD. (265)

6.2. Areas for future research

With such a wide-reaching thesis, there are a number of different areas where my findings suggests areas for possible future research. Here, I first indicate research questions around gender, smoking and blood pressure and then more far-reaching ideas.

Given the heterogeneity of association between gender and the initial presentations of cardiovascular disease, it would be useful to evaluate patients' and doctors' views of CVD risk, as affected by age and gender, and the ways in which those views affect care-seeking or provision of care. Given the indications I found of changes over time in the rate ratios for a number of different presentations, mainly driven by reductions in men, further research to identify trends in gender differences in primary prevention linked to changes in onset of disease is both feasible and desirable. This seems particularly important given that I also found that women were more likely to present with fatal presentations than men.

A better understanding of the age-period-cohort effect of smoking behaviour in men and women as related to cardiovascular disease prevalence would add to the understanding of changes in gender differences in risk over time. This could be particularly useful in understanding the apparent gender modification of the risk of smoking for AAA, as well as trends in AAA identified elsewhere. Where there appears to be a decline in men in the incidence of AAA(251) which may be linked to decline in men's smoking in previous decades, the pattern in women is less clear. An increase in the incidence in AAA in women might be expected, because the smoking prevalence in women did not peak until after the 1960s, two decades after the peak in men.(359) Further research to model the effect of smoking on expected AAA incidence rates to inform decisions about screening for AAA in women would be important.

With a median of eight blood pressure measurements in men and 10 in women, this EHR cohort could be used to study regression dilution in a real world cohort, to compare to the findings using the Whitehall II cohort, used to identify regression dilution.(314) In addition to the academic interest such a comparison would hold, information from EHR cohort could be useful in assessing risk in patients where recent measurements may be lacking. Planned future research of the Clinical Epidemiology Group will look at the changing profile of risk factors over time and its relationship to prognosis, initially

focussing on blood pressure. The present findings do suggest further work to disaggregate the elements of raised blood pressure which are associated with initial presentation with AAA would be worthwhile, particularly given the conflicting evidence from previous studies. Further research could help to determine the causes of raised blood pressure in the young, whether this be obesity, salt intake or lack of exercise, which would indicate the most appropriate interventions to contribute to primary prevention of CVD. Discrepancies between the findings of the current analyses on the strength of association between raised blood pressure and initial presentation of CVD and published studies on first presentation show the importance of further research which follows the course of the disease from initial presentation to subsequent presentations to mortality. By segmenting each stage, important insights can be gained which can be used to improve health outcomes.

There are a number of different directions in which the current research could be extended. The present analyses have been limited to three factors associated with onset of cardiovascular disease: gender, smoking and blood pressure; future analyses planned for the Clinical Epidemiology Group include extending this research to investigate the association of diabetes, lipids, obesity and alcohol with a wide range of CVD presentations. Fatal and non-fatal presentations have been combined in these analyses. Distinguishing between the two would allow comparability with a wider range of published studies. Investigation of the progression after onset of CVD to subsequent events, and the potentially changing role of risk factors in that progression in individuals, poses an interesting and worthwhile challenge. Trends in mortality of CVD are changing rapidly so modelling the impact of changes in the prevalence of risk factors over time, particularly the age period cohort effects of smoking, would add to a more nuanced understanding of the nature of cardiovascular disease.

6.3. Conclusions

In summary, I built with my colleagues in CALIBER some of the tools necessary to use linked EHRs for this specific project, but also for building programmes of research. In my substantive analyses, I demonstrated heterogeneous associations between gender, smoking and, to a lesser extent, blood pressure across a wide range of initial CVD presentations. Such heterogeneity indicates the importance of investigation of the association of risk factors with more granular endpoints in understanding the aetiology of cardiovascular disease, as well as the importance of studying onset of the disease, rather than mortality. My findings have important implications for risk assessment, screening,

quality of care and public health interventions to reduce risk in several disease areas, as well as indicating promising areas for future research.

7. Summary of recommendations for public health and clinical practice

- There is an urgent need to create – and use - a UK risk score for onset of CVD which includes both a broader range of endpoints and risk factors than any of the risk equations currently in use.
- The possibility of personalising primary prevention treatment, to take account of the weight of risk factors for specific endpoint or patient preference, should be considered.
- Doctors should recognise that the substantially increased risk in men compared to women at younger ages is restricted to myocardial infarction and coronary death and not discount the possibility of other CVD presentations in younger women.
- Where women present with increased risks, those risk factors should be treated as aggressively as they would be in men.
- GPs should record more detailed information on smoking behaviour, so that more accurate assessments of risk can be made. Changes to incentives such as QoF should be considered.
- Quitting smoking should be encouraged as much as possible, but where individuals cannot or will not quite, harm reduction treatment to additionally reduce their risk of myocardial infarction should be considered.
- Further research should be undertaken to determine whether including women who are current or ex-smokers in the AAA screening programme is warranted.
-

8. Summary of recommendations for research

- The possibility of linked EHR research to replace investigator-led cohorts should continue to be explored.
- Approaches to address missingness in EHRs should continue to be explored, including treating missing data as informative, for example in risk stratification.
- The accuracy of algorithms using multiple EHR sources to identify multiple endpoints should be investigated.
- The views of patients and doctors on CVD risk, as affected by gender and age, and the ways in which those views affect care-seeking / giving should be investigated.
- Further research to model the effect of age-period-cohort effects of smoking on expected AAA incidence rates to inform decisions about screening for AAA in both women and men should be undertaken.
- The changing profile of risk from changes in blood pressure should be investigated, to improve understanding of prognosis from this risk factor and to provide a comparison to regression dilution identified in investigator lead cohort.
- Further research to distinguish between fatal and non-fatal presentations of cardiovascular disease should be undertaken.
- Additional risk factors, including lipids, BMI, diabetes and alcohol and their association with initial presentations of CVD should be undertaken.

Appendix A: Acronyms

Acronym	Full Text
AAA	abdominal aortic aneurysm
ACE	Angiotensin-converting enzyme
ACEI	Angiotensin-converting enzyme inhibitor
ACS	Acute coronary syndrome
AHA	American Heart Association
AMI	Acute myocardial infarction
ARB	Angiotensin receptor blocker
BMI	Body mass index
BP	Blood pressure
CA	Calcium
CABG	Coronary artery bypass graft
CALIBER	CArdiovascular disease research using LIinked Bespoke studies and Electronic health Records
CCAD	Central Cardiac Audit Database
CHD	Coronary heart disease
CI	Confidence interval
CPRD	Clinical Practice Research Datalink
CT	Computed tomography
CVA	Cerebrovascular accident
CVD	Cardiovascular disease
DBP	Diastolic blood pressure
DC	Death certificate
ECG	Electrocardiogram
EHR	Electronic health record
Gb	Gigabyte
GHZ	Gigahertz
GPRD	General Practice Research Database
HDL	High-density lipoprotein
HES	Hospital Episode Statistics
HF	Heart failure
HR	Hazard ratio
ICD	International of the Classification of Diseases
ICD-10	International of the Classification of Diseases, Version 10
ICD-9	International of the Classification of Diseases, Version 9
ICVD	Ischaemic cerebrovascular disease
IHD	Ischemic heart disease
IMD	Index of multiple deprivation
IQR	Inter-quartile range

Acronym	Full Text
kg	kilograms
L	litre
LDL	Low-density lipoprotein
LSHTM	London School of Hygiene and Tropical Medicine
LVA	Left ventricular arrhythmia
m	metre
MI	Myocardial infarction
MINAP	Myocardial Ischaemia National Audit Project
mm Hg	Millimetre of mercury
mmol	Micromole
MRI	Magnetic resonance imaging
NICE	National Institute for Health and Clinical Excellence
NOS	Not otherwise specified
NRT	Nicotine replacement therapy
nSTEMI	Non-ST-elevation myocardial infarction
ONS	Office for National Statistics
OPCS	Office of Population Censuses and Surveys
PAD	Peripheral arterial disease
PbR	Payment by results
PCI	Percutaneous coronary intervention
PEDW	Patient Episode Database for Wales
PPV	Positive predictive value
QoF	Quality and Outcomes Framework
RAM	Rapid access memory
RR	Rate Ratio
SA	Stable angina
SAIL	Secure anonymised information linkage
SBP	Systolic blood pressure
SCD	Sudden cardiac death
sd	standard deviation
SHIP	Scottish Health Informatics Programme
STEMI	ST-elevation myocardial infarction
THIN	The Health Improvement Network
TIA	Transient ischaemic attack
UA	Unstable angina
UCD	Unheralded coronary death
UCL	University College London
UCOD	Underlying cause of death
UTS	Up to standard

Appendix B: Literature Review of Studies Validating Death Certificates

Death Certificate Studies: Coronary Heart Disease

Study	Data Years	Country	ICD	Rate of Autopsy (where given)	ICD Code on DC	Patient details	Cases reviewed	Certified as IHD	True positives	Measure of Accuracy	Significant gender differences?
Type 1 Studies: Death Certificates compared to autopsy results											
Modelmog, 1992(360)	1986-1987	Germany	9	97%	390-459		1023	511	229	Sensitivity: 83% Specificity: 69% lower sensitivity and specificity for specific ICD codes	Not assessed
Type 2 Studies: Death Certificates compared to conclusions of adjudication panel											
Alperovitch, 2009(361)	1999-2004	France	10	--	I00-I99	Male and female (45%), 65+	625	203	129	Kappa for CVD = 0.46	No
Coady, 2001(362)	1987-1995	USA	9	7% to 25%[8]	410-414, 429.2	Male,35-74, IH deaths	1712	--	36.4%	Sensitivity: 77% PPV: 85%	No
						Male, 35-74, OOH deaths	2331	--	56.0%	Sensitivity: 88% PPV: 64%	No
						Female,35-74, IH deaths	1204	--	30.1%	Sensitivity: 71% PPV: 79%	No
						Female, 35-74, OOH deaths	1069	--	43.9%	Sensitivity: 81% PPV: 52%	No
de Henauw, 1998(363)	1983-1991	Belgium	9	--	410-414	Male and female, 25-69, IH and OOH deaths	15559	1987	1135	Sensitivity: 53% Specificity: 99% PPV: 91%	Not assessed
Folsom, 1987(364)	1979	USA	9	24%	410-414, 427	Male and female (31%), 30-74, OOH deaths	328	237	223	Sensitivity: 90% Specificity: 83% PPV: 94%	Yes, greater sensitivity in men

[8] Proportion of deaths autopsied varied by county.

Study	Data Years	Country	ICD	Rate of Autopsy (where given)	ICD Code on DC	Patient details	Cases reviewed	Certified as IHD	True positives	Measure of Accuracy	Significant gender differences?
Goraya, 2000(119)	1981-1994	USA	9	38%	410-414	Male and female, age range not given, OOH deaths	174	131	126	Sensitivity: 91% Specificity: 86% PPV: 96%	Not assessed
Ives, 2009(365)	1989-2004	USA	9	--	410-411, 413-414, 428, 429 (CHD & HF)	Male and female, 65+	3194	877	612	Sensitivity: 74% Specificity: 89% PPV: 70% kappa=0.61 (0.58-0.64)	Not assessed
Lahti, 2001(366)	1995	Finland	9	34%	390-459	Male and female (7.4%), no age restriction, death certificates requiring further adjudication by panel	3478	926	739	False –ve rate: 23% False +ve rate: 20%	Yes. 4% more cases in women referred to panel than expected
					410-414			216	155	False –ve rate: 37% False +ve rate: 8%	Yes. 4% more cases in women referred to panel than expected
Lloyd-Jones, 1998(367)	1948-1988	USA	9	12%	410-414, 427	Male and female, 45+	2683	942	635	Sensitivity: 84% Specificity: 84% PPV: 67%	Yes, PPV lower in women because fewer deaths

Study	Data Years	Country	ICD	Rate of Autopsy (where given)	ICD Code on DC	Patient details	Cases reviewed	Certified as IHD	True positives	Measure of Accuracy	Significant gender differences?
Pajunen, 2005(118)	1998-2002	Finland	10	--	I20-25, I46, R96, R98	Male, 35-74	--	1281	--	Sensitivity: 92% PPV:92%	No
						Female, 35-74	--	490	--	Sensitivity: 92% PPV:91%	No
						Male, 75+	--	1054	--	Sensitivity: 89% PPV: 93%	Slight difference
						Female, 75+	--	1984	--	Sensitivity: 90% PPV: 89%	Slight difference
Rapola, 1997(368)	1985-1993	Finland	8/9	--	410-414	Male and female, age range not given	191	191	181	PPV: 95%	Not assessed
Type 3: Death Certificates compared to contemporaneous prior hospitalisations											
Goldacre, 1993(114)	1979-1986	England	9	--	410-4, 427-9	Male and female, age range not given, dying within 4 weeks of hospitalisation	55318	10635	8271	86% of people with admission for circulatory disease had circulatory disease certified as underlying cause of death.	Not assessed
Goldacre, 2003(369)	1979-1998	England	9	--	410	Male and female, age range not given	69333	--	--	Ratio of mentioned to underlying is 1.088	No
Goldacre, 2004(116)	1979-1998	England	9	--	410-4	Male and female, 35-74, dying within 1 year of hospitalisation	7964	--	--	90% of deaths occurring within 30 days of admission for MI coded IHD on DC	Not assessed

Study	Data Years	Country	ICD	Rate of Autopsy (where given)	ICD Code on DC	Patient details	Cases reviewed	Certified as IHD	True positives	Measure of Accuracy	Significant gender differences?
Johansson, 2009(370)	1995	Sweden	9	26%	410-414	Male and female, age range not given, death within 1 year of hospitalisation	1094	115	--	87% of patients admitted for IHD had same cause on DC	No
Johansson, 2002(371)	1995	Sweden	9	--	390-459	Male, age range not given, death within 1 year of hospitalisation	35,836	16370	5868	32.3% (3395) had diagnosis incompatible with original underlying cause of death--	Not assessed
						Female, age range not given, death within 1 year of hospitalisation	33,982	15989	5458	37.0% (3897) had diagnosis incompatible with original underlying cause of death.	Not assessed

Death Certificate Studies: Sudden Cardiac Death

Study	Data Years	Country	ICD	Rate of Autopsy (where given)	ICD Code on DC	Patient details	Cases reviewed	Certified as SCD	True positives	Measure of Accuracy	Significant gender differences?
Type 2 Studies: Death Certificates compared to conclusions of adjudication panel											
Chugh, 2004(202)	2002-2003	USA	10	12%	I00-I09, I11, I20-I51, Q20-Q24, R95-R99	Male and female (43%), cases of sudden cardiac death based on WHO definition	1007	325	193	Sensitivity: 59% Specificity: 86% PPV: 19%	Not assessed
Fox, 2005(120)	1948-1999	USA	9	--	410-414, 429.2	Male and female (45%), 28-62 (at study enrolment), OOH death	1797	592	187	Sensitivity: 46% Specificity: 71% PPV: 32%	Yes, twofold overestimation in women
	1948-1999	USA	9	--	410-414, 429.2	Male, 28-62 (at study enrolment), OOH death	960	324	141	Sensitivity: 48% Specificity: 73% PPV: 44%	Yes
	1948-1999	USA	9	--	410-414, 429.2	Female, 28-62 (at study enrolment), OOH death	837	268	46	Sensitivity: 41% Specificity: 69% PPV: 17%	Yes
Goraya, 2000(119)	1981-1994	USA	9	38%	410-414	Male and female, age range not given, OOH deaths	174	131	113	Sensitivity: 89% Specificity: 51% PPV: 77%	Not assessed
Iribarren, 1998(121)	1985-1990	USA	9	19%	410-414	Male and female, 30-74, OOH	2035	835	222	Sensitivity: 87%	Not assessed

Study	Data Years	Country	ICD	Rate of Autopsy (where given)	ICD Code on DC	Patient details	Cases reviewed	Certified as SCD	True positives	Measure of Accuracy	Significant gender differences?
						deaths				Specificity: 66% PPV: 27%	
					427.5			322	60	Sensitivity: 23% Specificity: 85% PPV: 19%	Not assessed
Traven, 1996(372)	1984-1989	USA	9	56%	410-414, 429.2	Male and female, black and white, aged 35-44 years	--	--	--	52% of sudden deaths validated as caused by CHD	Presented but not assessed
WOSCOPS, 1995(168)	1989-1991	Scotland	9	--	all codes	Male, aged 45-64	58	--		56% agreement between cause of death from patient follow-up and linked computerised records	Not assessed

Appendix C: CALIBER Data Portal and Developmental Tools

A key feature of the CALIBER research platform is the algorithms for research-ready data which incorporate data definitions, including meta-data, and the code lists required to produce research-ready datasets. These algorithms are now accessed electronically via the CALIBER web data portal at https://www.caliberresearch.org/portal. The data definitions are open-source, but accessing the code lists requires researchers to register and login. A time-limited account (Login: JG_Thesis, Password: AMxtsoqv4) has been created to provide access to this information. For further information, email Spiros Denaxas at s.denaxas@ucl.ac.uk .

To ensure consistency in approaches to allocating response categories across multiple variables, principles for response category allocation for types of Read codes were agreed. These principles are shown in Table 28 below. The coding programs which provide the auditable algorithms used to generate the coding lists employed by the CALIBER research platform are not currently uploaded routinely on the data portal. An exemplar STATA do-file is shown in Table 28 below. As part of the process for ensuring accurate variable algorithms and code lists, we produced entity code reports, which identified which entity codes were used with specific coding lists. An exemplar entity code report is shown in Table 30. We also conducted reverse entity reports, which cross-referenced coding lists used with specific entity codes (Table 31).

Table 28: Principles for Allocation of Response Categories

Codes about ...	Phrase	Example	Rule	Rationale	Exceptions
Patient treatment / measurement	[condition] [medication] contraindicated	66U6. – HRT contraindicated	Definite diagnosis (if treatment indicates diagnosis)	Indicates GP's intention to treat which in turn indicates patient has relevant condition	
	not [condition] [medication] tolerated	8I71.00 – Warfarin not tolerated	Definite diagnosis (if treatment indicates diagnosis)	Indicates GP's intention to treat which in turn indicates patient has relevant condition	
	adverse reaction to [condition] [medication]	TJC7. – adverse reaction to other antihypertensives	Definite diagnosis (if treatment indicates diagnosis)	Indicates patient has received treatment which in turn indicates that patient has relevant condition	adverse reaction to anti-depressants, as can be used for anxiety – coded as possible.
	[condition] [medication] poisoning	SL234 – insulin poisoning	Exclude, except where the medication is hard to obtain (e.g. streptokinase)	Not known whether drug taken by person for whom it was prescribed, so no conclusions can be drawn	SL44200 – Streptokinase poisoning
	[condition] [medication] not indicated	8I68. – calcium channel blocker not indicated	Decide on case by case basis	depends on condition and treatment so must be decided on case by case basis	
	[condition] treatment stopped	667A.00 – epilepsy treatment stopped	Decide on case by case basis	depends on condition and treatment so must be decided on case by case basis	
	target [measurement]...	246K.00 - target systolic blood pressure	Not used to indicate specific measurement. For diagnosis decide on a case by case basis.	depends on condition and measurement so must be decided on case by case basis	
Patient behaviour relating to healthcare	… [condition] [treatment] refused	8I3X. – diabetic retinopathy screening refused	Definite diagnosis (if treatment indicates diagnosis)	Indicates GP's intention to treat which in turn indicates patient has relevant condition	

Codes about ...	Phrase	Example	Rule	Rationale	Exceptions
Patient education / advice	Education score - [condition]	3881. Education score - diabetes	Possible diagnosis of [condition]	Indicates possibility of diagnosis but if only code in patient record definite diagnosis of [condition] inappropriate	
	Attends … [condition] education	9OLB. Attended diabetes structured education programme	Possible diagnosis of [condition]	Indicates possibility of diagnosis but if only code in patient record definite diagnosis of [condition] inappropriate	
	Referred/ Referral to … [condition] education	8Hj4. Referral to DESMOND diabetes structured education programme	Possible diagnosis of [condition]	Indicates possibility of diagnosis but if only code in patient record definite diagnosis of [condition] inappropriate	
	Did not attend / did not complete … [condition] education	9NiD. Did not attend DESMOND diabetes structured education programme	Possible diagnosis of [condition]	Indicates possibility of diagnosis but if only code in patient record definite diagnosis of [condition] inappropriate	
	Advice … [condition]	67I9. Advice about weight	Decide on case by case basis	Don't know nature of advice and whether predicated in prevention or treatment	
	Leaflet ... [condition]	8CE1.00 – Alcohol leaflet given	Decide on case by case basis	Don't know why given leaflet and whether predicated in prevention or treatment	
	Health education - [condition]	679L. Health education - diabetes	Possible diagnosis of [condition]	Indicates possibility of diagnosis but if the only code in patient record definite diagnosis of [condition] inappropriate	

Codes about ...	Phrase	Example	Rule	Rationale	Exceptions
	[condition] education ... completed	9OLJ. DAFNE diabetes structured education programme completed	Possible diagnosis of [condition]	Indicates possibility of diagnosis but if the only code in patient record definite diagnosis of [condition] inappropriate	
	member of [condition] society	13Y5.00 Epilepsy society member	Possible diagnosis of [condition]	Indicates possibility of diagnosis but if the only code in patient record definite diagnosis of [condition] inappropriate	
Referral or care	Seen in ... [condition] clinic	9N1Q. Seen in diabetic clinic	Possible diagnosis of [condition]	Indicates possibility of diagnosis but if the only code in patient record definite diagnosis of [condition] inappropriate	
	[condition] D.V. done	8Hl4.00 Cardiology D.V. done	Possible diagnosis of [condition]	Indicates possibility of diagnosis but if the only code in patient record definite diagnosis of [condition] inappropriate	
	Seen by [condition] clinician	9N2k. Seen by smoking cessation advisor	Possible diagnosis of [condition]	Indicates possibility of diagnosis but if the only code in patient record definite diagnosis of [condition] inappropriate	
	Referred to [condition] clinician	8H4C. Referred to chest physician	Exclude	do not know reason for referral	referral to diabetic foot clinic – count as possible diabetes
	[condition] D.V. requested	8HKE.00 Diabetology D.V. requested	Exclude	do not know reason for referral.	

