Using electronic health records to investigate blood biomarkers and onset of cardiovascular diseases: example of differential white cell count

Anoop Dinesh Shah

UCL

PhD

I, Anoop Dinesh Shah, 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.

Abstract

Background: Electronic health records are invaluable for studying clinically recorded biomarkers and outcomes, but conversion of raw datasets to a research-ready format and replication of analyses can be challenging. The differential white cell count is a common blood test which may reflect inflammation and cardiovascular risk, but previous epidemiological studies have been small or in selected populations.

Aim: To develop methods to assist in using electronic health record databases for research, and apply them to an exemplar study investigating differential white cell counts and onset of cardiovascular diseases.

Methods: The data source was CALIBER: linked electronic health records from primary care, hospitalisation, acute coronary syndrome registry and death registry. Software was developed to assist selection of relevant diagnostic codes, data manipulation, and multiple imputation of missing data. Individuals in CALIBER without prior cardiovascular disease were followed up for a range of specific initial cardiovascular presentations. Survival models were used to investigate the associations of type 2 diabetes and differential white cell counts with initial cardiovascular presentations. The association of total white cell count with mortality was investigated in CALIBER and replicated in the New Zealand PREDICT cohort.

Results: Add-on software packages for the R statistical system were developed and published in an open access repository. Type 2 diabetes was associated with higher risk of coronary disease and ischaemic stroke but lower risk of abdominal aortic aneurysm and subarachnoid haemorrhage. Among 775 231 individuals with a record of differential leukocyte count followed up for median 3.8 years, 54 980 experienced an initial presentation of cardiovascular disease. High neutrophil counts were strongly associated with heart failure and myocardial infarction, but not angina. Low eosinophil counts were associated with heart failure and unheralded coronary death. Total white cell count showed a 'J' shaped association with mortality in both CALIBER and PREDICT.

Conclusions: White cell subtypes are differentially associated with specific initial cardiovascular presentations. A range of tools were developed to assist researchers using CALIBER and other electronic health record data sources.

Publications arising from this thesis:

1. Shah AD, Langenberg C, Rapsomaniki E, Denaxas S, Pujades-Rodriguez M, Gale CP, Deanfield J, Smeeth L, Timmis A, Hemingway H. Type 2 diabetes and incidence of a wide range of cardiovascular diseases: a cohort study in 1·9 million people. Lancet Diabetes Endocrinol. 2014. doi: 10.1016/S2213-8587(14)70219-0

2. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of Random Forest and parametric imputation models when imputing missing data using MICE: a CALIBER study. Am J Epidemiol. 2014; 179(6): 764-774. doi: 10.1093/aje/kwt312

3. Herrett E*, Shah AD*, Boggon R, Denaxas S, Timmis A, Smeeth L, Van Staa T, Hemingway H. 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. doi: 10.1136/bmj.f2350 * joint first authors

Manuscripts in preparation or under review arising from this thesis:

1. Shah AD, Thornley S, Chung S-C, Denaxas S, Jackson R, Hemingway H. A two-country electronic health record cohort study in England (CALIBER) and New Zealand (PREDICT): association of clinically measured white cell count with all-cause mortality.

2. Shah AD, Denaxas S, Nicholas O, Hingorani A, Hemingway H. Clinical relevance of neutrophil counts in the normal range and the incidence of twelve cardiovascular diseases: a CALIBER cohort study.

3. Shah AD, Denaxas S, Nicholas O, Hingorani A, Hemingway H. Eosinophil counts and the incidence of twelve cardiovascular diseases: a CALIBER cohort study.

Acknowledgements

Financial support

I am in receipt of a Wellcome Trust Clinical Research Fellowship (0938/30/Z/10/Z). CALIBER is a collaboration between UCL, the London School of Hygiene and Tropical Medicine (LSHTM) and the Clinical Practice Research Datalink (CPRD). CALIBER is funded by a Wellcome Trust project grant (086091/Z/08/Z) and a National Institute of Health Research programme grant (RP-PG-0407-10314). This study was supported by the Farr Institute of Health Informatics Research, funded by the Medical Research Council (K006584/1), in partnership with Arthritis Research UK, the British Heart Foundation, Cancer Research UK, the Economic and Social Research Council, the Engineering and Physical Sciences Research Council, the National Institute of Health Research, the National Institute for Social Care and Health Research (Welsh Assembly Government), the Chief Scientist Office (Scottish Government Health Directorates), and the Wellcome Trust.

Contribution of other researchers

The CALIBER data portal (chapter 3) was developed by Spiros Denaxas, who also implemented the cohort creation process and extraction of phenotypes in the master CALIBER database. Phenotyping algorithms for identifying patients with the cardiovascular diseases studied in this PhD were generated by a group of CALIBER researchers, principally Julie George, Emily Herrett and myself; I also converted the algorithms into a standard format for uploading to the data portal. The validation study of myocardial infarction recording in CALIBER (section 4.2) was a joint project with Emily Herrett to which we contributed equally. Eleni Rapsomaniki carried out the multiple imputation for the diabetes study (chapter 5). The CALIBER-PREDICT study (chapter 9) was a collaboration with Simon Thornley in New Zealand, who carried out all analyses in the New Zealand data. The remaining work in this thesis is my own.

Other support

I would like to thank my PhD supervisors Prof. Harry Hemingway, Prof. Aroon Hingorani and Dr. Owen Nicholas for their advice and support in preparing my fellowship application and during my PhD.

I would particularly like to thank Spiros Denaxas, who has developed CALIBER as a research platform and made my research much easier, Emily Herrett, with whom I worked on the validation study for myocardial infarction recording in CALIBER, and Simon Thornley, with whom I worked in an international collaboration studying white blood cell counts and mortality in England and New Zealand. I would like to thank Jonathan Bartlett, my MSc project tutor who advised on the new multiple imputation method that was inspired

by the findings of my Medical Statistics MSc. Adam Timmis and Liam Smeeth have given me useful clinical and epidemiological advice.

I would like to thank my colleagues at UCL and LSHTM: Eleni Rapsomaniki, Mar Pujades, Riyaz Patel, Claudia Langenberg, Katherine Morley, Joshua Wallace, Olga Archangelidi, Bram Duyx, Sheng-Chia Chung, Julie George, Dimitris Stogiannis, Giovanna Ceroni, Denise Beales, Carla Logon, Natalie Fitzpatrick, Dorothea Nitsch, Laurie Tomlinson, Harriet Forbes, Catriona Shaw, Hannah Evans, Rosalind Eggo, Arturo Gonzales, Victoria Allan and others I have worked with. I would like to thank Tjeerd van Staa, Tim Williams, Nick Wilson, Arlene Gallagher, Shivani Padmanabhan and the other staff at the Clinical Practice Research Datalink.

Finally I would like to thank my wife Divya ♡, my family and friends for supporting me during this time.

Contents

List of Figures

List of Tables

List of abbreviations

AAA abdominal aortic aneurysm

BP blood pressure

CALIBER CArdiovascular disease research using LInked BEspoke studies and electronic Records

CART classification and regression trees

CI confidence interval

COPD chronic obstructive pulmonary disease

CPRD Clinical Practice Research Datalink

CRP C-reactive protein

CVD cardiovascular disease

ECG electrocardiogram

eGFR estimated glomerular filtration rate

FBC full blood count

FCS full conditional specification

GP general practitioner

GPRD General Practice Research Database

HDL high density lipoprotein cholesterol

HES Hospital Episode Statistics

HMG-CoA 3-hydroxy-3-methylglutaryl coenzyme A

HR hazard ratio

IBD inflammatory bowel disease

ICD-10 International Classification of Diseases, 10th Edition

ICPC International Classification of Primary Care

IL-5 interleukin-5

IL-6 interleukin-6

IPCI Integrated Primary Care Information database

IQR interquartile range

ISAC Independent Scientific Advisory Committee

MAR missing at random

MCAR missing completely at random

MeSH medical subject headings

MHRA Medicines and Healthcare products Regulatory Agency

MI myocardial infarction

MICE multivariate imputation by chained equations

MINAP Myocardial Ischaemia National Audit Project

NHI National Health Index (New Zealand)

NHS National Health Service

NK natural killer

NSTEMI non ST elevation myocardial infarction

ONS Office for National Statistics

OPCS Office for Population Censuses and Surveys

RR relative risk

SD standard deviation

SNP single nucleotide polymorphism

STEMI ST elevation myocardial infarction

T2D type 2 diabetes

TIA transient ischaemic attack

UK United Kingdom of Great Britain and Northern Ireland

WBC white cell count

1 Introduction

1.1 Outline of thesis

Electronic health records and advances in computing technology provide an exciting opportunity for improving health and healthcare by harnessing the power of big data. This thesis explores the use of electronic health record databases for investigating blood biomarkers and patient outcomes, with the example of differential white blood cell count and onset of different cardiovascular diseases in initially healthy people.

Electronic health record data are 'big data' in every sense of the phrase and can yield useful insights into every aspect of health and care, from aetiologic studies to clinical trial design to monitoring the quality and safety of healthcare. In this thesis, as well as using electronic health records to answer a set of clinical questions, I examine the research process itself from raw data to results. The thesis describes new tools I have developed to assist in data preparation, variable definition and statistical analysis.

The first chapter of this thesis introduces the use of electronic health records for research, the cardiovascular diseases of interest, the biological roles of white blood cells, and why white blood cell counts are relevant to the development of cardiovascular diseases. Chapter 2 reviews previous epidemiological studies investigating white blood cell counts and incidence of cardiovascular diseases. Chapter 3 describes CALIBER [1], a research platform comprised of linked primary care, hospitalisation, mortality and acute coronary syndrome registry data in England. CALIBER was the main database used for the studies that comprise this thesis. Chapter 4 describes validation work and methods I have developed to assist in the use of CALIBER data for epidemiological studies. Chapter 5 describes the identification of the initial presentation of cardiovascular diseases in a cohort of individuals in CALIBER. In this chapter, I investigate type 2 diabetes and the blood marker of diabetic control, HbA1c, as risk factors for cardiovascular diseases.

Chapter 6, describes the extraction and exploration of differential white blood cell counts recorded in CALIBER. The main study is reported in chapters 7 and 8. This is an investigation of the association of the differential white blood cell count with the onset of twelve cardiovascular diseases, particularly focussing on neutrophil count and eosinophil count. Replication of study findings in other populations is important to verify the generalisability of the conclusions, so the following chapter (chapter 9) reports an analysis of total white blood cell count and mortality in CALIBER and the New Zealand PREDICT cohort.

The final discussion (chapter 10) brings together the findings of this thesis and includes recommendations for clinical practice, research and future development of electronic health records.

1.2 Electronic health records for epidemiology

Electronic health records (EHR) are being increasingly used in many countries around the world. Research databases based on electronic health records are invaluable for studying disease epidemiology, pharmacoepidemiology and quality of care [2]. They record 'real-world' clinical information and treatments, outside the formal structure of clinical trials or bespoke cohort studies. EHR will become increasingly important in the era of precision medicine [3], where detailed genotype and phenotype information will inform the development of new specialised therapies. Non-healthcare data (such as geographic data, social care data and data from mobile devices) are increasingly being linked to healthcare data in order to answer new questions and inform public health interventions [2].

1.2.1 Electronic health record data sources

Information in these 'real-world' data sources can be generated in a number of ways:

Information entered by clinicians. Symptoms, diagnoses, examination findings, letters, prescriptions and reports of investigations are examples of reports generated by clinicians, which may be entered into electronic health records in a variety of formats. Typically diagnoses and prescribed medication are coded using a dictionary, but much information is entered as unstructured free text, and some is available only as scanned text or handwriting. In the UK, general practices have been using computers for over 20 years [4], and pseudonymised patient data have been compiled into a number of research databases. The Clinical Practice Research Datalink (CPRD, `www.cprd.com`), formerly known as the General Practice Research Database, is owned by the Medicines and Healthcare products Regulatory Agency and is described in more detail in subsection 3.3.1 on page 53. The Health Improvement Network (THIN, `www.thin-uk.net`), QResearch (`www.qresearch.org`) and ResearchOne (`www.researchone.org`) are collaborations between academic groups and the system manufacturer. These databases generally include only coded and numerical data; limited amounts of free text are available for research use in CPRD and THIN but they have to be manually anonymised before being released to researchers. CALIBER contains data from a subset of CPRD practices which consented to linkage of patient records with other data sources.

Because general practitioners (GPs) look after patients over the long term and receive letters from other care providers, the GP record is the most complete longitudinal health care record that exists for most people in the UK. However, it does not include details of acute conditions or investigations performed in secondary care, such as acute hospital care or mental health care.

In contrast to general practice, adoption of electronic health record systems in UK hospitals has been slow [5], and studies using hospital electronic records rarely involve multiple sites. This is in contrast with the US, where research studies can draw

on the electronic health records of multiple hospitals, such as the eMERGE Network https://emerge.mc.vanderbilt.edu/ [6]. An initiative by the National Institute for Health Research, the 'Health Informatics Collaborative' aims to addess this by setting up a collaboration between five hospitals in England with large biomedical research centres (www.nihr.ac.uk/about/nihr-hic.htm) in order to share patient data for research.

Laboratory results. Local and regional databases of laboratory results can be invaluable for research when linked to other data sources, as has been done for the Test Safe database in New Zealand [7]. Where laboratory results are electronically transferred into electronic health records, such as in UK general practice, this information can be conveniently extracted for research alongside detailed clinical information.

Prescribing information. Electronic prescribing systems can provide detailed information on drug prescription and drug administration, and will be increasingly useful for research as more hospitals adopt electronic prescribing. Individual level data on prescriptions dispensed in the community are available in some countries such as New Zealand [8], and can form a useful research resource when linked to other individual patient data sources such as hospitalisation, laboratory results and mortality.

Registries, audits, mandatory reporting databases There are numerous registries, audits and mandatory patient-level reports collected on a regional or national level in the UK and other countries. Linkage of these data sources to outcomes such as mortality can be a vital component of clinical audit to improve the quality of care. CALIBER contains data from MINAP, the Myocardial Ischaemia National Audit Project, which collects information on acute coronary syndromes in England and Wales; it is described in more detail in subsection 3.3.3 on page 57.

Registries can collect information in a variety of ways, and information submitted to the registry may or may not also be part of the patient's medical record as used for clinical decision making. Information can be abstracted from the clinical record by clerks or audit nurses (as in MINAP [9]), submitted directly by clinicians using a web-based form (e.g. SWEDEHEART, the Swedish heart attack register [10]) or extracted automatically from the electronic health record (e.g. the English National Diabetes Audit [11]). National death registries, while not part of the EHR themselves, are an important data source for many long-term studies using EHR or registry data sources.

Secondary use data Many countries maintain databases of hospital admission data for administrative and billing purposes, such as Hospital Episode Statistics (HES) in England (www.hscic.gov.uk/hes), Patient Episode Database in Wales (www.infoandstats.wales.nhs.uk/page.cfm?orgid=869&pid=40977), the Scottish Morbidity Record (www.ndc.scot.nhs.uk/Data-Dictionary/SMR-Datasets/), and the SPARCS database in New York State (https://www.health.ny.gov/diseases/chronic/hospdata.htm). HES

submissions from hospitals contain admission dates, purpose of admission (emergency or elective), specialty, coded primary and secondary diagnoses, and procedures performed. Hospitals in England are reimbursed according to tariffs associated with the diagnoses and procedures coded. However, the information in the HES submissions is not used for direct patient care, and involvement of clinicians in the coding process is variable [12].

1.2.2 Linkage of data sources

Patient records in different data sources can be linked in order to provide a richer dataset for research and overcome some of the limitations of an individual data source. For example, linkage of a general practice database to a cancer registry can allow detailed clinical information about the cancer to be combined with longitudinal information on a patient's other medical conditions [13]. For some research questions, outcomes and exposures may be recorded in different datasets, so linkage is the only way to answer the question, whereas in other cases linkage may enrich the main data source or enable validation of the data.

Linkage algorithms aim to identify which records in different datasets (or within a single dataset) belong to the same patient, using identifiers such as NHS number, sex, date of birth and address or postcode [14]. If strong identifiers are well recorded in both source datasets the linkage may be considered 'deterministic', with matches considered to be 'exact'. However, if each record may have a number of potential matches with different probabilities, records can be linked probabilistically in order to take into account the uncertainty in the linkage and obtain unbiased results [15].

Linkage of data sources has to be performed carefully to ensure that linkage errors (missed matches and false matches) do not bias the results [16]. Where different datasets are used for numerator and denominator, bias can be caused by variations in the probability of linkage between subgroups, such as between ethnic groups [17]. It is important to cross-reference linked datasets and try to understand the reasons why clinical events may not be recorded in one or more of the data sources [18]. New statistical approaches allow the uncertainty in matches to be propagated into analysis results, for example using multiple imputation [19], in order to obtain unbiased results even in the presence of linkage error.

1.2.3 Electronic health records versus bespoke studies

Historically, many epidemiological findings have been based on bespoke cohort studies, as this was the only way to obtain reliable follow-up data on a large number of people. However, if the study only requires information about a person's health which is already captured in usual clinical practice, EHR databases may be preferable to a bespoke cohort study, for the following reasons:

- Larger size – EHR databases can potentially contain many more patients than any consented cohort, and this will grow with time. This is essential for comparing the strength of association of risk factors with rare outcomes. As an illustration I carried out a sample size simulation for a cohort study investigating an endpoint with incidence 8 per 100 000 patient-years, described in more detail in section 6.3 (**Table 6.1 on page 134**). In this example, a sample size of 700 000 with mean 3 year follow-up is required to detect with 80% power that hazard ratios 1.2 and 1.8 are significantly different.

- More diseases – Comprehensive EHR systems record information about all diseases and conditions affecting a patient, rather than the limited set of conditions or information that an investigator decides to collect.

- Less selection bias – EHR databases based on usual clinical care are more representative of the general population than a bespoke study cohort, which may have a low response rate for recruitment [20].

- Lower cost – Re-using information which is already collected for clinical care allows data on large numbers of people to be aggregated cheaply.

- Wide coverage – EHR data sources can potentially cover entire regions or countries. In contrast, cohort studies may be geographically isolated to facilitate recruitment and follow-up of participants. For example, one of the best-known cohorts for investigating the onset of cardiovascular diseases was based in Framingham, a small town of 28 000 inhabitants in the US state of Massachusetts [21].

- Contemporary – EHR cohorts provide information on current patterns of morbidity and healthcare, compared to historical cohort studies.

- General population denominator – UK general practice databases such as CPRD can provide a population denominator, because the majority of the UK population is registered with a general practitioner and the entire population is eligible for care within the same system (the NHS) [22].

- Link with clinical care – EHR captures information on 'real-world' clinical events, and can provide valuable insights into the quality of care. There is also the exciting potential for clinical trials to be performed within EHR cohort populations, although there are still many challenges to overcome [23].

EHR data sources also have some disadvantages, which means that consented cohort studies are necessary for answering some research questions. A key disadvantage is that the completeness of data and granularity of clinical information in EHR can be variable. Some information may be 'missing' in the sense that it was generated but not recorded in the EHR in a way that can be extracted (e.g. as free text instead of a coded value), but much information is 'missing' because it was never generated in the first place, if it was

not clinically relevant. Clinical parameters may be measured only for a subset of patients and at variable times; this needs to be considered carefully when planning studies using these data sources in order to avoid introducing bias. For example, HbA1c, a blood test for diagnosing diabetes and assessing disease control, is measured only in patients who already have diabetes, those with suspected diabetes and those in whom the clinician wishes to rule out diabetes (see chapter 5).

1.3 The CALIBER research platform – electronic health records for cardiovascular research

In this thesis, I used linked data from CALIBER (Cardiovascular Research using Linked Bespoke Studies and Electronic Records). The CALIBER research platform includes primary care data from CPRD, hospitalisation data from HES, detailed information on acute coronary syndromes from MINAP and date and cause of death from the national death registry [1]. CALIBER contains data from 244 CPRD general practices in England which consented to the linkage, covering about 4% of the English population in 2006 [18], and is curated by the Clinical Epidemiology Group in the UCL Institute of Health Informatics. The data sources and validation studies for CALIBER are described in detail in the next two chapters. This section compares CALIBER with primary care data sources available in other countries.

CALIBER has a number of unique features that set it apart internationally from other cardiovascular research datasets, and make it the ideal data source for this thesis [1]. England has detailed electronic primary care records, national cardiovascular registries, a national database of hospitalisation records and a national mortality registry, with a unique person identifier (the NHS number) facilitating linkage between the data sources. There are a number of general practice databases in England based on different EHR system manufacturers, but currently CPRD is the only one linked to a cardiovascular registry and other data sources (the CALIBER platform).

Scotland and Wales also have strong primary care data sources, with some practices contributing to CPRD. They also have their own primary care databases which include a large proportion of practices in the country: the Welsh SAIL databank `www.saildatabank.com` and the Scottish Primary Care Information Resource `www.spire.scot.nhs.uk/`. There is ongoing work to link these databases to other data sources, but at the time of writing this thesis, CALIBER was the only primary care database linked to a cardiovascular registry, hospitalisation and mortality records.

Small collections of primary care data have been used for research in a number of countries, but the only large database (with over a million patients) outside the UK is in the Netherlands: the Integrated Primary Care Information (IPCI) database [24]. IPCI contains records of diagnoses (coded using the International Classification of Primary Care, ICPC), referrals, laboratory test results, surgical procedures, medication and hospitalizations, and is used for postmarketing surveillance [25] and epidemiology [26]. However it

is not linked to other data sources. In New Zealand, the PREDICT programme gathers information on cardiovascular risk assessments in primary care and links them with laboratory values, dispensed medication, hospitalisation and mortality data. Full longitudinal general practice records are not available in PREDICT, but for this thesis the information on cardiovascular risk factors and outcomes was valuable for a replication study (chapter 9).

Differences in the way healthcare is organised in other countries may preclude aggregation of ambulatory care information into a research database. Unlike the UK and Scandinavian countries, in many countries it is usual for patients to present directly to a specialist, bypassing the gatekeeper role of primary care providers, meaning that they may not have a comprehensive longitudinal patient record curated in a single place. Some US managed care organisations (e.g. Kaiser Permanente) curate research databases spanning primary and secondary care, but the lack of a national healthcare system means that US healthcare databases are limited to a selected segment of society, such as persons with a particular health insurance plan, older persons or veterans.

A strength of UK primary care data is that it can provide a population denominator for calculation of incidence rates, as the majority of people are registered with a general practitioner [22] and it is only possible to register permanently with a single practice (temporary registrations are excluded from research databases). It is not possible to obtain a population denominator from primary care data in countries that allow people to register with more than one primary care practice. Deficiencies in primary care information systems are another source of limitation; a conference workshop in 2006 noted that some countries had a multitude of incompatible primary care systems (e.g. Germany, Croatia) and some used free text extensively rather than coded data (e.g. Czech republic) [27]. In contrast, the Netherlands and the UK have almost complete coverage of primary care with only a few EHR systems covering the entire country [27].

Scandinavian countries have extensively linked disease registries and other healthcare datasets, but lack large primary care research databases. For example, the CopDiff database in Copenhagen (`http://almenpraksis.ku.dk/english/research/copdiff/`) links differential white cell counts performed in general practice for any reason with outcomes including cancer, hospitalisations and mortality [28]. The absence of primary care data limits the utility of these databases for cardiovascular research, as important risk factors such as smoking and body mass index are not recorded [29].

In summary, for research purposes UK primary care data sources have a number of key advantages over data from other countries:

- Population-based – most people are registered with a general practice, and patients are permitted to register with only one primary care practice (e.g. UK, the Netherlands, Scandinavian countries)
- Strong primary care – the GP maintains a longitudinal care record and is gatekeeper for secondary care (e.g. UK, the Netherlands, Scandinavian countries)

- Small number of clinical systems used in primary care – this facilitates data extraction for research (e.g. UK, the Netherlands)
- Important diagnoses recorded using a coding system (e.g. UK, the Netherlands)
- Linkage of primary care data to other sources – as far as I know, this is only done in the UK at scale

Only the UK and the Netherlands have population-based primary care data available on over a million patients. The CPRD database in England is much larger than IPCI, and is linked to other data sources to form the CALIBER research platform. The richness of the source datasets make CALIBER ideal for the studies in this thesis.

1.4 Recording of blood biomarkers in electronic health records

Many blood tests are performed in clinical practice, and they are commonly stored in a structured electronic format in electronic patient records, making them amenable to research. Research on blood biomarkers is important for understanding how they are associated with disease and how they can best be used to assist clinical decision making. For a number of years UK general practices have received results electronically from the laboratory, and CPRD contains these results. However, tests requested by hospital clinicians are generally not included, only tests requested by the GP.

The differential leukocyte count is a component of the full blood count, one of the most common blood tests in medical practice. A wide range of parameters are reported by blood count analysers and are included as part of the report, but clinicians are typically interested in only a handful of parameters, or may be interested only in gross perturbations of some of the parameters. The richness of the continuous values of these parameters are not fully used clinically. This thesis will explore the associations of the differential leukocyte count with cardiovascular diseases, and seeks to contribute to discussions on how these tests should be interpreted and used clinically.

1.5 Initial presentation of cardiovascular diseases

Cardiovascular disease is the most common cause of death worldwide and a significant cause of morbidity [30] and healthcare costs (21% of overall NHS expenditure in the UK in 2004 [31]). Myocardial infarction and stroke are acute cardiovascular events responsible for much of this morbidity and mortality, but they have declined in incidence in developed countries [32, 33]. Chronic cardiovascular conditions such as stable angina and heart failure comprise a significant proportion of the contemporary cardiovascular disease burden [34]. These conditions have in common cardiovascular risk factors such as smoking [35] and hypertension [36], and many of them also share aetiological mechanisms, such as atherosclerosis.

Atherosclerosis is a pathological process in which lipid-laden deposits in arterial walls cause arteries to become narrower, restricting the blood flow and increasing the risk of thrombosis. In coronary arteries this can lead to myocardial infarction (MI); in the cerebral circulation it can cause strokes and in the peripheral arteries it can cause intermittent claudication, ischaemic legs and gangrene [37]. Damage to arterial walls is another important pathological process which can lead to arterial stiffness, weakness and propensity to rupture [38]. The myocardium can also be affected – chronic hypertension can lead to left ventricular hypertrophy, with increased stiffness and impaired function, eventually leading to heart failure [39].

For this thesis I chose to study the following twelve initial presentations of cardiovascular diseases: stable angina, unstable angina, non-fatal myocardial infarction, unheralded coronary death (fatal myocardial infarction as the initial presentation of cardiovascular disease), transient ischaemic attack, ischaemic stroke, haemorrhagic stroke, subarachnoid haemorrhage, a composite of ventricular arrhythmia or sudden cardiac death, heart failure, peripheral arterial disease and abdominal aortic aneurysm.

These diseases are responsible for much morbidity and mortality, and share a number of pathological processes that involve leukocytes; they are summarised in **Table 1.1 on page 29** and described in more detail in the rest of this chapter.

Table 1.1: Summary of key pathological processes involving leukocytes that may contribute to the twelve cardiovascular diseases of interest

Initial presentation of cardiovascular disease	**Pathological process**		
	Atherosclerosis	**Inflammation**	**Thrombosis**
Coronary artery disease presentations			
Stable angina	●	●	
Unstable angina	●	●	○
Non-fatal myocardial infarction	●	●	●
Unheralded coronary death	●	●	●
Cerebrovascular disease presentations			
Transient ischaemic attack	○	○	●
Ischaemic stroke	○	○	●
Haemorrhagic stroke	?	?	
Subarachnoid haemorrhage	?	?	
Other cardiovascular diseases			
Ventricular arrhythmia / sudden cardiac death	○	○	○
Heart failure	○	○	
Peripheral arterial disease	●	●	○
Abdominal aortic aneurysm		●	

Legend: ● always, ○ sometimes, ? possibly

Coronary artery disease presentations. Narrowing of coronary arteries without thrombosis causes stable angina. Few studies estimate separate associations for subtypes of coronary disease, but I wished to do so because the processes of atherosclerosis and thrombosis may have differential importance in these presentations and the underlying risk factors may be differentially associated with angina and myocardial infarction. I separated stable and unstable angina, and for myocardial infarction I differentiated between fatal and non-fatal events, as they give rise to very different patient experiences. Dropping dead with a heart attack 'out of the blue' ('unheralded coronary death'), in a person with no known history of cardiovascular disease, is a particularly devastating event which deserves special consideration.

Cerebrovascular disease presentations. Among cerebrovascular diseases, I sought to differentiate between different causes of stroke, namely ischaemic, haemorrhagic (intracerebral haemorrhage) and subarachnoid haemorrhage.

Subarachnoid haemorrhage is usually due to rupture of an intracranial arterial aneurysm [40], but it has been suggested that inflammation and atherosclerosis also play a role. One cohort study found an association between total leukocyte counts and incidence of subarachnoid haemorrhage [40], and histological studies of ruptured aneurysms found inflammatory infiltrates near the site of rupture as well as atherosclerotic lesions [41].

I included transient ischaemic attack as a separate endpoint because it commonly leads to ischaemic stroke but the patient experience is very different, as they receive an initial warning rather than an unheralded disabling stroke.

Other cardiovascular diseases. I included a composite of ventricular arrhythmias and sudden cardiac death (presumably due to ventricular fibrillation) as an 'arrhythmia' endpoint, in order to differentiate between mechanisms acting via atherosclerosis and ischaemic heart disease versus those increasing the risk of arrhythmias by direct influence on the myocardium. Heart failure was another endpoint; as the initial presentation of cardiovascular disease, it may be due to chronic hypertension or silent ischaemic heart disease. Abdominal aortic aneurysms are characterised by an inflammatory process in the aortic wall with leukocyte infiltration, destruction of elastin and collagen and loss of smooth muscle [42, 43]; ruptured or leaking aneurysms can be fatal [44]. Peripheral arterial disease commonly presents initially with chronic symptoms (intermittent claudication) but can progress to critical limb ischaemia and gangrene; smoking is a strong risk factor [35].

I included non-specific stroke and non-specific coronary disease as additional endpoints, as some patients did not have a more specific cardiovascular diagnosis recorded in the EHR. For some analyses these endpoints may be grouped with the specific endpoint they are most similar to; I have described these decisions in chapter 5 where I report the phenotype definitions in CALIBER for these cardiovascular diseases.

Cardiovascular diseases not included. I did not include infective conditions such as endocarditis or inflammatory conditions such as vasculitis as initial presentations of cardiovascular disease, as the risk factors and preventative strategies are different. Similarly I chose not to include congenital or degenerative conditions such as valve disease. I also excluded conditions that may be asymptomatic and are predominantly cardiovascular risk factors rather than diseases in their own right, such as diabetes, hypertension and obesity. As I was studying only the initial presentation of cardiovascular disease in this project, I disregarded events occurring after the initial presentation, and patient's follow-up ended at that point.

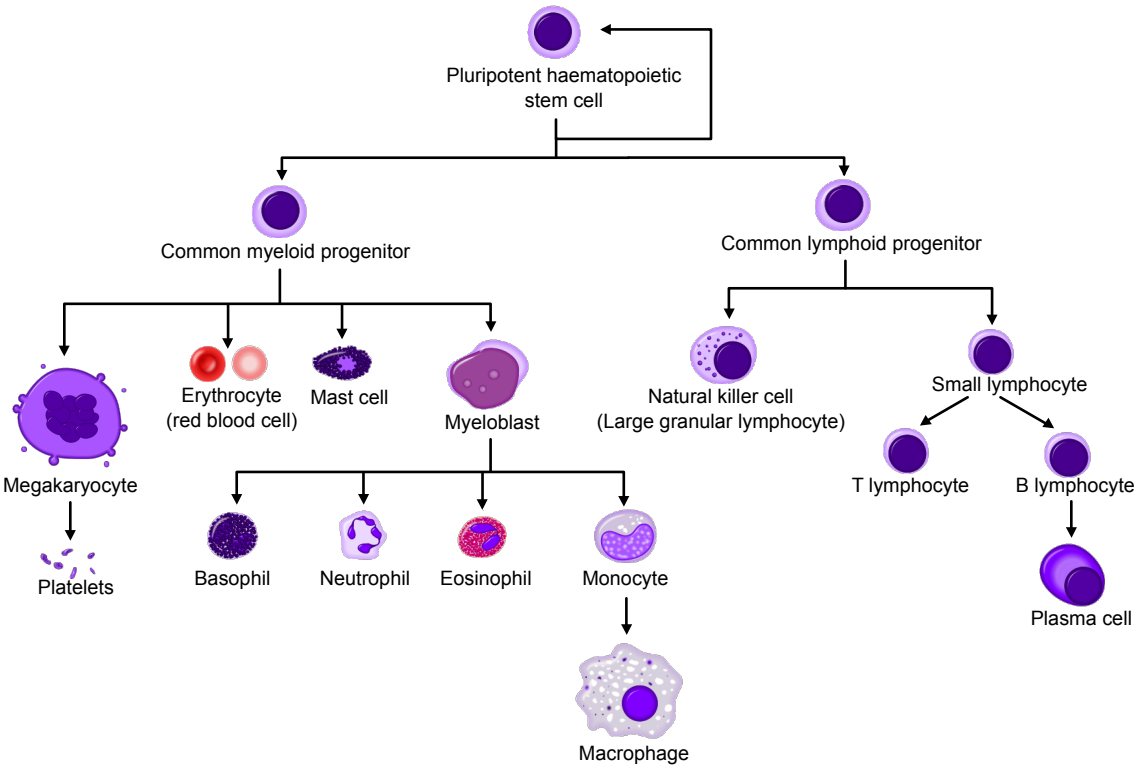
1.6 Cardiovascular diseases and inflammation

Many cohort studies have demonstrated that total white blood cell counts are associated with greater incidence [45] and worse prognosis [46] of coronary disease [47]. White blood cells play a major role in atherosclerosis, which is an inflammatory process [48], and some cardiovascular diseases such as abdominal aortic aneurysms are a result of other inflammatory processes which weaken blood vessel walls [44]. However, although leukocyte counts are included in the full blood count (complete blood count), one of the most commonly performed blood tests, these numbers are not currently used in cardiovascular decision making or risk prediction [30]. Part of the reason might be that differential leukocyte counts were not widely available at the time that many of the risk factors for atherosclerosis were being discovered. In 1948, when the Framingham heart study commenced, automated cell counters were not available so leukocyte counts could only be obtained by laborious manual counting [21]. Regression dilution bias due to imprecision in manual counting may have resulted in under-estimation of associations between leukocyte components and coronary heart disease [49].

Why focus on leukocyte counts, when there are other inflammatory biomarkers available to measure? C reactive protein (CRP) and interleukin-6 (IL-6) are associated with cardiovascular disease [50, 51] and measurable in peripheral blood. However, the advantage of studying leukocyte counts is that they are already measured in a large number of patients, making the study possible in clinically collected data. Associations of leukocyte counts can be studied in relation to the large number of diseases recorded in clinical data, rather than a narrow subset chosen as endpoints of a bespoke study. If a risk prediction model based on leukocyte counts were found to be clinically useful, it could be incorporated in risk prediction models for minimal cost. A further reason is that the leukocyte subtypes have well-known biological roles which are potentially causal in atherosclerotic disease, unlike CRP, for which Mendelian randomisation studies suggest it is not causative in cardiovascular disease [52].

Manipulation of the immune system is now considered a potentially useful therapeutic mechanism in cardiovascular disease, and leukocyte counts could be important for guiding or monitoring such therapy. Trials are currently underway to examine whether

Figure 1.1: Simplified diagram of haematopoiesis



Modified from image in Wikimedia Commons by A. Rad.

anti-inflammatory agents such as methotrexate can prevent cardiovascular disease in high risk individuals [53]. This is described further in section 1.8 on page 36.

1.7 Leukocyte biology

1.7.1 Leukocyte production and differentiation

All blood cells are derived from haematopoetic stem cells which reside in the bone marrow. Pluripotent stem cells give rise to lymphoid and myeloid lineages (**Figure 1.1 on page 32**). Lymphocytes are derived from a lymphoid progenitor; basophils, neutrophils, monocytes and eosinophils are derived from a granulocyte / myeloid progenitor. Monocytes migrate into tissues and differentiate into macrophages. 'Granulocytes' include neutrophils, basophils, eosinophils and mast cells [54]. Neutrophils comprise the majority of granulocytes and are also the most numerous type of white blood cell in most patients.

Lymphocytes, monocytes, basophils, eosinophils and neutrophils are measured in peripheral blood of patients as part of full blood count analyses. This test is performed for a variety of indications including suspected disorders of erythrocytes (e.g. anaemia, bleeding, haemolysis or polycythaemia), platelets (e.g. bruising) or leukocytes (e.g. infection, leukaemia), or any general medical condition.

Leukocyte counts can be affected by many factors such as acute and chronic infections, autoimmune disease, medication and haematological conditions. Knowing whether a patient is clinically unwell at the time of a blood test is important in clinical decision making, and is likely to be important for interpreting tests in a research context. EHR data sources that record a wide range of general medical conditions (such as infections) are essential to identify tests performed when a patient was clinically 'stable'.

The underlying stable leukocyte count for a patient is thought to be genetically determined. The eMERGE consortium has developed standardised phenotypic definitions for identifying acute conditions at the time of blood testing in order to identify stable leukocyte count values that can be correlated with genotype [55, 56].

1.7.2 Physiological roles of leukocytes

Neutrophils

Neutrophils are the most abundant type of white blood cell in humans and are an essential part of the innate immune system. They are essential for defence against bacterial infection, and have several modes of action: they can phagocytose pathogens, they can release cytotoxic free radicals and proteolytic enzymes [57], or they can destroy pathogens using 'extracellular traps' consisting of chromatin and granules which form an extracellular fibril matrix [58].

Neutrophils become activated during myocardial ischaemia [57]. Neutrophils have been detected in atherosclerotic plaques [59] and might increase the risk of atherosclerotic plaque rupture and thrombosis by damaging or obstructing blood vessels, or by activating the coagulation system [60]. Overactive neutrophil extracellular traps and neutrophil-platelet interactions can also play a part in causing thrombosis [58].

Neutrophils are thought to be involved in the development of abdominal aortic aneurysms, which are characterised by an inflammatory process in the aortic wall with leukocyte infiltration, destruction of elastin and collagen and loss of smooth muscle [42, 43]. Evidence for the importance of neutrophils comes from experiments in mice, in which neutrophil depletion reduced the formation of aortic aneurysms [61].

Higher neutrophil counts are associated with smoking [62] and increased incidence of type 2 diabetes [63]. Hence it seems likely that they are associated with cardiovascular diseases, and their roles in inflammation and thrombosis suggest that this may be a causal relationship rather than merely a marker of an inflammatory state.

Lymphocytes

Lymphocytes comprise T, B and natural killer (NK) cells, each of which consists of a number of subclasses with different functions. B lymphocytes include B2 cells, which are associated with adaptive immunity. They respond to antigen presentation in a T-cell dependent manner and produce antibodies against foreign pathogens [64]. B1 cells are

a component of the innate immune system, and are thought to be protective in atherosclerosis because they spontaneously produce natural antibodies that bind to oxidised LDL cholesterol and apoptotic cells [64]. Regulatory B lymphocytes can suppress the immune response [64].

T lymphocytes include a range of different subtypes. T helper cells are required for the promotion of certain types of immune response. T regulatory cells induce immune tolerance and may be atheroprotective. CD8+ cytotoxic T cells infiltrate atherosclerotic plaques and promote cell death [65].

NK cells are a component of the innate immune system. Their function is to cause lysis of virally infected cells and tumour cells [66], but they have been detected in atherosclerotic plaques, suggesting that they may play a role in atherosclerosis [67].

Associations between lymphocyte count and cardiovascular disease may not be as straightforward as for the other cell types because different lymphocyte subtypes are involved in different immune responses and may have opposing influences on atherosclerosis and inflammation. Lymphocyte subsets and their association with cardiovascular diseases have been investigated only in small studies (several hundred patients) [65], and are measured only for specific indications rather than in usual care.

Lymphocyte counts may be relevant to cardiovascular diseases by mechanisms other than atherosclerosis. Low total lymphocyte count is associated with malnutrition and immune suppression [68] and observational studies have shown that it is a poor prognostic factor in patients with heart failure [69] and ST elevation myocardial infarction [70].

Monocytes

Monocytes migrate across the epithelium of the blood vessel into the atheroma and differentiate into macrophages [71]. Macrophages phagocytose and accumulate lipids, secrete cytokines that attract other leukocytes, digest the extracellular matrix, disturb smooth muscle cell function, and can influence vasodilatation [71]. After myocardial infarction in a mouse model, increased sympathetic nervous signalling leads to increased production of leukocytes including monocytes and acceleration of atherosclerotic plaque growth, which may contribute to reinfarction [37].

There are two subsets of monocytes: $CD14^{hi}CD16^{-}$ monocytes are inflammatory and have been associated with myocardial infarction and stroke in patients undergoing elective angiography [72], whereas $CD14^{+}CD16^{+}$ monocytes are thought to be primarily reparative. In the $ApoE^{-/-}$ mouse model of hypercholesterolaemia, mice fed a high-fat diet had increased monocyte and granulocyte counts and increased aortic atherosclerosis [73].

Eosinophils

Eosinophils are multifunctional leukocytes which are more abundant in parasitic infestations [74] but their main role is thought to be in maintaining tissue homeostasis through

'Local Immune And Remodelling / Repair' activities [75]. Eosinophil counts increase acutely after myocardial infarction [76] and histological studies found large numbers of eosinophils in coronary thrombi [77] and post-mortem hearts with cardiac rupture [78]. Genome wide association studies found that a non-synonymous single nucleotide polymorphism (SNP) in SH2B3 was associated both with eosinophil count and myocardial infarction [79], suggesting a potential causal link.

Interleukin-5 (IL-5) is the principal cytokine mediating the release of eosinophils into the bloodstream [80], and is strongly associated with eosinophil counts in cross-sectional studies [81]. Eosinophil counts are raised in diseases characterised by an overactive immune response such as asthma [82], for which anti-IL5 monoclonal antibodies are being investigated as a potential therapy [83].

There is some evidence that IL-5 may be atheroprotective – a cross-sectional study found that higher IL-5 was correlated with decreased subclinical atherosclerosis [84] and variants in the IL-5 gene are associated with coronary artery disease [85]. In mouse studies, macrophage-specific overexpression of IL-5 led to increased secretion of IgM antibodies that inhibit uptake of oxidized low-density lipoprotein, with consequent inhibition of atherosclerosis [86].

Basophils

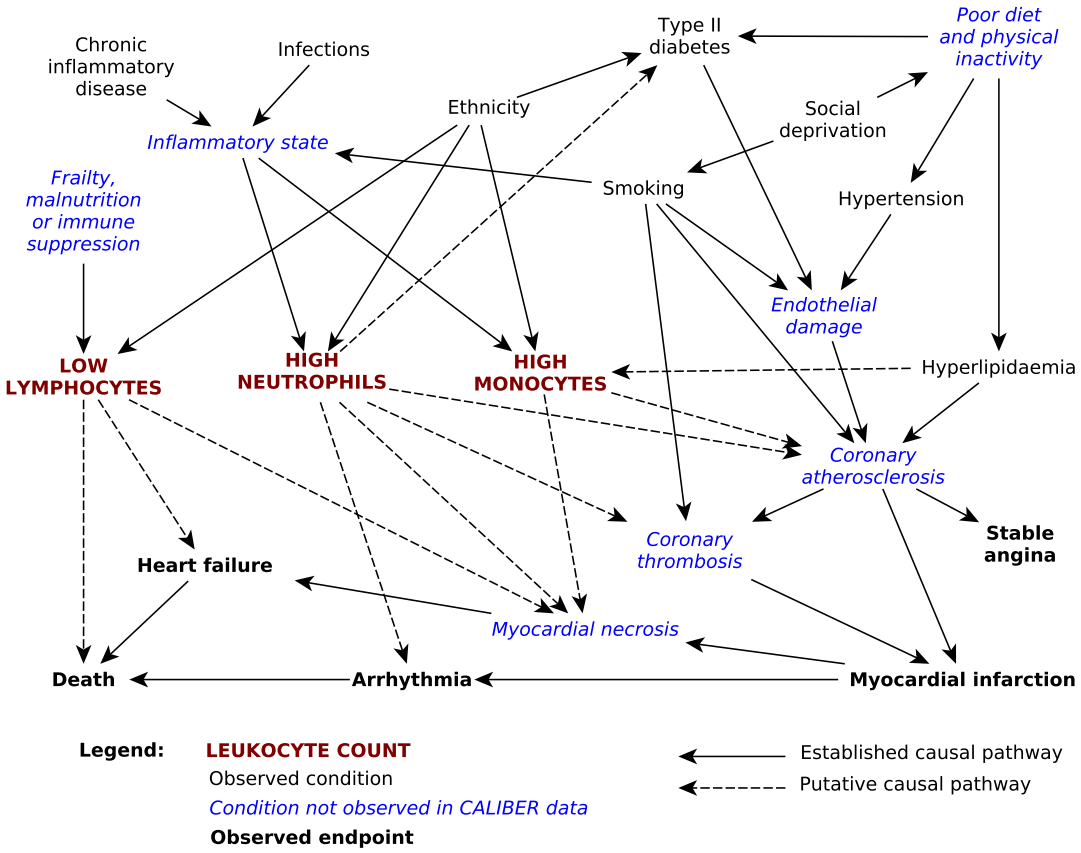
Basophils are involved in allergic and other inflammatory reactions. They release a number of biologically active mediator substances (such as histamine) in response to IgE receptor activation. Cerivastatin and atorvastatin inhibit IgE-mediated histamine release in human basophils in vitro [87], and it is thought that this may contribute to their anti-inflammatory effect.

1.7.3 Leukocytes and atherosclerosis

The process of atherosclerosis involves the accumulation of low-density lipoproteins containing cholesterol in the artery wall. These deposits activate the endothelium, triggering the recruitment of monocytes and T cells. Monocytes migrate into the intima and differentiate into macrophages, which internalise the lipoproteins to form large 'foam cells'. T lymphocytes in lesions recognize local antigens and secrete pro-inflammatory cytokines that contribute to local inflammation and growth of the plaque. Inflammation may lead to local proteolysis, plaque rupture, and thrombus formation, which causes ischaemia and infarction [88].

These processes may give rise to causal relationships between circulating leukocytes and specific cardiovascular diseases, which I have illustrated in **Figure 1.2 on page 36**. Smoking causes atherosclerosis and thrombosis by a variety of mechanisms [89], and is associated with higher neutrophil, lymphocyte and monocyte counts [62]. Higher neutrophil count is associated with increased incidence of type 2 diabetes [63]. Death

Figure 1.2: Directed acyclic graph showing how neutrophils, monocytes and lymphocytes may be involved in causal pathways for coronary artery disease



following myocardial infarction is considered to occur either via heart failure or arrhythmia [90], although there are other less common causes such as myocardial rupture.

1.8 Effects of drugs on leukocytes

There are many drugs such as steroids, chemotherapy agents and biologic agents which can affect leukocyte activity or counts. Most of these drugs are not currently used therapeutically in atherosclerotic cardiovascular disease, except for statins. Statins may have greater benefit in patients with higher total leukocyte counts, as shown in the LIPID study of pravastatin [91]. The differential leukocyte count was not measured in this study, but the total leukocyte count is largely a reflection of neutrophil count as they are the most numerous type of white blood cell.

Statins are commonly used for treatment of hypercholesterolemia and prevention of atherosclerotic diseases, but they have pleiotropic effects, one of which is to reduce inflammation. Their primary mechanism of action in lowering cholesterol is competitive inhibition of the enzyme 3-hydroxy-3-methylglutaryl coenzyme A (HMG-CoA) reductase which catalyses a rate-limiting step in cholesterol biosynthesis [92]. Mevalonate,

the product of HMG-CoA reductase, is also the the precursor of isoprenoid compounds that regulate immune activation [93]. In vitro studies suggest that statins alter neutrophil migration [94], and a human study on 16 individuals with dyslipidaemia showed that a year of statin therapy decreased the production of reactive oxygen species by neutrophils [95]. Simvastatin reduced T lymphocyte proliferation and natural killer cell cytotoxicity in a healthy volunteer study [96], and inhibited pro-inflammatory cytokines expressed by monocytes in patients with hypercholesterolaemia [97].

The anti-inflammatory effects of statins suggest that they may have benefit in non-atherosclerotic diseases. Atorvastatin has shown a clinical benefit in a randomised clinical trial for rheumatoid arthritis [98] and statins are now being trialled in patients with multiple sclerosis [99].

Another drug with potential anti-inflammatory properties is metformin, which reduced the neutrophil / lymphocyte ratio in patients with polycystic ovary syndrome [100] and reduced CRP and other inflammatory markers in patients with impaired fasting glucose [101].

Colchicine is a drug used for treating gout, and has an antitubulin effect that inhibits neutrophil function. In an observational study, colchicine use was associated with lower incidence of myocardial infarction among patients with gout [102], and a randomised clinical trial of colchicine 0.5mg/day reduced the rate of acute coronary syndrome or stroke among patients with stable coronary disease [103].

There are two clinical trials underway to determine whether reduction of inflammation can reduce the risk of cardiovascular events. The Canakinumab Anti-Inflammatory Thrombosis Outcomes Study (CANTOS) is evaluating whether interleukin-1β inhibition reduces cardiovascular events among patients with stable coronary artery disease and elevated CRP ($\geq$ 2 mg/L). The Cardiovascular Inflammation Reduction Trial is comparing low dose methotrexate with placebo among patients with a history of myocardial infarction and either diabetes or metabolic syndrome [53].

1.9 Areas of clinical uncertainty

Although there has been much interest in the inflammatory hypothesis of cardiovascular diseases (either as a component of atherosclerosis or having effects such as weakening blood vessel walls), epidemiological studies have focussed on a few areas such as prognosis of acute coronary syndromes or heart failure. There are still substantial gaps in our knowledge; the association between leukocyte counts and non-coronary atherosclerotic diseases such as stroke and abdominal aortic aneurysm have not been investigated extensively, and the strength of associations have not been compared between different cardiovascular phenotypes.

It is also unclear whether leukocyte counts add value in risk prediction. Total white blood cell count did not improve the prediction of cardiovascular death among patients with acute coronary syndrome when incorporated into the Global Registry of Acute Coro-

nary Events (GRACE) score [104]. However, a prediction model which incorporates all available leukocyte subtype counts has not been tested.

Although there is some evidence that drugs such as metformin [101], statins [91] and colchicine [103] may have beneficial effects on leukocyte counts, it is unclear how far this explains their biological effects and whether there is justification for further trials. The true effect size is unclear, as many of the published studies have been small and there is the potential for publication bias.

1.10 Aim and objectives

1.10.1 Aim

To develop methods to assist in using electronic health record databases for research, and apply them to an exemplar study investigating differential white cell counts and onset of cardiovascular diseases.

1.10.2 Objectives

1. Develop tools to assist in analysis of electronic health record data
2. Investigate the association of differential leukocyte count with incidence of cardiovascular diseases

1.10.3 Approach to fulfilling aim and objectives

The first part of this thesis focusses on the methods: the CALIBER database, the validity and provenance of the information within, and the process of converting raw data into a research-ready format. It is vital to assess the validity of the linked data in order to be able to use it appropriately for research (section 4.2). Identification of disease phenotypes is time-consuming and is best performed once, with re-usable algorithms. Software is required to facilitate this process, and I have developed a suite of R packages to do so (section 4.3). Missing data is a pervasive problem in electronic health record datasets but these datasets contain much information that can be used to assist in the imputation of missing data. However, conventional parametric imputation models are limited in the number of predictor variables they can handle, and they do not automatically account for non-linearities and interactions between the predictors. I therefore developed an imputation method using the Random Forest machine learning algorithm (section 4.4).

I describe the phenotype definitions for initial presentations of cardiovascular diseases in chapter 5, and demonstrate their association with type 2 diabetes. In the following chapter, chapter 6 I explore differential white blood cell counts recorded in CALIBER, and then investigate their associations with initial presentations of cardiovascular diseases in chapters 7 and 8.

I replicate an analysis of total white blood cell count and all-cause mortality in CALIBER and the New Zealand PREDICT database (chapter 9) and finally discuss the project as a whole.

2 Literature review

2.1 Chapter outline

This chapter describes a systematic literature search for cohort studies investigating the association between differential leukocyte counts (neutrophils, lymphocytes, monocytes, eosinophils, basophils) and incidence of any of the twelve cardiovascular diseases of interest in healthy populations.

2.2 Abstract

Background: There are biological mechanisms which suggest that leukocyte counts might be associated with onset of cardiovascular diseases in healthy populations. This literature review aims to identify prior cohort studies which have investigated such associations in order to inform further studies.

Methods: I searched EMBASE and MEDLINE using the OvidSP web interface for studies relating neutrophil, lymphocyte, monocyte, eosinophil, basophil, granulocyte or total white blood cell counts with incidence or prognosis of cardiovascular diseases, with over 2000 patients and follow-up of at least 1 year.

Results: The initial search returned 3099 articles, but on review only 15 publications reporting results of studies in 14 cohorts fulfilled the criteria for inclusion in this review. There was some evidence of a positive association of total white cell count and neutrophil count with coronary artery disease, but weak, conflicting or non-significant associations with other cardiovascular diseases or leukocyte subtypes. All studies retrieved investigated leukocyte counts taken from a consented cohort as part of a bespoke study.

Discussion: There were no large-scale population-based cohort studies investigating the associations with clinically recorded white cell counts. Neutrophil counts are likely to be associated with coronary artery disease, but associations with other cardiovascular diseases or leukocyte subtypes are unclear and warrant further research.

2.3 Introduction

White blood cells play a major role in atherosclerosis, which is an inflammatory process [48], and some cardiovascular diseases such as abdominal aortic aneurysms are a

result of other inflammatory processes which weaken blood vessel walls [44]. Although white blood cell (leukocyte) counts are included in the full blood count, one of the most commonly performed blood tests, these numbers are not currently used in cardiovascular decision making [30]. This is because the association of leukocyte counts and cardiovascular diseases is not well known. In contrast to lipids [105] and C reactive protein [106], there are no recent meta-analyses or individual participant consortia investigating differential leukocyte counts and cardiovascular diseases.

I therefore carried out a systematic review to test the hypothesis that leukocyte subtypes are differentially associated with onset of different cardiovascular diseases. I aimed to identify large cohort studies with long-term follow-up (at least 1 year) investigating the association of baseline leukocyte counts with incidence of cardiovascular diseases in healthy populations. This literature review would clarify gaps in current knowledge and help to inform the design of the studies to be performed as part of this PhD.

2.4 Methods

I searched EMBASE and MEDLINE using the OvidSP web interface (`https://ovidsp.ovid.com/`) for studies relating neutrophil, lymphocyte, monocyte, eosinophil, basophil, granulocyte or total white blood cell counts with incidence or prognosis of cardiovascular diseases (ischaemic heart disease, transient ischaemic attack, subarachnoid haemorrhage, ischaemic or haemorrhagic stroke, abdominal aortic aneurysm, peripheral vascular disease, heart failure, cardiac arrest or sudden cardiac death). I searched on phrases in the text or abstract and also used Medical Subject Headings (MeSH terms) as follows:

1. (leukocytes or eosinophils or neutrophils or lymphocytes or granulocytes or monocytes or basophils or "leukocyte count" or "lymphocyte count").mp. [mp=ti, ab, ot, nm, hw, kf, ps, rs, ui, sh, tn, dm, mf, dv, kw]
2. (("white blood cell" or "white blood cells" or "white cells" or "white cell" or leukocyte* or eosinophil* or neutrophil* or lymphocyte* or granulocyte* or monocyte* or basophil*) and (count or counts or total or differential or measur* or index or indices)).m_titl.
3. (occur* or association or incidence or onset).m_titl.
4. incidence.mp. [mp=ti, ab, ot, nm, hw, kf, ps, rs, ui, sh, tn, dm, mf, dv, kw]
5. (prognosis or mortality or survival).m_titl.
6. (prognosis or mortality or survival).mp. [mp=ti, ab, ot, nm, hw, kf, ps, rs, ui, sh, tn, dm, mf, dv, kw]
7. exp Myocardial Ischemia/
8. ("ischaemic heart" or "ischemic heart" or angina or "myocardial infarction" or "myocardial infarct" or "acute coronary syndrome" or "acute coronary syndromes" or

```
    "coronary disease" or "coronary artery disease" or
    "coronary heart disease").m_titl.
 9. exp heart arrest/
10. exp "tachycardia, ventricular"/
11. ("ventricular fibrillation" or "ventricular flutter").mp. [mp=ti, ab,
    ot, nm, hw, kf, ps, rs, ui, sh, tn, dm, mf, dv, kw]
12. ("heart arrest" or "cardiac arrest" or "sudden cardiac death" or
    ("ventricular" and ("flutter" or "fibrillation" or "tachycardia"))).m_titl.
13. exp "heart failure"/
14. ("heart failure" or "cardiac failure" or "ventricular failure").m_titl.
15. exp stroke/
16. exp "brain ischemia"/
17. (stroke or "cerebrovascular accident" or "cerebrovascular disease" or
    "brain ischemia" or "transient ischaemic attack" or
    "transient ischemic attack").m_titl.
18. ("cerebral hemorrhage" or "intracranial hemorrhage, hypertensive" or
    "basal ganglia hemorrhage" or "subarachnoid hemorrhage").mp. [mp=ti, ab,
    ot, nm, hw, kf, ps, rs, ui, sh, tn, dm, mf, dv, kw]
19. (("intracerebral" or "intra-cerebral" or "subarachnoid") and ("bleed"
    or "haemorrhage" or "hemorrhage")).m_titl.
20. "peripheral arterial disease".mp. [mp=ti, ab, ot, nm, hw, kf, ps, rs,
    ui, sh, tn, dm, mf, dv, kw]
21. (("peripheral arterial disease" or "peripheral vascular disease" or
    "intermittent claudication" or (leg or "lower limb")) and (ischemia
    or ischaemia or "arterial disease" or "vascular disease")).m_titl.
22. "aortic aneurysm, abdominal".mp. [mp=ti, ab, ot, nm, hw, kf, ps, rs,
    ui, sh, tn, dm, mf, dv, kw]
23. "abdominal aortic aneurysm".m_titl.
24. 1 or 2
25. 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or
    19 or 20 or 21 or 22 or 23
26. 24 and 25
27. 3 or 4 or 5 or 6
28. 26 and 27
29. remove duplicates from 28
30. limit 29 to human
```

This search returned 2403 articles when initially performed on 13 November 2013. When re-run on 23 March 2015, an additional 696 titles were retrieved.

On reviewing titles and abstracts I found 328 articles which may contain epidemiological studies of interest. I retrieved the full text of these articles and identified other relevant articles through searches of reference lists. As the purpose of this systematic review was

to identify research questions for the primary analysis, I did not attempt to contact study authors or identify unpublished work, and I did not perform a formal assessment of the quality of the studies.

I limited the results to studies in English with more than 2000 participants and follow-up of at least 1 year, which investigated leukocyte counts and incidence of cardiovascular disease in the general population or a healthy cohort. I collated the most highly adjusted measure of association (generally a hazard ratio) reported in each study. Where studies reported both a linear association and a comparison of categories (e.g. quartiles), I reported the comparison of categories as it would be easier to interpret.

The oldest study [107] reported a main analysis with log hazard ratios (beta coefficients) adjusted for age, sex and blood pressure, as well as a secondary analysis with further adjustment for smoking and cholesterol in a small subset of the patients. I used the results adjusted for age, sex and blood pressure because these included all the patients, and converted the effect size into a hazard ratio for 10^9/L higher leukocyte count (this involved dividing the effect size by 10, as it was expressed per 100 cells/mm^3 which is equivalent to per 10^8 cells/L higher leukocyte count).

As the studies were heterogenous (in the way they analysed leukocyte counts as categorical or continuous variables, in their choice and definitions of endpoints, and the level of adjustment for risk factors), it was not considered worthwhile to attempt a meta-analysis. I referred to the MOOSE guidelines [108] in writing this report.

2.5 Results

I found 15 publications reporting results of studies in 14 cohorts investigating total white cell counts and incidence of cardiovascular diseases. There were fewer studies reporting associations with the differential leukocyte count. A systematic review in 2004 found 3 studies investigating the differential leukocyte count and incidence of coronary heart disease [90]. A subsequent meta-analysis included unpublished data from additional cohorts [49], and along with my literature search this resulted in a total of 9 publications reporting on 8 cohorts.

All studies investigating differential leukocyte count investigated granulocyte or neutrophil counts (**Table 2.2 on page 47**). Most of these studies also investigated lymphocyte counts (**Table 2.4 on page 49**) and monocyte counts (**Table 2.5 on page 50**). Relatively few studies reported on eosinophil counts (**Table 2.6 on page 51**) or basophil counts (**Table 2.7 on page 51**).

All studies reported here were based on follow-up of consented investigator-led cohorts. None looked at more than one or two endpoints in the same study, or used clinically recorded leukocyte counts. One recent large study was based on routine records from a medical insurance cohort [109], but the white blood count was measured specifically for the purpose of the study, rather than as part of clinical care.

2.5.1 Total leukocyte count and incidence of cardiovascular diseases

There are 14 previous large cohort studies investigating total white cell counts and incidence of cardiovascular diseases (**Table 2.1 on page 44**). The majority of studies (9/14) focussed on the association with coronary disease, rather than other cardiovascular diseases. Five studies reported aggregate cardiovascular disease or cardiovascular mortality and only 6 studies reported non-coronary outcomes (4 reports for stroke, 1 for subarachnoid haemorrhage and 2 for heart failure).

Most studies reported a positive association between higher total leukocyte count and future risk of coronary artery disease, heart failure and cerebral infarction. One study reported a null association with haemorrhagic stroke [110]. Two studies reported hazard ratios by sex; in the Malmo Preventive Project there was no significant difference in the association between the sexes for stroke or coronary disease [111], but in the largest study (based in the Korean Medical Insurance scheme) the association with mortality from atherosclerotic disease was stronger in men than in women, and stronger among non-smokers than smokers [109]. Higher leukocyte count was also associated with increased risk of subarachnoid haemorrhage [40].

Table 2.1: Total white cell count and incidence of cardiovascular diseases: previous cohort studies with over 2000 participants

Endpoints other than coronary disease are shaded: pink for non-coronary endpoints and grey for cardiovascular disease aggregates.

Author, year, study	Population	Endpoint	N events	Adjusted measure of association
Jee, 2005, Korean Cancer Prevention Study [109]	438 500 government employees, teachers, and their dependants insured with Korea Medical Insurance Corporation	Mortality from atherosclerotic disease	11 676	HR (95% CI) for WBC ≥ 9 vs $< 5 \times 10^9$/L: male non-smokers 2.33 (1.71, 3.17) male smokers 1.44 (1.25, 1.67) female non-smokers 1.34 (1.17, 1.53) female smokers 1.20 (0.95, 1.52)
Li, 2010, Malmo Preventive Project [111]	17 131 men and 2932 women with average follow-up 24 years	MI or coronary death	2600	HR (95% CI) for top vs bottom quartile: men 1.31 (1.16, 1.48) women 1.46 (0.87, 2.46)
		Stroke	1333	HR (95% CI) for top vs bottom quartile: men 1.11 (0.93, 1.33) women 1.12 (0.63, 2.02)
Söderholm, 2014, Malmo Preventive Project [40]	17 083 men and 2711 women	Subarachnoid haemorrhage	95	HR 2.05 (95% CI 1.06, 3.96) for top vs bottom quartile

Author, year, study	Population	Endpoint	N events	Adjusted measure of association
Margolis, 2005, Women's Health Initiative Observational Study [112]	66 261 women with mean follow-up 6.1 years	Coronary death, MI or stroke	1510	HR (95% CI) for top vs bottom quartile: overall 1.47 (1.26, 1.72) non-Hispanic Whites 1.47 (1.25, 1.74) Blacks 1.52 (0.91, 2.53)
		Coronary death	187	HR 2.36 (95% CI 1.51, 3.68) for top vs bottom quartile
		Nonfatal MI	701	HR 1.41 (95% CI 1.12, 1.78) for top vs bottom quartile
		Stroke	738	HR 1.46 (95% CI 1.17, 1.81) for top vs bottom quartile
Pfister, 2012, EPIC-Norfolk [113]	16 011 people without prior MI, cancer, stroke or heart failure	Incident heart failure	935	HR 1.24 (95% CI 0.96, 1.60) for top vs bottom quartile
Wheeler, 2004, NHEFS [49]	4625 people with average follow-up 18 years	Hospitalised angina, MI or coronary death	914	HR 1.01 (95% CI 0.85, 1.20) for top vs bottom tertile
Tamakoshi, 2007, NIPPON DATA90 [114]	6756 Japanese residents followed up from nationwide random survey	Cardiovascular mortality	161	RR 1.79 (95% CI 0.97, 3.71) for WBC $9-10\times10^9$/L vs $4-4.9\times10^9$/L
Shankar, 2007, Blue Mountains Eye Study [115]	2904 people with no prior cardiovascular disease or cancer	Cardiovascular mortality	242	HR 2.01 (95% CI 1.40, 2.90) for top vs bottom quartile
Sweetnam, 1997, Caerphilly and Speedwell [116]	4615 men aged 45 to 63 living in two neighbourhoods	MI or coronary death	565	Relative odds 2.10 (95% CI 1.51, 2.92) for top vs bottom quintile
Lee, 2001, Atherosclerosis Risk in Communities (ARIC) [117]	13 555 men and women recruited from 4 US communities	MI or coronary death	488	RR 1.8 (95% CI 1.32, 2.43) for top vs bottom quartile
		Cardiovascular death	258	RR 2.3 (95% CI 1.58, 3.44) for top vs bottom quartile
		Stroke	220	RR 2.0 (95% CI 1.29, 2.99) for top vs bottom quartile
Engstrom, 2009, Malmo [118]	16 940 middle-aged men without prior MI or stroke	Hospitalised heart failure	436	HR 1.73 (95% CI 1.3, 2.3) for top vs bottom quartile
Kavousi, 2012, Rotterdam Study [119]	5933 people without prior coronary disease	MI or coronary death	347	HR 1.8 (95% CI 1.3, 2.5) for top vs bottom quartile
Twig, 2012, MELANY [45]	29 120 young men in the Israeli Army	Coronary artery disease on angiography	188	HR ratio 1.84 (95% CI 1.00, 3.37) for top vs bottom quintile

Author, year, study	Population	Endpoint	N events	Adjusted measure of association
Prentice, 1982, Hiroshima & Nagasaki [107]	12 858 people in Adult Health Study cohort	Angina, MI or coronary death	154	HR 1.09 per 10^9/L higher WBC (P = 0.02) adjusted for age, sex and blood pressure
Prentice, 1982, Hiroshima & Nagasaki [110]	13 040 people in Adult Health Study cohort	Cerebral infarction	336	HR 1.07 per 10^9/L higher WBC (P = 0.01)
		Intracerebral haemorrhage	73	HR 0.92 per 10^9/L higher WBC (P = 0.21)
Olivares, 1993, Paris Prospective Study II [120]	2856 White men with no previous coronary disease working for SNCF or civil service	Angina, MI or coronary death	46	RR 1.09 per 10^9/L higher WBC adjusted for age and smoking

2.5.2 Neutrophil counts and incidence of cardiovascular diseases

Ten publications reported associations of neutrophil counts with cardiovascular diseases in 8 cohorts (**Table 2.2 on page 47**). The majority of the studies found that higher neutrophil count was associated with increased risk of coronary artery disease. The two studies which investigated stroke subtypes showed that neutrophil counts were positively associated with cerebral infarction but not intracerebral haemorrhage [110, 122].

Smoking is strongly associated with neutrophil counts and with cardiovascular disease, but the association between cardiovascular disease and neutrophil count has been shown to be present among never-smokers as well as people with a smoking history [116].

Some studies investigated associations with granulocyte counts. Granulocytes include neutrophils, eosinophils, basophils and mast cells. The majority of granulocytes are neutrophils; hence the granulocyte count for any individual patient is numerically similar to the neutrophil count. Higher granulocyte counts were associated with a moderately increased risk of coronary disease and heart failure (**Table 2.3 on page 48**).

2.5.3 Lymphocyte counts and incidence of cardiovascular diseases

Lymphocyte counts showed no consistent association with onset of coronary disease, heart failure or stroke in 9 cohorts (**Table 2.4 on page 49**). Most of the associations reported were not significantly different from null.

2.5.4 Monocyte counts and incidence of cardiovascular diseases

Previous studies investigating monocyte counts and onset of cardiovascular diseases suggest a weak association with higher monocyte counts in some cohorts, but the only statistically significant result was in the Paris Prospective Study II [120] (**Table 2.5 on page 50**). There were no significant associations with heart failure or stroke. In the

Table 2.2: Neutrophil counts and incidence of cardiovascular diseases: previous cohort studies with over 2000 participants

Author, year	Study population	N patients	Follow up	Endpoint	N events	Adjusted measure of association
Adamsson Eryd, 2012 [121]	Malmo Diet and Cancer Study	27 085	Mean 13.6 years	Coronary death or MI	1965	HR 1.13 (95% CI 1.08, 1.18) per SD (1.33 $\times 10^9$/L) higher
Zia, 2012 [122]	Malmo Diet and Cancer Study	26 927	Mean 13.6 years	Cerebral infarction	1314	HR 1.3 (95% CI 1.1, 1.5) for top vs bottom quartile
				Intracerebral haemorrhage	201	HR 0.8 (95% CI 0.5, 1.2) for top vs bottom quartile
Wheeler, 2004 [49]	ARIC	11 305	Mean 10 years	Coronary death or MI	527	RR 2.0 for top vs bottom tertile (read off graph in meta-analysis)
Shah, 2014 [123]	National Health and Nutrition Examination Survey-III	7250	Mean 14.1 years	Nonfatal MI or coronary death	231	HR 1.57 (95% CI 0.72, 3.44) for $>$ 7.7 vs $\leq 7.7 \times 10^9$/L
Gillum, 2005 [124]	NHEFS	4625	Median 18.3 years	Coronary artery disease or coronary death	914	HR 1.09 (0.93, 1.29) for top vs bottom tertile
Olivares, 1993 [120]	Paris Prospective Study II	3659	Mean 66.5 months	Coronary heart disease	46	RR 1.03 (95% CI 0.88, 1.19) per 10^9/L higher
Karino, 2015 [125]	Honolulu Heart Program	2879	8 years	ACS or coronary death	279	HR 1.57 (95% 1.57, 2.34) for top vs bottom quartile (P for trend 0.02)
Prentice, 1982 [107]	Hiroshima and Nagasaki	12 858	Mean 11 years	Angina, MI or coronary death	154	HR 1.12 per 10^9/L higher (P = 0.05)
Prentice, 1982 [110]	Hiroshima and Nagasaki	13 040	Mean 11 years	Cerebral infarction	336	HR 1.14 per 10^9/L higher (P <0.001)
				Intracerebral haemorrhage	73	HR 0.95 per 10^9/L higher (P = 0.61)
Sweetnam, 1997 [116]	Caerphilly (men only)	2163	5 years	Coronary death or MI	143	HR 3.54 for top vs bottom quintile (P for trend 0.003)

ACS = acute coronary syndrome; CI = confidence interval; HR = hazard ratio; MI; myocardial infarction; RR = relative risk.

Table 2.3: Granulocyte counts and incidence of cardiovascular diseases: cohort studies with over 2000 participants

Author, year	Study population	N patients	Follow up	Endpoint	N events	Adjusted measure of association
Bekwelem, 2011 [126]	ARIC	14 485	Mean 15.5 years	Heart failure	1647	HR 2.19 (95% 1.83, 2.61) for top vs bottom quintile
Pfister, 2012 [113]	EPIC-Norfolk	16 011	Mean 12.4 years	Heart failure	935	HR (95% CI) per 10^9/L higher: men 1.16 (1.09, 1.24) women 1.05 (0.97, 1.13)
Karino, 2015 [125]	Honolulu Heart Program	2879	8 years	ACS or coronary death	279	HR 1.66 (95% CI 1.11, 2.48) for top vs bottom quartile (P for trend 0.01)

ACS = acute coronary syndrome; CI = confidence interval; HR = hazard ratio; MI; myocardial infarction; RR = relative risk.

ARIC study the associations were not reported directly but had to be read off a graph or otherwise deduced from information in the paper, so I was unable to ascertain confidence intervals [49, 126].

2.5.5 Eosinophil counts and incidence of cardiovascular diseases

The Hiroshima and Hagasaki study [107] and Caerphilly study [116] showed that higher eosinophil count was associated with greater incidence of coronary heart disease (**Table 2.6 on page 51**). There was no significant association with ischaemic or haemorrhagic stroke in the Hiroshima and Hagasaki study [110].

2.5.6 Basophil counts and incidence of cardiovascular diseases

I found only three studies investigating the association of basophils with cardiovascular disease in the healthy population (**Table 2.7 on page 51**), and there were no statistically significant associations reported with coronary disease or stroke.

2.6 Discussion

This literature review did not find any large-scale population-based cohort studies investigating the association of *clinically recorded* white cell counts with incidence of cardiovascular diseases. There was a study in a Korean medical insurance database involving over 400 000 participants with over 10 000 events, but even in this study the leukocyte counts were obtained specifically for the study [109]. The lack of previous studies is surprising, as white cell counts are commonly performed and frequently recorded electronically, but

Table 2.4: Lymphocyte counts and incidence of cardiovascular diseases: previous cohort studies with over 2000 participants

Author, year	Study population	N patients	Follow up	Endpoint	N events	Adjusted measure of association
Adamsson Eryd, 2012 [121]	Malmo Diet and Cancer Study	27 085	Mean 13.6 years	Coronary death or MI	1965	HR 1.05 (95% CI 1.00, 1.11) per SD (0.88 $\times 10^9$/L) higher
Zia, 2012 [122]	Malmo Diet and Cancer Study	26 927	Mean 13.6 years	Cerebral infarction	1314	HR 1.2 (95% CI 0.98, 1.4) for top vs bottom quartile
				Intracerebral haemorrhage	201	HR 0.7 (95% CI 0.4, 1.01) for top vs bottom quartile
Pfister, 2012 [113]	EPIC-Norfolk	16 011	Mean 12.4 years	Heart failure	935	HR (95% CI) per 10^9/L higher: men 0.97 (0.83, 1.13) women 0.92 (0.77, 1.10)
Bekwelem, 2011 [126]	ARIC	14 485	Mean 15.5 years	Heart failure	1647	HR 0.86 for top vs bottom quintile
Wheeler, 2004 [49]	ARIC	11 337	Mean 10 years	Coronary death or MI	531	RR 1.2 for top vs bottom tertile (read off graph in meta-analysis)
Shah, 2014 [123]	National Health and Nutrition Examination Survey-III	7250	Mean 14.1 years	Nonfatal MI, fatal CHD	231	HR 1.33 (95% CI 0.81, 2.18) for $>$ 1.5 vs $\leq 1.5 \times 10^9$/L
Gillum, 2005 [124]	NHEFS	4625	Median 18.3 years	Coronary artery disease or coronary death	914	HR 1.05 (0.89, 1.24) for top vs bottom tertile
Olivares, 1993 [120]	Paris Prospective Study II	3659	Mean 66.5 months	Coronary heart disease	46	RR 1.06 (95% CI 0.73, 1.55) per 10^9/L higher
Karino, 2015 [125]	Honolulu Heart Program	2879	8 years	ACS or coronary death	279	HR 1.14 (95% 0.80, 1.62) for top vs bottom quartile (P for trend 0.94)
Prentice, 1982 [107]	Hiroshima and Nagasaki	12 858	Mean 11 years	Angina, MI or coronary death	154	HR 1.05 per 10^9/L higher (P = 0.62)
Prentice, 1982 [110]	Hiroshima and Nagasaki	13 040	Mean 11 years	Cerebral infarction	336	HR 1.01 per 10^9/L higher (P = 0.89)
				Intracerebral haemorrhage	73	HR 0.73 per 10^9/L higher (P = 0.07)
Sweetnam, 1997 [116]	Caerphilly (men only)	2163	5 years	Coronary death or MI	143	HR 1.08 for top vs bottom quintile (P for trend 0.21)

CI = confidence interval; HR = hazard ratio; MI; myocardial infarction; RR = relative risk.

Table 2.5: Monocyte counts and incidence of cardiovascular diseases: previous cohort studies with over 2000 participants

Author, year	Study population	N patients	Follow up	Endpoint	N events	Adjusted measure of association
Pfister, 2012 [113]	EPIC-Norfolk	16 011	Mean 12.4 years	Heart failure	935	HR (95% CI) per 10^9/L higher: men 0.71 (0.53, 0.93) women 1.10 (0.82, 1.48)
Bekwelem, 2011 [126]	ARIC	14 485	Mean 15.5 years	Heart failure	1647	HR 0.89 for top vs bottom quintile
Wheeler, 2004 [49]	ARIC	11 293	Mean 10 years	Coronary death or MI	527	RR 1.3 for top vs bottom tertile (read off graph in meta-analysis)
Prentice, 1982 [107]	Hiroshima and Nagasaki	12 858	Mean 11 years	Coronary heart disease	153	RR 1.26 (P = 0.26) for monocytes $\geq$ 0.6 $\times 10^9$/L vs <0.36 $\times 10^9$/L
Prentice, 1982 [110]	Hiroshima and Nagasaki	13 040	Mean 11 years	Cerebral infarction	336	HR 0.87 per 10^9/L higher (P = 0.58)
				Intracerebra haemorrhage	73	HR 0.43 per 10^9/L higher (P = 0.16)
Gillum, 2005 [124]	NHEFS	4625	Median 18.3 years	Coronary artery disease or coronary death	914	HR 0.96 (0.82, 1.13) for top vs bottom tertile
Olivares, 1993 [120]	Paris Prospective Study II	3659	Mean 66.5 months	Coronary heart disease	46	RR 1.15 (95% CI 1.07, 1.24) per 10^8/L higher
Karino, 2015 [125]	Honolulu Heart Program	2879	8 years	ACS or coronary death	279	HR 1.22 (95% 0.86, 1.74) for top vs bottom quartile (P for trend 0.28)
Sweetnam, 1997 [116]	Caerphilly (men only)	2163	5 years	Coronary death or MI	143	HR 1.34 for top vs bottom quintile (P for trend 0.22)

CI = confidence interval; HR = hazard ratio; MI; myocardial infarction; RR = relative risk.

Table 2.6: Eosinophil counts and incidence of cardiovascular diseases: previous cohort studies with over 2000 participants

Author, year	Study population	N patients	Follow up	Endpoint	N events	Adjusted measure of association
Prentice, 1982 [107]	Hiroshima and Nagasaki	12 858	Mean 11 years	Coronary heart disease	153	RR 1.78 (P = 0.01) for eosinophils $\geq$ 0.3 $\times 10^9$/L vs <0.12 $\times 10^9$/L
Prentice, 1982 [110]	Hiroshima and Nagasaki	13 040	Mean 11 years	Cerebral infarction	336	HR 0.92 per 10^9/L higher (P = 0.79)
				Intracerebral haemorrhage	73	HR 1.14 per 10^9/L higher (P = 0.84)
Olivares, 1993 [120]	Paris Prospective Study II	3659	Mean 66.5 months	Coronary heart disease	46	RR 1.03 (95% CI 0.87, 1.21) per 10^8/L higher
Sweetnam, 1997 [116]	Caerphilly (men only)	2163	5 years	Coronary death or MI	143	HR 2.15 for top vs bottom quintile (P for trend 0.05)

CI = confidence interval; HR = hazard ratio; MI; myocardial infarction; RR = relative risk.

Table 2.7: Basophil counts and incidence of cardiovascular diseases: previous cohort studies with over 2000 participants

Author, year	Study population	N patients	Follow up	Endpoint	N events	Adjusted measure of association
Prentice, 1982 [107]	Hiroshima and Nagasaki	12 858	Mean 11 years	Angina, MI or coronary death	154	HR 1.12 per 10^9/L higher (P = 0.53)
Prentice, 1982 [110]	Hiroshima and Nagasaki	13 040	Mean 11 years	Cerebral infarction	336	HR 1.54 per 10^9/L higher (P = 0.69)
				Intracerebral haemorrhage	73	HR 4.13 per 10^9/L higher (P = 0.51)
Olivares, 1993 [120]	Paris Prospective Study II	3659	Mean 66.5 months	Coronary heart disease	46	RR 1.14 (95% CI 0.94, 1.38) per 10^8/L higher
Sweetnam, 1997 [116]	Caerphilly (men only)	2163	5 years	Coronary death or MI	143	HR 1.50 for top vs bottom quintile (P for trend 0.36)

CI = confidence interval; HR = hazard ratio; MI; myocardial infarction; RR = relative risk.

perhaps attention has been diverted in recent years to measurement of acute phase proteins, cytokines and other novel inflammatory markers instead. There were no published results of recent meta-analyses or large-scale consortia.

The only cardiovascular diseases investigated for their associations with the leukocyte differential were coronary artery disease, cerebral infarction, intracerebral haemorrhage and heart failure. There were no studies investigating other cardiovascular diseases responsible for significant morbidity and mortality such as abdominal aortic aneurysm or peripheral arterial disease, and only one study (with just 95 events) investigating subarachnoid haemorrhage, a devastating but poorly understood condition.

Previous studies have generally not studied more than one endpoint in the same analysis, so the relative strengths of associations with different cardiovascular presentations is not known. The findings had varying levels of adjustment and some studies were not able to adjust for smoking (e.g. in the Hiroshima and Nagasaki study, where only a limited subset of participants had smoking status recorded [107]). Therefore the finding of increased cardiovascular risk with higher eosinophil count may be confounded; indeed studies in other settings such as patients undergoing percutaneous intervention [127] and patients with acute heart failure [128], *low* eosinophil count was found to be associated with worse prognosis.

2.6.1 Limitations

This literature review may have missed some relevant studies, because some relative risks of interest may be reported only in the text of articles if they are not the main focus of the paper (e.g. if they are in a multivariable model), and thus may not be flagged in the title. Other limitations of this review are that I did not attempt to contact authors for unpublished studies and I limited the review to large studies published in English. This results in a restricted and less systematic review of the literature than would be optimal, but the larger studies are likely to be of higher quality and and are more likely to show a true association if one exists. In any case the purpose of the review was to highlight important research questions and verify that they have not been answered previously; the literature is so scarce that the new CALIBER study is fully justified.

2.6.2 Conclusion

This literature review has highlighted important gaps in our knowledge of associations of leukocyte counts with onset of cardiovascular diseases. Knowledge of these associations can provide insight into aetiological mechanisms and guide potential future therapies.

This justifies the development of the CALIBER resource for the investigation of the differential leukocyte count and onset of cardiovascular diseases in a large, contemporary population-based cohort. The next few chapters describe the data sources that comprise CALIBER, development of research tools for CALIBER and the definition of a cohort for answering the research question.

3 The CALIBER database

3.1 Chapter outline

This chapter describes the CALIBER programme, which is the data source for the majority of the work in this thesis, and the following chapter (chapter 4 on page 65) describes my contributions to validation and enhancement of CALIBER.

3.2 Introduction

CALIBER is a database of linked routinely collected electronic health records from England [1], comprising data from primary care (Clinical Practice Research Datalink, CPRD) [129], hospital admissions [130], the Myocardial Ischaemia National Audit Project (MINAP) [9] and the national death registry at the Office for National Statistics (ONS). The data sources complement each other in providing different types of information about a patient's medical history (**Figure 3.1 on page 54**). CALIBER also contains small-area indices of deprivation from ONS (index of multiple deprivation) [131] linked by the patient's postcode. The index of multiple deprivation is a score calculated for each patient's neighbourhood on the basis of social indices such as income, education, and employment. The CALIBER dataset has pseudonymised identifiers and does not identify the location of a patient or general practice except at a very crude level (one of 10 regions in England).

3.3 Source datasets

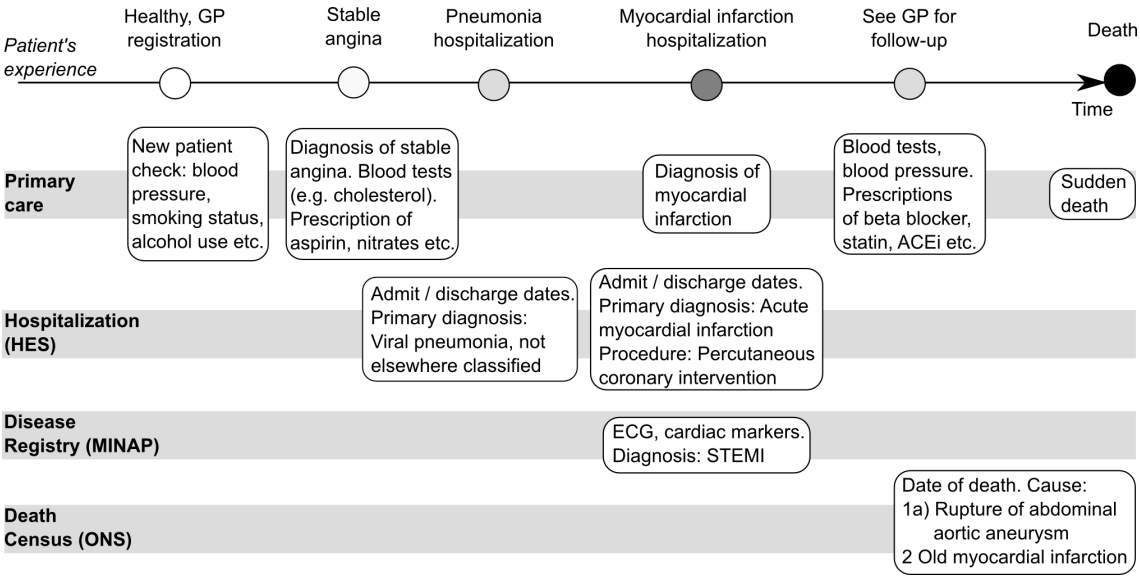
3.3.1 Primary care data: Clinical Practice Research Datalink

The Clinical Practice Research Datalink (CPRD) contains pseudonymised primary care electronic patient records from UK general practices, currently covering 6.9% of the UK population with 4.4 million currently registered patients [129]. CPRD has been used extensively for epidemiological research with over a thousand research studies published in a wide range of disease areas [129].

History of the CPRD

The original name for what would become the CPRD was the 'VAMP research databank' (Value Added Medical Products). General practitioners were given free computers in return for contributing data to the research databank. The VAMP Medical clinical system used the OXford Medical Information System (OXMIS) dictionary to encode clinical entries in order to minimise the use of hard disk space. Anonymised patient records were

Figure 3.1: How a patient's medical history may be recorded in the CALIBER data sources



collected at regular intervals from the practices, initially on tape and later electronically. The text-based VAMP Medical software was replaced by Vision, which had a graphical user interface, and practices switched from OXMIS to the Read Clinical Terminology at varying dates in the 1990s.

The research database was renamed the General Practice Research Database (GPRD) in 1993 [129]. It was transferred to the Office for National Statistics and subsequently to the Medicines and Healthcare products Regulatory Agency (MHRA) around the year 2000. A subset of GPRD practices in England consented to record linkage with other datasets, and were linked to Hospital Episode Statistics and the death registry. It was this subset that was further linked to MINAP to constitute the CALIBER dataset.

The recent change of name to 'Clinical Practice Research Datalink' reflects the aspiration of further linkages with other registries, clinical datasets and cohort studies in the UK [129]. The current CPRD contains only Read terms, with OXMIS terms having been converted to the Read equivalents. CPRD has now developed a system for collecting data from EMIS systems, which is the most commonly used GP system supplier in the UK [132].

Validity and representativeness of CPRD

Many studies using the CPRD performed validation of the coded diagnoses against anonymised paper records requested from the GP or anonymised electronic free text. A systematic review of CPRD validation studies found that diagnoses were generally reliable [133]. However, because the only information in the database is that recorded during usual clinical care, clinical parameters or tests would not be recorded in individuals for whom it is not indicated.

Patients can opt out of data collection by informing their GP, but the majority of patients in CPRD practices contribute data, and CPRD patients are broadly representative of the UK population in terms of age, sex [129] and ethnicity [134]. However the practices in CPRD may not be representative of all practices in the UK by geographic distribution or size [135], and it is likely that the quality of data is also not representative, as CPRD practices are given data recording guidelines and regular feedback on the completeness and quality of data.

Structure of information recorded in CPRD

Information in the CPRD is recorded in a number of tables, which can be linked by the pseudonymised patient identifier in order to build up a complete picture of a patient's healthcare experience.

Patients – one row per patient, with demographic details such as year of birth, date of death and registration dates.

Practices – one row per practice, giving details such as region of the UK and the date when the practice achieved a good standard of data completeness for research purposes ('up-to-standard date').

Consultations – each patient episode is considered a 'consultation' and all data are entered in consultations (face-to-face, telephone or administrative). This table allows diagnoses and prescriptions entered in the same consultation to be identified. It can also be useful for measuring healthcare utilisation and costs, and for linking to the staff table.

Staff – one row per staff member, with gender and role.

Events – there are a number of event tables, and a patient can have any number of events. Each event is linked to a single consultation and has an event date, a medical dictionary code (Read code) or product dictionary code (Multilex) and associated information.

- **Clinical** – Read coded diagnoses entered by the GP, additional data such as blood pressure measurements
- **Referrals** – referrals to secondary care, with the indication recorded as a Read code
- **Immunisations** – records of immunisations
- **Therapy** – prescriptions
- **Test** – results of laboratory tests, each with a Read code

The type of information in the Test and Clinical tables is specified by the 'entity code' of each record. This defines what additional data are included with the record; for example

a haemoglobin (blood test) result will have a value, units and a normal range reported, whereas a blood pressure record will contain systolic and diastolic measurements and a record of the Korotkoff sound. Entity codes usually, but not always, correspond to Read codes; usually the Read code is more specific (as there are thousands of Read codes but only a few hundred entity codes). The entity codes can also be thought of as referring to different information models, or 'archetypes', which I discuss further in the last chapter (section 10.1.2 on page 235).

Read terms were designed for coding by GPs and incorporate synonymous terms with variations in the way doctors may express common diagnoses. Apart from diagnoses, the Read terminology includes codes for other categories of information such as history, examination findings, procedures and test results [136].

Considerations when using general practice data

UK general practice databases have the advantage that individuals are registered with a general practitioner (GP) over a defined time period, and the majority of the population are registered. This provides a denominator and enables the database to be used for estimating population incidence and prevalence of diseases [137, 138]. These estimates are based on the assumption that the research database population is representative and the condition of interest is completely recorded in general practice data.

Patients can only register permanently with a single GP in the UK, so when constructing a general practice database cohort, patients with temporary registration should be ignored.

3.3.2 Hospital Episode Statistics

Hospital Episode Statistics (HES) is the database of hospital admissions in England [130]. It contains coded information about each admission such as whether it was an emergency or elective admission, the specialty, dates of admission and discharge, and diagnoses (primary and secondary) coded using the International Classification of Diseases, 10th Edition (ICD-10) [139]. HES also contains records of procedures, coded using the Office of Population Censuses and Surveys codes (OPCS-4). Data are entered by clinical coding clerks at the end of an admission. Outpatient information is available in the main HES database but has not been linked in CALIBER.

HES provides anonymised datasets for research, with patients linked between admissions and hospitals by an algorithm involving NHS number, date of birth, sex, postcode, hospital code and local patient identifier (hospital number) [140]. Episodes which cannot be linked will be considered to belong to a different patient. The completeness of the linkage is higher for recent years as a greater proportion of records have an NHS number (97% of inpatient activity in 2007–2008, compared to just 83% in 2000–2001).

Numerous epidemiological studies have been performed using HES, and validation studies found that for severe vascular disease, including myocardial infarction, stroke and

pulmonary embolism, HES records appeared to be both reliable and complete [141]. However, studies must exercise caution because it is an administrative dataset. The utility of HES for research is limited by the granularity of the data and the limitations of the ICD-10 coding system. For example, there are no specific codes for ST elevation (STEMI) and non-ST elevation myocardial infarctions (NSTEMI), although the probability the the event is coded using particular codes is different between STEMI and NSTEMI (**Table 4.3 on page 74**). Investigations and minor procedures may be inconsistently recorded. Uncommon or unusual conditions such as recreational drug toxicity may not be consistently recorded [142].

3.3.3 Acute coronary syndrome registry: MINAP

MINAP (the Myocardial Ischaemia National Audit Project) is the national registry of acute coronary syndromes, to which all acute hospitals in England and Wales contribute data. The primary purpose of MINAP was to improve quality of care but it has also become invaluable for research. MINAP comprises one entry per acute coronary syndrome admission, with 123 fields containing detailed clinical information such as coded electrocardiogram (ECG) findings, troponin results and interventions performed. Information is abstracted from clinical notes by audit nurses or clerks [9].

MINAP contains rich information on the clinical characteristics and treatment of patients with acute coronary syndromes; for example it includes information on smoking status and ECG findings which are not recorded electronically in most hospital systems. However, as it is a voluntary registry there may be some selection bias in patients that are included, and this can vary by hospital. MINAP is thought to particularly under-report non-ST elevation myocardial infarction, but some myocardial infarctions are included twice if the patient was transferred between hospitals acutely [9].

The information in MINAP is not used for direct clinical care, but as a clinical audit it has been used to drive improvements in quality of care. Availability of MINAP data to researchers is governed by the MINAP Academic Group [9].

3.3.4 Death registry

The death registry for England and Wales curated by the Office for National Statistics (ONS) includes the date of death and the causes entered on the death certificate. A single underlying cause of death is allocated according to the WHO ICD-10 algorithm based on the information recorded on the death certificate, likely causal sequence and International Classification of Diseases selection rules [143].

Deaths in England and Wales have been coded using ICD-10 since 2001 [139] and ICD-9 in previous years. There was a change in the rules for selecting the underlying cause from ICD-9 to ICD-10 which means that causes of death are not directly comparable between 2001 onwards and previous years. However, the effect of these changes varied by cause of death; some causes of death such as pnemonia were particularly

affected but there was little effect on ischaemic heart disease and cerebrovascular disease [143].

3.3.5 Deprivation index

The index of multiple deprivation is a composite measure of deprivation calculated using indicators in the following domains:

- Income deprivation
- Employment deprivation
- Health deprivation and disability
- Education, skills and training deprivation
- Barriers to housing and services
- Crime
- Living environment deprivation

The index is calculated for super output areas (postcode areas) and is available from the Office for National Statistics. In CALIBER it was linked to patient records via the postcode, but postcode is not included in the final pseudonymised CALIBER dataset [131].

3.4 Linkage

CALIBER contains data from 244 CPRD general practices in England which consented for their data to be linked to other sources. The linkage was carried out in October 2010 by a Trusted Third Party, using a deterministic match between NHS number, date of birth and gender. 96% of patients with a valid NHS number were successfully matched. Unfortunately the CALIBER group does not have access to further information about the linkage which would ideally be provided, such as the linkage criteria used for matching each pair of records or other measures of match quality.

3.4.1 Recording of death in linked data

As a simple check for errors in the CALIBER linkage I compared the recording of death in CPRD and ONS, as both data sources should record the date of death. Death can be recorded in CPRD as a Read code for death, transfer out of the practice with a reason of death or a structured death notification entry.

For this investigation I used a sample of patients with unstable angina or myocardial infarction recorded in any of the CALIBER data sources, and limited the set of patients to those who had a record of death in ONS or CPRD while they were actively registered with a CPRD practice.

Of the 23 474 patients with a record of death in either CPRD or ONS, 22 975 (97.9%) had a record in both sources, 422 (1.80%) had a record in CPRD but not ONS and 77

Table 3.1: Days difference between dates of death in death registry and primary care for patients with acute coronary syndromes in CALIBER

Number of days death registry date is before primary care death date	Frequency	Percentage (95% CI)
>30 days before	122	0.53 (0.44, 0.64)
8–30 days before	99	0.43 (0.35, 0.53)
1–7 days before	475	2.07 (1.89, 2.26)
Same day	18 363	80.0 (79.4, 80.4)
1–7 days after	1926	8.38 (8.03, 8.75)
8–30 days after	1728	7.52 (7.19, 7.87)
>30 days after	262	1.14 (1.01, 1.29)

(0.33%) had a record in ONS but not CPRD. I compared the difference in the date of death between the two data sources (**Table 3.1 on page 59**).

This shows that the vast majority of deaths in CALIBER are recorded in both the death registry and primary care. Deaths might not be recorded in the death registry if the patient died abroad, or there was a linkage problem (missed match). Deaths might be recorded in the death registry but not primary care if the GP was not informed of the death or there was a false match on linkage.

The date of death in primary care data was usually on the same day as the death registry date of death, rarely before, but was up to 30 days after in almost 16% of patients. The date of death might be later in CPRD than the death registry if the general practitioner records the death as occurring on the date he or she finds out rather than the actual date of death. If the date of death is earlier in CPRD than in the death registry it suggests that the CPRD date of death is incorrect – there were a few deaths which were out by exactly 1 year, suggesting a data entry error.

3.5 Access to CALIBER data

Access to CALIBER data operates by a 'safe haven' model, where researchers have access to the data only under the terms of a collaborative project with the CALIBER group at UCL. Researchers have to be physically located at University College London, the London School of Hygiene and Tropical Medicine or another of the collaborating institutions in order to access the data. The raw data are stored in a MySQL database which is accessible only to the data manager.

The CALIBER database is stored with a pseudonymised identifier linking records from the same patient in the different datasets. CALIBER has defined a collection of variables and clinical phenotypes which have been converted to SQL scripts for rapid extraction from the master database.

CALIBER data for a specific project can be extracted either in raw form or as derived CALIBER variables depending on the researcher's preference, and is limited to the cohort

of patients used for that project. For the work described in this thesis I required both the raw form (for a subset of patients) to explore the data, and derived CALIBER variables for ease of processing datasets with millions of patients.

3.6 Ethics and approvals

Ethics and governance. CPRD was set up as a research database with ethical approval, and individual studies using CPRD data require local CPRD approval but not full ethical review. Data collection and re-use is permitted as the data are anonymous to researchers, with identifiers such as date of birth, name and address removed. The free text associated with coded data is not routinely available to researchers, although it can be requested (with a cost for manual anonymisation) and has been used for validation studies [18], but may not be routinely collected in the future.

Use of identifiable data (e.g. for linkage) requires application for an exemption under section 251 of the NHS Act 2006. This legislation permits the Secretary of State for Health to make regulations to set aside the common law duty of confidentiality for defined medical purposes [144]. Identifiable data should only be used where essential and would be accessible only to those who need it, i.e. researchers performing the linkage.

Patients in participating practices are opted in for data collection by default but have the opportunity to inform their GP that they wish their data not to be uploaded. The number of patients that opt out is small, so CPRD patients are broadly representative of the general population [129].

The CALIBER dataset comprises CPRD data linked to HES, ONS and MINAP by a trusted third party [1] and the final dataset is held in a pseudonymised form. Although direct identifiers such as name and date of birth are not contained within the data, the amount of information about individual patients is quite detailed so it is treated as sensitive. CALIBER researchers are provided with pseudonymised data and have to commit to not disclosing any information that may be able to identify a patient. Datasets should be stored and analysed on secure servers and only aggregate data should be exported or published.

CPRD has Multi-centre Research Ethics Committee approval for all purely observational research using CPRD data. The CALIBER record linkage study has had separate ethical approval (09/H0810/16). Individual studies using CALIBER data have been approved by the CPRD Independent Scientific Advisory Committee (ISAC) and the MINAP Academic Group (MAG). CALIBER also has a Scientific Oversight Committee which approves researchers' analytic protocols to ensure that the research being performed using the dataset is appropriate. It is CALIBER policy that analytic protocols have to be circulated among the group and approved, and the study registered with `www.clinicaltrials.gov` before data are released to the researcher.

The validation study on myocardial infarction in CALIBER (section 4.2) was approved by ISAC 'Comparison of the information recorded in MINAP, GPRD and HES', protocol

11_088, and is registered on clinicaltrials.gov (unique identifier NCT01569139, https://www.clinicaltrials.gov/ct2/show/NCT01569139).

The study investigating the association of type 2 diabetes with initial presentation of cardiovascular diseases (chapter 5) falls within the remit of the ISAC: 'Initial presentation of cardiovascular diseases' protocol 12_153R, and is registered on clinicaltrials.gov (unique identifier NCT01804439, https://www.clinicaltrials.gov/ct2/show/NCT01804439).

The studies investigating the association of white cell counts with initial presentation of cardiovascular diseases and all-cause mortality (chapters 6, 7, 8 and 9) are registered on clinicaltrials.gov (unique identifier unique identifier NCT02014610, https://www.clinicaltrials.gov/ct2/show/NCT02014610), and also fall within the remit of the ISAC: 'Initial presentation of cardiovascular diseases' protocol 12_153R.

The study on missing data methods (section 4.4) falls within the remit of ISAC 'Cardiometabolic biomarkers in the prognosis of specific coronary disease phenotypes', protocol 10_160. It is not registered on clinicaltrials.gov because it is a methodological study.

3.7 CALIBER data portal

The CALIBER data portal (**Figure 3.2 on page 62**) is a web portal for researchers to access descriptions of contributed CALIBER variable definitions (phenotypes), the underlying algorithms and lists of Read, ICD-10 or OPCS codes used to define research variables ('codelists'). Each phenotype page has a link to web pages describing each of the component variables. Flow diagrams are used where necessary to help users understand the creation of complex phenotypes (example for stable angina: **Figure 3.3 on page 63**). These show how different data sources are combined, and how diagnoses can be inferred if there are records of relevant investigations or procedures without an explicit diagnosis (for example, coronary artery bypass grafting or coronary angioplasty imply a diagnosis of stable angina if there is no recent record of acute coronary syndrome).

Information on the CALIBER data portal is licensed under a Creative Commons non-commercial licence. Algorithms are freely accessible on the website but users have to register to access the codelists. There is no charge for registration because its purpose is purely for tracking the use of the site.

Underlying the data portal is a GitHub repository, which stores the codelists and SQL code for querying the database in order to extract the variable. GitHub (https://github.com/) is a website which can host a set of files under control of the Git distributed version control system. Git tracks changes to multiple files and stores a list of user-written commit messages alongside a record of the changes. This makes it possible to view the changes in a file over time, and revert files to any previous state. It can merge changes from different versions of files, which enables multiple users to work on the same set of files simultaneously.

Figure 3.2: Screenshot of the CALIBER web portal

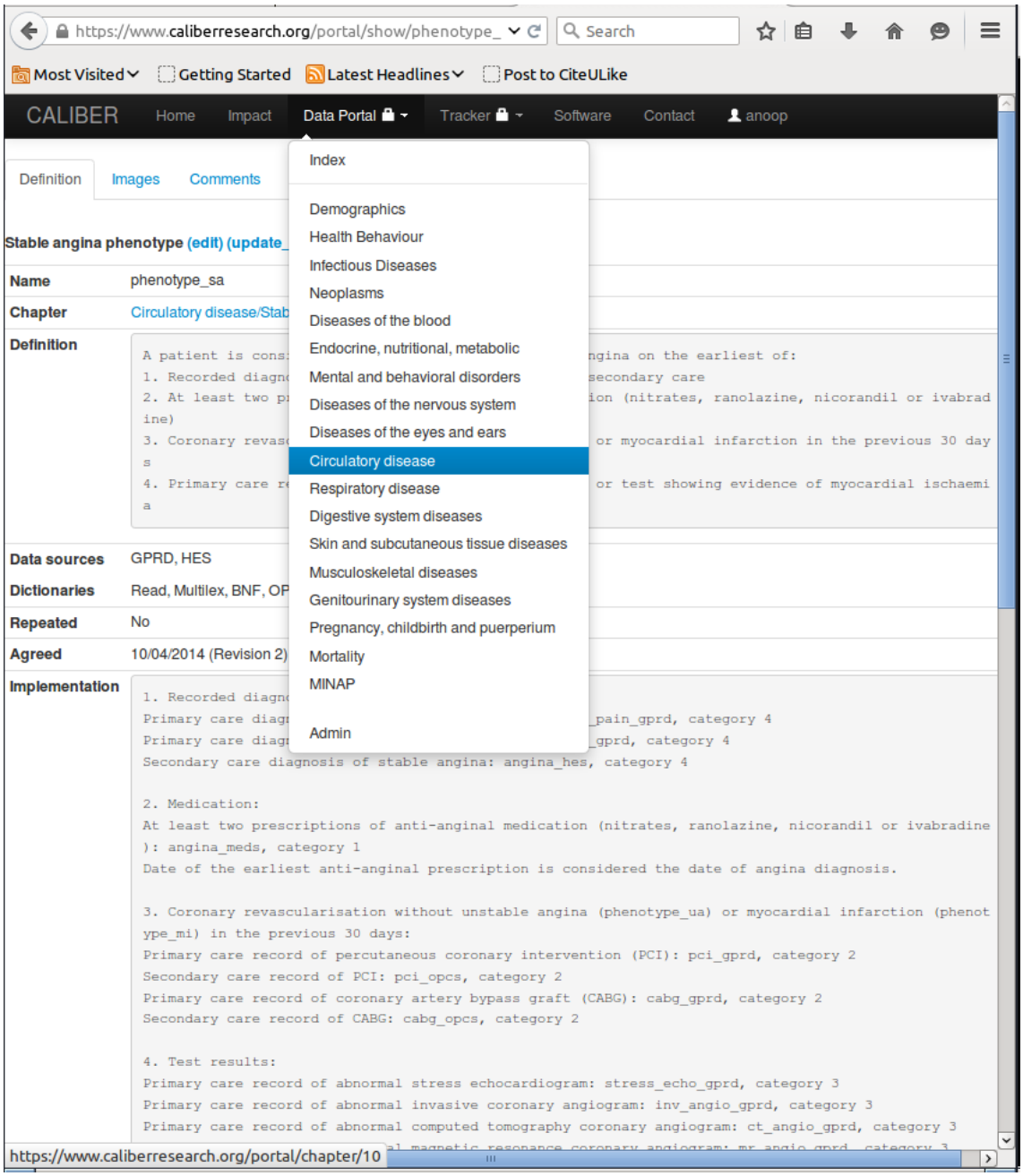
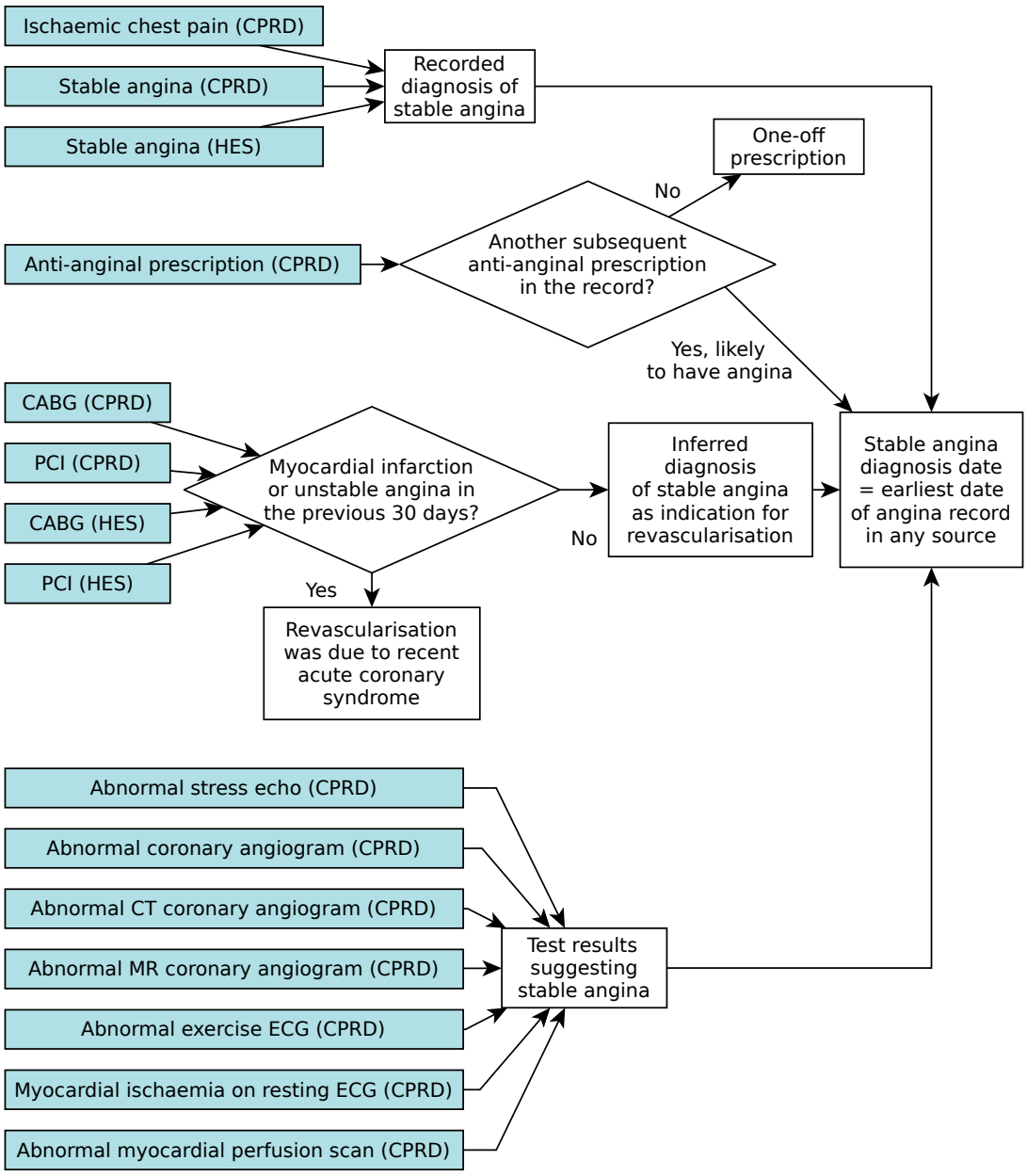


Figure 3.3: Flowchart showing data sources for stable angina phenotype



3.8 Summary and conclusions

CALIBER is a rich and exciting resource for large-scale epidemiology spanning primary and secondary care. It harnesses the wealth of data in clinical general practice systems and supplements it with information on hospitalisation, acute coronary syndromes and deaths from national registries.

However, there are many steps to be taken before CALIBER data are usable in a research dataset. These steps could have been done in an *ad hoc* manner for this single project, but I sought to develop reusable tools and resources as far as possible in order to aid reproducibility of research findings and facilitate a range of different future studies.

The next chapter will describe these tools and processes. The first project involved cross-referencing and validation of the source datasets for myocardial infarction; this is essential in order to be able to obtain unbiased inferences from the data. Secondly, I developed software to assist in generating and processing the datasets in a reproducible way. Thirdly, I sought to address the problem of missing data in epidemiological datasets by developing and testing a novel imputation method.

Following on from these methodological projects in chapter 4, I will specify the study cohort and endpoint definitions in chapter 5.

4 Methods development of the CALIBER resource

4.1 Chapter outline

CALIBER contains a wealth of data which can yield profound insights into disease processes and healthcare, but converting clinical data into a research dataset is not straightforward. The process can be facilitated at a number of steps by the three projects described in this chapter.

The first project was to validate myocardial infarction diagnoses in the linked data sources (section 4.2). This was a joint project with Emily Herrett (London School of Hygiene and Tropical Medicine) and has been published in *BMJ* [18].

The second project (section 4.3 on page 80) describes common methods for converting raw CALIBER data into a research-ready format. This includes a suite of packages for the R statistical system to manipulate CALIBER data and clinical terminologies [145, 146]. I also describe the process of defining a clinical phenotype using coded diagnoses and other information in the electronic health record database.

The third project (section 4.4 on page 90) addresses the issue of missing data, one of the most common problems in analysing clinically-collected data. I developed a new method for multiple imputation using the Random Forest algorithm within the framework of Multivariate Imputation by Chained Equations (MICE) [147, 148].

4.2 Validation of acute myocardial infarction diagnoses in CALIBER

This is a condensed version of a paper in *BMJ* [18] cross-referencing acute myocardial infarction records in the electronic health record data sources that comprise CALIBER. We also produced a video abstract (`http://www.bmj.com/content/346/bmj.f2350?view=long&pmid=23692896`).

4.2.1 Abstract

Background: Linked electronic health records are increasingly used in health research. For a common acute medical condition, myocardial infarction, no previous study has compared four electronic health record sources.

Aim: To determine the completeness and diagnostic validity of myocardial infarction recording across four national health record sources: primary care, hospital care, a disease registry, and mortality register.

Methods: We used a sample of patients with acute myocardial infarction in England between January 2003 and March 2009, identified in four prospectively collected, linked electronic health record sources: the primary care Clinical Practice Research Datalink (CPRD), Hospital Episode Statistics (HES), acute coronary syndrome registry (Myocardial Ischaemia National Audit Project, MINAP) and cause-specific mortality data from the Office for National Statistics (ONS). The main outcome measures were recording of acute myocardial infarction, all-cause mortality within one year of acute myocardial infarction, diagnostic validity of acute myocardial infarction compared to electrocardiographic and troponin findings in the disease registry (gold standard).

Results: We found 21 482 patients with a record of acute myocardial infarction in one or more sources. Risk factors and non-cardiovascular co-existing conditions were similar across patients identified in CPRD, HES and MINAP. Immediate all-cause mortality was highest among acute myocardial infarction patients recorded in CPRD, which (unlike HES and MINAP) includes patients who did not reach hospital, but at one year mortality rates in cohorts from each source were similar. 5561 patients with non-fatal acute myocardial infarction (31.0%) were recorded in all three sources and 11 482 (63.9%) in at least two sources. Crude incidence of acute myocardial infarction was underestimated by 25–50% using one source compared to using all three. Compared to acute myocardial infarction defined in MINAP, the positive predictive value of acute myocardial infarction recorded in CPRD was 92.2% (95% confidence interval (CI) 91.6, 92.8) and in HES was 91.5% (95% CI 90.8, 92.1).

Conclusion: Failure to use linked electronic health records from primary care, hospital admission, disease registry and death certificates may lead to biased estimates of myocardial infarction incidence and outcome.

4.2.2 Introduction

Linked electronic health records are increasingly used in research, in measuring outcomes of health care, and in health policy. However, for acute myocardial infarction the overlap between these four electronic health record sources, the patient risk factors and subsequent mortality have not previously been compared. Previous cross-referencing studies have typically compared one or two electronic sources, such as coded hospital discharge diagnoses and cause of death, to case note review [149–151], questionnaire to general practitioners [141] or active case finding in a prospective consented study [152–154].

Linkages with the national, ongoing acute coronary syndrome registry allowed detailed myocardial infarction diagnoses (with coded electrocardiographic findings and markers of myocardial necrosis, not available in other sources) to be compared with primary care and hospitalisation diagnoses. Linkages with primary care allowed evaluation of risk factors

in patients with a record of acute myocardial infarction in any source. Linkages with the death record allowed evaluation of cause specific mortality of myocardial infarction recorded in any source, including among cases not admitted to hospital.

Our objective therefore was to compare incidence, recording, agreement of dates and codes, risk factors and all-cause mortality of acute myocardial infarction recorded in primary care, hospital discharges, national acute coronary syndrome registry and the national death registry.

4.2.3 Methods

We used the CALIBER four-source linked dataset (primary care data from CPRD, hospital admission data from HES, mortality data from ONS and detailed information about acute coronary syndromes from MINAP) as described in chapter 3. We identified patients with acute myocardial infarction in any of the CALIBER data sources between January 2003 and March 2009 and compared the characteristics of patients recorded in each source.

We identified myocardial infarction in CPRD by searching for any of 62 Read codes for myocardial infarction in the patient's record, in HES by ICD-10 codes I21, I22 or I23 recorded as the primary discharge diagnosis for the first episode of the hospital admission, and in MINAP using the recorded discharge diagnosis, markers of myocardial necrosis and coded electrocardiogram findings.

We studied the first myocardial infarction per patient during the study period. We calculated the incidence of myocardial infarction and all cause mortality within one year for different combinations of sources, with the denominator being a patient's registration period with a CPRD general practice during the time that the practice met CPRD recording standards.

We assessed the diagnostic validity of myocardial infarction recorded in CPRD or HES compared to the MINAP gold standard. Further details are given in the publication [18].

The study period was 1 January 2003 to 31 March 2009 (when all record sources were concurrent) and confined to patients who had been registered with their general practice for at least a year and the practice had been submitting data for at least one year that met Clinical Practice Research Datalink data quality standards for continuity and plausibility of data recording. In the main analysis we included only patients with at least one record of admission to hospital in Hospital Episode Statistics at any time (for any cause) as these patients were shown to be linkable, but we conducted a sensitivity analysis including all patients. We selected the first record of myocardial infarction during the patient's study period as the index event and considered myocardial infarction records in the other data sources as representing the same event if they were dated within 30 days of the index event.

Table 4.1: Recording of risk factors in CPRD before myocardial infarction recorded in any of the CALIBER data sources from 1st January 2003 to 31st March 2009

	CPRD	HES	MINAP	ONS
Number of patients	15 819	13 831	10 351	4017
Age in years, median (IQR)	73 (61–81)	73 (61–82)	72 (61–81)	81 (73–87)
N women (%)	5810 (36.7)	5072 (36.7)	3649 (35.3)	1752 (43.6)
Most deprived quintile, n (%)	3211 (20.3)	2641 (19.1)	1997 (19.3)	849 (21.1)
Smoking, n (%)				
Current	4147 (26.2)	3608 (26.1)	2729 (26.4)	638 (15.9)
Ex	9414 (59.5)	8176 (59.1)	6194 (59.8)	2622 (65.3)
Non	1933 (12.2)	1745 (12.6)	1341 (13.0)	521 (13.0)
Missing	325 (2.1)	302 (2.2)	87 (0.8)	236 (5.9)
Systolic blood pressure mmHg, mean (SD)	145 (15.4)	145 (15.6)	145 (15.2)	146 (16.1)
Missing n (%)	385 (2.4)	351 (2.5)	198 (1.9)	64 (1.6)
Use of blood pressure lowering drugs, n (%)	9149 (57.8)	7907 (57.2)	5950 (57.5)	2919 (72.7)
Total serum cholesterol mmol/L, mean (SD)	5.4 (0.9)	5.4 (0.9)	5.4 (0.9)	5.2 (0.9)
Missing n (%)	4646 (29.3)	4291 (31.0)	2927 (28.3)	1101 (27.4)
HDL cholesterol mmol/L, mean (SD)	1.3 (0.3)	1.3 (0.3)	1.3 (0.3)	1.4 (0.3)
Missing n (%)	6985 (44.2)	6214 (44.9)	4443 (42.9)	1703 (42.4)
Use of lipid lowering medications, n (%)	5632 (35.6)	4686 (33.9)	3669 (35.4)	1757 (43.7)
Framingham hard coronary heart disease risk score, n (%)				
<10%	1273 (8.0)	1019 (7.4)	851 (8.2)	186 (4.6)
10–20%	4718 (29.8)	4121 (29.8)	3181 (30.7)	1248 (31.1)
>20%	2799 (17.7)	2439 (17.6)	1841 (17.8)	872 (21.7)
Missing	7029 (44.4)	6252 (45.2)	4478 (43.3)	1711 (42.6)
Diabetes, n (%)	2885 (18.2)	2467 (17.8)	1858 (17.9)	927 (23.1)
Charlson index, mean (SD)	2.5 (1.7)	2.4 (1.6)	2.4 (1.6)	3.2 (1.9)
Primary care consultation rate per year, median (IQR)	3.7 (1.6–7.9)	3.5 (1.5–7.7)	3.6 (1.6–7.8)	5.0 (2.3–9.8)

Deprivation was assessed by the index of multiple deprivation. Blood pressure and cholesterol values are the mean of measurements before the date of myocardial infarction. The Framingham score is based on patients with complete data for blood pressure and cholesterol.
The total number of patients was 21 482. Patients might be represented in more than one column if their myocardial infarction was recorded in more than one source.

Cardiovascular risk factors and non-cardiovascular coexisting conditions

For patients with acute myocardial infarction, we identified risk factors recorded in primary care, including age, sex, social deprivation [131], smoking, use of antihypertensives or lipid lowering drugs, diabetes mellitus, Charlson comorbidity index [155], and primary care consultation rate before the event. We used mean measures of systolic blood pressure, total cholesterol and high density lipoprotein cholesterol before myocardial infarction along with age, sex and smoking status (where recorded) to estimate the 10 year Framingham risk for acute myocardial infarction or coronary death [156].

Follow-up for mortality

We followed all patients with a record of myocardial infarction in any source for one year for death as recorded in the ONS death registry. We categorised patients as having fatal or non-fatal myocardial infarction by whether they died of any cause within seven days of the myocardial infarction. If a patient had a myocardial infarction record in CPRD, HES, or MINAP after their date of death, we considered that they died on the day of their myocardial infarction.

Agreement in recording

If the time difference between the earliest date of acute myocardial infarction in one source and the date in another source was no more than 30 days we considered that the records of acute myocardial infarction in the different sources agreed. A myocardial infarction recorded more than 30 days after the earliest date was considered a new event and was not included, ensuring that each patient appeared only once in the analysis. We chose 30 days to account for any delay in recording of myocardial infarction in primary care, assuming that any record within 30 days of a hospital admission was likely to represent the same event and anything after 30 days could feasibly be a subsequent myocardial infarction. We carried out a sensitivity analysis using a 90 day threshold.

Statistical analysis

Incidence. We estimated population based incidence rates of fatal and non-fatal acute myocardial infarction using the denominator of all adults in the CALIBER primary care population aged 18 and over (2.2 million), followed up for a mean 4.1 years between 2003 and 2009. We used each of the data sources separately and together to identify incident myocardial infarction, ending the follow-up period for a patient on the date of their first myocardial infarction during the study period, death, or de-registration from the general practice.

Cardiovascular risk factors and non-cardiovascular coexisting conditions. We compared patients with fatal and non-fatal acute myocardial infarction identified in the four data sources for risk factors and coexisting conditions recorded in primary care.

Death after acute myocardial infarction. We produced cumulative incidence curves for coronary and non-coronary mortality for patients recorded in each data source and compared mortality using a Cox proportional hazards model adjusted for age and sex.

Agreement in recording. We would expect patients who survived seven days after myocardial infarction to be recorded in the primary care, disease registry, and hospital admissions sources, and we assessed agreement between these three sources in a Venn diagram. For patients who died within seven days, we examined the proportion recorded by each source but did not compare agreement across all four sources as we would not expect the hospital discharge data and disease registry to record patients who died before reaching hospital.

We calculated the positive predictive value of primary care or hospital discharge diagnoses of acute myocardial infarction among patients who also had a record in the acute coronary syndrome registry. For each Read code for myocardial infarction, we calculated the proportion of patients with a myocardial infarction also recorded in HES or MINAP, and used this information to refine the phenotyping algorithm for myocardial infarction for use in the studies in this thesis (see section 5.4.4).

Data were analysed using Stata 12 and R 2.14 [157].

4.2.4 Results

We identified 21 482 patients with acute myocardial infarction in any of the four data sources.

Incidence

Among the single source crude estimates for incidence of myocardial infarction, primary care data (CPRD) gave the highest estimate, of 187 per 100 000 patient years (95% confidence interval (CI) 184, 190), followed by hospital discharge data (HES) with 154 per 100 000 patient years (95% CI 152, 157), acute coronary syndrome registry (MINAP) with 115 per 100 000 patient years (95% CI 113, 118), and death registry with 45 per 100 000 patient years (95% CI 43, 46). Combining these three sources yielded an estimate of 243 per 100 000 patient years (95% CI 239, 246, **Figure 4.3 on page 74**). The crude incidence of acute myocardial infarction was 25% lower using only CPRD and 50% lower using only MINAP compared with using all three sources.

Cardiovascular risk factors and comorbidity

Overall, the cohorts identified from CPRD, HES and MINAP had a similar prevalence of cardiovascular risk factors and comorbidities. However, compared with those recorded in MINAP or HES, patients with fatal or non-fatal myocardial infarction recorded only in CPRD were on average two years younger and more likely to be current smokers and in the most deprived fifth (P <0.001 for these comparisons). Patients recorded by the death registry were older than patients recorded in the other sources and had a higher burden of risk factors reflecting their age. However, other demographic characteristics and cardiovascular risk factors were broadly similar across patients recorded in the different sources (**Table 4.1 on page 68**).

Death after acute myocardial infarction

Immediate all cause mortality was highest among patients with acute myocardial infarction recorded in CPRD, which (unlike HES and MINAP) included patients who did not reach hospital. Patients with myocardial infarction identified in MINAP had a lower crude 30 day mortality (10.8%, 95% CI 10.2%, 11.4%) than those identified in HES (13.9%, 95% CI 13.3%, 14.4%) or in CPRD (14.9%, 95% CI 14.4%, 15.5%, **Figure 4.1 on page 73**). At one year, however, mortality was similar amongst patients with myocardial infarction identified in any of the three sources, at around 20%.

Non-fatal acute myocardial infarction: agreement between record sources

Among the 17 964 patients with at least one record of non-fatal acute myocardial infarction, 13 380 (74.5%) were recorded by CPRD, 12 189 (67.9%) by HES, and 9438 (52.5%) by MINAP. Overall, 5561 (31.0%) of patients had the event recorded in all three sources and 11 482 (63.9%) in at least two sources (**Figure 4.2 on page 73**). When we extended the recording window from 30 days to 90 days, the proportion recorded in all three sources increased only slightly, to 32.0% (n=5747). When we included patients who had never had a record of a hospital admission in HES, the proportion of non-fatal myocardial infarctions recorded in all three sources decreased slightly to 30.0% (5561/18 536) and the proportion recorded only in primary care increased from 17.7% (3188/17 964) to 20.3% (3760/18 536). A sensitivity analysis in which the HES case definition included secondary diagnoses of myocardial infarction, where myocardial infarction was not the reason for admission, produced a slight increase in the proportion recorded in all three sources (5812/18 283, 32.0%), and identified 306 additional myocardial infarctions that were not in any other source.

Positive predictive value

For patients with myocardial infarction in CPRD or HES who also had an associated record in MINAP, the positive predictive value of the myocardial infarction diagnosis (the

Table 4.2: Information recorded in the disease registry (MINAP) within 30 days for non-fatal myocardial infarction (MI) recorded in primary care (CPRD) or hospital admission (HES)

Positive predictive value of HES or CPRD MI is calculated considering a MINAP diagnosis of MI to be the gold standard. Numbers in parentheses are percentages unless stated otherwise.

	Whether MI is recorded in CPRD, HES or both		
Information in MINAP	**CPRD**	**HES**	**CPRD and HES**
Number of patients	7224	7489	6006
ECG findings			
ST elevation	3373 (46.7)	3455 (46.1)	3062 (51.0)
Other abnormality	2337 (32.4)	2485 (33.2)	1816 (30.2)
Normal	389 (5.4)	429 (5.7)	306 (5.1)
Not recorded	1125 (15.6)	1120 (15)	822 (13.7)
Cardiac markers			
Raised	6149 (85.1)	6358 (84.9)	5121 (85.3)
Normal	368 (5.1)	411 (5.5)	299 (5.0)
Missing	707 (9.8)	720 (9.6)	586 (9.8)
Peak troponin			
N (%) with peak troponin recorded	6109 (84.6)	6357 (84.9)	5029 (83.7)
Median (inter-quartile range)	2.03 (0.47-10.0)	2.04 (0.45-10.2)	2.34 (0.53-11.6)
CALIBER diagnosis			
ST elevation MI	3386 (46.9)	3441 (46.0)	3064 (51.0)
Non ST elevation MI	3274 (45.3)	3410 (45.5)	2497 (41.6)
Unstable angina	384 (5.3)	425 (5.7)	312 (5.2)
Other	180 (2.5)	213 (2.8)	133 (2.2)
Positive predictive value for MI, with MINAP MI as the gold standard (95% CI)	92.2 (91.6, 92.8)	91.5 (90.8, 92.1)	92.6 (91.9, 93.3)

probability that the diagnosis recorded in MINAP was myocardial infarction rather than unstable angina or a non-cardiac diagnosis) was 92.2% (6660/7224, 95% CI 91.6%, 92.8%) for CPRD and 91.5% (6851/7489, 95% CI 90.8%, 92.1%) for HES (**Table 4.2 on page 72**).

Unlike MINAP, HES did not record the type of myocardial infarction (ST elevation myocardial infarction, STEMI or non ST elevation myocardial infarction, NSTEMI), because ICD-10 does not contain specific codes for the type of myocardial infarction. However, the pattern of ICD-10 codes differed between STEMI and NSTEMI, with STEMI more likely to be recorded using codes I210 or I211 and NSTEMI more likely to be recorded as I214 or I219 (**Table 4.3 on page 74**). Read codes denoting a myocardial infarction record in CPRD varied in how likely they were to be associated with myocardial infarction in HES or MINAP, as shown in **Table 4.4 on page 75**.

For patients with a record in MINAP and either CPRD or HES, the positive predictive value of the acute myocardial infarction diagnosis (i.e. the probability that the diagnosis recorded in MINAP was myocardial infarction rather than unstable angina or a non-cardiac diagnosis) was 92.2% (95% CI 91.6%, 92.8%) for myocardial infarction in CPRD and 91.5% (95% CI 90.8%, 92.1%) for myocardial infarction in HES (**Table 4.2 on page 72**).