Codes about ...	Phrase	Example	Rule	Rationale	Exceptions
	Under care of [condition] clinician	ZL183 Under care of cardiologist	Possible diagnosis of [condition]	Indicates possibility of diagnosis but if the only code in patient record definite diagnosis of [condition] inappropriate	
	Discharged from [condition] care …	8Hg4. Discharged from care of diabetes specialist nurse	Possible diagnosis of [condition]	Indicates possibility of diagnosis but if the only code in patient record definite diagnosis of [condition] inappropriate	
	Waiting / awaited	5531. - Angiocardiography awaited	Exclude	Waiting for test or admission does not indicate [condition] diagnosis	
	DNA ... [condition] clinic		Possible diagnosis of [condition]	Indicates possibility of diagnosis but if the only code in patient record definite diagnosis of [condition] inappropriate	
Chronic disease identification & management	... [condition] monitoring/ monitoring admin	662.. – hypertension monitoring	Possible diagnosis of [condition]	Indicates possibility of diagnosis but if only code in patient record definite diagnosis of [condition] inappropriate	662o.00 – Haemorrhagic stroke monitoring – code as history of ...
	...[condition] review/annual review	66AS.00 – Diabetic annual review OR 662e.00 Stroke/CVA annual review	Definite diagnosis OR history of	Regular review of chronic condition (definite) OR review following acute event (history of)	
	... [condition] register	8HHy. – Referral to diabetic register	Possible diagnosis of [condition]	Indicates possibility of diagnosis but if only code in patient record definite diagnosis of [condition] inappropriate	

Codes about ...	Phrase	Example	Rule	Rationale	Exceptions
	... deleted/removed from [condition] register	9HA1.00 – Removed from depression register	Exclude		
	... [condition] initial assessment	66U1. Menopause initial assessment	Possible diagnosis of [condition]	Indicates possibility of diagnosis but if only code in patient record definite diagnosis of [condition] inappropriate	
	Screening for [condition]	68X.. Screening for cardiovascular system disease	Exclude	Screening does not indicate [condition], unless screening is for complication of [condition]	68A7.00 – Diabetic retinopathy
	Follow-up [condition] assessment	66A2. follow-up diabetic assessment	Possible diagnosis of [condition]	Indicates possibility of diagnosis but if only code in patient record definite diagnosis of [condition] inappropriate	
	... [condition] monitoring not required	hypertension monitoring not required	Exclude	No conclusions can be drawn from these codes	
	... [condition] resolved	hypertension resolved	history of ... diagnosis	Indicates had [condition] in the past	
	... [condition] monitoring deleted	9OI9.00 - hypertens. monitor deleted	history of ... diagnosis	Indicates had [condition] in the past	
	exception reporting: [condition]...	9h3..00 - Exception reporting: hypertension quality indicators	Possible diagnosis of [condition]	Indicates possibility of diagnosis but if only code in patient record definite diagnosis of [condition] inappropriate	
	excepted from [condition] quality indicators ...	9h31.00 Excepted from hypertension qual indicators: Patient unsuit	Possible diagnosis of [condition]	Indicates possibility of diagnosis but if only code in patient record definite diagnosis of [condition] inappropriate	

Codes about ...	Phrase	Example	Rule	Rationale	Exceptions
	refuses ... [condition] monitoring	9Ob1.00 – Refuses coronary heart disease monitoring	Possible diagnosis of [condition]	Indicates possibility of diagnosis but if only code in patient record definite diagnosis of [condition] inappropriate	
	[condition] care plan	8CS0.00 – Diabetes care plan agreed	Possible diagnosis of [condition]	Indicates possibility of diagnosis but if only code in patient record definite diagnosis of [condition] inappropriate	
Neonatal/infant/child	neonatal/infant/child [condition]	Q492.00 – neonatal hypertension	case by case	CALIBER only includes adults, so if code include indicates a history of [condition]	

Date created: 15.3.2011

Agreed: 8.4.2011

Amended: 12.4.2011

Table 29: STATA do-file to create code-list for Stable Angina (GPRD)

```
clear all
macro drop _all
set linesize 100
version 10.0
// Program:      do file to create coding list for diagnosis of stable angina
//                NB excludes ischaemia chest pain which is included in chest pain code list
// Name:         stable_angina.do
// Authors:      Julie George, harry Hemingway, UCL; Liam Smeeth & Emily Herrett,
LSHTM
// Date created: 13.3.2011, amended 24.3.2011, 25.3.2011, 1.4.2011, 11.4.2011

// STEP 1: DEFINE SEARCH TERMS
loc interm " "*angina*" "*stenocard*" "*coronary*artery*spasm*" "*spasm*cor*artery*"  "
loc interm "`interm' "*card*syndrome*x*" "*cor*syndrome*x*"  "

/// STEP 2: WORD SEARCH OF GPRD MEDICAL CODE DICTIONARY
** GPRD Gold Medical Browser, Version 1.3.0
** using Read information 27th November 2009
** Read code chapters relating to occupation (Chapter 0) have been removed
use "C:\Users\JulieG\Dropbox\Coding_lists\_GPRD Pegasus
dictionary\medical_with_counts.dta", clear

*** update the marker where read term matches search terms
foreach word in `interm'  {
        replace marker = marker |strmatch(lc_readterm, "`word'")
        }
*
// STEP 3: CODE SEARCH OF GPRD MEDICAL CODE DICTIONARY
*** search for specific terms not found by word search using read codes
        ** G33z200      Syncope anginosa
        ** G34y000      chronic coronary insufficiency
foreach word in "G33z200" "G34y000"{
                replace marker = marker | strpos(readcode,"`word'")
}
*
// STEP 4: DROP ALL TERMS NOT CAPTURED BY SEARCH TERMS
keep if marker

// STEP 5: EXCLUDE TERMS TO MAKE SEARCH MORE SPECIFIC
*** remove terms that rated by clinician as not relevant
  ** %%%
loc exterm " "fh*" "no fh*" "*score*" "vincent*" "strepto*" "herp*" "*unstable*" "*at rest*""
loc exterm "`exterm'"*worsening*" "*abdominal*" "*ludwig*" "
foreach word in `exterm' {
        drop if strmatch(lc_readterm,"`word'")
        }
*
// STEP 6: EXCLUDE ANY REMAINING UNWANTED CODES
*** remove codes that rated by clinician as not relevant
**388E.00       Canadian Cardiovascular Society classification of angina
**G311.11       Crescendo angina
**G311300       Refractory angina
**ZR37.00       Canadian Cardiovascular Society classification of angina
local excode ""388E.00"  "G311.11" "G311300" "ZR37.00""
```

```
foreach word in `excode' {
	drop if strmatch(readcode, "`word'")
}
*
//STEP 7: ALLOCATE CODES TO SPECIFIC CATEGORIES
gen int sa_cat = 4
lab var sa_cat "Category"
lab def sa_cat 0 "0.Not diagnosed" 1 "1.History of" 2 "2. Vasospastic" 3 "3. Cardiac
syndrome X" ///
4 "4.Stable angina", modify
lab val sa_cat sa_cat

** History of codes
loc interm " "*h/o*" "
foreach word in `interm'  {
	replace sa_cat = 1 if strmatch(lc_readterm, "`word'")
	}
*
local incode ""%%%" "
foreach word in `incode' {
	replace sa_cat = 1 if strmatch(readcode, "`word'")
}
*
** Vasospastic codes
loc interm " "*prinzmetal*" "*variant*" "*spasm*" "
foreach word in `interm'  {
	replace sa_cat = 2 if strmatch(lc_readterm, "`word'")
	}
*
local incode ""%%%" "
foreach word in `incode' {
	replace sa_cat = 2 if strmatch(readcode, "`word'")
}
*** Cardiac syndrome x codes
loc interm " "*syndrome x*""
foreach word in `interm'  {
	replace sa_cat = 3 if strmatch(lc_readterm, "`word'")
	}
local incode ""%%%" "
foreach word in `incode' {
	replace sa_cat = 3 if strmatch(readcode, "`word'")
}
*
//STEP 7: SAVE LIST OF CODES
*** drop redundant variables
drop marker lc_readterm immun_events db_date
order medcode readcode readterm sa_cat
sort sa_cat readcode
*** save to STATA file and text file
save "stable_angina_codes.dta", replace
outsheet using "stable_angina_codes.txt", replace
```

Table 30: Entity Code report for HDL/LDL ratio and Total: HDL ratio

GPRD variable report: HDL LDL ratio
Generated on: Fri May 6 11:18:26 BST 2011

*** Test
Total records: 107058

Readcode breakdown:
14369 - HDL : LDL ratio - 95022 records
19853 - Serum LDL/HDL ratio - 7537 records
49695 - Plasma LDL/HDL ratio - 4499 records

Entity type breakdown:

Entity: 338 - HDL/LDL ratio - 106976
data1 - Operator
3 - '=' - 106802 records
0 - 'Data Not Entered' - 172 records
1 - '<' - 2 records
data3 - Unit of measure
151 - 'ratio' - 71691 records
0 - 'No Data Entered' - 23925 records
96 - 'mmol/L' - 10423 records
161 - '1/1' - 649 records
1 - '%' - 238 records
104 - 'ng' - 26 records
26 - '1' - 15 records
156 - 'mmol' - 3 records
124 - 's' - 1 records
185 - 'L/L' - 1 records
97 - 'mmol/mol' - 1 records
122 - 'rad' - 1 records
99 - 'mol/L' - 1 records
126 - 'u' - 1 records
data4 - Qualifier
0 - 'Data Not Entered' - 90269 records
9 - 'Normal' - 7263 records
41 - 'High' - 4609 records
8 - 'High' - 3978 records
10 - 'Low' - 651 records
40 - 'Low' - 95 records
7 - 'Very High' - 74 records
12 - 'Abnormal' - 32 records
13 - 'Potential Abnormal' - 4 records
11 - 'Very Low' - 1 records
data7 - Normal range basis
0 - 'Data Not Entered' - 106921 records
5 - 'Generic normal range' - 53 records
4 - 'Gestational age based' - 2 records

Entity: 288 - Other laboratory tests - 82
data1 - Operator
0 - 'Data Not Entered' - 78 records
3 - '=' - 4 records
data3 - Unit of measure
0 - 'No Data Entered' - 82 records
data4 - Qualifier
0 - 'Data Not Entered' - 82 records
data7 - Normal range basis
0 - 'Data Not Entered' - 82 records

*** Referral
Total records: 80

Readcode breakdown:
14369 - HDL : LDL ratio - 80 records
19853 - Serum LDL/HDL ratio - 0 records
49695 - Plasma LDL/HDL ratio - 0 records

*** Clinical

Total records: 18
Total records with data:

Readcode breakdown:
14369 - HDL : LDL ratio - 17 records
19853 - Serum LDL/HDL ratio - 1 records
49695 - Plasma LDL/HDL ratio - 0 records

GPRD variable report: Total HDL ratio
Generated on: Fri May 6 11:13:04 BST 2011

*** Test

Total records: 1993866

Readcode breakdown:
14105 - Cholesterol/HDL ratio - 16312 records
14108 - HDL : total cholesterol ratio - 321582 records
14371 - Serum cholesterol/HDL ratio - 888987 records
14372 - Total cholesterol:HDL ratio - 736790 records
40935 - Plasma cholesterol/HDL ratio - 30195 records

Entity type breakdown:

Entity: 163 - Serum cholesterol - 600
data1 - Operator
3 - '=' - 600 records
data3 - Unit of measure
0 - 'No Data Entered' - 356 records
96 - 'mmol/L' - 179 records
1 - '%' - 48 records

26 - '1' - 12 records
151 - 'ratio' - 5 records
data4 - Qualifier
0 - 'Data Not Entered' - 577 records
9 - 'Normal' - 12 records
8 - 'High' - 11 records
data7 - Normal range basis
0 - 'Data Not Entered' - 600 records

Entity: 265 - Sinus x-ray - 1
data1 - Qualifier
0 - 'Data Not Entered' - 1 records

Entity: 214 - Blood lipids - 93
data1 - Qualifier
0 - 'Data Not Entered' - 93 records

Entity: 338 - HDL/LDL ratio - 1969029
data1 - Operator
3 - '=' - 1968679 records
0 - 'Data Not Entered' - 339 records
4 - '>' - 6 records
2 - '<=' - 3 records
1 - '<' - 1 records
5 - '>=' - 1 records

data3 - Unit of measure
0 - 'No Data Entered' - 1737307 records
151 - 'ratio' - 150422 records
161 - '1/1' - 63005 records
1 - '%' - 14930 records
26 - '1' - 3147 records
96 - 'mmol/L' - 129 records
104 - 'ng' - 59 records
126 - 'u' - 4 records
146 - 'us' - 4 records
156 - 'mmol' - 3 records
187 - 'mmol/mmol' - 3 records
124 - 's' - 2 records
21 - '/mL' - 2 records
41 - 'd' - 2 records
45 - 'feet' - 2 records
142 - 'umol/L' - 1 records
125 - 'stone' - 1 records
110 - 'nmol/L' - 1 records
82 - 'mg/dL' - 1 records
27 - '1(tot)' - 1 records
185 - 'L/L' - 1 records
61 - 'iu/L' - 1 records

86 - 'mg/mmol' - 1 records

data4 - Qualifier
0 - 'Data Not Entered' - 1856216 records
41 - 'High' - 96733 records
9 - 'Normal' - 6737 records
8 - 'High' - 4924 records
44 - 'Abnormal' - 1874 records
12 - 'Abnormal' - 1184 records
7 - 'Very High' - 634 records
40 - 'Low' - 194 records
38 - 'Potentially abnormal' - 183 records
10 - 'Low' - 133 records
43 - 'Significantly High' - 124 records
13 - 'Potential Abnormal' - 87 records
11 - 'Very Low' - 3 records
14 - 'Outside ref range' - 3 records

data7 - Normal range basis
0 - 'Data Not Entered' - 1969016 records
5 - 'Generic normal range' - 13 records

Entity: 288 - Other laboratory tests - 12965
data1 - Operator
3 - '=' - 12785 records
0 - 'Data Not Entered' - 180 records
data3 - Unit of measure
0 - 'No Data Entered' - 11949 records
1 - '%' - 611 records
151 - 'ratio' - 246 records
161 - '1/1' - 155 records
96 - 'mmol/L' - 2 records
123 - 'rem' - 1 records
26 - '1' - 1 records
data4 - Qualifier
0 - 'Data Not Entered' - 12829 records
25 - 'Normal' - 106 records
26 - 'Abnormal' - 30 records
data7 - Normal range basis
0 - 'Data Not Entered' - 12965 records

Entity: 438 - Other biochemistry test - 11178
data1 - Operator
3 - '=' - 11175 records
0 - 'Data Not Entered' - 2 records
4 - '>' - 1 records
data3 - Unit of measure
151 - 'ratio' - 6150 records
96 - 'mmol/L' - 2484 records
0 - 'No Data Entered' - 2234 records
161 - '1/1' - 222 records

1 - '%' - 72 records
123 - 'rem' - 12 records
185 - 'L/L' - 1 records
122 - 'rad' - 1 records
187 - 'mmol/mmol' - 1 records
100 - 'mol/s' - 1 records

data4 - Qualifier
0 - 'Data Not Entered' - 9964 records
41 - 'High' - 926 records
8 - 'High' - 171 records
9 - 'Normal' - 99 records
7 - 'Very High' - 6 records
10 - 'Low' - 6 records
13 - 'Potential Abnormal' - 5 records
12 - 'Abnormal' - 1 records

data7 - Normal range basis
0 - 'Data Not Entered' - 11178 records

*** Referral

Total records: 195

Readcode breakdown:
14105 - Cholesterol/HDL ratio - 31 records
14108 - HDL : total cholesterol ratio - 7 records
14371 - Serum cholesterol/HDL ratio - 13 records
14372 - Total cholesterol:HDL ratio - 144 records
40935 - Plasma cholesterol/HDL ratio - 0 records

*** Clinical

Total records: 809

Readcode breakdown:
14105 - Cholesterol/HDL ratio - 329 records
14108 - HDL : total cholesterol ratio - 215 records
14371 - Serum cholesterol/HDL ratio - 20 records
14372 - Total cholesterol:HDL ratio - 245 records
40935 - Plasma cholesterol/HDL ratio - 0 records

Table 31: Reverse Entity Code Report for 217 Electrocardiogram: Read codes, and their frequency, used with entity code

Read code	Read Term	Frequency in CALIBER
32...12	ECG	292000
3212.00	Standard ECG	204872
32...00	Electrocardiography	91476
3216.00	ECG normal	33647
32M..00	24 Hour ECG	23698
321B.00	12 lead ECG	17536
3272.00	ECG: atrial fibrillation	16746
32...11	Cardiography - ECG	16582
3217.00	ECG abnormal	10886
321..00	ECG - general	10196
3211.00	ECG requested	3712
324..00	ECG:left ventricle hypertrophy	1645
3242.00	ECG: shows LVH	1186
329A.00	ECG: left bundle branch block	1130
3299.00	ECG: right bundle branch block	1031
32Z..00	Electrocardiography NOS	955
3241.00	ECG: no LVH	897
3213.00	Exercise ECG	822
322..00	ECG: myocardial ischaemia	781
3214.00	Ambulatory ECG	778
321C.00	ECG sinus rhythm	740
3273.00	ECG: atrial flutter	733
3219.00	ECG equivocal	478
321A.00	ECG - no new changes	439
3213000	Exercise ECG normal	366
329..00	ECG: heart block	353
321Z.00	ECG - general - NOS	324
3282.00	ECG: ventricular tachycardia	235
3263.00	ECG: ventricular ectopics	203
3274.00	ECG: paroxysmal atrial tachy.	172
32F4.00	ECG: T wave inverted	157
323..00	ECG: myocardial infarction	154
32J5.00	Left axis deviation	147
327..00	ECG: supraventricular arrhythmia	141
3214000	Ambulatory ECG normal	135
3232.00	ECG: old myocardial infarction	133
3264.00	ECG: atrial ectopics	126
32E2.00	ECG: S-T interval abnormal	113
3222.00	ECG:shows myocardial ischaemia	105
328..00	ECG: ventricular arrhythmia	103
329Z.00	ECG: heart block NOS	98
3213100	Exercise ECG abnormal	96
326..00	ECG: ectopic beats	90
327Z.00	ECG: supraventric. arryth. NOS	88
3215.00	ECG not done	87
3262.00	ECG: extrasystole	85
32L..00	ECG: left ventricular strain	74
32...13	EKG	69
32F2.00	ECG: T wave abnormal	65
3294.00	ECG:partial A-V block-long P-R	62
32E4.00	ECG: S-T depression	57
3233.00	ECG: antero-septal infarct.	52
3214100	Ambulatory ECG abnormal	51

Read code	Read Term	Frequency in CALIBER
32J1.00	ECG: QRS complex normal	49
32E3.00	ECG: S-T elevation	46
324Z.00	ECG: LVH NOS	46
32J2.00	ECG: QRS complex abnormal	45
3297.00	ECG: Wenckebach phenomenon	36
3234.00	ECG:posterior/inferior infarct	35
32F3.00	ECG: T wave flattened	32
3298.00	ECG: complete A-V block	27
3213011	Negative exercise ECG test	27
32K3.00	ECG: Q-T interval prolonged	25
325..00	ECG:right ventricle hypertrop.	25
R143100	[D]Electrocardiogram (ECG) abnormal	24
3221.00	ECG: no myocardial ischaemia	21
328Z.00	ECG: ventricular arrhythmia NOS	21
32B2.00	ECG: Q wave abnormal	20
32B..00	ECG: Q wave	20
32F..00	ECG: T wave	19
326Z.00	ECG: ectopic beats NOS	19
33B9100	Exercise tolerance test done	18
32I3.00	ECG: P-R interval prolonged	15
32C2.00	ECG: R wave abnormal	14
32J6.00	Right axis deviation	14
323Z.00	ECG: myocardial infarct NOS	14
3271.00	ECG: no supraventric. arryth.	12
3231.00	ECG: no myocardial infarction	12
32FZ.00	ECG: T wave NOS	11
3293.00	ECG:complete sinu-atrial block	10
3218.00	ECG - improved	9
3283.00	ECG: ventricular fibrillation	9
32A2.00	ECG: P wave abnormal	9
3213111	Positive exercise ECG test	8
3292.00	ECG: partial sinu-atrial block	8
3236.00	ECG: lateral infarction	7
322Z.00	ECG: myocardial ischaemia NOS	7
3295.00	ECG: partial A-V block - 2:1	6
32C..00	ECG: R wave	6
32I4.00	ECG: P-R interval shortened	6
3291.00	ECG: no heart block	6
32JZ.00	ECG: QRS complex NOS	5
32E1.00	ECG: S-T interval normal	5
3251.00	ECG: no RVH	5
3261.00	ECG: no ectopic beats	5
32KZ.00	ECG: Q-T interval NOS	5
32C1.00	ECG: R wave normal	5
8HR1.00	Refer for ECG recording	4
32IZ.00	ECG: P-R interval NOS	4
32C3.00	ECG: R wave tall	4
32I1.00	ECG: P-R interval normal	4
32A..00	ECG: P wave	4
32BZ.00	ECG: Q wave NOS	4
32AZ.00	ECG: P wave NOS	4
32B3.00	ECG: Q wave pathological	4
32B1.00	ECG: Q wave normal	3
32A3.00	ECG: P mitrale	3
32m..00		3

Read code	Read Term	Frequency in CALIBER
32A4.00	ECG: P pulmonale	3
3252.00	ECG: shows RVH	3
G20..00	Essential hypertension	2
32I..00	ECG: P-R interval	2
32K1.00	ECG: Q-T interval normal	2
32EZ.00	ECG: S-T interval NOS	2
32F1.00	ECG: T wave normal	2
325Z.00	ECG: RVH NOS	2
32K..00	ECG: Q-T interval	2
662P.00	Hypertension monitoring	2
662..12	Hypertension monitoring	2
32E..00	ECG: S-T interval	2
3235.00	ECG: subendocardial infarct	2
32A1.00	ECG: P wave normal	2
32I2.00	ECG: P-R interval abnormal	2
G2...00	Hypertensive disease	2
32D1.00	ECG: S wave normal	1
3281.00	ECG: no ventricular arrhythmia	1
56F..00	Diagnostic procedure declined	1
4145.00	Blood sample -> Lab NOS	1
G580.00	Congestive heart failure	1
8A52.11	ECG monitoring	1
32K4.00	ECG: Q-T interval shortened	1
32K2.00	ECG: Q-T interval abnormal	1
32D3.00	ECG: S wave deep	1
32J..00	ECG: QRS complex	1
J10y400	Oesopheal reflux without mention of oesophagitis	1
G65..00	Transient cerebral ischaemia	1
32J3.00	ECG: QRS complex prolonged	1

Appendix D: Medications

Table 32: Nitrates and specific anti-anginal medications

BNF chapter	Drug category	Specific drugs
2.6.1	Nitrates	Glyceryl trinitrate, isosorbide dinitrate, isosorbide mononitrate
2.6.3	Other antianginal drugs	Ivabradine, nicorandil, ranolazine

Table 33: Blood-pressure -lowering medication

BNF chapter	Drug category	Specific drugs
2.2.1	Thiazides & related diuretics	bendroflumethiazide, chlortalidone, cyclopenthiazide, indapamide, metolazone, xipamide
2.2.3	Potassium-sparing diuretics and aldosterone antagonists	Amiloride hydrochloride, triamterene, aldosterone antagonists (eplerenone, spironolactone)
2.2.4	Potassium-sparing diuretics with other diuretics	Co-amilofruse, co-amilozide, triamterene with furosemide, amiloride with cyclopenthiazide, amiloride with bumatanide, triamterene with thiazides, spironolactone with thiazides, spironolactone with loop diuretics
2.4	Beta-adrenoceptor blocking drugs	Propranolol hydrochloride, acebutolol, atenolol, bisoprolol fumarate, carvedilol, celiprolol hydrochloride, co-tenidone, esmolol hydrochloride labetalol hydrochloride, metoprolol tartrate, nadolol, nebivolol, oxprenolol hydrochloride, pindolol, sotalol hydrochloride. (Not timolol maleate eye drops)
2.5	Hypertension and heart failure	(All sections listed below)
2.5.1	Vasodilator antihypertensive drugs	Hydralazine hydrochloride, minoxidil. (Not sildenafil, sodium nitroprusside, tadalafil Ambrisentan, bosentan, diazoxide, iloprost)
2.5.2	Centrally acting antihypertensive drugs	Moxonidine, methyldopa (not clonidine)
2.5.4	Alpha-adrenoceptor blocking drugs	Doxazosin, indoramin, prazosin, terazosin

2.5.5	Drugs affecting the renin-angiotensin system	(All sections listed below)
2.5.5.1	Angiotensin-converting enzyme inhibitors (ACE inhibitors)	Captopril, cilazapril, enalapril maleate, fosinopril sodium imidapril hydrochloride, lisinopril, moexipril hydrochloride, perindopril erbumine, perindopril arginine, quinapril, ramipril, ramipril with felodipine, trandolapril
2.5.5.2	Angiotensin-II receptor antagonists	Candesartan cilexetil, eprosartan, irbesartan, losartan potassium, olmesartan medoxomil, telmisartan, valsartan
2.5.5.3	Renin inhibitors	aliskiren
2.6.2	Calcium-channel blockers	Amlodipine, diltiazem hydrochloride, felodipine, isradipine, lacidipine, lercanidipine hydrochloride, nicardipine hydrochloride, nifedipine, nimodipine, verapamil hydrochloride

Excluded drugs:

- Timolol eye drops (BNF 2.4 Beta-adrenoceptor blocking drugs)
- Sildenafil, sodium nitroprusside, tadalafil, bosentan, diazoxide, ambrisentan, iloprost (BNF 2.5.1 Vasodilator antihypertensive drugs)
- Clonidine (BNF 2.5.2 Centrally acting antihypertensive drugs)
- Guanethidine monosulphate (BNF 2.5.3 Adrenergic neurone blocking drugs)

Table 34: Statins

BNF chapter	Drug category	Specific drugs
2.12.1	Statins	Atorvastatin, fluvastatin, pravastatin sodium, rosuvastatin, simvastatin

Appendix E: Hierarchy of diagnoses used to select initial presentation

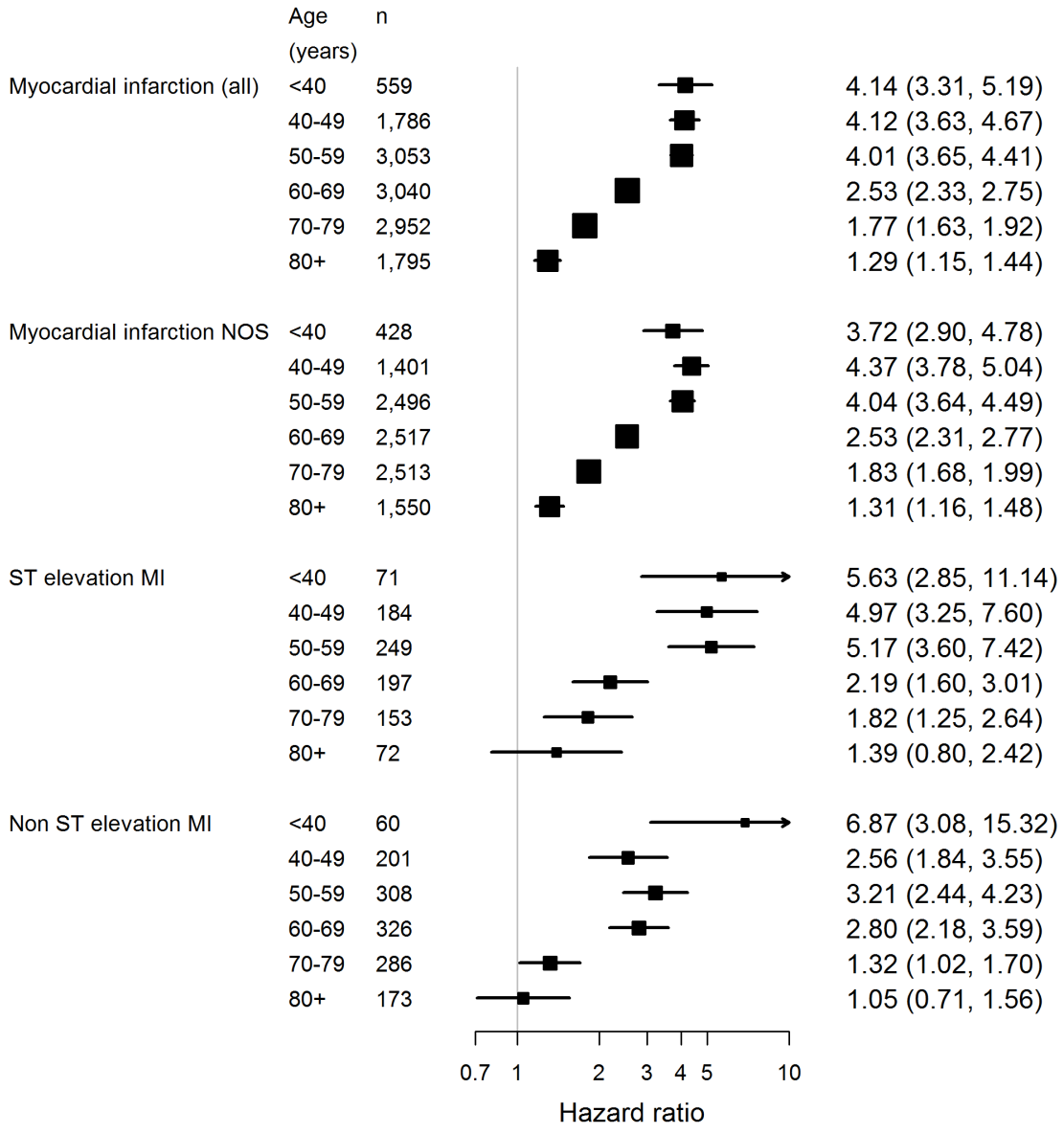
category	variable	event_code	order
End of study		400	0.5
Left practice	tod_date	401	0.5
Last practice download	lcd_date	402	0.5
Specified other deaths	ons_mortality	500	0.8
CHD death	ons_mortality	501	0.8
Sudden cardiac death	ons_mortality	502	0.8
Cardiac arrest death	ons_mortality	503	0.8
Heart failure death	ons_mortality	504	0.8
Cerebrovascular death (Stroke)	ons_mortality	505	0.8
Peripheral arterial disease death	ons_mortality	506	0.8
Abn cerebro test	cerebro_test_gprd	88	1
Ischaem Stroke	ischaem_stroke_gprd	69	2
ischaemic stroke (HES)	ishc_stroke_hes	141	3
Stroke NOS	stroke_nos_gprd	72	4
Stroke NOS (HES)	stroke_nos_hes	146	5
isch cerebro disease NEC	cerebro_NEC_gprd	74	6
ischaemic cvd (HES)	ischaem_cvd_hes	151	7
Cerebrovascular ops	cerebro_ops_gprd	75	8
cerebrovascular ops (OPCS)	cerebro_procs_opcs	215	9
PAD test abnormal	pad_test_gprd	99	10
AAA	aaa_gprd	77	11
AAA (HES)	aaa_hes	156	12
AAA ops	aaa_ops_gprd	79	13
AAA ops (OPCS)	aaa_procs_opcs	221	14
PAD	pad_gprd	78	15
PAD (HES)	pad_hes	158	16
PAD ops	pad_ops_gprd	80	17
PAD ops (OPCS)	pad_procs_opcs	225	18
CHD NOS (HES)	chd_nos_hes	120	19
CHD NOS	chd_nos_gprd	56	20

category	variable	event_code	order
abnormal	itests_gprd	22	21
abnormal	atest_gprd	27	22
beginning of at least two	nitrate_meds_gprd	30	23
Stable angina diagnosis	SA_diagnosis	13	24
CABG	CABG_gprd	16	25
CABG revision	CABG_gprd	17	26
PCI	pci_gprd	19	27
PCI during admission (MINAP)	coronary_intervention	331	28
CABG done	cabg_opcs	200	29
CABG revision	cabg_opcs	201	30
PCI done	pci_opcs	202	31
Acute coronary syndrome	acs_gprd	58	32
Acute IHD (HES)	acute_ihd_hes	125	33
unstable angina	UA_gprd	35	34
angina pectoris (HES)	angina_hes	105	35
Unstable angina (HES)	uangina_hes	115	36
UA	MI_phenotype	322	37
transluminal coronary thrombolysis	lysis_gprd	50	38
coronary thrombolysis, mode unspecified	lysis_gprd	51	39
transluminal coronary thrombolysis	lysis_opcs	204	40
Troponin abnormal	gprd_troponins_cat	95	41
Cardiac markers abnormal	gprd_cardiac_markers_cat	97	42
CKMB abnormal	gprd_ckmb_cat	98	43
MI, timing uncertain (HES)	mi_hes	112	44
MI NOS	mi_nos_gprd	45	45
Definite MI (HES)	mi_hes	111	46
nSTEMI	nstemi_gprd	41	47
nSTEMI	MI_phenotype	321	48
STEMI	stemi_gprd	38	49
STEMI (MINAP)	MI_phenotype	320	50
Ventricular tachycardia	arrest_gprd	63	51
Ventricular fibrillation	arrest_gprd	64	52
Implanted cardiac defibrillation device	arrest_gprd	65	53

category	variable	event_code	order
Asystole/EMD/cardiac arrest/resusc	arrest_gprd	66	54
ventricular tachycardia (HES)	arrest_hes	133	55
V Fib (HES)	arrest_hes	134	56
ICD (HES)	arrest_hes	135	57
asystole/EMD/cardiac arrest/resusc (HES)	arrest_hes	136	58
ventricular arryhthmia (HES)	arrest_hes	137	59
ICD device (OPCS)	arrest_opcs	206	60
asystole/EMD/cardiac arrest/resusc	arrest_opcs	207	61
HF NOS	hf_gprd	85	62
HF, specified other cause	hf_gprd	86	63
Echo abnormal: LVD	echo	87	64
HF NOS (HES)	hf_hes	167	65
HF: hypertension cause	hf_gprd	84	66
HF hypertension cause (HES)	hf_hes	166	67
Sudden cardiac death	sudden_death_gprd	92	68
sudden cardiac death (HES)	sudden_death_hes	172	69

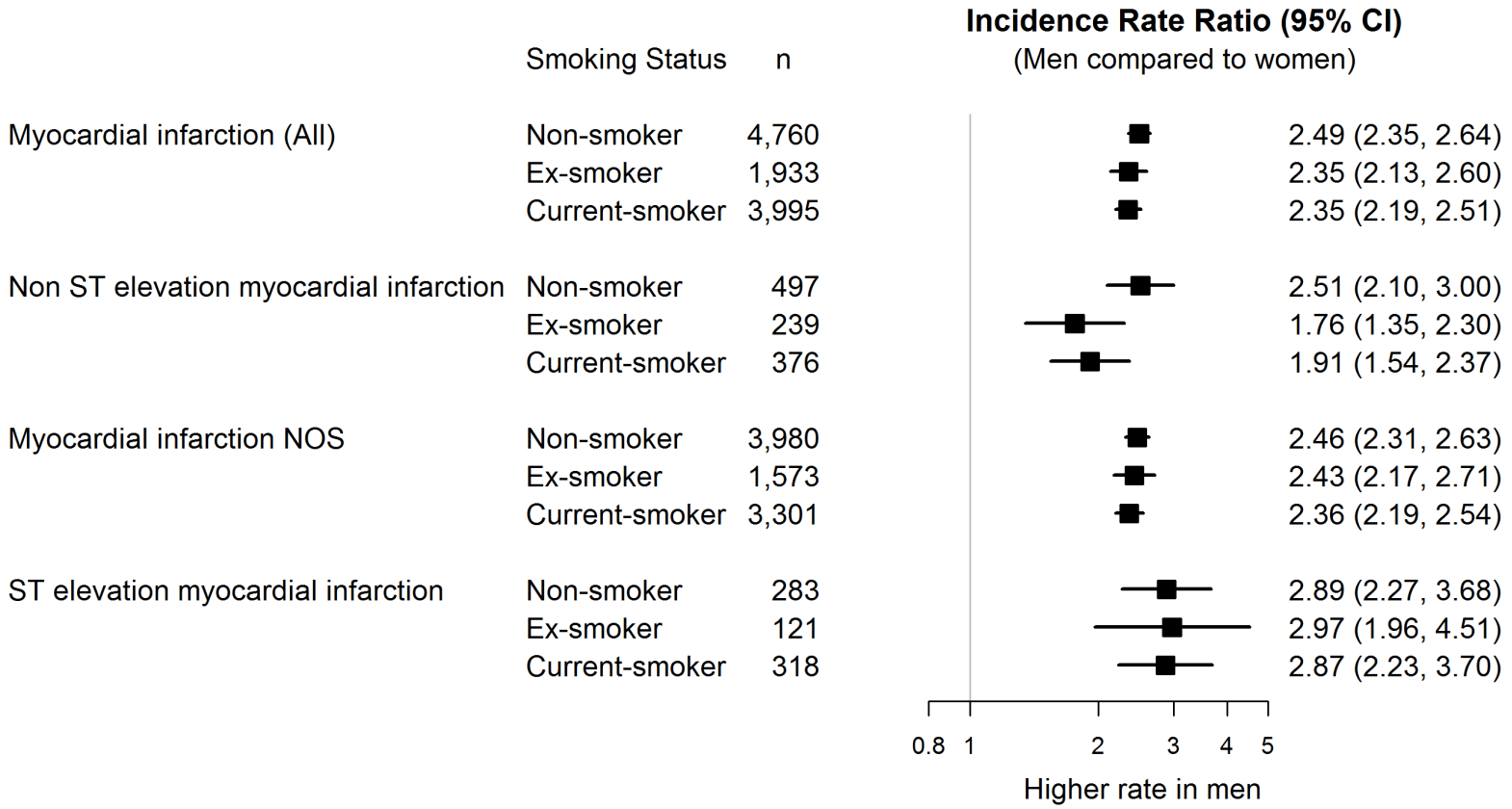
Appendix F: Additional Tables and Figures for Chapter 5 Gender and the Initial Presentation of a Wide Range of Cardiovascular Diseases

Figure 43: Incidence rate ratios (men compared to women) for initial presentation of myocardial infarction types, stratified by age group



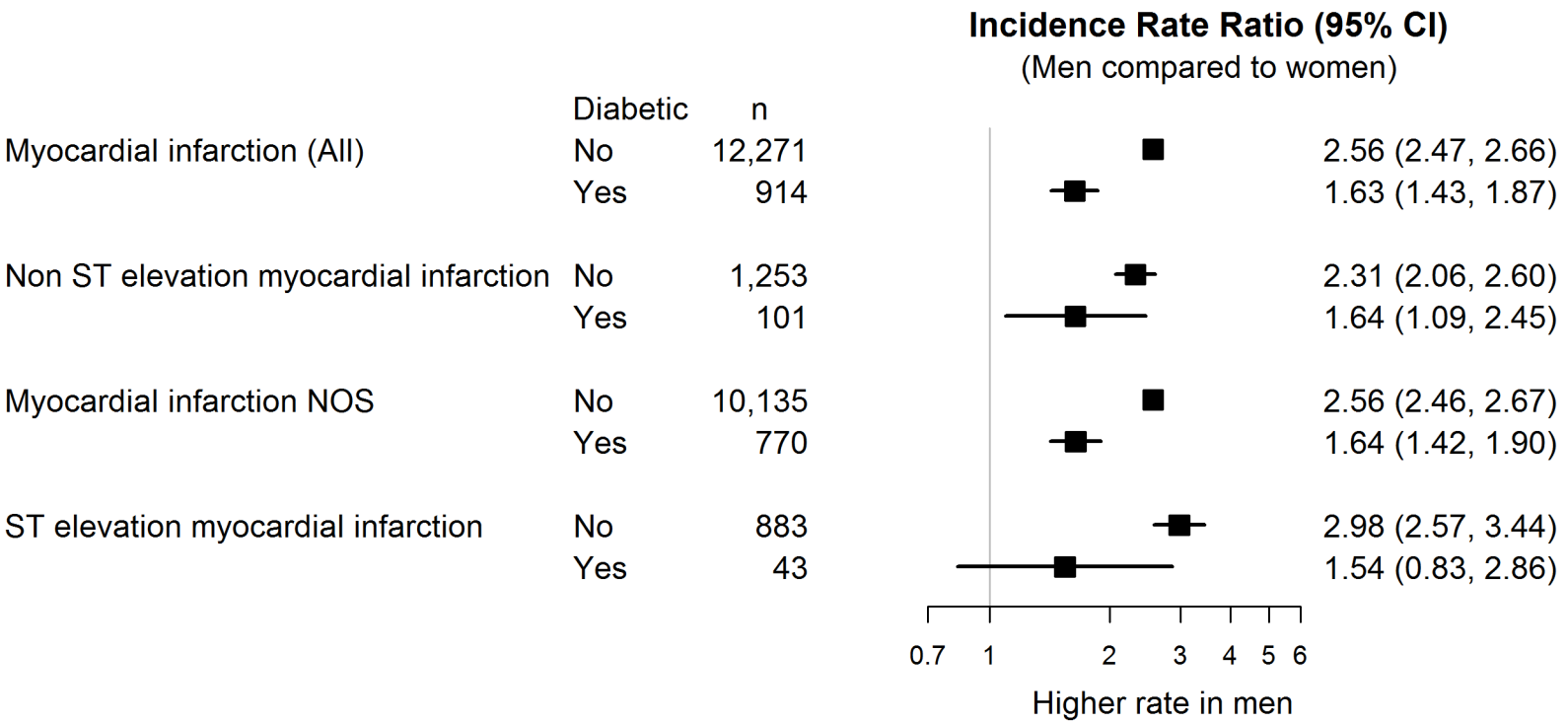
N=1,758,584. NOS indicates not otherwise specified; MI, myocardial infarction

Figure 44: Incidence rate ratios (men compared to women) for initial presentation of specific myocardial infarction phenotypes, stratified by smoking status



N=1,453,635. NOS indicates not otherwise specified; MI, myocardial infarction

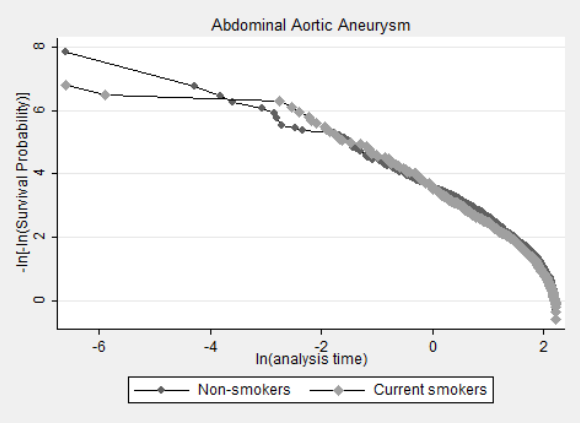
Figure 45: Incidence rate ratios (men compared to women) for initial presentation of myocardial infarction types, stratified by diabetes status

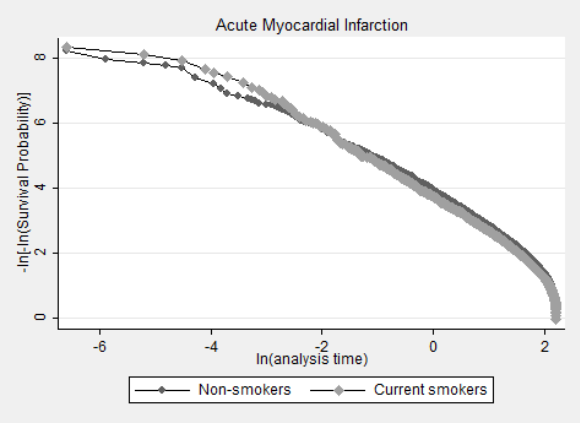


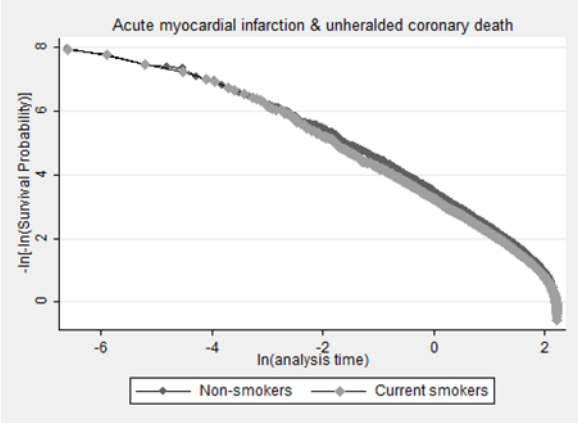
N=1,758,584. NOS indicates not otherwise specified; MI, myocardial infarction

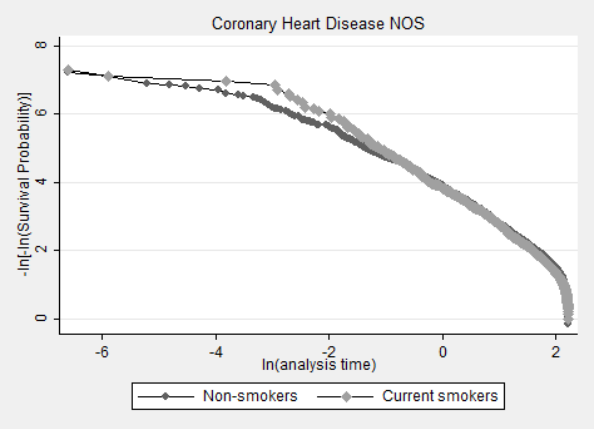
Appendix G: Additional Tables and Figures for Chapter 6 - Association of smoking with initial presentation of cardiovascular disease across a wide range of presentations

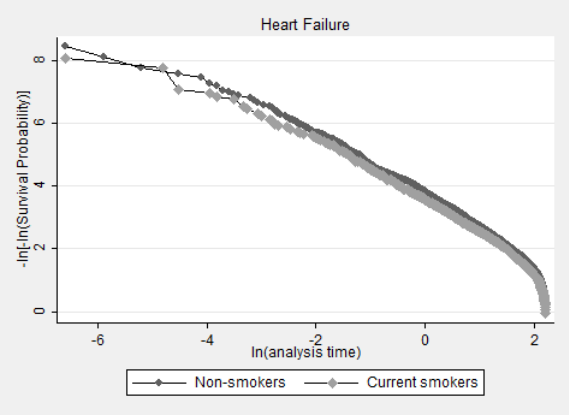
Figure 46: Proportional hazard of smoking compared to not smoking for range of initial presentations of CVD (in alphabetical order)

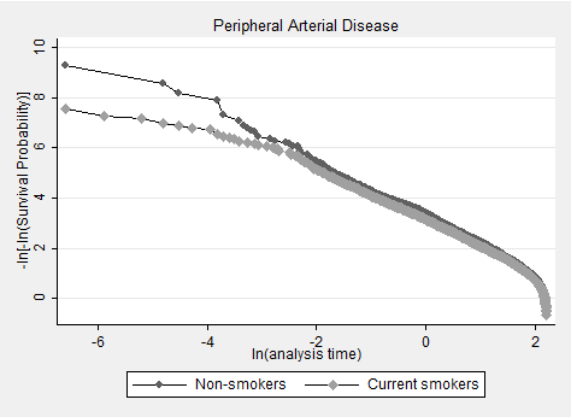


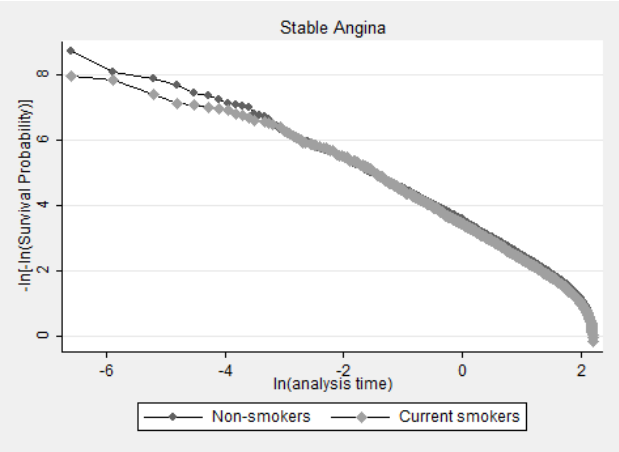


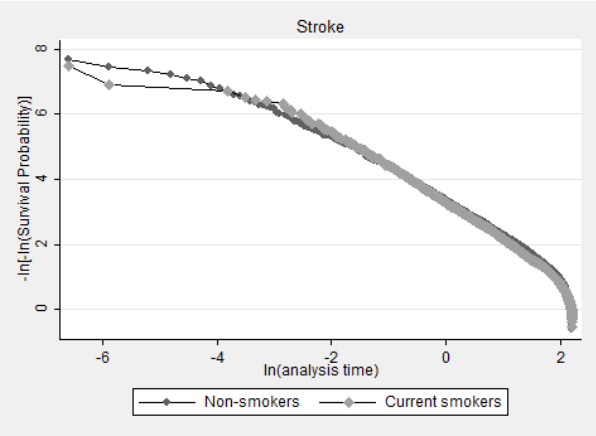


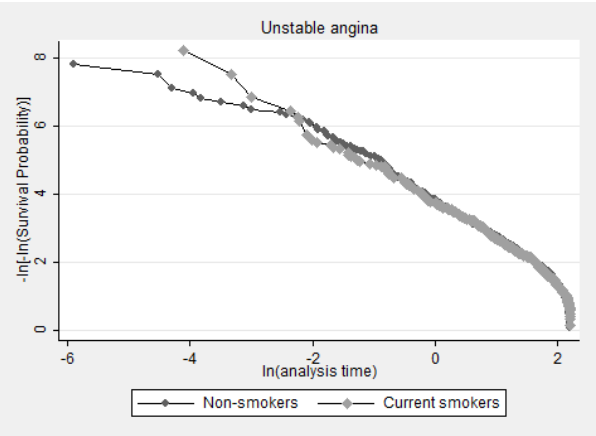


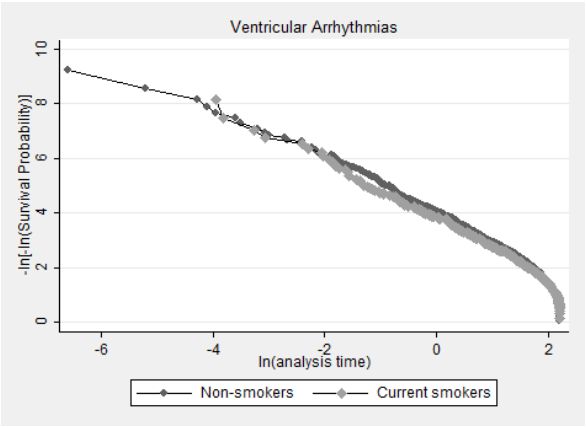


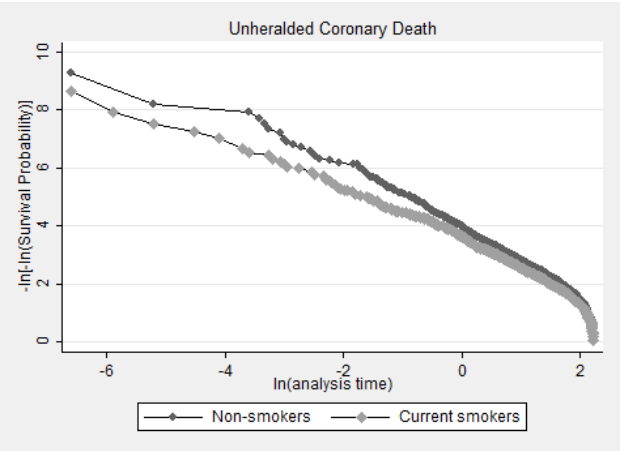






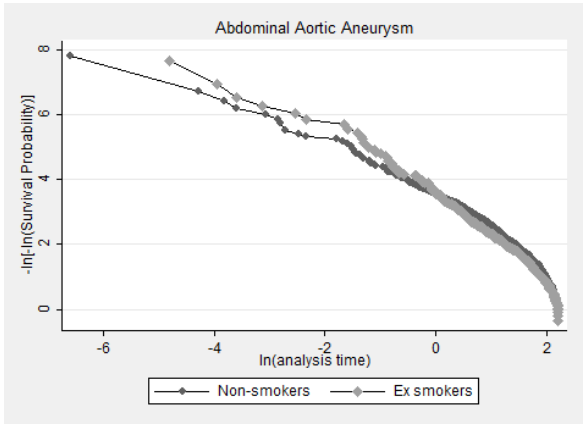


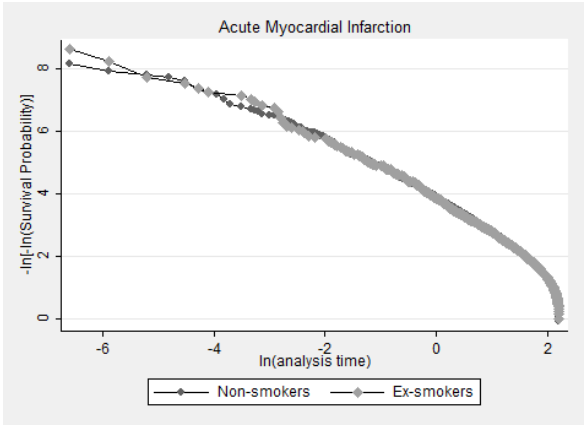


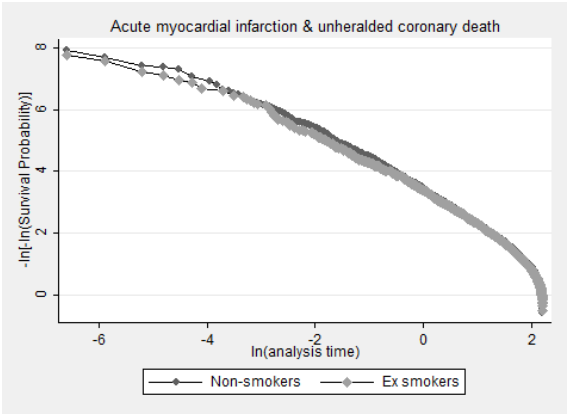


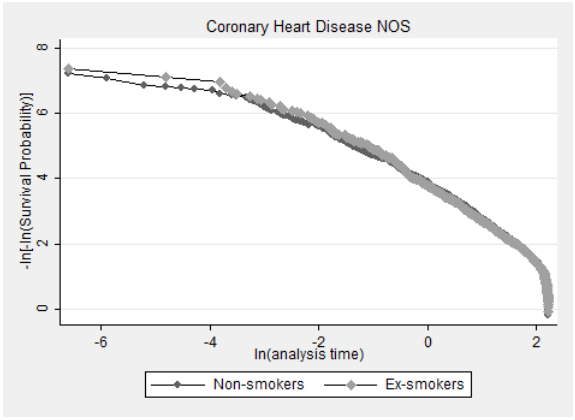
Log log graph of current smokers compared to non-smokers, adjusted for age at entry. Ln indicates log.