Figure 4.1: Kaplan Meier curves showing all cause mortality, stratified by record source
CPRD N=15,819, HES N=13,831, MINAP N=10,351; total 20,819 patients

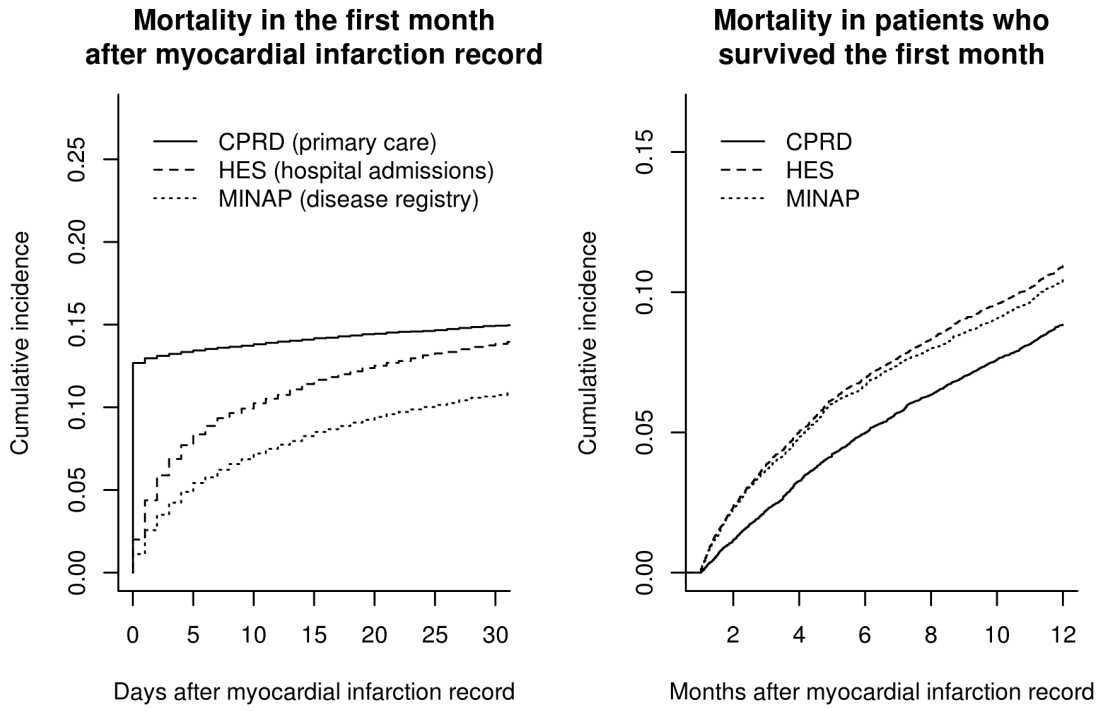


Figure 4.2: Number and percentage of records from primary care (CPRD), hospital data (HES) and the disease registry (MINAP) for non-fatal myocardial infarction
N=17,964 patients

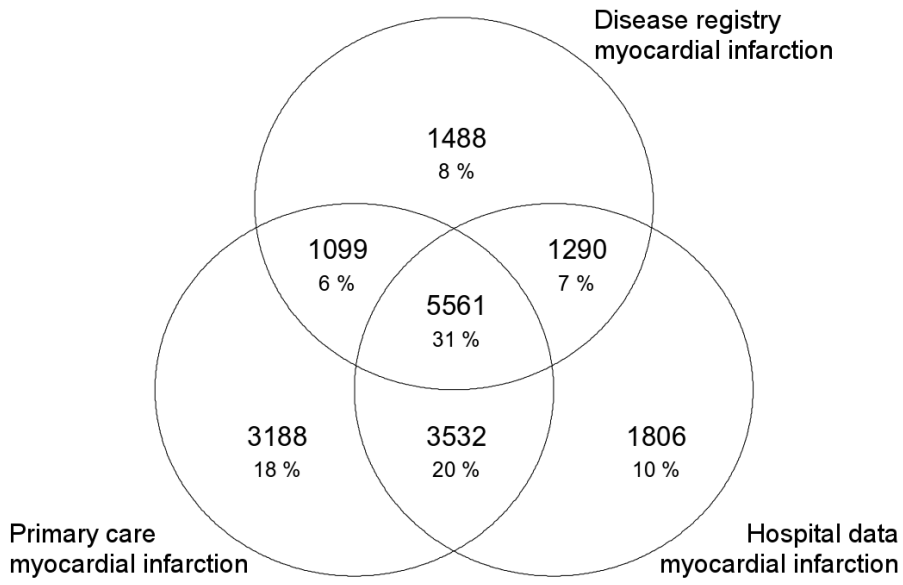


Figure 4.3: Crude incidence of acute fatal and non-fatal myocardial infarction estimated using different combinations of primary care data (CPRD), hospital admission data (HES), disease registry (MINAP) and death registry (ONS)

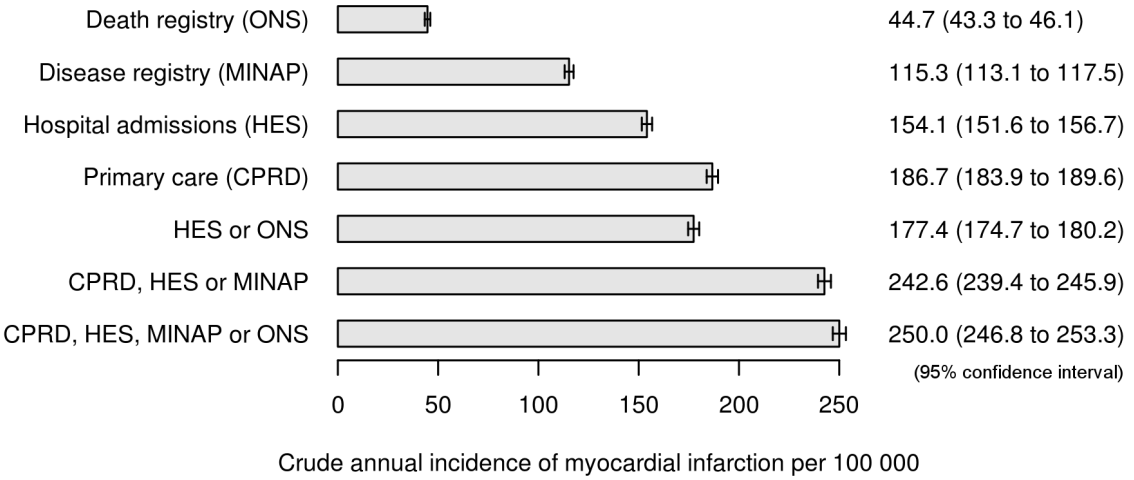


Table 4.3: Distribution of ICD-10 codes for ST and non-ST elevation myocardial infarction

ICD-10 code	ICD-10 term	STEMI N	STEMI %	NSTEMI N	NSTEMI %
I210	Acute transmural myocardial infarction of anterior wall	1122	30.2	171	4.8
I211	Acute transmural myocardial infarction of inferior wall	1472	39.6	228	6.4
I212	Acute transmural myocardial infarction of other sites	119	3.2	75	2.1
I213	Acute transmural myocardial infarction of unspecified site	19	0.5	6	0.2
I214	Acute subendocardial myocardial infarction	119	3.2	836	23.5
I219	Acute myocardial infarction, unspecified	606	16.3	1803	50.7
I220	Subsequent myocardial infarction of anterior wall	63	1.7	21	0.6
I221	Subsequent myocardial infarction of inferior wall	114	3.1	24	0.7
I228	Subsequent myocardial infarction of other sites	17	0.5	84	2.4
I229	Subsequent myocardial infarction of unspecified site	61	1.6	309	8.7
I230	Haemopericardium as current complication following acute myocardial infarction	0	0.0	1	0.0
I232	Ventricular septal defect as current complication following acute myocardial infarction	0	0.0	1	0.0
I238	Other current complications following acute myocardial infarction	2	0.1	0	0.0

STEMI, ST elevation myocardial infarction; NSTEMI, non-ST elevation myocardial infarction

Table 4.4: Proportion of CPRD myocardial infarctions associated with HES primary diagnosis of myocardial infarction and MINAP diagnosis of myocardial infarction, by Read term

Read term	N patients	HES primary	MINAP STEMI	MINAP NSTEMI	New category
G30..00 Acute myocardial infarction	5220	3458 (66.2%)	1425 (27.3%)	1120 (21.5%)	Acute MI
G30..15 MI - acute myocardial infarction	3312	2166 (65.4%)	920 (27.8%)	726 (21.9%)	Acute MI
G307100 Acute non-ST segment elevation myocardial infarction	2653	1908 (71.9%)	103 (3.9%)	1205 (45.4%)	NSTEMI
G30z.00 Acute myocardial infarction NOS	754	500 (66.3%)	162 (21.5%)	183 (24.3%)	Acute MI
G30X000 Acute ST segment elevation myocardial infarction	714	616 (86.3%)	463 (64.8%)	62 (8.7%)	STEMI
G308.00 Inferior myocardial infarction NOS	450	375 (83.3%)	256 (56.9%)	29 (6.4%)	Acute MI
G301z00 Anterior myocardial infarction NOS	115	93 (80.9%)	67 (58.3%)	11 (9.6%)	Acute MI
G301.00 Other specified anterior myocardial infarction	79	50 (63.3%)	28 (35.4%)	6 (7.6%)	Acute MI
G30..14 Heart attack	74	32 (43.2%)	13 (17.6%)	14 (18.9%)	Uncertain MI
G300.00 Acute anterolateral infarction	58	48 (82.8%)	29 (50%)	4 (6.9%)	Acute MI
G302.00 Acute inferolateral infarction	57	45 (78.9%)	31 (54.4%)	2 (3.5%)	Acute MI
G307000 Acute non-Q wave infarction	40	25 (62.5%)	4 (10%)	11 (27.5%)	Acute MI
G301100 Acute anteroseptal infarction	34	28 (82.4%)	16 (47.1%)	6 (17.6%)	Acute MI
323..00 ECG: myocardial infarction	31	13 (41.9%)	6 (19.4%)	2 (6.5%)	MI on ECG
G307.00 Acute subendocardial infarction	30	22 (73.3%)	1 (3.3%)	11 (36.7%)	Acute MI
G304.00 Posterior myocardial infarction NOS	23	14 (60.9%)	8 (34.8%)	4 (17.4%)	Acute MI
G30..17 Silent myocardial infarction	22	2 (9.1%)	0 (0%)	1 (4.5%)	Uncertain MI
G38..00 Postoperative myocardial infarction	18	1 (5.6%)	1 (5.6%)	3 (16.7%)	Acute MI
G30A.00 Mural thrombosis	17	3 (17.6%)	0 (0%)	0 (0%)	Uncertain MI
G303.00 Acute inferoposterior infarction	16	13 (81.3%)	8 (50%)	3 (18.8%)	Acute MI
G305.00 Lateral myocardial infarction NOS	14	8 (57.1%)	6 (42.9%)	1 (7.1%)	Acute MI
G310.11 Dressler's syndrome	13	0 (0%)	0 (0%)	0 (0%)	MI complications
G30y.00 Other acute myocardial infarction	9	7 (77.8%)	2 (22.2%)	2 (22.2%)	Acute MI
G30y200 Acute septal infarction	8	2 (25%)	3 (37.5%)	0 (0%)	Acute MI
G30..11 Attack - heart	7	2 (28.6%)	2 (28.6%)	0 (0%)	Uncertain MI

Results shown for 25 most common Read terms for MI. STEMI, ST elevation myocardial infarction; NSTEMI, non-ST elevation myocardial infarction
'New category' is a suggested classification of myocardial infarction Read codes based on this work.

4.2.5 Discussion

We compared electronic health records on one major disease event – acute myocardial infarction – across four English, ongoing sources of health record data: primary care (CPRD), hospital admissions (HES), a quality improvement disease registry (MINAP), and the death registry (ONS). We found that each data source missed a substantial proportion of myocardial infarction events, but that patients with a myocardial infarction record in any source had risk factor profiles and one-year mortality consistent with true myocardial infarction. Taken together, these findings support the wider use of linkage of multiple record sources by clinicians, policy makers, and researchers.

Fatal myocardial infarction

Both primary care and death registry data can be used to capture fatal myocardial infarction occurring out of hospital among people without a record of myocardial infarction in HES or MINAP. The death registry is a useful source of fatal acute myocardial infarction for research, as most (83.0%) patients who were identified as having acute myocardial infarction in any of the data sources and died within seven days had myocardial infarction recorded as their underlying cause of death. These figures agree with results from the Oxford Record Linkage Study, where among 5686 patients admitted to hospital with myocardial infarction 85.2% who died within 30 days had myocardial infarction recorded as the underlying cause of death [158].

Non-fatal myocardial infarction

We found that each record source misses cases, with only one third of non-fatal myocardial infarctions being recorded in all three data sources (CPRD, HES and MINAP). Two thirds of non-fatal myocardial infarctions were recorded in at least two sources. CPRD was the single most complete source of non-fatal myocardial infarction records (one quarter of all non-fatal myocardial infarction events not recorded), HES missed one third, and MINAP missed nearly half (**Figure 4.2 on page 73**). Previous two-source comparison studies in Scotland [159], Australia [160], Denmark [161], and the Netherlands [154] have shown that hospital records alone underestimate the true incidence of myocardial infarction. Despite the low sensitivity of these data sources, in our study the positive predictive value of myocardial infarction records in primary care and hospital admission sources were over 90% compared with the disease registry gold standard (**Table 4.2 on page 72**). However, some of the myocardial infarctions recorded only in primary care are likely to be historical diagnoses because the mortality rate in the first month is much lower than among those also recorded in hospital sources.

Our results using cross referencing of electronic health records in 20 000 patients are consistent with previous manual approaches to validation in primary care, which validated Read coded diagnoses in a few hundred patients against anonymised free text, death

certificates, paper medical records, hospital discharge summaries, or questionnaires to general practitioners [133, 162–167]. Our much larger sample size, however, allowed us to evaluate individual Read terms that are used to record myocardial infarction (**Table 4.4 on page 75**). This type of validation has not been done previously for myocardial infarction and may be relevant to other common conditions that can be recorded using a variety of codes, such as stroke [168].

Hospital admission data

To our knowledge, no studies have examined the positive predictive value of ICD-10 coded myocardial infarction diagnosis in hospital admission data against an ongoing disease registry. We found that 62.8% of non-fatal myocardial infarctions recorded in CPRD and 72.6% recorded in MINAP were also recorded in HES.

Disease registry and maximising true positives

The strength of the disease registry MINAP is that it contains detailed diagnostic records: troponin values, coded electrocardiographic findings, and cardiologist diagnosis of STEMI or NSTEMI. These data items provide validated endpoints from all hospitals in England and Wales, and are not available in other sources. An acute myocardial infarction recorded in MINAP can be considered an electronic health record 'gold standard', as a myocardial infarction recorded by a registry is likely to fulfil international diagnostic criteria [9]. Validation of myocardial infarctions in CPRD or HES against those recorded by MINAP showed a positive predictive value of over 90%, making them suitable for detecting endpoints in cohort studies and trials, where poor endpoint resolution can dilute any observed effect and reduce the power of a study [169]. The positive predictive value was not 100% because some myocardial infarction records in primary care may actually have been related to unstable angina or chest pain of an unknown cause.

Limitations of this study

One limitation is that our data were from a sample of English general practices rather than the entire population, but patients included in CPRD have been shown to be representative of the UK population (see section 3.3.1). A second limitation concerns the generalisability of our findings for the quality of primary care data. Practices contributing data to the CPRD are advised of recording guidelines and their data are accepted only when they meet standards of data completeness, so they are likely to record disease events better than general practices that do not contribute. Our estimates of agreement from this study may therefore be higher than for practices that do not contribute to CPRD. Thirdly, our validation of HES and CPRD myocardial infarctions against MINAP was limited to the subset of patients with a MINAP record, and caution must be exercised in extending these conclusions to patients with myocardial infarctions without a MINAP record.

Clinical and policy implications

With the current emphasis on measuring clinical outcomes in health systems, our study has important implications for practice and policy. Firstly, we propose much wider use of linked record sources in commissioning and in research to estimate disease occurrence and outcome, because of the biases inherent in using only one source of records. Changing the estimates for incidence of myocardial infarction could potentially alter the modelled effect of population based healthcare interventions. Our findings underscore the importance of international initiatives to accelerate availability of linked data [170–172].

Secondly, a national strategy for biomedical informatics is required to tackle manifest system failings: a single health event should, ideally, have a single record that is propagated in multiple record systems. Disease registries such as MINAP could be improved if embedded in real time clinical care of all patients with myocardial infarction, rather than the current situation in which hospitals employ audit staff to retrospectively enter records on patients in coronary care units. This is applicable to many disease areas, not just coronary care. In section 10.1.2 (page 235) I outline how this may be achieved in a standardised way using openEHR archetypes.

Thirdly, primary care records could be improved by recording precise dates of events, and avoidance of repeat recording of acute events. This can be achieved by having user-friendly problem-based medical records and ensuring that healthcare staff can use it correctly and easily.

Future research

Several lines of research are warranted by our findings. Firstly, research is required to understand how electronic health record data are coded, and how this can be improved. Secondly, more extensive cross referencing is required against additional sources of information on myocardial infarction. These include self reported myocardial infarction (which may be less dependent on specific setting in the health system), manual review of all the available local case records (paper and electronic), and investigation of electronic free text recorded by general practitioners (for diagnoses that are not recorded using a Read code). Such efforts are underway in the UK Biobank cohort (n=500 000) [169]. There is a need for investigator led cohorts and trials to link with the primary care record [173]. It is important to evaluate how other registries, such as cancer registries, compare with primary care and hospital data sources [174]. This is essential for large studies where manual review of case records is not feasible.

Conclusion

Failure to use linked electronic health records from primary care, hospital care, disease registry, and death certificates may lead to biased estimates of the incidence and outcome of myocardial infarction. The use of linked data helps to overcome limitations of the

individual sources, but it is important to cross-reference and understand the data sources in order to be able to use them effectively in research [18].

4.3 Converting raw data to research-ready variables in CALIBER

4.3.1 Abstract

Introduction. Researchers using electronic health record databases have had to develop complex algorithms and compile lists of diagnosis terms in order to define cohorts and extract clinical characteristics or phenotypes of the patients in the study. Typically these methods have not been shared among research communities, which has led to duplication of effort and difficulties in reproducing research findings.

Methods. I wrote a set of packages for the R statistical system was developed to assist in generation of codelists (lists of diagnosis, procedure or other terms that are used to define a research variable), data management and handling missing data. A process was developed for developing phenotype algorithms, documenting them and publishing them on the CALIBER web portal.

Results. Three R packages have been published and are currently being used by other CALIBER researchers, and are accessible to any researcher to download and use. The CALIBER team has developed 25 disease phenotypes and over 400 other variables, which are curated on the CALIBER portal.

Discussion. The existence of the CALIBER portal and other tools has greatly facilitated research using the CALIBER database. However, looking beyond CALIBER there is a need for a broader repository of open access methods for using electronic health records.

4.3.2 Introduction

Researchers using electronic health record databases such as the CPRD have had to develop algorithms, some of which may be quite complex, to identify patients with particular phenotypes. These involve identifying diagnoses and other coded information. Lists of diagnoses codes are typically shared informally within a research group, and sometimes published along with papers, but the lack of systematic sharing of algorithms can result in duplicated effort, lack of transparency and non-comparability of studies that used different definitions.

It is now being recognised that this is an inefficient way to do research, and there are new initiatives to encourage sharing of code and algorithms. One notable example is the Electronic Medical Records and Genomics (eMERGE) network [6], which aims to bring together electronic health record datasets from a number of US hospitals to explore phenotype-genotype associations in larger number of patients. Members of the eMERGE consortium publish their phenotyping algorithms (methods for deducing a patient's clinical characteristics based on information in the electronic health record) on the eMERGE Phenotype Knowledgebase (PheKB) website `https://phekb.org/phenotypes?`.

The slow steps taken towards sharing methods in electronic health record research contrast with the early adoption of standardisation and sharing of methods in bioinformatics. The CALIBER programme aims to take further steps in this direction, by providing a portal for sharing and distributing data preparation algorithms and phenotype definitions (`https://caliberresearch.org/portal`).

Creating a research variable from CALIBER raw data typically requires the researcher to first define the diagnoses or procedures of interest by looking for terms in a relevant terminology (e.g. Read, ICD-10 or OPCS), then to extract records with the terms of interest, and finally to process the data to arrive at a workable definition for a study. For example, creation of a variable for body mass index requires extraction of height and weight records, data cleaning to convert all the measurements into the same units and remove erroneous results, a method for handling measurements recorded at different times, and calculation of the final body mass index from the cleaned data [175].

4.3.3 R packages

R is an increasingly popular open source statistical system and programming language [157]. A key strength of R is the large number of user-contributed extension packages which can extend its functionality and interface with other programs. R has a well-documented, structured system for creating packages which mandates the creation of comprehensive help documentation. Creating re-usable code in a package is initially more work than writing an *ad hoc* function, but the code can be published and used more easily by other researchers so it is beneficial in the long run – both in terms of saving time and improving the transparency of research.

I created a set of R packages which are available from the 'caliberanalysis' project page on R-Forge (`www.caliberanalysis.r-forge.r-project.org`):

CALIBERcodelists – functions to create, modify, compare, merge, export and convert codelists. A codelist is a list of Read, OPCS and ICD-10 codes for identifying a particular condition in raw CALIBER data.

CALIBERdatamanage – data management and display tools, including functions for loading databases with automatic date conversion, for creating a variable per patient from data at multiple time points, and producing summary tables and forest plots.

CALIBERrfimpute – Random Forest multiple imputation for MICE (see section 4.4 on page 90). This package is published on the peer-reviewed Comprehensive R Archive Network repository, which is an online resource of quality-assured packages that can be installed directly by anyone running R.

In addition I created a package 'CALIBERlookups' containing dictionaries and lookup tables for CALIBER, for use within our research group or other NHS / UK academic collaborations. CALIBERlookups is a compilation of CPRD lookups, ICD-10 and OPCS

dictionaries from the UK Terminology Centre, and ICD-9-CM codes from the US National Centre for Health Statistics. By providing all these resources within a single convenient package it spares researchers from having to download and prepare each individual dictionary.

CALIBERlookups is not published on R-Forge. It cannot be released generally because the WHO has granted a restricted licence to the Department of Health for use of the ICD-10 dictionary solely for healthcare and health research in the UK, which means that I am not permitted to publish it. (In contrast, Read and OPCS are available on the Open Government Licence, which does permit redistribution and reuse.)

The R packaging system enables R code, help files, test cases, examples and documentation to be distributed in an organised way, so that other researchers can understand and use the code. The source code and history of modifications to the code are curated on R-Forge.

4.3.4 CALIBER data management

The CALIBER data management R package provides convenient interfaces for certain data management tasks. In particular it streamlines the handling of large datasets, which can either be handled in memory as 'data.table' objects (very fast and flexible, but limited by the amount of computer memory), or on disk as 'ffdf' objects (which can be much larger, and leave more memory free for analysis, but have a less flexible programming interface). My CALIBERdatamanage package provides convenient functions for converting between the 'data.table' and 'ffdf' dataset formats, and defines data management functions with a common interface to data in either format.

These are some of the functions included in the package:

importDT Imports file to data.table, automatically detecting date format and extracting files from a zip file.

importFFDF Imports a single file or multiple files to ffdf, automatically detecting date format and extracting files from a zip file.

cohort Designates a dataset as a 'cohort' object, with one row per patient, a defined patient identifier and the facility to include a data dictionary.

extractCodes Subsets a data.table or ffdf object according to ICD-10, Read or OPCS codes in a codelist.

transferVariables Transfers variables from one dataset to another based on a key column. This makes it easier to work with multiple datasets each containing a few variables, saving computer memory.

addCodelistToCohort Creates a one-row-per-patient extract of a file based on a codelist (e.g. earliest diagnosis within a particular time window, based on an index date) and

merges it onto a cohort. This single function can replace several lines of code and make scripts more intuitive.

addToCohort Creates a one-row-per-patient extract of a file based on a value (e.g. lab result within a particular time window, based on an index date) and merges it onto a cohort.

multiforest A flexible function to create multiple forest plots with annotations, based on a specification in a spreadsheet format (which can include formatting information as well as values and text).

4.3.5 Codelists

Before describing the CALIBER system for deploying codelists, I will briefly define some relevant words and phrases:

Controlled vocabulary – an authoritative list of terms (words or phrases) to be used in indexing [176].

Terminology – a type of controlled vocabulary which covers the set of terms in a professional domain. For example, OPCS and the Read Clinical Terms Version 3 are terminologies [177]. Terminologies may be classifications or ontologies.

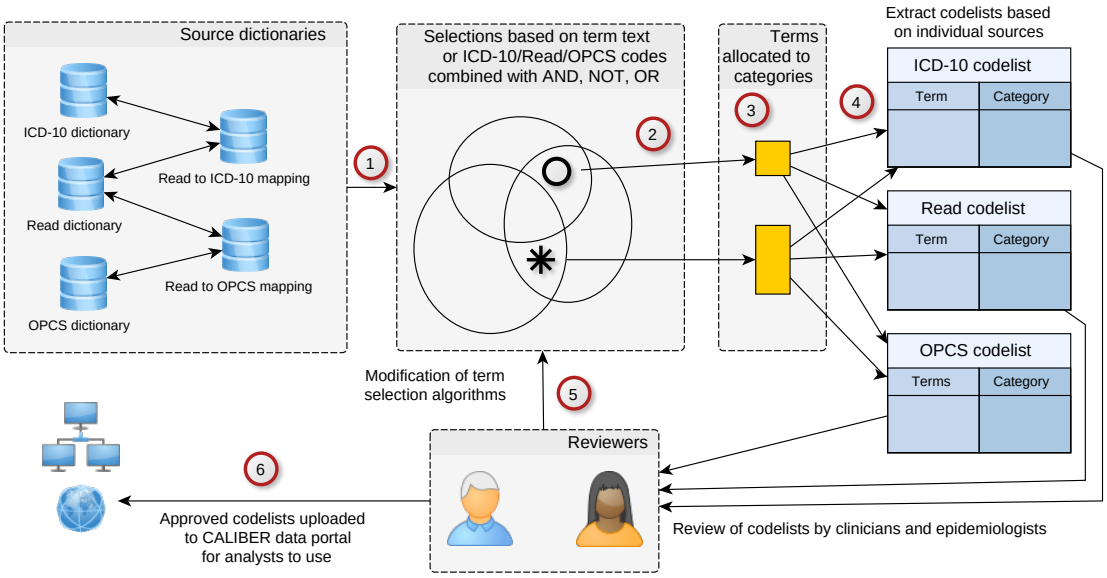
Classification / **taxonomy** – arranges all the concepts in the domain into groups or a hierarchical structure; the International Classification of Disease is an example [139].

Ontology – a set of concepts with definitions of the relationship between different concepts, such as 'is a part of' or 'is a type of' [176]. SNOMED-CT is a large terminology which contains the Read and ICD-10 terms and will eventually replace Read in the NHS. It is supposed to be an ontology, with each term defined in terms of relationships with other terms, although currently many of the terms do not have such relationships defined.

Algorithms for extracting research variables from electronic health record (EHR) data sources frequently need to identify records according to a terminology (e.g. selecting all ICD-10 terms for myocardial infarction). There are thousands of potential terms included in the terminologies, and for aggregation purposes one may require a handful of terms or hundreds of terms.

Codelists are developed by an iterative process which involves searching for terms in the dictionaries, combining selections of terms to derive a final set of chosen terms, and assigning a category to these terms (**Figure 4.4 on page 84**). In CALIBER, codelists are stored in a custom human-readable text file format which includes a column for metadata such as the author's name, version number and date. This file format can be opened in a simple text editor, a spreadsheet program, a statistical package or custom software

Figure 4.4: Process for generating a codelist using the R CALIBERcodelists package



1. Decide which source dictionaries to use, e.g. Read, ICD-10 and/or OPCS.
2. Create a selection of terms from the source dictionaries. e.g. all terms containing the word 'angina', or all ICD-10 codes beginning with 'I20'. Combine selections using Boolean operators such as AND, OR or NOT to identify exactly which terms are of interest.
3. Allocate a category to terms in a particular selection, e.g. 'history of angina'.
4. Set the metadata for the codelists under construction (version number, category descriptions, author name, date). Extract the Read, ICD-10 and OPCS codelists and save them in a standard file format.
5. Ask clinicians and epidemiologists to review the codelists and suggest any changes. Also it may be useful to validate the codelist by exploring CALIBER data, e.g. comparing myocardial infarction records extracted using a Read codelist versus records in MINAP. The algorithm can be changed and the results compared with the previous version if necessary.
6. Approved codelists with documentation can be shared on the CALIBER data portal (section 3.7 on page 61) for use in subsequent studies.

developed by the CALIBER data manager to automatically extract records of interest from the CALIBER master database. All codelists stored on the data portal have a version number in order to track which versions of each codelist were used for a particular project.

4.3.6 Phenotyping algorithms

The phenotype of an individual is the set of observable characteristics about that person, which may include clinical measurements such as height and blood pressure, or disease states such as diabetes or coronary artery disease. In electronic health record research, one is interested in studying the phenotype but the information about the phenotype is what is recorded in the electronic health record. It is crucial to understand the process by which information about a subject enters the record in order to be able to interpret it and infer the phenotype correctly [178].

For example, suppose one is interested in the phenotype of type 2 diabetes. If the only data source is hospitalisation data (HES) and the patient is never hospitalised, one cannot determine their diabetes status. If the patient is hospitalised and there is no diagnosis code for diabetes, it could mean any of the following:

- The patient is not diabetic
- The patient is diabetic, but was not tested for diabetes, and neither the patient nor the healthcare team know the patient is diabetic
- The patient is diabetic and the healthcare team know about it, but it is not entered in the patient record or not coded
- The patient knows their diabetes status, but the doctor does not

Ideally evidence should be gathered to inform the probabilities of each of these scenarios. Cross-referencing with other data sources, or looking at other information which might give clues (e.g. test results or medication in linked data) [179] may help. Frequently it is necessary to make some assumptions based on clinical knowledge, and develop a strategy for allocating a diagnostic label to the patient (e.g. diabetes confirmed, probable diabetes). This type of strategy is called a phenotyping algorithm.

There are few diseases or conditions for which it is possible to create a 'perfect' phenotype (identifying all cases of the disease with perfect accuracy), because it relies on having accurate information in the clinic and for that information to be entered consistently and accurately in the electronic health record. Instead a phenotyping algorithm may choose to maximise either diagnostic specificity or sensitivity, or aim to obtain an unbiased estimate of incidence. Studies using electronic health record data can include sensitivity analyses using different phenotyping algorithms (e.g. using different sets of data sources, or using restricted or expanded sets of diagnoses codes). This can be used to show that the limitations of the phenotyping process have not introduced bias into the results.

Cross-referencing between linked data sources can help inform the best use of each data source, as demonstrated in section 4.2.

Example: phenotyping algorithm for type 2 diabetes

This is an example of a CALIBER phenotyping algorithm for type 2 diabetes. It uses CPRD and HES data sources to classify whether a patient is diabetic at a particular point in time, and attempts to identify the type of diabetes. The algorithm was developed for a cohort study comparing the prognosis of people with type 2 diabetes or no diabetes, which is reported in chapter 5. It was assumed that the majority of diabetic patients of the age group included in the study (i.e. adults of all ages) would have type 2 diabetes, which becomes more prevalent with older age [180].

The aim of the algorithm was to identify patients with type 2 diabetes. It was not essential to exclude people with diabetes from the 'normal' comparison population with absolute certainty, because the normal population is so large that a small amount of mis-

classification would not substantially change the nature of the population. The algorithm was not designed to identify patients with type 1 diabetes, nor to accurately calculate the incidence of type 2 diabetes, because of the limitations on classifying the type of diabetes.

In developing the algorithm I reviewed previous algorithms for identifying diabetes in UK primary care datasets [181] and undertook exploratory work cross-referencing diabetes diagnoses with medication, glucose tests, HbA1c, date of diagnosis and patient registration date at the practice.

Type of diabetes can be classified using a combination of the following types of information in the electronic health record, each of which has potential limitations:

Diagnoses codes Assuming the patient and healthcare professional know what type of diabetes the patient has, it can theoretically be coded using the specific Read codes for type 1, type 2, or other specific types of diabetes. However, the recommended diagnosis codes can change over time, as they have done for myocardial infarction (as discussed in section 4.2), and I found that patients diagnosed in earlier years were less likely to have a specific code for diabetes. An additional problem is that some Read terms are ambiguous, for example 'insulin-dependent diabetes' is supposed to mean 'type 1 diabetes' but may be inappropriately applied to patients with type 2 diabetes on insulin. Some patients had conflicting diagnostic codes.

Medication Patients with type 1 diabetes are almost always treated with insulin, whereas patients with type 2 diabetes may have no medication (diet control), oral medication or insulin. Patients on oral glucose lowering medication are most likely to have type 2 diabetes (except for women on metformin, who may have polycystic ovary syndrome).

History Age at diagnosis, body mass index (BMI) at diagnosis, initial medication and other features of a patient's diabetes history can give important clues as to the type of diabetes, but this relies on patients having been registered at their GP throughout that time. CPRD only contains data from a patient's current period of registration, and only important summary diagnoses are entered for their previous history.

Clinical measurements Glucose and HbA1c are recorded inconsistently for patients without diabetes. High glucose in the absence of a formal diagnosis of diabetes may mean that the patient has diabetes but it is not coded; however interpretation should be cautious in the absence of the full clinical information that would be available to a clinician seeing the patient (e.g. whether the patient has symptoms).

Although others have developed detailed algorithms for classifying type of diabetes based on medication, history of medication (e.g. whether the patient was given insulin from the outset), age at diagnosis, body mass index etc. [181], the limitations in the recording of these additional pieces of information would limit any benefit to be gained

by a more complex algorithm. It would also risk introducing more selection bias, if only patients with detailed records can be classified accurately. Instead I reviewed the characteristics of patients identified by a simple algorithm based solely on diagnosis codes (**Figure 4.5 on page 89**) and found that they were reasonable for a population with type 2 diabetes (**Table 5.2 on page 114**). I also performed a sensitivity analysis by different levels of certainty of the type 2 diabetes diagnosis. The majority of diabetic patients in this cohort had type 2 diabetes, so the population with 'any diabetes' should be quite similar to the population with type 2 diabetes, and if similar associations with cardiovascular disease are observed in both populations, the lack of precision of diagnosis coding is not of concern.

This algorithm identifies patients with type 2 diabetes on a specified index date (e.g. study entry date) based solely on diagnostic information recorded in CPRD and HES. It does not use any information after study entry to classify people, in order to reduce immortal time bias (i.e. patients who survived would be more likely to be classified in a particular category), but such information could be included appropriately as auxiliary information if the multiple imputation approach is used. However this bias is not eliminated completely because some codes specifying the type of diabetes may have been entered retrospectively only for patients who survived.

The algorithm was designed to identify patients with type 2 diabetes or any diabetes, and was not designed to identify all patients with type 1 diabetes (because many were left with an unclassified type of diabetes). Ideally the type of diabetes for the unclassified patients could be inferred from other clinical information, and the uncertainty propagated in the analysis (e.g. using multiple imputation).

I calculated the age-specific prevalence of diabetes in CALIBER to help to validate this algorithm. These estimates (shown in **Table 4.5 on page 88**) are broadly similar to the prevalence in the Health Survey for England 2003, in which the percentage of people reporting doctor-diagnosed diabetes was 3.5% in men and 2.5% in women aged 45–54, increasing to 8.1% and 4.6% respectively in those aged 55–64 [182]. The prevalence of diabetes has increased considerably over the past decade; in 2010 it was estimated that 11.1% of men and 8.0% of women aged 55–64 in England were diabetic [180].

4.3.7 Discussion

In this section I have described tools to assist reproducible research using CALIBER including the definition of phenotypes, and management of datasets and codelists.

The R packages have been successfully used by a number of researchers within CALIBER, which has been extremely useful for suggesting improvements and finding errors. However there is a need for more researchers to be aware of the utility of such packages, and the practice of making a package from ad-hoc functions enforces discipline such as good documentation and testing, which improves the quality of code and can be a good exercise even if the aim is not to share the code with the wider community.

Table 4.5: Age and sex-specific prevalence of diabetes recorded in CALIBER

Sex	Age group	N patients	Percentage of patients with diabetes mellitus at baseline (95% CI)			
			Any diabetes	Type 1 diabetes	Type 2 diabetes	Diabetes of unspecified type
Men	30–40	420 975	0.87 (0.85, 0.90)	0.47 (0.45, 0.49)	0.28 (0.26, 0.30)	0.13 (0.12, 0.14)
	40–50	215 607	2.25 (2.19, 2.32)	0.50 (0.47, 0.53)	1.37 (1.32, 1.42)	0.38 (0.36, 0.41)
	50–60	170 679	4.41 (4.32, 4.51)	0.45 (0.42, 0.48)	3.24 (3.16, 3.33)	0.73 (0.69, 0.77)
	60–70	119 507	8.31 (8.15, 8.46)	0.52 (0.48, 0.56)	6.32 (6.18, 6.46)	1.47 (1.40, 1.54)
	70–80	80 614	10.5 (10.3, 10.8)	0.53 (0.48, 0.58)	7.89 (7.71, 8.08)	2.12 (2.03, 2.23)
	$\geq$ 80	38 991	9.96 (9.66, 10.3)	0.39 (0.33, 0.46)	7.32 (7.06, 7.58)	2.25 (2.10, 2.40)
Women	30–40	411 996	0.77 (0.74, 0.80)	0.35 (0.33, 0.37)	0.25 (0.23, 0.26)	0.17 (0.16, 0.19)
	40–50	198 619	1.52 (1.47, 1.57)	0.34 (0.31, 0.37)	0.90 (0.86, 0.95)	0.28 (0.25, 0.30)
	50–60	166 467	2.90 (2.82, 2.98)	0.34 (0.31, 0.37)	2.03 (1.96, 2.10)	0.53 (0.49, 0.56)
	60–70	123 553	5.79 (5.66, 5.93)	0.36 (0.33, 0.40)	4.37 (4.26, 4.49)	1.06 (1.00, 1.11)
	70–80	103 284	7.82 (7.66, 7.98)	0.42 (0.38, 0.46)	5.70 (5.56, 5.84)	1.70 (1.62, 1.78)
	$\geq$ 80	85 325	7.69 (7.51, 7.87)	0.31 (0.27, 0.35)	5.59 (5.43, 5.74)	1.79 (1.71, 1.88)

The CALIBER data portal is an unusual resource because it exposes and shares research methods that have previously been kept within research groups. This has led to duplication of effort and has hampered the reproducibility of research using EHR. However, the portal is limited to methods for using CALIBER data; it would be useful to have a repository which was more widely applicable for research using EHR data. Some algorithms can be quite generic and some may require specialisation to handle the idiosyncrasies of particular data sources.

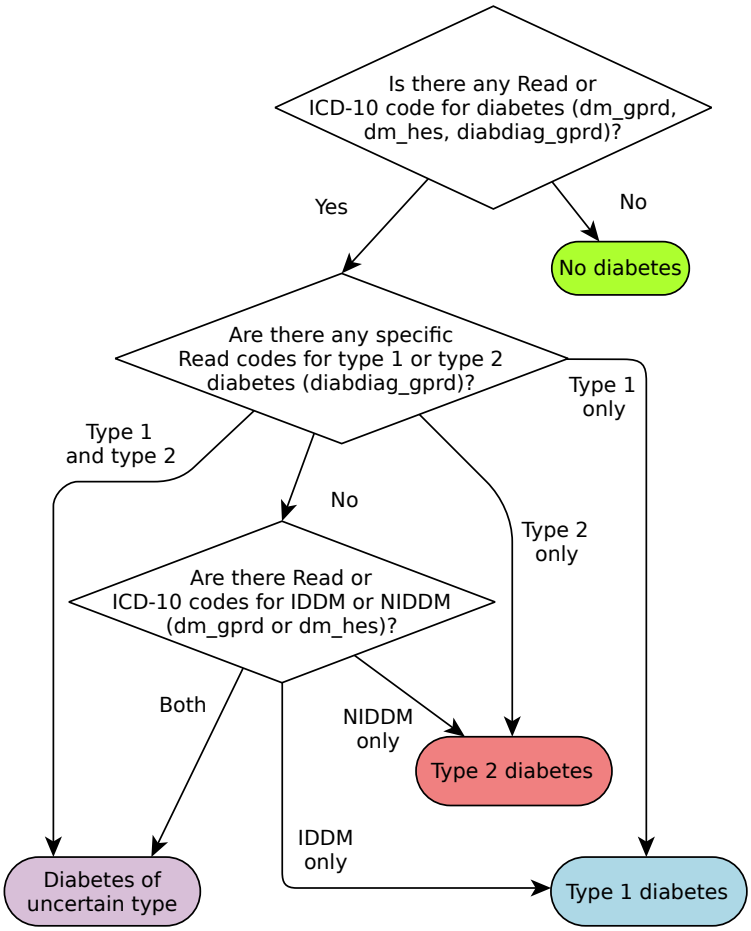
Comparison with other resources

There have been a number of recent initiatives to encourage and facilitate the sharing of algorithms within the EHR research community. A notable example is the Phenotype Knowledgebase of the eMERGE consortium [6]. It has a similar number of public phenotypes to CALIBER, about 30 as of June 2015 (`https://phekb.org/phenotypes?`). Phenotyping algorithms are proposed by one centre in the consortium and can be adapted and validated in other centres. The technicalities of the algorithms are different (e.g. the terminology used is ICD-9 rather than Read or ICD-10) and documents are stored as PDF files to be read by humans rather than in a computable standardised format. The advantage of having CALIBER codelists in a standardised format is that it is easy and automatic to extract relevant records from the master CALIBER database.

The phenotyping algorithms used in eMERGE are in some ways more advanced than those in CALIBER because they use more detailed hospital data. For example, they may incorporate algorithms based on echocardiogram or other imaging, or natural language processing using the free text in the medical record, which is not routinely available in CPRD or CALIBER.

The clinical codes website run by the University of Manchester (`https://clinicalcodes.rss.mhs.man.ac.uk/`) has some similarities with the CALIBER portal in

Figure 4.5: CALIBER phenotype algorithm for diabetes



that it stores codelists in a standardised format, but there are also some differences. The clinical codes website is open and anyone is free to download a codelist, unlike CALIBER which requires registration. The clincal codes website curates codelists in isolation without algorithms, and the codelists are not organised in a structured way. In contrast, the CALIBER portal is organised in a hierarchical way similar to ICD-10; this requires ongoing manual curation by the data management team.

These early steps should be built upon to create a wider open repository of EHR methods and codelists, which would greatly facilitate reproducible research using EHR.

Conclusions

The existence of the CALIBER portal and other tools has greatly facilitated data preparation using the CALIBER database, particularly for the research reported in this thesis. The next section describes a new method I have developed to assist in imputing missing data.

4.4 Methods for handling missing data

This section is a condensed version of a manuscript published in *American Journal of Epidemiology* [148], and describes the development and testing of a new method for multiple imputation which is particularly suited to epidemiological datasets. This Random Forest imputation method was used in the studies investigating differential leukocyte counts and incidence of cardiovascular diseases and mortality in this thesis (chapters 7, 8 and 9).

The program is published in the CALIBERrfimpute package [147] on the Comprehensive R Archive Network `http://cran.r-project.org/web/packages/CALIBERrfimpute/index.html`, and is starting to be used by other researchers. Random Forest imputation methods are starting to be used in diverse research areas such as imagery enhancements in cognitive behavioural group therapy [183], and exploring factors associated with pressure ulcers [184].

4.4.1 Abstract

Introduction: Multivariate Imputation by Chained Equations (MICE) is commonly used for imputing missing data in epidemiological research. The 'true' imputation model may contain non-linearities which are not included in default imputation models. Random Forest imputation is a machine learning technique which can accommodate non-linearities and interactions and does not require a particular regression model to be specified.

Aims: To develop a new Random Forest based MICE algorithm and compare it to parametric MICE.

Methods: A simulation study was performed by taking 1000 random samples of 2000 patients drawn from the 10 128 patients with stable angina in the CALIBER database and completely recorded covariates, artificially making some variables missing at random and comparing parameter estimates obtained using different imputation methods.

Results: Both MICE methods produced unbiased estimates of (log) hazard ratios but Random Forest was more efficient and produced narrower confidence intervals.

Conclusions: Random Forest imputation may be useful for imputing complex epidemiological datasets in which some patients have missing data.

4.4.2 Introduction

Missing data are a pervasive problem in epidemiological studies, particularly for research using clinically collected health records [185] such as CALIBER. Incomplete datasets are frequently analysed using multiple imputation, which involves creating multiple complete

versions of the dataset with missing values imputed by random draws from distributions inferred from observed data [186].

Multivariate Imputation by Chained Equations (MICE), also called Full Conditional Specification (FCS), is a common method of generating imputed values by drawing from estimated conditional distributions of each variable given all the others [187]. Imputation models must be appropriately specified for analyses based on imputed datasets to yield unbiased parameter estimates and associated standard errors. The default setting in implementations of MICE is for imputation models to include continuous variables as linear terms only with no interactions, but omission of important non-linear terms may lead to biased results [188]. Other potential problems with parametric regression models are that they cannot include more predictor variables than the number of observations without recourse to prior information [189], and inclusion of highly correlated variables may cause problems due to collinearity.

Random Forest is an extension of Classification And Regression Trees (CART) [190], which are predictive models that recursively subdivide the dataset based on values of the predictor variables. They do not rely on distributional assumptions and can accommodate non-linear relations and interactions [191]. Stekhoven et al. have developed a Random Forest-based algorithm for missing data imputation called missForest [192], but it aims to predict individual missing values accurately rather than taking random draws from a distribution, so it would not be appropriate for imputing values for use in further analysis.

The aim of this study was to develop a MICE imputation method using Random Forest, and test it in a realistically complex survival analysis based on patients with stable angina in the CALIBER database [1].

4.4.3 Methods

Missing data imputed using MICE where each variable is imputed using Random Forest

Within the MICE framework, missing values of continuous variables are conventionally imputed by fitting a linear regression model for the observed values, predicting the conditional mean for each missing value and randomly imputing a value from a normal distribution centred on this conditional mean. My new method uses Random Forest regression on a bootstrap sample of records with observed values of the variable to be imputed. Records with missing values in the dependent variable are imputed by random draws from independent normal distributions centred on conditional means predicted using Random Forest. I used the 'out-of-bag' mean square error [191] as the estimator of residual variance (which I assumed to be normally distributed).

For binary or unordered categorical variables, I used Random Forest to fit individual regression trees to a bootstrap sample of the data, and imputed each missing value as the prediction of a randomly chosen tree.

The Random Forest imputation functions are available on the Comprehensive R Archive Network [147].

Study population

The cohort for the current study was derived from an existing CALIBER study on stable angina. Patients were eligible to enter the cohort after 2001 when they had been registered for at least a year with a participating general practice. An 'index date' was defined for each patient as the date of the first record of stable angina during the period that they were eligible to enter the cohort. Stable angina was defined as two or more prescriptions of anti-anginal medication in CPRD, a record of percutaneous coronary intervention or coronary artery bypass grafting, or a diagnostic code for stable angina in CPRD or HES. Patients were followed up until the earliest of: non-fatal myocardial infarction (as recorded in HES or MINAP), death (as recorded in ONS) or transfer out of the general practice. Patients who died or experienced unstable angina or myocardial infarction within 7 days of cohort entry were excluded.

Measurement of variables

For continuous measures such as blood pressure and blood biomarkers, the mean measurement over the 2 years prior to the index date was used as the variable in modelling, and an auxiliary variable was generated containing the nearest value after the index date to use in imputation models. Serum creatinine, neutrophil counts and lymphocyte counts were log transformed using base 2 logarithms. Comorbidities were considered to be present if there was an appropriate diagnostic Read code in CPRD on or before the index date (or a record in MINAP or HES for previous myocardial infarction). Smoking status was identified in CPRD using the most recent smoking record prior to the index date if available, otherwise the earliest record after the index date. The year in which general practices started to receive blood results electronically was estimated by the annual number of lymphocyte count blood tests recorded.

Endpoints

The composite endpoint consisted of non-fatal myocardial infarction identified in HES or MINAP, or death from any cause identified in ONS. The type of event at the endpoint was not used in the survival model, but was used as an auxiliary variable in missingness and imputation models.

Survival analysis

The substantive analysis aimed to investigate the association between blood biomarkers and prognosis among patients with stable angina in CALIBER. The fully observed

predictor variables were: age, age squared, gender, previous myocardial infarction, diabetes, previous stroke, peripheral arterial disease, and heart failure. Smoking status was a partially observed 3-category variable (never, ex, current), and I included the following partially observed continuous variables: systolic blood pressure, log neutrophil count, log lymphocyte count, high density lipoprotein (HDL) cholesterol, haemoglobin concentration and log serum creatinine. For each of these variables I took the mean of any observed values in the 2 years prior to the start of follow-up. I used the Efron approximation for ties. I did not investigate alternative models and ignored clustering by general practice.

Generation of sample datasets for simulation study

I created datasets with missing data for which I knew the 'true' values, with a missingness pattern similar to that observed in the actual dataset but which was Missing At Random (MAR), such that the MAR assumption underlying most multiple imputation approaches was satisfied. I denoted the entire cohort of 52 576 stable angina patients as dataset 'A'. Patients with no missing values in any of the variables in the survival model were denoted dataset 'B' (13 308 patients) (**Figure 4.6 on page 94**).

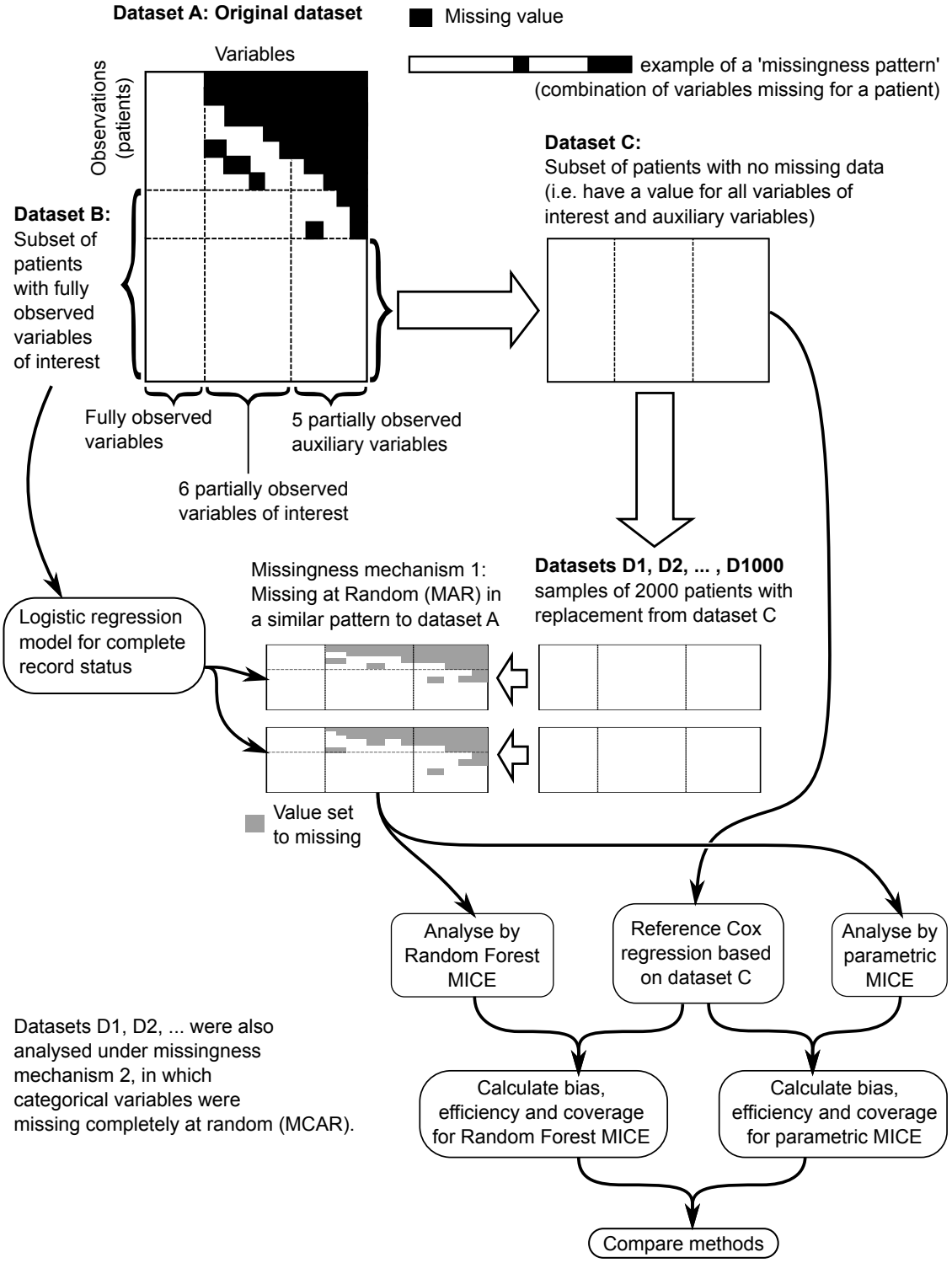
I used logistic regression based on completely observed variables to investigate factors associated with a patient having a complete record (i.e. whether a patient was in dataset B). The subset of patients with complete recording of all analysis and auxiliary variables were denoted dataset 'C' (10 128 patients), and they formed the basis for the resampling study. I drew 1000 random samples from dataset C, each with 2000 patients. I carried out sampling with replacement. I carried out simulations with one of two missingness mechanisms: (1) missing at random (MAR) in a similar pattern to dataset A, or (2) an artificial pattern of missingness completely at random (MCAR) in which only categorical variables were missing.

Multiple imputation of test datasets

Imputation models included all the variables in the substantive Cox model, event status, marginal Nelson-Aalen cumulative hazard [193] and the following auxiliary variables: type of endpoint, whether the practice was receiving electronic lab results, and the earliest recorded value after the index date of blood pressure and the five blood biomarkers. I imputed continuous variables using MICE with normal-based linear regression, predictive mean matching with 3 nearest neighbours ('PMM'), and my new Random Forest method with 10 or 100 trees ('MICE RF 10', 'MICE RF 100'). I imputed categorical variables using MICE with logistic or polytomous regression, or MICE with Random Forest (choice of 10 or 100 trees).

I generated 10 MICE imputations, each drawn from a separate chain with a different random seed, with 10 cycles of imputation before drawing the imputed dataset. In addition to MICE, I also evaluated missForest [192], generating multiple imputations by running missForest using different random seeds.

Figure 4.6: Schematic diagram showing generation of datasets with artificial missingness from a population of patients with stable angina in the CALIBER database.



Regardless of the imputation method, all datasets were analysed using the same multivariable semi-parametric Cox model described above. The log hazard ratios from the Cox model were combined using Rubin's rules [194], which assumes that imputed values were drawn from the appropriate Bayesian posterior. I carried out analyses using R 2.12.1 [157].

Comparison of results obtained by different methods

I considered the Cox proportional hazards model fitted to the entire dataset C as the 'true' result for the assessment of bias and confidence interval coverage of hazard ratios. I compared the length of 95% confidence intervals between two methods using paired sample t tests. I compared coverage of 95% confidence intervals for each coefficient separately using McNemar's test, defining discordant pairs as datasets in which the 95% confidence interval included the 'true' value for one method but not the other. I compared the efficiency of the estimators by calculating their empirical standard deviations.

4.4.4 Results

The prevalence of missingness among partially observed variables ranged from 1.5% for smoking to 56.7% for lymphocyte counts and 56.8% for neutrophil counts (**Table 4.6 on page 96**). The logistic regression model showed that patients with diabetes, peripheral arterial disease and previous stroke were more likely to have complete records (**Table 4.7 on page 97**). Coefficients from Cox models for patients with complete data (dataset C) were similar to the average results from full data analysis of the subsamples (as expected in the absence of small sample bias), so I compared imputation estimates to those from dataset C (**Table 4.8 on page 97**).

Estimates from complete record analysis were biased for some parameters under MAR (missingness mechanism 1); this may be expected because I had introduced missingness dependent on the outcome. The Random Forest MICE estimate for smoking (categorical) was biased towards the null (**Table 4.10 on page 99**, **Figure 4.8 on page 101**), but there was no material bias in other parameters estimated by Random Forest or parametric MICE. However, imputation using missForest produced materially biased estimates for all continuous variables missing at random (**Table 4.9 on page 98**, **Figure 4.7 on page 100**).

All the imputation methods tested produced more efficient parameter estimates than complete record analysis. MICE with Random Forest produced slightly more efficient estimates than parametric MICE, and the average between-imputation variance was also lower. Parametric MICE yielded confidence intervals with approximately 93–95% coverage. The mean lengths of confidence intervals were shorter using Random Forest MICE than parametric MICE ($P < 0.001$ for each comparison) but coverage was either equal or greater using Random Forest MICE (**Table 4.9 on page 98**, **Table 4.10 on page 99**).

Table 4.6: Characteristics of patients with complete records (dataset B) and patients with missing data in a CALIBER study of patients with stable angina (N = 52 576)

	Patients with complete records (in dataset B, N=13 308)			**Patients with missing data** (in dataset A but not B, N=39 268)		
Variable	Mean (SD)	%	% missing	Mean (SD)	%	% missing
Predictors						
Age (years)	68 (11)		0	69 (12)		0
Female gender		43.4	0		44.1	0
Smoking status			0			2.0
Never smoker		15.0			12.8	
Ex smoker		57.1			46.7	
Current smoker		27.9			38.5	
Medical history *						
Diabetes		22.3	0		13.9	0
Previous myocardial infarction		19.2	0		23.8	0
Previous stroke		10.0	0		9.0	0
Previous peripheral arterial disease		8.3	0		7.5	0
Previous heart failure		9.0	0		11.2	0
Mean of continuous variables over 2 years before index date:						
Mean creatinine (μmol/L)	98 (34)		0	101 (40)		48.4
Mean haemoglobin (g/dL)	14 (1.5)		0	14 (1.6)		63.3
Mean HDL cholesterol (mmol/L)	1.4 (0.43)		0	1.3 (0.45)		76.0
Mean lymphocytes ($\times 10^9$/L)	2.1 (1.1)		0	2.1 (1.6)		75.9
Mean neutrophils ($\times 10^9$/L)	4.4 (1.6)		0	4.7 (1.9)		76.0
Mean systolic blood pressure (mmHg)	142 (16)		0	144 (18)		13.3
Endpoints						
Follow-up in years	3.6 (2.5)			5.2 (2.9)		
Composite endpoint		13.4			26.7	
Non-fatal myocardial infarction		2.2			5.8	
Fatal coronary heart disease		3.4			6.2	
Death due to other cause		7.9			14.8	
Auxiliary						
Whether general practice receiving lab data		86.1	0		48.0	0
Earliest measurement of continuous variables after index date:						
Creatinine (μmol/L)	99 (40)		7.8	102 (44)		9.1
Haemoglobin (g/dL)	13 (1.7)		14.4	14 (1.7)		15.4
HDL cholesterol (mmol/L)	1.3 (0.41)		15.9	1.4 (0.51)		25.7
Lymphocyte count ($\times 10^9$/L)	2.1 (1.2)		15.3	2.1 (1.3)		20.2
Neutrophil count ($\times 10^9$/L)	4.5 (1.9)		15.3	4.6 (2.0)		20.1
Systolic blood pressure (mmHg)	138 (20)		4.6	140 (21)		5.3

* Medical history variables were completely observed because the absence of a recorded diagnosis was considered to represent absence of the condition

Table 4.7: Logistic regression model for factors associated with the outcome of having a complete record

Variable	Odds ratio	95% confidence interval
Age, per 10 years older	4.81 ***	4.02, 5.75
Age squared (10 years)^2	0.89 ***	0.87, 0.90
Female gender	1.08 **	1.03, 1.12
Diabetes	1.74 ***	1.64, 1.84
Peripheral arterial disease	1.24 ***	1.15, 1.35
Previous stroke	1.26 ***	1.17, 1.36
Heart failure	0.96	0.89, 1.04
Previous myocardial infarction	1.03	0.98, 1.09
Whether practice receives electronic lab results	4.74 ***	4.48, 5.01
Endpoint: fatal coronary heart disease	0.40 ***	0.36, 0.45
Endpoint: non-fatal myocardial infarction	0.29 ***	0.26, 0.34
Endpoint: non-coronary death	0.42 ***	0.39, 0.45
Cumulative hazard	0.03 ***	0.02, 0.03

Significance level: * P <0.05, ** P <0.01, *** P <0.001

Table 4.8: Cox proportional hazards model for death or non-fatal myocardial infarction among patients with stable angina in the CALIBER database: results from complete data (dataset C) and empirical average from 1000 samples of 2000 patients

	Complete dataset C		**Average from samples**		
Variable	HR	95% CI	HR	95% CI	Bias of log HR
Age, per 10 years older	1.07	0.53, 2.16	1.14	0.22, 5.89	0.0589
Age squared (10 years)^2	1.05	1.00, 1.10	1.05	0.93, 1.17	-0.0028
Female gender	0.80	0.69, 0.92	0.79	0.57, 1.09	-0.0143
Diabetes	1.60	1.40, 1.81	1.61	1.20, 2.17	0.0103
Peripheral arterial disease	1.51	1.29, 1.77	1.53	1.07, 2.21	0.0164
Previous stroke	1.34	1.15, 1.56	1.35	0.95, 1.92	0.0046
Heart failure	2.13	1.85, 2.45	2.16	1.56, 3.00	0.0143
Previous myocardial infarction	1.33	1.17, 1.51	1.34	1.00, 1.79	0.0059
Neutrophils per doubling	1.47	1.30, 1.66	1.47	1.11, 1.95	0.0029
Lymphocytes per doubling	0.81	0.72, 0.90	0.80	0.61, 1.04	-0.0068
Haemoglobin per g/dL	0.90	0.87, 0.94	0.90	0.81, 1.00	-0.0034
HDL cholesterol per mmol/L	1.07	0.91, 1.25	1.05	0.74, 1.51	-0.0158
Creatinine per doubling	1.80	1.54, 2.09	1.80	1.27, 2.57	0.0043
Systolic blood pressure per 10mmHg	1.00	0.97, 1.04	1.01	0.93, 1.09	0.0023
Smoking status:					
Never smoker	1	Reference	1	Reference	
Ex smoker	1.21	0.99, 1.47	1.22	0.77, 1.92	0.0105
Current smoker	1.39	1.13, 1.71	1.41	0.87, 2.29	0.0185

Table 4.9: Comparisons between imputation methods in 1000 samples with artificially introduced missingness in a pattern similar to the original dataset (missingness mechanism 1), with 25% complete records

Coefficient	Method	Mean bias	Z-score for bias	SD of estimate	Mean CI length	SD CI length	Coverage (%)	λ	σ_B^2
Neutrophils per doubling	Full data	0.0021	0.43	0.158	0.564	0.030	92.3	-	-
True log HR = 0.382	MICE normal	−0.038	−5.15	0.233	0.884	0.141	93.4	0.533	0.0244
	MICE PMM	−0.041	−5.68	0.231	0.889	0.144	93.4	0.528	0.0245
Empirical standard error	missForest	0.27	27.72	0.303	0.781	0.057	63.2	0.038	0.0014
of log HR = 0.144	MICE RF 10	−0.024	−4.55	0.165	0.798	0.097	97.9	0.368	0.0143
	MICE RF 100	0.05	8.33	0.190	0.804	0.083	96.0	0.294	0.0116
Lymphocytes per doubling	Full data	-6×10^{-3}	−1.23	0.155	0.526	0.028	91.7	-	-
True log HR = −0.217	MICE normal	8.4×10^{-4}	0.13	0.204	0.760	0.105	93.1	0.459	0.0157
	MICE PMM	0.0065	0.99	0.206	0.769	0.113	92.2	0.459	0.0162
Empirical standard error	missForest	−0.19	−22.21	0.270	0.724	0.050	72.5	0.036	0.0011
of log HR = 0.134	MICE RF 10	0.0028	0.56	0.156	0.727	0.081	97.8	0.338	0.0109
	MICE RF 100	−0.032	−5.61	0.178	0.721	0.066	95.6	0.253	0.0080
Haemoglobin per g/dL	Full data	−0.0036	−1.99	0.057	0.202	0.011	91.6	-	-
True log HR = −0.1	MICE normal	−0.0066	−2.73	0.076	0.279	0.036	92.6	0.410	0.0019
	MICE PMM	−0.0036	−1.47	0.077	0.279	0.037	92.7	0.412	0.0019
Empirical standard error	missForest	−0.056	−19.96	0.089	0.255	0.016	77.3	0.021	0.0001
of log HR = 0.0515	MICE RF 10	−0.01	−5.61	0.059	0.261	0.025	97.2	0.288	0.0012
	MICE RF 100	−0.022	−10.42	0.066	0.258	0.021	94.1	0.200	0.0008
HDL per mmol/L	Full data	−0.015	−2.44	0.199	0.718	0.069	92.5	-	-
True log HR = 0.0676	MICE normal	−0.019	−2.33	0.261	1.048	0.181	94.6	0.492	0.0321
	MICE PMM	−0.021	−2.43	0.272	1.063	0.206	94.5	0.486	0.0333
Empirical standard error	missForest	−0.092	−8.46	0.344	1.005	0.117	84.2	0.034	0.0021
of log HR = 0.183	MICE RF 10	−0.024	−4.06	0.188	0.954	0.122	98.2	0.336	0.0184
	MICE RF 100	−0.017	−2.41	0.220	0.968	0.116	97.0	0.263	0.0150
Creatinine per doubling	Full data	4×10^{-3}	0.66	0.193	0.708	0.057	93.0	-	-
True log HR = 0.585	MICE normal	0.035	4.85	0.231	0.843	0.098	92.5	0.236	0.0104
	MICE PMM	0.04	5.51	0.230	0.839	0.101	92.9	0.254	0.0111
Empirical standard error	missForest	0.028	3.33	0.262	0.825	0.084	88.2	0.013	0.0006
of log HR = 0.181	MICE RF 10	0.022	3.39	0.207	0.854	0.090	96.1	0.191	0.0086
	MICE RF 100	0.034	4.93	0.220	0.838	0.083	94.4	0.123	0.0053
Systolic BP per 10mmHg	Full data	0.0021	1.52	0.045	0.159	0.007	92.0	-	-
True log HR = 0.00412	MICE normal	-7.3×10^{-4}	−0.50	0.046	0.172	0.010	94.6	0.122	0.0002
	MICE PMM	−0.0013	−0.89	0.046	0.173	0.010	94.8	0.125	0.0002
Empirical standard error	missForest	−0.0035	−2.27	0.048	0.167	0.008	92.0	0.004	0.0000
of log HR = 0.0407	MICE RF 10	−0.0039	−2.80	0.044	0.170	0.009	95.7	0.091	0.0002
	MICE RF 100	−0.0045	−3.16	0.045	0.169	0.008	94.3	0.061	0.0001

Legend: λ, ratio of between-imputation to total variance; σ_B^2, between-imputation variance; SD, standard deviation

Table 4.10: Comparisons between methods for handling missing data in 1000 samples with categorical variables Missing Completely at Random (missingness mechanism 2), with 26% complete records

Coefficient	Method	Mean bias	Z-score for bias	SD of estimate	Mean CI length	SD CI length	Coverage (%)	λ	σ_B^2	% falsely classified
Previous myocardial infarction	Full data	6×10^{-3}	1.22	0.155	0.587	0.025	94.2	-	-	-
True log HR = 0.285	MICE logistic	−0.013	−2.46	0.169	0.682	0.053	95.4	0.229	0.0065	29.6
	missForest	0.0015	0.27	0.180	0.625	0.031	91.6	0.028	0.0007	17.3
Empirical standard error	MICE RF 10	−0.02	−4.21	0.149	0.663	0.046	97.3	0.197	0.0053	28.5
of log HR = 0.15	MICE RF 100	−0.02	−4.16	0.150	0.664	0.046	97.5	0.201	0.0054	28.5
Diabetes	Full data	0.011	2.30	0.157	0.592	0.025	93.6	-	-	-
True log HR = 0.467	MICE logistic	0.017	3.21	0.171	0.685	0.053	95.7	0.228	0.0065	32.0
	missForest	0.016	2.73	0.182	0.627	0.031	90.7	0.040	0.0010	19.7
Empirical standard error	MICE RF 10	−0.02	−4.25	0.149	0.668	0.048	97.4	0.206	0.0056	30.7
of log HR = 0.151	MICE RF 100	−0.022	−4.57	0.151	0.668	0.047	97.6	0.205	0.0056	30.7
Previous stroke	Full data	0.0054	0.86	0.198	0.707	0.044	94.1	-	-	-
True log HR = 0.295	MICE logistic	−0.0038	−0.58	0.207	0.828	0.079	95.5	0.245	0.0103	17.9
	missForest	0.0043	0.65	0.211	0.764	0.055	93.0	0.005	0.0002	8.4
Empirical standard error	MICE RF 10	−0.01	−1.79	0.183	0.808	0.071	97.9	0.216	0.0087	16.7
of log HR = 0.18	MICE RF 100	−0.0086	−1.49	0.183	0.806	0.071	97.8	0.213	0.0085	16.7
Peripheral arterial disease	Full data	0.016	2.59	0.198	0.730	0.045	93.7	-	-	-
True log HR = 0.412	MICE logistic	−0.0015	−0.21	0.218	0.857	0.080	94.9	0.254	0.0114	15.5
	missForest	0.029	4.18	0.223	0.787	0.058	91.7	0.006	0.0002	7.0
Empirical standard error	MICE RF 10	0.0057	0.94	0.192	0.834	0.076	97.1	0.219	0.0094	14.5
of log HR = 0.186	MICE RF 100	0.0047	0.78	0.192	0.830	0.073	96.9	0.213	0.0090	14.5
Heart failure	Full data	0.015	2.47	0.190	0.653	0.034	91.8	-	-	-
True log HR = 0.756	MICE logistic	0.015	2.22	0.207	0.759	0.065	93.9	0.230	0.0081	14.6
	missForest	5.7×10^{-4}	0.08	0.216	0.696	0.042	89.4	0.015	0.0004	7.2
Empirical standard error	MICE RF 10	−0.035	−5.78	0.189	0.747	0.059	95.5	0.204	0.0069	13.7
of log HR = 0.167	MICE RF 100	−0.032	−5.38	0.190	0.747	0.060	95.4	0.205	0.0070	13.7
Current smoker vs. never	Full data	0.022	2.62	0.263	0.970	0.077	94.0	-	-	-
True log HR = 0.327	MICE logistic	0.024	2.65	0.291	1.094	0.108	94.1	0.194	0.0141	52.4
	missForest	−0.035	−3.56	0.307	1.062	0.096	91.6	0.027	0.0018	35.0
Empirical standard error	MICE RF 10	−0.097	−12.92	0.237	1.072	0.104	95.6	0.224	0.0158	50.0
of log HR = 0.247	MICE RF 100	−0.096	−12.68	0.239	1.072	0.106	95.3	0.222	0.0157	50.0
Ex smoker vs. never	Full data	0.013	1.66	0.246	0.909	0.080	93.7	-	-	-
True log HR = 0.187	MICE logistic	−0.0069	−0.82	0.265	1.023	0.110	94.2	0.193	0.0123	52.4
	missForest	0.045	5.34	0.269	0.981	0.102	93.3	0.027	0.0016	35.0
Empirical standard error	MICE RF 10	−0.059	−8.81	0.211	1.001	0.105	97.1	0.229	0.0141	50.0
of log HR = 0.232	MICE RF 100	−0.059	−8.71	0.212	1.001	0.109	97.1	0.228	0.0142	50.0

Legend: λ, ratio of between-imputation to total variance; σ_B^2, between-imputation variance; SD, standard deviation

Figure 4.7: Boxplots showing bias of estimates of log hazard ratios for partially observed continuous variables with data missing at random (missingness mechanism 1) in 1000 samples of patients with stable angina in the CALIBER database.
The solid horizontal line is the 'true' log hazard ratio from the full dataset C; the dashed lines are ± 1 empirical standard error.