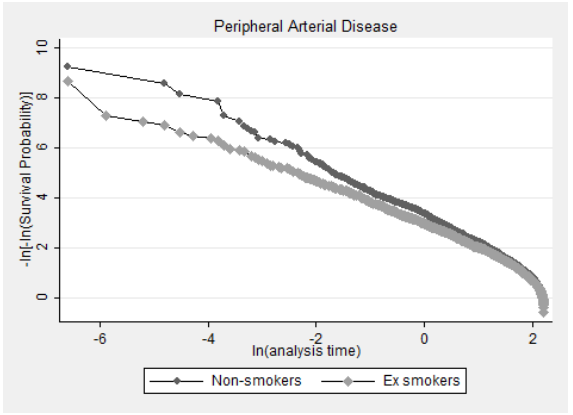
Figure 47: Proportional hazard of ex-smoking compared to not smoking for range of initial presentations of CVD (in alphabetical order)

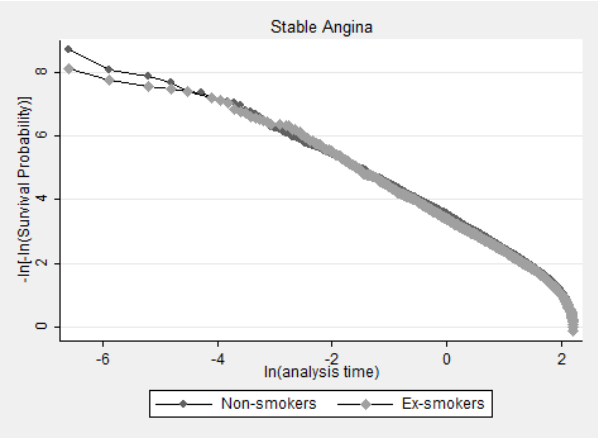


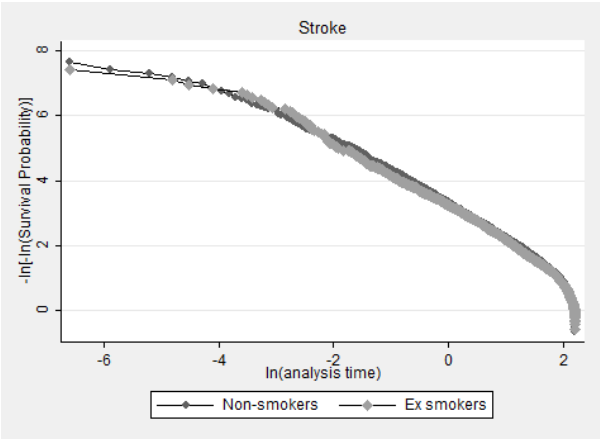


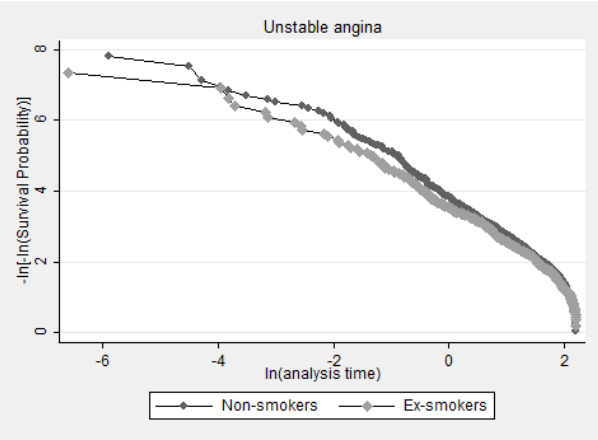


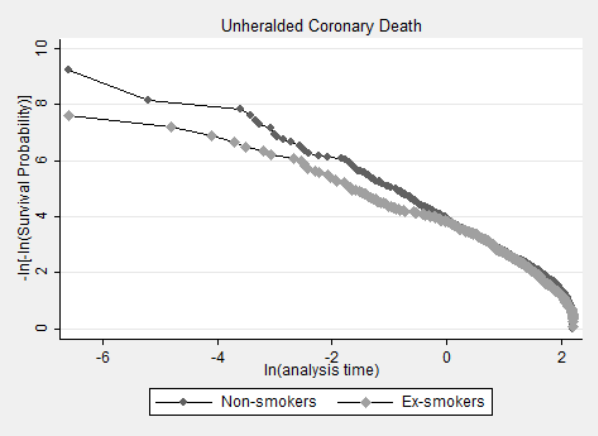


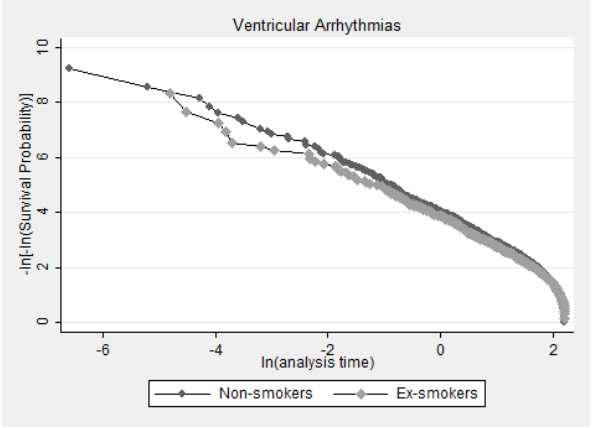






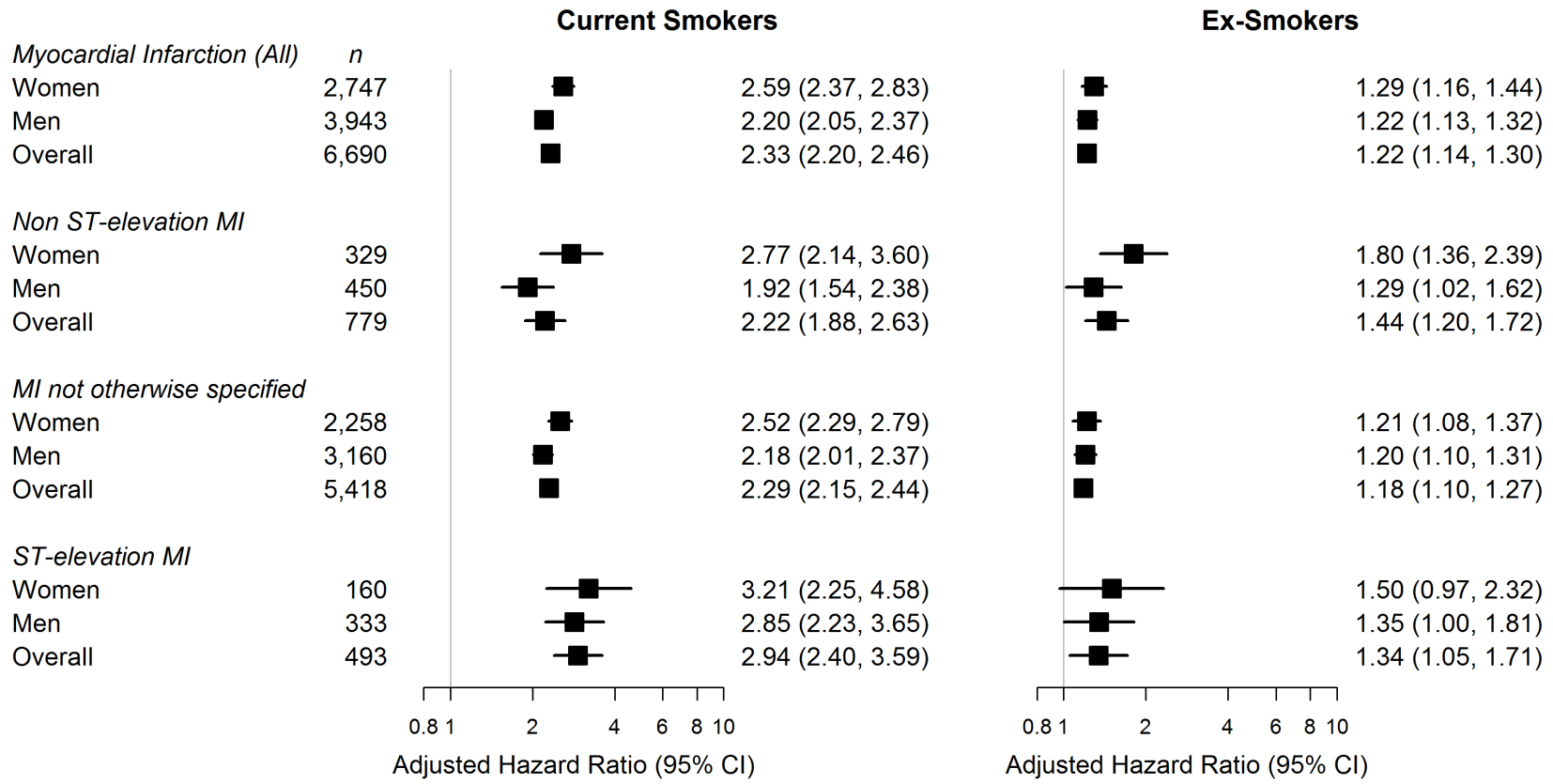






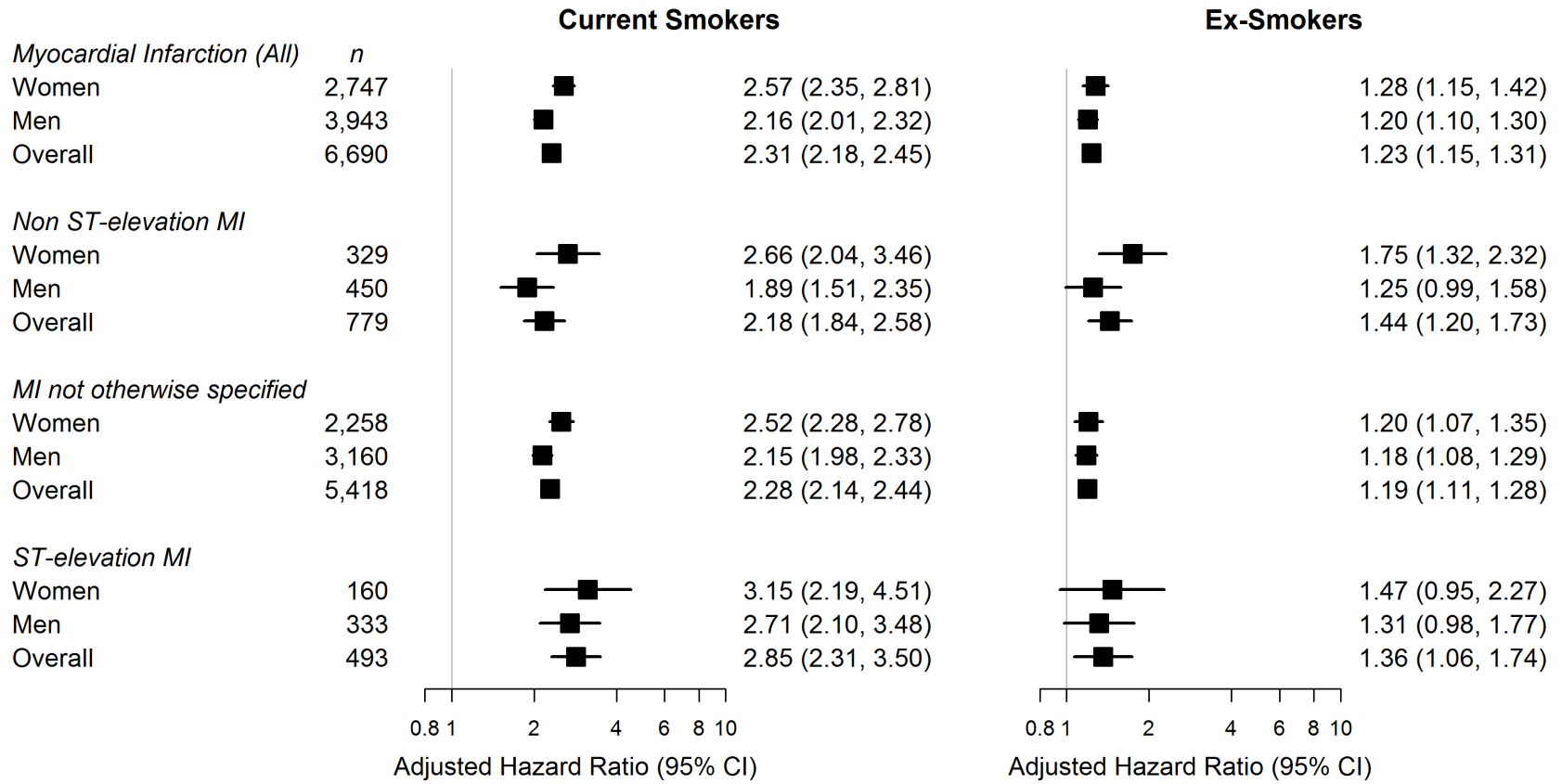
Log log graph of current smokers compared to non-smokers, adjusted for age at entry. Ln indicates log.

Figure 48: Age-adjusted hazard ratios for initial presentations of acute myocardial infarction overall and for AMI subtypes associated with being current smoker or ex-smoker compared to non-smokers, overall and in women and men



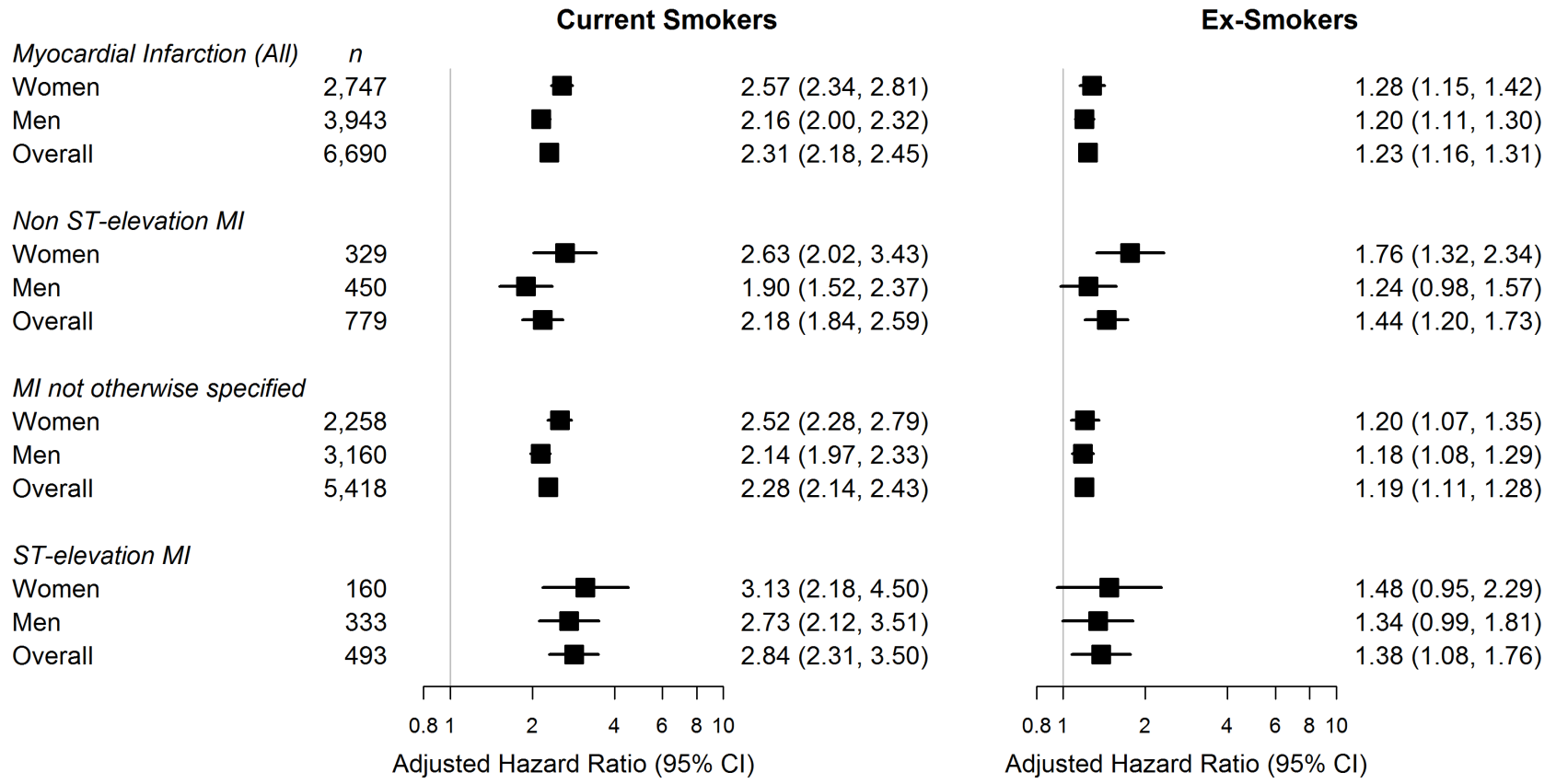
Hazard ratios adjusted for age at baseline for men and women, and additionally for sex for all patients, in complete cases. N= 897,892. MI indicates myocardial infarction; CI, confidence interval; HR, hazard ratio.

Figure 49: Multivariable adjusted hazard ratios for initial presentations of cardiac disease associated with being current smoker or ex-smoker compared to non-smokers, overall and in women and men



Hazard ratios adjusted for age at baseline, sex, deprivation (index of multiple deprivation quintile), smoking status (current smoker, ex-smoker, non-smoker), diabetes mellitus (yes/no), mean systolic blood pressure at baseline, blood pressure medication (yes/no) and statin use at baseline (yes/no) , in complete cases. N= 897,892. MI indicates myocardial infarction; CI, confidence interval; HR, hazard ratio.

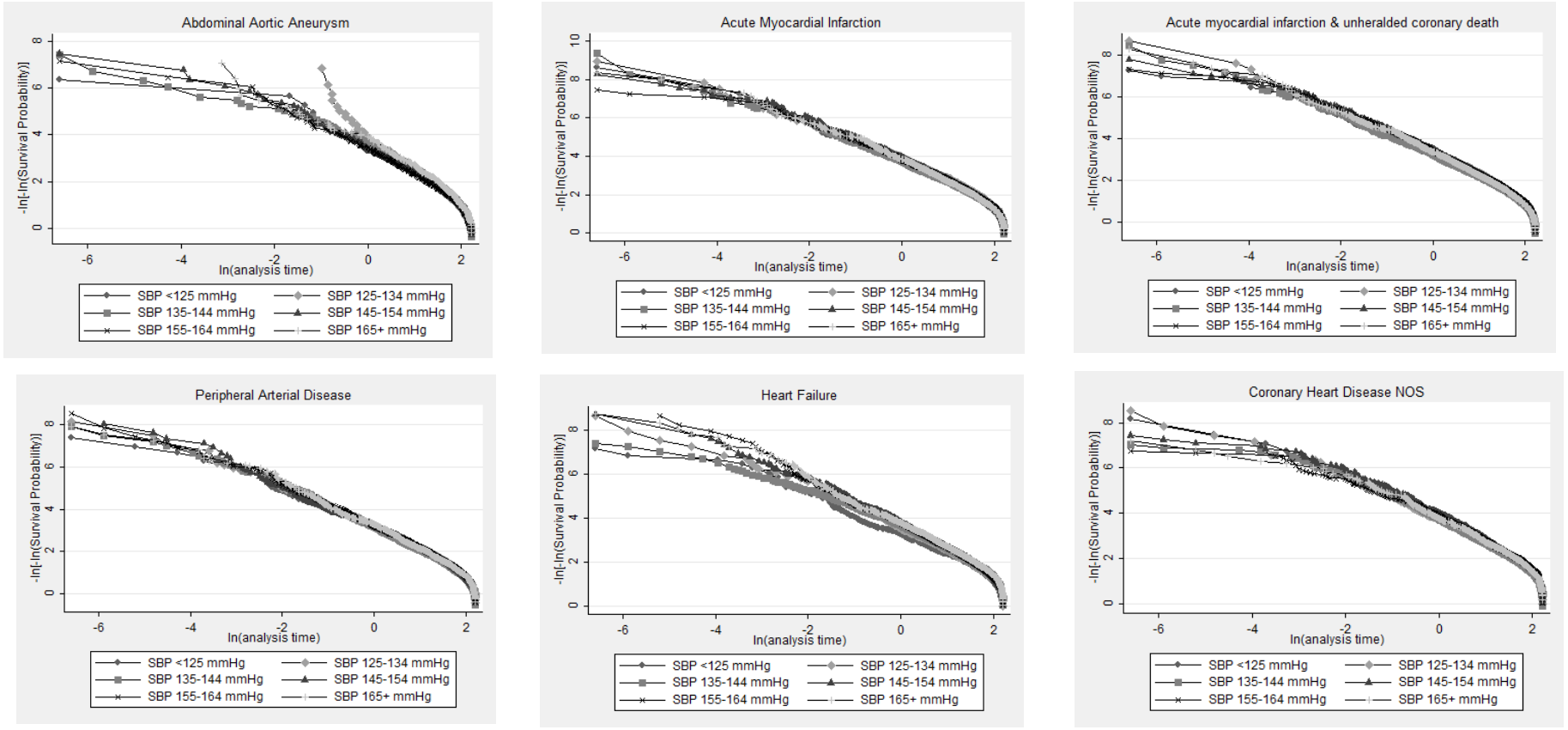
Figure 50: Multivariable adjusted hazard ratios (with random effects for GP practice) for initial presentations of cardiac disease associated with being current smoker or ex-smoker compared to non-smokers, overall and in women and men

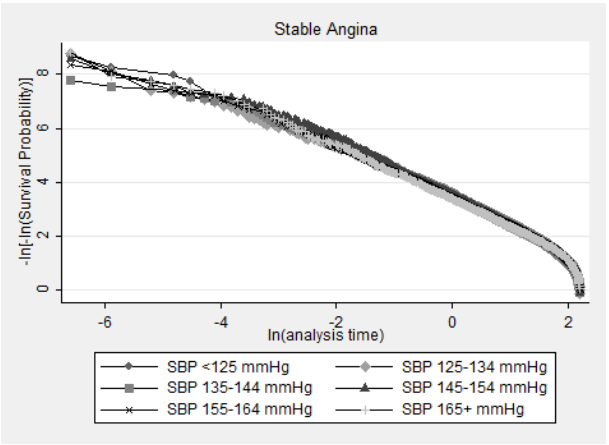


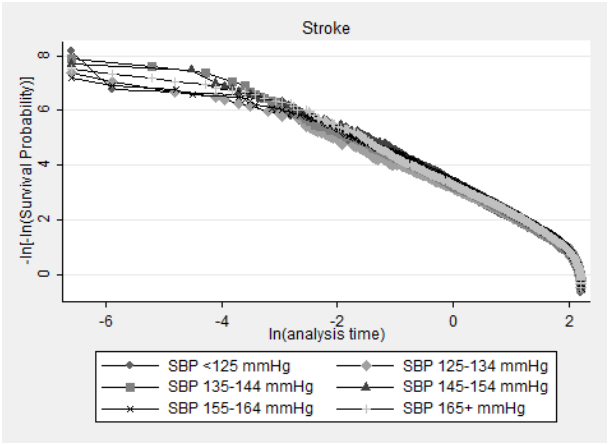
Hazard ratios adjusted for age at baseline, sex, deprivation (index of multiple deprivation quintile), smoking status (current smoker, ex-smoker, non-smoker), diabetes mellitus (yes/no), mean systolic blood pressure at baseline, blood pressure medication (yes/no) and statin use at baseline (yes/no), with frailty term to take account of clustering at practice level, in complete cases. N= 897,892. MI indicates myocardial infarction; CI, confidence interval; HR, hazard ratio.

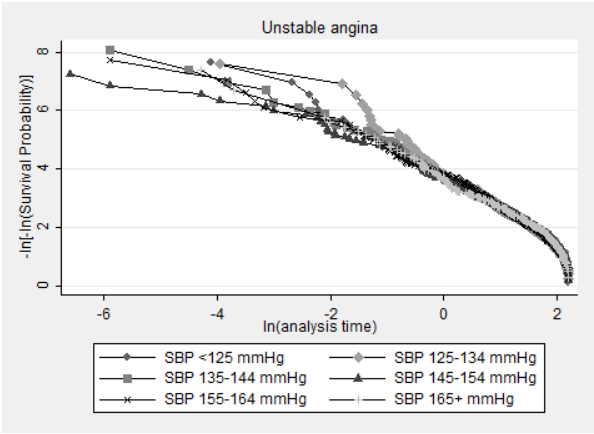
Appendix H: Additional Tables and Figures for Chapter 7 - Association of blood pressure with initial presentation of cardiovascular disease across a wide range of presentations

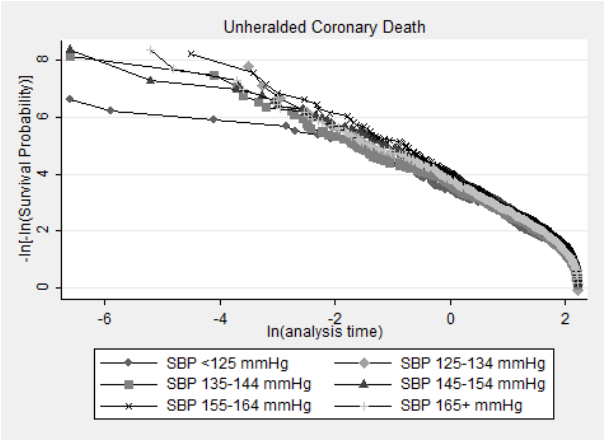
Figure 51: Proportional hazards of SBP categories (10mmHg) for range of initial presentations of CVD (in alphabetical order)

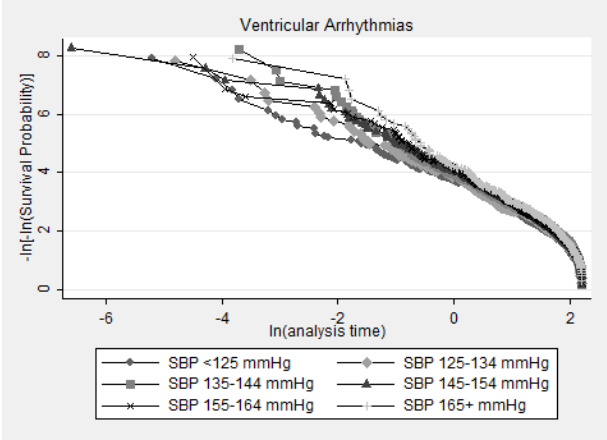






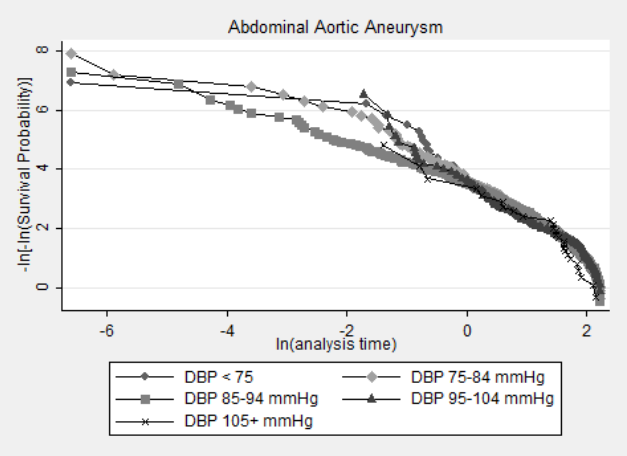


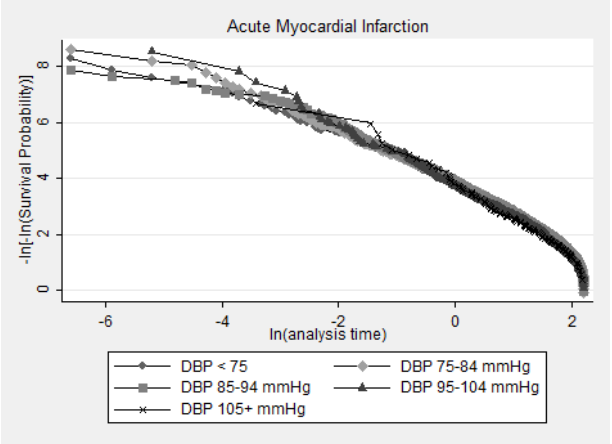


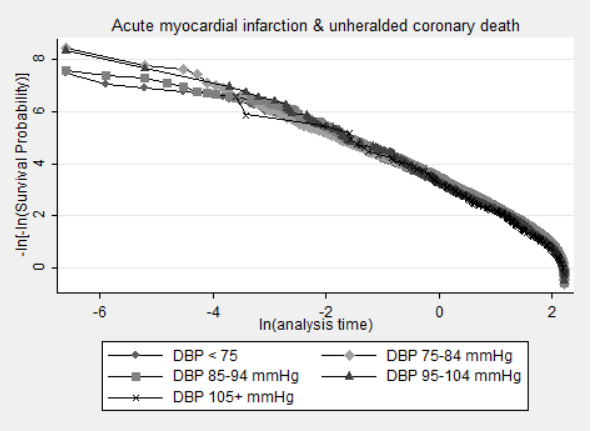


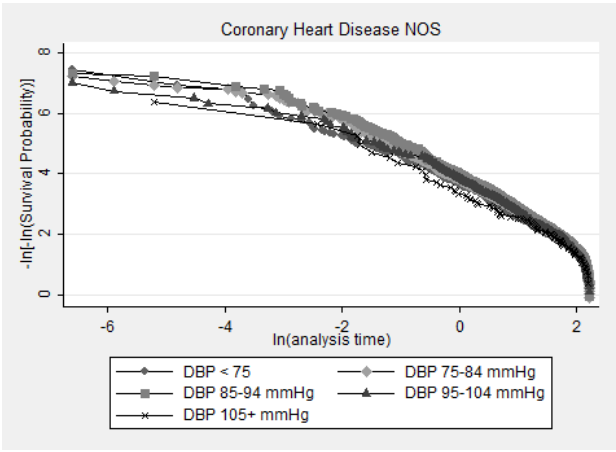
Log log graph of categorical SBP variable, adjusted for age at entry. Ln indicates log.

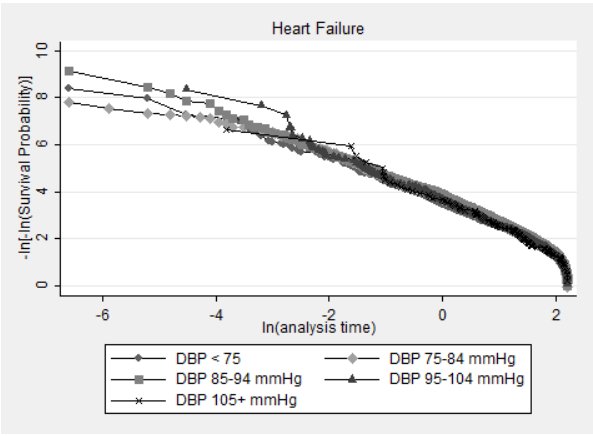
Figure 52: Proportional hazard of DBP categories (10mmHg) for range of initial presentations of CVD (in alphabetical order)

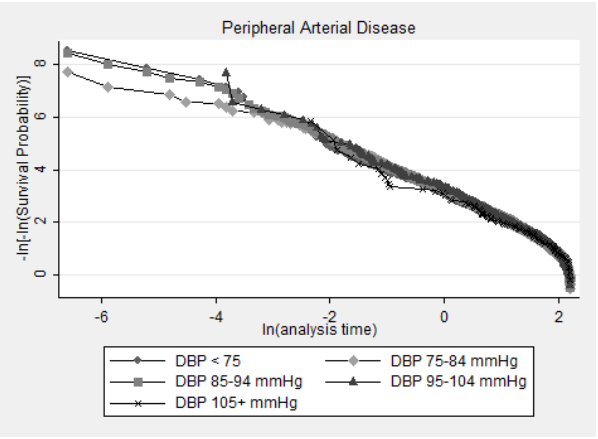


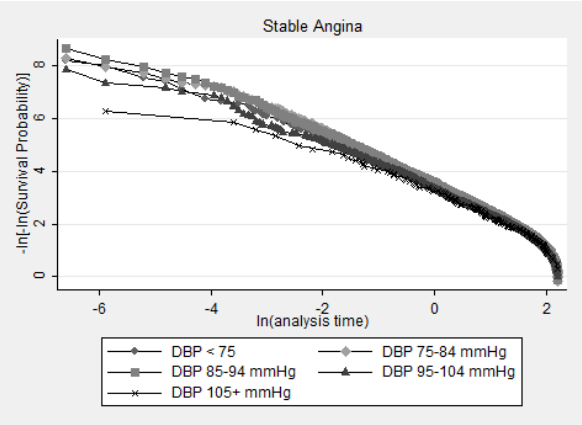


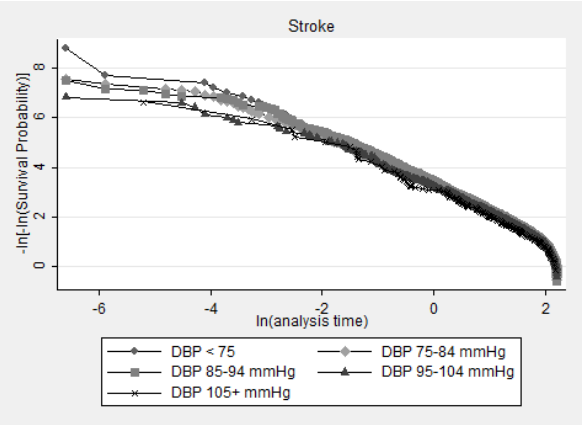


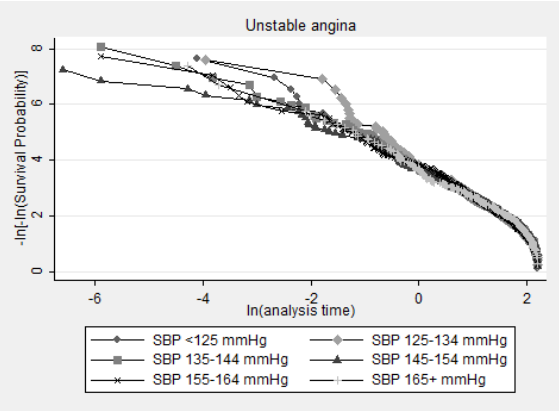


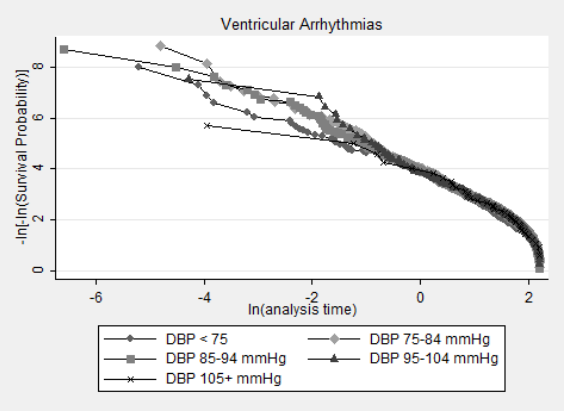




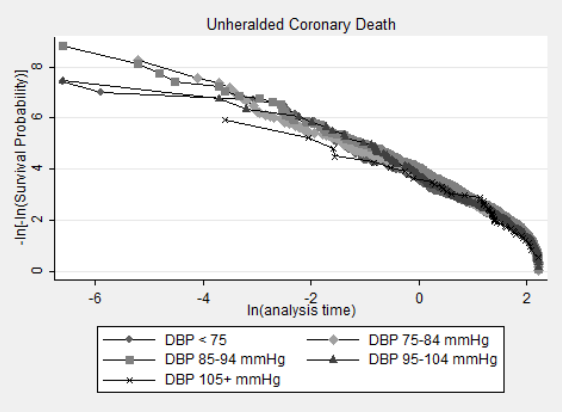








at



Log log graphs of categorical DBP variable, adjusted for age entry. Ln indicates log.

Table 35: Hazard ratio for gender-BP interaction for AAA, Acute MI / Coronary death, PAD and Stroke

	HR*	p	LR chi2(1)	Prob > chi2
SBP				
AMI + UCD	1.01 (0.98, 1.04)	0.617	0.25	0.6167
Stroke	0.95 (0.92, 0.98)	0.000	12.82	0.0003
AAA	0.95 (0.88, 1.02)	0.181	1.78	0.1824
PAD	0.98 (0.94, 1.01)	0.220	1.50	0.2202
DBP				
AMI + UCD	0.99 (0.94, 1.03)	0.565	0.33	0.5649
Stroke	0.93 (0.88, 0.97)	0.001	11.51	0.0007
AAA	0.97 (0.86, 1.10)	0.652	0.20	0.6519
PAD	0.98 (0.93, 1.04)	0.543	0.37	0.5425

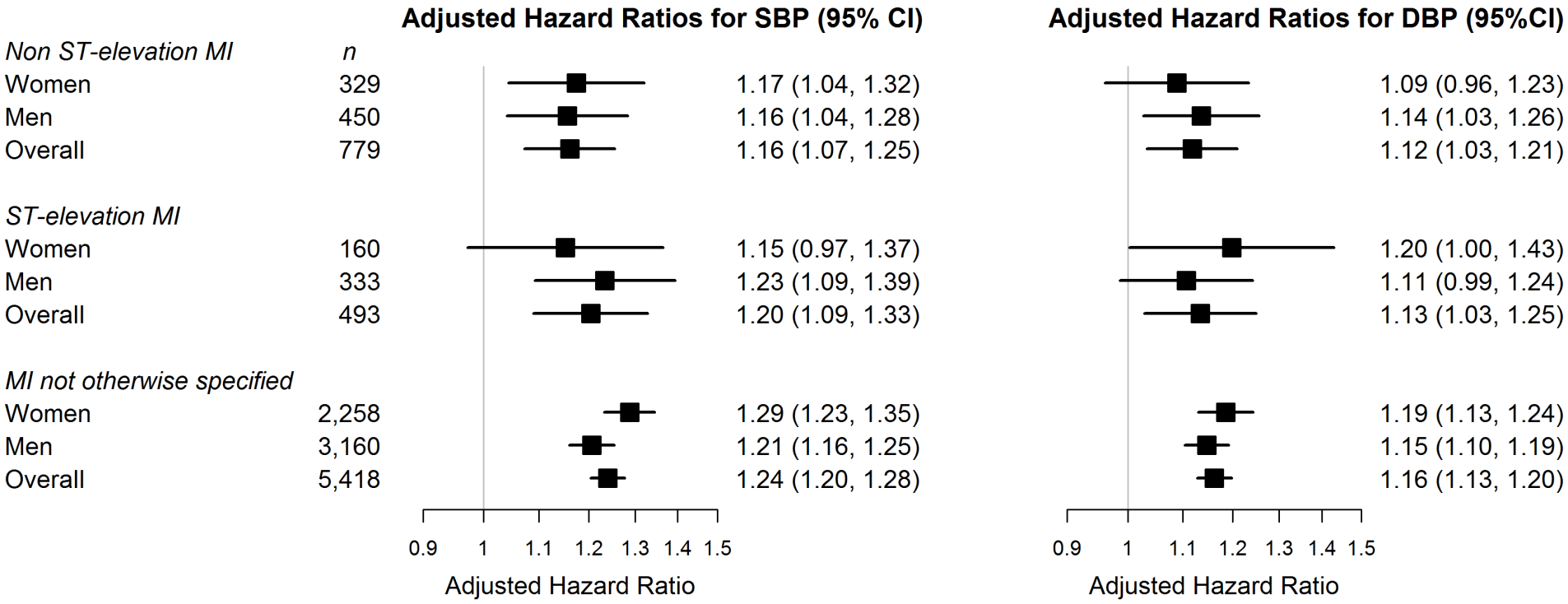
**Interaction HR in model adjusted for age, age^2, gender, age-gender interaction, deprivation, diabetes, statin use, BP, and interaction between BP and gender for specific CVD endpoints.*

Table 36: Hazard ratios for gender-BP interaction for cardiac endpoints

	HR*	p	LR test	P
SBP				
Stable angina	1.00 (0.97, 1.04)	0.830	0.05	0.8289
Unstable angina	1.02 (0.92, 1.13)	0.720	0.13	0.7190
CHD NOS	1.04 (0.98, 1.11)	0.238	1.39	0.2377
AMI	1.04 (0.99, 1.10)	0.121	2.40	0.1215
Heart failure	1.01 (0.96, 1.06)	0.715	0.13	0.7146
Ventricular arrhythmias	1.19 (1.08, 1.30)	0.000	13.44	0.0002
Unheralded coronary death	0.92 (0.84, 1.00)	0.040	4.22	0.0398
DBP				
Stable angina	1.02 (0.98, 1.07)	0.233	1.42	0.2329
Unstable angina	0.96 (0.87, 1.06)	0.435	0.61	0.4348
CHD NOS	1.01 (0.95, 1.08)	0.684	0.17	0.6837
AMI	1.00 (0.95, 1.06)	0.930	0.01	0.9303
Heart failure	0.97 (0.92, 1.03)	0.324	0.97	0.3241
Ventricular arrhythmia	1.10 (1.00, 1.21)	0.039	4.24	0.0395
Unheralded coronary death	0.94 (0.86, 1.02)	0.145	2.13	0.1448

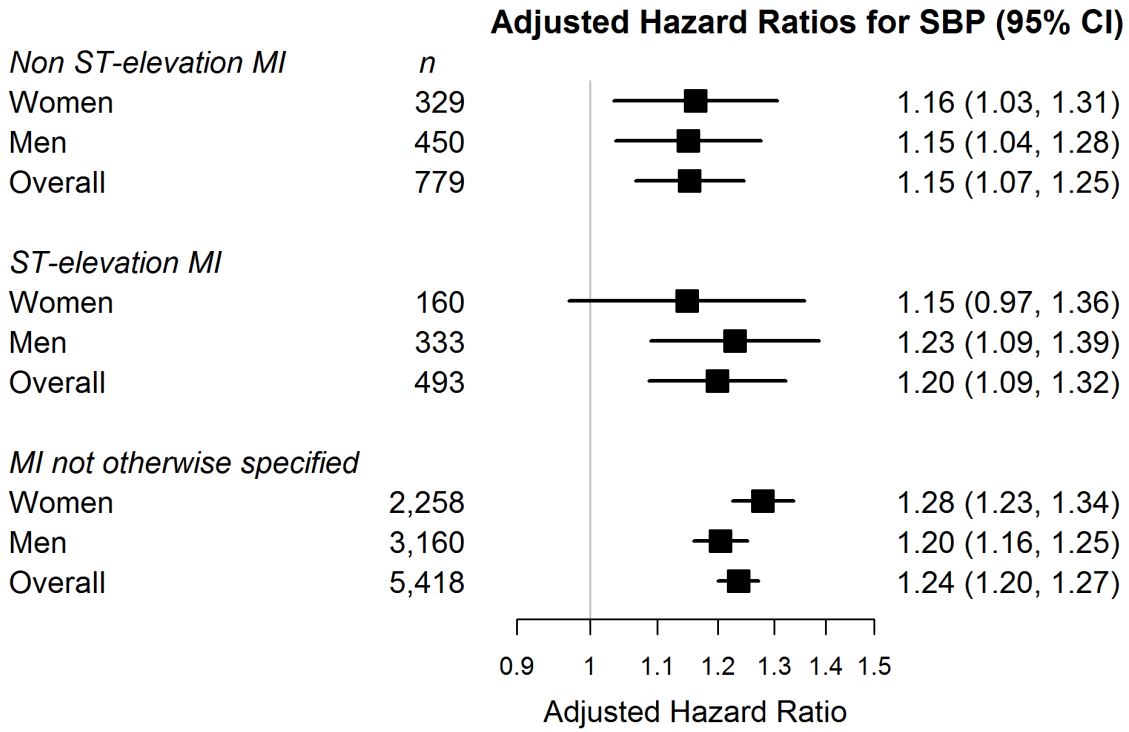
**Interaction HR in model adjusted for age, age^2, gender, age-gender interaction, deprivation, diabetes, statin use, BP, and interaction between BP and gender.*

Figure 53: Age-adjusted hazard ratios for initial presentations of ST elevation myocardial infarction, non ST elevation myocardial infarction and myocardial infarction not otherwise specified associated with 1 standard deviation increase in systolic and diastolic blood pressure overall and in women and men



Hazard ratios for mean systolic or diastolic blood pressure at baseline (per 1 standard deviation (sd)), adjusted for age at baseline, and sex in complete cases. N= 897,892. For SBP, 1 sd = 18.9 mmHg; for DBP, 1 sd=10 mmHg. MI indicates myocardial infarction; CI, confidence interval; DBP, diastolic blood pressure; HR, hazard ratio; SBP, systolic blood pressure.

Figure 54: Multivariable adjusted hazard ratios for initial presentations of ST-elevation myocardial infarction, non ST-elevation myocardial infarction and myocardial infarction not otherwise specified associated with 1 standard deviation increase in systolic and diastolic blood pressure overall and in women and men



Hazard ratios for mean systolic or diastolic blood pressure at baseline (per 1 standard deviation (sd)), adjusted for age at baseline, sex, deprivation (index of multiple deprivation quintile), smoking status (current smoker, ex-smoker, non-smoker), diabetes mellitus (yes/no) and statin use at baseline (yes/no), in complete cases. N= 897,892. For SBP, 1 sd = 18.9 mmHg; for DBP, 1 sd=10 mmHg. MI indicates myocardial infarction; CI, confidence interval; DBP, diastolic blood pressure; HR, hazard ratio; SBP, systolic blood pressure.

Table 37: Hazard ratios for gender-BP interaction for specific myocardial infarction endpoints

	HR*	p	LR test	P
SBP				
STEMI	0.93 (0.76, 1.14)	0.484	0.49	0.4838
NSTEMI	1.00 (0.86, 1.17)	0.986	0.00	0.9864
MI NOS	1.06 (1.00, 1.12)	0.060	3.53	0.0604
DBP				
STEMI	1.01 (0.95, 1.07)	0.643	0.21	0.6432
NSTEMI	0.93 (0.79, 1.08)	0.340	0.91	0.3398
MI NOS	1.05 (0.85, 1.29)	0.776	0.08	0.7765

**Interaction HR in model adjusted for age, age^2, gender, age-gender interaction, deprivation, diabetes, statin use, BP, and interaction between BP and gender.*

References

1. Murabito JM, Evans JC, Larson MG, Levy D. Prognosis after the onset of coronary heart disease. An investigation of differences in outcome between the sexes according to initial coronary disease presentation. *Circulation* 1993;88:2548–55.

2. Huxley RR, Woodward M. Cigarette smoking as a risk factor for coronary heart disease in women compared with men: a systematic review and meta-analysis of prospective cohort studies. *Lancet* 2011;378:1297–305.

3. Harding S, Rosato M, Teyhan A. Trends for coronary heart disease and stroke mortality among migrants in England and Wales, 1979-2003: slow declines notable for some groups. *Heart* 2008;94:463–70.

4. Aune E, Røislien J, Mathisen M, Thelle DS, Otterstad JE. The "smoker's paradox" in patients with acute coronary syndrome: a systematic review. *BMC Med* 2011;9:97.

5. Zaman MJ, Shipley MJ, Stafford M, Brunner EJ, Timmis AD, Marmot MG, *et al.* Incidence and prognosis of angina pectoris in South Asians and Whites: 18 years of follow-up over seven phases in the Whitehall-II prospective cohort study. *JPublic Heal* 2011;33:430–8.

6. Daly CA, De Stavola B, Sendon JLL, Tavazzi L, Boersma E, Clemens F, *et al.* Predicting prognosis in stable angina--results from the Euro heart survey of stable angina: prospective observational study. *BMJ* 2006;332:262–7.