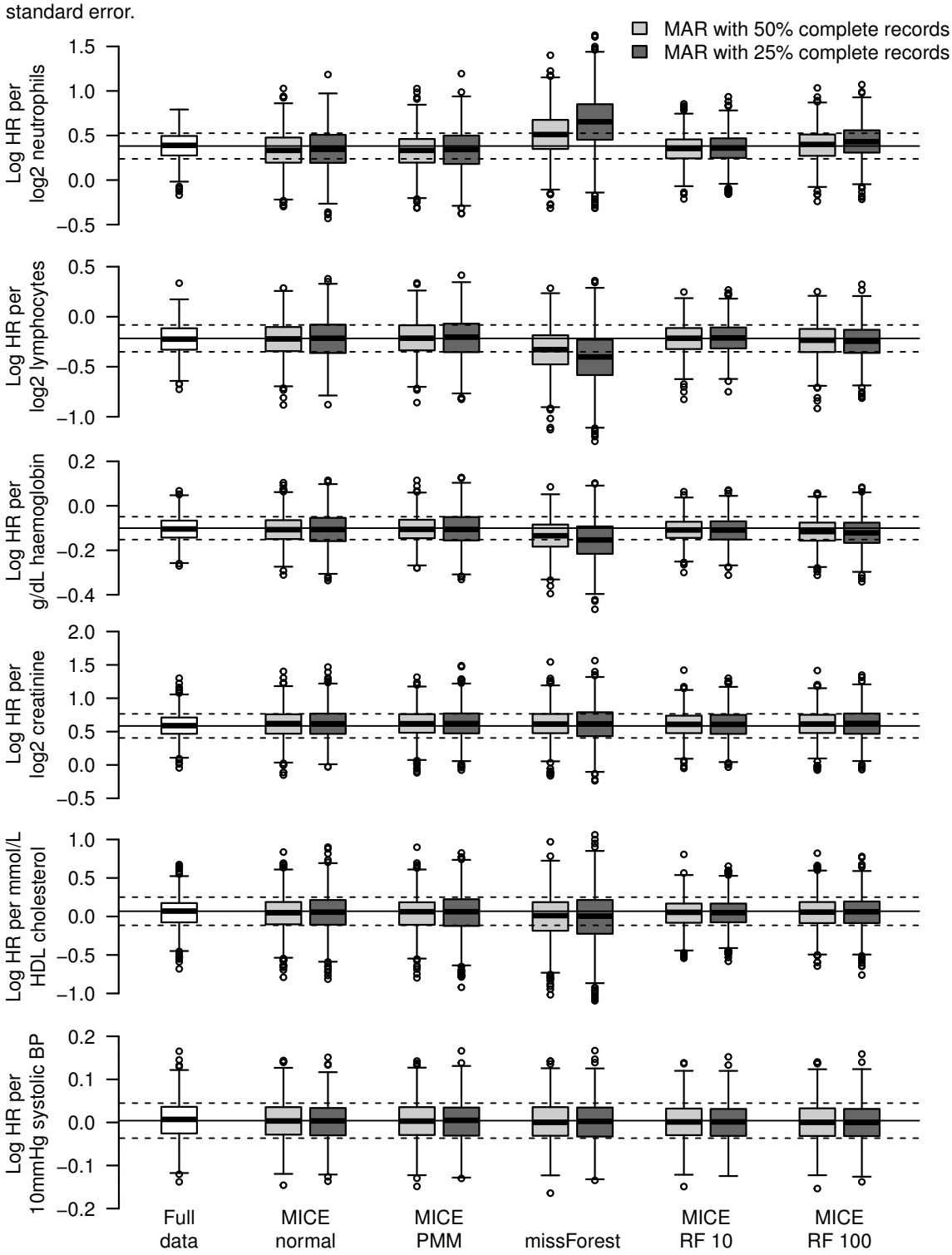
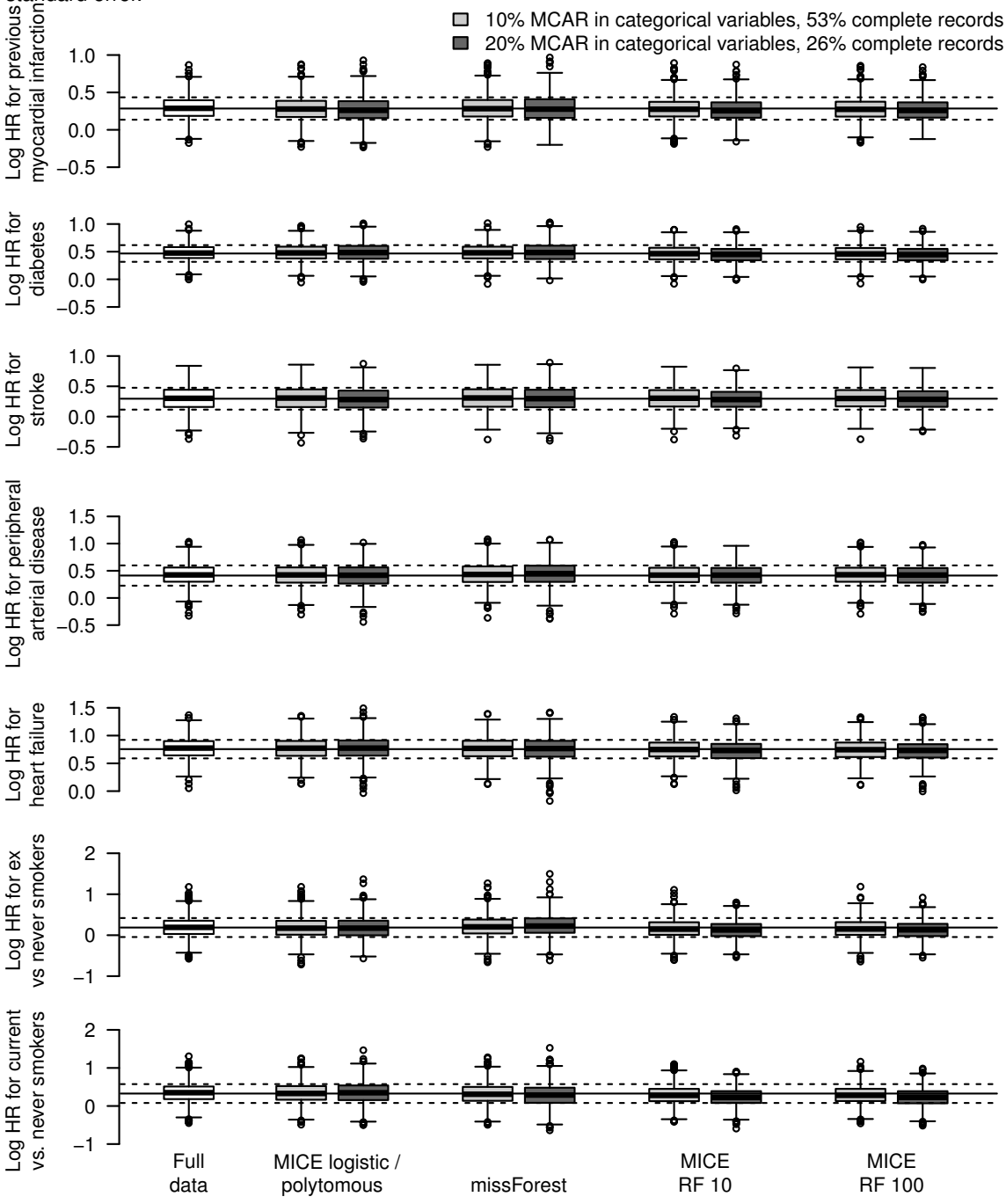


Figure 4.8: Boxplots showing bias of estimates of log hazard ratios for partially observed categorical variables missing completely at random (missingness mechanism 2) in 1000 samples of patients with stable angina in the CALIBER database. The solid horizontal line is the 'true' log hazard ratio from dataset C; the dashed lines are ± 1 empirical standard error.



For categorical variables, missForest produced imputed values which were more likely to be equal to the 'true' (observed) value than the MICE methods, but confidence intervals were too small with below nominal coverage, and between-imputation variance was very small (**Table 4.10 on page 99**, **Figure 4.8 on page 101**). There was no difference in bias, precision or coverage between normal-based MICE and predictive mean matching (**Table 4.9 on page 98**). Random Forest MICE with 100 trees for continuous variables produced estimates with slightly narrower confidence intervals than with 10 trees, but with greater bias, worse coverage of 95% confidence intervals and 10 times the computational cost (**Table 4.9 on page 98**). For categorical variables, Random Forest MICE with 10 or 100 trees produced almost identical results (**Table 4.10 on page 99**).

4.4.5 Discussion

Summary of main findings

In this resampling study of methods for handling missing data, parametric and Random Forest MICE produced estimates with no material bias for a Cox model on data with artificially introduced MAR missingness. Random Forest-based MICE produced more efficient estimates and narrower confidence intervals than parametric MICE, yet in some cases coverage probability was greater than 95%, suggesting that some confidence intervals may be conservative. A possible explanation for the efficiency gain with Random Forest MICE is that it was able to make better use of the available information by accommodating non-linearities among the predictors. Random Forest imputation may be useful for imputing complex epidemiological datasets in which some patients have missing data.

Imputation methods for MICE

It is important that imputation models are correctly specified for analyses to yield unbiased estimates, and Random Forest may help avoid the bias that can occur with parametric MICE if the latter's imputation models are mis-specified. Standard parametric MICE performed well, suggesting that the true imputation models did not contain significant non-linearities or interactions, and hence Random Forest did not confer an advantage from the perspective of bias. The default settings for MICE do not include interactions between the variables and it is routine practice to include only those interactions that are in the substantive model, rather than actively searching for all possible interactions and non-linearities. Imputation models should also be compatible with the substantive model [195], and Random Forest avoids the need to specify how the outcome should be conditioned on in the imputation models for covariates.

A disadvantage of Random Forest is that the 'models' are complex and not easily interpretable, although arguably this is not a shortcoming for the purpose of imputation. My method assumes that the residuals from the Random Forest regression are normally distributed with constant variance.

On these 2000-patient datasets, computation time was 3 times as long for Random Forest MICE with 10 trees compared to parametric MICE, but Random Forest may yield a saving in analyst time because there is theoretically less need for transformation of fully observed variables or investigation of non-linearities and interactions. It is also possible to include a large number of related predictor variables in Random Forest models without encountering problems due to collinearity.

MissForest was included as an example of algorithms for completing single datasets [192], but it replaces missing values with predicted values rather than drawing from a distribution, leading to biased parameter estimates.

Limitations

Although this study has strengths (it was based on a real dataset, and the analysis was realistically complex) it also has important limitations. The most important limitation in producing general recommendations is that it was based on a single analysis of a single study, so results should be generalised to other datasets with caution.

A limitation of the resampling methodology was that in order to avoid excessive computing time I used only ten imputations for most of the comparisons, leading to noisy estimates of between-imputation variability. To save time I also restricted the number of cycles of MICE to ten, and although I looked at plots of chain means and standard deviations between cycles for a few runs, this is a crude way of assessing chain convergence and it is possible that the chains may not have converged by the end of every run.

I ignored practice-level clustering at the imputation and analysis stage, for simplicity, but could have properly accounted for this by hierarchical models [196].

Recommendations for further development

Random Forest multiple imputation should be tested on a larger range of datasets and in simulations to explore whether it gives unbiased estimates where there are non-trivial non-linearities or interactions in imputation models, such that a standard parametric MICE which ignores them gives biased results. Random Forest tuning parameters (such as the number of trees and number of nodes) should be further investigated.

4.4.6 Conclusions

Standard parametric MICE and my new Random Forest MICE method work reasonably well under artificially introduced missingness at random in a realistically complex dataset. Random Forest imputation should be further investigated in situations where MICE with default parametric imputation models produces biased results.

4.5 Overall summary of chapter

In this chapter I have presented work on validation and novel tools for data management and statistical analysis. Although all these projects were based on the CALIBER dataset, the methods and principles have much wider application.

Specifically for this project, the work on cross-referencing myocardial infarction records was used to refine the CALIBER phenotyping algorithm for myocardial infarction, such as the selection of Read codes to use. The data management and codelist packages were used for defining variables and manipulating data. The Random Forest missing data method was applied to the dataset used in the main analysis alongside conventional parametric imputation in chapters 7, 8 and 9; the similarity of the results obtained using Random Forest and parametric MICE helps to validate the method, and suggests that it should be further developed and deployed in the future.

The next chapter describes the cohort specification for investigating the onset of cardiovascular disease in CALIBER, and applies this to a study investigating type 2 diabetes and the twelve initial presentations of cardiovascular disease.

5 Investigating the initial presentation of cardiovascular diseases in linked electronic health records: an example of type 2 diabetes

5.1 Chapter outline

This chapter describes the approach to studying the initial presentation of cardiovascular disease in CALIBER, including the cohort definition and ascertainment of endpoints in linked record sources. This is applied to the example of type 2 diabetes. The work described in this chapter has been published in *Lancet Diabetes and Endocrinology* [197].

5.2 Abstract

Background: The contemporary associations of type 2 diabetes with a wide range of incident cardiovascular diseases have not been compared. I aimed to study associations between type 2 diabetes and 12 initial manifestations of cardiovascular disease.

Methods: I used linked primary care, hospitalisation, disease registry and death certificate records from 1997 to 2010 in the CALIBER programme. The cohort included 1.92 million people in England aged 30 and older who were free from cardiovascular disease at baseline. I compared cumulative incidence curves for the initial presentation of cardiovascular disease and used Cox models to estimate cause-specific hazard ratios. This study is registered at clinicaltrials.gov (NCT01804439).

Results: 113 638 first presentations of cardiovascular disease occurred during a median follow-up of 5.5 years. Among the 34 198 people with type 2 diabetes, 6137 experienced a first cardiovascular presentation, of which the most common were peripheral arterial disease (16.2%) and heart failure (14.1%). Type 2 diabetes was positively associated with peripheral arterial disease (adjusted hazard ratio (HR) 2.98; 95% confidence interval (CI) 2.76, 3.22), ischaemic stroke (HR 1.72, 95% CI 1.52, 1.95), stable angina (HR 1.62, 95% CI 1.49, 1.77), heart failure (HR 1.56, 95% CI 1.45, 1.69), and non-fatal myocardial infarction (HR 1.54, 95% CI 1.42, 1.67), but was inversely associated with abdominal aortic aneurysm (HR 0.46, 95% CI 0.35, 0.59) and subarachnoid haemorrhage (HR 0.48, 95% CI 0.26, 0.89), and not associated with arrhythmia or sudden cardiac death (HR 0.95, 95% CI 0.76, 1.19). With the exception of non-fatal myocardial infarction in younger patients, where the association with T2D was stronger in women, there were no other significant sex differences in these associations.

Conclusions: Heart failure and peripheral arterial disease are the most common initial manifestations of cardiovascular disease in type 2 diabetes. The differences between relative risks of different cardiovascular diseases in patients with type 2 diabetes have implications for clinical risk assessment and trial design.

5.3 Introduction

Type 2 diabetes has become an increasingly common disease estimated to affect 380 million people worldwide by 2025 [198]. People with type 2 diabetes (T2D) are at increased risk of cardiovascular diseases and their clinical complications [199].

Previous studies have not compared the relationship between T2D and a wide range of cardiovascular outcomes such as heart failure, peripheral arterial disease, abdominal aortic aneurysm, and ventricular arrhythmias in the same study, but focussed instead on a narrower range of disease outcomes, usually just one or two [200]. Such comparisons require large studies to reliably estimate associations for rare outcomes. Several small studies have shown an impact of T2D on cardiovascular manifestation by outcome [44] and sex [201], but these studies were not designed to reliably assess associations with multiple cardiovascular diseases or across population subgroups.

In this project I investigated twelve cardiovascular diseases that are important causes of morbidity and mortality: coronary artery disease (stable angina, unstable angina, fatal and non-fatal myocardial infarction), cerebrovascular disease (transient ischaemic attack, ischaemic stroke, haemorrhagic stroke, subarachnoid haemorrhage), arrhythmias, heart failure, peripheral arterial disease and abdominal aortic aneurysm. The rationale for choosing these conditions is described earlier (section 1.5 on page 28), but briefly they are a set of serious, symptomatic cardiovascular conditions that share some aspects of pathogenesis (atherosclerosis, thrombosis and inflammation).

The aim of this study was to investigate and compare associations between T2D and future risk of twelve of the most common initial cardiovascular presentations in men and women in the CALIBER dataset.

5.4 Methods

5.4.1 Study population

The study population was drawn from the CALIBER programme [1], as described in chapter 3. The study period was January 1997 to March 2010, and individuals were eligible for inclusion when they were aged 30 or over and had been registered for at least one year in a practice which met research data recording standards.

Follow-up was censored at the occurrence of an endpoint, death, de-registration from the practice or the last data collection for the practice, whichever occurred first. Individuals with a prior history of cardiovascular disease (defined as any of the endpoints, or a non-

specific entry of 'cardiovascular disease' or 'atherosclerotic disease' in any of the data sources) were excluded.

5.4.2 Diabetes status

I developed a phenotyping algorithm to identify patients with type 2 diabetes at baseline based on coded diagnoses recorded in CPRD or HES on or before study entry; for more details see section 4.3.6 on page 85. Lists of Read and ICD-10 diagnosis codes which convey diabetic status are curated on the CALIBER data portal (`https://www.caliberresearch.org/portal/chapter/6`).

Patients who developed new onset diabetes during follow-up were analysed according to their baseline status of no diabetes. This study compared people with T2D to those without diabetes, excluding people with type 1 diabetes or diabetes of uncertain type.

5.4.3 Covariates

For continuous variables (body mass index (BMI), high density lipoprotein (HDL) cholesterol, total cholesterol, and systolic blood pressure) I used the most recent measurement in CPRD in the year prior to study entry as the baseline value, but included measurements outside this time window in imputation models. Social deprivation was included in models as quintiles of the index of multiple deprivation [131], described in subsection 3.3.5 on page 58. Data recorded prior to study entry was used to classify participants as never, ex or current smokers at baseline.

5.4.4 Endpoints

The primary endpoint was the first record of one of twelve cardiovascular presentations in any of the data sources. Any events occurring after the first cardiovascular presentation were ignored. If more than one endpoint occurred on the same day, the top endpoint in **Table 5.1 on page 108** was chosen. Secondary outcomes included cardiovascular mortality and all-cause mortality.

The definition of each endpoint was formalised as a CALIBER phenotype, similar to the diabetes example (section 4.3.6 on page 85)

Endpoint definitions

Cardiovascular phenotype definitions based on the CALIBER data sources are curated on the CALIBER data portal (www.caliberresearch.org/portal).

1. Arrhythmia or sudden cardiac death. `https://caliberresearch.org/portal/show/phenotype_scd`

Table 5.1: Data sources for initial presentation of cardiovascular diseases (study endpoints)

Priority	Endpoint	Type	Source
1	Arrhythmia, cardiac arrest or sudden cardiac death	Fatal or non-fatal	CPRD, HES, ONS
2	Heart failure	Fatal or non-fatal	CPRD, HES, ONS
3	Unheralded coronary death	Fatal	ONS
4	Myocardial infarction	Non-fatal	CPRD, MINAP, HES
5	Unstable angina	Non-fatal	CPRD, MINAP, HES
6	Stable angina	Non-fatal	CPRD
7	Coronary disease not otherwise specified	Non-fatal	CPRD, HES
8	Abdominal aortic aneurysm	Fatal or non-fatal	CPRD, HES, ONS
9	Peripheral arterial disease	Fatal or non-fatal	CPRD, HES, ONS
10	Subarachnoid haemorrhage	Fatal or non-fatal	CPRD, HES, ONS
11	Intracerebral haemorrhage	Fatal or non-fatal	CPRD, HES, ONS
12	Ischaemic stroke	Fatal or non-fatal	CPRD, HES, ONS
13	Stroke not otherwise specified	Fatal or non-fatal	CPRD, HES, ONS
14	Transient ischaemic attack	Non-fatal	CPRD, HES
15	Death due to other cause	Fatal	ONS

A composite of ventricular arrhythmias, implantable cardioverter defibrillator, and sudden cardiac death. It was defined using diagnoses and procedure codes in primary care, secondary care and death certificates.

The definition used in this type 2 diabetes study also included procedure codes for cardioversion. The codes for cardioversion were subsequently removed as the majority of patients undergoing cardioversion had atrial fibrillation rather than a ventricular arrhythmia. The updated definition, as used in the leukocyte study (chapters 7 and 8), does not include cardioversion as a component of the endpoint, and has been renamed 'Ventricular arrhythmia or sudden cardiac death'.

2. Heart failure. `https://caliberresearch.org/portal/show/phenotype_hf`

Defined by coded diagnoses in primary care, secondary care and death certificates.

3. Unheralded coronary death. `https://caliberresearch.org/portal/show/phenotype_ucd`

Death with the primary cause certified as coronary heart disease, and no prior history of cardiovascular disease. Patients with myocardial infarction who died on the day of their infarct were considered to have unheralded coronary death.

4. Non-fatal myocardial infarction. `https://caliberresearch.org/portal/show/phenotype_mi`

Defined as a disease registry diagnosis of an acute coronary syndrome with elevated troponin, or a primary or secondary care diagnosis of myocardial infarction.

The definition used in this type 2 diabetes study was an older definition, which was updated based on the results of the validation study (section 4.2). The leukocyte study (chapters 7 and 8) used the updated definition.

5. Unstable angina. `https://caliberresearch.org/portal/show/phenotype_ua`

Defined as a primary or secondary care diagnosis of unstable angina, or an acute coronary syndrome without myocardial infarction recorded in the disease registry.

6. Stable angina. `https://caliberresearch.org/portal/show/phenotype_sa`

Defined by a coded diagnosis in primary or secondary care of ischaemic chest pain or stable angina, a positive myocardial ischaemia test, two or more prescriptions of antianginal medication, or coronary revascularisation. Information from primary and secondary care data sources were combined as shown in the flowchart **Figure 3.3 on page 63**.

7. Coronary disease not further specified. `https://caliberresearch.org/portal/show/phenotype_chd_nos`

A non-specific diagnosis of ischaemic or coronary heart disease in primary or secondary care that does not fall into one of the more specific categories. It was not included among the twelve diseases in the main displays of hazard ratios, but for cumulative incidence calculations it was combined with unstable angina.

8. Abdominal aortic aneurysm. `https://caliberresearch.org/portal/show/phenotype_aaa`

Defined by coded diagnoses and procedures in primary care, secondary care and death certificates.

9. Peripheral arterial disease. `https://caliberresearch.org/portal/show/phenotype_pad`

This includes intermittent claudication, limb ischemia or gangrene due to atherosclerotic disease in the arteries of the legs. It was defined by coded diagnoses and procedures in primary care, secondary care and death certificates.

10. Subarachnoid haemorrhage. `https://caliberresearch.org/portal/show/phenotype_stroke_subarachnoid`

Defined by coded diagnoses in primary care, secondary care and death certificates.

11. Intracerebral haemorrhage. https://caliberresearch.org/portal/show/phenotype_stroke_intracerebral_haem

Defined by coded diagnoses in primary care, secondary care and death certificates.

12. Ischaemic stroke. https://caliberresearch.org/portal/show/phenotype_stroke_ischaemic

Defined using coded diagnoses in primary care, secondary care and death certificates. Patients with a procedure code for carotid endarterectomy within 90 days of a stroke of unspecified type were considered to have ischaemic stroke.

13. Stroke not further specified. https://caliberresearch.org/portal/show/phenotype_stroke_nos

Diagnosis of stroke which does not state it is ischaemic or haemorrhagic. This clinical event was not included among the twelve diseases in the main displays of hazard ratios, but for cumulative incidence calculations it was combined with ischaemic stroke.

14. Transient ischaemic attack. https://caliberresearch.org/portal/show/phenotype_tia

Defined by coded diagnoses in primary or secondary care.

5.4.5 Statistical analysis

To describe the incidence of each initial presentation over time I constructed cumulative incidence curves, taking into account the other possible initial presentations as competing events. Normal-based confidence intervals were constructed based on Greenwood's variance formula, as implemented in the R prodlim package [202].

I carried out survival analysis to describe and model the first occurrence of any cardiovascular disease. A patient's follow-up ended when they experienced one of the cardiovascular endpoints or when they were censored. Subsequent events (e.g. myocardial infarction occurring after stable angina) were not analysed.

I considered using either Cox models (for modelling cause-specific hazards) or the Fine and Gray model (to compare cumulative incidence curves by modelling subdistribution hazards). As the aim of this study was observational epidemiology – to explore associations rather than predict risk – I decided to use the Cox model, as cause specific hazard ratios were appropriate quantities to estimate. The Fine and Gray model compares cumulative incidence curves but the subdistribution hazard ratios that it estimates do not have a straightforward interpretation, unlike cause-specific hazards. However, cause specific hazards cannot be used to predict cumulative incidence.

I used follow-up time as the timescale for the Cox models in order to investigate interactions with age, and I verified the proportional hazards assumption by plotting Schoenfeld residuals against transformed time.

In the primary analysis, I compared individuals with T2D to those without diabetes, adjusting for age, sex, body mass index, deprivation, HDL cholesterol, total cholesterol, systolic blood pressure, smoking status, and prescription of statins or antihypertensive medication in the year before study entry. The baseline hazard function of each model was stratified by general practice and sex, and missing covariate data were handled using multiple imputation. I also carried out analyses adjusted for age and sex only, and adjusted for age, sex and cardiovascular risk factors. I evaluated interactions with age and sex.

Among patients with type 2 diabetes, I carried out a secondary analysis of the association of glycaemic control with cardiovascular endpoints. Glycaemic control was measured by the mean HbA1c from 3 years before to 3 years after study entry, ignoring values occurring after an endpoint. I categorised HbA1c with cutpoints at 48 mmol/mol (= 6.5%, the threshold for diagnosis of diabetes) [203], and 58 mmol/mol (= 7.5%, the recommended threshold for initiation of thiazolidinedione or insulin in T2D in UK guidelines) [204].

5.4.6 Multiple imputation

Missing data were handled by Multiple Imputation by Chained Equations implemented in the mice package in R [187]. Imputation models were estimated separately for men and women and included:

- All the baseline covariates used in the main analysis (age, quadratic age, sex, body mass index, deprivation, high density lipoprotein cholesterol, total cholesterol, systolic blood pressure, smoking status, statins and antihypertensive prescriptions).
- Prior (between 14 and 4 years before study entry) and post (between 0 and 1 year after study entry) averages of continuous covariates in the main analysis.
- Baseline measurements of covariates not considered in the main analysis (diastolic blood pressure, alcohol intake, white cell count, haemoglobin, creatinine and alanine aminotransferase).
- Baseline medications (low-dose aspirin, loop diuretics, oral contraceptives and hormone replacement therapy).
- Coexisting medical conditions (history of depression, cancer, renal disease, liver disease and chronic obstructive pulmonary disease).
- The Nelson-Aalen hazard and the event status for each endpoint analysed in the data [193].

Non-normally distributed variables were log-transformed for imputation and exponentiated back to their original scale for analysis. Five multiply imputed datasets were generated, and Cox models were fitted to each dataset. Coefficients were combined using Rubin's rules.

The Random Forest imputation method described in section 4.4 was being developed in parallel and was not ready for use at the time of this study. However, it was used in the remaining studies in this thesis (chapters 7, 8 and 9).

5.4.7 Sensitivity analyses

Fifteen percent of people with diabetes did not have a code for their type of diabetes (**Figure 5.1 on page 113**) and were excluded from the primary analyses. To assess the impact of this potential misclassification bias, I performed a sensitivity analysis comparing individuals with any diabetes diagnosis to those without diabetes (the majority of diabetic patients in a cohort of this age would have T2D). I performed sensitivity analyses ignoring endpoints recorded only in primary care (from CPRD), restricted to fatal endpoints, or restricted to individuals who entered the study after 2004 (when recording of covariates would be expected to be more complete following introduction of the Quality and Outcomes Framework) [205]. Data were analysed using R 2.15 [157].

5.5 Results

5.5.1 Baseline characteristics of study participants

The study cohort included 1 887 062 participants without diabetes and 34 198 with type 2 diabetes (**Figure 5.1 on page 113**). People with T2D had lower mean HDL cholesterol and higher mean BMI than people without diabetes. The use of statins and antihypertensive medication was higher in people with T2D and increased over time. People with T2D were approximately twice as likely to have Black or South Asian ethnicity as those without diabetes (**Table 5.2 on page 114**). During follow-up, 51 690 participants were newly diagnosed with diabetes, with a median time to diagnosis of 4.9 years.

5.5.2 Comparison of associations between type 2 diabetes and twelve initial cardiovascular presentations

I analysed 113 638 cardiovascular events over 11.6 million person-years of follow-up (median follow-up 5.5 years). Of these, 8056 events were in people with type 2 diabetes, who accrued 162 756 person-years of follow-up. Peripheral arterial disease and heart failure were the most common endpoints among people with T2D (16.2% and 14.1%) (**Figure 5.2 on page 115**); the proportions of these endpoints in people without diabetes were 9.4% and 12.2% (**Table 5.3 on page 115**). Cumulative incidence curves showed substantial differences in the direction and strength of the associations between each of the cardiovascular manifestations and T2D (**Figure 5.3 on page 116** and **Table 5.4 on page 117**). Peripheral arterial disease showed the largest difference of all the diseases in cumulative incidence; a woman with T2D at age 40 had 9.7% risk (95% CI 8.4, 11.1) of developing peripheral arterial disease as her first presentation of cardiovascular disease by age 80; a woman without diabetes had a 3.2% risk (95% CI 3.1, 3.3). The corresponding figures for men were 11.7% (95% CI 10.5, 13.0) and 4.5% (95% CI 4.4, 4.7).

Figure 5.1: Flow of patients through the study

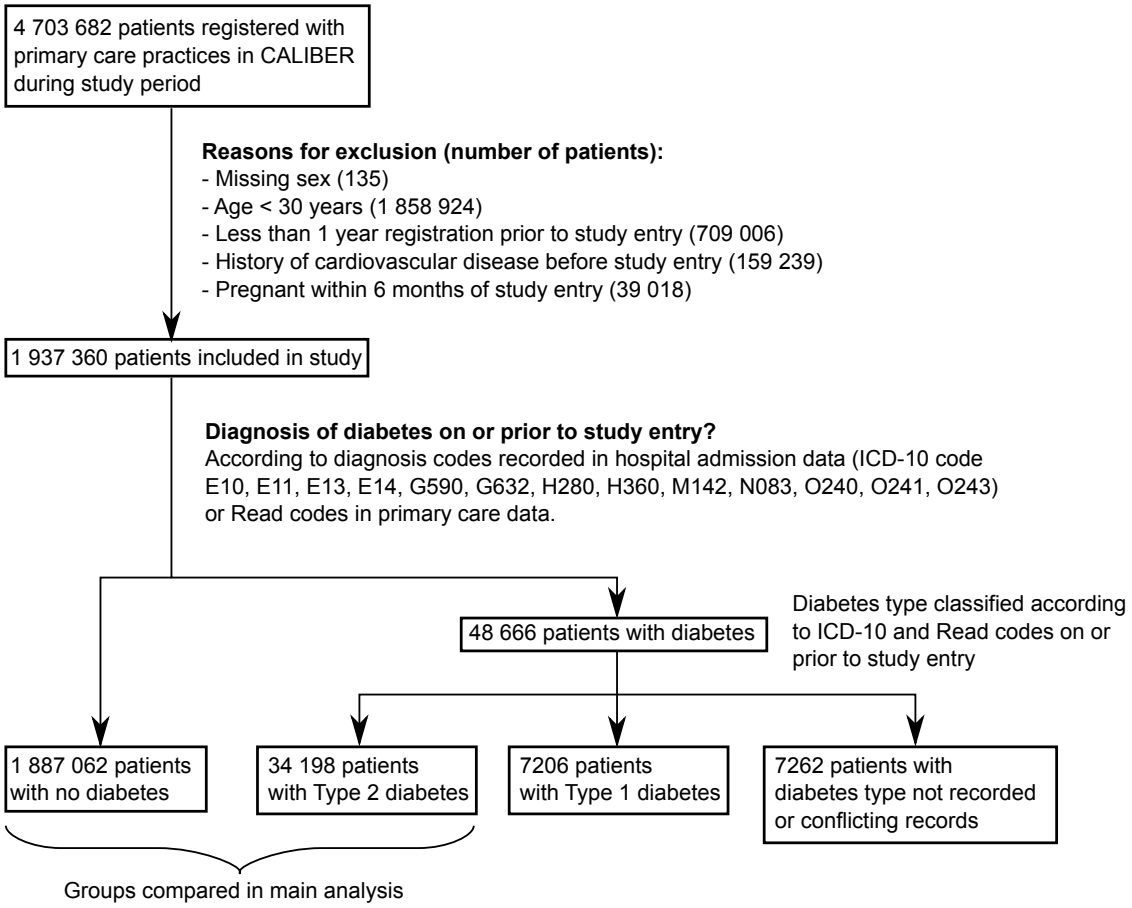


Table 5.2: Baseline patient characteristics among 1 921 260 people without cardiovascular diseases at baseline, according to type 2 diabetes status

	Women		**Men**	
	No diabetes	Type 2 diabetes	No diabetes	Type 2 diabetes
Number of patients	955 655	15 805	931 407	18 393
Age in years, mean (SD)	48.2 (16.2)	64.7 (14.4)	45.4 (14)	60.8 (13)
Most deprived quintile, n (%)	184 942 (19.4%)	4313 (27.3%)	189 247 (20.3%)	4285 (23.3%)
Ethnicity, n (%): White	503 274 (90.8%)	9641 (81.8%)	388 763 (90.5%)	10 716 (82.4%)
South Asian	14 673 (2.6%)	912 (7.7%)	12 865 (3.0%)	1041 (8.0%)
Black	16 722 (3.0%)	748 (6.3%)	13 148 (3.1%)	670 (5.2%)
Other	19 328 (3.5%)	491 (4.2%)	14 640 (3.4%)	579 (4.5%)
Smoking, n (%): Current	130 946 (17.6%)	2028 (14.4%)	148 020 (23.7%)	3327 (20.6%)
Ex	108 593 (11.4%)	2927 (18.5%)	109 538 (11.8%)	5507 (29.9%)
Never	506 383 (67.9%)	9158 (64.9%)	366 729 (58.7%)	7327 (45.3%)
Values of continuous measurements, most recent within 1 year prior to study entry, mean (SD)				
Systolic blood pressure*, mmHg	127 (19.6)	141 (19.7)	133 (17.2)	139 (17.6)
Diastolic blood pressure*, mmHg	77.1 (10.3)	79.5 (10.1)	80.3 (10.0)	80.5 (10.0)
Total cholesterol*, mmol/l	5.54 (1.14)	5.08 (1.22)	5.39 (1.11)	4.77 (1.16)
HDL cholesterol*, mmol/l	1.57 (0.45)	1.34 (0.39)	1.28 (0.37)	1.14 (0.33)
Body mass index*, kg/m^2	26.0 (5.59)	30.5 (6.76)	26.5 (4.43)	29.3 (5.43)
Random glucose*, mmol/l	5.35 (1.16)	10.4 (4.93)	5.56 (1.29)	10.7 (4.78)
Fasting glucose*, mmol/l	5.14 (0.75)	8.72 (3.53)	5.31 (0.79)	9.1 (3.51)
Glycaemic control**				
HbA1c recorded, n (%)	12 835 (1.3%)	12 187 (77.1%)	12 031 (1.3%)	14 435 (78.5%)
HbA1c, mean (SD) mmol/mol	44.3 (14.7)	59.3 (18)	46.9 (16.3)	60.5 (18.1)
HbA1c $<$ 48 mmol/mol, n (%)	9468 (73.8%)	3445 (28.3%)	8030 (66.7%)	3671 (25.4%)
HbA1c 48-58 mmol/mol, n (%)	1854 (14.4%)	3458 (28.4%)	2004 (16.7%)	4075 (28.2%)
HbA1c $\geq$ 58 mmol/mol, n (%)	1513 (11.8%)	5284 (43.4%)	1997 (16.6%)	6689 (46.3%)
Diabetes treatment in year before study entry, %				
Diet alone		4591 (29.0%)		5467 (29.7%)
Metformin	884 (0.1%)	7714 (48.8%)	106 (0.0%)	8764 (47.6%)
Sulphonylurea	133 (0.0%)	6241 (39.5%)	124 (0.0%)	7526 (40.9%)
Insulin	189 (0.0%)	1939 (12.3%)	77 (0.0%)	1882 (10.2%)
Thiazolidinedione	5 (0.0%)	962 (6.1%)	6 (0.0%)	1225 (6.7%)
DPP4 inhibitor, meglitinide derivative or GLP-1 agonist	1 (0.0%)	153 (1.0%)	1 (0.0%)	182 (1.0%)
Cardiovascular preventive treatment in year prior to study entry, %				
Statin	18 233 (1.9%)	5636 (35.7%)	19 537 (2.1%)	6680 (36.3%)
Any antihypertensive	156 098 (16.3%)	9404 (59.5%)	99 179 (10.6%)	9898 (53.8%)
ACE inhibitor	59 412 (6.2%)	5278 (33.4%)	44 176 (4.7%)	6333 (34.4%)
Angiotensin receptor blocker	7324 (0.8%)	1357 (8.6%)	5276 (0.6%)	1263 (6.9%)
Beta blocker	54 861 (5.7%)	2755 (17.4%)	34 455 (3.7%)	2730 (14.8%)

* Most recent measurement within 1 year of study entry

** Mean HbA1c over 3 years before or after study entry

Percentage of patients with non-missing values of covariates: ethnicity 52.5%, smoking 72.9%, systolic blood pressure 42.2%, diastolic blood pressure 42.2%, total cholesterol 8.3%, HDL cholesterol 5.4%, body mass index 30.3%.

Figure 5.2: Distribution of initial presentations of cardiovascular disease among patients with type 2 diabetes and no prior history of cardiovascular disease

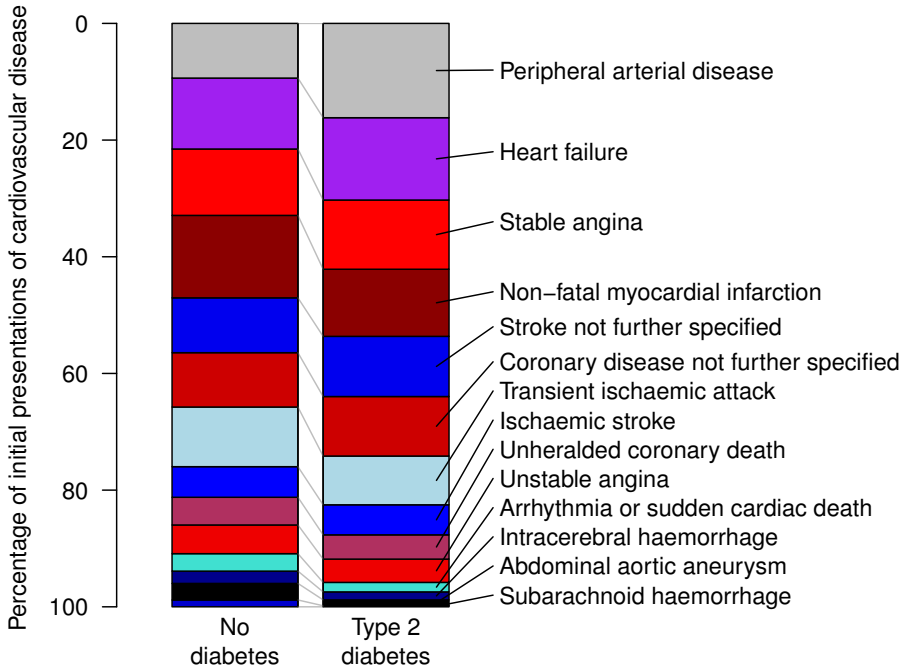


Table 5.3: Distribution of events, time to event and age at event for initial presentation of cardiovascular disease among patients with no diabetes or type 2 diabetes

	No diabetes (107 501 events)			**Type 2 diabetes (6137 events)**		
Initial presentation of cardiovascular disease	% of events	Median time to event (years)	Median age at event (years)	% of events	Median time to event (years)	Median age at event (years)
Stable angina	11.4	3.2	67.0	11.9	2.5	68.6
Unstable angina	4.9	4.5	63.1	4.0	2.8	64.7
Coronary disease not further specified	9.3	4.4	67.3	10.2	3.3	67.4
Non-fatal myocardial infarction	14.1	4.7	66.7	11.5	3.4	71.1
Unheralded coronary death	4.7	5.1	76.2	4.2	4.1	77.0
Heart failure	12.2	3.6	79.8	14.1	3.1	76.9
Arrhythmia or sudden cardiac death	3.0	5.0	67.1	1.6	4.0	68.3
Transient ischaemic attack	10.2	4.0	74.3	8.4	3.4	75.2
Ischaemic stroke	5.2	5.5	76.4	5.1	4.0	76.0
Stroke not further specified	9.4	3.4	79.7	10.3	2.5	78.3
Subarachnoid haemorrhage	1.2	4.4	57.3	0.2	2.8	74.4
Intracerebral haemorrhage	2.1	4.8	74.0	1.4	4.2	74.7
Peripheral arterial disease	9.4	4.4	70.1	16.2	3.5	71.7
Abdominal aortic aneurysm	2.8	5.0	76.2	1.0	4.5	77.8
Overall	100.0	4.2	72.0	100.0	3.2	72.6

Figure 5.3: Cumulative incidence of twelve cardiovascular diseases by diabetes status
Crude cumulative incidence curves taking into account competing risks, for patients with type 2 diabetes or without diabetes at age 40.

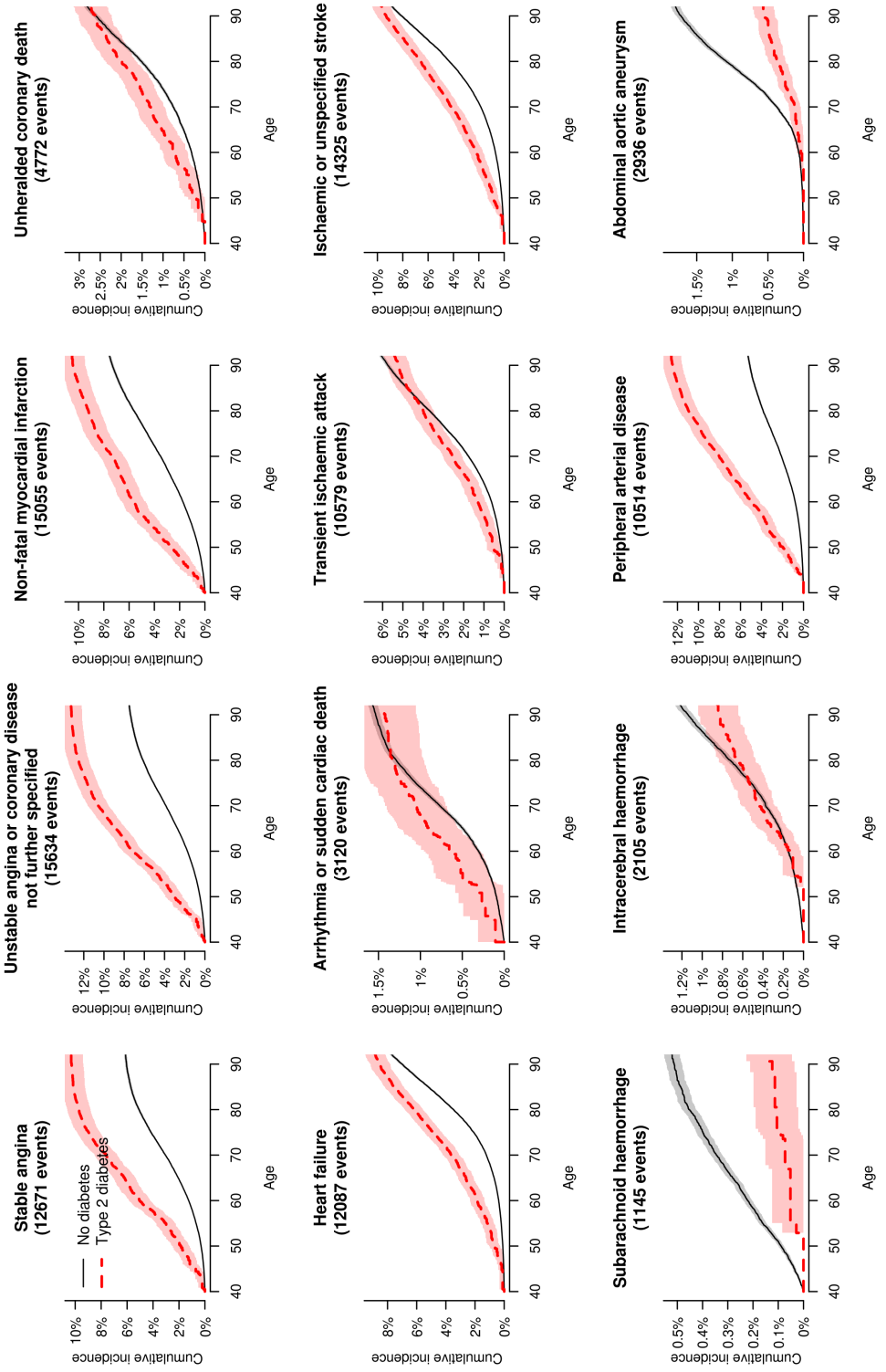


Table 5.4: Cumulative incidence of initial presentation of cardiovascular diseases at age 80 for patients with type 2 diabetes or no diabetes at age 40

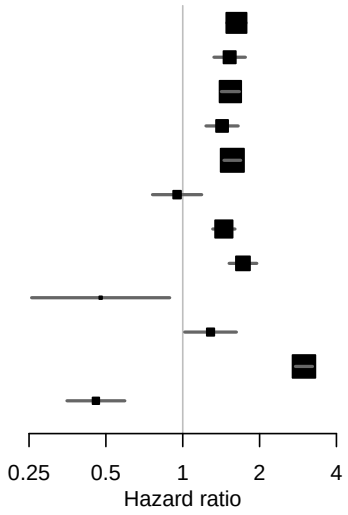
Cumulative incidences expressed as percentages (95% CI), taking into account risk of competing events

	Women		**Men**	
Initial presentation	No diabetes	Type 2 diabetes	No diabetes	Type 2 diabetes
Stable angina	4.5 (4.4, 4.6)	10.1 (8.7, 11.5)	5.7 (5.5, 5.8)	9.5 (8.4, 10.7)
Unstable angina or coronary disease not further specified	5.2 (5.1, 5.3)	11.6 (10.0, 13.2)	7.1 (6.9, 7.3)	13.1 (11.7, 14.5)
Non-fatal myocardial infarction	3.4 (3.3, 3.5)	6.0 (4.9, 7.2)	8.0 (7.8, 8.1)	11.4 (9.9, 12.8)
Unheralded coronary death	0.9 (0.9, 1.0)	1.5 (0.9, 2.0)	2.2 (2.1, 2.3)	2.4 (1.9, 2.9)
Heart failure	3.3 (3.1, 3.4)	6.4 (5.4, 7.3)	3.7 (3.6, 3.9)	6.1 (5.2, 6.9)
Arrhythmia or sudden cardiac death	0.9 (0.8, 1.0)	0.7 (0.4, 1.0)	1.7 (1.6, 1.8)	1.8 (1.2, 2.4)
Transient ischaemic attack	3.6 (3.5, 3.7)	4.1 (3.3, 4.9)	3.8 (3.7, 3.9)	3.9 (3.3, 4.6)
Ischaemic or unspecified stroke	3.9 (3.8, 4.1)	7.3 (6.2, 8.5)	4.7 (4.6, 4.9)	6.1 (5.3, 6.9)
Subarachnoid haemorrhage	0.6 (0.5, 0.6)	0.2 (0.0, 0.4)	0.3 (0.3, 0.4)	0.0 (0.0, 0.1)
Intracerebral haemorrhage	0.7 (0.6, 0.7)	0.4 (0.2, 0.6)	0.8 (0.7, 0.8)	0.8 (0.5, 1.1)
Peripheral arterial disease	3.2 (3.1, 3.3)	9.7 (8.4, 11.1)	4.5 (4.4, 4.7)	11.7 (10.5, 13.0)
Abdominal aortic aneurysm	0.5 (0.5, 0.6)	0.1 (0.0, 0.2)	1.7 (1.6, 1.8)	0.6 (0.4, 0.8)
All cardiovascular presentations	30.7 (30.3, 31.0)	58.2 (54.9, 61.4)	44.3 (43.8, 44.7)	67.4 (64.4, 70.4)
Death from other causes without any cardiovascular disease prior to death	14.7 (14.5, 15.0)	12.4 (11.0, 13.8)	16.4 (16.2, 16.6)	11.9 (10.7, 13.1)

Figure 5.4: Hazard ratios for association of type 2 diabetes with twelve cardiovascular diseases

Hazard ratios for different initial presentations of cardiovascular diseases associated with type 2 diabetes, adjusted for age, sex, smoking, body mass index, deprivation, high density lipoprotein cholesterol, total cholesterol, systolic blood pressure, smoking status, statins and antihypertensive prescriptions. P values *** <0.001, ** <0.01, * <0.05.

	Number of events		
Initial presentation of cardiovascular disease	*No diabetes*	*Type 2 diabetes*	**Hazard ratio (95% CI)**
Stable angina	12232	728	1·62 (1·49−1·77) ***
Unstable angina	5286	245	1·53 (1·32−1·76) ***
Non−fatal myocardial infarction	15191	706	1·54 (1·42−1·67) ***
Unheralded coronary death	5101	255	1·43 (1·23−1·65) ***
Heart failure	13072	866	1·56 (1·45−1·69) ***
Arrhythmia or sudden cardiac death	3218	100	0·95 (0·76−1·19)
Transient ischaemic attack	10990	513	1·45 (1·31−1·60) ***
Ischaemic stroke	5643	316	1·72 (1·52−1·95) ***
Subarachnoid haemorrhage	1260	11	0·48 (0·26−0·89) *
Intracerebral haemorrhage	2265	84	1·28 (1·02−1·62) *
Peripheral arterial disease	10074	992	2·98 (2·76−3·22) ***
Abdominal aortic aneurysm	3051	62	0·46 (0·35−0·59) ***



In multiply adjusted Cox models, there were strong positive associations between T2D and peripheral arterial disease (HR 2.98, 95% CI 2.76, 3.22), ischaemic stroke (HR 1.72, 95% CI 1.52, 1.95), stable angina (HR 1.62, 95% CI 1.49, 1.77), heart failure (HR 1.56, 95% CI 1.45, 1.69), and non-fatal myocardial infarction (HR 1.54, 95% CI 1.42, 1.67) (**Figure 5.4 on page 118** and **Table 5.5 on page 121**).

In contrast, T2D was inversely associated with abdominal aortic aneurysm (HR 0.46, 95% CI 0.35, 0.59) and subarachnoid haemorrhage (HR 0.48, 95% CI 0.26, 0.89). There was no statistically significant association with arrhythmia / sudden cardiac death (HR 0.95, 95% CI 0.76, 1.19). The strengths of associations were slightly attenuated by adjustment for risk factors and medication compared to the model adjusted for age and sex only (**Table 5.6 on page 123**, **Figure 5.5 on page 119**).

Figure 5.5: Association of type 2 diabetes with initial presentation of cardiovascular diseases using different adjustments

'Risk factors' adjustment includes age, sex, body mass index, deprivation, high density lipoprotein cholesterol, total cholesterol, systolic blood pressure, and smoking status. 'Treatment' includes statins and antihypertensive prescription in the year before study entry. P values *** <0.001, ** <0.01, * <0.05

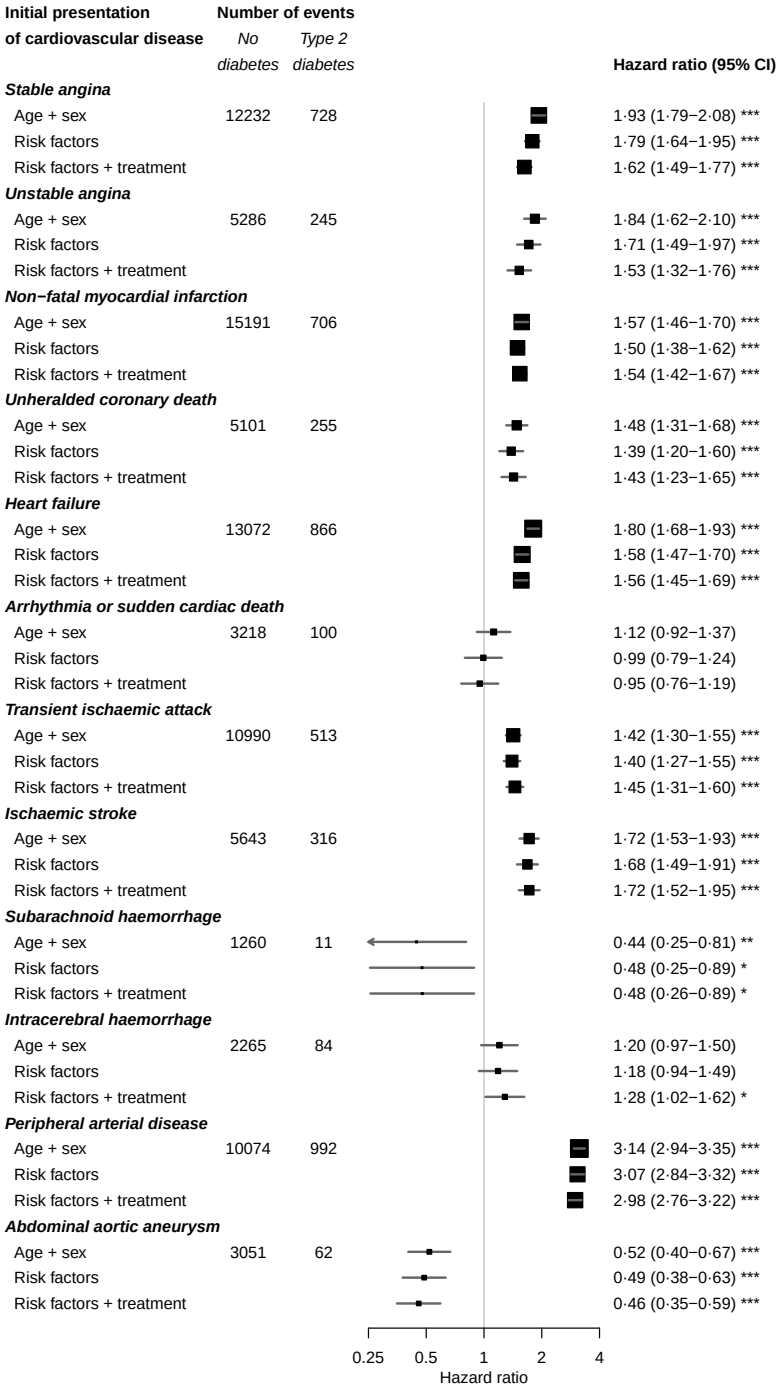


Figure 5.6: Hazard ratios for association of type 2 diabetes with twelve cardiovascular diseases by sex and age

Hazard ratios by sex and age group for the association of different initial presentations of cardiovascular disease with type 2 diabetes, adjusted for age, smoking, body mass index, deprivation, high density lipoprotein cholesterol, total cholesterol, systolic blood pressure, smoking status, statins and antihypertensive prescriptions. P values *** <0.001, ** <0.01, * <0.05

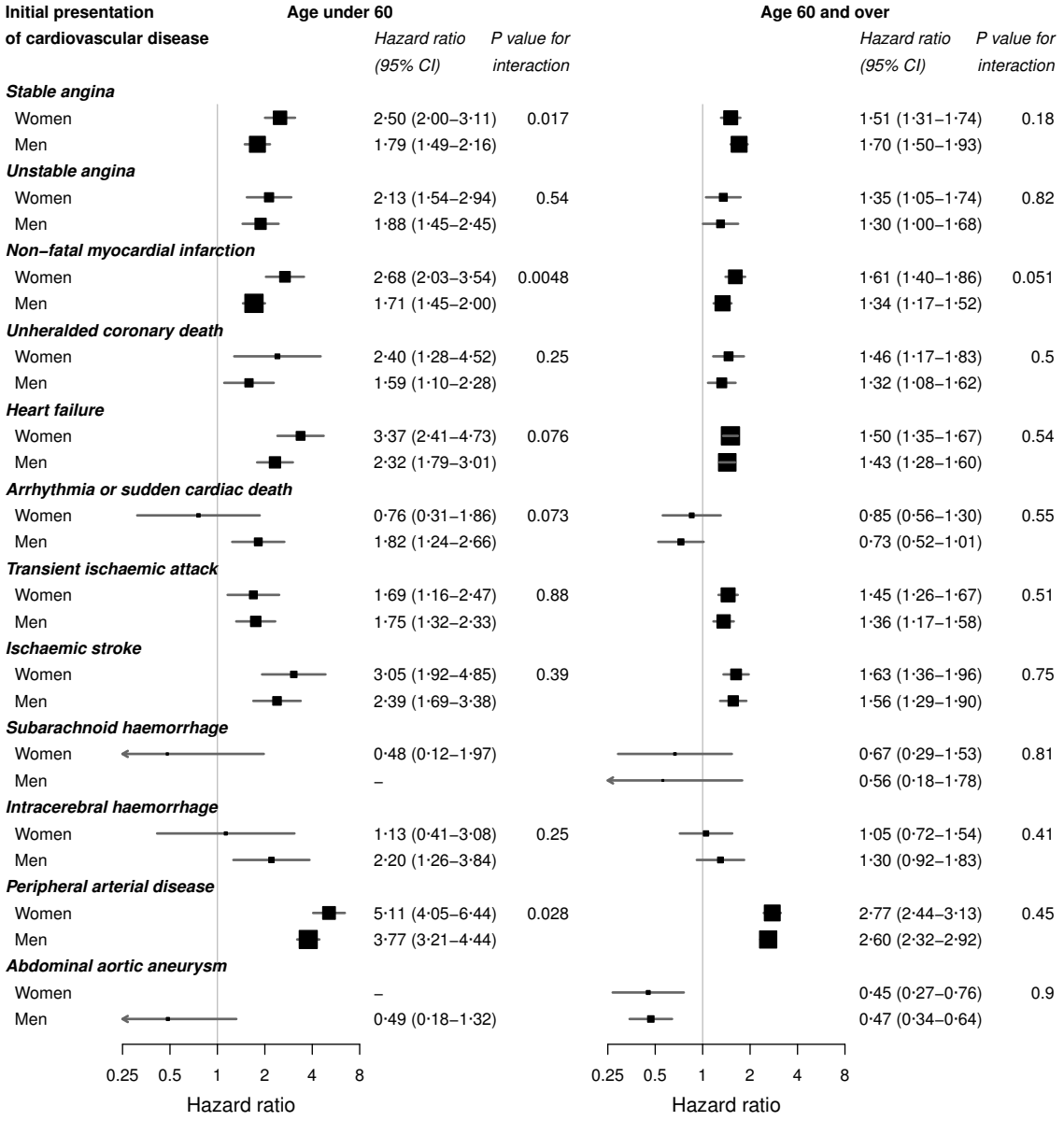


Table 5.5: Hazard ratios for association of type 2 diabetes with cardiovascular disease incidence using different levels of adjustment

Both analyses used the complete dataset with multiple imputation of missing covariate values. 'Risk factors' comprised body mass index, deprivation, high density lipoprotein cholesterol, total cholesterol, systolic blood pressure and smoking status.

	Primary analysis (adjusted for age, sex, risk factors, statins and antihypertensives)		**Adjusted for age, sex and risk factors**	
Initial presentation	Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value
Stable angina	1.62 (1.49, 1.77)	< 0.0001	1.79 (1.64, 1.95)	< 0.0001
Unstable angina	1.53 (1.32, 1.76)	< 0.0001	1.71 (1.49, 1.97)	< 0.0001
Coronary disease not further specified	1.58 (1.45, 1.73)	< 0.0001	1.93 (1.76, 2.11)	< 0.0001
Non-fatal myocardial infarction	1.54 (1.42, 1.67)	< 0.0001	1.50 (1.38, 1.62)	< 0.0001
Unheralded coronary death	1.43 (1.23, 1.65)	< 0.0001	1.39 (1.20, 1.60)	< 0.0001
Heart failure	1.56 (1.45, 1.69)	< 0.0001	1.58 (1.47, 1.70)	< 0.0001
Arrhythmia or sudden cardiac death	0.95 (0.76, 1.19)	0.65	0.99 (0.80, 1.24)	0.94
Transient ischaemic attack	1.45 (1.31, 1.60)	< 0.0001	1.40 (1.27, 1.55)	< 0.0001
Ischaemic stroke	1.72 (1.52, 1.95)	< 0.0001	1.68 (1.49, 1.91)	< 0.0001
Stroke not further specified	1.64 (1.48, 1.81)	< 0.0001	1.64 (1.48, 1.82)	< 0.0001
Subarachnoid haemorrhage	0.48 (0.26, 0.89)	0.020	0.48 (0.26, 0.89)	0.020
Intracerebral haemorrhage	1.28 (1.02, 1.62)	0.035	1.18 (0.94, 1.49)	0.15
Peripheral arterial disease	2.98 (2.76, 3.22)	< 0.0001	3.07 (2.84, 3.32)	< 0.0001
Abdominal aortic aneurysm	0.46 (0.35, 0.59)	< 0.0001	0.49 (0.38, 0.63)	< 0.0001
Other death	1.10 (1.05, 1.17)	0.0004	1.01 (0.96, 1.07)	0.70

5.5.3 Associations in women and men, by age group and ethnicity

For individuals aged 40 years without cardiovascular disease, the overall estimated risk of developing any cardiovascular disease by age 80 was 30.7% for women and 44.3% for men without diabetes, compared to 58.2% for women and 67.4% for men with T2D (**Table 5.4 on page 117**). I observed slightly greater risk of non-fatal myocardial infarction associated with T2D in women under 60 years of age (HR 2.68, 95% CI 2.03, 3.54) compared to men under 60 years of age (HR 1.71, 95% CI 1.45, 2.00; p = 0.0048 for interaction) but no other statistically significant sex differences (**Figure 5.6 on page 120**). Associations between T2D and coronary endpoints, transient ischaemic attack, ischaemic stroke and peripheral arterial disease were significantly stronger among younger compared to older age groups (**Figure 5.7 on page 125**). Hazard ratios for T2D and cardiovascular endpoints were consistent across ethnic groups for the common endpoints such as heart failure, stable angina and peripheral arterial disease, but the results for less common endpoints were imprecise because of the small number of events (**Figure 5.8 on page 126**).

5.5.4 Glycaemic control in patients with type 2 diabetes

The risk of cardiovascular diseases associated with the presence of T2D was highest for those with an HbA1c $\geq$ 58 mmol/mol or those in whom the HbA1c had not been recorded (**Figure 5.9 on page 127**). Individuals with T2D and an HbA1c <48 mmol/mol had an increased risk of peripheral arterial disease (HR 1.82, 95% CI 1.52, 2.18) and ischaemic stroke (HR 1.53, 95% CI 1.18, 1.98), but no greater risk of any of the other cardiovascular diseases.

5.5.5 Sensitivity analyses and secondary outcomes

Overall results were similar among the subset of individuals with completely recorded covariates (**Table 5.7 on page 123**), but effect estimates were imprecise because the number of patients was much smaller. Hazard ratios for any diabetes versus no diabetes were similar to the main results (T2D versus no diabetes), and restricting the patient population to those entering the study after 2004 also made little difference to the results (**Table 5.8 on page 124**). Apart from a slightly stronger association with heart failure (HR 2.15; 95% CI 1.96, 2.36 versus 1.56; 95% CI 1.45, 1.69), results did not differ when primary care data was omitted as a source of endpoint information (**Figure 5.10 on page 128**). T2D was strongly associated with composite cardiovascular mortality and all cause mortality (**Table 5.9 on page 124**). Plots of Schoenfeld residuals showed that hazard ratios were constant over time, i.e. the proportional hazards assumption of the Cox model was not violated (**Figure 5.11 on page 129**).

Table 5.6: Age and sex adjusted hazard ratios for association of type 2 diabetes with initial presentation of cardiovascular diseases

	All patients		**Patients with completely recorded covariates**	
Initial presentation	Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value
Stable angina	1.93 (1.79, 2.08)	< 0.0001	2.10 (1.67, 2.64)	< 0.0001
Unstable angina	1.84 (1.62, 2.10)	< 0.0001	1.19 (0.83, 1.72)	0.34
Coronary disease not further specified	2.11 (1.95, 2.29)	< 0.0001	1.75 (1.41, 2.16)	< 0.0001
Non-fatal myocardial infarction	1.57 (1.46, 1.70)	< 0.0001	1.48 (1.17, 1.88)	0.0011
Unheralded coronary death	1.48 (1.31, 1.68)	< 0.0001	1.73 (1.16, 2.59)	0.0070
Heart failure	1.80 (1.68, 1.93)	< 0.0001	3.29 (2.48, 4.35)	< 0.0001
Arrhythmia or sudden cardiac death	1.12 (0.92, 1.37)	0.26	0.83 (0.46, 1.51)	0.55
Transient ischaemic attack	1.42 (1.30, 1.55)	< 0.0001	1.20 (0.89, 1.61)	0.23
Ischaemic stroke	1.72 (1.53, 1.93)	< 0.0001	1.47 (0.98, 2.19)	0.060
Stroke not further specified	1.77 (1.63, 1.92)	< 0.0001	1.22 (0.88, 1.69)	0.23
Subarachnoid haemorrhage	0.45 (0.25, 0.81)	0.0077	–	–
Intracerebral haemorrhage	1.20 (0.97, 1.50)	0.099	0.93 (0.45, 1.93)	0.85
Peripheral arterial disease	3.14 (2.94, 3.35)	< 0.0001	2.80 (2.24, 3.49)	< 0.0001
Abdominal aortic aneurysm	0.52 (0.40, 0.67)	< 0.0001	0.31 (0.14, 0.66)	0.0026
Other death	1.04 (0.99, 1.09)	0.11	1.45 (1.24, 1.69)	< 0.0001

Table 5.7: Hazard ratios from complete case analyses for association of type 2 diabetes with initial presentation of cardiovascular diseases

These analyses were limited to the subset of individuals with completely recorded covariate information: 59 116 with no diabetes and 12 411 with type 2 diabetes. 'Risk factors' comprised body mass index, deprivation, high density lipoprotein cholesterol, total cholesterol, systolic blood pressure and smoking status.

	Adjusted for age, sex, risk factors, statins and antihypertensives		**Adjusted for age, sex and risk factors**	
Initial presentation	Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value
Stable angina	2.03 (1.59, 2.60)	< 0.0001	2.10 (1.65, 2.67)	< 0.0001
Unstable angina	1.16 (0.78, 1.73)	0.45	1.28 (0.87, 1.88)	0.22
Coronary disease not further specified	1.64 (1.30, 2.07)	< 0.0001	1.69 (1.35, 2.13)	< 0.0001
Non-fatal myocardial infarction	1.50 (1.17, 1.93)	0.0016	1.49 (1.16, 1.90)	0.0018
Unheralded coronary death	2.22 (1.43, 3.46)	0.0004	1.93 (1.24, 3.02)	0.0038
Heart failure	3.01 (2.22, 4.07)	< 0.0001	2.96 (2.19, 3.99)	< 0.0001
Arrhythmia or sudden cardiac death	0.92 (0.49, 1.73)	0.79	0.88 (0.47, 1.63)	0.68
Transient ischaemic attack	1.33 (0.97, 1.81)	0.076	1.27 (0.93, 1.73)	0.13
Ischaemic stroke	1.61 (1.05, 2.46)	0.028	1.57 (1.03, 2.40)	0.034
Stroke not further specified	1.16 (0.83, 1.64)	0.39	1.14 (0.81, 1.60)	0.46
Subarachnoid haemorrhage	–	–	–	–
Intracerebral haemorrhage	1.20 (0.55, 2.63)	0.66	1.15 (0.53, 2.50)	0.73
Peripheral arterial disease	3.05 (2.40, 3.88)	< 0.0001	3.16 (2.50, 4.01)	< 0.0001
Abdominal aortic aneurysm	0.30 (0.14, 0.67)	0.0034	0.30 (0.14, 0.65)	0.0026
Other death	1.63 (1.38, 1.93)	< 0.0001	1.53 (1.29, 1.80)	< 0.0001

Table 5.8: Hazard ratios from secondary analyses for association of type 2 diabetes with initial presentation of cardiovascular diseases

Hazard ratios were adjusted for age, sex, risk factors (body mass index, deprivation, high density lipoprotein cholesterol, total cholesterol, systolic blood pressure, smoking status), and statin and antihypertensive prescription in the year before study entry.

	Patients with any diabetes versus no diabetes		**Type 2 diabetes versus no diabetes, restricted to patients entering the study after 2004**	
Initial presentation	Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value
Stable angina	1.70 (1.57, 1.83)	< 0.0001	1.62 (1.17, 2.23)	0.0030
Unstable angina	1.66 (1.47, 1.87)	< 0.0001	1.30 (0.86, 1.95)	0.21
Coronary disease not further specified	1.73 (1.60, 1.87)	< 0.0001	1.33 (1.03, 1.73)	0.031
Non-fatal myocardial infarction	1.74 (1.62, 1.86)	< 0.0001	1.48 (1.14, 1.91)	0.0031
Unheralded coronary death	1.81 (1.60, 2.04)	< 0.0001	1.93 (1.31, 2.84)	0.00098
Heart failure	1.76 (1.65, 1.87)	< 0.0001	2.02 (1.55, 2.63)	< 0.0001
Arrhythmia or sudden cardiac death	1.16 (0.97, 1.39)	0.11	1.10 (0.58, 2.06)	0.78
Transient ischaemic attack	1.54 (1.42, 1.68)	< 0.0001	1.24 (0.91, 1.69)	0.18
Ischaemic stroke	1.72 (1.54, 1.92)	< 0.0001	1.42 (1.00, 2.01)	0.051
Stroke not further specified	1.80 (1.65, 1.95)	< 0.0001	1.24 (0.99, 1.56)	0.059
Subarachnoid haemorrhage	0.54 (0.33, 0.89)	0.016	–	–
Intracerebral haemorrhage	1.40 (1.15, 1.70)	0.0007	1.10 (0.53, 2.27)	0.80
Peripheral arterial disease	3.33 (3.12, 3.55)	< 0.0001	2.70 (2.12, 3.44)	< 0.0001
Abdominal aortic aneurysm	0.49 (0.39, 0.62)	< 0.0001	0.33 (0.15, 0.72)	0.0056
Other death	1.39 (1.33, 1.45)	< 0.0001	1.46 (1.29, 1.64)	< 0.0001

Table 5.9: Hazard ratios for composite endpoints

All analyses used the complete dataset with multiple imputation of missing covariate values. 'Risk factors' comprised body mass index, deprivation, high density lipoprotein cholesterol, total cholesterol, systolic blood pressure and smoking status.

	Adjusted for age, sex, risk factors, statins and antihypertensives		**Adjusted for age, sex and risk factors**	
Initial presentation	Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value
Cardiovascular death	1.53 (1.45, 1.62)	< 0.0001	1.62 (1.53, 1.71)	< 0.0001
Non-cardiovascular death	1.29 (1.24, 1.34)	< 0.0001	1.22 (1.17, 1.26)	< 0.0001
Death due to any cause	1.35 (1.31, 1.39)	< 0.0001	1.31 (1.27, 1.36)	< 0.0001

Figure 5.7: Adjusted hazard ratios for type 2 diabetes and initial presentations of cardiovascular diseases by age group

Hazard ratios by age group for the association of different initial presentations of cardiovascular disease with type 2 diabetes, adjusted for sex, body mass index, deprivation, high density lipoprotein cholesterol, total cholesterol, systolic blood pressure, smoking status, statins and antihypertensive prescriptions. P values *** <0.001, ** <0.01, * <0.05

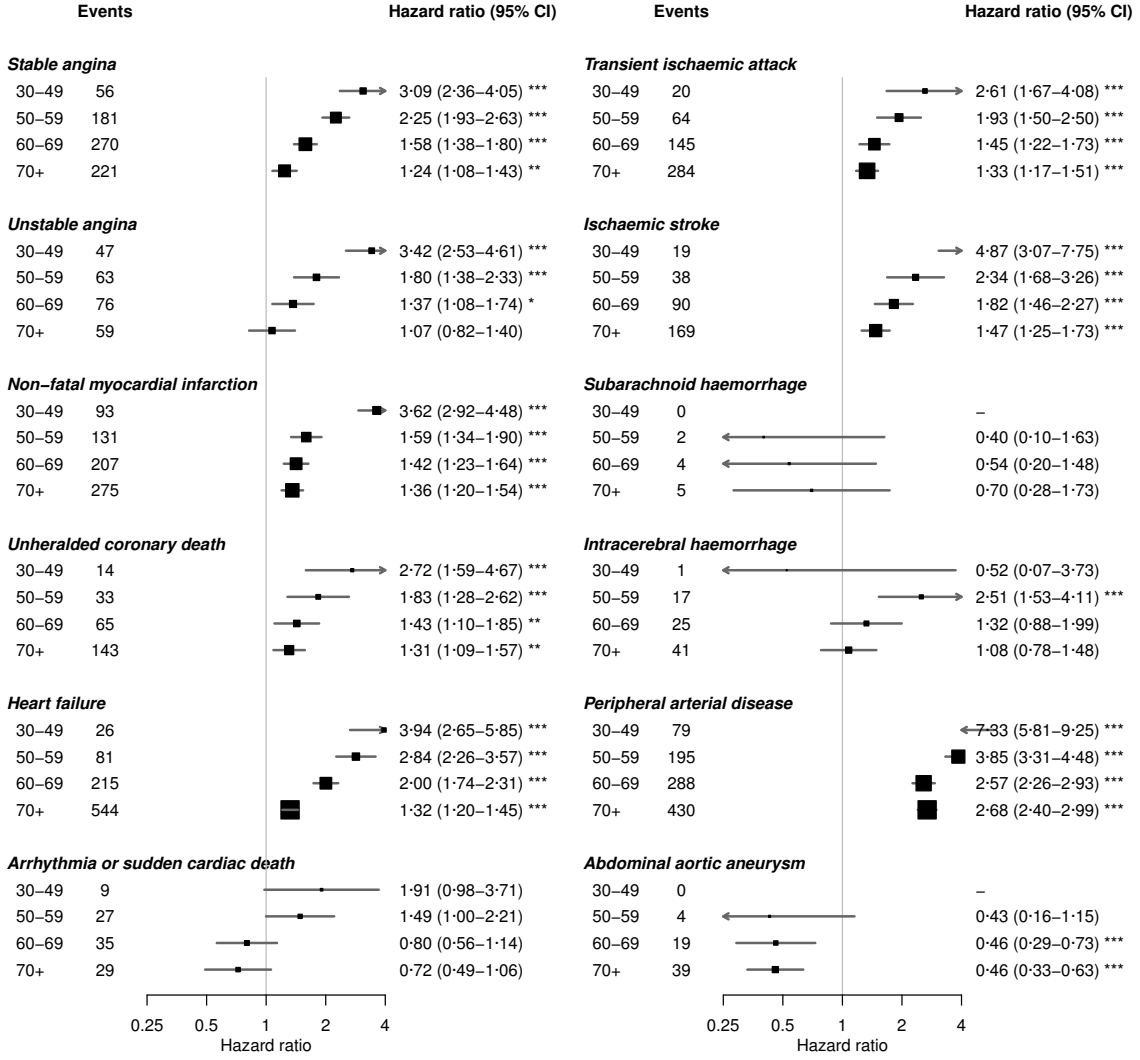


Figure 5.8: Adjusted hazard ratios for type 2 diabetes and initial presentations of cardiovascular diseases by ethnicity

Hazard ratios by ethnicity for the association of different initial presentations of cardiovascular disease with type 2 diabetes, adjusted for age, sex, body mass index, deprivation, high density lipoprotein cholesterol, total cholesterol, systolic blood pressure, smoking status, statins and antihypertensive prescriptions. P values *** <0.001, ** <0.01, * <0.05

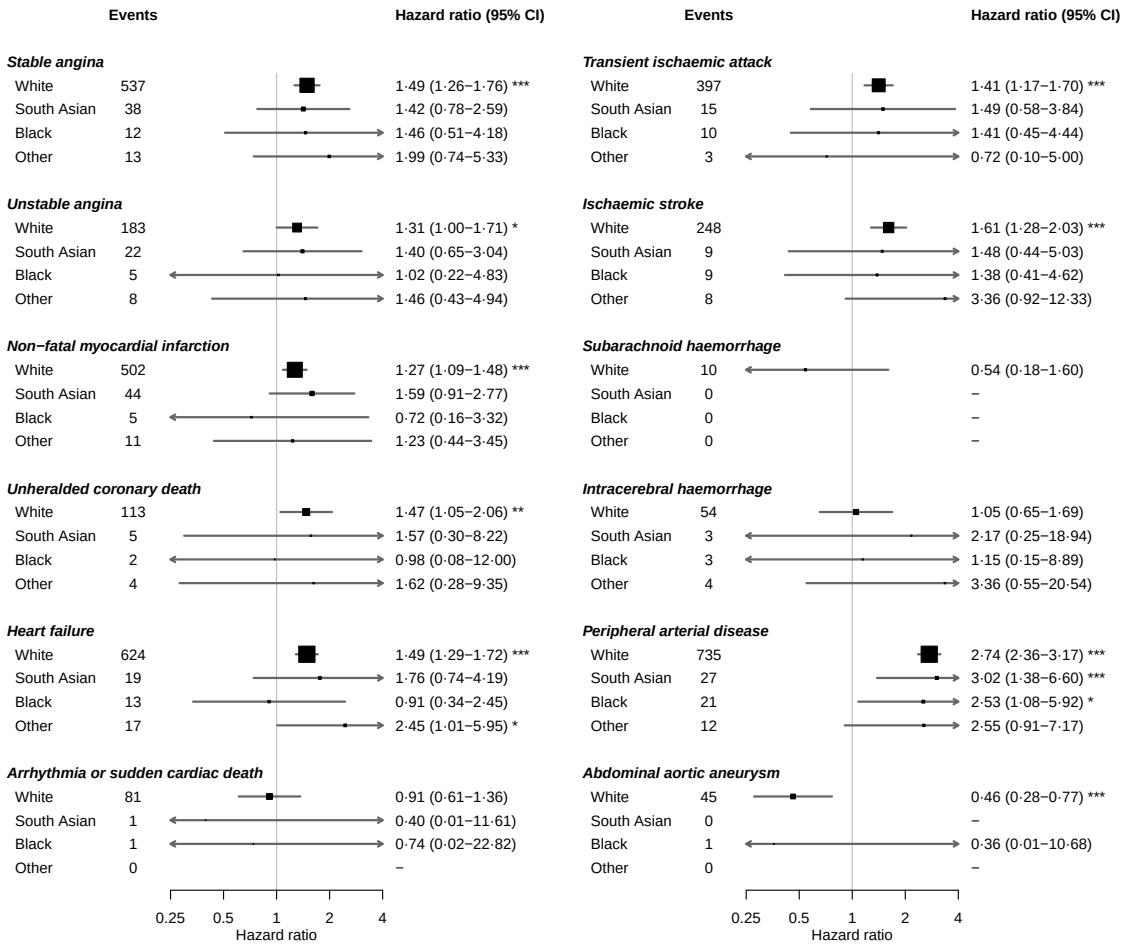


Figure 5.9: Association of type 2 diabetes with initial presentation of cardiovascular diseases by level of glycaemic control

Hazard ratios for initial presentations of cardiovascular diseases associated with type 2 diabetes with different levels of HbA1c (mean in mmol/mol within 3 years of baseline), compared with no diabetes, adjusted for age, sex, body mass index, deprivation, high density lipoprotein cholesterol, total cholesterol, systolic blood pressure, smoking status, statins and antihypertensive prescriptions. P values *** <0.001, ** <0.01, * <0.05

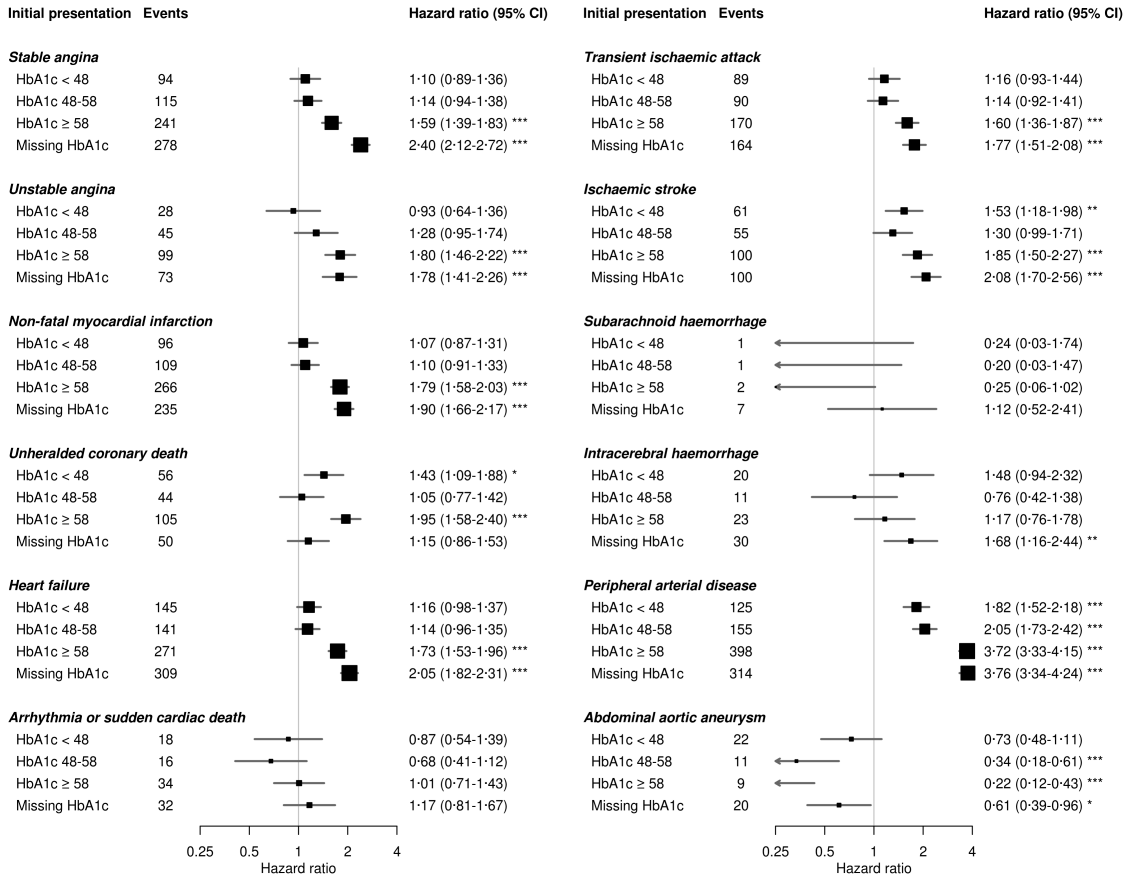


Figure 5.10: Association of type 2 diabetes with initial presentation of cardiovascular diseases using different subsets of the data sources

Hazard ratios adjusted for age, sex, body mass index, deprivation, high density lipoprotein cholesterol, total cholesterol, systolic blood pressure, smoking status, statins and antihypertensive prescriptions. P values *** <0.001, ** <0.01, * <0.05

Initial presentation of cardiovascular disease	Number of events: No diabetes	Number of events: Type 2 diabetes	Hazard ratio (95% CI)
Stable angina			
Using all data sources (main analysis)	12232	728	1·62 (1·49–1·77) ***
Ignoring primary care	1204	50	1·46 (1·08–1·97) *
Unstable angina			
Using all data sources (main analysis)	5286	245	1·53 (1·32–1·76) ***
Ignoring primary care	9092	471	1·70 (1·54–1·88) ***
Non−fatal myocardial infarction			
Using all data sources (main analysis)	15191	706	1·54 (1·42–1·67) ***
Ignoring primary care	14316	683	1·58 (1·46–1·72) ***
Heart failure			
Using all data sources (main analysis)	13072	866	1·56 (1·45–1·69) ***
Ignoring primary care	6582	572	2·15 (1·96–2·36) ***
Using death endpoints only	1835	127	1·59 (1·27–1·98) ***
Arrhythmia or sudden cardiac death			
Using all data sources (main analysis)	3218	100	0·95 (0·76–1·19)
Ignoring primary care	3505	112	0·95 (0·78–1·17)
Transient ischaemic attack			
Using all data sources (main analysis)	10990	513	1·45 (1·31–1·60) ***
Ignoring primary care	3437	175	1·63 (1·39–1·92) ***
Ischaemic stroke			
Using all data sources (main analysis)	5643	316	1·72 (1·52–1·95) ***
Ignoring primary care	7547	427	1·75 (1·57–1·94) ***
Using death endpoints only	896	48	1·61 (1·15–2·24) **
Subarachnoid haemorrhage			
Using all data sources (main analysis)	1260	11	0·48 (0·26–0·89) *
Ignoring primary care	1579	14	0·48 (0·28–0·84) **
Using death endpoints only	607	6	0·42 (0·18–0·94) *
Intracerebral haemorrhage			
Using all data sources (main analysis)	2265	84	1·28 (1·02–1·62) *
Ignoring primary care	2498	86	1·14 (0·91–1·43)
Using death endpoints only	1274	46	1·03 (0·75–1·42)
Peripheral arterial disease			
Using all data sources (main analysis)	10074	992	2·98 (2·76–3·22) ***
Ignoring primary care	3296	297	2·92 (2·56–3·33) ***
Using death endpoints only	330	7	0·60 (0·27–1·30)
Abdominal aortic aneurysm			
Using all data sources (main analysis)	3051	62	0·46 (0·35–0·59) ***
Ignoring primary care	2638	47	0·40 (0·30–0·54) ***
Using death endpoints only	1855	30	0·34 (0·23–0·50) ***

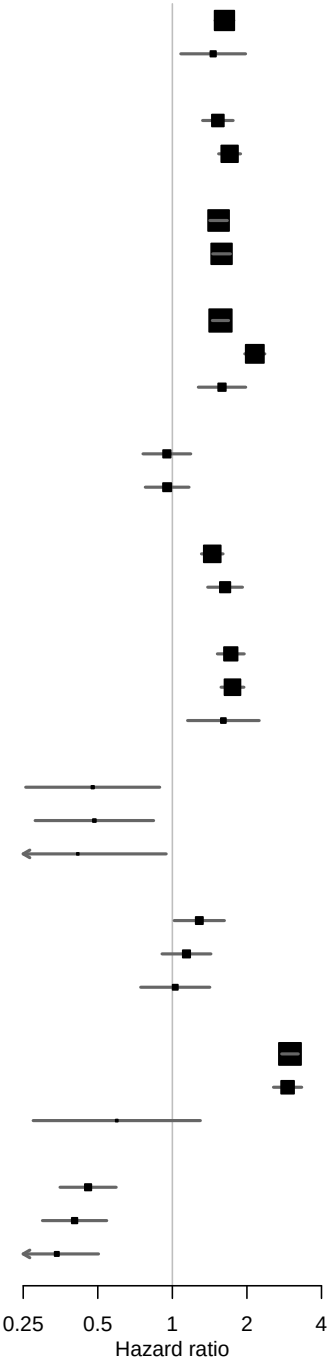
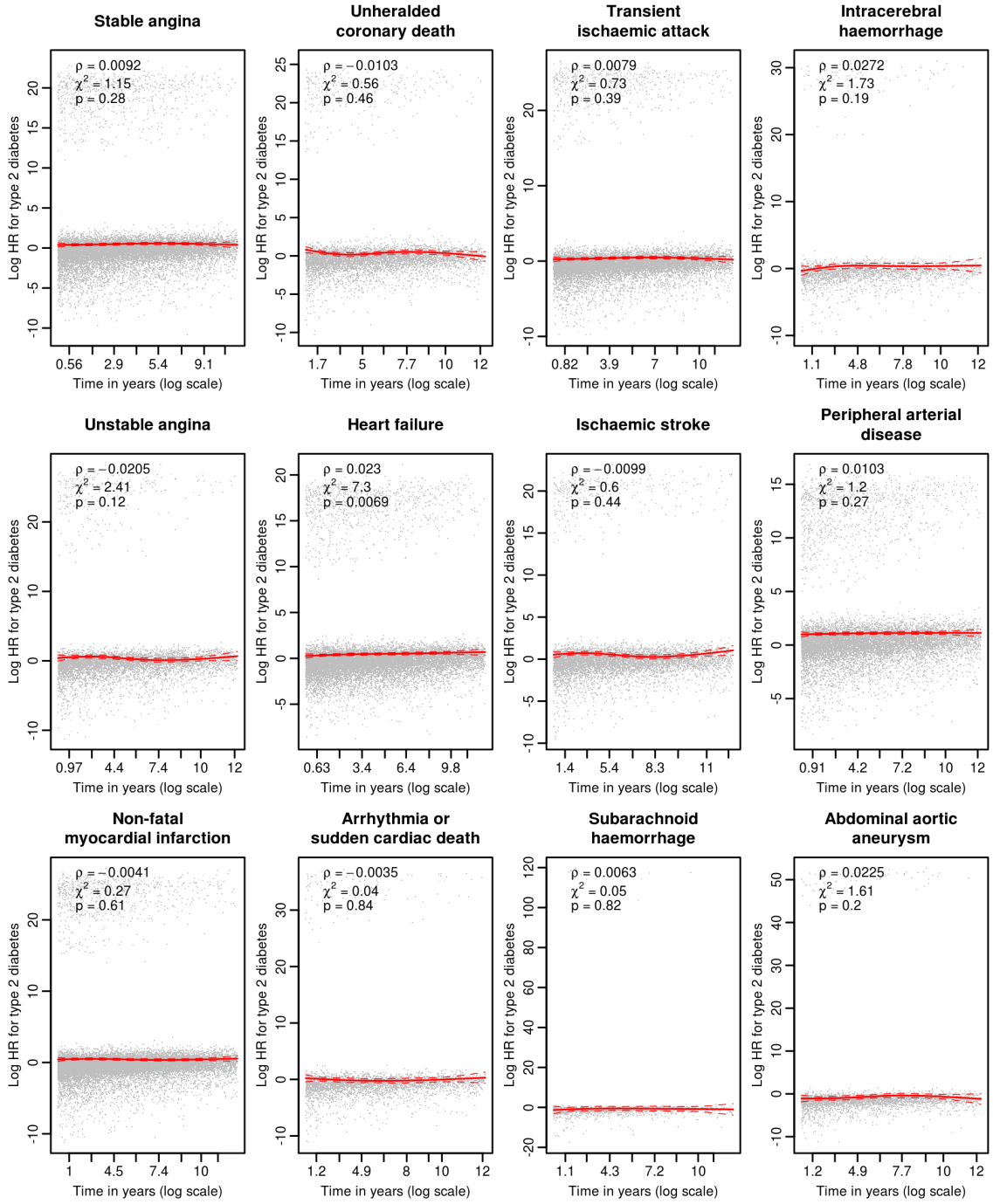


Figure 5.11: Schoenfeld residuals and tests for proportional hazards

Scaled Shoenfeld residuals for type 2 diabetes versus no diabetes, adjusted for age, sex, body mass index, deprivation, high density lipoprotein cholesterol, total cholesterol, systolic blood pressure, smoking status, statins and antihypertensive prescriptions. The lines show the estimate and 95% confidence interval for the beta coefficient for type 2 diabetes over time. The top left corner of each graph contains ρ, χ^2 and p value for the correlation between transformed survival time and scaled Schoenfeld residuals. For clarity, points plotted on the graph are for a random sample of 20 000 patients rather than all 1.9 million.



5.6 Discussion

5.6.1 Summary of main findings

This study demonstrates the value of the CALIBER dataset in investigating the initial presentation of cardiovascular disease. Findings of increased risk of many cardiovascular diseases in patients with T2D provides a reassuring check on the validity of the dataset for this type of study. However, this study also provides new information that the associations with different cardiovascular presentations differ in terms of direction and magnitude of effects.

5.6.2 Outcomes with increased risk in type 2 diabetes

A group of atherosclerotic diseases including coronary heart disease and peripheral arterial disease have a similar association of increased risk in individuals with type 2 diabetes. Peripheral arterial disease is among the most common first presentations of cardiovascular disease, consistent with the increase in prevalence between 2000 and 2010 found in a recent systematic review [206]. Furthermore, of all the twelve diseases studied, peripheral arterial disease showed the strongest association with T2D, with an adjusted hazard ratio of 3, consistent with previous studies [206]. Heart failure was also among the most common first presentations (accounting for 14.4% of events among people with T2D). The definition used in this study excludes heart failure occurring after acute myocardial infarction, but this result must be interpreted with caution because it is known that all data sources miss some cases of myocardial infarction (section 4.2). Possible mechanisms may involve prolonged hypertension, chronic hyperglycemia, microvascular disease, glycosylation of myocardial proteins, diabetic nephropathy and autonomic neuropathy [198]. Similar to the meta-analysis of cohort studies in the Emerging Risk Factors Collaboration [200], I found a weaker association of cardiovascular disease with T2D among older people.

5.6.3 Outcomes with reduced risk in type 2 diabetes

A novel finding was the association of T2D with reduced incidence of two major aneurysmal diseases: abdominal aortic aneurysm and subarachnoid haemorrrhage. An inverse association between diabetes AAA has been suggested by cross-sectional studies of abdominal aortic aneurysm screening and case-control studies [44, 207, 208], but evidence from large prospective cohort studies has been mixed [209, 210].

No previous cohort study has been large enough to investigate the association of type 2 diabetes with subarachnoid haemorrhage, but there have been a number of case control studies which are consistent with my results [211]. These findings support the need to identify potential mechanisms which might have therapeutic implications. Such mechanisms may include diabetes-mediated changes in the vascular extracellular matrix, such

as glycation of the extracellular matrix and covalent cross-linking between elastin and collagen in the aortic wall [212].

5.6.4 Outcomes not significantly associated with type 2 diabetes

Several studies have shown that T2D is associated with out of hospital cardiac arrest and sudden cardiac death [213], and an important unanswered question has been the extent to which this effect is mediated by incident diseases known to increase arrhythmic risk such as myocardial infarction and heart failure. In this study I looked at the initial presentation of cardiovascular disease, i.e. arrhythmia with no prior myocardial infarction or heart failure, and my finding of no association suggests that the higher risk of cardiac arrest with T2D is mediated by atherosclerotic coronary disease.

5.6.5 Sex differences

The absolute risk for coronary heart disease morbidity and mortality is lower in women compared to men, but the relative risk associated with diabetes has generally been reported to be higher in women than men [201,214,215]. The most recent meta-analysis of prospective studies reported relative risks for incident coronary heart disease of 2.63 for women and 1.85 for men with diabetes [201]. Sex differences in the association of T2D with outcomes other than coronary heart disease have less often been reported, with conflicting results for stroke [216, 217]. I observed weak evidence of a slightly stronger association of T2D with myocardial infarction in women compared to men among those aged under 60, which is consistent with previous reports, but no other sex differences in associations between T2D and cardiovascular outcomes [218].

5.6.6 HbA1c and glycaemic control

The finding that levels of HbA1c were associated with the risk of many cardiovascular diseases strengthens the case for the causal relevance of blood glucose. These associations were particularly strong for peripheral arterial disease, and persisted even among those with HbA1c <48 mmol/mol, suggesting that risk is not controlled in this group. Individuals in whom HbA1c had not been measured were at higher risk, consistent with measurement being a marker of overall quality of care.

5.6.7 Clinical and research implications

These results will help inform drug development and the design of trials in T2D. To date, trials often include acute myocardial infarction and acute stroke as their primary event endpoints, because of a previous perception that these were the main disease burdens or that they are easier to ascertain and validate than chronic disease endpoints. The results of this study suggest that it is also important to consider chronic disease endpoints such

as heart failure, peripheral arterial disease and stable angina because they are common, have a high morbidity burden and may have different treatment effects from myocardial infarction or stroke.

5.6.8 Limitations

Although a strength of this study is the ability to resolve a wide range of cardiovascular diseases in a validated electronic health record (EHR) linkage, EHRs are limited in the accuracy and detail they include. Residual confounding, due to diet, physical activity, stress and environmental factors, may be a potential source of bias. Although T2D coding has error in clinical practice [181], this is unlikely to be an important source of bias in this study because analyses based on the diagnostic marker of HbA1c were consistent. Error may occur in recording the cardiovascular outcomes; one might expect those which are more reliably coded (such as non-fatal myocardial infarction) to show stronger associations than diseases which are less reliably coded (such as stable angina). The fact that this was not observed suggests that the heterogeneity is real. It was also reassuring that similar associations with T2D were seen when endpoint ascertainment was restricted to hospital admissions and deaths. The lifetime risks were somewhat artificial, as they were based on extrapolation of event rates observed in people entering the cohort at different ages, but this meant they were more contemporaneous than estimates from long term follow-up of cohorts diagnosed decades ago.

5.6.9 Conclusion

This CALIBER study showed that type 2 diabetes was associated with increased risk of many cardiovascular diseases. Heart failure and peripheral arterial disease are the most common initial manifestations of cardiovascular disease in type 2 diabetes. The differences between relative risks of different cardiovascular diseases in patients with type 2 diabetes may have implications for clinical risk assessment and trial design.

5.7 Summary

In this chapter I described the definition of a cohort for studying initial presentation of cardiovascular diseases in CALIBER, and the identification of twelve initial presentations of cardiovascular disease. These definitions will be used in the remainder of the thesis for investigating leukocyte counts and cardiovascular diseases.

6 Full blood count data in CALIBER

6.1 Chapter outline

Full blood counts are commonly performed in general practice, and are available within the CPRD data in CALIBER. This chapter contains preparatory work for a cohort study in CALIBER investigating differential leukocyte counts and risk of incident cardiovascular diseases. The results of this study are reported in the subsequent two chapters.

6.2 Abstract

Objective: To develop an analysis plan for investigating the association of the leukocyte differential with initial presentation of cardiovascular diseases in CALIBER.

Methods: I performed a sample size and power calculation to find out the minimum difference in effect sizes that a CALIBER study would be able to detect. I developed algorithms to extract leukocyte counts from the CALIBER master database. I defined a study cohort of patients with leukocyte counts and explored the association of the differential leukocyte count with cardiovascular risk factors, in order to inform an analysis plan.

Results: For an endpoint with frequency 8 per 100 000 patient-years, a sample size of 700 000 patients with mean follow-up 3 years has approximately 80% power to detect that hazard ratios of 1.2 and 1.8 are significantly different. I identified a cohort of 1 811 724 individuals in CALIBER with no prior cardiovascular disease, of whom 776 861 had a full blood count measurement. Current smokers had higher neutrophil, lymphocyte and monocyte counts than non or ex smokers.

Discussion: These findings inform the analysis protocol for the main study investigating leukocyte counts and cardiovascular diseases, reported in chapters 7 and 8. Smoking was found to have strong associations with leukocyte counts and will be an important confounder to adjust for in analyses.

6.3 Sample size and power calculation

I performed a sample size calculation by simulation to estimate the power to detect a difference in the hazard ratio per standard deviation higher white blood cell count between two event types. This calculation was based on a baseline incidence of 8 per 100 000

Table 6.1: Sample size estimation for detection of difference in hazard ratio for two endpoints

Hazard ratio per 1 SD higher biomarker level		Baseline incidence of event per 100 000 patient years		Power (%) to detect difference between two hazard ratios (mean 3 year follow-up) by sample size			
Event 1	Event 2	Event 1	Event 2	100 000	300 000	500 000	700 000
1.2	1.6	8	8	5.4	10.7	10.0	23.7
1.2	1.8	8	8	18.4	41.0	41.4	80.1
1.2	1.6	16	16	31.6	68.4	68.3	96.6
1.2	1.8	16	16	49.1	84.2	85.6	100.0

patient-years (as for subarachnoid haemorrhage, the endpoint expected to have the lowest incidence [219]).

I created 1000 simulated datasets, each containing 100 000 to 700 000 patients. I assumed white cell counts were distributed according to a standard normal distribution, and event times were distributed exponentially. The cause-specific hazard was dependent on white cell count with specified hazard ratios. Follow-up times were uniformly distributed between 0 and 6 years (i.e. mean 3 year follow-up).

I analysed the simulated data using cause-specific Cox models, considering event 1 as censoring when analysing event 2 and vice versa. The two cause-specific hazard ratios were considered to be significantly different if their 95% confidence intervals did not overlap.

The sample size calculation estimates that for an endpoint with frequency 8 per 100 000 patient-years, a sample size of 700 000 patients with mean follow-up 3 years has approximately 80% power to detect that hazard ratios of 1.2 and 1.8 are significantly different (**Table 6.1 on page 134**). This sample size is significantly larger than the UK Biobank cohort (n=500 000) and is generally only achievable in EHR cohorts or large consortia.

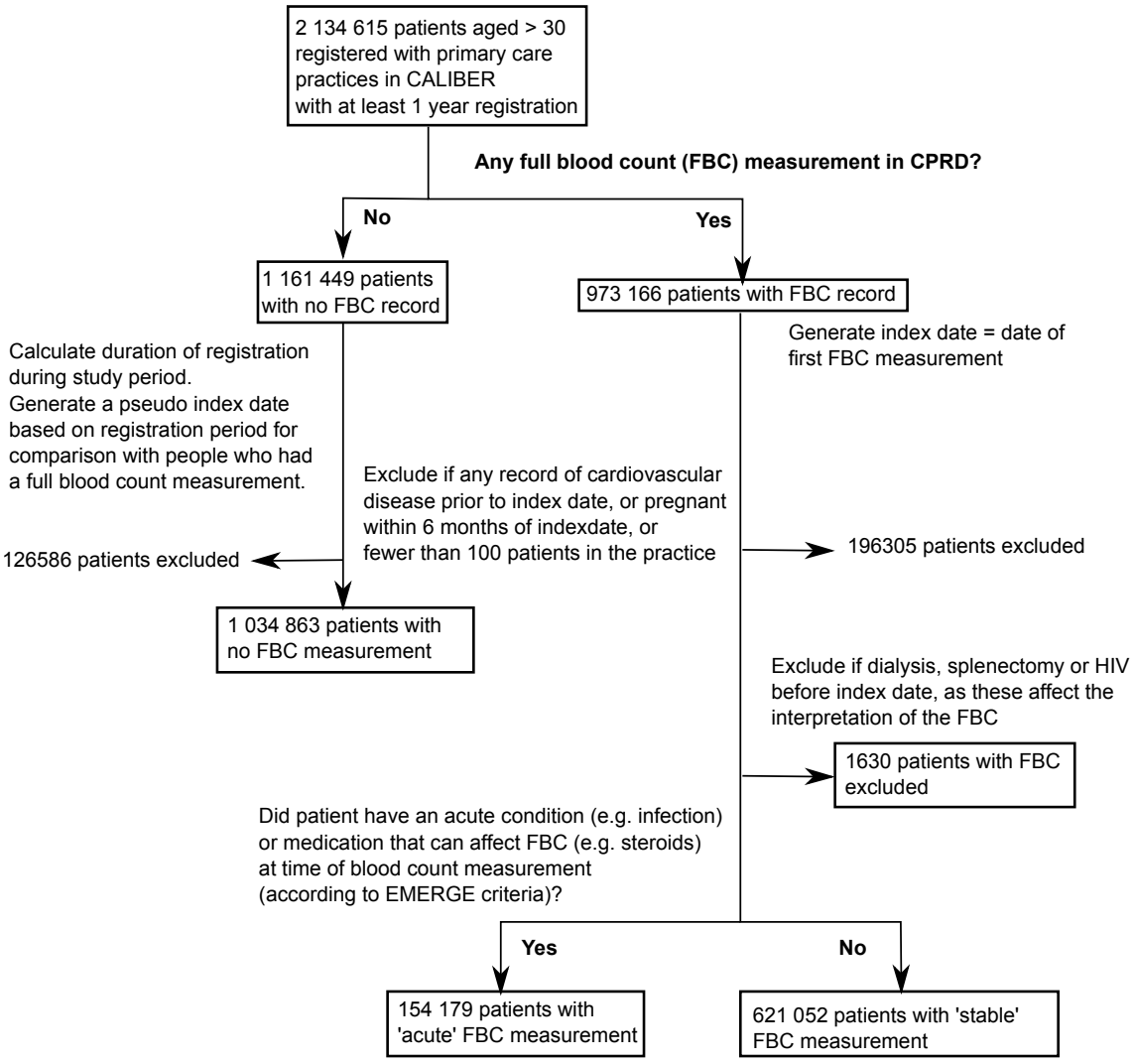
6.4 Definition of CALIBER study population

The patient flow diagram for this study is shown in **Figure 6.1 on page 135**.

The start and end dates for the study were decided based on the quality and completeness of data available in the CALIBER data sources. Although CPRD practices have been collecting data since the late 1980s, linkage to Hospital Episode Statistics commenced in January 1997. Each CPRD practice has an 'up-to-standard' date, defined as the earliest date that set of data completeness indicators are fulfilled. The eligibility start date was therefore chosen as January 1998, and individuals were eligible for inclusion when they were at least 30 years of age and had been registered for at least one year in an 'up-to-standard' practice. The lower age limit of 30 was chosen because below this age the incidence of cardiovascular endpoints was very low.

I extracted all full blood count measurements (including haemoglobin concentration,

Figure 6.1: Patient flow diagram



platelet count, total leukocyte count and differential leukocyte count) for all patients in the cohort. Details of the algorithm for extracting full blood counts are given in section 6.7 on page 139. The 'index date' (study start date) for each participant was the date of the first full blood count recorded in CPRD while the participant was eligible. Patients with a prior history of cardiovascular disease and women with a pregnancy record within 6 months of the start of the study were excluded, as in other CALIBER studies investigating the initial presentation of cardiovascular diseases [35, 36, 197, 220].

The study end date was March 2010, the last collection date of CPRD data. Patients were censored if they transferred out of the practice before the end of the study.

6.5 Classification of patient status at time of blood tests

White cell counts can be affected by many factors such as acute and chronic infections, autoimmune diseases, medication and haematological conditions. The electronic Medical

Records and Genomics (eMERGE) consortium has proposed an algorithm to identify the stable underlying white cell count in electronic health records using information about intercurrent illnesses and prescriptions [55]. This algorithm was developed for genetic studies investigating determinants of a patient's genetically determined stable white cell count. Studies using these methods do not aim to predict prognosis based on the white cell count; hence they can use all the information in the medical record regardless of time [56, 221].

My aim was slightly different; I wished to investigate the prognostic significance of white blood cell counts according to whether the patient had an acute medical condition that may affect white cell counts. Using information after the blood test (such as a diagnosis of infection) to classify patient status could create immortal time bias, as patients who survived would be more likely to have such information recorded. Therefore I limited the time window for information pertaining to patient status to be on or before the date of the blood test.

I converted the ICD-9, CPT4 and HCPC codes as used in the eMERGE algorithm to ICD-10, OPCS and Read codes for use with CALIBER data, and identified the relevant drug preparations (e.g. chemotherapy agents, methotrexate and steroids) in the CPRD drug dictionary.

The criteria for an 'acute' patient state were: in hospital on the date of blood test, vaccination in the previous 7 days, anaemia diagnosis within the previous 30 days, symptoms or diagnosis of infection within the previous 30 days, prior diagnosis of myelodysplastic syndrome, prior diagnosis of haemoglobinopathy, cancer chemotherapy or G-CSF within 6 months before index date, or the use of drugs affecting the immune system such as methotrexate or steroids within the previous 3 months.

In secondary analyses, I explored associations between onset of cardiovascular diseases and the mean of the first two 'stable' leukocyte count measurements taken since the start of eligibility.

6.6 Extracting covariate data

I extracted demographic variables, cardiovascular risk factors, comorbidities and acute conditions and prescriptions around the time of the blood test from CPRD. I extracted hospitalisation records and comorbidities additionally from HES.

For continuous covariates I used the most recent value in the year prior or up to 1 day after the full blood count measurement. I also extracted the first measurement after this time window and the most recent measurement before the time window, along with the timing of these measurements relative to the index date, to use as auxiliary variables for multiple imputation. The CALIBER variable definitions for continuous variables specified a valid range for each continuous variable; values outside this range (which were biologically implausible and almost certainly errors) were ignored.

I calculated body mass index from height and weight records in CPRD, handling outliers

and errors according to the algorithm of Bhaskaran et al. [175]. For smoking, I used all information prior to the index date in order to determine smoking status. Patients whose most recent record stated they were a smoker were considered to be current smokers. Of the remainder, those who had any previous record of being a current or ex smoker were considered to be ex smokers, and those with at least one non smoking record and no ex or current smoking records were considered to be never smokers.

Characteristics of patients included and excluded from the study are shown in **Table 6.2 on page 138**. For patients with a full blood count recorded, covariate information is based on the date of the blood count. For other patients, I generated a 'pseudo' index date for this comparison. I generated this pseudo index date as follows. First, I set a duration of registration prior to the index date for each patient without a full blood count recorded. I randomly selected this duration from the distribution of durations between registration and index date among patients with a full blood count recorded and a similar total duration of registration. The pseudo index date was generated by adding this duration to the date the patient registered with the practice.

Table 6.2: Characteristics of individuals in CALIBER with no prior history of cardiovascular disease

		FBC recorded while eligible for study		
	No FBC	**Patients excluded**	**'Acute' FBC**	**'Stable' FBC**
N patients	1 034 863	1630	154 179	621 052
Women, n (%)	455 290 (44.0%)	780 (47.9%)	94 383 (61.2%)	367 573 (59.2%)
Age, median (IQR)	41.5 (33.9-53.8)	53.1 (42.7-65.0)	56.2 (43.1-68.9)	51.6 (41.1-63.4)
Most deprived quintile, n (%)	220 256 (21.4%)	328 (20.2%)	27 196 (17.7%)	105 775 (17.1%)
Duration of registration before index date (years), median (IQR)	5.63 (1.98-13.6)	11.3 (4.09-19.1)	11.1 (4.21-20.3)	10.9 (4.22-19.6)
Total duration of registration during study period (years), median (IQR)	8.18 (3.47-17.3)	15.8 (8-23.7)	15.9 (8.34-24.8)	15.7 (8.39-24.1)
Ethnicity, n (%):				
White	389 590 (87.1%)	1086 (91.3%)	96 591 (92.7%)	348 443 (92.8%)
South Asian	14 697 (3.3%)	12 (1.0%)	3015 (2.9%)	10 435 (2.8%)
Black	20 471 (4.6%)	68 (5.7%)	2301 (2.2%)	7673 (2.0%)
Other	22 511 (5.0%)	24 (2.0%)	2282 (2.2%)	9038 (2.4%)
Missing	587 594 (56.8%)	440 (27.0%)	49 990 (32.4%)	245 463 (39.5%)
Smoking status, n(%):				
Never	456228 (53.2%)	754 (49.1%)	75432 (51.6%)	311270 (52.9%)
Ex	146577 (17.1%)	386 (25.1%)	37055 (25.4%)	140144 (23.8%)
Current	254115 (29.7%)	395 (25.7%)	33600 (23.0%)	137421 (23.3%)
Missing	177943 (17.2%)	95 (5.8%)	8092 (5.2%)	32217 (5.2%)
Recording within 1 year before index date:				
HDL and total cholesterol	43 157 (4.2%)	448 (27.5%)	40 520 (26.3%)	204 507 (32.9%)
Blood pressure	267 706 (25.9%)	1062 (65.2%)	97 069 (63.0%)	409 775 (66.0%)
eGFR	48 033 (4.6%)	892 (54.7%)	79 971 (51.9%)	284 628 (45.8%)
Body mass index	137 518 (13.3%)	540 (33.1%)	47 194 (30.6%)	195 144 (31.4%)
Most recent value within one year prior to index date:				
Systolic blood pressure, mmHg	129 (118-140)	134 (122-146)	136 (122-150)	136 (121-150)
Body mass index, kg/m^2	26.1 (23.1-30.1)	26.4 (23.4-30.3)	26.9 (23.5-31.1)	27.1 (23.8-31.1)
Total cholesterol, mmol/L	5.3 (4.54-6)	5.32 (4.6-6.2)	5.4 (4.7-6.2)	5.5 (4.8-6.21)
HDL cholesterol, mmol/L	1.32 (1.1-1.6)	1.35 (1.1-1.7)	1.4 (1.11-1.7)	1.4 (1.13-1.7)
eGFR, mL/min/1.73m^2	86.8 (72.4-100)	80.1 (63.5-94.6)	79.8 (66.2-93.5)	82.5 (69.6-95.5)
Diagnoses on or before index date:				
Atrial fibrillation	3777 (0.4%)	35 (2.1%)	2221 (1.4%)	5501 (0.9%)
Cancer	30 216 (2.9%)	355 (21.8%)	37 565 (24.4%)	9956 (1.6%)
Diabetes	18 733 (1.8%)	140 (8.6%)	8361 (5.4%)	28 766 (4.6%)
Atopy (hayfever, asthma or atopic dermatitis)	130 120 (12.6%)	334 (20.5%)	31 951 (20.7%)	114 716 (18.5%)
Chronic obstructive pulmonary disease	9294 (0.9%)	50 (3.1%)	6396 (4.1%)	8625 (1.4%)
Connective tissue disease	9295 (0.9%)	81 (5.0%)	11 208 (7.3%)	10 651 (1.7%)
Inflammatory bowel disease	4723 (0.5%)	26 (1.6%)	3489 (2.3%)	5157 (0.8%)
Medication use in the year before index date:				
Antihypertensives	96 323 (9.3%)	825 (50.6%)	43 779 (28.4%)	153 502 (24.7%)
Statins	19 388 (1.9%)	185 (11.3%)	9442 (6.1%)	37 364 (6.0%)

FBC, full blood count

6.7 Extraction of leukocyte counts from CPRD data

Laboratory values are generally transferred automatically to the general practice system and are recorded in CPRD as numerical values, with additional data fields for the operator (equal, greater than or less than), units and reference range. Each record is associated with an 'entity type' defining the type of result contained in the record (e.g. 'Neutrophil count') and defining the meaning of the other data fields, and a Read code (e.g. 'Percentage neutrophils', specifying the test in more detail. There are a few hundred entity types in the CPRD information model but thousands of potential Read terms.

A common format for the majority of blood test results was as follows:

Column	Description
patid	Patient identifier
eventdate	Date of the blood test
sysdate	System date (date the information was entered on the GP system)
enttype	Entity type
medcode	CPRD medical code corresponding to Read code / Read term
data1	Operator (coded entry: =, $<$, $>$)
data2	Value (numeric)
data3	Unit of measure (coded entry: 10*9/L, %, etc.)
data4	Qualifier (coded entry: High, Low, Normal, Abnormal, etc.)
data5	Normal range from (numeric)
data6	Normal range to (numeric)
data7	Normal range basis (coded entry: age based, age and sex etc.)

The meaning of the data fields could vary between entity types. I used a sample CALIBER dataset (100,000 patients with stable coronary disease) to investigate how white cell counts were recorded and what steps would be required to convert the raw values into a variable suitable for use in modelling.

6.7.1 Example: extraction of total white cell count

First I generated a codelist for total white cell count:

Read code	Read term
42H..00	Total white cell count
42H..11	White blood count
42H..12	White cell count
42H1.00	White cell count normal
42H2.00	Leucopenia - low white count
42H2.11	Leucopenia
42H3.00	Leucocytosis -high white count
42H3.11	Leucocytosis
42H5.00	White cell count abnormal
42H7.00	Total white blood count
42H8.00	Total WBC (IMM)
42HZ.00	Total white cell count NOS
D400A00	Leucopenia
D40y.11	Leucocytosis
D40y.12	Leukocytosis
D40yz11	Leucocytosis

Next I tabulated the entity types associated with codes in this codelist.

Entity code	Entity description	Number of records
207	Total White Blood cell count	708810
288	Other laboratory tests	24
346	Other autoantibodies	1
363	Lipoprotein electrophoresis	1

I decided to extract only entity type 207; there may be a few white cell counts recorded with entity type 288, but including this entity type in the extraction algorithm would make it more complicated and increase the risk of error without much benefit.

This is a tabulation of the 8 most frequent units associated with entity type 207:

Unit of measure	Percentage of records
10*9/L	94.2
(Missing)	2.4
10*9	1.3
/L	0.7
10*12/L	0.3
10*9/mL	0.3
10*6/L	0.2
(ph)	0.2

I plotted the distribution of values associated with each of these units to verify that they seemed sensible, and identify if any units had the wrong scale (**Figure 6.2 on page 142**). Some values had missing units, and in order to decide whether to use these values I reviewed their distribution to see if it was possible to assume what the unit should be. I decided how to handle zero values based on whether it is biologically plausible for a value to be zero; this was the case for white cell counts (severely immunocompromised patients might have a white cell count of zero) but not for measures such as haemoglobin or HbA1c.

In the case of total white cell count I found that values with missing units were much more likely to be zero, suggesting that many of the zero values are actually missing. However, as it is impossible to differentiate between a missing and true zero value, I excluded all values without units.

Based on this exploratory work and a consensus on the range of biologically plausible values (based on the literature and clinical experience), I derived an algorithm for extraction of each white cell count and other clinically recorded biomarkers. The algorithm pseudocode was converted into SQL by the data manager in order to extract the values from the master CALIBER database. This is the algorithm pseudocode for total white cell count:

```
If Read code is in the total_wbc_gprd codelist
AND enttype = 207
AND data2 [value] < 50
AND data3 [units] IN (37 [10*9/L], 153 [10*9], 17 [/L])
THEN total_wbc_gprd = data2
```

The algorithms are published on the CALIBER data portal (`https://www.caliberresearch.org/portal/`). The algorithm for total white cell count is here: `https://www.caliberresearch.org/portal/show/total_wbc_gprd`

Figure 6.2: Distribution of values associated with different units for raw CPRD white cell count data

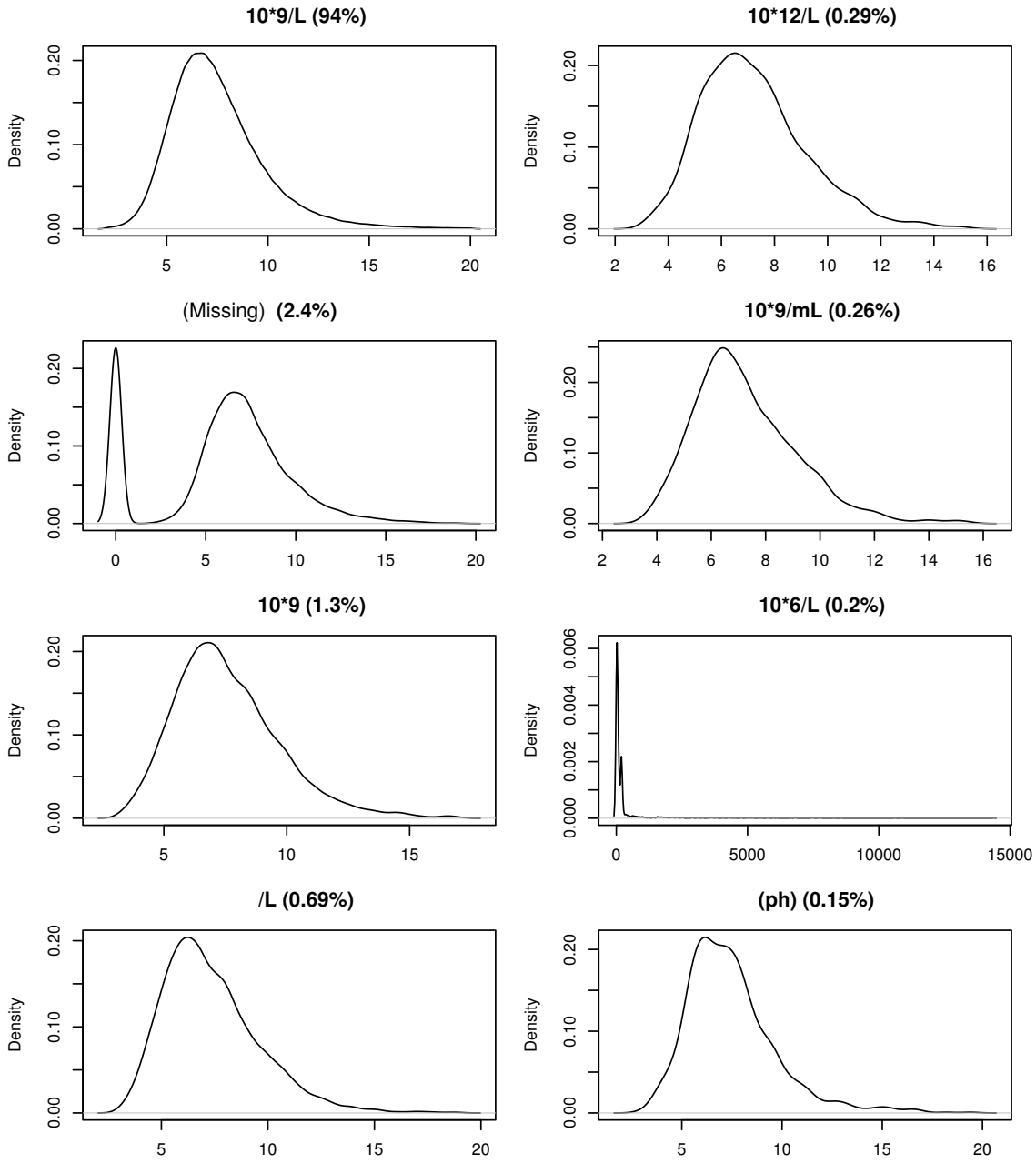
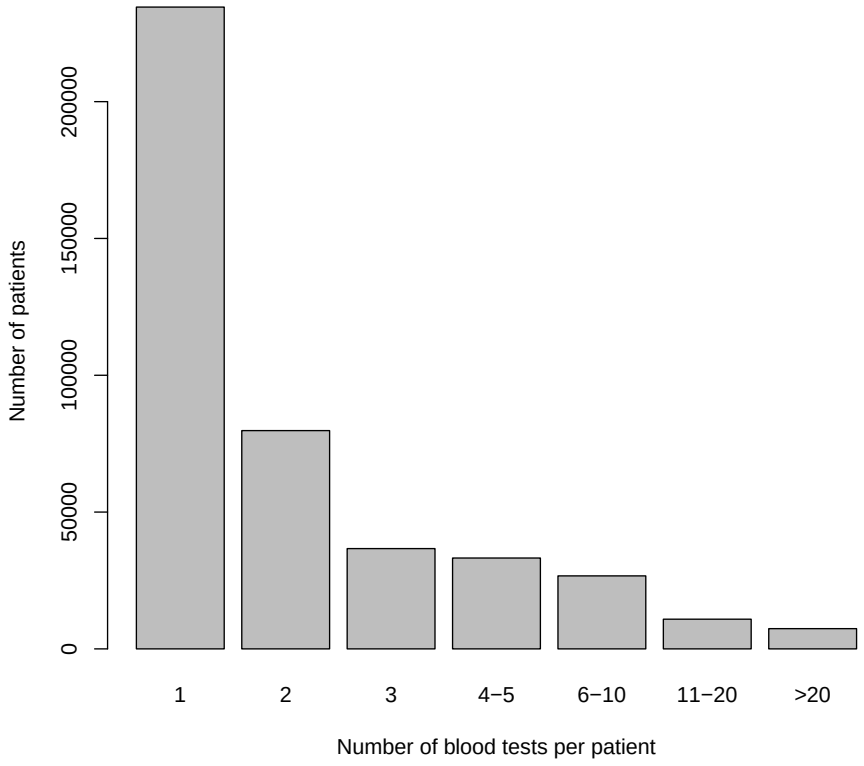


Figure 6.3: Number of full blood count tests per patient in CALIBER



The data extraction algorithm was implemented on the main CALIBER server to extract the components of the full blood count as recorded in CPRD: haemoglobin concentration, platelet count, total white cell count, neutrophil count, eosinophil count, basophil count, monocyte count and lymphocyte count. If a patient had more than one measurement on a given day, the values were aggregated by taking the mean.