7. Scarborough P, Bhatnagar P, Wickramasinghe K, Smolina K, Mitchell C, Rayner M. Coronary heart disease statistics 2010 edition. London: : British Heart Foundation 2010. http://www.heartstats.org/datapage.asp?id=9075

8. Timmis AD, Feder G, Hemingway H. Prognosis of stable angina pectoris: why we need larger population studies with higher endpoint resolution. *Heart* 2007;93:786–91.

9. Carrol L. *Alice's Adventures in Wonderland*. London: : Macmillan 1865.

10. Hemingway H, Croft P, Perel P, Hayden J, Abrams K, Timmis A, *et al.* Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *BMJ* 2013;:forthcoming.

11. Champney KP, Frederick PD, Bueno H, Parashar S, Foody J, Merz CN, *et al.* The joint contribution of sex, age and type of myocardial infarction on hospital mortality following acute myocardial infarction. *Heart* 2009;95:895–9.

12. Terkelsen CJ, Lassen JF, Norgaard BL, Gerdes JC, Jensen T, Gotzsche LB, *et al.* Mortality rates in patients with ST-elevation vs. non-ST-elevation acute myocardial infarction: observations from an unselected cohort. *Eur Heart J* 2005;26:18–26.

13. Abbott JD, Ahmed HN, Vlachos HA, Selzer F, Williams DO. Comparison of outcome in patients with ST-elevation versus non-ST-elevation acute myocardial infarction treated with percutaneous coronary intervention (from the National Heart, Lung, and Blood Institute Dynamic Registry). *Am J Cardiol* 2007;100:190–5.

14. Hsia J, Aragaki A, Bloch M, LaCroix AZ, Wallace R. Predictors of angina pectoris versus myocardial infarction from the Women's Health Initiative Observational Study. *Am J Cardiol* 2004;93:673–8.

15. Merry AHH, Boer JMA, Schouten LJ, Feskens EJM, Verschuren WMM, Gorgels APM, *et al.* Smoking, alcohol consumption, physical activity, and family history and the risks of acute myocardial infarction and unstable angina pectoris: a prospective cohort study. *BMC Cardiovasc Disord* 2011;11:13.

16. Capewell S, Livingston B., MacIntyre K, Chalmers JW., Boyd J, Finlayson A, *et al.* Trends in case-fatality in 117718 patients admitted with acute myocardial infarction in Scotland. *Eur Heart J* 2000;21:1833–40.

17. Rosamond WD, Chambless LE, Folsom AR, Cooper LS, Conwill DE, Clegg L, *et al.* Trends in the incidence of myocardial infarction and in mortality due to coronary heart disease, 1987 to 1994. *N Engl J Med* 1998;339:861–7.

18. Abildstrom SZ, Rasmussen S, Madsen M. Significant decline in case fatality after acute myocardial infarction in Denmark--a population-based study from 1994 to 2001. *Scand Cardiovasc J SCJ* 2002;36:287–91.

19. Rogers WJ, Frederick PD, Stoehr E, Canto JG, Ornato JP, Gibson CM, *et al.* Trends in presenting characteristics and hospital mortality among patients with ST elevation and non-ST elevation myocardial infarction in the National Registry of Myocardial Infarction from 1990 to 2006. *Am Heart J* 2008;156:1026–34.

20. Hellermann JP, Reeder GS, Jacobsen SJ, Weston SA, Killian JM, Roger VL. Longitudinal trends in the severity of acute myocardial infarction: a population study in Olmsted County, Minnesota. *Am J Epidemiol* 2002;156:246–53.

21. Lampe FC, Morris RW, Walker M, Shaper AG, Whincup PH. Trends in rates of different forms of diagnosed coronary heart disease, 1978 to 2000: prospective, population based study of British men. *BMJ* 2005;330:1046.

22. Capewell S, Murphy NF, MacIntyre K, Frame S, Stewart S, Chalmers JW, *et al.* 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* 2006;92:1563–70.

23. Ohman EM, Bhatt DL, Steg PG, Goto S, Hirsch AT, Liau C-S, *et al.* The REduction of Atherothrombosis for Continued Health (REACH) Registry: an international, prospective, observational investigation in subjects at risk for atherothrombotic events-study design. *Am Heart J* 2006;151:786.e1–10.

24. Nedkoff L, Briffa TG, Knuiman M, Hung J, Norman PE, Hankey GJ, *et al.* Temporal trends in the incidence and recurrence of hospitalised atherothrombotic disease in an Australian population, 2000-07: data linkage study. *Heart* Published Online First: 21 July 2012.http://www.ncbi.nlm.nih.gov/pubmed/22821274 (accessed 20 Aug 2012).

25. Stokes J, Kannel WB, Wolf PA, Cupples LA, D'Agostino RB. The relative importance of selected risk factors for various manifestations of cardiovascular disease among men and women from 35 to 64 years old: 30 years of follow-up in the Framingham Study. *Circulation* 1987;75:V65–73.

26. Appelros P, Stegmayr B, Terént A. Sex differences in stroke epidemiology: a systematic review. *Stroke* 2009;40:1082–90.

27. Roquer J, Campello AR, Gomis M. Sex differences in first-ever acute stroke. *Stroke* 2003;34:1581–5.

28. Egorova N, Vouyouka AG, Quin J, Guillerme S, Moskowitz A, Marin M, *et al.* Analysis of gender-related differences in lower extremity peripheral arterial disease. *J Vasc Surg* 2010;51:372–8.e1; discussion 378–9.

29. Pilote L, Dasgupta K, Guru V, Humphries KH, McGrath J, Norris C, *et al.* A comprehensive view of sex-specific issues related to cardiovascular disease. *Can Med Assoc J* 2007;176:S1–44.

30. Kardys I, Vliegenthart R, Oudkerk M, Hofman A, Witteman JCM. The Female Advantage in Cardiovascular Disease: Do Vascular Beds Contribute Equally? *Am J Epidemiol* 2007;166:403–12.

31. Huxley R, Barzi F, Woodward M. Excess risk of fatal coronary heart disease associated with diabetes in men and women: meta-analysis of 37 prospective cohort studies. *BMJ* 2006;332:73–8.

32. Vaccarino V, Badimon L, Corti R, De Wit C, Dorobantu M, Hall A, *et al.* Ischaemic heart disease in women: Are there sex differences in pathophysiology and risk factors? *Cardiovasc Res* 2011;90:9–17.

33. Blauwet LA, Hayes SN, McManus D, Redberg RF, Walsh MN. Low rate of sex-specific result reporting in cardiovascular trials. *Mayo ClinProc* 2007;82:166–70.

34. Johnson SM, Karvonen CA, Phelps CL, Nader S, Sanborn BM. Assessment of analysis by gender in the Cochrane reviews as related to treatment of cardiovascular disease. *J Women's Heal* 2003;12:449–57.

35. Oertelt-Prigione S, Wiedmann S, Endres M, Nolte CH, Regitz-Zagrosek V, Heuschmann P. Stroke and myocardial infarction: a comparative systematic evaluation of gender-specific analysis, funding and authorship patterns in cardiovascular research. *Cerebrovasc Dis* 2011;31:373–81.

36. Lim SS, Vos T, Flaxman AD, Danaei G, Shibuya K, Adair-Rohani H, *et al.* A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* 2012;380:2224–60.

37. *Health Survey for England - 2011, Health, social care and lifestyles*. Leeds: : The Health and Social Care Information Centre 2012. http://www.hscic.gov.uk/catalogue/PUB09300

38. Kenfield SA, Wei EK, Rosner BA, Glynn RJ, Stampfer MJ, Colditz GA. Burden of smoking on cause-specific mortality: application to the Nurses' Health Study. *Tob Control* 2010;19:248–54.

39. Shinton R, Beevers G. Meta-analysis of relation between cigarette smoking and stroke. *BMJ* 1989;298:789–94.

40. Murray CJ, Richards MA, Newton JN, Fenton KA, Anderson HR, Atkinson C, *et al.* UK health performance: findings of the Global Burden of Disease Study 2010. *Lancet* 2013;381:997–1020.

41. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, *et al.* General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* 2008;117:743–53.

42. Lewington S, Clarke R, Qizilbash N, Peto R, Collins R. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *Lancet* 2002;360:1903–13.

43. Huxley R, Woodward M, Barzi F, Wong JW, Pan WH, Patel A. Does sex matter in the associations between classic risk factors and fatal coronary heart disease in populations from the Asia-Pacific region? *J Women's Heal* 2005;14:820–8.

44. Office for Life Sciences. Strategy for UK Life Science. London: 2011. http://www.bis.gov.uk/assets/biscore/innovation/docs/s/11-1429-strategy-for-uk-life-sciences

45. UK e-health records research capacity and capability. London: 2011. http://www.mrc.ac.uk/Utilities/Documentrecord/index.htm?d=MRC007896

46. UK Clinical Research Collaboration and the Wellcome Trust. Frontiers Meeting: Use of Electronic Patient Records for Research and Health Benefit. London: 2007. http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtd038686.pdf

47. Office for strategic co-ordination of health research. A Strategic Framework for Health Informatics in Support of Research. London: 2010.

48. Academy of Medical Sciences. Personal data for public good: Using health information in medical research. London: 2006.

49. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, *et al.* Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Ann Intern Med* 2010;153:600–6.

50. Holman CD, Bass AJ, Rouse IL, Hobbs MS. Population-based linkage of health records in Western Australia: development of a health services research linked database. *Aust N Z J Public Health* 1999;23:453–9.

51. Canadian Institute for Health Information. *The Difference Data Makes—Canadian Institute for Health Information Annual Report, 2010–2011*. Ottawa: 2011. https://secure.cihi.ca/free_products/annual_report_2010-2011_en.pdf (accessed 30 May 2012).

52. Mandl KD, Kohane IS. Tectonic shifts in the health information economy. *N Engl J Med* 2008;358:1732–7.

53. A Blueprint for Health Records Research in Scotland. Dundee: 2011. http://www.scot-ship.ac.uk/sites/default/files/Reports/SHIP_BLUEPRINT_DOCUMENT_consultation_draft_081211.pdf (accessed 4 Apr 2012).

54. Trutwein B, Holman CDJ, Rosman DL. Health Data Linkage Conserves Privacy in a Research-Rich Environment. *Ann Epidemiol* 2006;16:279–80.

55. Thomas R, Walport M. *Data Sharing Review*. London: : NHS Information Centre 2008. http://www.connectingforhealth.nhs.uk/systemsandservices/infogov/links/datasharingreview.pdf

56. Lowrance WW. *Learning from Experience: Privacy and the secondary use of data in health research*. London: : Nuffield Trust 2002. http://www.nuffieldtrust.org.uk/sites/files/nuffield/publication/learning-from-experience-nov02.pdf

57. Dokholyan RS, Muhlbaier LH, Falletta JM, Jacobs JP, Shahian D, Haan CK, *et al.* Regulatory and ethical considerations for linking clinical and administrative databases. *Am Heart J* 2009;157:971–82.

58. Kho ME, Duffett M, Willison DJ, Cook DJ, Brouwers MC. Written informed consent and selection bias in observational studies using medical records: systematic review. *BMJ* 2009;338:b866.

59. Tu J V, Willison DJ, Silver FL, Fang J, Richards J a, Laupacis A, *et al.* Impracticability of informed consent in the Registry of the Canadian Stroke Network. *N Engl J Med* 2004;350:1414–21.

60. MacLeod MC, Bray CA, Kendrick SW, Cobbe SM. Enhancing the power of record linkage involving low quality personal identifiers: use of the best link principle and cause of death prior likelihoods. *Comput Biomed Res* 1998;31:257–70.

61. Karmel R, Gibson D. Event-based record linkage in health and aged care services data: a methodological innovation. *BMC Health Serv Res* 2007;7:154.

62. Durham E, Xue Y, Kantarcioglu M, Malin B. Private medical record linkage with approximate matching. *AMIA Annu Symp Proc* 2010;2010:182–6.

63. Potz N, Powell D, Lamagni TL, Pebody R, Bridger D, Duckworth G. Probabilistic record linkage of infection records and death registrations: a tool strengthen surveillance. *Stat Commun Infect Dis* 2010;2:article 6.

64. Lyons RA, Jones KH, John G, Brooks CJ, Verplancke J-P, Ford D V, *et al.* The SAIL databank: linking multiple health and social care datasets. *BMC Med Inform Decis Mak* 2009;9:3.

65. Black N. Secondary use of personal data for health and health services research: why identifiable data are essential. *J Heal ServRes Policy* 2003;8 Suppl 1:S1–40.

66. Kaye J. From single biobanks to international networks: developing e-governance. *Hum Genet* 2011;130:377–82.

67. Gulliford MC, Charlton J, Ashworth M, Rudd AG, Toschke AM. Selection of medical diagnostic codes for analysis of electronic patient records. Application to stroke in a primary care database. *PLoS One* 2009;4:e7168.

68. Dave S, Petersen I. Creating medical and drug code lists to identify cases in primary care databases. *Pharmacoepidemiol Drug Saf* 2009;18:704–7.

69. Brooks CJ, Lyons RA, Jones KH, Hutton AJ, Walker R, Evans KA, *et al.* The prescribed duration algorithm: utilising "free text" from multiple primary care electronic systems. *Pharmacoepidemiol Drug Saf* 2010;19:983–9.

70. Atreja A, Achkar J-P, Jain AK, Harris CM, Lashner BA. Using technology to promote gastrointestinal outcomes research: a case for electronic health records. *Am J Gastroenterol* 2008;103:2171–8.

71. Yip YL. Unlocking the potential of electronic health records for translational research. Findings from the section on bioinformatics and translational informatics. *Yearb Med Inform* 2012;7:135–8.

72. James S, Fröbert O, Lagerqvist B. Cardiovascular registries: a novel platform for randomised clinical trials. *Heart* 2012;98:1329–31.

73. Dean BB, Lam J, Natoli JL, Butler Q, Aguilar D, Nordyke RJ. Review: use of electronic medical records for health outcomes research: a literature review. *Med Care Res Rev* 2009;66:611–38.

74. Developing STROBE guidelines for secondary data research. In: *Primary Care Databases Symposium*. 2012. http://www.idrn.org/events/upcoming/primarycaredatabases.php

75. Bohensky MA, Jolley D, Sundararajan V, Evans S, Ibrahim J, Brand C. Development and validation of reporting guidelines for studies involving data linkage. *Aust N Z J Public Health* 2011;35:486–9.

76. Walley T, Mantgani A. The UK General Practice Research Database. *Lancet* 1997;350:1097–9.

77. Gallagher AM, Puri S, Staa TV. Linkage of the General Practice Research Database (GPRD) with other data sources. *Pharmacoepidemiol Drug Saf* 2011;:S230–S364.

78. Chisholm J. The Read clinical classification. *BMJ* 1990;300:1092.

79. Benson T. The history of the Read Codes: the inaugural James Read Memorial Lecture 2011. *Inform Prim Care* 2011;19:173–82.

80. Wood L, Martinez C. The general practice research database: role in pharmacovigilance. *Drug Saf* 2004;27:871–81.

81. General Practitioners Committee British Medical Association. Investing in general practice. The new general medical services contract. London: 2003. http://www.nhsemployers.org/SiteCollectionDocuments/gms_contract_cd_130209.pdf

82. Department of Health. Quality and Outcomes Framework guidance for GMS contract 2011/12. UK: : The NHS Confederation (Employers) Company Ltd 2011. www.nhsemployers.org/SiteCollectionDocuments/QOFguidanceGMScontract_2011_12_FL 13042011.pdf

83. Steel N, Willems S. Research learning from the UK Quality and Outcomes Framework: a review of existing research. *Qual Prim Care* 2010;18:117–25.

84. Bhattarai N, Charlton J, Rudisill C, Gulliford MC. Coding, recording and incidence of different forms of coronary heart disease in primary care. *PLoS One* 2012;7:e29776.

85. McGovern MP, Boroujerdi MA, Taylor MW, Williams DJ, Hannaford PC, Lefevre KE, *et al.* The effect of the UK incentive-based contract on the management of patients with coronary heart disease in primary care. *Fam Pract* 2008;25:33–9.

86. Jeffries J, Fulton R. Making a population estimate in England and Wales. National Statistics Methodological Series No. 34. London: 2007.

87. Social Exclusion Task Force. Inclusion Health: Improving the way we meet the primary health care needs of the socially excluded. London: 2010. http://webarchive.nationalarchives.gov.uk/+/http://www.cabinetoffice.gov.uk/media/346571/inclusion-health.pdf

88. Wilkins D, Payne S, Granville G, Branney P. The Gender and Access to Health Services Study. London: : Department of Health 2009. http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_092042

89. Hunt K, Adamson J, Galdas P. Gender and Help-seeking: Towards Gender-comparative Studies. In: Annandale E, ed. *The Palgrave handbook of gender and healthcareA2 - Kuhlman,Ellen*. Basingstoke: : Palgrove Macmillan 2010.

90. Richards H, McConnachie A, Morrison C, Murray K, Watt G. Social and gender variation in the prevalence, presentation and general practitioner provisional diagnosis of chest pain. *J Epidemiol Community Health* 2000;54:714–8.

91. Hunt K, Adamson J, Hewitt C, Nazareth I. Do women consult more than men? A review of gender and consultation for back pain and headache. *J Health Serv Res Policy* 2011;16:108–17.

92. Hunt K, Ford G, Harkins L, Wyke S. Are women more ready to consult than men? Gender differences in family practitioner consultation for common chronic conditions. *J Heal Serv Res Policy* 1999;4:96–100.

93. Jick SS, Kaye JA, Vasilakis-Scaramozza C, Garcia Rodriguez LA, Ruigomez A, Meier CR, *et al.* Validity of the general practice research database. *Pharmacotherapy* 2003;23:686–9.

94. Langley TE, Szatkowski L, Gibson J, Huang Y, McNeill A, Coleman T, *et al.* Validation of The Health Improvement Network (THIN) primary care database for monitoring prescriptions for smoking cessation medications. *Pharmacoepidemiol Drug Saf* 2010;19:586–90.

95. Mabotuwana T, Warren J, Harrison J, Kenealy T. What can primary care prescribing data tell us about individual adherence to long-term medication?-comparison to pharmacy dispensing data. *Pharmacoepidemiol Drug Saf* 2009;18:956–64.

96. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. *BrJClinPharmacol* 2010;69:4–14.

97. Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. *Br J Gen Pract* 2010;60:e128–36.

98. Herrett E, Smeeth L, Walker L, Weston C. The Myocardial Ischaemia National Audit Project (MINAP). *Heart* 2010;96:1264–7.

99. Brophy S, Mannan S, John A, Cheung WI, Lyons R, Weston C, *et al.* Risk of further acute vascular events following an initial myocardial infarction or stroke. London: : Department for Transport 2006. http://www.dft.gov.uk/pgr/roadsafety/research/rsrr/theme6/

100. Lyons R, Williams R, Gravenor M, Brophy S, Weston C, Macey S, *et al.* Analysis of Risk Outcomes for Cardiac Conditions. London: : Department for Transport 2010. http://www.dft.gov.uk/pgr/roadsafety/research/rsrr/theme6/report107/

101. Herrett E, Shah AD, Boggon R, Denaxas S, Smeeth L, van Staa T, *et al.* Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. *BMJ* 2013;346:f2350.

102. WHO | International Classification of Diseases (ICD). http://www.who.int/classifications/icd/en/ (accessed 15 Feb 2012).

103. OPCS-4 Classification — NHS Connecting for Health. http://www.connectingforhealth.nhs.uk/systemsandservices/data/clinicalcoding/codingstandards/opcs4/index_html (accessed 15 Feb 2012).

104. Audit Commission. Introducing payment by results: Getting the balance right for the NHS and taxpayers. London: 2004. http://www.audit-commission.gov.uk/SiteCollectionDocuments/AuditCommissionReports/NationalStudies/PaymentByResults_report.pdf

105. Data quality. 2009.http://www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937&categoryID=97 (accessed 24 Feb 2012).

106. HES 2009-10 1 Inpatient Data Quality Note. London: 2010.

107. A decade in view. 2011.http://www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937&categoryID=451 (accessed 24 Feb 2012).

108. Suleman M, Clark MPA, Goldacre M, Burton M. Exploring the variation in paediatric tonsillectomy rates between English regions: a 5-year NHS and independent sector data analysis. *Clin Otolaryngol* 2010;35:111–7.

109. Mindell J, Klodawski E, Fitzpatrick J, Malhotra N, McKee M, Sanderson C. The impact of private-sector provision on equitable utilisation of coronary revascularisation in London. *Heart* 2008;94:1008–11.

110. Burns EM, Rigby E, Mamidanna R, Bottle A, Aylin P, Ziprin P, *et al.* Systematic review of discharge coding accuracy. *J Public Health (Bangkok)* 2011;:fdr054–.

111. Audit Commission. Improving data quality in the NHS: Annual report on the PbR assurance programme. London: 2010. http://www.audit-commission.gov.uk/SiteCollectionDocuments/Downloads/26082010pbrnhsdataqualityreport.pdf

112. Audit Commission. Improving coding, costing and commissioning. London: 2011. http://www.audit-commission.gov.uk/SiteCollectionDocuments/Downloads/pbrannualreport2011.pdf

113. Office for National Statistics. Mortality statistics : Metadata 2010. London: 2011.

114. Goldacre MJ. Cause-Specific Mortality - Understanding Uncertain Tips of the Disease Iceberg. *J Epidemiol Community Health* 1993;47:491–6.

115. Goldacre MJ, Roberts SE, Griffith M. Multiple-cause coding of death from myocardial infarction: population-based study of trends in death certificate data. *J Public Health Med* 2003;25:69–71.

116. Goldacre MJ, Roberts SE, Griffith M. Place, time and certified cause of death in people who die after hospital admission for myocardial infarction or stroke. *Eur J Public Health* 2004;14:338–42.

117. Mant J, Wilson S, Parry J, Bridge P, Wilson R, Murdoch W, *et al.* Clinicians didn't reliably distinguish between different causes of cardiac death using case histories. *J Clin Epidemiol* 2006;59:862–7.

118. Pajunen P, Koukkunen H, Ketonen M, Jerkkola T, Immonen-Raiha P, Karja-Koskenkari P, *et al.* The validity of the Finnish Hospital Discharge Register and Causes of Death Register data on coronary heart disease. *Eur J Cardiovasc Prev Rehabil* 2005;12:132–7.

119. Goraya TY, Jacobsen SJ, Belau PG, Weston SA, Kottke TE, Roger VL. Validation of death certificate diagnosis of out-of-hospital coronary heart disease deaths in Olmsted County, Minnesota. *Mayo Clin ProceedingsMayo Clin* 2000;75:681–7.

120. Fox CS, Evans JC, Larson MG, Lloyd-Jones DM, O'Donnell CJ, Sorlie PD, *et al.* A comparison of death certificate out-of-hospital coronary heart disease death with physician-adjudicated sudden cardiac death. *Am J Cardiol* 2005;95:5,9.

121. Iribarren C, Crow RS, Hannan PJ, Jacobs Jr. DR, Luepker R V. Validation of death certificate diagnosis of out-of-hospital sudden cardiac death. *Am J Cardiol* 1998;82:50–3.

122. Devis T, Rooney C. Death certification and the epidemiologist. *Heal StatQ* 1999;Spring:21–33.

123. Noble M, Mclennan D, Wilkinson K, Whitworth A. The English indices of deprivation 2007. *Communities* Published Online First: 2007.http://eprints.ioe.ac.uk/2461/ (accessed 15 Feb 2012).

124. Shaw M, Galobardes B, Lawlor D, Lynch J, Wheeler B, Davey Smith G. *The Handbook of inequality and socioeconomic position: Concepts and measures*. Bristol: : The Policy Press 2007.

125. Hemingway H, Shipley M, Macfarlane P, Marmot M. Impact of socioeconomic status on coronary mortality in people with symptoms, electrocardiographic abnormalities, both or neither: the original Whitehall study 25 year follow up. *J Epidemiol Community Health* 2000;54:510–6.

126. O'Flaherty M, Bishop J, Redpath A, McLaughlin T, Murphy D, Chalmers J, *et al.* Coronary heart disease mortality among young adults in Scotland in relation to social inequalities: time trend study. *BMJ* 2009;339:b2613.

127. Hippisley-Cox J, Coupland C, Robson J, Brindle P. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. *Br Med J* 2010;341:c6624.

128. NHS Number: Information for Staff. 2012.http://www.connectingforhealth.nhs.uk/systemsandservices/nhsnumber/staff

129. Rachel Boggon on behalf of GPRD. Personal communication. 2012.

130. Foreign Office sets out support for British nationals overseas. http://www.fco.gov.uk/en/news/latest-news/?view=News&id=659719382 (accessed 18 Mar 2012).

131. Cutrona SL, Toh S, Iyer A, Foy S, Daniel GW, Nair VP, *et al.* Validation of acute myocardial infarction in the Food and Drug Administration's Mini-Sentinel program. *Pharmacoepidemiol Drug Saf* Published Online First: 29 June 2012.http://www.ncbi.nlm.nih.gov/pubmed/22745038 (accessed 12 Sep 2012).