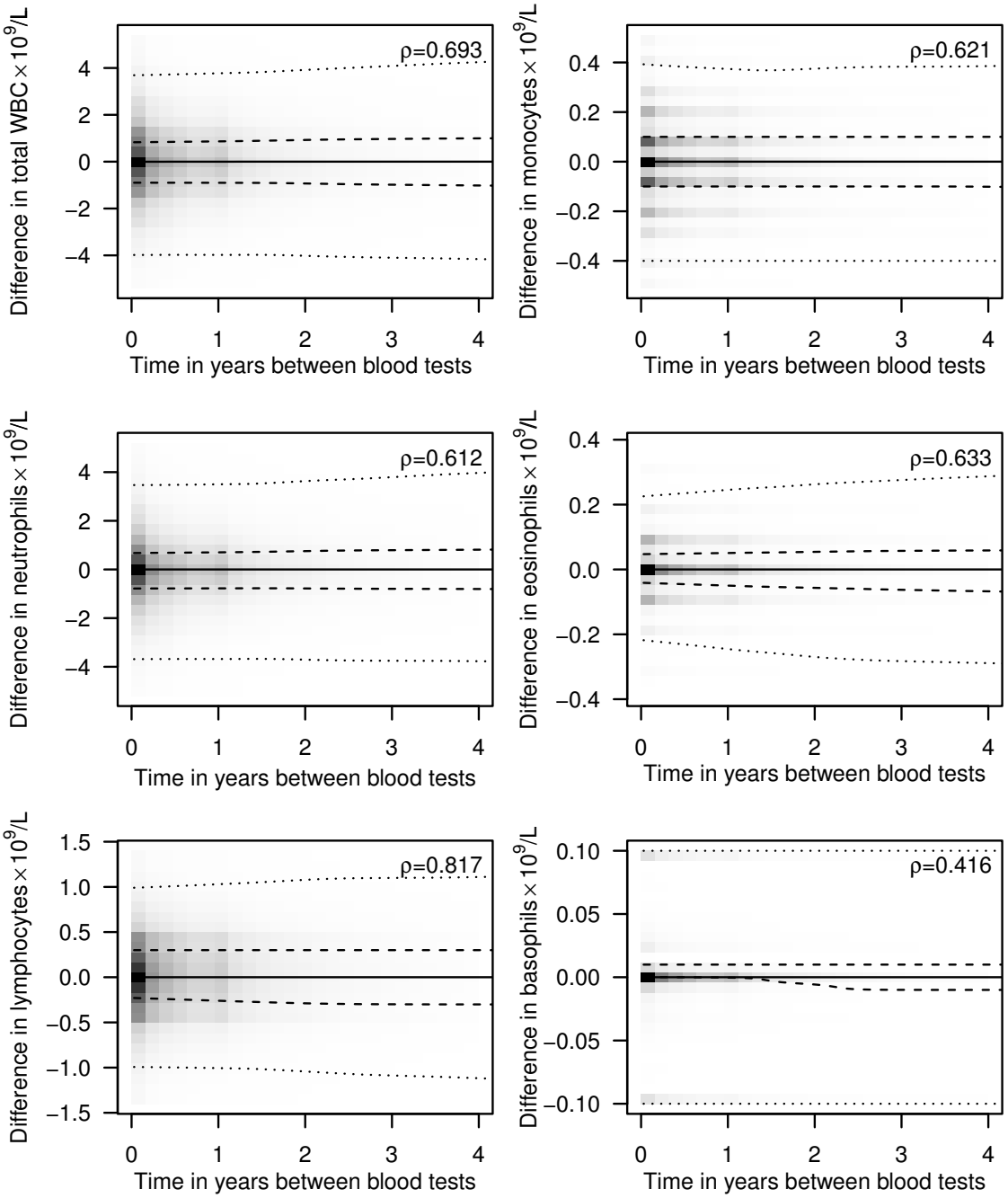
6.7.2 Results

The median number of full blood counts available per patient was 2 (mean 3.7, standard deviation 5.2, maximum 240). Two-thirds of patients (731 900 / 1 092 068, 67.0%) had more than one full blood count measurement, 513 009 (47.0%) had more than two and 52 137 (4.8%) had more than ten (**Figure 6.3 on page 143**).

The median interval between two full blood count blood tests was 0.50 years. There was no general trend for leukocyte counts over time (**Figure 6.4 on page 144**), but there was only moderate correlation between blood tests performed at different times in the same patient. For example, the within-person correlation of neutrophil counts taken under 'stable' conditions was 0.568. This compares to 0.68 for total cholesterol and 0.73 for HDL cholesterol [105].

Figure 6.4: Binned scatterplots of difference between any two consecutive leukocyte counts in the same patient, by time between measurements.

The central solid line is lowess smoothed mean; the dotted lines are lowess smoothed percentile lines at 2.5%, 25%, 75% and 97.5%. ρ is the correlation coefficient between two consecutive measurements.



6.8 Associations between differential leukocyte counts and cardiovascular risk factors

I plotted frequency polygons for the distribution of leukocyte counts by age, sex and levels of cardiovascular risk factors which are known or suspected to be associated with leukocyte counts. These included smoking [62], ethnicity [222], diabetes [63] and the patient's clinical condition at the time of blood testing. I using Mann-Whitney-Wilcoxon tests (for two groups) or Kruskal-Wallis tests (for multiple groups) to compare the distributions of leukocyte counts by values of cardiovascular risk factors.

Current smokers had higher neutrophil, lymphocyte and monocyte counts than non or ex smokers (P <0.001 by Kruskal-Wallis tests; **Figure 6.5 on page 146**).

Leukocyte counts were fairly similar across age groups (**Figure 6.6 on page 147**) and between genders (**Figure 6.7 on page 148**), apart from slightly higher neutrophil counts in women and higher monocyte counts in men in this population (both P <0.001).

Black people had lower neutrophil counts than other ethnicities (**Figure 6.8 on page 149**) (P <0.001). Individuals with diabetes and those with blood tests taken under 'acute' conditions tended to have higher neutrophil counts and lower lymphocyte counts than those taken under 'stable' conditions (**Figure 6.9 on page 150**, **Figure 6.10 on page 151**) (all P <0.001).

These associations of leukocyte counts with other cardiovascular risk factors show that they may be confounders and it is essential to adjust for them in multivariable models investigating leukocyte counts and incidence of cardiovascular diseases. Given the strong associations with smoking, it may also be necessary to examine interactions with smoking status. This is because imprecise recording of smoking severity in this dataset can cause residual confounding. Testing for interactions by smoking status can overcome this limitation; if the same associations are seen in never-smokers as in smokers, it cannot be due to confounding by smoking status.

Figure 6.5: Distribution of leukocyte counts by smoking status

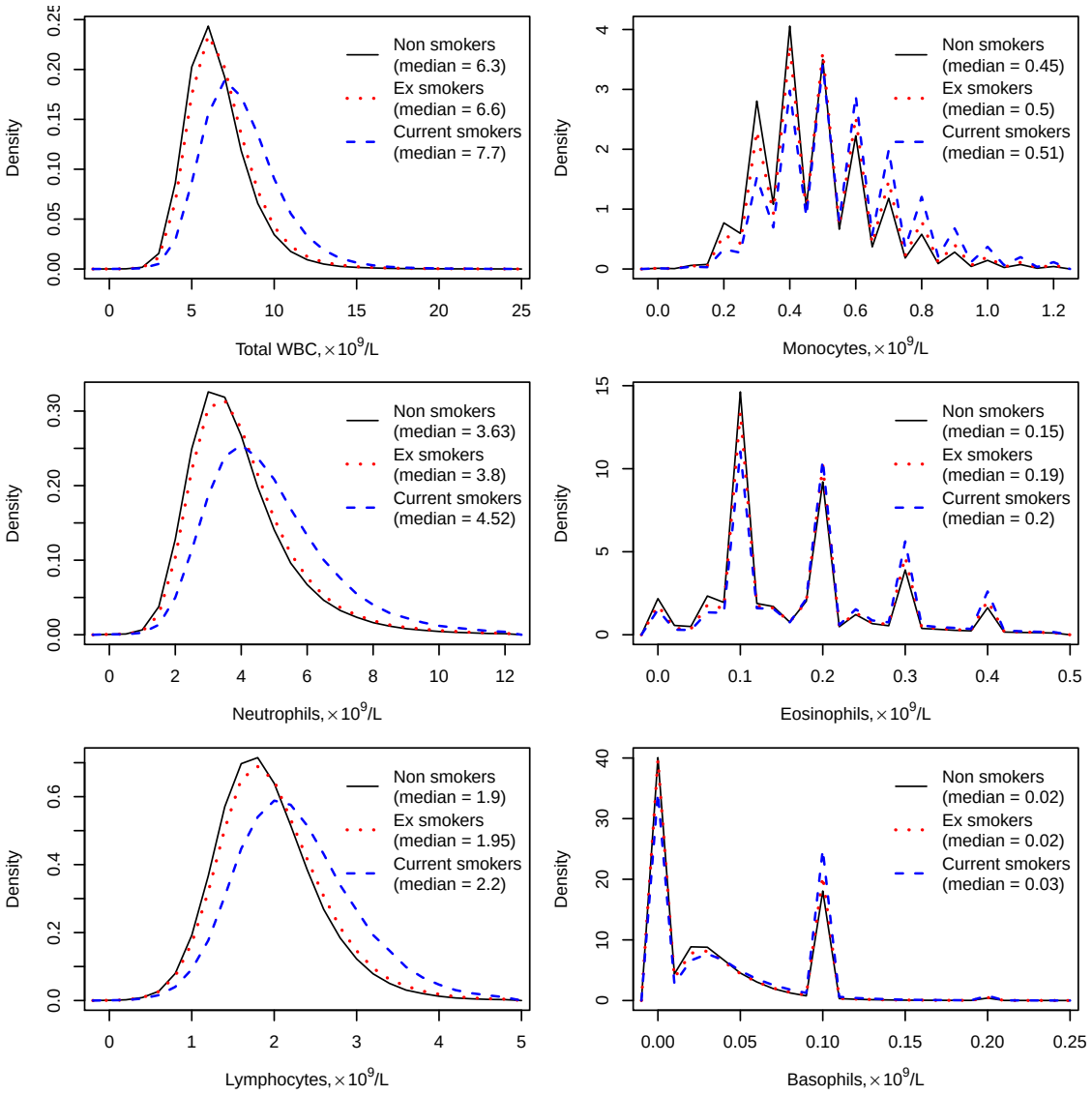


Figure 6.6: Distribution of leukocyte counts by age

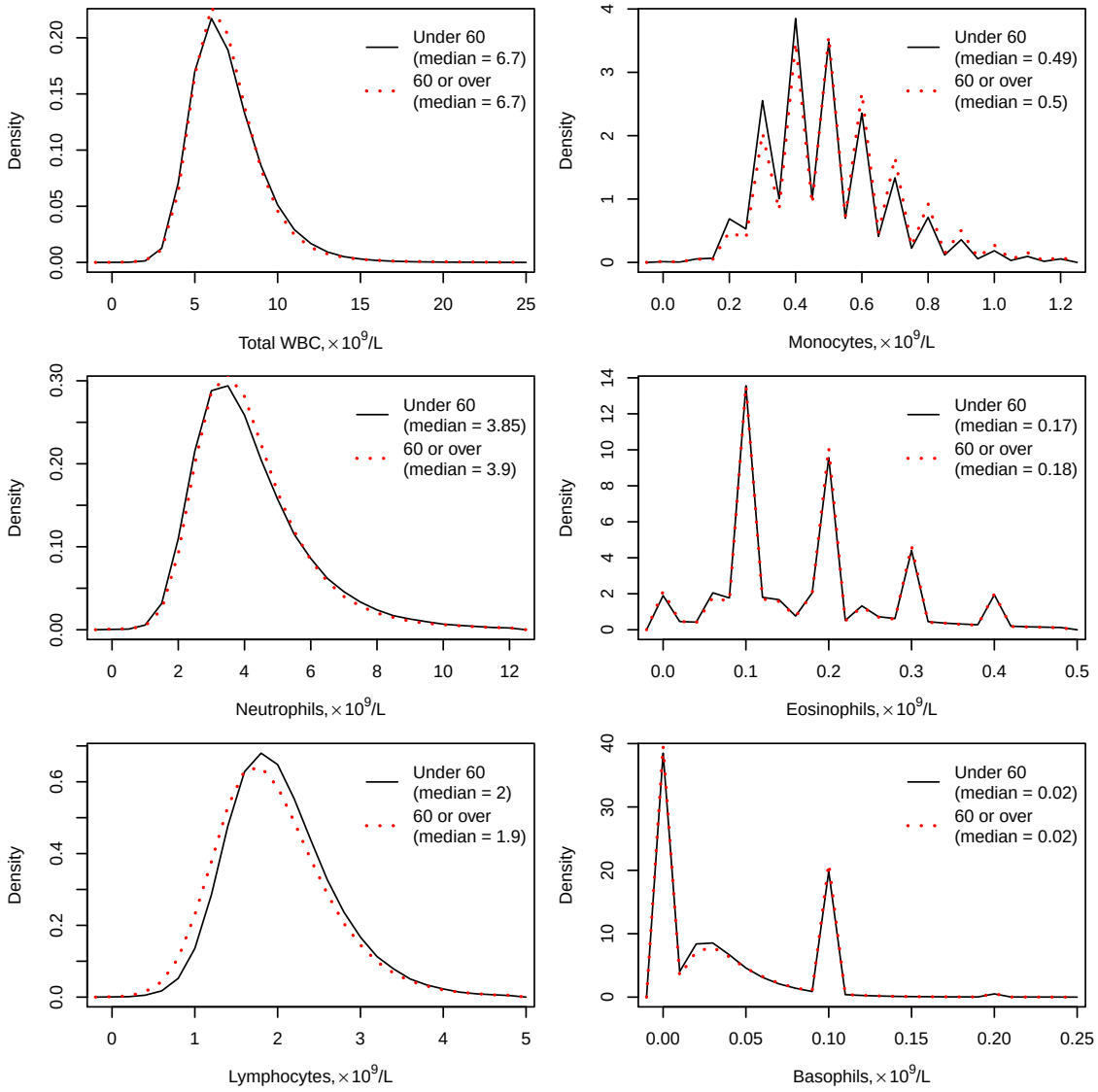


Figure 6.7: Distribution of leukocyte counts by sex

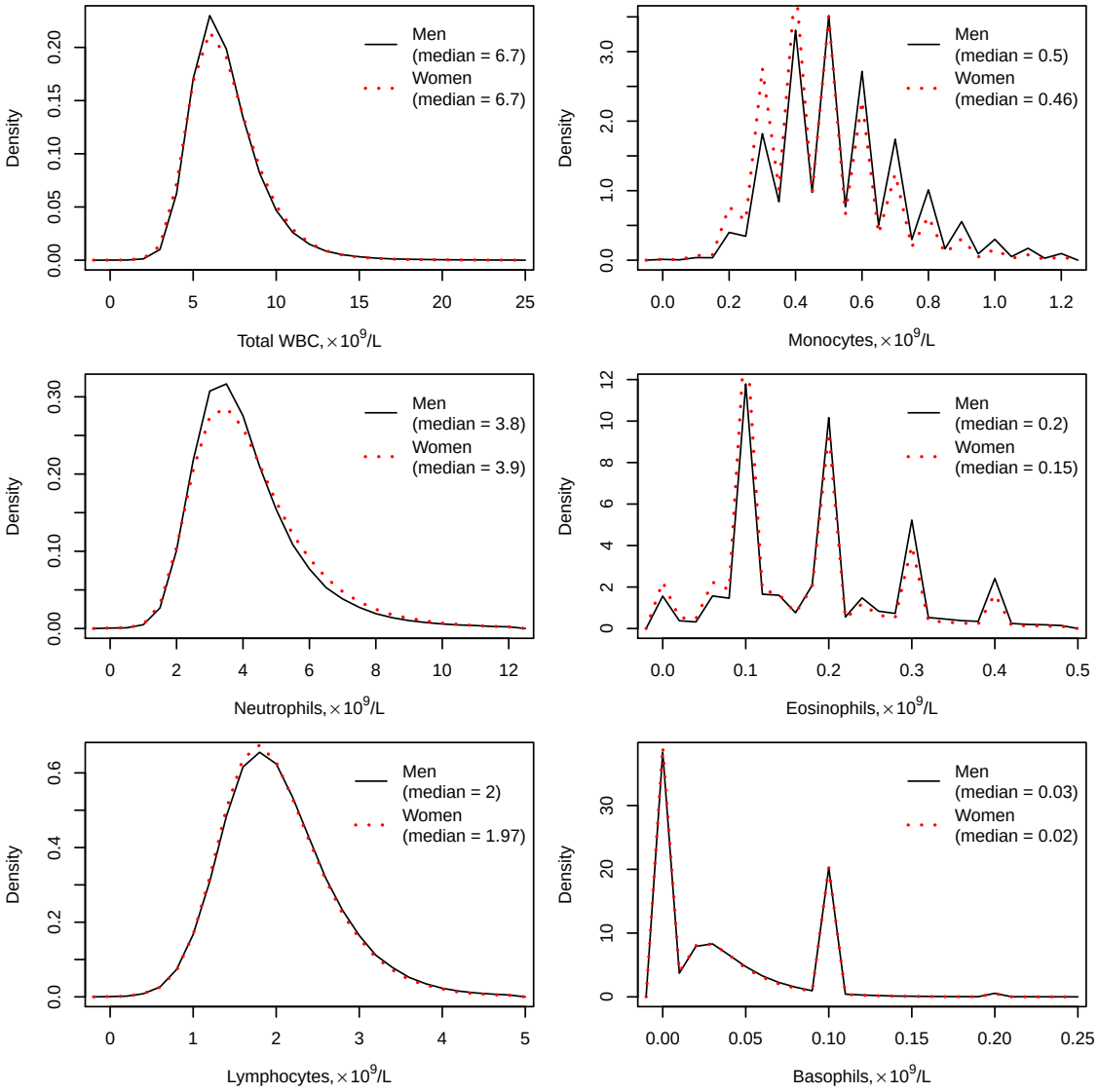


Figure 6.8: Distribution of leukocyte counts by ethnic group

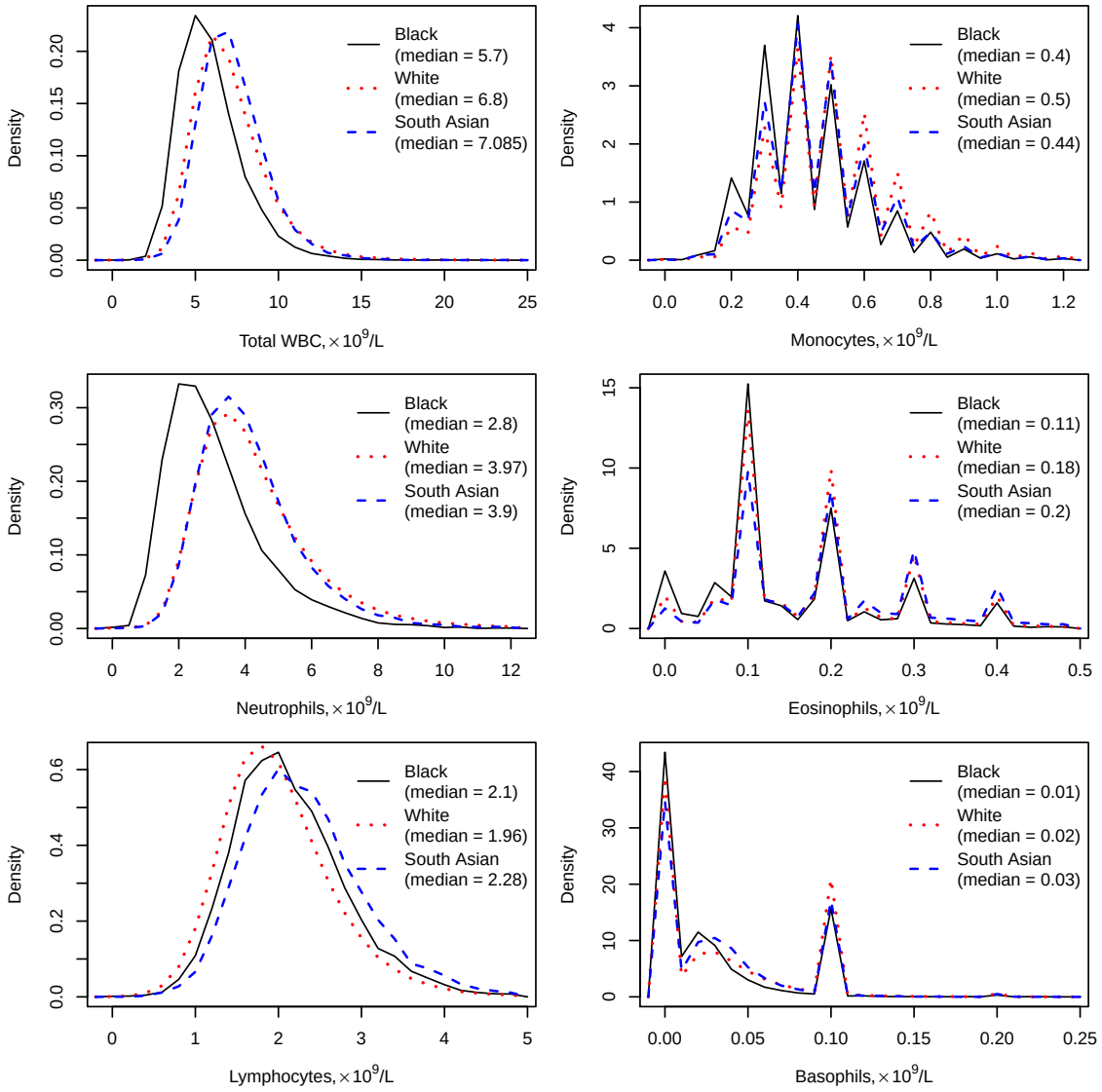


Figure 6.9: Distribution of leukocyte counts by patient condition at time of blood testing

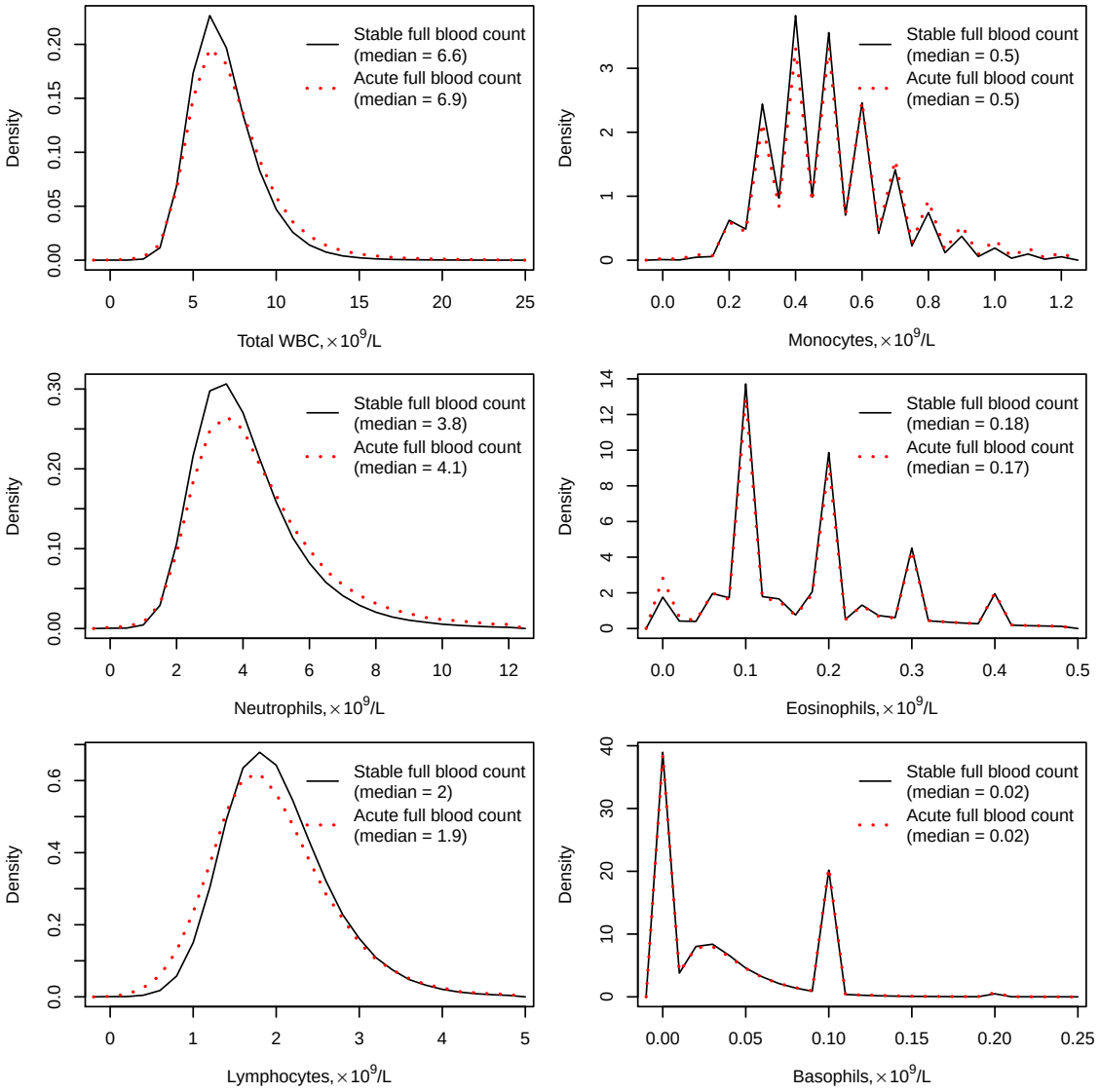
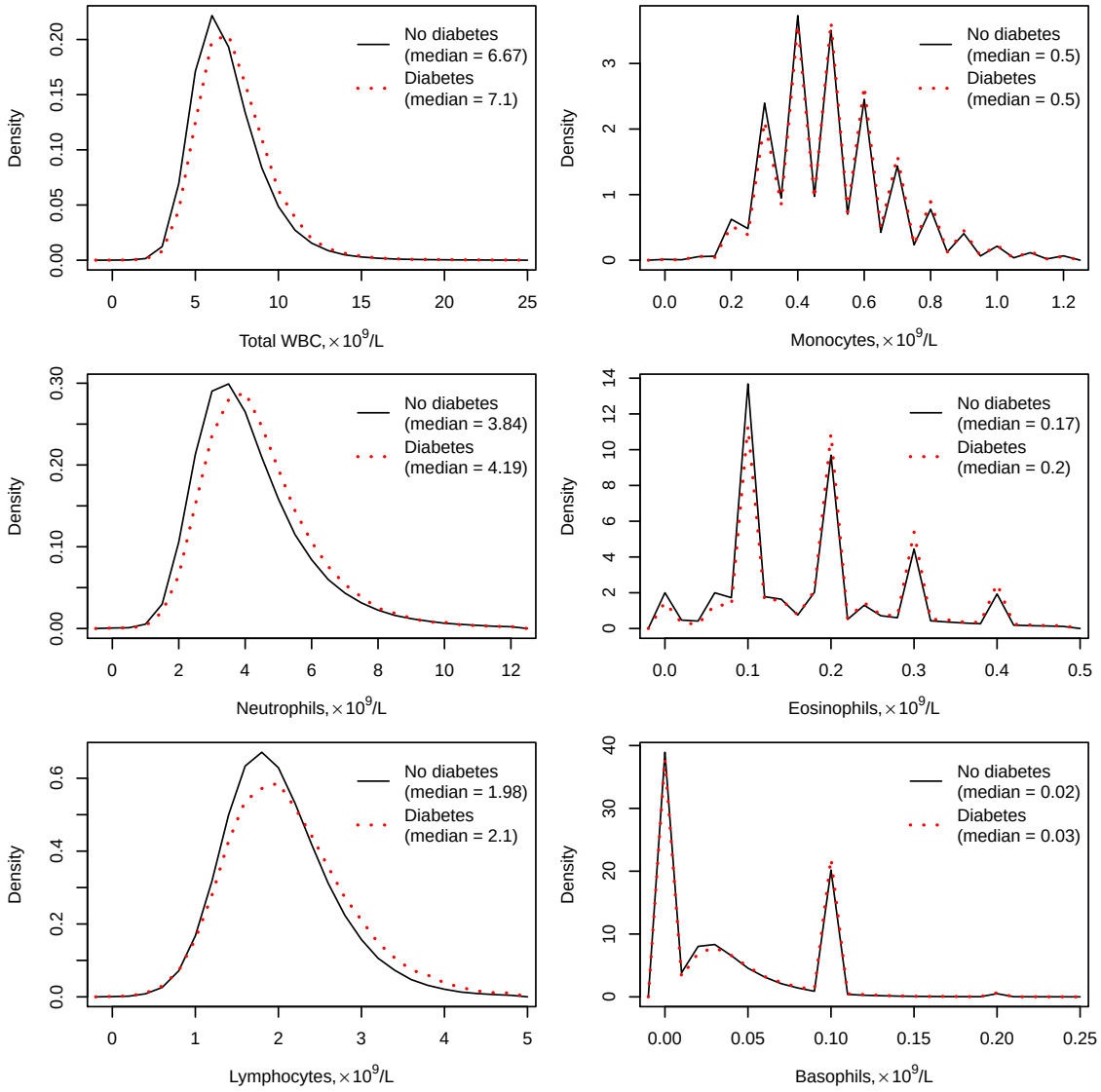


Figure 6.10: Distribution of leukocyte counts by diabetes status



6.9 Summary

For an endpoint with frequency 8 per 100 000 patient-years, a sample size of 700 000 patients with mean follow-up 3 years has approximately 80% power to detect that hazard ratios of 1.2 and 1.8 are significantly different.

I identified a cohort of 1 811 724 individuals in CALIBER patients with no prior cardiovascular disease, of whom 776 861 had a full blood count measurement while eligible. After excluding 1630 individuals with a history of dialysis, splenectomy or HIV, the cohort contained 621 052 individuals with a full blood count measured while clinically 'stable' and 154 179 with an 'acute' measurement.

Current smokers had higher neutrophil, lymphocyte and monocyte counts than non or ex smokers. Black people had lower neutrophil counts than other ethnicities. Individuals with diabetes and those with blood tests taken under 'acute' conditions tended to have higher neutrophil counts and lower lymphocyte counts than those taken under 'stable' conditions. Smoking has a strong association with both cardiovascular diseases and leukocyte counts, and is an important confounder to be adjusted for in the analysis.

This cohort will be used to investigate the association of differential leukocyte counts with the onset of cardiovascular diseases (reported in chapters 7 and 8). The following chapter, chapter 9 compares the association of total leukocyte count and mortality in CALIBER and the New Zealand PREDICT cohort.

7 Associations of neutrophils and other leukocyte counts with specific initial presentations of cardiovascular diseases

7.1 Chapter outline

This chapter investigates the association of neutrophil counts with incidence of cardiovascular diseases. The cohort definition is described in chapter 6.

7.2 Abstract

Background: The white blood cell (leukocyte) differential is one of the most common blood tests and some components, particularly neutrophils, reflect a patient's inflammatory state. Inflammation is a key process in atherosclerosis, which underlies many cardiovascular diseases, but little is known about the association of neutrophil counts with a wide range of cardiovascular diseases.

Methods: The data source was a linked primary care, hospital admission, disease registry, and mortality dataset for England (the CALIBER programme). The study population comprised people aged 30 or older who had a leukocyte differential recorded and were free from cardiovascular disease at baseline. Individuals were followed up for the first occurrence one of twelve cardiovascular presentations in any of the data sources: stable angina, unstable angina, myocardial infarction, unheralded coronary death, heart failure, ventricular arrhythmia / sudden cardiac death, transient ischaemic attack, ischaemic stroke, haemorrhagic stroke, subarachnoid haemorrhage, abdominal aortic aneurysm or peripheral arterial disease. Cox models were used to estimate cause-specific hazard ratios (HR) adjusted for smoking, other cardiovascular risk factors and acute conditions that could affect leukocyte counts (such as infections and immunosuppressant medication).

Results: Among 775 231 individuals in the cohort, there were 54 980 initial presentations of cardiovascular disease over median follow up of 3.8 years. Adjusted hazard ratios comparing neutrophil counts $6-7$ vs $2-3\times10^9$/L (within the normal range) showed strong associations with heart failure (HR 2.04; 95% confidence interval (CI) 1.82, 2.29), non-fatal myocardial infarction (HR 1.58; 95% CI 1.42, 1.76) and unheralded coronary death (HR 1.78; 95% CI 1.51, 2.10), but not stable angina (HR 0.97; 95% CI 0.88, 1.07) or unstable angina (HR 1.00; 95% CI 0.84, 1.19). The association with ischaemic stroke was weaker (HR 1.36; 95% CI 1.17, 1.57) and there was no association with haemorrhagic stroke, but there were strong associations with peripheral arterial disease (HR

1.95; 95% CI 1.72, 2.21) and abdominal aortic aneurysm (HR 1.72; 95% CI 1.34, 2.21). The associations of cardiovascular diseases with monocyte counts were similar to those with neutrophils but less strong. Low lymphocyte counts and high basophil counts were associated with increased incidence of heart failure.

Conclusions: Neutrophil counts are strongly associated with myocardial infarction, heart failure and peripheral arterial disease. Given the known causal relevance of neutrophils in cardiovascular diseases, the clinical relevance of this inflammatory marker should be further investigated in risk prediction.

7.3 Introduction

The differential white blood cell (leukocyte) count is one of the most common blood tests in medical practice. It comprises counts of different cell types associated with immunity and inflammation, such as neutrophils, lymphocytes, monocytes, eosinophils and basophils. These circulating blood cells play a major role in atherosclerosis, the pathological condition underlying coronary disease [48]. However, clinicians rarely use these cell counts in cardiovascular decision making, due to insufficient understanding how they relate to onset of cardiovascular diseases.

Previous studies found that neutrophil counts are associated with higher incidence of coronary disease [49, 116, 120, 121, 124, 125], heart failure [113] and stroke [122] (**Table 2.2 on page 47**). However, these studies failed to disaggregate the spectrum of coronary diseases, and tended to be small, with limited power to study interactions with age or sex. The association of neutrophil counts with acute as compared to chronic coronary initial presentations are not known, and no studies have investigated associations with vascular disease outside the heart and brain, such as abdominal aortic aneurysm or peripheral arterial disease. These associations cannot be extrapolated from other cardiovascular diseases; previous studies on other cardiovascular risk factors have found that the strength and direction of associations can vary considerably by cardiovascular presentation [35, 36, 197, 220].

The aim of this study was to investigate the association of counts of neutrophils and other leukocytes with the initial presentation of cardiovascular diseases in the CALIBER cohort [1].

7.4 Methods

7.4.1 Study population

The study population was drawn from the CALIBER linked electronic health record dataset [1] and is described in section 6.4. Briefly, individuals were eligible for inclusion when they were aged 30 or older and had been registered for at least one year in a

practice which met research data recording standards.

The study start date, ('index date') for each participant was the date of the first full blood count measurement recorded in CPRD while the participant was eligible. Patients with a prior history of cardiovascular disease and women with a pregnancy record within 6 months of the start of the study were excluded.

7.4.2 Exposure

The main exposure was the differential white cell count as recorded in CPRD, which was extracted as described in subsection 6.7.2. White cell counts can be affected by many factors such as acute and chronic infections, autoimmune diseases, medication and haematological conditions. Information recorded in CALIBER around the time of blood count measurement was used to allocate the patient state as 'stable' or 'acute' (section 6.5). Patients with HIV, splenectomy or on dialysis were excluded from this study, as their leukocyte counts may be difficult to interpret.

In secondary analyses, I explored associations between onset of cardiovascular diseases and the mean of the first two 'stable' measurements of neutrophil count taken since the start of eligibility.

7.4.3 Follow-up and endpoints

Patients were followed up while registered at the practice until the occurrence of an initial presentation of cardiovascular disease, death or transfer out of the practice. The primary endpoint was the first record of one of 12 initial presentations of cardiovascular disease in any of the data sources, as described earlier (section 5.4.4).

7.4.4 Covariates

I extracted demographic variables, cardiovascular risk factors, comorbidities and acute conditions and prescriptions around the time of the blood test from CPRD, as described in section 6.6.

7.4.5 Statistical analysis

I examined associations of cardiovascular diseases and quintiles of leukocyte subtypes in order to explore the shape of any associations. If the association was found to be linear, I also investigated models including the leukocyte subtype count as a linear variable. The precision with which leukocyte subtype counts were recorded varied from one to two decimal places (units: cells $\times 10^9$/L). In order to avoid biasing the category allocation by precision, I adjusted the category boundaries so that the second decimal place was 5. Basophil counts were very low relative to the precision of recording so I generated three categories instead of five. All category intervals were closed at the lower end and open

at the higher end; for example the category 3.45–4.25 includes 3.45 and all intermediate values up to but not including 4.25, and can be formally represented as "[3.45, 4.25)".

For neutrophil counts I sought to present results for clinically meaningful neutrophil counts, and separate values within and outside the reference range. Therefore although the initial exploratory analyses used quintiles of neutrophil count, I categorised neutrophils in units of 1×10^9 cells/L for the main results, to make them easier to interpret. For this study I defined 'normal' neutrophil counts as $2 - 7 \times 10^9$/L, which is a commonly quoted reference range [223] (although all laboratories quote slightly different ranges). This lent itself to convenient 5-level categorisation of neutrophil counts within the 'normal' range, with 2 additional categories for values lower or higher than the limits of the 'normal' range.

I generated cumulative incidence curves by category of neutrophil counts under a competing risks framework. I used the Cox proportional hazards model to generate cause-specific hazards for the different cardiovascular endpoints. Hazard ratios were adjusted for age (linear and quadratic), sex, age / sex interaction, index of multiple deprivation, ethnicity, smoking status, diabetes, body mass index, systolic blood pressure, estimated glomerular filtration rate (eGFR), high density lipoprotein cholesterol (HDL), total cholesterol, atrial fibrillation, inflammatory conditions (autoimmune conditions, inflammatory bowel disease or chronic obstructive pulmonary disease (COPD)), cancer, statin use, blood pressure medication and acute conditions at the time of blood testing. The baseline hazard was stratified by practice and sex. I plotted Schoenfeld residuals to assess the proportional hazards assumption, and split the follow-up time if hazard ratios changed over time. I handled missing baseline covariate data by multiple imputation as described in subsection 7.4.7.

7.4.6 Survival analysis and competing risks

I carried out survival analysis to describe and model the first occurrence of any cardiovascular disease. A patient's follow-up ended when they experienced one of the cardiovascular endpoints or when they were censored. Subsequent events (e.g. myocardial infarction occurring after stable angina) were not analysed.

To describe the incidence of each initial presentation over time we constructed cumulative incidence curves, taking into account the other possible initial presentations as competing events. Normal-based confidence intervals were constructed based on Greenwood's variance formula, as implemented in the R prodlim package [202].

For multivariable modelling I considered using Cox models (for modelling cause-specific hazards) or the Fine and Gray model (to compare cumulative incidence curves by modelling subdistribution hazards). As the aim of this study was observational epidemiology – to explore associations rather than predict risk – I decided that cause specific hazard ratios were the most appropriate quantity to estimate. Cause specific hazards should be interpreted together with cumulative incidence curves but cannot be used to

predict cumulative incidence. I experimented with the Fine and Gray model in R but found that it was computationally intensive, and would require significant software engineering or computing time to apply it to the large dataset used in these analyses.

I used follow-up time as the timescale for the Cox models in order to investigate how the strength of association between a leukocyte count measurement and outcome varies depending on time since measurement, and determine whether it is more useful for short / medium or long term prediction.

7.4.7 Multiple imputation

I used two methods of multiple imputation for imputing missing data: Random Forest and normal-based MICE. For the primary analysis I used Random Forest multiple imputation, as described in section 4.4, and implemented in the CALIBERrfimpute package [147]. Categorical variables were imputed using the 'rfcat' function, in which each imputed value is the prediction from a randomly chosen tree. Continuous variables were imputed using the 'rfcont' function, in which imputed values are randomly drawn from normal distributions centred on imputed means estimated using Random Forest.

As Random Forest imputation is a novel method, I also applied the established method of normal-based MICE imputation. I included the following variables in imputation models:

- Full blood count parameters: total leukcocyte count, neutrophil count, lymphocyte count, eosinophil count, basophil count, monocyte count, haemoglobin concentration, platelet count. For normal-based imputation, I included leukocyte subtype counts as categorical variables (quintiles) as they would be used in the analysis, in order to avoid imposing linearity in the imputation models.
- Cardiovascular risk factors including those in the substantive models: age, age squared, sex, body mass index, blood pressure, diabetes, smoking, total cholesterol, HDL cholesterol, triglycerides, eGFR, index of multiple deprivation
- Event indicator
- Type of event
- Time as the marginal Nelson-Aalen cumulative hazard [193]
- Use of statins or blood pressure lowering medication in the year before study entry
- Conditions potentially affecting blood counts, based on eMERGE criteria [55]: haemoglobinopathy, prior diagnosis of myelodysplasia, anaemia diagnosis within 30 days prior, anaemia diagnosis in folowing 30 days, prior diagnosis of cancer, cancer diagnosis in following 2 years, chemotherapy within 6 months prior, chemotherapy in following 3 months, renal dialysis prior, HIV diagnosis prior, steroid prescription within 3 months prior, methotrexate prescription within 3 months prior, prescription

for another drug affecting the immune system within 30 days prior, infection diagnosis within 30 days prior, infection diagnosis within 30 days after, infective symptoms within 30 days prior, infective symptoms within 30 days after, splenectomy prior, immunisation within 1-7 days prior

- Prior record of other comorbidities: atrial fibrillation, chronic obstructive pulmonary disease, inflammatory bowel disease, atopy, asthma, cancer, systemic autoimmune conditions

For imputing continuous measurements using Random Forest, I additionally used the last measurement before the 1-year time window before study entry, and the first measurement after study entry, along with the timing of these measurements relative to the study start date, as auxiliary variables for the imputation of that variable.

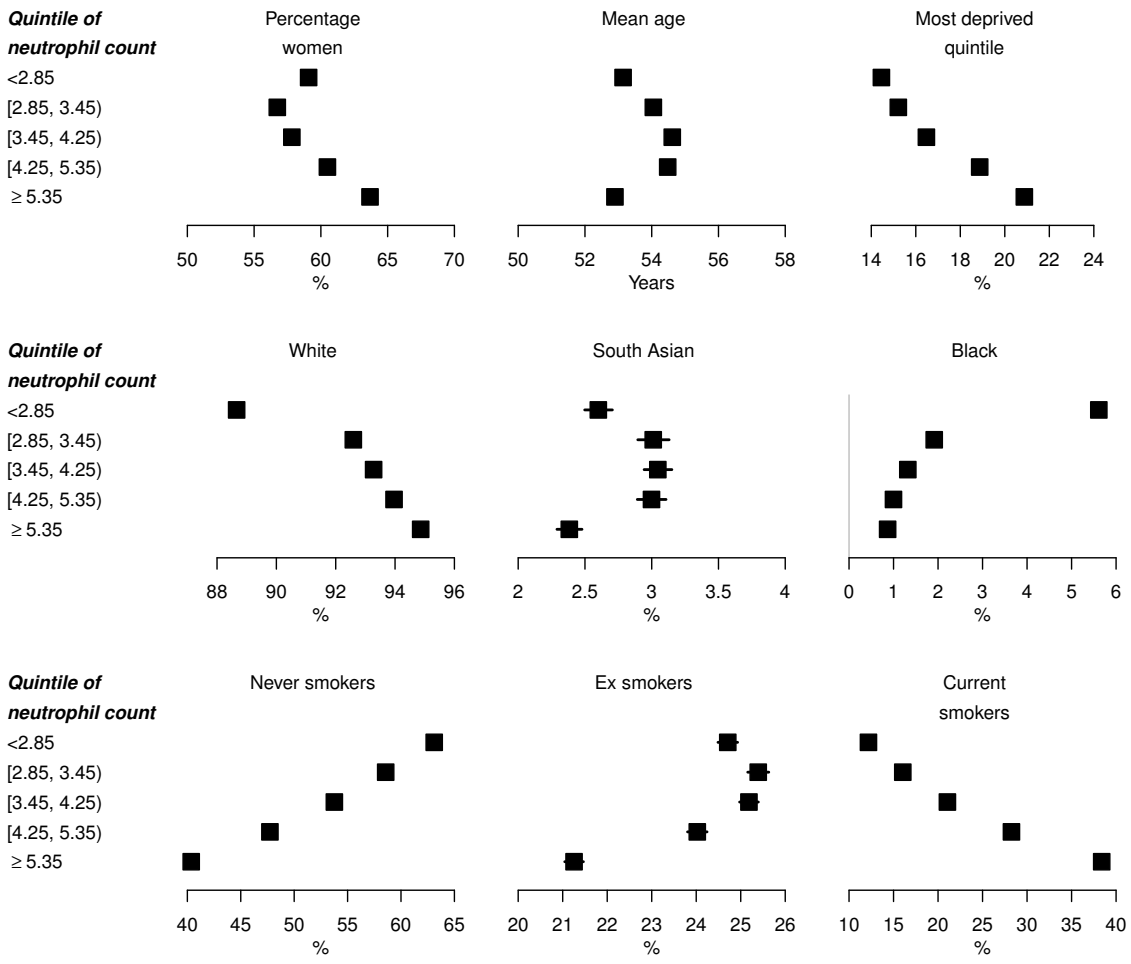
I log transformed skewed variables (HDL cholesterol, eGFR, total cholesterol and triglycerides) before imputation because both rfcont and normal-based multiple imputation functions assume a normal distribution of residuals. I exponentiated the imputed values and truncated the upper end of the distribution to the maximum value in the original data in order to remove extreme values (this affected only a small number of observations, but the values were very large and would affect any normal-based linear modelling). I split the dataset by gender and geographical region for multiple imputation, with general practice included as a categorical variable. I generated imputations in parallel on the CALIBER high performance computing cluster. I generated 20 imputations, each drawn from 20 iterations of MICE.

I reviewed plots of chain means and variances of imputed variables to check that chains were mixed adequately. I combined the results of Cox models using Rubin's rules [224]. I verified that the ratio of between-imputation variance to total variance ('fraction of missing information') for each parameter of interest was less than 20% in order to verify that 20 imputations were sufficient [187].

7.5 Results

The cohort included 621 052 patients with differential leukocyte counts recorded while clinically 'stable' and 154 179 patients with differential leukocyte counts performed during acute illness or treatment (**Figure 6.1 on page 135**). There were 54 980 initial presentations of cardiovascular disease over a median follow up of 3.8 years (interquartile range (IQR) 1.7 years to 6.0 years).

Figure 7.1: Selected characteristics of patients that vary significantly by neutrophil count
Bars represent 95% confidence intervals for the summary statistics



The overall standard deviation of neutrophil counts was 1.87 $\times$ 10^9 cells/L. Non-smokers, women and individuals with Black ethnicity tended to have lower neutrophil counts (**Figure 7.1 on page 159**). Patients with higher neutrophil count were more likely to smoke, live in a deprived area, and experience comorbidities such as asthma, COPD, diabetes, connective tissue diseases and inflammatory bowel disease (**Table 7.1 on page 160**). Higher neutrophil count was associated with higher counts of other leukocytes (**Figure 7.2 on page 164**) and higher platelet counts (**Table 7.1 on page 160**).

Individuals with higher total white cell counts tended to have higher counts of neutrophils, lymphocytes and platelets. Total white cell counts had a similar association with smoking and ethnicity as neutrophil counts (**Table 7.2 on page 161**).

Neutrophil counts outside the normal range were more likely to be associated with an acute condition at the time of blood testing (22 678 / 82 376, 27.5%) than those within the normal range (131 501 / 692 855, 19.0%) (P <0.0001). Symptoms or diagnosis of infection were the most frequent reason for the patient's condition to be classified as 'acute' (**Table 7.3 on page 162**).

Table 7.1: Characteristics of patients by quintiles of neutrophil count

Quintile	**1 (lowest)**	**2**	**3**	**4**	**5 (highest)**
Neutrophils, $\times 10^9$/L	**< 2.85**	**[2.85, 3.45)**	**[3.45, 4.25)**	**[4.25, 5.35)**	**≥ 5.35**
N patients	155 194	137 526	172 488	154 522	155 501
Women, n (%)	91 693 (59.1%)	78 026 (56.7%)	99 744 (57.8%)	93 470 (60.5%)	99 023 (63.7%)
Age, median (IQR)	52.7 (41.9-62.8)	53.2 (42.5-64.1)	53.4 (42.5-65.3)	52.5 (41.6-65.9)	49.8 (38.6-65.2)
Most deprived quintile, n (%)	25 060 (16.2%)	23 266 (17.0%)	31 830 (18.5%)	32 523 (21.1%)	36 004 (23.2%)
Ethnicity, n (%):					
White	81 153 (88.7%)	76 193 (92.6%)	98 708 (93.3%)	91 638 (94.0%)	97 342 (94.9%)
South Asian	2380 (2.6%)	2478 (3.0%)	3223 (3.0%)	2924 (3.0%)	2445 (2.4%)
Black	5137 (5.6%)	1575 (1.9%)	1398 (1.3%)	974 (1.0%)	890 (0.9%)
Other	2862 (3.1%)	2042 (2.5%)	2493 (2.4%)	1987 (2.0%)	1936 (1.9%)
Missing	63 662 (41.0%)	55 238 (40.2%)	66 666 (38.6%)	56 999 (36.9%)	52 888 (34.0%)
Full blood count parameters on index date, median (IQR):					
Eosinophils, $\times 10^9$/L	0.12 (0.1-0.2)	0.17 (0.1-0.22)	0.19 (0.1-0.26)	0.2 (0.1-0.30)	0.2 (0.1-0.3)
Lymphocytes, $\times 10^9$/L	1.8 (1.48-2.19)	1.9 (1.6-2.3)	2 (1.6-2.4)	2.1 (1.7-2.58)	2.1 (1.62-2.66)
Monocytes, $\times 10^9$/L	0.4 (0.3-0.5)	0.41 (0.35-0.5)	0.5 (0.4-0.6)	0.5 (0.4-0.63)	0.6 (0.5-0.8)
Basophils, $\times 10^9$/L	0.01 (0-0.04)	0.02 (0-0.06)	0.03 (0-0.07)	0.03 (0-0.09)	0.03 (0-0.1)
Haemoglobin, g/dL	13.8 (12.9-14.7)	14 (13.1-14.9)	14 (13.1-15)	14 (13-15)	13.8 (12.7-14.9)
Platelets, $\times 10^9$/L	237 (203-276)	250 (214-290)	259 (222-301)	271 (231-316)	288 (243-342)
Smoking status, n (%):					
Never	93 085 (63.1%)	76 531 (58.6%)	88 060 (53.8%)	69 871 (47.7%)	59 155 (40.4%)
Ex	36 445 (24.7%)	33 182 (25.4%)	41 251 (25.2%)	35 168 (24.0%)	31 153 (21.3%)
Current	17 970 (12.2%)	20 959 (16.0%)	34 483 (21.1%)	41 350 (28.2%)	56 259 (38.4%)
Missing	7694 (5.0%)	6854 (5.0%)	8694 (5.0%)	8133 (5.3%)	8934 (5.7%)
Most recent value within one year prior to index date, median (IQR):					
Systolic blood pressure, mmHg	134 (120-148)	137 (123-150)	138 (124-150)	138 (124-150)	134 (120-149)
Body mass index, kg/m^2	26.1 (23.2-29.6)	27 (23.9-30.7)	27.4 (24.2-31.3)	27.6 (24.1-32)	27 (23.4-31.9)
Total cholesterol, mmol/L	5.5 (4.8-6.3)	5.5 (4.8-6.3)	5.5 (4.8-6.3)	5.4 (4.7-6.2)	5.3 (4.6-6.1)
HDL cholesterol, mmol/L	1.5 (1.2-1.8)	1.4 (1.18-1.7)	1.36 (1.11-1.62)	1.3 (1.1-1.6)	1.3 (1.1-1.6)
eGFR, mL/min/1.73m^2	83.2 (71-95.7)	81.9 (69.5-94.4)	81.2 (68.3-94.1)	81.2 (67.8-94.7)	82.4 (67.4-96.6)
Diagnoses on or before index date, n (%):					
Atrial fibrillation	1038 (0.7%)	1157 (0.8%)	1746 (1.0%)	1798 (1.2%)	1983 (1.3%)
Cancer	9881 (6.4%)	7998 (5.8%)	10598 (6.1%)	9379 (6.1%)	9665 (6.2%)
Diabetes	5167 (3.3%)	5669 (4.1%)	8541 (5.0%)	8816 (5.7%)	8934 (5.7%)
Asthma	16 894 (10.9%)	15 902 (11.6%)	21 039 (12.2%)	20 319 (13.1%)	22 643 (14.6%)
COPD	1224 (0.8%)	1632 (1.2%)	2943 (1.7%)	3663 (2.4%)	5559 (3.6%)
Connective tissue disease	3081 (2.0%)	3044 (2.2%)	4501 (2.6%)	4762 (3.1%)	6471 (4.2%)
IBD	1274 (0.8%)	1350 (1.0%)	1798 (1.0%)	1819 (1.2%)	2405 (1.5%)
Medication use in the year before index date, n (%):					
Antihypertensives	33 405 (21.5%)	33 244 (24.2%)	46 025 (26.7%)	43 297 (28.0%)	41 310 (26.6%)
Statins	7801 (5.0%)	8264 (6.0%)	11 242 (6.5%)	10 358 (6.7%)	9141 (5.9%)

COPD, chronic obstructive pulmonary disease; eGFR, estimated glomerular filtration rate; HDL, high density lipoprotein cholesterol, IBD, inflammatory bowel disease; IQR, interquartile range.

Table 7.2: Characteristics of patients by quintiles of total white cell count

Total white cell count, $\times 10^9$/L	**1 (lowest) <5.35**	**2 [5.35, 6.25)**	**3 [6.25, 7.25)**	**4 [7.25, 8.65)**	**5 (highest) $\geq$8.65**
N patients	163 779	151 462	160 754	153 454	145 782
Women, n (%)	100 067 (61.1%)	87 602 (57.8%)	93 879 (58.4%)	91 450 (59.6%)	88 958 (61.0%)
Age, median (IQR)	52.8 (42.2-63.4)	53.3 (42.4-64.5)	53.2 (42.1-65.4)	52.4 (41.2-65.5)	50.1 (39-64.1)
Most deprived quintile, n (%)	23 135 (14.2%)	22 287 (14.8%)	26 385 (16.5%)	28 937 (18.9%)	32 227 (22.2%)
Acute condition at time of blood test, n (%)	30 772 (18.8%)	26 529 (17.5%)	29 495 (18.3%)	30 605 (19.9%)	36 778 (25.2%)
Ethnicity, n (%):					
White	88 603 (91.0%)	84 086 (92.6%)	92 034 (92.9%)	90 374 (93.2%)	89 937 (94.1%)
South Asian	1891 (1.9%)	2416 (2.7%)	3014 (3.0%)	3241 (3.3%)	2888 (3.0%)
Black	4228 (4.3%)	2045 (2.3%)	1621 (1.6%)	1151 (1.2%)	929 (1.0%)
Other	2635 (2.7%)	2284 (2.5%)	2361 (2.4%)	2174 (2.2%)	1866 (2.0%)
Missing	66 422 (40.6%)	60 631 (40.0%)	61 724 (38.4%)	56 514 (36.8%)	50 162 (34.4%)
Full blood count parameters on index date, median (IQR):					
Neutrophils, $\times 10^9$/L	2.5 (2.17-2.9)	3.3 (2.94-3.6)	3.9 (3.5-4.3)	4.76 (4.28-5.27)	6.5 (5.62-7.64)
Lymphocytes, $\times 10^9$/L	1.58 (1.3-1.81)	1.85 (1.58-2.14)	2.04 (1.7-2.4)	2.25 (1.83-2.7)	2.5 (1.91-3.1)
Eosinophils, $\times 10^9$/L	0.1 (0.1-0.2)	0.15 (0.1-0.2)	0.19 (0.1-0.26)	0.2 (0.1-0.3)	0.2 (0.1-0.3)
Monocytes, $\times 10^9$/L	0.38 (0.3-0.44)	0.4 (0.35-0.5)	0.5 (0.4-0.6)	0.53 (0.42-0.67)	0.65 (0.5-0.8)
Basophils, $\times 10^9$/L	0.01 (0-0.04)	0.02 (0-0.06)	0.03 (0-0.07)	0.03 (0-0.1)	0.04 (0-0.1)
Haemoglobin, g/dL	13.7 (12.8-14.7)	13.9 (13-14.9)	14 (13.1-15)	14 (13.1-15)	14 (12.9-15)
Platelets, $\times 10^9$/L	235 (201-274)	250 (215-289)	261 (224-303)	272 (233-317)	291 (247-345)
Smoking status, n (%):					
Never	100 561 (64.7%)	85 335 (59.4%)	82 201 (53.9%)	68 264 (46.9%)	50 341 (36.5%)
Ex	37 796 (24.3%)	36 733 (25.6%)	38 659 (25.4%)	35 083 (24.1%)	28 928 (21.0%)
Current	16 985 (10.9%)	21 647 (15.1%)	31 633 (20.7%)	42 121 (29.0%)	58 635 (42.5%)
Missing	8437 (5.2%)	7747 (5.1%)	8261 (5.1%)	7986 (5.2%)	7878 (5.4%)

IQR, interquartile range.

Cumulative incidence curves showed that there was a particularly high incidence of non-fatal myocardial infarction, unheralded coronary death, heart failure, peripheral arterial disease and abdominal aortic aneurysm among individuals in the top quintile of white cell count compared to those in the bottom quintile (**Figure 7.3 on page 165**). Plotting these curves for high and low neutrophil counts within the normal range revealed a similar pattern of associations with different cardiovascular presentations (**Figure 7.4 on page 166**). The risk appeared to be greatest for the first few months, but differences persisted for at least 8 years (**Figure 7.4 on page 166**).

The distribution of initial cardiovascular presentations varied by neutrophil count; individuals with higher neutrophil counts were more likely to present with unheralded coronary death, heart failure or peripheral arterial disease, and were relatively less likely to present with stable angina (**Table 7.4 on page 163**).

Table 7.3: Prevalence of acute conditions by category of neutrophil count within and outside the normal range

	Below normal range	**Within normal range**			**Above normal range**
Neutrophils, $\times 10^9$/L	**< 2**	**[2, 3)**	**[3, 5)**	**[5, 7)**	**≥ 7**
N patients	26 588	154 863	395 466	142 526	55 788
Women, n (%)	16 592 (62.4%)	89 956 (58.1%)	229 869 (58.1%)	89 817 (63.0%)	35 722 (64.0%)
Age, median (IQR)	51.3 (40.6–61.8)	53 (42.2–63.1)	53.2 (42.3–65.1)	51 (40.1–65.6)	48.2 (36.9–64.7)
Any acute condition at time of blood test, n (%)	5670 (21.3%)	26 932 (17.4%)	72 676 (18.4%)	31 893 (22.4%)	17 008 (30.5%)
In hospital on date of blood test	149 (0.6%)	263 (0.2%)	893 (0.2%)	732 (0.5%)	1043 (1.9%)
Vaccination within previous 7 days	285 (1.1%)	1575 (1.0%)	4383 (1.1%)	1463 (1.0%)	465 (0.8%)
Anaemia diagnosis within 30 days before	243 (0.9%)	752 (0.5%)	1659 (0.4%)	710 (0.5%)	393 (0.7%)
Infection diagnosis within 30 days before	1846 (6.9%)	9446 (6.1%)	25 280 (6.4%)	11 266 (7.9%)	6413 (11.5%)
Infective symptoms within 30 days before	1470 (5.5%)	7242 (4.7%)	20237 (5.1%)	9242 (6.5%)	4987 (8.9%)
Prior diagnosis of myelodysplastic syndrome	74 (0.3%)	80 (0.1%)	98 (0.0%)	38 (0.0%)	16 (0.0%)
Prior diagnosis of haemoglobinopathy	268 (1.0%)	668 (0.4%)	1161 (0.3%)	389 (0.3%)	163 (0.3%)
Chemotherapy or G-CSF within 6 months prior	364 (1.4%)	391 (0.3%)	692 (0.2%)	292 (0.2%)	317 (0.6%)
Methotrexate within 3 months prior	39 (0.1%)	346 (0.2%)	1423 (0.4%)	856 (0.6%)	413 (0.7%)
Steroid within 3 months prior	383 (1.4%)	2108 (1.4%)	8204 (2.1%)	5741 (4.0%)	4740 (8.5%)
Other immune drug within 3 months prior	376 (1.4%)	1611 (1.0%)	3829 (1.0%)	1725 (1.2%)	852 (1.5%)
Prior diagnosis of cancer	2089 (7.9%)	9343 (6.0%)	23 910 (6.0%)	8421 (5.9%)	3758 (6.7%)

Table 7.4: Endpoints by category of neutrophil count within and outside the normal range

	Below normal range	**Within normal range**			**Above normal range**
Neutrophils, $\times 10^9$/L	**< 2**	**[2, 3)**	**[3, 5)**	**[5, 7)**	**≥ 7**
Initial presentation of cardiovascular disease					
Stable angina	229 (20.7%)	1763 (22.0%)	5174 (18.2%)	1840 (14.5%)	542 (11.3%)
Unstable angina	70 (6.3%)	526 (6.6%)	1523 (5.4%)	574 (4.5%)	187 (3.9%)
Coronary disease not further specified	135 (12.2%)	943 (11.8%)	2648 (9.3%)	904 (7.1%)	293 (6.1%)
Non-fatal myocardial infarction	100 (9.0%)	874 (10.9%)	3465 (12.2%)	1514 (11.9%)	613 (12.8%)
Unheralded coronary death	38 (3.4%)	338 (4.2%)	1499 (5.3%)	823 (6.5%)	392 (8.2%)
Heart failure	105 (9.5%)	646 (8.1%)	3042 (10.7%)	1733 (13.7%)	698 (14.6%)
Ventricular arrhythmia or sudden cardiac death	20 (1.8%)	109 (1.4%)	373 (1.3%)	168 (1.3%)	77 (1.6%)
Transient ischaemic attack	109 (9.9%)	794 (9.9%)	2702 (9.5%)	1094 (8.6%)	377 (7.9%)
Ischaemic stroke	80 (7.2%)	532 (6.7%)	1907 (6.7%)	873 (6.9%)	298 (6.2%)
Stroke not further specified	87 (7.9%)	519 (6.5%)	2143 (7.5%)	1035 (8.2%)	407 (8.5%)
Subarachnoid haemorrhage	22 (2.0%)	107 (1.3%)	269 (0.9%)	117 (0.9%)	62 (1.3%)
Intracerebral haemorrhage	29 (2.6%)	176 (2.2%)	548 (1.9%)	224 (1.8%)	83 (1.7%)
Peripheral arterial disease	66 (6.0%)	528 (6.6%)	2468 (8.7%)	1384 (10.9%)	620 (13.0%)
Abdominal aortic aneurysm	16 (1.4%)	143 (1.8%)	683 (2.4%)	388 (3.1%)	136 (2.8%)
Total	1106	7998	28444	12671	4785
Other deaths					
Cancers	675 (63.4%)	2107 (55.2%)	7443 (52.4%)	4308 (52.8%)	3060 (57.4%)
Dementia	22 (2.1%)	157 (4.1%)	825 (5.8%)	414 (5.1%)	187 (3.5%)
Pneumonia	41 (3.8%)	201 (5.3%)	931 (6.6%)	560 (6.9%)	331 (6.2%)
Chronic obstructive pulmonary disease	11 (1.0%)	78 (2.0%)	470 (3.3%)	444 (5.4%)	323 (6.1%)
Liver disease	65 (6.1%)	192 (5.0%)	354 (2.5%)	144 (1.8%)	114 (2.1%)
Other causes of death	251 (23.6%)	1084 (28.4%)	4184 (29.5%)	2295 (28.1%)	1320 (24.7%)
Total	1065	3819	14 207	8165	5335

Figure 7.2: Binned scatterplots showing correlations between counts of neutrophils and other leukocyte subtypes

The lines are based on linear regression models for neutrophil count on the count of another leukocyte subtype.