132. Saczynski JS, Andrade SE, Harrold LR, Tjia J, Cutrona SL, Dodd KS, *et al.* A systematic review of validated methods for identifying heart failure using administrative data. *Pharmacoepidemiol Drug Saf* 2012;21 Suppl 1:129–40.

133. Andrade SE, Harrold LR, Tjia J, Cutrona SL, Saczynski JS, Dodd KS, *et al.* A systematic review of validated methods for identifying cerebrovascular accident or transient ischemic attack using administrative data. *Pharmacoepidemiol Drug Saf* 2012;21 Suppl 1:100–28.

134. Tamariz L, Harkins T, Nair V. A systematic review of validated methods for identifying ventricular arrhythmias using administrative and claims data. *Pharmacoepidemiol Drug Saf* 2012;21 Suppl 1:148–53.

135. Lewis JD, Brensinger C. Agreement between GPRD smoking data: a survey of general practitioners and a population-based survey. *Pharmacoepidemiol Drug Saf* 2003;13:437–41.

136. MacDonald TM, Morant S V, Pettitt D. Prevalence of clinically recognised hypertension (HT) and dyslipidaemia (DL) in the UK. *Pharmacoepidemiol Drug Saf* 2003;12:S25–26.

137. Stone MA, Camosso-Stefinovic J, Wilkinson J, de Lusignan S, Hattersley AT, Khunti K. Incorrect and incomplete coding and classification of diabetes: a systematic review. *Diabet Med* 2010;27:491–7.

138. Whitelaw FG, Nevin SL, Milne RM, Taylor RJ, Taylor MW, Watt AH. Completeness and accuracy of morbidity and repeat prescribing records held on general practice computers in Scotland. *Br J Gen Pract* 1996;46:181–6.

139. Van Staa T, Abenheim L. The quality of information recorded on a UK database of primary care records: A study of hypoglycemia and other conditions. *Pharmacoepidemiol Drug Saf* 1994;3:15–21.

140. Hassey A, Gerrett D, Wilson A. A survey of validity and utility of electronic patient records in a general practice. *Br Med J* 2001;322:1401–5.

141. Samy AK, Whyte B, MacBain G. Abdominal aortic aneurysm in Scotland. *Br J Surg* 1994;81:1104–6.

142. Johal A, Mitchell D, Lees T, Cromwell D, van der Meulen J. Use of Hospital Episode Statistics to investigate abdominal aortic aneurysm surgery. *Br J Surg* 2012;99:66–72.

143. Heckbert SR, Kooperberg C, Safford MM, Psaty BM, Hsia J, McTiernan A, *et al.* Comparison of self-report, hospital discharge codes, and adjudication of cardiovascular events in the Women's Health Initiative. *Am J Epidemiol* 2004;160:1152–8.

144. Merry AHH, Boer JMA, Schouten LJ, Feskens EJM, Verschuren WMM, Gorgels APM, *et al.* Validity of coronary heart diseases and heart failure based on hospital discharge and mortality data in the Netherlands using the cardiovascular registry Maastricht cohort study. *Eur J Epidemiol* 2009;24:237–47.

145. Cannon PJ, Connell PA, Stockley IH, Garner ST, Hampton JR. Prevalence of angina as assessed by a survey of prescriptions for nitrates. *Lancet* 1988;1:979–81.

146. Kiyota Y, Schneeweiss S, Glynn RJ, Cannuscio CC, Avorn J, Solomon DH. Accuracy of Medicare claims-based diagnosis of acute myocardial infarction: estimating positive predictive value on the basis of review of hospital records. *Am Hear J* 2004;148:99–104.

147. Rosamond WD, Chambless LE, Sorlie PD, Bell EM, Weitzman S, Smith JC, *et al.* Trends in the sensitivity, positive predictive value, false-positive rate, and comparability ratio of hospital discharge diagnosis codes for acute myocardial infarction in four US communities, 1987-2000. *Am J Epidemiol* 2004;160:1137–46.

148. Hammad TA, McAdams MA, Feight A, Iyasu S, Dal Pan GJ. Determining the predictive value of Read/OXMIS codes to identify incident acute myocardial infarction in the General Practice Research Database. *Pharmacoepidemiol Drug Saf* 2008;17:1197–201.

149. Johansson S, Wallander M-A, Ruigómez A, García Rodríguez LA. Is there any association between myocardial infarction, gastro-oesophageal reflux disease and acid-suppressing drugs? *Aliment Pharmacol Ther* 2003;18:973–8.

150. Andersohn F, Suissa S, Garbe E. Use of first- and second-generation cyclooxygenase-2-selective nonsteroidal antiinflammatory drugs and risk of acute myocardial infarction. *Circulation* 2006;113:1950–7.

151. García Rodríguez LA, Varas-Lorenzo C, Maguire A, González-Pérez A. Nonsteroidal antiinflammatory drugs and the risk of myocardial infarction in the general population. *Circulation* 2004;109:3000–6.

152. Hall GC, Brown MM, Mo J, MacRae KD. Triptans in migraine: The risks of stroke, cardiovascular disease, and death in practice. *Neurology* 2004;62:563–8.

153. Meier CR, Jick SS, Derby LE, Vasilakis C, Jick H. Acute respiratory-tract infections and risk of first-time acute myocardial infarction. *Lancet* 1998;351:1467–71.

154. Gray J, Majeed A, Kerry S, Rowlands G. Identifying patients with ischaemic heart disease in general practice: cross sectional study of paper and computerised medical records. *Br Med J* 2000;321:548–50.

155. Johansson S, Wallander MA, Ruigómez A, García Rodríguez LA. Incidence of newly diagnosed heart failure in UK general practice. *Eur J Heart Fail* 2001;3:225–31.

156. Maru S, Koch GG, Stender M, Clark D, Gibowski L, Petri H, *et al.* Antidiabetic drugs and heart failure risk in patients with type 2 diabetes in the U.K. primary care setting. *Diabetes Care* 2005;28:20–6.

157. Van Staa T-P, Abenhaim L. The quality of information recorded on a UK database of primary care records: A study of hospitalizations due to hypoglycemia and other conditions. *Pharmacoepidemiol Drug Saf* 1994;3:15–21.

158. Hasan M, Meara RJ, Bhowmick BK. The quality of diagnostic coding in cerebrovascular disease. *Int J Qual Heal Care* 1995;7:407–10.

159. Davenport RJ, Dennis MS, Warlow CP. The accuracy of Scottish Morbidity Record (SMR1) data for identifying hospitalised stroke patients. *Health Bull (Raleigh)* 1996;54:402–5.

160. Hall GC, Brown MM, Mo J, MacRae KD. Triptans in migraine: the risks of stroke, cardiovascular disease, and death in practice. *Neurology* 2004;62:563–8.

161. De Bruin ML, van Hemel NM, Leufkens HGM, Hoes AW. Hospital discharge diagnoses of ventricular arrhythmias and cardiac arrest were useful for epidemiologic research. *J Clin Epidemiol* 2005;58:1325–9.

162. Huerta C, Lanes SF, García Rodríguez LA. Respiratory medications and the risk of cardiac arrhythmias. *Epidemiology* 2005;16:360–6.

163. De Abajo FJ, Rodríguez LA. Risk of ventricular arrhythmias associated with nonsedating antihistamine drugs. *Br J Clin Pharmacol* 1999;47:307–13.

164. Hoyle BL, Castillo F, Clark B, Perpich D, Wackerow J. Metadata for the Longitudinal Data Life Cycle. 2011. doi:http://dx.doi.org/10.3886/DDILongitudinal03

165. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J ChronicDis* 1987;40:373–83.

166. Khan NF, Perera R, Harper S, Rose PW. Adaptation and validation of the Charlson Index for Read/OXMIS coded databases. *BMC Fam Pract* 2010;11:1.

167. Denaxas SC, George J, Herrett E, Shah AD, Kalra D, Hingorani AD, *et al.* Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). *Int J Epidemiol* 2012;41:1625–38.

168. Group TW of SCPS. Computerised record linkage: Compared with traditional patient follow-up methods in clinical trials and illustrated in a prospective epidemiological study. *J Clin Epidemiol* 1995;48:1441–52.

169. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ* 2007;335:136.

170. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Brindle P. Performance of the QRISK cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice: a validation study. *Heart* 2008;94:34–9.

171. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, *et al.* Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* 2008;336:1475–82.

172. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Brindle P. Performance of the QRISK cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice: a validation study. *Heart* 2008;94:34–9.

173. Collins GS, Altman DG. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. *BMJ* 2010;340:c2442.

174. Simpson CR, Hippisley-Cox J, Sheikh A. Trends in the epidemiology of smoking recorded in UK general practice. *Br J Gen Pract* 2010;60:e121–7.

175. Bottomley A. Methodology for assessing the prevalence of angina in primary care using practice based information in northern England. *J Epidemiol Community Health* 1997;51:87–9.

176. Maitland-van der Zee AH, Klungel OH, Stricker BH, van der Kuip DA, Witteman JC, Hofman A, *et al.* Repeated nitrate prescriptions as a potential marker for angina pectoris. A comparison with medical information from the Rotterdam Study. *Pharm World Sci* 2003;25:70–2.

177. Alpert JS, Thygesen K, Jaffe A, White HD. The universal definition of myocardial infarction: a consensus document: ischaemic heart disease. *Heart* 2008;94:1335–41.

178. Rubin DB. Inference and Missing Data. *Biometrika* 1976;63:581–90.

179. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, *et al.* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Br Med J* 2009;339.ISI:000267678300003

180. Marston L, Carpenter JR, Walters KR, Morris RW, Nazareth I, Petersen I. Issues in multiple imputation of missing data for large general practice clinical databases. *Pharmacoepidemiol Drug Saf* 2010;19:618–26.

181. Capewell S, Livingston BM, MacIntyre K, Chalmers JW, Boyd J, Finlayson A, *et al.* Trends in case-fatality in 117 718 patients admitted with acute myocardial infarction in Scotland. *Eur Heart J* 2000;21:1833–40.

182. Bjorck L, Rosengren A, Bennett K, Lappas G, Capewell S. Modelling the decreasing coronary heart disease mortality in Sweden between 1986 and 2002. *EurHeart J* 2009;30:1046–56.

183. Dignam JJ, Kocherginsky MN. Choice and interpretation of statistical tests used when competing risks are present. *J Clin Oncol* 2008;26:4027–34.

184. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med* 2007;26:2389–430.

185. Pepe MS, Mori M. Kaplan-Meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Stat Med* 1993;12:737–51.

186. Lunn M, McNeil D. Applying Cox regression to competing risks. *Biometrics* 1995;51:524–32.

187. Bakoyannis G, Touloumi G. Practical methods for competing risks data: A review. *Stat Methods Med Res* 2011;:0962280210394479–.

188. Cox DR. Regression Models and Life-Tables. *J R Stat Soc Ser B* 1972;34:187–220.

189. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003;327:557–60.

190. Hertz-Picciotto I, Rockhill B. Validity and efficiency of approximation methods for tied survival times in Cox regression. *Biometrics* 1997;53:1151–6.

191. Therneau T, Grambsch P. *Modelling survival data: Extending the Cox Model*. 2nd ed. New York: : Springer 2001.

192. Office for National Statistics. Population Estimates for UK, England and Wales, Scotland and Northern Ireland, mid 2005. 2010.http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm:77-213624 (accessed 24 Nov 2012).

193. Information Centre for Health and Social Care. Health Survey for England 2005 Latest Trends. 2006.http://www.ic.nhs.uk/statistics-and-data-collections/health-and-lifestyles-related-surveys/health-survey-for-england/health-survey-for-england-2005-latest-trends (accessed 24 Nov 2012).

194. World Health Organisation. The global burden of disease: 2004 update. Geneva: 2004.

http://www.who.int/healthinfo/global_burden_disease/2004_report_update/en/index.html

195. Haan MN, Selby J V, Rice DP, Quesenberry CP, Schofield KA, Liu J, *et al.* Trends in cardiovascular disease incidence and survival in the elderly. *Ann Epidemiol* 1996;6:348–56.

196. Arnold AM, Psaty BM, Kuller LH, Burke GL, Manolio TA, Fried LP, *et al.* Incidence of cardiovascular disease in older Americans: the cardiovascular health study. *J Am Geriatr Soc* 2005;53:211–8.

197. Ishikawa S, Kayaba K, Gotoh T, Nago N, Nakamura Y, Tsutsumi A, *et al.* Incidence of total stroke, stroke subtypes, and myocardial infarction in the Japanese population: the JMS Cohort Study. *J Epidemiol* 2008;18:144–50.

198. Scott RAP, Bridgewater SG, Ashton HA. Randomized clinical trial of screening for abdominal aortic aneurysm in women. *Br J Surg* 2002;89:283–5.

199. Lederle FA, Johnson GR, Wilson SE, Chute EP, Littooy FN, Bandyk D, *et al.* Prevalence and associations of abdominal aortic aneurysm detected through screening. Aneurysm Detection and Management (ADAM) Veterans Affairs Cooperative Study Group. *Ann Intern Med* 1997;126:441–9.

200. Goyal A, Norton CR, Thomas TN, Davis RL, Butler J, Ashok V, *et al.* Predictors of incident heart failure in a large insured population: A one million person-year follow-up study. *Circ Hear Fail* 2010;3:698–705.

201. Johansson S, Wallander M a, Ruigómez A, García Rodríguez L a. Incidence of newly diagnosed heart failure in UK general practice. *Eur J Heart Fail* 2001;3:225–31.

202. Chugh SS, Jui J, Gunson K, Stecker EC, John BT, Thompson B, *et al.* Current burden of sudden cardiac death: multiple source surveillance versus retrospective death certificate-based review in a large U.S. community. *J Am Coll Cardiol* 2004;44:1268–75.

203. Byrne R, Constant O, Smyth Y, Callagy G, Nash P, Daly K, *et al.* Multiple source surveillance incidence and aetiology of out-of-hospital sudden cardiac death in a rural population in the West of Ireland. *Eur Heart J* 2008;29:1418–23.

204. Higgins JP, Higgins JA. Epidemiology of peripheral arterial disease in women. *J Epidemiol* 2003;13:1–14.

205. Anand SS, Islam S, Rosengren A, Franzosi MG, Steyn K, Yusufali AH, *et al.* Risk factors for myocardial infarction in women and men: insights from the INTERHEART study. *Eur Heart J* 2008;29:932–40.

206. Tunstall-Pedoe H, Kuulasmaa K, Amouyel P, Arveiler D, Rajakangas AM, Pajak A. Myocardial infarction and coronary deaths in the World Health Organization MONICA Project. Registration procedures, event rates, and case-fatality rates in 38 populations from 21 countries in four continents. *Circulation* 1994;90:583–612.

207. Jousilahti P, Vartiainen E, Tuomilehto J, Puska P. Sex, age, cardiovascular risk factors, and coronary heart disease: a prospective follow-up study of 14,786 middle-aged men and women in Finland. *Circulation* 1999;99:1165–72.

208. Tunstall-Pedoe H. Myth and paradox of coronary risk and the menopause. *Lancet* 1998;351:1425–7.

209. Goyal A, Norton CR, Thomas TN, Davis RL, Butler J, Ashok V, *et al.* Predictors of incident heart failure in a large insured population: a one million person-year follow-up study. *Circ Hear Fail* 2010;3:698–705.

210. Hemingway H, McCallum A, Shipley M, Manderbacka K, Martikainen P, Keskimäki I. Incidence and prognostic implications of stable angina pectoris among women and men. *JAMA* 2006;295:1404–11.

211. Filipovic M, Seagroatt V, Goldacre MJ. Differences between women and men in surgical treatment and case fatality rates for ruptured aortic abdominal aneurysm in England. *Br J Surg* 2007;94:1096–9.

212. Chugh SS, Uy-Evanado A, Teodorescu C, Reinier K, Mariani R, Gunson K, *et al.* Women have a lower prevalence of structural heart disease as a precursor to sudden cardiac arrest: The Ore-SUDS (Oregon Sudden Unexpected Death Study). *J Am Coll Cardiol* 2009;54:2006–11.

213. Lerner DJ, Kannel WB. Patterns of coronary heart disease morbidity and mortality in the sexes: a 26-year follow-up of the Framingham population. *Am Heart J* 1986;111:383–90.

214. Lee WL, Cheung AM, Cape D, Zinman B. Impact of diabetes on coronary artery disease in women and men: a meta-analysis of prospective studies. *Diabetes Care* 2000;23:962–8.

215. Kanaya AM, Grady D, Barrett-Connor E. Explaining the sex difference in coronary heart disease mortality among patients with type 2 diabetes mellitus: a meta-analysis. *Arch Intern Med* 2002;162:1737–45.

216. Shaw LJ, Bairey Merz CN, Pepine CJ, Reis SE, Bittner V, Kelsey SF, *et al.* Insights from the NHLBI-Sponsored Women's Ischemia Syndrome Evaluation (WISE) Study: Part I: gender differences in traditional and novel risk factors, symptom evaluation, and gender-optimized diagnostic strategies. *J Am Coll Cardiol* 2006;47:S4–S20.

217. Reeves MJ, Bushnell CD, Howard G, Gargano JW, Duncan PW, Lynch G, *et al.* Sex differences in stroke: epidemiology, clinical presentation, medical care, and outcomes. *Lancet Neurol* 2008;7:915–26.

218. Kucharska-Newton AM, Couper DJ, Pankow JS, Prineas RJ, Rea TD, Sotoodehnia N, *et al.* Diabetes and the risk of sudden cardiac death, the Atherosclerosis Risk in Communities Study. *Acta Diabetol* 2010;47:S161–S168.

219. Scarborough P, Bhatnagar P, Wickramasinghe K, Smolina K, Mitchell C. Coronary heart disease statistics: 2010 edition. London: 2010.

220. O'Flaherty M, Ford E, Allender S, Scarborough P, Capewell S. Coronary heart disease trends in England and Wales from 1984 to 2004: concealed levelling of mortality rates among young adults. *Heart* 2008;94:178–81.

221. Baena-Díez JM, Vidal-Solsona M, Byram AO, González-Casafont I, Ledesma-Ulloa G, Martí-Sans N. The epidemiology of cardiovascular disease in primary care. the Zona Franca Cohort study in Barcelona, Spain. *Rev Esp Cardiol* 2010;63:1261–9.

222. Canoui-Poitrine F, Luc G, Juhan-Vague I, Morange P-E, Arveiler D, Ferrieres J, *et al.* Respective contribution of conventional risk factors and antihypertensive treatment to stable angina pectoris and acute coronary syndrome as the first presentation of coronary heart disease: the PRIME Study. *Eur J Cardiovasc Prev Rehabil* 2009;16:550–5.

223. Glynn RJ, Rosner B. Comparison of risk factors for the competing risks of coronary heart disease, stroke, and venous thromboembolism. *Am J Epidemiol* 2005;162:975–82.

224. Ducimetière P, Ruidavets JB, Montaye M, Haas B, Yarnell J. Five-year incidence of angina pectoris and other forms of coronary heart disease in healthy men aged 50-59 in France and Northern Ireland: the Prospective Epidemiological Study of Myocardial Infarction (PRIME) Study. *Int J Epidemiol* 2001;30:1057–62.

225. Dagenais GR, Robitaille NM, Lupien PJ, Christen A, Gingras S, Moorjani S, *et al.* First coronary heart disease event rates in relation to major risk factors: Quebec cardiovascular study. *Can J Cardiol* 1990;6:274–80.

226. Silventoinen K, Magnusson PKE, Neovius M, Sundström J, Batty GD, Tynelius P, *et al.* Does obesity modify the effect of blood pressure on the risk of cardiovascular disease? A population-based cohort study of more than one million Swedish men. *Circulation* 2008;118:1637–42.

227. Rothwell PM, Coull AJ, Silver LE, Fairhead JF, Giles MF, Lovelock CE, *et al.* Population-based study of event-rate, incidence, case fatality, and mortality for all acute vascular events in all arterial territories (Oxford Vascular Study). *Lancet* 2005;366:1773–83.

228. Kanamasa K, Ishikawa K, Hayashi T, Hoshida S, Yamada Y, Kawarabayashi T, *et al.* Increased cardiac mortality in women compared with men in patients with acute myocardial infarction. *Intern Med* 2004;43:911–8.

229. Jha AK, Orav EJ, Li Z, Epstein AM. The inverse relationship between mortality rates and performance in the Hospital Quality Alliance measures. *Health Aff* 2007;26:1104–10.

230. Centre TH and SCI. Hospital Episodes Statistics (HES). 2011.http://www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937

231. Denaxas S, George J, Herrett E, Shah A, Kalra D, Hingorani A, *et al.* Data Resource Profile: Cardiovascular disease research using Linked Bespoke studies and Electronic Records (CALIBER). *IntJ Epidemiol* 2012.

232. World Health Organisation. International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10). Chapter V: Mental and Behavioural Disorders. Clinical descriptions and diagnostic guidelines. http://apps.who.int/classifications/apps/icd/icd10online/. 2006.http://apps.who.int/classifications/apps/icd/icd10online/

233. Breslow NE, Day NE. *Statistical Methods in Cancer Research, Volume II: The Design and Analysis of Cohort Studies*. Lyon: : International Agency for Research on Cancer, WHO 1987.

234. Maas AHEM, van der Schouw YT, Regitz-Zagrosek V, Swahn E, Appelman YE, Pasterkamp G, *et al.* Red alert for women's heart: the urgent need for more research and knowledge on cardiovascular disease in women: proceedings of the workshop held in Brussels on gender differences in cardiovascular disease, 29 September 2010. *Eur Heart J* 2011;32:1362–8.

235. Aggarwal S, Qamar A, Sharma V, Sharma A. Abdominal aortic aneurysm: A comprehensive review. *Exp Clin Cardiol* 2011;16:11–5.

236. Sweeting MJ, Thompson SG, Brown LC, Powell JT. Meta-analysis of individual patient data to examine factors affecting growth and rupture of small abdominal aortic aneurysms. *Br J Surg* 2012;99:655–65.

237. Jónsdóttir LS, Sigfússon N, Gudnason V, Sigvaldason H, Thorgeirsson G. Do lipids, blood pressure, diabetes, and smoking confer equal risk of myocardial infarction in women as in men? The Reykjavik Study. *J Cardiovasc Risk* 2002;9:67–76.

238. Regitz-Zagrosek V, Brokat S, Tschope C. Role of gender in heart failure with normal left ventricular ejection fraction. *Prog Cardiovasc Dis*;49:241–51.

239. Almdal T, Scharling H, Jensen JS, Vestergaard H. The independent effect of type 2 diabetes mellitus on ischemic heart disease, stroke, and death: a population-based study of 13,000 men and women with 20 years of follow-up. *Arch Intern Med* 2004;164:1422–6.

240. Hart CL, Hole DJ, Smith GD. Risk factors and 20-year stroke mortality in men and women in the Renfrew/Paisley study in Scotland. *Stroke* 1999;30:1999–2007.

241. Lederle FA, Larson JC, Margolis KL, Allison MA, Freiberg MS, Cochrane BB, *et al.* Abdominal aortic aneurysm events in the women's health initiative: cohort study. *BMJ* 2008;337:a1724.

242. Kent KC, Zwolak RM, Egorova NN, Riles TS, Manganaro A, Moskowitz AJ, *et al.* Analysis of risk factors for abdominal aortic aneurysm in a cohort of more than 3 million individuals. *J Vasc Surg* 2010;52:539–48.

243. Cosford PA, Leng GC. Screening for abdominal aortic aneurysm. *Cochrane Database Syst Rev* 2007;:CD002945.

244. Sigvant B., Wiberg-Hedman K., Bergqvist D., Rolandsson O., Wahlberg E. Risk factor profiles and use of cardiovascular drug prevention in women and men with peripheral arterial disease. *Eur J Cardiovasc Prev Rehabil* 2009;16:39–46.

245. Smolina K, Wright FL, Rayner M, Goldacre MJ. Determinants of the decline in mortality from acute myocardial infarction in England between 2002 and 2010: linked national database study. *BMJ* 2012;344:d8059.

246. Towfighi A, Markovic D, Ovbiagele B. National gender-specific trends in myocardial infarction hospitalization rates among patients aged 35 to 64 years. *Am J Cardiol* 2011;108:1102–7.

247. Koek H, de Bruin A, Gast A, Gevers E, Kardaun J, Reitsma J, *et al.* Decline in incidence of hospitalisation for acute myocardial infarction in the Netherlands from 1995 to 2000. *Heart* 2006;92:162–5.

248. Yeung DF, Boom NK, Guo H, Lee DS, Schultz SE, Tu J V. Trends in the incidence and outcomes of heart failure in Ontario, Canada: 1997 to 2007. *Can Med Assoc J* Published Online First: 20 August 2012. doi:10.1503/cmaj.111958

249. Towfighi A, Markovic D, Ovbiagele B. Recent patterns of sex-specific midlife stroke hospitalization rates in the United States. *Stroke* 2011;42:3029–33.

250. Choke E, Vijaynagar B, Thompson J, Nasim A, Bown MJ, Sayers RD. Changing epidemiology of abdominal aortic aneurysms in England and Wales: older and more benign? *Circulation* 2012;125:1617–25.

251. Anjum A, Powell JT. Is the incidence of abdominal aortic aneurysm declining in the 21st century? Mortality and hospital admissions for England & Wales and Scotland. *Eur J Vasc Endovasc Surg* 2012;43:161–6.