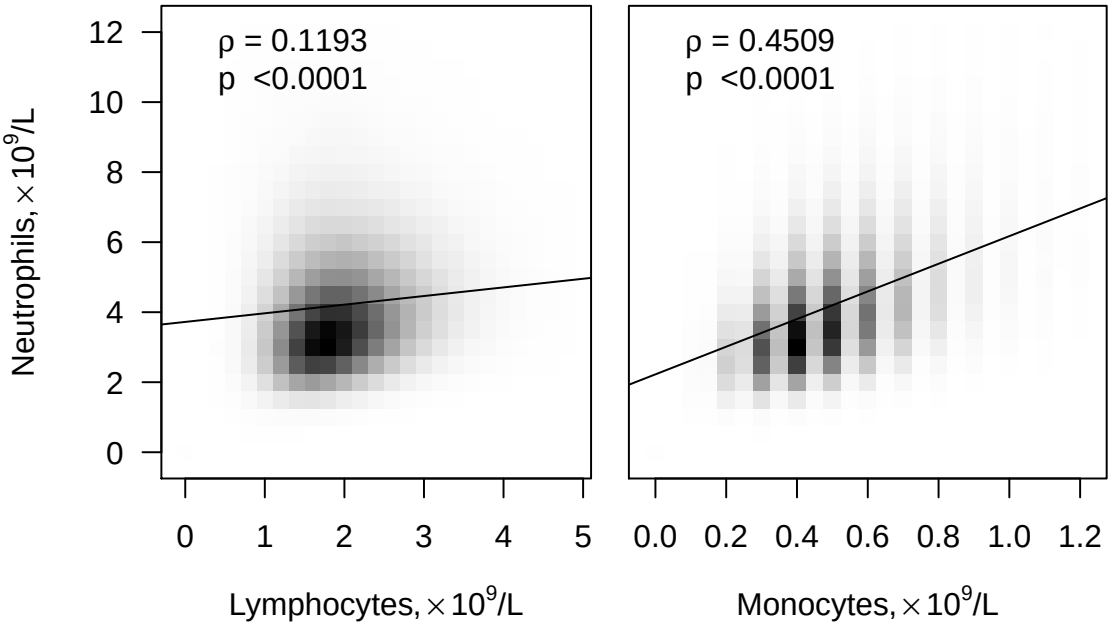


Figure 7.3: Cumulative incidence curves for different initial presentations of cardiovascular disease by total white cell count

Curves are shown for the highest and lowest quintiles of total white cell count (WBC).

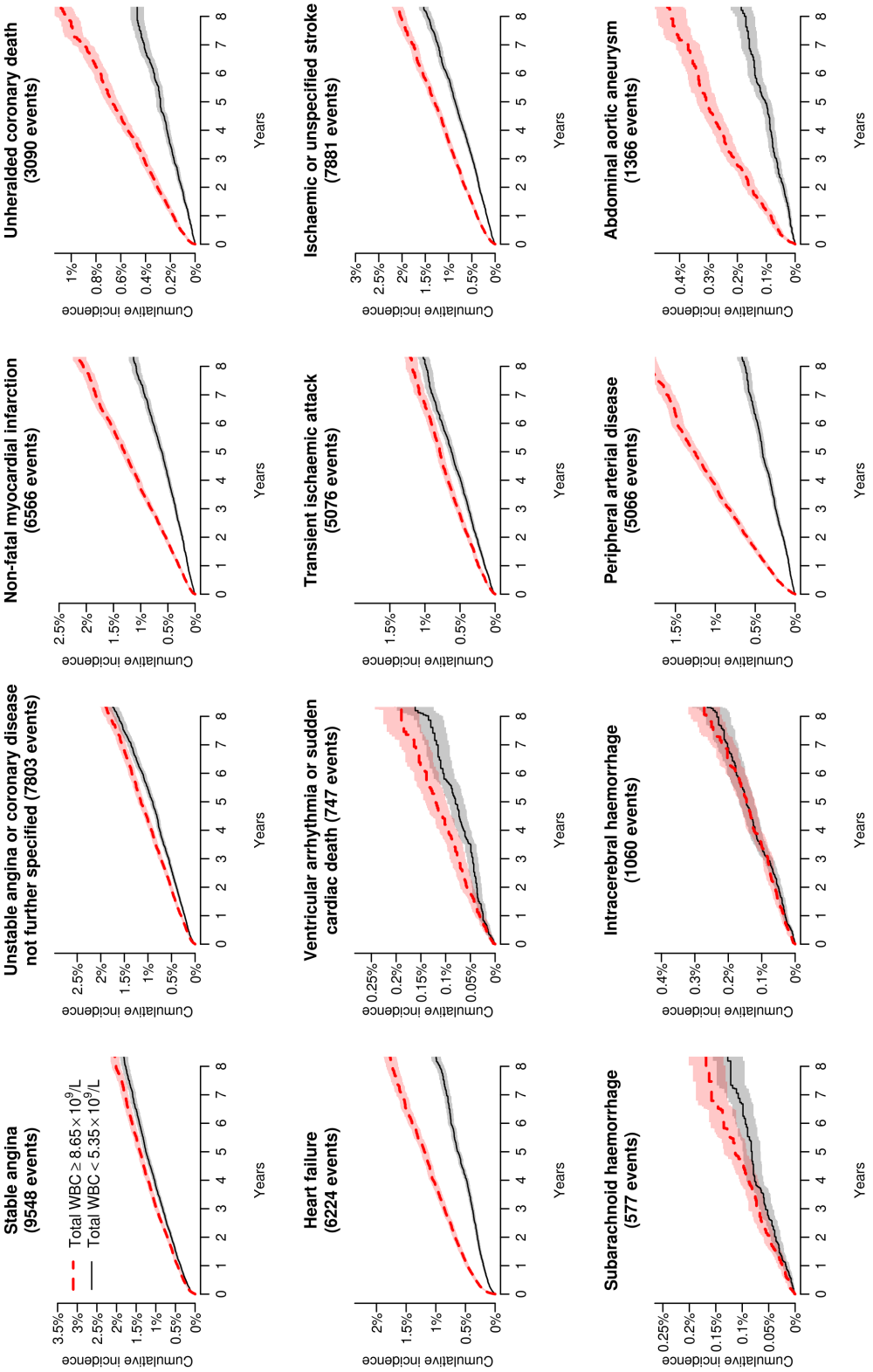


Figure 7.4: Cumulative incidence curves for different initial presentations of cardiovascular disease by level of neutrophil count within the normal range

Curves are shown for the highest and lowest categories of neutrophil count within the normal range.

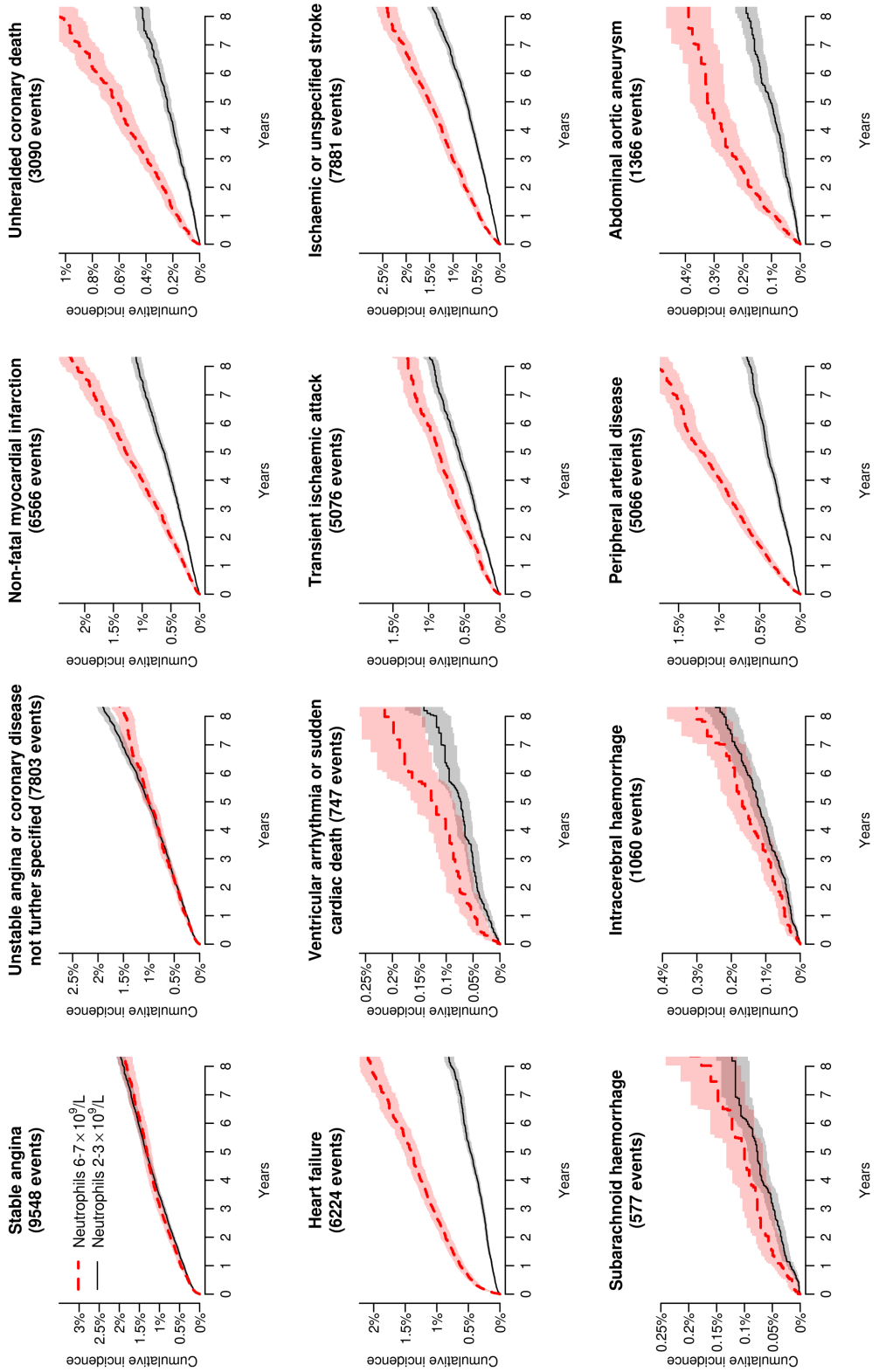


Figure 7.5: Hazard ratios for association of neutrophil quintiles with different initial presentations of cardiac disease, by level of adjustment

'Multiple adjustment' includes adjustment for age, sex, deprivation, ethnicity, smoking, diabetes, body mass index, systolic blood pressure, eGFR, HDL, total cholesterol, atrial fibrillation, inflammatory conditions, cancer, statin use, blood pressure medication and acute conditions at the time of blood testing. P values * < 0.05, ** < 0.01, *** < 0.001

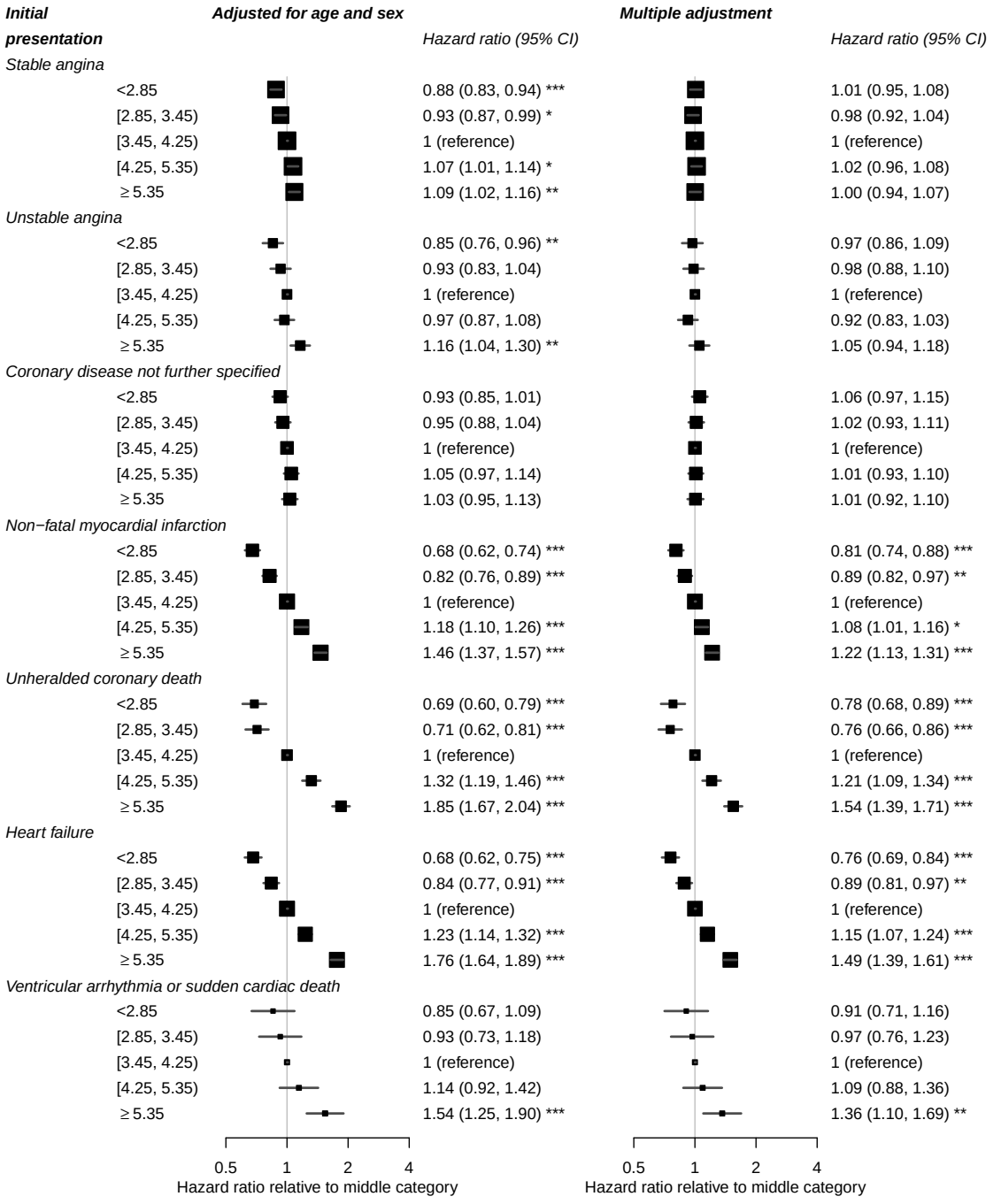


Figure 7.6: Adjusted hazard ratios for association of neutrophil quintiles with different initial presentations of non-cardiac cardiovascular endpoints, by level of adjustment

'Multiple adjustment' includes adjustment for age, sex, deprivation, ethnicity, smoking, diabetes, body mass index, systolic blood pressure, eGFR, HDL, total cholesterol, atrial fibrillation, inflammatory conditions, cancer, statin use, blood pressure medication and acute conditions at the time of blood testing. P values * < 0.05, ** < 0.01, *** < 0.001

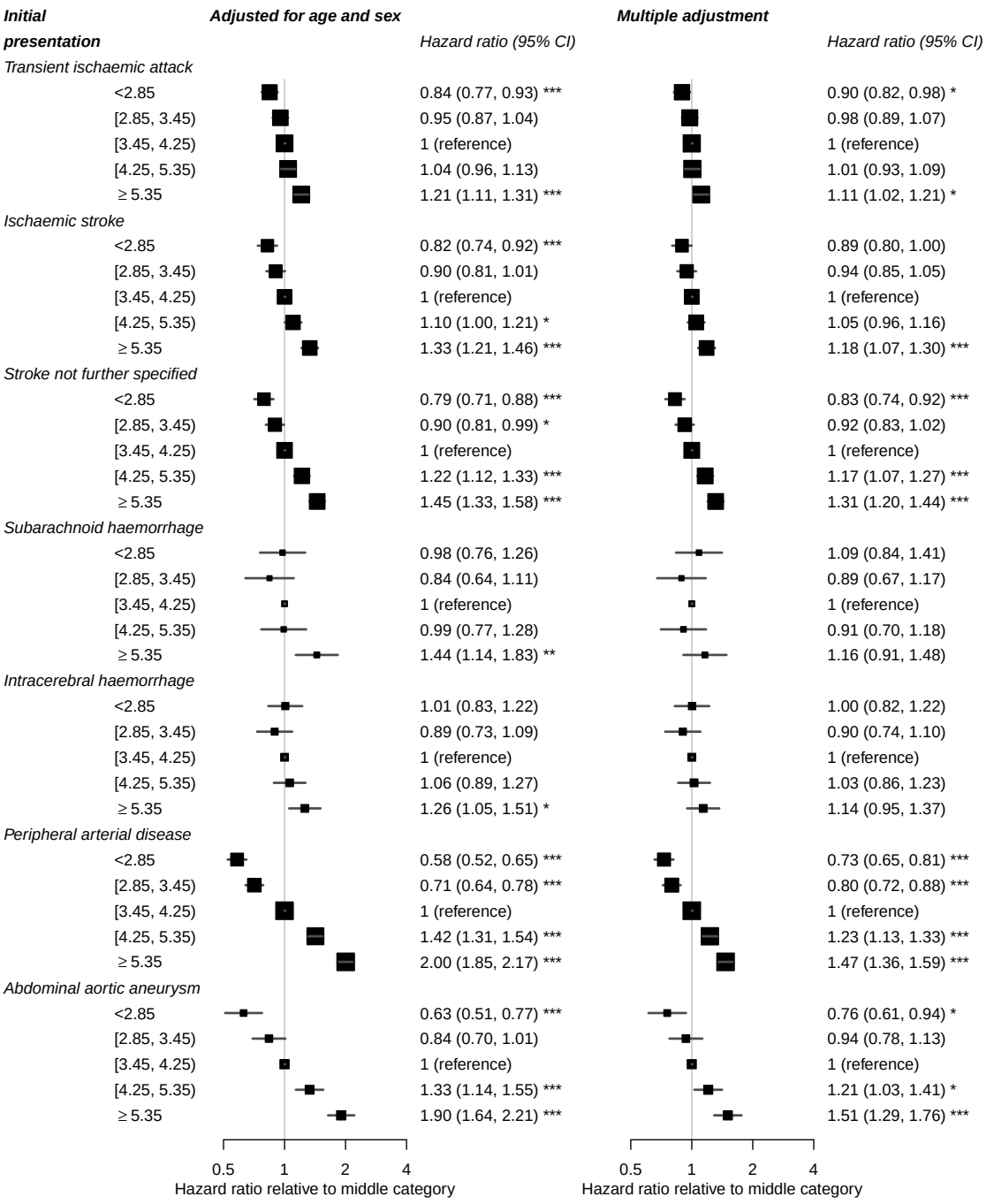
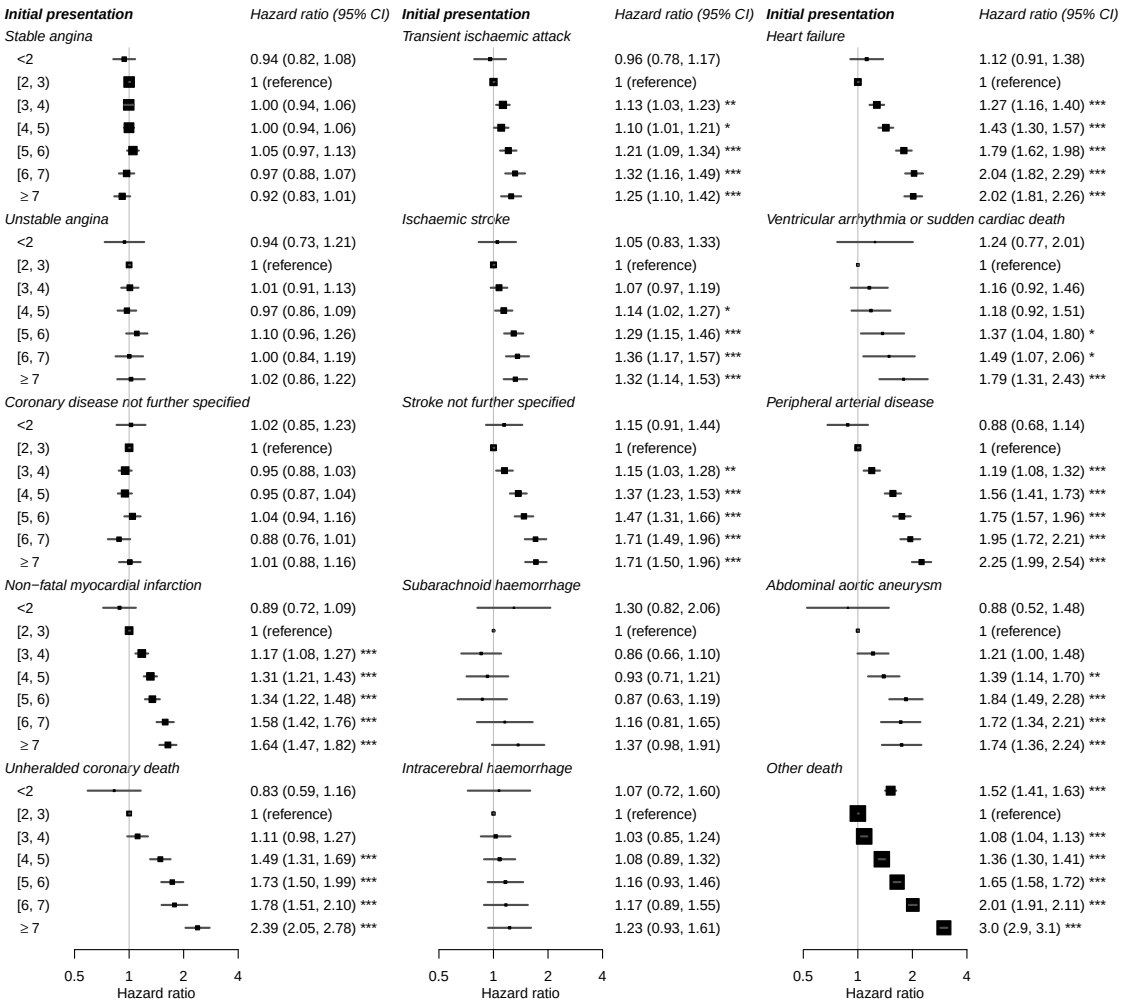


Figure 7.7: Association of neutrophil categories with different initial presentations of cardiovascular diseases

Hazard ratios are adjusted for age, sex, deprivation, ethnicity, smoking, diabetes, systolic blood pressure, blood pressure medication, body mass index, total cholesterol, HDL cholesterol, statin use, eGFR, atrial fibrillation, autoimmune conditions, inflammatory bowel disease, COPD, cancer, and acute conditions at the time of blood testing. P values * < 0.05, ** < 0.01, *** < 0.001



Age-adjusted hazard ratios from cause-specific Cox models showed heterogeneous associations between neutrophil counts and different cardiac (**Figure 7.5 on page 167**) and non-cardiac (**Figure 7.6 on page 168**) initial presentations of cardiovascular disease. The directions of associations were the same between age-sex adjusted and multiply adjusted models, but the associations were generally stronger in age-sex adjusted models.

Using separate categories for neutrophil counts outside the reference range demonstrated that the linear association was strongest among neutrophil counts within the normal range (**Figure 7.7 on page 169**). Adjusted hazard ratios (HR) comparing neutrophil counts $6-7$ vs $2-3\times10^9$/L (within the normal range) showed strong associations with heart failure (HR 2.04, 95% confidence interval (CI) 1.82–2.29), non-fatal myocardial infarction (HR 1.58, 95% CI 1.42–1.76) and unheralded coronary death (HR 1.78, 95% CI 1.51–2.10), but not stable angina (HR 0.97, 95% CI 0.88–1.07) or unstable angina (HR 1.00, 95% CI 0.84–1.19). The association with ischaemic stroke was weaker (HR 1.36, 95% CI 1.17–1.57) and there was no association with haemorrhagic stroke, but there were strong associations with peripheral arterial disease (HR 1.95, 95% CI 1.72–2.21) and abdominal aortic aneurysm (HR 1.72, 95% CI 1.34–2.21) (**Figure 7.7 on page 169**). There was a J-shaped association between neutrophil counts and non-cardiovascular death (**Figure 7.7 on page 169**).

As the associations with cardiovascular endpoints were monotonic and approximately linear, I treated neutrophil count as a linear variable in subsequent modelling. Associations were stronger among the subgroup of patients with neutrophil counts within the normal range (**Figure 7.8 on page 172**).

The strength of some associations were stronger for neutrophil counts taken when the patient was clinically stable, particularly for peripheral arterial disease and heart failure (**Figure 7.9 on page 173**). Some associations were further enhanced by using the mean of two consecutive stable neutrophil counts, which were available in 393 543 patients and were taken median 1.4 years apart (IQR 0.6, 2.7 years) (**Figure 7.9 on page 173**). The within-patient correlation between these two values was only moderate (ρ = 0.568), and the standard deviation of the differences was 1.67×10^9/L, with no variation over time (**Figure 7.10 on page 174**).

Plots of scaled Schoenfeld residuals showed evidence of non-proportional hazards, with stronger associations in the first 6 months for many endpoints (**Figure 7.11 on page 175**). I therefore split the follow-up time by 6 months, and found an inverse association of neutrophil count and stable angina as the initial presentation of cardiovascular disease in the first 6 months but a stronger positive association with heart failure, unheralded coronary death and ischaemic stroke thereafter (**Figure 7.12 on page 176**).

Associations with coronary endpoints, heart failure and peripheral arterial disease were stronger among younger patients (**Figure 7.13 on page 177**). The association between neutrophil count and initial presentation with heart failure was stronger in men (HR per 10^9 higher neutrophil count 1.10 vs 1.07, p = 0.001). There was an association between neutrophil count and initial presentation with transient ischaemic attack in women but not

men (HR 1.05 vs 1.00, p = 0.007) (**Figure 7.14 on page 178**), but no other significant interactions with sex, and no interaction by smoking status (**Figure 7.15 on page 179**).

Other leukocyte subtypes exhibited different associations with incidence of cardiovascular diseases. The associations with monocyte count were similar to those with neutrophils but less strong **Figure 7.16 on page 180**. Low lymphocyte counts and high basophil counts were associated with increased incidence of heart failure, unheralded coronary death and other death (**Figure 7.17 on page 181**, **Figure 7.18 on page 182**). Detailed results for eosinophil counts are reported in chapter 8.

The two methods of multiple imputation, Random Forest and normal-based MICE, yielded almost identical estimates (**Figure 7.19 on page 183**).

Figure 7.8: Linear association of neutrophil count with different initial presentations of cardiovascular disease

Hazard ratios are adjusted for age, sex, deprivation, ethnicity, smoking, diabetes, systolic blood pressure, blood pressure medication, body mass index, total cholesterol, HDL cholesterol, statin use, eGFR, atrial fibrillation, autoimmune conditions, inflammatory bowel disease, COPD, cancer, and acute conditions at the time of blood testing. P values * < 0.05, ** < 0.01, *** < 0.001

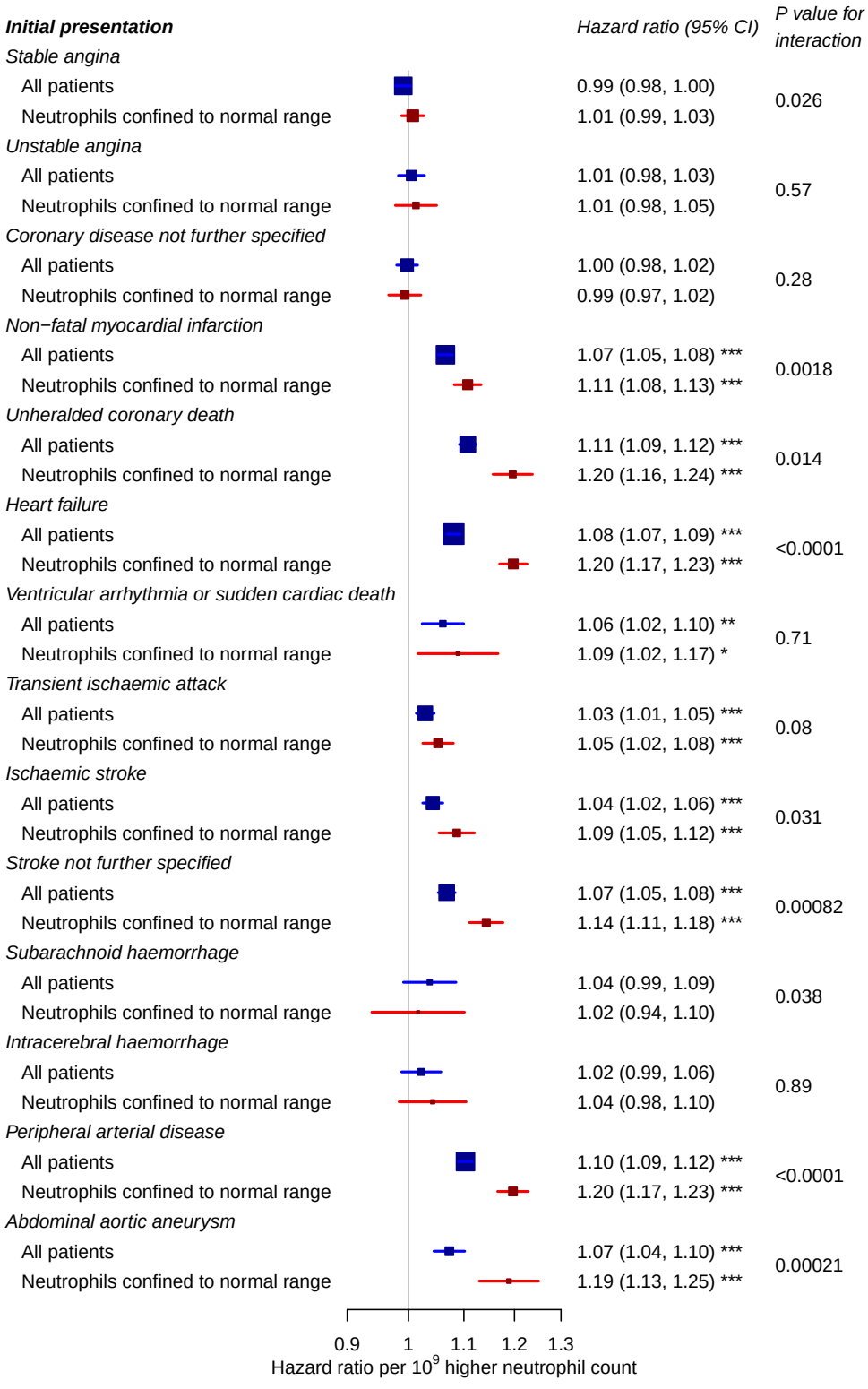
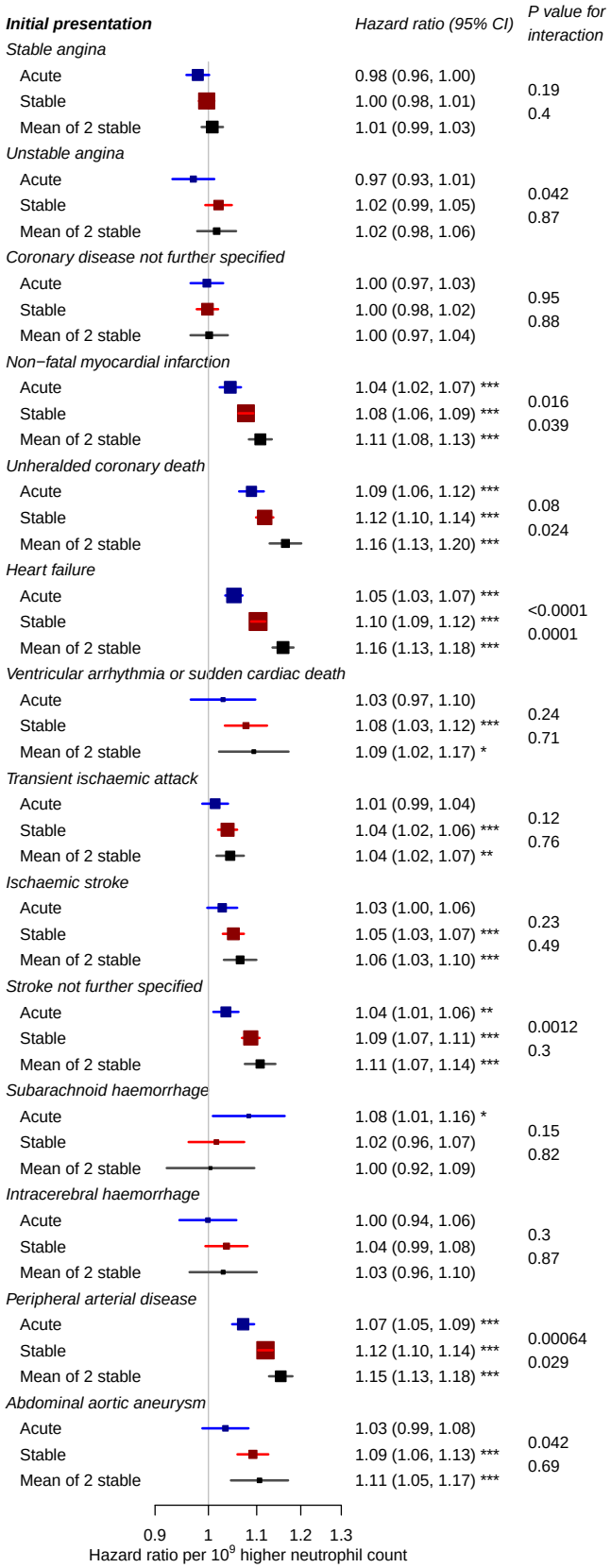


Figure 7.9: Linear association of neutrophil count with different initial presentations of cardiovascular disease, by patient's clinical state at the time of testing



'Mean of 2 stable' refers to the mean of two consecutive neutrophil counts performed in a stable clinical state. Hazard ratios are adjusted for age, sex, deprivation, ethnicity, smoking, diabetes, systolic blood pressure, blood pressure medication, body mass index, total cholesterol, HDL cholesterol, statin use, eGFR, atrial fibrillation, autoimmune conditions, inflammatory bowel disease, COPD and cancer. P values * < 0.05, ** < 0.01, *** < 0.001

Figure 7.10: Binned scatterplot showing difference between two consecutive neutrophil counts taken when a patient was clinically 'stable'

N = 374 460. Correlation coefficient = 0.568. The solid line is lowess smoothed mean and the dotted lines are lowess smoothed 2.5% and 97.5% centiles.

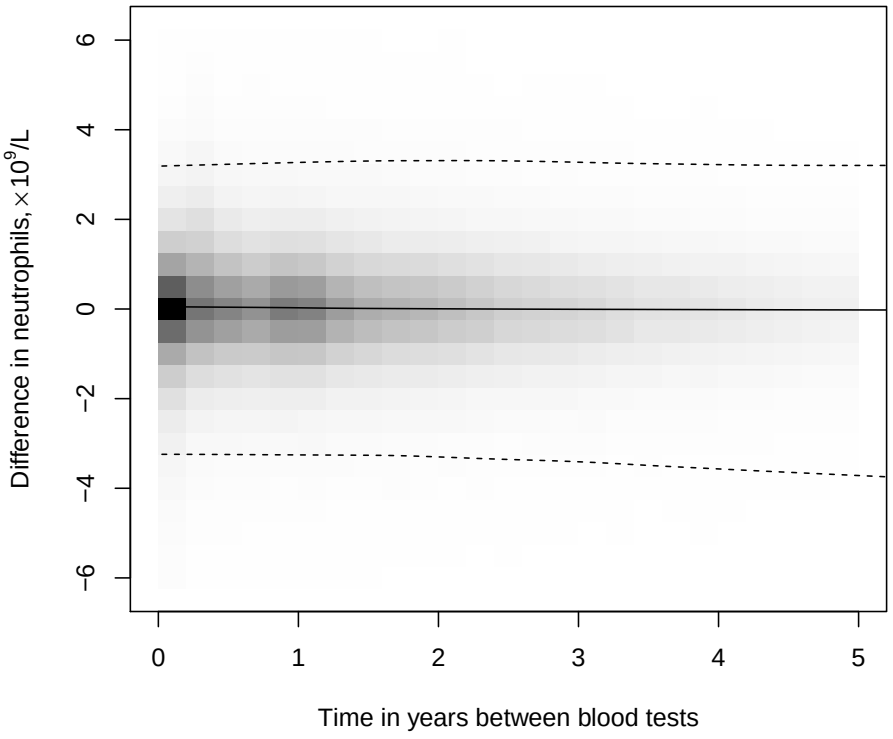


Figure 7.11: Scaled Schoenfeld residuals for adjusted hazard ratios comparing highest versus middle quintile of neutrophil count, for different initial presentations of cardiovascular disease

Hazard ratios adjusted for age, sex, deprivation, ethnicity, smoking, diabetes, body mass index, systolic blood pressure, eGFR, HDL, total cholesterol, atrial fibrillation, inflammatory conditions, cancer, statin use, blood pressure medication and acute conditions at the time of blood testing. The lines show the estimate and 95% confidence interval for the beta coefficient. The top left corner of each graph contains ρ, χ^2 and p value for the correlation between log survival time and scaled Schoenfeld residuals. For clarity, points plotted on the graph are for a random sample of 20 000 patients.

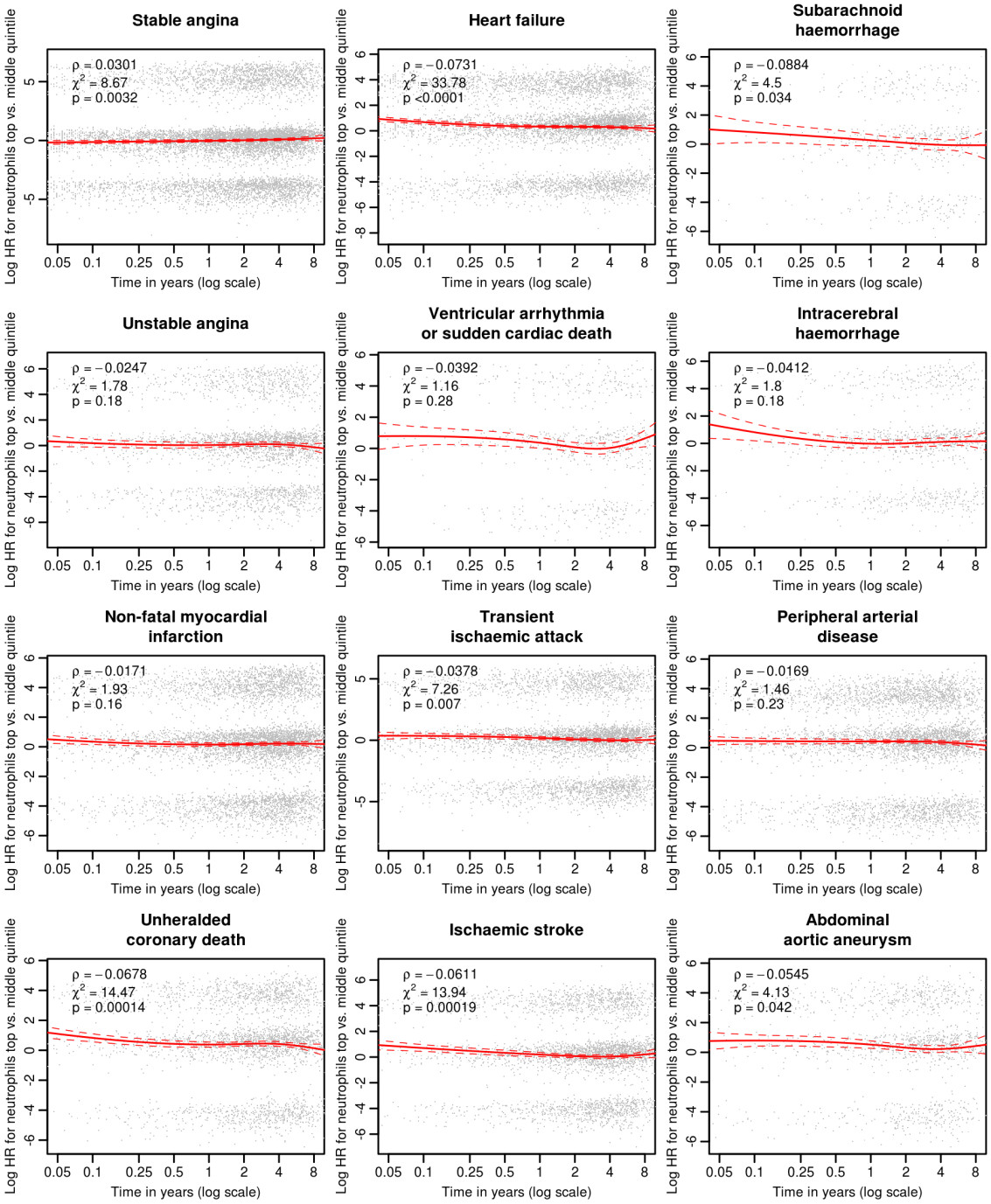
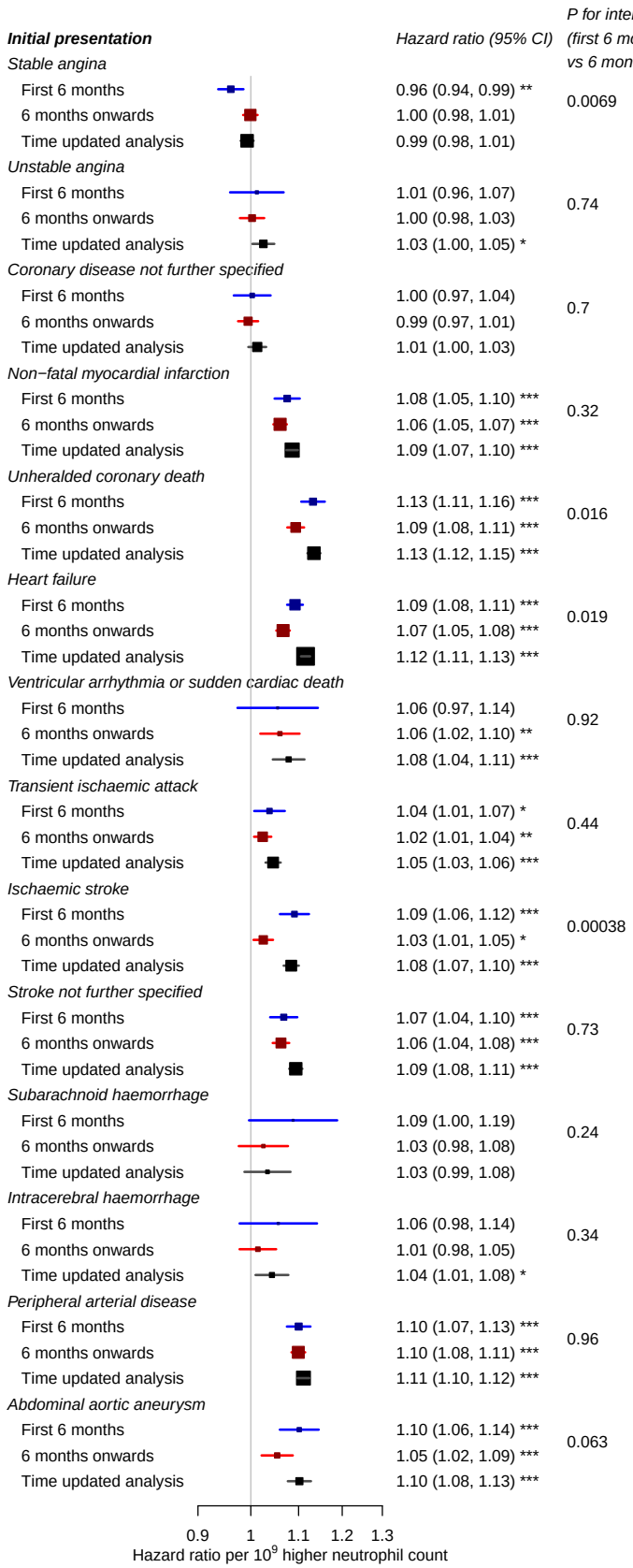


Figure 7.12: Linear association of neutrophil count with different initial presentations of cardiovascular disease, by time since measurement



Hazard ratios are adjusted for age, sex, deprivation, ethnicity, smoking, diabetes, systolic blood pressure, blood pressure medication, body mass index, total cholesterol, HDL cholesterol, statin use, eGFR, atrial fibrillation, autoimmune conditions, inflammatory bowel disease, COPD, cancer, and acute conditions at the time of blood testing. P values * < 0.05, ** < 0.01, *** < 0.001

Figure 7.13: Linear association of neutrophil count with different initial presentations of cardiovascular disease, by age group

Hazard ratios are adjusted for age, sex, deprivation, ethnicity, smoking, diabetes, systolic blood pressure, blood pressure medication, body mass index, total cholesterol, HDL cholesterol, statin use, eGFR, atrial fibrillation, autoimmune conditions, inflammatory bowel disease, COPD, cancer, and acute conditions at the time of blood testing. P values * < 0.05, ** < 0.01, *** < 0.001

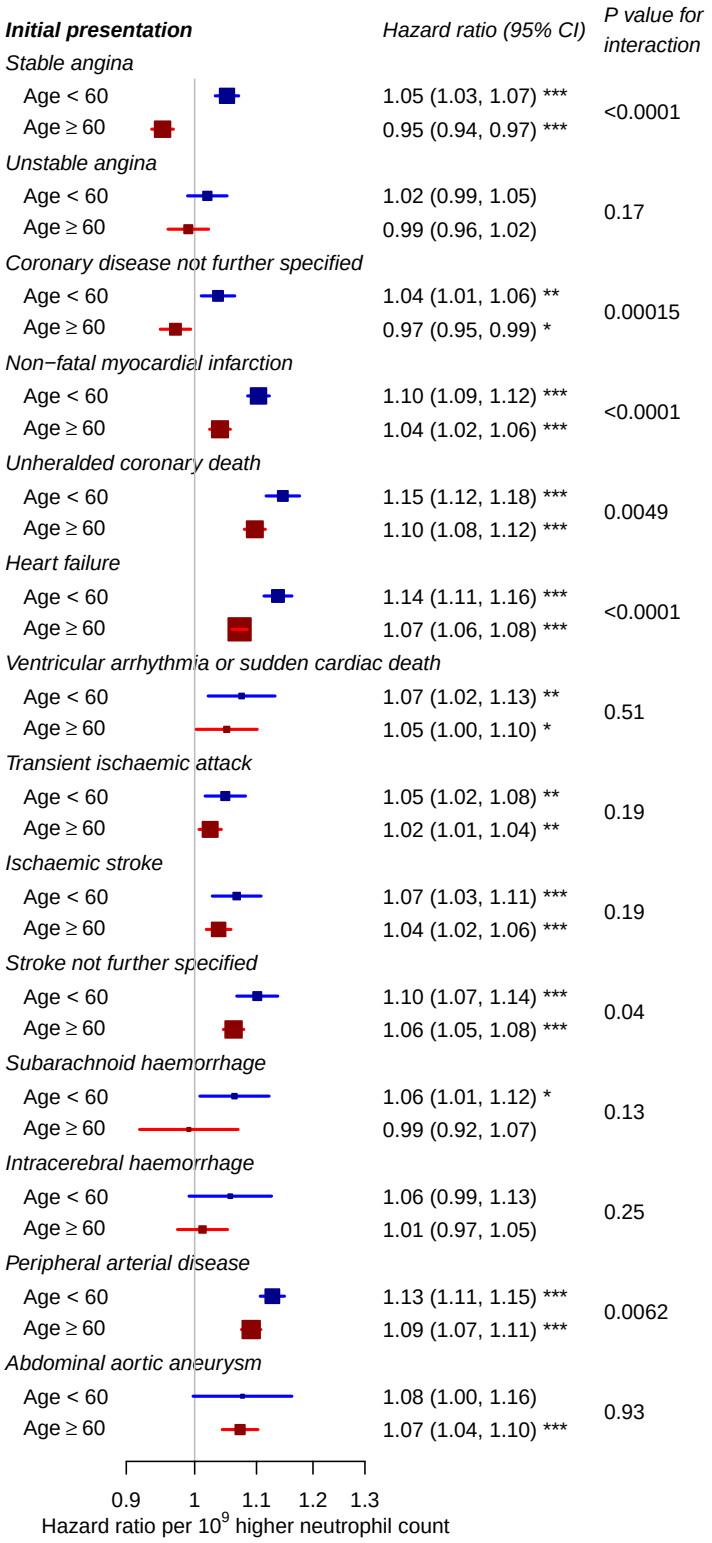


Figure 7.14: Linear association of neutrophil count with different initial presentations of cardiovascular disease, by sex

Hazard ratios are adjusted for age, sex, deprivation, ethnicity, smoking, diabetes, systolic blood pressure, blood pressure medication, body mass index, total cholesterol, HDL cholesterol, statin use, eGFR, atrial fibrillation, autoimmune conditions, inflammatory bowel disease, COPD, cancer, and acute conditions at the time of blood testing. P values * < 0.05, ** < 0.01, *** < 0.001

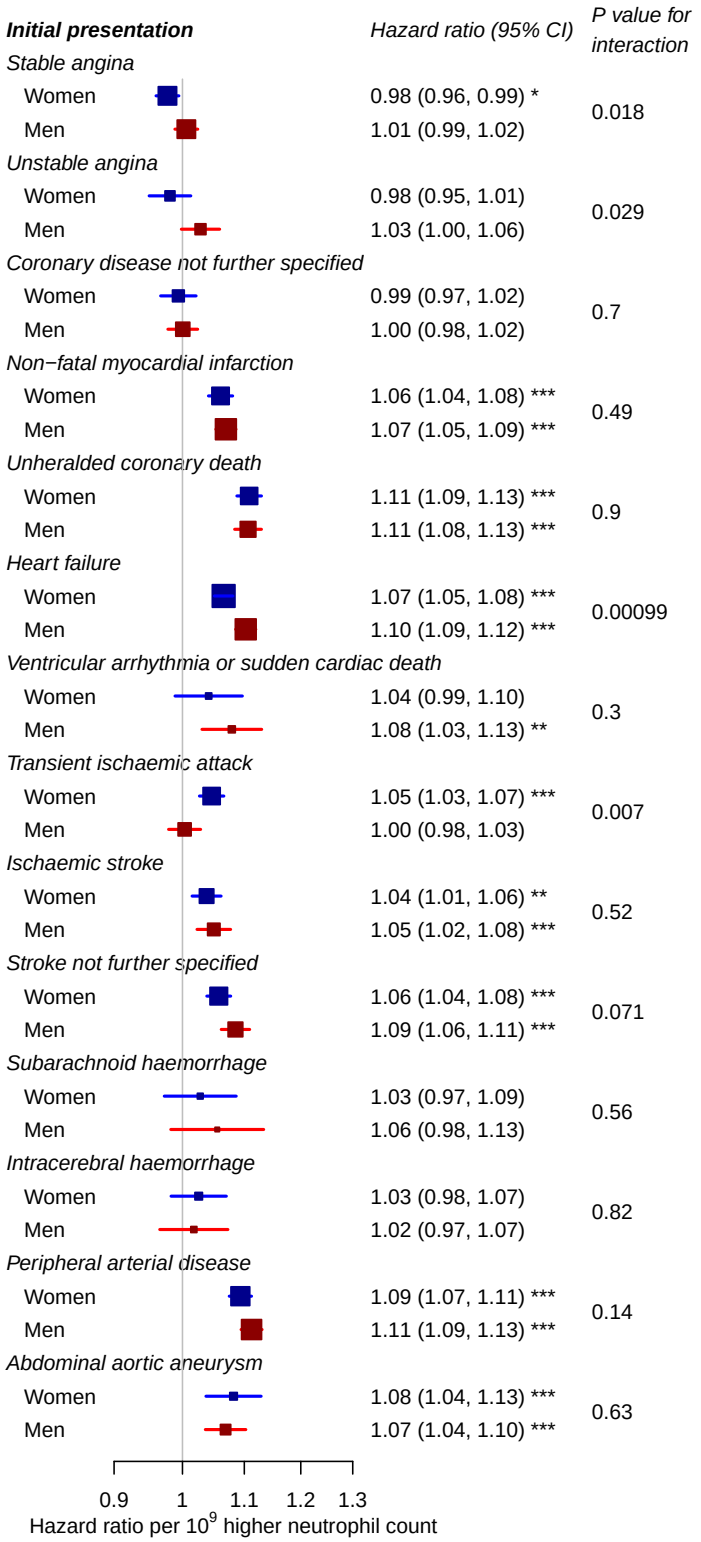


Figure 7.15: Linear association of neutrophil count with different initial presentations of cardiovascular disease, by smoking status

Hazard ratios are adjusted for age, sex, deprivation, ethnicity, smoking, diabetes, systolic blood pressure, blood pressure medication, body mass index, total cholesterol, HDL cholesterol, statin use, eGFR, atrial fibrillation, autoimmune conditions, inflammatory bowel disease, COPD, cancer, and acute conditions at the time of blood testing. P values * < 0.05, ** < 0.01, *** < 0.001

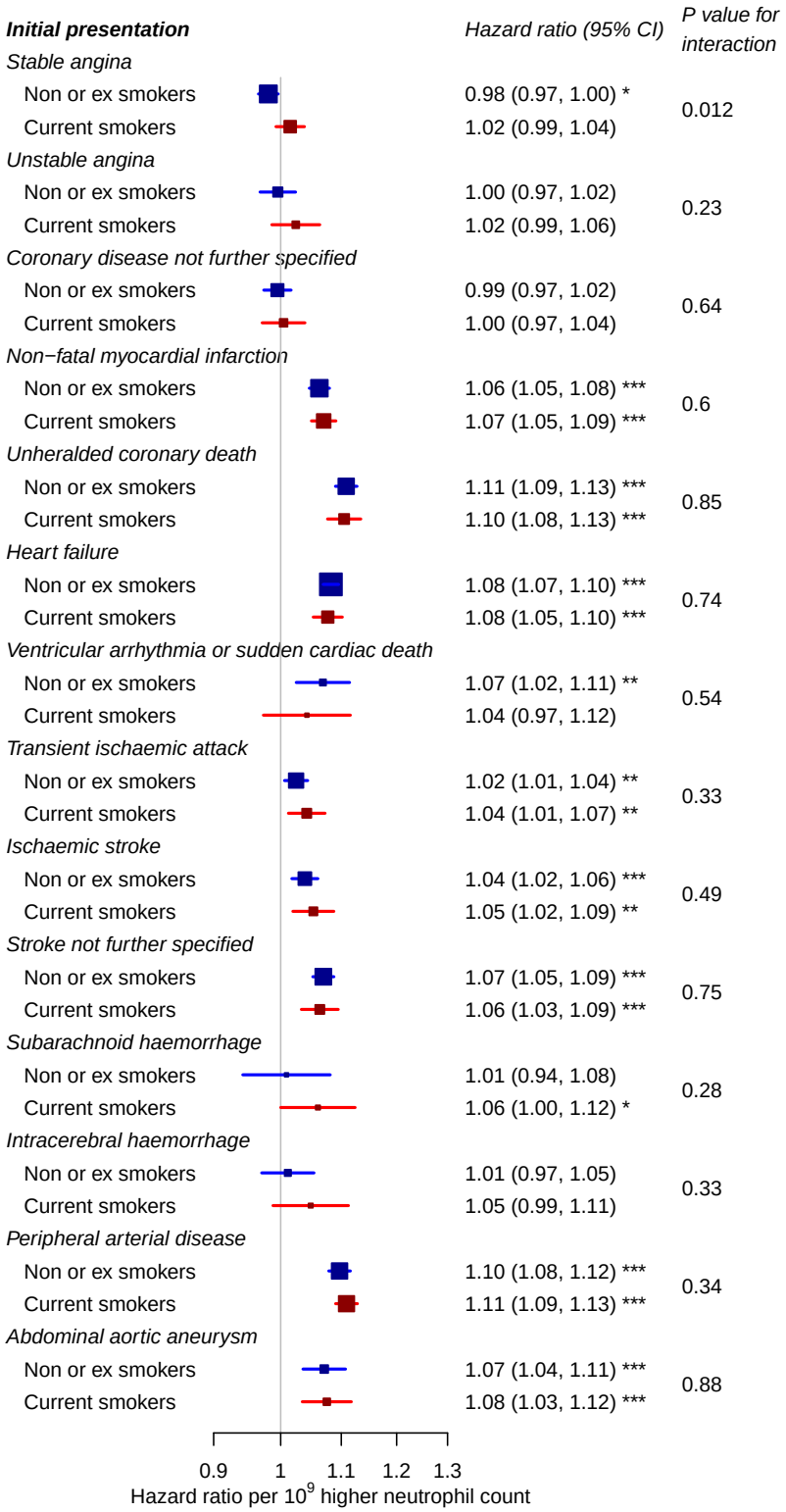


Figure 7.16: Association of quintiles of monocyte counts with different initial presentations of cardiovascular disease

Hazard ratios are adjusted for age, sex, deprivation, ethnicity, smoking, diabetes, systolic blood pressure, blood pressure medication, body mass index, total cholesterol, HDL cholesterol, statin use, eGFR, atrial fibrillation, autoimmune conditions, inflammatory bowel disease, COPD, cancer, and acute conditions at the time of blood testing. P values * < 0.05, ** < 0.01, *** < 0.001

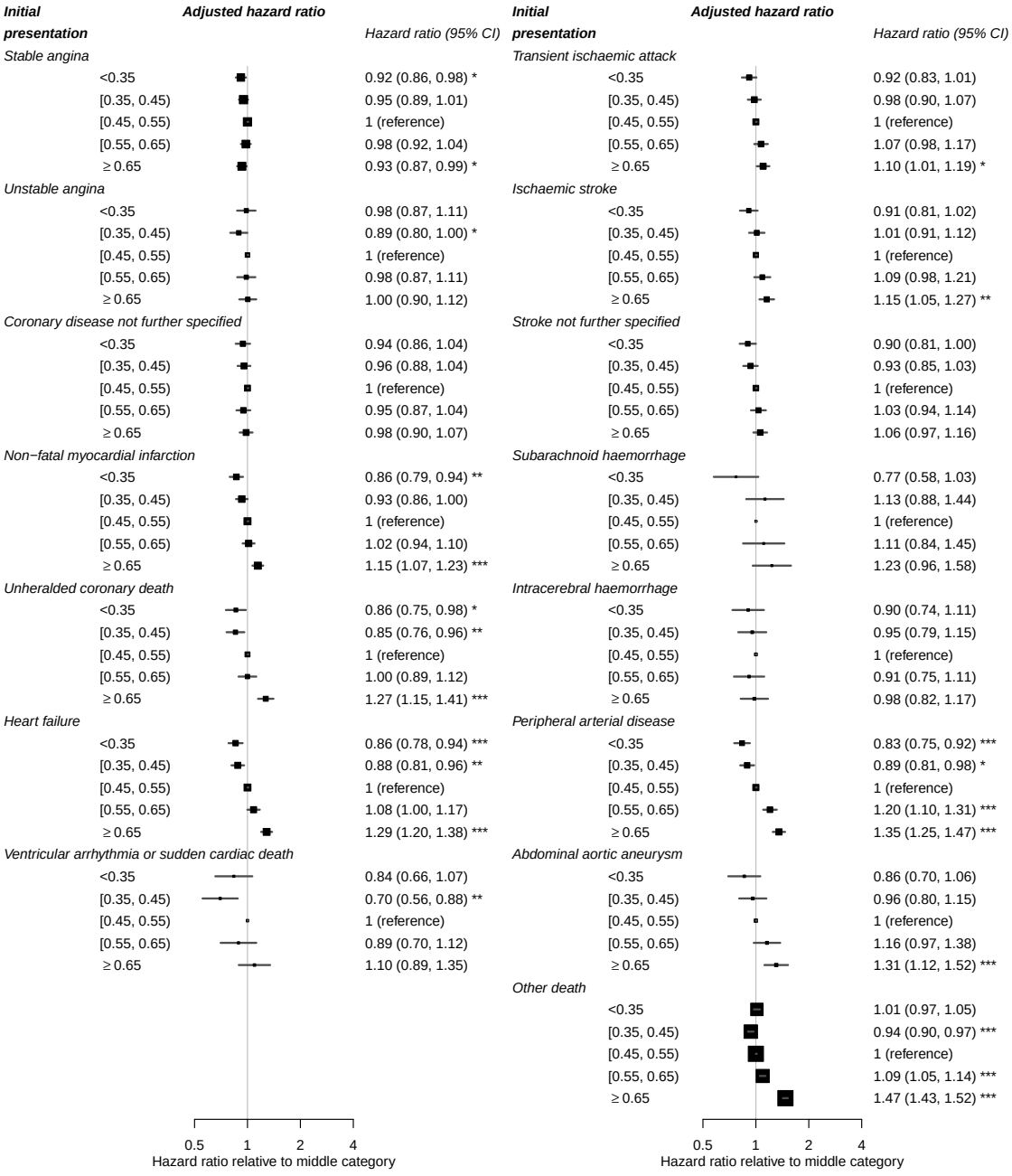


Figure 7.17: Association of quintiles of lymphocyte counts with different initial presentations of cardiovascular disease

Hazard ratios are adjusted for age, sex, deprivation, ethnicity, smoking, diabetes, systolic blood pressure, blood pressure medication, body mass index, total cholesterol, HDL cholesterol, statin use, eGFR, atrial fibrillation, autoimmune conditions, inflammatory bowel disease, COPD, cancer, and acute conditions at the time of blood testing. P values * < 0.05, ** < 0.01, *** < 0.001