252. Inglis SC, Lewsey JD, Chandler D, Byrne DS, Lowe GDO, MacIntyre K. Sex-specific time trends in first admission to hospital for peripheral artery disease in Scotland 1991-2007. *Br J Surg* 2012;99:680–7.

253. Mant J, Painter R, Vessey M. Risk of myocardial infarction, angina and stroke in users of oral contraceptives: an updated analysis of a cohort study. *Br J Obstet Gynaecol* 1998;105:890–6.

254. Marjoribanks J, Farquhar C, Roberts H, Lethaby A. *Long term hormone therapy for perimenopausal and postmenopausal women (Review)*. London: : John Wiley & Sons, Ltd. 2012. http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD004143.pub4/pdf/standard

255. Owen-Smith V, Hannaford PC, Elliott AM. Increased mortality among women with Rose angina who have not presented with ischaemic heart disease. *Br J Gen Pract* 2003;53:784–9.

256. Croft PR, Thomas E. Chest pain and subsequent consultation for coronary heart disease: a prospective cohort study. *Br J Gen Pr* 2007;57:40–4.

257. Healy B. The Yentl syndrome. *N Engl J Med* 1991;325:274–6.

258. Merz CN. The Yentl syndrome is alive and well. *Eur Heart J* 2011;32:1313–5.

259. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;97:1837–47.

260. Price JF, Mowbray PI, Lee AJ, Rumley A, Lowe GD, Fowkes FG. Relationship between smoking and cardiovascular risk factors in the development of peripheral arterial disease and coronary artery disease: Edinburgh Artery Study. *Eur Heart J* 1999;20:344–53.

261. Chugh SS, Reinier K, Teodorescu C, Evanado A, Kehr E, Al S, *et al.* Epidemiology of sudden cardiac death: clinical and research implications. *Prog Cardiovasc Dis* 2008;51:8,28.

262. Thorgeirsson G, Thorgeirsson G, Sigvaldason H, Witteman J. Risk factors for out-of-hospital cardiac arrest: the Reykjavik Study. *Eur Heart J* 2005;26:1499–505.

263. Escobedo LG, Caspersen CJ. Risk factors for sudden coronary death in the United States. *Epidemiology* 1997;8:175–80.

264. Jouven X, Desnos M, Guerot C, Ducimetière P. Predicting sudden death in the population: the Paris Prospective Study I. *Circulation* 1999;99:1978–83.

265. Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, *et al.* A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* 2008;452:638–42.

266. Ambrose JA, Barua RS. The pathophysiology of cigarette smoking and cardiovascular disease: an update. *J Am Coll Cardiol* 2004;43:1731–7.

267. Salahuddin S, Prabhakaran D, Roy A. Pathophysiological mechanisms of tobacco-related CVD. *Glob Heart* 2012;7:113–20.

268. Woodward M, Lam TH, Barzi F, Patel A, Gu D, Rodgers A, *et al.* Smoking, quitting, and the risk of cardiovascular disease among women and men in the Asia-Pacific region. *Int J Epidemiol* 2005;34:1036–45.

269. Pirie K, Peto R, Reeves GK, Green J, Beral V. The 21st century hazards of smoking and benefits of stopping: a prospective study of one million women in the UK. *Lancet* 2012;null. doi:10.1016/S0140-6736(12)61720-6

270. Lawlor DA, Song Y-M, Sung J, Ebrahim S, Smith GD. The association of smoking and cardiovascular disease in a population with low cholesterol levels: a study of 648,346 men from the Korean national health system prospective cohort study. *Stroke* 2008;39:760–7.

271. Kondo T, Osugi S, Shimokata K, Honjo H, Morita Y, Maeda K, *et al.* Smoking and smoking cessation in relation to all-cause mortality and cardiovascular events in 25,464 healthy male Japanese workers. *Circ J* 2011;75:2885–92.

272. Woodward M, Lam TH, Barzi F, Patel A, Gu D, Rodgers A, *et al.* Smoking, quitting, and the risk of cardiovascular disease among women and men in the Asia-Pacific region. *Int J Epidemiol* 2005;34:1036–45.

273. Delaney JAC, Moodie EEM, Suissa S. Validating the effects of drug treatment on blood pressure in the General Practice Research Database. *Pharmacoepidemiol Drug Saf* 2008;17:535–45.

274. Glynn RJ, Rosner B. Methods to evaluate risks for composite end points and their individual components. *J Clin Epidemiol* 2004;57:113–22.

275. Björck L, Rosengren A, Wallentin L, Stenestrand U. Smoking in relation to ST-segment elevation acute myocardial infarction: findings from the Register of

Information and Knowledge about Swedish Heart Intensive Care Admissions. *Heart* 2009;95:1006–11.

276. Kennon S, Suliman A, MacCallum PK, Ranjadayalan K, Wilkinson P, Timmis AD. Clinical characteristics determining the mode of presentation in patients with acute coronary syndromes. *J Am Coll Cardiol* 1998;32:2018–22.

277. Dunder K, Lind L, Lagerqvist B, Zethelius B, Vessby B, Lithell H. Cardiovascular risk factors for stable angina pectoris versus unheralded myocardial infarction. *Am Heart J* 2004;147:502–8.

278. Go AS, Iribarren C, Chandra M, Lathon P V, Fortmann SP, Quertermous T, *et al.* Statin and beta-blocker therapy and the initial presentation of coronary heart disease. *Ann Intern Med* 2006;144:229–38.

279. Gaspardone A, Crea F, Perino M, Iamele M, Tomai F, Versaci F, *et al.* Risk factors in patients with different clinical and angiographic manifestations of ischemic heart disease. *Cardiologia* 1995;40:679–84.

280. Sagastagoitia JD, Sáez Y, Vacas M, Narváez I, de Lafuente JPS, Molinero E, *et al.* Acute versus chronic myocardial ischemia: a differential biological profile study. *Pathophysiol Haemost Thromb* 2008;36:91–7.

281. Nakamura K, Huxley R, Ansary-Moghaddam A, Woodward M. The hazards and benefits associated with smoking and smoking cessation in Asia: a meta-analysis of prospective studies. *Tob Control* 2009;18:345–53.

282. Godtfredsen NS, Osler M, Vestbo J, Andersen I, Prescott E. Smoking reduction, smoking cessation, and incidence of fatal and non-fatal myocardial infarction in Denmark 1976-1998: a pooled cohort study. *J Epidemiol Community Heal* 2003;57:412–6.

283. Lewington S, Clarke R, Qizilbash N, Peto R, Collins R. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *Lancet* 2002;360:1903–13.

284. Lawes CMM, Bennett DA, Parag V, Woodward M, Whitlock G, Lam TH, *et al.* Blood pressure indices and cardiovascular disease in the Asia Pacific region: a pooled analysis. *Hypertension* 2003;42:69–75.

285. Sauvaget C, Ramadas K, Thomas G, Thara S, Sankaranarayanan R. Prognosis criteria of casual systolic and diastolic blood pressure values in a prospective study in India. *J Epidemiol Community Heal* 2010;64:366–72.

286. Gu D, Chen J, Wu X, Duan X, Jones DW, Huang J, *et al.* Prehypertension and risk of cardiovascular disease in Chinese adults. *J Hypertens* 2009;27:721–9.

287. Antikainen R, Jousilahti P, Tuomilehto J. Systolic blood pressure, isolated systolic hypertension and risk of coronary heart disease, strokes, cardiovascular disease and all-cause mortality in the middle-aged population. *J Hypertens* 1998;16:577–83.

288. Franklin SS, Wong ND, Kannel WB. Age-specific relevance of usual blood pressure to vascular mortality. *Lancet* 2003;361:1389; author reply 1391–2.

289. Franklin SS, Lopez VA, Wong ND, Mitchell GF, Larson MG, Vasan RS, *et al.* Single versus combined blood pressure components and risk for cardiovascular disease: the Framingham Heart Study. *Circulation* 2009;119:243–50.

290. Farnett L, Mulrow CD, Linn WD, Lucey CR, Tuley MR. The J-curve phenomenon and the treatment of hypertension. Is there a point beyond which pressure reduction is dangerous? *JAMA* 1991;265:489–95.

291. Wills AK, Lawlor DA, Matthews FE, Sayer AA, Bakra E, Ben-Shlomo Y, *et al.* Life course trajectories of systolic blood pressure using longitudinal data from eight UK cohorts. *PLoS Med* 2011;8:e1000440.

292. Singh GM, Danaei G, Pelizzari PM, Lin JK, Cowan MJ, Stevens GA, *et al.* The Age Associations of Blood Pressure, Cholesterol and Glucose: Analysis of Health Examination Surveys from International Populations. *Circulation* Published Online First: 4 April 2012. doi:10.1161/CIRCULATIONAHA.111.058834

293. Asia Pacific Cohort Studies Collaboration. The impact of cardiovascular risk factors on the age-related excess risk of coronary heart disease. *Int J Epidemiol* 2006;35:1025–33.

294. Franklin SS, Larson MG, Khan SA, Wong ND, Leip EP, Kannel WB, *et al.* Does the relation of blood pressure to coronary heart disease risk change with aging? The Framingham Heart Study. *Circulation* 2001;103:1245–9.

295. Franco OH, Peeters A, Bonneux L, de Laet C. Blood pressure in adulthood and life expectancy with cardiovascular disease in men and women: life course analysis. *Hypertension* 2005;46:280–6.

296. Turnbull F, Woodward M, Neal B, Barzi F, Ninomiya T, Chalmers J, *et al.* Do men and women respond differently to blood pressure-lowering treatment? Results of prospectively designed overviews of randomized trials. *Eur Heart J* 2008;29:2669–80.

297. Jónsdóttir LS, Sigfússon N, Gudnason V, Sigvaldason H, Thorgeirsson G. Do lipids, blood pressure, diabetes, and smoking confer equal risk of myocardial infarction in women as in men? The Reykjavik Study. *J Cardiovasc Risk* 2002;9:67–76.

298. Yusuf S, Hawken S, Ounpuu S, Dans T, Avezum A, Lanas F, *et al.* Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *Lancet* 2004;364:937–52.

299. Conen D, Ridker PM, Buring JE, Glynn RJ. Risk of cardiovascular events among women with high normal blood pressure or blood pressure progression: prospective cohort study. *BMJ* 2007;335:432.

300. Mattace-Raso FUS, van der Cammen TJM, van Popele NM, van der Kuip DAM, Schalekamp MADH, Hofman A, *et al.* Blood pressure components and cardiovascular events in older adults: the Rotterdam study. *J Am Geriatr Soc* 2004;52:1538–42.

301. Vaccarino V, Berkman LF, Krumholz HM. Long-term outcome of myocardial infarction in women and men: a population perspective. *Am J Epidemiol* 2000;152:965–73.

302. Vaccarino V, Holford TR, Krumholz HM. Pulse pressure and risk for myocardial infarction and heart failure in the elderly. *J Am Coll Cardiol* 2000;36:130–8.

303. Lida M, Ueda K, Okayama A, Kodama K, Sawai K, Shibata S, *et al.* Impact of elevated blood pressure on mortality from all causes, cardiovascular diseases, heart disease and stroke among Japanese: 14 year follow-up of randomly selected population from Japanese -- Nippon data 80. *J Hum Hypertens* 2003;17:851–7.

304. O'Donnell CJ, Ridker PM, Glynn RJ, Berger K, Ajani U, Manson JE, *et al.* Hypertension and borderline isolated systolic hypertension increase risks of cardiovascular disease and mortality in male physicians. *Circulation* 1997;95:1132–7.

305. Artac M, Dalton ARH, Majeed A, Huckvale K, Car J, Graley C, *et al.* Assessment of cardiovascular risk factors prior to NHS Health Checks in an urban setting: cross-sectional study. *J R Soc Med* 2012;3:17.

306. Greenland S, Schlesselman JJ, Criqui MH. The fallacy of employing standardized regression coefficients and correlations as measures of effect. *Am J Epidemiol* 1986;123:203–8.

307. Schocken DD, Benjamin EJ, Fonarow GC, Krumholz HM, Levy D, Mensah GA, *et al.* Prevention of heart failure: a scientific statement from the American Heart Association Councils on Epidemiology and Prevention, Clinical Cardiology, Cardiovascular Nursing, and High Blood Pressure Research; Quality of Care and Outcomes Research Interdisc. *Circulation* 2008;117:2544–65.

308. Eagle KA, Hirsch AT, Califf RM, Alberts MJ, Steg PG, Cannon CP, *et al.* Cardiovascular ischemic event rates in outpatients with symptomatic atherothrombosis or risk factors in the united states: insights from the REACH Registry. *Crit Pathw Cardiol* 2009;8:91–7.

309. Albert CM, Chae CU, Grodstein F, Rose LM, Rexrode KM, Ruskin JN, *et al.* Prospective study of sudden cardiac death among women in the United States. *Circulation* 2003;107:2096–101.

310. Kannel WB, Wilson PW, D'Agostino RB, Cobb J. Sudden coronary death in women. *Am Heart J* 1998;136:205–12.

311. Lloyd-Jones DM, Leip EP, Larson MG, Vasan RS, Levy D. Novel approach to examining first cardiovascular events after hypertension onset. *Hypertension* 2005;45:39–45.

312. Greenland S, Maclure M, Schlesselman JJ, Poole C, Morgenstern H. Standardized regression coefficients: a further critique and review of some alternatives. *Epidemiology* 1991;2:387–92.

313. Newman TB, Browner WS. In defense of standardized regression coefficients. *Epidemiology* 1991;2:383–6.

314. Clarke R, Shipley M, Lewington S, Youngman L, Collins R, Marmot M, *et al.* Underestimation of risk associations due to regression dilution in long-term follow-up of prospective studies. *Am J Epidemiol* 1999;150:341–53.

315. Messerli FH, Mancia G, Conti CR, Hewkin AC, Kupfer S, Champion A. Dogma Disputed : Can Aggressively Lowering Blood Pressure in Hypertensive Patients with Coronary Artery Disease Be Dangerous ? *Ann Intern Med* 2006.

316. Law MR, Morris JK, Wald NJ. Use of blood pressure lowering drugs in the prevention of cardiovascular disease: meta-analysis of 147 randomised trials in the context of expectations from prospective epidemiological studies. *BMJ* 2009;338:b1665.

317. Gandini S, Botteri E, Iodice S, Boniol M, Lowenfels AB, Maisonneuve P, *et al.* Tobacco smoking and cancer: a meta-analysis. *Int J Cancer* 2008;122:155–64.

318. Intercollegiate Stroke Working Party. National Sentinel Stroke Clinical Audit 2010 Round 7: Public Report for England, Wales and Northern Ireland. London: 2012. http://www.rcplondon.ac.uk/sites/default/files/national-sentinel-stroke-audit-2010-public-report-and-appendices_0.pdf

319. NICOR. National Heart Failure Audit: April 2010-March 2011. London: 2012. http://www.hqip.org.uk/assets/NCAPOP-Library/Heart-Failure-Audit-Report-NICOR-2010-2011.pdf

320. Nolan J, Gallagher E, Lloyd-Scott L, Rowan K. National Cardiac Arrest Audit (NCAA). *J Intensive Care Soc* 2009;10:313–5.

321. Pearce N. A Short Introduction to Epidemiology. Wellington, NZ: 2005. www.publichealth.ac.nz

322. Weed DL, Trock B. Interactions and public health decisions. *J Clin Epidemiol* 1988;41:207–9.

323. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Lippincott Williams & Wilkins 2008.

324. Siemiatycki J, Thomas DC. Biological models and statistical interactions: an example from multistage carcinogenesis. *Int J Epidemiol* 1981;10:383–7.

325. Weinberg CR. Interaction and exposure modification: are we asking the right questions? *Am J Epidemiol* 2012;175:602–5.

326. Walter SD, Holford TR. Additive, multiplicative, and other models for disease risks. *Am J Epidemiol* 1978;108:341–6.

327. Pearce N. Analytical implications of epidemiological concepts of interaction. *Int J Epidemiol* 1989;18:976–80.

328. Spiegelhalter D. Using speed of ageing and "microlives" to communicate the effects of lifetime habits and environment. *BMJ* 2012;345:e8223.

329. Rothman KJ, Greenland S, Walker AM. Concepts of interaction. *Am J Epidemiol* 1980;112:467–70.

330. Easton DF, Peto J, Babiker AG. Floating absolute risk: an alternative to relative risk in survival and case-control analysis avoiding an arbitrary reference group. *Stat Med* 1991;10:1025–35.

331. Austin PC. Absolute risk reductions and numbers needed to treat can be obtained from adjusted survival models for time-to-event outcomes. *J Clin Epidemiol* 2010;63:46–55.

332. Gerds TA, Scheike TH, Andersen PK. Absolute risk regression for competing risks: interpretation, link functions, and prediction. *Stat Med* 2012;31:3921–30.

333. Benichou J, Gail MH. Estimates of absolute cause-specific risk in cohort studies. *Biometrics* 1990;46:813–26.

334. Rapsomaniki E, George J, PUjades-Rodriguez M, Shah A, Denaxas S, Smeeth L, *et al.* Initial presentation of a wide range of cardiovascular diseases: heterogeneity of lifetime risk and age and sex associations in 1,937,360 people. *Prep* 2013.

335. National Institute for Health and Clinical Excellence. Cardiovascular risk assessment: the modification of blood lipids for the primary and secondary prevention of cardiovascular disease (draft for consultation). London: : National Institute for Health and Clinical Excellence 2007. http://guidance.nice.org.uk/page.aspx?o=438182

336. Cooper A, Nherera L, Calvert N, O'Flynn N, Turnbull N, Robson J, *et al.* Clinical Guidelines and Evidence Review for Lipid Modification: Cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease. London: : National Collaborating Centre for Primary Care and Royal College of General Practitioners 2009. http://www.nice.org.uk/CG67

337. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, *et al.* Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* 2008;336:1475–82.

338. Woodward M, Brindle P, Tunstall-Pedoe H. Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). *Heart* 2007;93:172–6.

339. Caro J, Migliaccio-Walle K, Ishak KJ, Proskorovsky I. The morbidity and mortality following a diagnosis of peripheral arterial disease: long-term follow-up of a large database. *BMC Cardiovasc Disord* 2005;5:14.

340. Sprengers RW, Janssen KJM, Moll FL, Verhaar MC, van der Graaf Y. Prediction rule for cardiovascular events and mortality in peripheral arterial disease patients: data from the prospective Second Manifestations of ARTerial disease (SMART) cohort study. *J Vasc Surg* 2009;50:1369–76.

341. Turnbull F. Effects of different blood-pressure-lowering regimens on major cardiovascular events: results of prospectively-designed overviews of randomised trials. *Lancet* 2003;362:1527–35.

342. Anderson KM, Odell PM, Wilson PW, Kannel WB. Cardiovascular disease risk profiles. *Am Heart J* 1991;121:293–8.

343. Expert panel on detection evaluation and treatment of high blood cholesterol in adults. Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, And Treatment

of High Blood Cholesterol In Adults (Adult Treatment Panel III). *JAMA* 2001;285:2486–97.

344. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, *et al.* Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* 2008;336:1475–82.

345. Conroy RM, Pyorala K, Fitzgerald AP, Sans S, Menotti A, De Backer G, *et al.* Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *Eur Heart J* 2003;24:987–1003.

346. Matheny M, McPheeters M, Glasser A, Mercaldo N, Weaver R, Jerome R, *et al.* Systematic Review of Cardiovascular Disease Risk Assessment Tools. Evidence Synthesis No. 85. AHRQ Publication No. 11-05155-EF-1. Rockville, MD: 2011.

347. National Institute for Health and Care Excellence. Tobacco: harm-reduction approaches to smoking (PH45). London: 2013. http://publications.nice.org.uk/tobacco-harm-reduction-approaches-to-smoking-ph45

348. Blanchard JF, Armenian HK, Friesen PP. Risk factors for abdominal aortic aneurysm: results of a case-control study. *Am J Epidemiol* 2000;151:575–83.

349. Rodin MB, Daviglus ML, Wong GC, Liu K, Garside DB, Greenland P, *et al.* Middle age cardiovascular risk factors and abdominal aortic aneurysm in older age. *Hypertension* 2003;42:61–8.

350. Cornuz J, Sidoti Pinto C, Tevaearai H, Egger M. Risk factors for asymptomatic abdominal aortic aneurysm: systematic review and meta-analysis of population-based screening studies. *Eur J Public Health* 2004;14:343–9.

351. Wanhainen A, Bergqvist D, Boman K, Nilsson TK, Rutegård J, Björck M. Risk factors associated with abdominal aortic aneurysm: a population-based study with historical and current data. *J Vasc Surg* 2005;41:390–6.

352. Forsdahl SH, Singh K, Solberg S, Jacobsen BK. Risk factors for abdominal aortic aneurysms: a 7-year prospective study: the Tromsø Study, 1994-2001. *Circulation* 2009;119:2202–8.

353. National Institute for Health and Clinical Excellence. Hypertension: Clinical management of primary hypertension in adults. Manchester: 2011. http://publications.nice.org.uk/hypertension-cg127

354. Scientific Advisory Committee on Health. Salt and Health. London: 2003.

355. De Vries CS, Bromley SE, Farmer RD. Myocardial infarction risk and hormone replacement: differences between products. *Maturitas* 2006;53:343–50.

356. Hammad TA, Graham DJ, Staffa JA, Kornegay CJ, Dal Pan GJ. Onset of acute myocardial infarction after use of non-steroidal anti-inflammatory drugs. *Pharmacoepidemiol Drug Saf* 2008;17:315–21.

357. Shah AD, Martinez C, Hemingway H. The Freetext Matching Algorithm: a computer program to extract diagnoses and causes of death from unstructured text in

electronic health records. *BMC Med Inform Decis Mak* 2012;12. doi:10.1186/1472-6947-12-88

358. Wang Z, Shah AD, Tate AR, Denaxas S, Shawe-Taylor J, Hemingway H. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. *PLoS One* 2012;7:e30412.

359. Graham H. Smoking prevalence among women in the European Community 1950–1990. *Soc Sci Med* 1996;43:243–54.

360. Modelmog D, Rahlenbeck S, Trichopoulos D. Accuracy of death certificates: a population-based, complete-coverage, one-year autopsy study in East Germany. *Cancer Causes Control CCC* 1992;3:541–6.

361. Alperovitch A, Bertrand M, Jougla E, Vidal JS, Ducimetiere P, Helmer C, *et al.* Do we really know the cause of death of the very old? Comparison between official mortality statistics and cohort study classification. *Eur J Epidemiol* 2009;24:669–75.

362. Coady SA, Sorlie PD, Cooper LS, Folsom AR, Rosamond WD, Conwill DE. Validation of death certificate diagnosis for coronary heart disease: the Atherosclerosis Risk in Communities (ARIC) Study. *J Clin Epidemiol* 2001;54:40–50.

363. De Henauw S, de Smet P, Aelvoet W, Kornitzer M, De Backer G. Misclassification of coronary heart disease in mortality statistics. Evidence from the WHO-MONICA Ghent-Charleroi Study in Belgium. *J Epidemiol Community Heal* 1998;52:513–9.

364. Folsom AR, Gomez-Marin O, Gillum RF, Kottke TE, Lohman W, Jacobs Jr. DR. Out-of-hospital coronary death in an urban population--validation of death certificate diagnosis. The Minnesota Heart Survey. *Am J Epidemiol* 1987;125:1012–8.

365. Ives DG, Samuel P, Psaty BM, Kuller LH. Agreement between nosologist and cardiovascular health study review of deaths: implications of coding differences. *J Am Geriatr Soc* 2009;57:133–9.

366. Lahti RA, Penttila A. The validity of death certificates: routine validation of death certification and its effects on mortality statistics. *Forensic SciInt* 2001;115:15–32.

367. Lloyd-Jones DM, Martin DO, Larson MG, Levy D. Accuracy of death certificates for coding coronary heart disease as the cause of death. *Ann Intern Med* 1998;129:1020–6.

368. Rapola JM, Virtamo J, Korhonen P, Haapakoski J, Hartman AM, Edwards BK, *et al.* Validity of diagnoses of major coronary events in national registers of hospital diagnoses and deaths in Finland. *Eur J Epidemiol* 1997;13:133–8.

369. Goldacre MJ, Duncan ME, Cook-Mozaffari P, Griffith M. Trends in mortality rates comparing underlying-cause and multiple-cause coding in an English population 1979-1998. *J Public Health Med* 2003;25:249–53.

370. Johansson LA, Bjorkenstam C, Westerling R. Unexplained differences between hospital and mortality data indicated mistakes in death certification: an investigation of 1,094 deaths in Sweden during 1995. *J Clin Epidemiol* 2009;62:1202–9.

371. Johansson LA, Westerling R. Comparing hospital discharge records with death certificates: can the differences be explained? *J Epidemiol Community Heal* 2002;56:301–8.

372. Traven ND, Kuller LH, Ives DG, Rutan GH, Perper JA. Coronary heart disease mortality and sudden death among the 35-44-year age group in Allegheny County, Pennsylvania. *AnnEpidemiol* 1996;6:130–6.