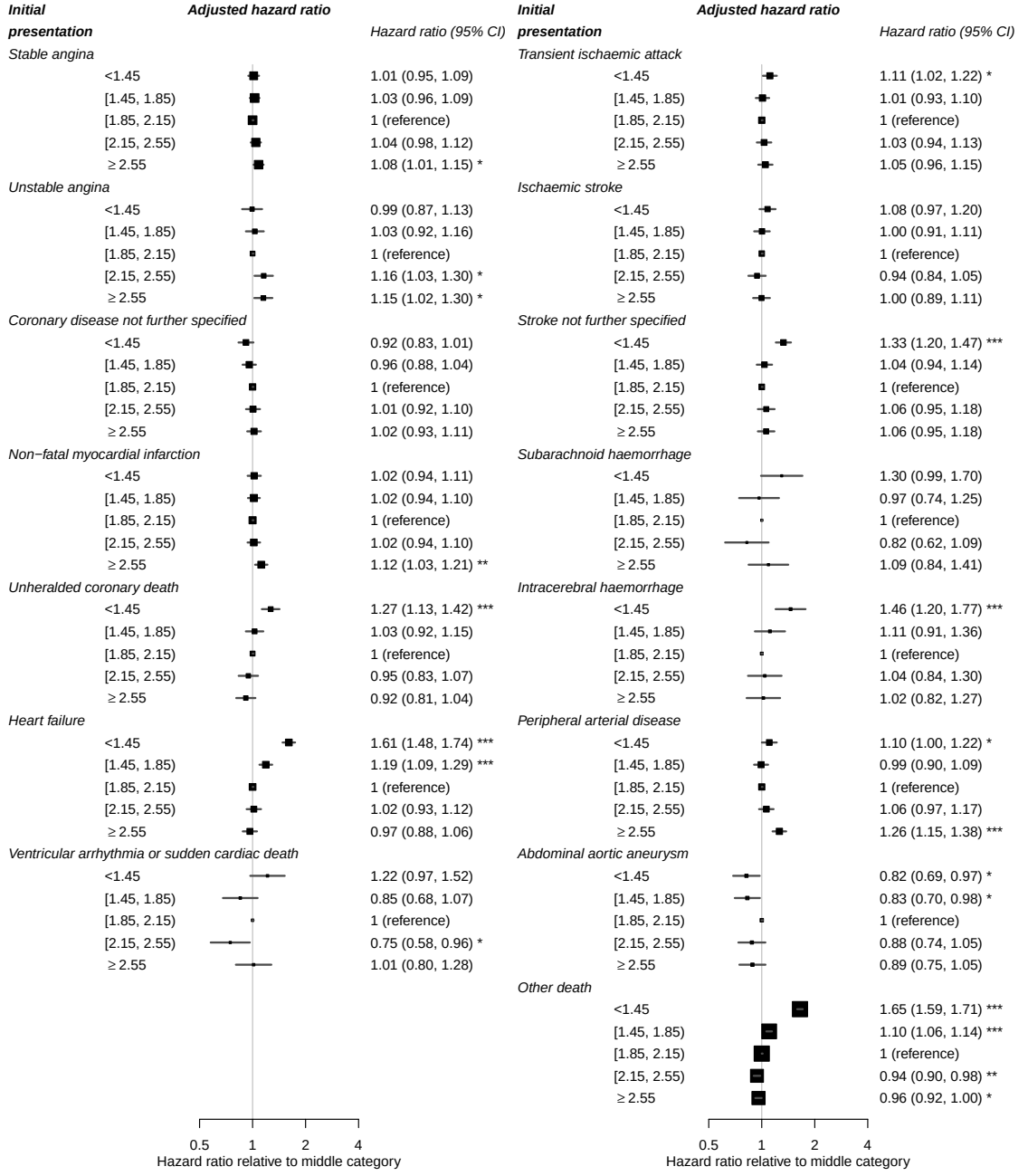


Figure 7.18: Association of basophil counts with different initial presentations of cardiovascular disease

Hazard ratios are adjusted for age, sex, deprivation, ethnicity, smoking, diabetes, systolic blood pressure, blood pressure medication, body mass index, total cholesterol, HDL cholesterol, statin use, eGFR, atrial fibrillation, autoimmune conditions, inflammatory bowel disease, COPD, cancer, and acute conditions at the time of blood testing. P values * < 0.05, ** < 0.01, *** < 0.001

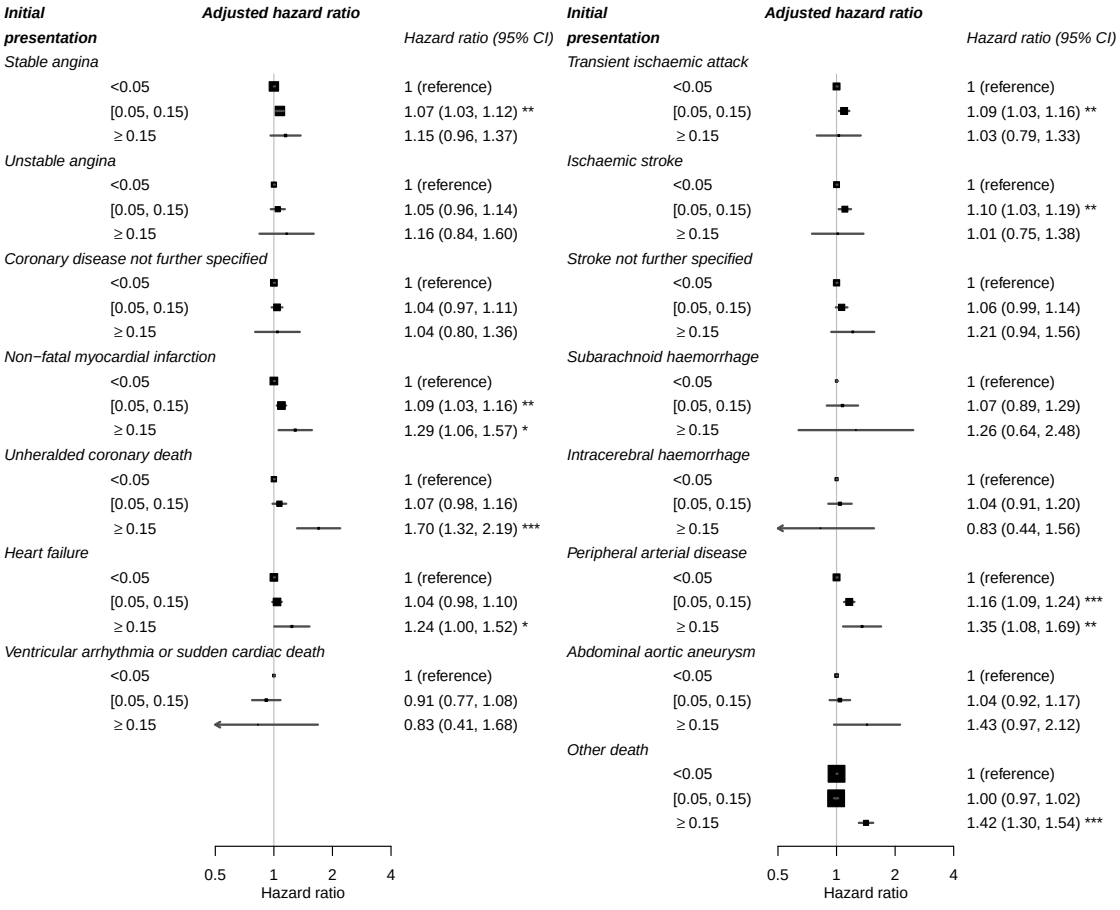
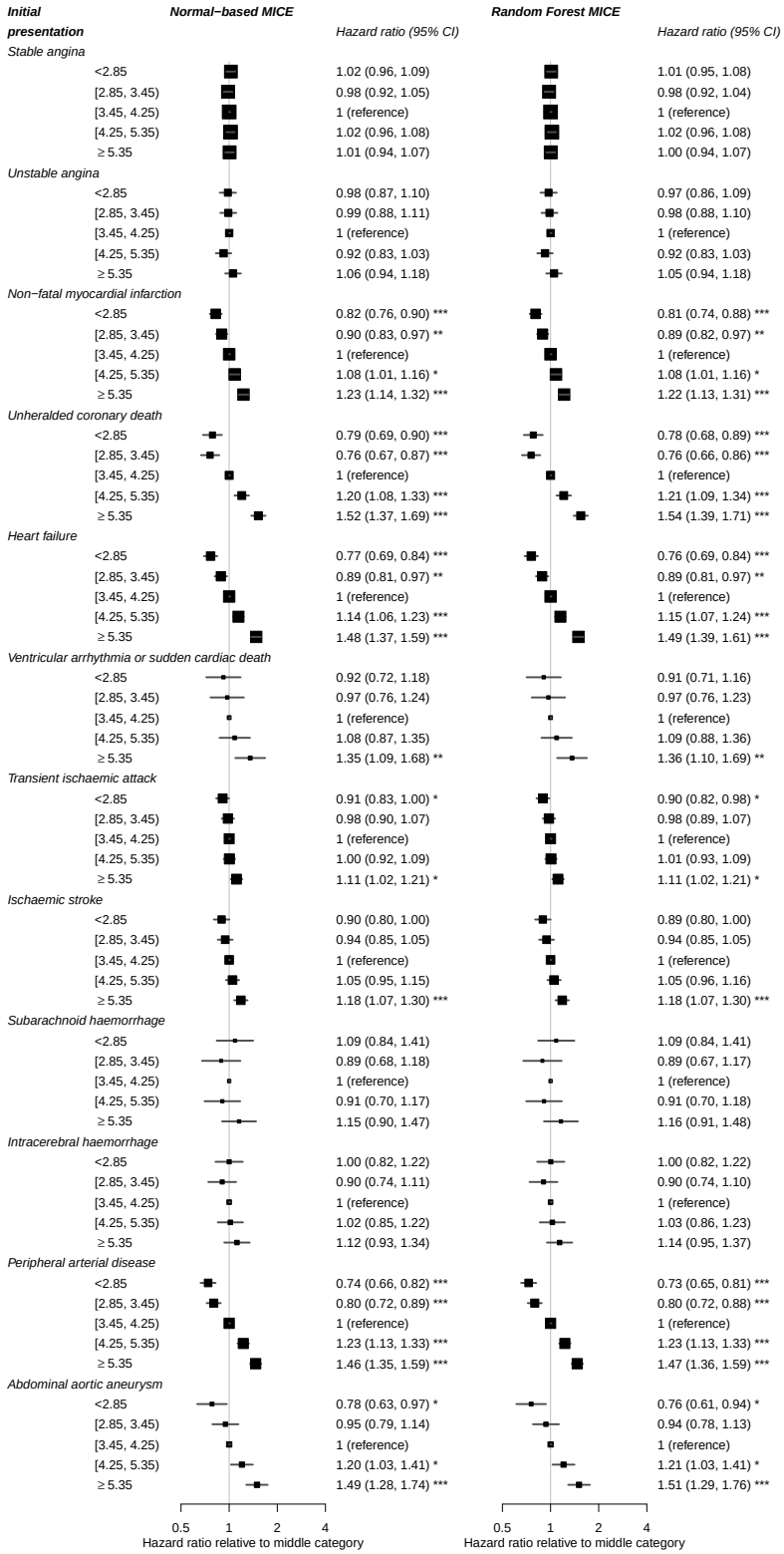


Figure 7.19: Association of neutrophil quintiles with different initial presentations of cardiovascular disease, using different methods of multiple imputation

Hazard ratios are adjusted for age, sex, deprivation, ethnicity, smoking, diabetes, systolic blood pressure, body mass index, total cholesterol, HDL cholesterol, eGFR, atrial fibrillation, autoimmune conditions, inflammatory bowel disease, COPD, cancer, statin use, blood pressure medication and acute conditions at the time of blood testing. P values * < 0.05, ** < 0.01, *** < 0.001



7.6 Discussion

Neutrophil counts within the normal range had strong linear associations with specific initial presentations of cardiovascular diseases in a large population-based cohort. Unlike previous studies which did not disaggregate coronary endpoints, I found that neutrophil counts had strong monotonic positive associations with myocardial infarction, unheralded coronary death and ventricular arrhythmia / sudden cardiac death, but no association with stable or unstable angina. Consistent with previous studies (**Table 2.2 on page 47**), I also found a strong association of neutrophil count with heart failure [113, 126], a moderate association with cerebral infarction [122] and no association with intracerebral haemorrhage [122]. Low neutrophil count (below the lower limit of the reference range) was not associated with increased risk of cardiovascular diseases, but was associated with significantly increased risk of non-cardiovascular death.

7.6.1 Role of neutrophils in thrombosis, inflammation and atherosclerosis

Neutrophils are the most numerous type of white blood cell in peripheral blood, and neutrophil counts are known to be raised in acute infections and other inflammatory states. Inflammation is a key process in the development of atherosclerotic plaques [48], which is why inflammatory biomarkers are of much interest in cardiovascular risk prediction. Many modifiable factors can increase the level of chronic low-grade inflammation, such as obesity [225, 226], periodontal disease [227], smoking [89], lack of exercise [228] and air pollution [229]. Air pollution is associated with increased incidence of heart failure [230], ischaemic heart disease and stroke [231]. Regular exercise suppresses pro-inflammatory cytokine production [228], whereas central obesity is thought to cause chronic inflammation via pro-inflammatory cytokines produced by adipose tissue [225, 226].

The pattern of associations seen in this study suggests that as well as their role in inflammation and atherosclerosis, neutrophils may also increase the risk of arterial thrombosis. Possible mechanisms include interactions with the endothelium and platelets [57] and overactivity of neutrophil extracellular traps [58]. This would explain why neutrophil counts are more strongly associated with myocardial infarction than stable or unstable angina.

This study cannot ascertain whether it is circulating neutrophils per se that confer the additional risk, or merely the underlying inflammatory state (with circulating cytokines) of which the neutrophil count is a surrogate marker [47]. Evidence for the causal relevance of neutrophils comes from a clinical trial of colchicine among patients with stable coronary disease [103]. Colchicine acts by inhibiting microtubule formation in neutrophils [232] and reduced the incidence of acute coronary syndrome and stroke in the trial [103].

Reducing chronic inflammation is thus a potentially important novel therapeutic avenue in atherosclerotic disease. Trials are currently underway to examine whether anti-inflammatory agents such as methotrexate [233] or canakinumab, a human monoclonal

antibody that selectively neutralizes interleukin-1β [234], can prevent cardiovascular disease in high risk individuals.

7.6.2 Neutrophils and cardiovascular risk

Neutrophil counts in relation to cardiovascular risk have been relatively under-studied in the general population (**Table 2.2 on page 47**), in contrast to circulating cytokines (e.g. interleukin-6 [51]) and acute phase proteins (e.g. C reactive protein, fibrinogen [50]). These biomarkers of inflammation are strongly associated with increased risk of coronary disease [50, 51] and US guidelines recommend that measurement of high-sensitivity C reactive protein can be considered to inform a cardiovascular risk treatment decision in uncertain cases [30]. The advantage of using neutrophil count over other biomarkers of inflammation in clinical risk prediction is that it is already performed, and would not incur any additional costs for testing. In contrast, C reactive protein is rarely measured in general practice and only in a highly selected subset of patients with suspected acute inflammation or infection, and the other biomarkers are not measured in clinical practice.

The strength of the association between neutrophil count and cardiovascular diseases suggest that it should be considered a cardiovascular risk factor in a similar league to systolic blood pressure (SBP). Comparing the strength of associations for heart failure with neutrophil count or SBP across the normal range (i.e. 90–140 mmHg for SBP), the hazard ratio of 2.04 for neutrophil count is significantly greater than 1.51 (95% CI 1.36, 1.64) for SBP (scaled to a 50mmHg difference from Supplementary Figure S5 in Rapsomaniki et al. [36]). Neutrophil count and SBP have similar strengths of association with unheralded coronary death (HR 1.78 and 1.71 respectively). The finding that neutrophil count is associated with myocardial infarction but not stable angina implies that neutrophil count may be related to thrombosis as well as atherosclerosis.

Neutrophils are short-lived cells and are affected by acute illnesses; I found there was significant variation between repeat tests on the same individual, giving a within-person correlation of neutrophil counts taken under 'stable' conditions of 0.568. This compares to 0.68 for total cholesterol and 0.73 for HDL cholesterol [105]. Hence inter-test variability could affect the usefulness of the neutrophil count as a cardiovascular risk marker. I did not adjust for this regression dilution bias in these analyses, because although the adjusted results would convey more accurately the strength of the biological association with average neutrophil count (or the underlying inflammatory process that it represents), they would give a misleading impression of the predictive value of the single or small number of differential leukocyte measurements that would be performed in practice. In an analysis adjusted for regression dilution bias the exposure would effectively be the unobservable mean neutrophil count. I found that associations were stronger among patients with neutrophil counts taken under 'stable' conditions, and were further strengthened when the mean of two such measurements were used, which would provide a more precise estimate of the level of chronic inflammation.

Smoking causes elevation of neutrophil counts [89,235]. The finding of similar strength of associations in non-smokers and smokers shows that residual confounding due to imprecise recording of smoking severity cannot explain away these results. I adjusted for smoking in the main analysis in order to demonstrate the independent association of neutrophil counts with cardiovascular diseases unconfounded by other risk factors. However, if a raised neutrophil count (or the underlying chronic inflammation of which it is a proxy) is on one of the causal pathways linking smoking to cardiovascular disease [89], adjustment for smoking may result in 'over-adjustment', with under-estimation of the component of cardiovascular risk conveyed by chronic inflammation or neutrophils.

7.6.3 Limitations

Although this study has strengths – large size, population base, extensive adjustment for potential confounders – it also has important limitations. As with any observational study, the results cannot be taken to imply causation because there is the possibility of residual confounding. Another limitation is selection bias, as participants were only included if they had a full blood count performed in routine clinical care, and the indications for this test could vary widely. However I investigated conditions and medication that may acutely affect the leukocyte count and obtained consistent results in a range of sensitivity analyses.

The measurement of leukocyte counts was undertaken in routine clinical care, by a large number of different laboratories without study-wide protocols. However, this would be expected to lead to random error rather than systematic bias. My results are consistent with previous studies which recruited healthy individuals and measured leukocyte counts under standardised conditions.

As the study was based on electronic health records, some values of baseline variables were missing for some patients. However, I obtained similar results by imputing missing data using two different methods of multiple imputation, and I used rich datasets for imputation with many clinical variables as auxiliary variables in imputation models, which would help to ensure that data were Missing at Random.

The ascertainment of endpoints was in routinely coded clinical data, without endpoint adjudication. It is known that all of the data sources used in this study miss some events; however the validation study of myocardial infarction showed that events recorded in each source had the expected prognosis and patient characteristics of myocardial infarction, suggesting that the sources have good specificity (see section 4.2). Validation studies of the CPRD have also shown that diagnoses recorded in primary care are likely to be correct [133]. Any failures in endpoint recording are likely to be non-differential in relation to the leukocyte count.

7.6.4 Research implications

There are two broad avenues for further research: investigating potential causal mechanisms and using neutrophil counts in risk prediction. Investigation of causal mechanisms could involve epidemiological studies of associations involving upstream determinants of neutrophil counts, such as granulocyte colony stimulating factor, interleukin-17 and interleukin-23 [236]. Mendelian randomisation studies using SNPs for genes associated with neutrophil count, such as those identified in the 17q21 region [56, 237], may also help to determine whether the neutrophil count itself is causal in atherosclerotic disease. This would strengthen the case for immune modulation to be considered as a therapy to prevent or treat cardiovascular diseases, and may suggest novel therapeutic targets.

Neutrophil counts are almost always measured as part of a full leukocyte differential, meaning that they are usually associated with measures of other cell types (lymphocytes, monocytes, eosinophils and basophils). Monocytes are part of the immune response and migrate into atherosclerotic plaques to form foam cells [48]; they showed similar associations with cardiovascular diseases to neutrophils (**Figure 7.16 on page 180**). Low counts of lymphocytes were associated with increased heart failure, unheralded coronary death and other death (**Figure 7.17 on page 181**). Previous studies [113, 126] have not been large enough to demonstrate a statistically significant association. There was no clear pattern to the remaining results, consistent with previous studies (**Table 2.4 on page 49**), probably because the overall lymphocyte count comprises a number of subtypes with different roles (e.g. B cells, killer T cells, T regulatory cells). This is in contrast to neutrophils, which comprise a single class of cell all with the same function. Low lymphocyte count is associated with malnutrition [238] and may also be due to bone marrow insufficiency and general frailty.

A recommendation arising from this research is that the components of the leukocyte differential should be tested in predictive models for cardiovascular diseases [30, 47]. Given the variability in neutrophil counts, more accurate predictions may be obtained using repeat measurements.

7.6.5 Clinical implications

Leukocyte counts are commonly performed in clinical practice, but they tend to inform clinical management only if the results are outside the reference range. Much prognostic information is conveyed by neutrophil counts within the reference range, and it should be considered as a continuous measure like cholesterol or blood pressure. Neutrophil count is affected by behavioural cardiovascular risk factors such as smoking and lack of exercise, so it may be useful for monitoring a patient's progress with a cardiovascular risk reduction programme; for example smoking cessation reduced the mean neutrophil count by 1.0×10^9/L [239], and a clinical trial of moderate regular exercise (8 kcal per kg body weight per week) reduced the mean neutrophil count by 0.1×10^9/L in overweight postmenopausal women [240].

7.6.6 Conclusion

Neutrophil counts are differentially associated with the short and long term risk of initial presentations of cardiovascular diseases. Clinicians should be aware of the prognostic information conveyed by neutrophil counts even within the reference range. The leukocyte differential should be investigated for use in cardiovascular risk prediction models.

7.7 Summary

This study uses the large sample size and rich clinical information in the CALIBER dataset to show that neutrophil counts are differentially associated specific initial presentations of cardiovascular diseases. There were also interesting findings in relation to eosinophil counts, which are described in the next chapter, chapter 8. These novel findings should be replicated in other settings, so I present a replication study in chapter 9.

8 Associations of eosinophil counts with specific initial presentations of cardiovascular diseases

8.1 Chapter outline

This chapter investigates the association of eosinophil counts with incidence of cardiovascular diseases. The cohort definition is described in chapter 6 and is the same that which yielded the results for neutrophils (chapter 7).

8.2 Abstract

Background: The eosinophil count is a component of the white blood cell (leukocyte) differential, a common blood test. Eosinophils may be implicated in coronary disease, but the association of eosinophil counts with the onset of cardiovascular diseases has not previously been investigated in large population-based studies.

Methods: The data source was a linked primary care, hospital admission, disease registry, and mortality dataset for England (the CALIBER programme). The study included people aged 30 or older with a leukocyte differential recorded, who were free from cardiovascular disease as baseline. Participants were followed for the first occurrence of stable angina, unstable angina, myocardial infarction, unheralded coronary death, heart failure, ventricular arrhythmia / sudden cardiac death, transient ischaemic attack, ischaemic stroke, haemorrhagic stroke, subarachnoid haemorrhage, abdominal aortic aneurysm or peripheral arterial disease. Cox models were used to estimate cause-specific hazard ratios (HR). This study is registered at ClinicalTrials.gov (NCT02014610).

Results: The cohort comprised 775 231 individuals, of whom 54 980 experienced an initial presentation of cardiovascular disease over a median follow up of 3.8 years. Low eosinophil count was associated with low levels of cardiovascular risk factors such as smoking and diabetes. There was a strong association of low eosinophil counts (<0.05 compared to 0.15–0.25 $\times 10^9$/L) over the first 6 months with heart failure (adjusted HR 2.05; 95% confidence interval (CI) 1.72, 2.43), and unheralded coronary death (adjusted HR 1.94, 95% CI 1.40, 2.69), but not stable angina or non-fatal myocardial infarction. Associations beyond the first 6 months were weaker.

Conclusions: Low eosinophil counts are strongly associated with increased short-term incidence of heart failure and coronary death.

8.3 Introduction

Eosinophils are multifunctional leukocytes (white blood cells) which have various roles in inflammatory processes [74], and are thought to be particularly important for maintenance of tissue homeostasis through 'Local Immune And Remodelling / Repair' activities [75]. Eosinophil counts are routinely measured in peripheral blood as part of the full blood count (complete blood count), one of the most widely performed blood tests in medical practice. The exact nature of their roles in relation to cardiovascular diseases are unclear. Eosinophil counts increase acutely after myocardial infarction [76] and histological studies found large numbers of eosinophils in coronary thrombi [77] and post-mortem hearts with cardiac rupture [78], suggesting that eosinophils may be implicated in the pathogenesis of myocardial infarction. This hypothesis is supported by the finding of a non-synonymous SNP associated both with eosinophil count and myocardial infarction [79].

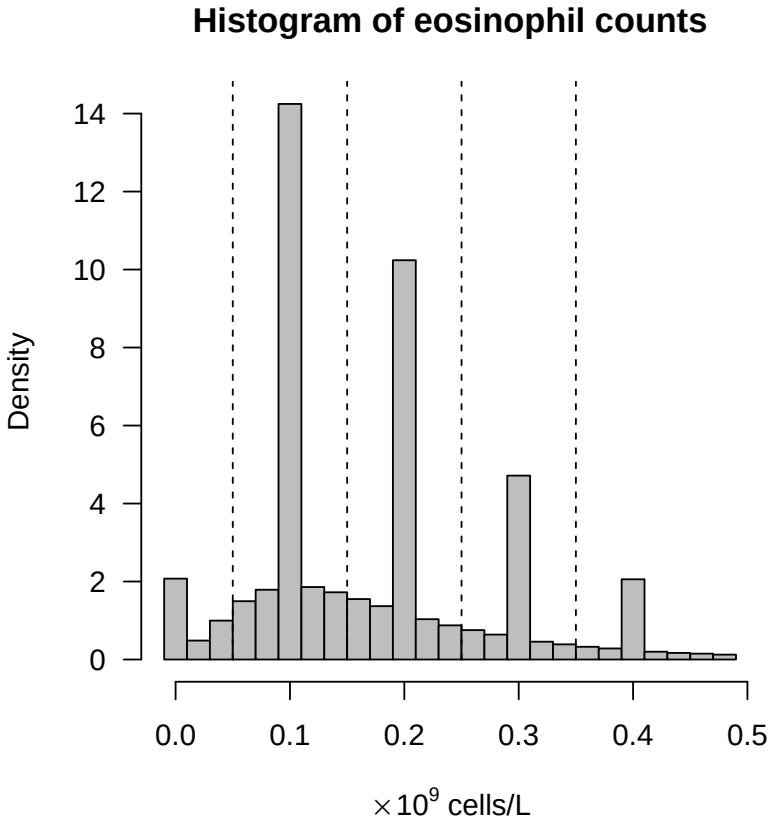
Prospective cohort studies investigating the association of baseline eosinophil counts with development of cardiovascular disease in healthy people have been few in number and small in size (**Table 2.6 on page 51**). In the Hiroshima and Nagasaki study [107] and Caerphilly study [116], higher eosinophil counts were associated with greater incidence of coronary heart disease. Conversely, in populations with cardiovascular disease (e.g. patients undergoing percutaneous intervention [127] or with acute heart failure [128]), lower eosinophil counts were associated with worse prognosis. Low eosinophil count was also found to be associated with worse short-term mortality in critical care patients [241].

The existing evidence relating eosinophil counts to cardiovascular diseases is thus scarce and contradictory. Previous studies were small, had limited adjustment for cardiovascular risk factors, and investigated a narrow range of cardiac endpoints. It is unknown whether eosinophil counts are associated with other cardiovascular diseases such as stroke, heart failure or peripheral arterial disease, and whether associations differ by short or long term risk. The aim of this study was to address these gaps in knowledge by means of a large population-based cohort study in the CALIBER linked electronic health record database [1].

8.4 Methods

The cohort definition, covariates and multiple imputation methods were as described in section 7.4. Briefly, patients aged 30 or over who had at least one full blood count recorded while registered at a CPRD practice were eligible for the study. The patient state at the time of the blood test was classified as 'acute' or 'stable', and patients were followed up for the initial presentation of one of twelve cardiovascular diseases. Baseline covariate data were extracted as described in **Section 6.6**.

Figure 8.1: Histogram showing distribution of eosinophil counts and category cutpoints



8.4.1 Exposure

The main exposure was the eosinophil count as recorded in CPRD. If a patient had more than one measurement on a given day, the values were aggregated by taking the mean. I examined quintiles of eosinophil count in order to avoid presuming a particular shape for the association with cardiovascular diseases. The precision with which eosinophil counts were recorded (in cells $\times 10^9$/L) varied from one to two decimal places (**Figure 8.1 on page 191**). In order to avoid biasing the category allocation by precision, I manually adjusted the category boundaries so that the second decimal place was 5, thus ensuring that any value recorded to two decimal places would end up in the same category as if it were recorded to only one decimal place. All category intervals were closed at the lower end and open at the higher end.

In secondary analyses, I explored associations between onset of cardiovascular diseases and the mean of the first two 'stable' measurements of eosinophil count taken since the start of eligibility. I also carried out a sensitivity analysis excluding patients with a history of prior loop diuretic use, who might have symptoms of heart failure without a formal diagnosis.

8.4.2 Statistical analysis

I generated cumulative incidence curves by category of eosinophil counts under a competing risks framework. I used the Cox proportional hazards model to generate cause-specific hazards for the different cardiovascular endpoints. Hazard ratios were adjusted for age (linear and quadratic), sex, age / sex interaction, index of multiple deprivation, ethnicity, smoking status, diabetes, body mass index, systolic blood pressure, eGFR, HDL, total cholesterol, atrial fibrillation, inflammatory conditions (autoimmune conditions, inflammatory bowel disease or COPD), cancer, statin use, blood pressure medication and acute conditions at the time of blood testing. The baseline hazard was stratified by practice and sex. I plotted Schoenfeld residuals to assess the proportional hazards assumption, and split the follow-up time if hazard ratios changed over time. I handled missing baseline covariate data by multiple imputation using chained equations [187], using Random Forest multiple imputation models [148], as described previously (subsection 7.4.7).

8.5 Results

The cohort included 621 052 patients with differential leukocyte counts while clinically 'stable' and 154 179 patients with differential leukocyte counts performed during acute illness or treatment (**Figure 6.1 on page 135**). 54 980 initial presentations of cardiovascular disease were observed over a median of 3.8 (IQR 1.7, 6.0) years follow-up.

Eosinophil counts were low and measured imprecisely, so the categories contain unequal numbers of patients (**Table 8.1 on page 193**). Amongst patients in the lowest category, 28 791 had 0 eosinophils. Of these, only 104 (0.4%) also had zero neutrophils. Patients in the lowest eosinophil category were most likely to have an acute state at the time of the blood test (**Table 8.2 on page 194**).

Lower eosinophil count was associated with cancer, being female, and Black ethnicity, whereas higher eosinophil count was associated with male sex, deprivation, South Asian ethnicity, higher body mass index, and atopy (**Figure 8.2 on page 195**).

Crude cumulative incidence curves showed that individuals with eosinophil counts less than 0.05 $\times 10^9$/L had a greater incidence of heart failure as the initial presentation of cardiovascular disease, at least in the first few months (**Figure 8.3 on page 196**).

In multiply adjusted analyses there was a strong association of low eosinophil counts (<0.05 compared to 0.15–0.25 $\times 10^9$/L) with incident heart failure (adjusted hazard ratio (HR) 1.39; 95% confidence interval (CI) 1.26, 1.54), unheralded coronary death (HR 1.29, 95% CI 1.12, 1.50) and ventricular arrhythmia or sudden cardiac death (HR 1.49, 95% CI 1.12, 2.00), but not stable angina (HR 0.93, 95% CI 0.84, 1.03), non-fatal myocardial infarction (HR 1.00, 95% CI 0.89, 1.12) or ischaemic stroke (HR 1.02, 95% CI 0.88, 1.19) (**Figure 8.4 on page 197**).

Table 8.1: Characteristics of patients by category of eosinophil count

Category	**1 (lowest)**	**2**	**3**	**4**	**5 (highest)**
Eosinophils, $\times 10^9$/L	**< 0.05**	**[0.05, 0.15)**	**[0.15, 0.25)**	**[0.25, 0.35)**	**≥ 0.35**
N patients	44 112	307 668	228 639	106 092	88 720
Women, n (%)	29 990 (68.0%)	199 375 (64.8%)	132 275 (57.9%)	55 902 (52.7%)	44 414 (50.1%)
Age, median (IQR)	52 (39.8-66.8)	52.1 (41-64.3)	53 (42.1-64.6)	52.8 (42-64.5)	51.9 (41.1-64.4)
Most deprived quintile, n (%)	7805 (17.7%)	54 017 (17.6%)	44 765 (19.6%)	22 237 (21.0%)	19 859 (22.5%)
Ethnicity, n (%):					
White	25 511 (91.7%)	175 418 (93.0%)	131 908 (93.5%)	61 286 (92.8%)	50 911 (90.7%)
South Asian	552 (2.0%)	4187 (2.2%)	3774 (2.7%)	2238 (3.4%)	2699 (4.8%)
Black	1051 (3.8%)	4496 (2.4%)	2356 (1.7%)	1017 (1.5%)	1054 (1.9%)
Other	714 (2.6%)	4564 (2.4%)	3056 (2.2%)	1501 (2.3%)	1485 (2.6%)
Missing	16 284 (36.9%)	119 003 (38.7%)	87 545 (38.3%)	40 050 (37.8%)	32 571 (36.7%)
Full blood count parameters on index date, median (IQR):					
Neutrophils, $\times 10^9$/L	3.98 (2.8-5.7)	3.7 (2.9-4.8)	3.9 (3.1-4.97)	4 (3.2-5.12)	4.2 (3.3-5.36)
Lymphocytes, $\times 10^9$/L	1.6 (1.2-2)	1.87 (1.5-2.3)	2 (1.66-2.5)	2.1 (1.71-2.6)	2.2 (1.8-2.73)
Monocytes, $\times 10^9$/L	0.42 (0.3-0.6)	0.44 (0.35-0.6)	0.5 (0.4-0.6)	0.5 (0.4-0.64)	0.53 (0.4-0.7)
Basophils, $\times 10^9$/L	0 (0-0.02)	0.02 (0-0.05)	0.03 (0-0.08)	0.04 (0-0.1)	0.05 (0-0.1)
Haemoglobin, g/dL	13.5 (12.5-14.5)	13.8 (12.9-14.8)	14 (13.1-15)	14.1 (13.1-15.1)	14.1 (13.1-15.1)
Platelets, $\times 10^9$/L	249 (206-297)	255 (217-299)	261 (223-306)	266 (227-312)	273 (233-321)
Smoking status, n (%):					
Never	25 034 (61.0%)	167 598 (57.6%)	109 365 (50.3%)	46 699 (46.2%)	38 006 (45.1%)
Ex	8840 (21.5%)	68 595 (23.6%)	53 712 (24.7%)	25 193 (24.9%)	20 859 (24.7%)
Current	7158 (17.4%)	54 997 (18.9%)	54 292 (25.0%)	29 089 (28.8%)	25 485 (30.2%)
Missing	3080 (7.0%)	16 478 (5.4%)	11 270 (4.9%)	5111 (4.8%)	4370 (4.9%)
Most recent value within one year prior to index date, median (IQR):					
Systolic blood pressure, mmHg	133 (120-148)	135 (120-150)	138 (123-150)	138 (124-150)	136 (122-150)
Body mass index, kg/m^2	25 (22-28.5)	26.5 (23.3-30.4)	27.6 (24.3-31.7)	27.8 (24.4-32)	27.5 (24.2-31.7)
Total cholesterol, mmol/L	5.36 (4.6-6.1)	5.5 (4.8-6.27)	5.5 (4.8-6.3)	5.5 (4.8-6.2)	5.4 (4.7-6.2)
HDL cholesterol, mmol/L	1.5 (1.2-1.8)	1.4 (1.2-1.71)	1.36 (1.1-1.63)	1.3 (1.1-1.6)	1.3 (1.1-1.6)
eGFR, mL/min/1.73m^2	81.4 (67.6-95.4)	82.1 (69.1-95.2)	81.7 (68.8-94.6)	81.9 (68.7-95.1)	82.4 (68.7-95.7)
Diagnoses on or before index date, n (%):					
Atrial fibrillation	555 (1.3%)	3137 (1.0%)	2163 (0.9%)	1027 (1.0%)	840 (0.9%)
Cancer	3895 (8.8%)	19 459 (6.3%)	13 523 (5.9%)	5950 (5.6%)	4694 (5.3%)
Diabetes	1502 (3.4%)	12 125 (3.9%)	11 941 (5.2%)	6022 (5.7%)	5537 (6.2%)
Asthma or atopy	9783 (22.2%)	73 708 (24.0%)	63 179 (27.6%)	33 355 (31.4%)	32 875 (37.1%)
COPD	863 (2.0%)	4417 (1.4%)	4445 (1.9%)	2484 (2.3%)	2812 (3.2%)
Connective tissue disease	1680 (3.8%)	8526 (2.8%)	6236 (2.7%)	2916 (2.7%)	2501 (2.8%)
IBD	704 (1.6%)	3077 (1.0%)	2400 (1.0%)	1249 (1.2%)	1216 (1.4%)
Medication use in the year before index date, n (%):					
Antihypertensives	9988 (22.6%)	73 943 (24.0%)	60 755 (26.6%)	28 752 (27.1%)	23 843 (26.9%)
Statins	1654 (3.7%)	15 864 (5.2%)	15 207 (6.7%)	7503 (7.1%)	6578 (7.4%)

COPD, chronic obstructive pulmonary disease; eGFR, estimated glomerular filtration rate; HDL, high density lipoprotein cholesterol, IBD, inflammatory bowel disease; IQR, interquartile range.

Table 8.2: Prevalence of acute conditions on date of blood testing by category of eosinophil count

Category **Eosinophils, $\times 10^9$/L**	**1 (lowest)** **< 0.05**	**2** **[0.05, 0.15)**	**3** **[0.15, 0.25)**	**4** **[0.25, 0.35)**	**5 (highest)** **≥ 0.35**
In hospital on date of blood test	662 (1.5%)	1177 (0.4%)	670 (0.3%)	291 (0.3%)	280 (0.3%)
Vaccination within previous 7 days	399 (0.9%)	2998 (1.0%)	2446 (1.1%)	1276 (1.2%)	1052 (1.2%)
Anaemia diagnosis within 30 days before	393 (0.9%)	1617 (0.5%)	940 (0.4%)	407 (0.4%)	400 (0.5%)
Infection diagnosis within 30 days before	3859 (8.7%)	20 129 (6.5%)	15 622 (6.8%)	7543 (7.1%)	7098 (8.0%)
Infective symptoms within 30 days before	3439 (7.8%)	16 166 (5.3%)	11 902 (5.2%)	5719 (5.4%)	5952 (6.7%)
Prior diagnosis of myelodysplastic syndrome	79 (0.2%)	117 (0.0%)	55 (0.0%)	31 (0.0%)	24 (0.0%)
Prior diagnosis of haemoglobinopathy	200 (0.5%)	1082 (0.4%)	688 (0.3%)	352 (0.3%)	327 (0.4%)
Chemotherapy or G-CSF within 6 months prior	561 (1.3%)	780 (0.3%)	390 (0.2%)	153 (0.1%)	172 (0.2%)
Methotrexate within 3 months prior	203 (0.5%)	1175 (0.4%)	966 (0.4%)	391 (0.4%)	342 (0.4%)
Steroid within 3 months prior	2304 (5.2%)	7244 (2.4%)	5473 (2.4%)	2849 (2.7%)	3306 (3.7%)
Other immune drug within 3 months prior	993 (2.3%)	3308 (1.1%)	2182 (1.0%)	992 (0.9%)	918 (1.0%)
Prior diagnosis of cancer	3895 (8.8%)	19 459 (6.3%)	13 523 (5.9%)	5950 (5.6%)	4694 (5.3%)
Any acute condition	12 068 (27.4%)	58 821 (19.1%)	43 324 (18.9%)	20 619 (19.4%)	19 347 (21.8%)

Figure 8.2: Selected characteristics of patients that varied significantly by eosinophil count

Bars represent 95% confidence intervals for the summary statistics

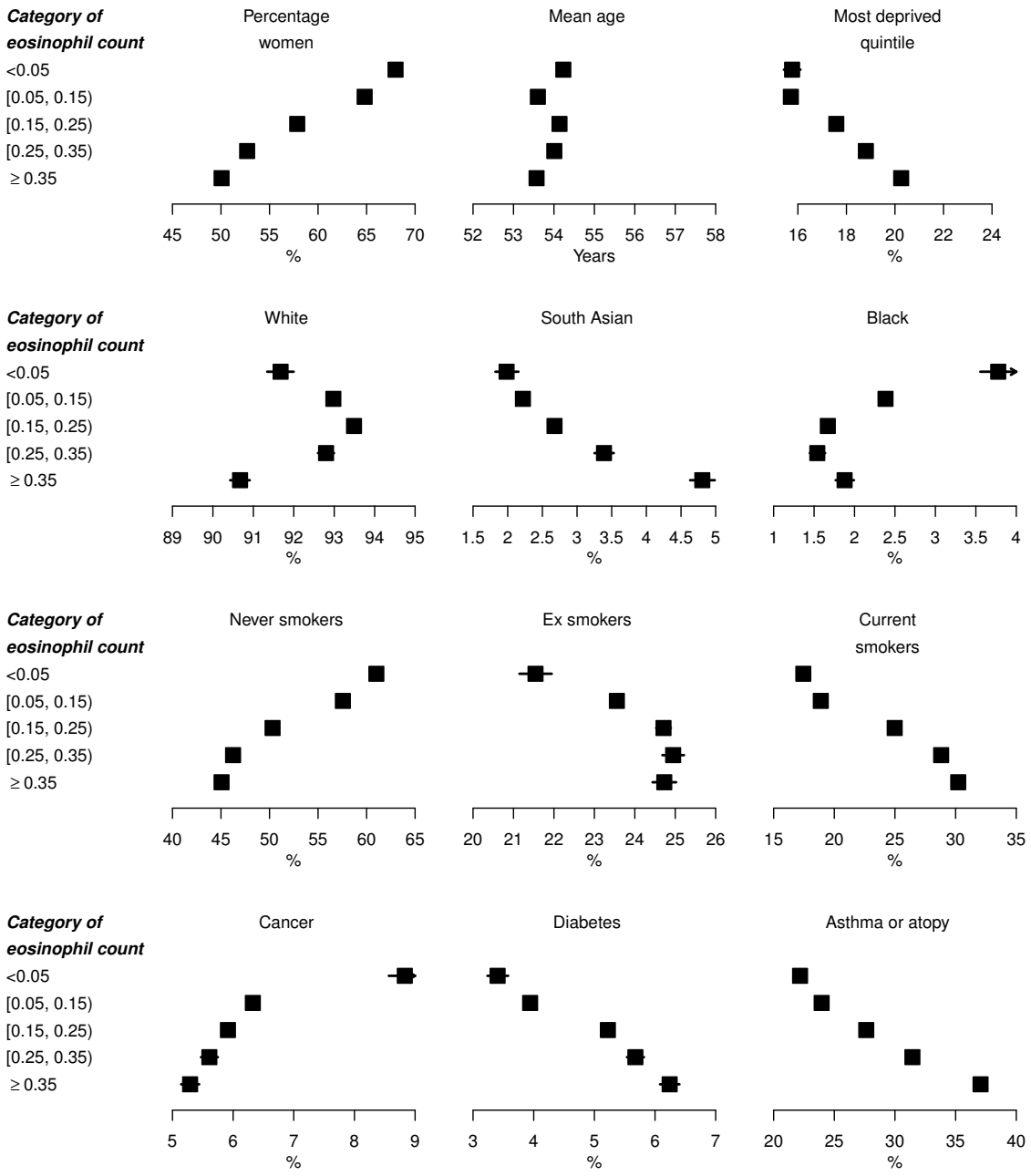


Figure 8.3: Cumulative incidence curves for different initial presentations of cardiovascular disease by category of eosinophil count

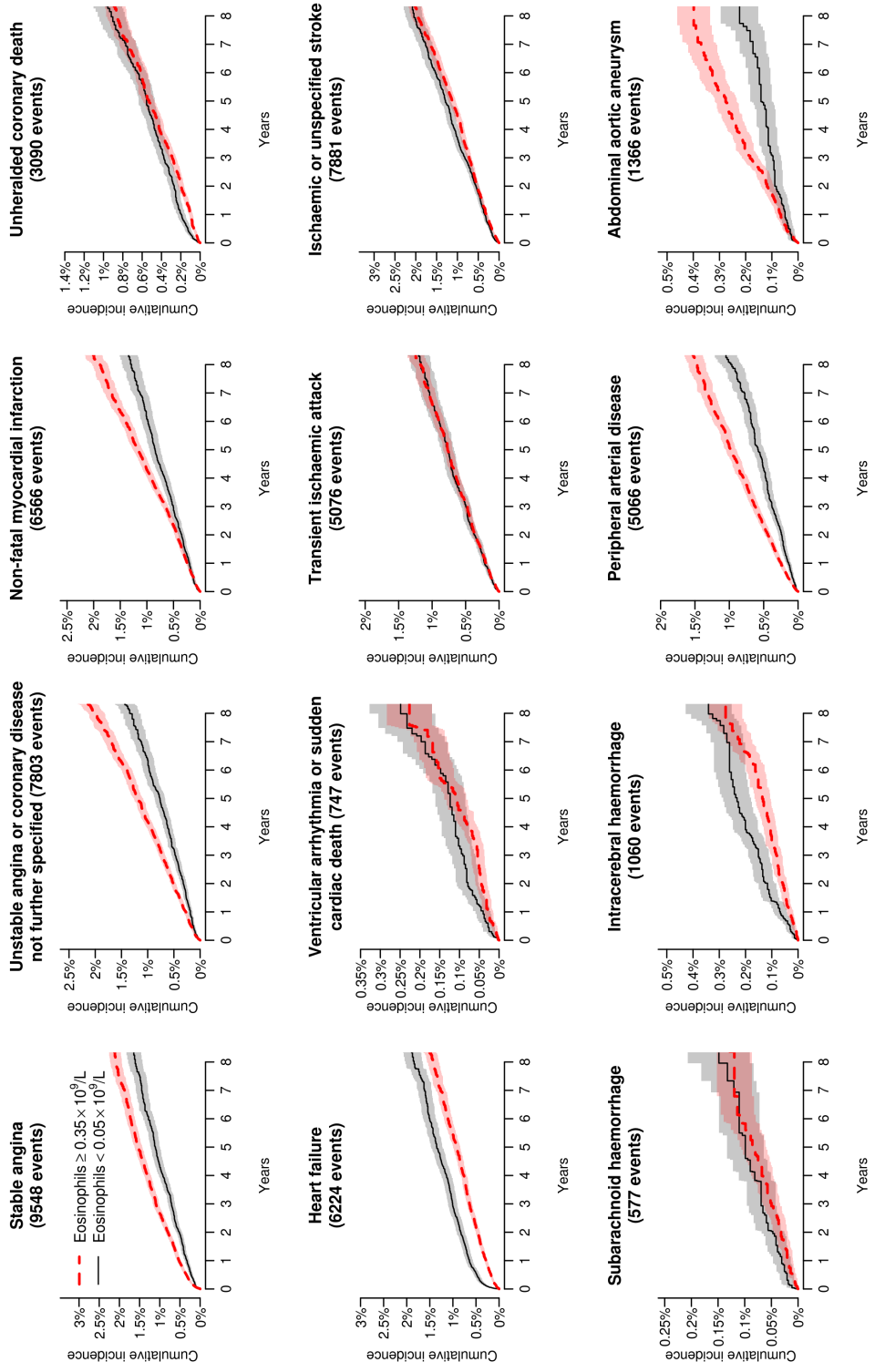


Figure 8.4: Association of eosinophil categories with different initial presentations of cardiovascular disease

Hazard ratios are adjusted for age, sex, deprivation, ethnicity, smoking, diabetes, systolic blood pressure, body mass index, total cholesterol, HDL cholesterol, eGFR, atrial fibrillation, autoimmune conditions, inflammatory bowel disease, COPD, cancer, statin use, blood pressure medication and acute conditions at the time of blood testing. P values * < 0.05, ** < 0.01, *** < 0.001

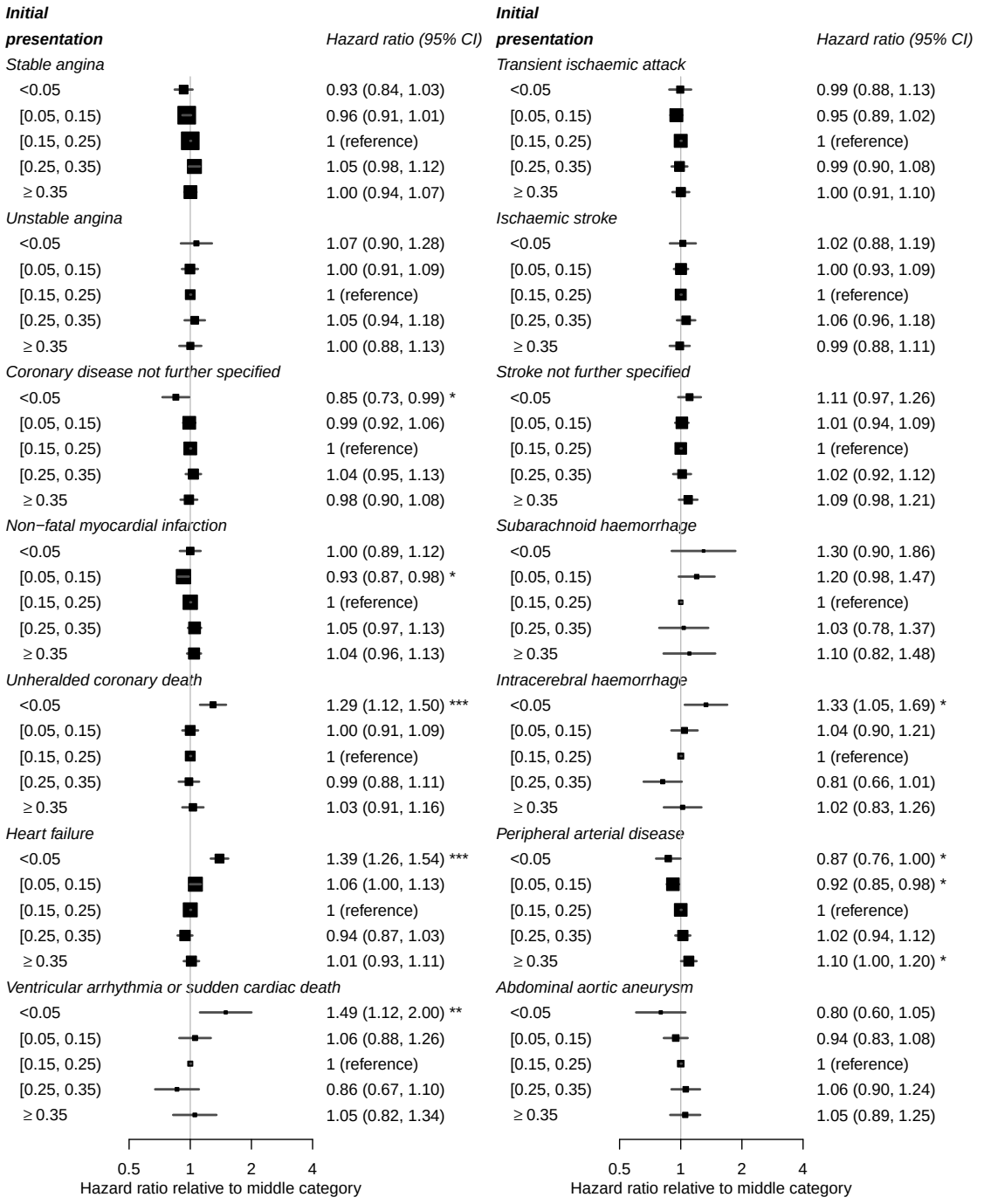
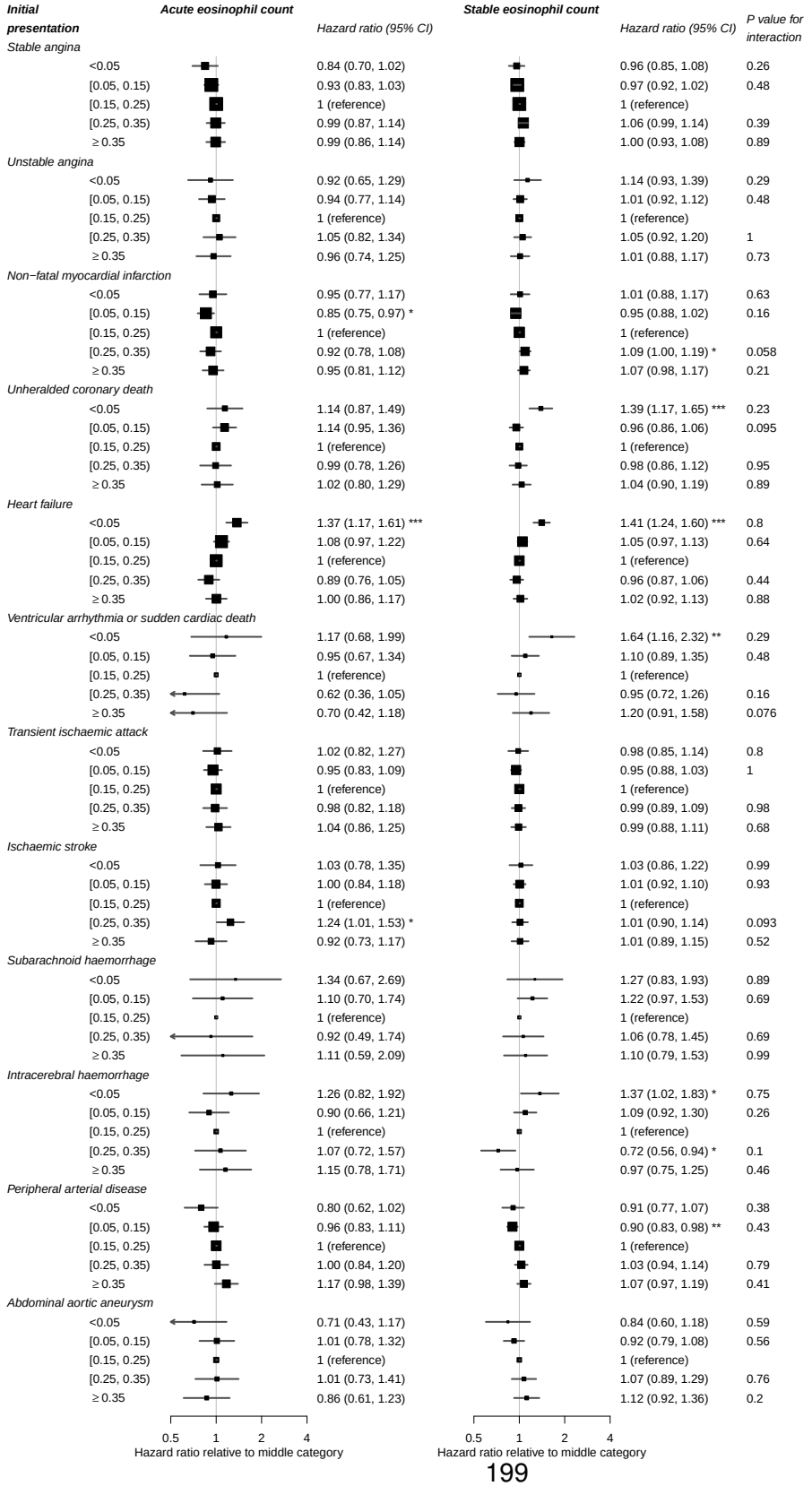


Table 8.3: Endpoints by category of eosinophil count

Category	**1 (lowest)**	**2**	**3**	**4**	**5 (highest)**
Eosinophils, $\times 10^9$/L	**< 0.05**	**[0.05, 0.15)**	**[0.15, 0.25)**	**[0.25, 0.35)**	**≥ 0.35**
Initial presentation of cardiovascular disease					
Stable angina	440 (13.6%)	3494 (17.3%)	2975 (18.0%)	1475 (18.2%)	1164 (16.8%)
Unstable angina	151 (4.7%)	1061 (5.3%)	876 (5.3%)	441 (5.4%)	351 (5.1%)
Coronary disease not further specified	193 (6.0%)	1805 (8.9%)	1546 (9.3%)	772 (9.5%)	607 (8.8%)
Non-fatal myocardial infarction	335 (10.4%)	2219 (11.0%)	2039 (12.3%)	1062 (13.1%)	911 (13.2%)
Unheralded coronary death	240 (7.4%)	1119 (5.5%)	895 (5.4%)	436 (5.4%)	400 (5.8%)
Heart failure	545 (16.9%)	2391 (11.8%)	1761 (10.6%)	794 (9.8%)	733 (10.6%)
Ventricular arrhythmia or sudden cardiac death	60 (1.9%)	288 (1.4%)	215 (1.3%)	91 (1.1%)	93 (1.3%)
Transient ischaemic attack	311 (9.6%)	1931 (9.6%)	1538 (9.3%)	704 (8.7%)	592 (8.6%)
Ischaemic stroke	223 (6.9%)	1418 (7.0%)	1082 (6.5%)	542 (6.7%)	425 (6.1%)
Stroke not further specified	307 (9.5%)	1638 (8.1%)	1181 (7.1%)	558 (6.9%)	507 (7.3%)
Subarachnoid haemorrhage	38 (1.2%)	246 (1.2%)	155 (0.9%)	74 (0.9%)	64 (0.9%)
Intracerebral haemorrhage	91 (2.8%)	432 (2.1%)	303 (1.8%)	114 (1.4%)	120 (1.7%)
Peripheral arterial disease	240 (7.4%)	1708 (8.5%)	1576 (9.5%)	799 (9.9%)	743 (10.7%)
Abdominal aortic aneurysm	58 (1.8%)	448 (2.2%)	420 (2.5%)	231 (2.9%)	209 (3.0%)
Total	3232	20198	16562	8093	6919
Other deaths					
Cancers	2132 (54.9%)	6858 (53.7%)	4578 (55.3%)	2105 (53.9%)	1920 (51.1%)
Dementia	167 (4.3%)	682 (5.3%)	370 (4.5%)	204 (5.2%)	182 (4.8%)
Pneumonia	277 (7.1%)	790 (6.2%)	498 (6.0%)	240 (6.2%)	259 (6.9%)
Chronic obstructive pulmonary disease	129 (3.3%)	479 (3.8%)	352 (4.3%)	171 (4.4%)	195 (5.2%)
Liver disease	127 (3.3%)	367 (2.9%)	212 (2.6%)	86 (2.2%)	77 (2.1%)
Other causes of death	1052 (27.1%)	3597 (28.2%)	2267 (27.4%)	1096 (28.1%)	1122 (29.9%)
Total	3884	12773	8277	3902	3755

Figure 8.5: Association of categories of eosinophil count with different initial presentations of cardiovascular disease, by patient's clinical state at time of blood test

Hazard ratios are adjusted for age, sex, deprivation, ethnicity, smoking, diabetes, systolic blood pressure, body mass index, total cholesterol, HDL cholesterol, eGFR, atrial fibrillation, autoimmune conditions, inflammatory bowel disease, COPD, cancer, statin use and blood pressure medication. P values * < 0.05, ** < 0.01, *** < 0.001



Low eosinophil counts were also associated with significantly increased risk of non-cardiovascular death (HR 1.79, 95% CI 1.72, 1.86), without any clear preponderance for a particular cause of death (**Table 8.3 on page 198**). There was no significant difference in these associations between eosinophil counts measured when a patient was 'acute' or 'stable' (**Figure 8.5 on page 199**).

In models adjusted only for age and sex, low eosinophil count was associated with reduced hazard of stable angina (HR 0.83, 95% CI 0.75, 0.92) and peripheral arterial disease (HR 0.81, 95% CI 0.70, 0.92) (**Figure 8.6 on page 201**), but these associations were no longer present with multiple adjustment (**Figure 8.4 on page 197**).

Plots of scaled Schoenfeld residuals showed evidence of non-proportional hazards over time, particularly for the heart failure endpoint (**Figure 8.7 on page 202**). I therefore split the follow-up time at 6 months and found that all associations were stronger in the first 6 months and weak or null thereafter (**Figure 8.8 on page 203**). In the first 6 months the hazard ratio associating low eosinophil counts (<0.05 compared to 0.15–0.25 $\times 10^9$/L) with incident heart failure was 2.05 (95% CI 1.72, 2.43). The corresponding hazard ratio for unheralded coronary death was 1.94 (95% CI 1.40, 2.69), and for ventricular arrhythmia or sudden cardiac death was 3.05 (95% CI 1.48, 6.28) (**Figure 8.8 on page 203**).

Considering linear associations with eosinophil count, there was some evidence that higher stable eosinophil counts were weakly associated with increased risk of non-fatal myocardial infarction (HR per 0.1×10^9 higher eosinophil count 1.02, 95% CI 1.01, 1.04, p = 0.0088) and peripheral arterial disease (HR 1.03, 95% CI 1.01, 1.04, p = 0.003); see **Figure 8.9 on page 204**. These associations were weak relative to the distribution of eosinophil counts in the study population (standard deviation 0.16×10^9/L). Splitting follow-up time revealed a similar picture: higher eosinophil counts were associated with slightly increased risk of non-fatal myocardial infarction or peripheral arterial disease from 6 months onwards (**Figure 8.9 on page 204**).

There were no significant interactions with age (**Figure 8.10 on page 205**) or sex (**Figure 8.11 on page 206**). Results from sensitivity analysis in which patients with a history of loop diuretic use were excluded were consistent with the main results (data not shown). Results obtained using the two different methods of multiple imputation, Random Forest MICE and normal-based MICE, were also very similar (**Figure 8.12 on page 207**).

Figure 8.6: Association of eosinophil categories with different initial presentations of cardiovascular diseases, adjusted for age and sex

P values * < 0.05, ** < 0.01, *** < 0.001

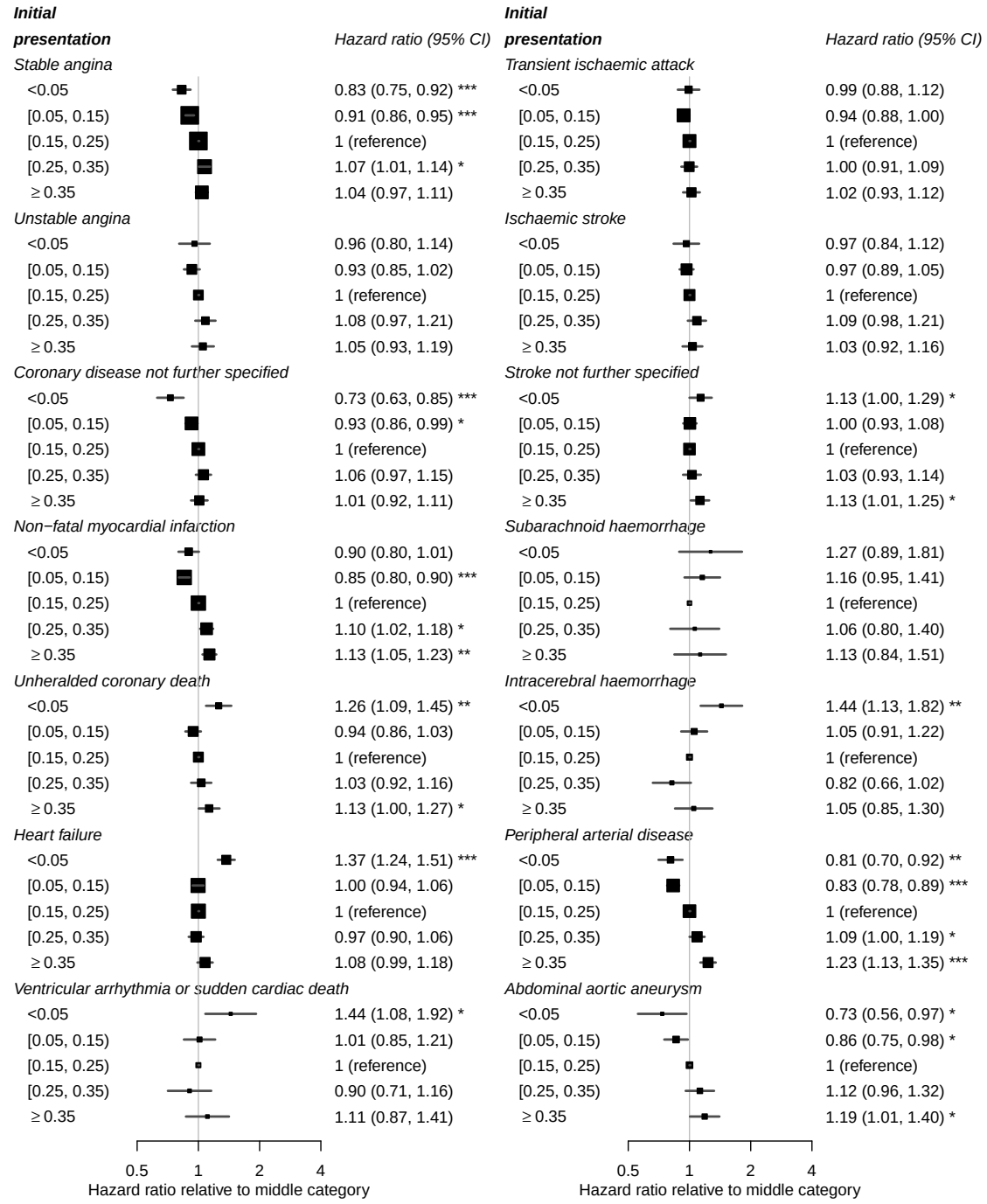


Figure 8.7: Scaled Schoenfeld residuals for adjusted hazard ratios comparing lowest versus middle category of eosinophil count, for different initial presentations of cardiovascular disease

Hazard ratios adjusted for age, sex, deprivation, ethnicity, smoking, diabetes, body mass index, systolic blood pressure, eGFR, HDL, total cholesterol, atrial fibrillation, inflammatory conditions, cancer, statin use, blood pressure medication and acute conditions at the time of blood testing. The lines show the estimate and 95% confidence interval for the beta coefficient. The top left corner of each graph contains ρ, χ^2 and p value for the correlation between log survival time and scaled Schoenfeld residuals. For clarity, points plotted on the graph are for a random sample of 20 000 patients.

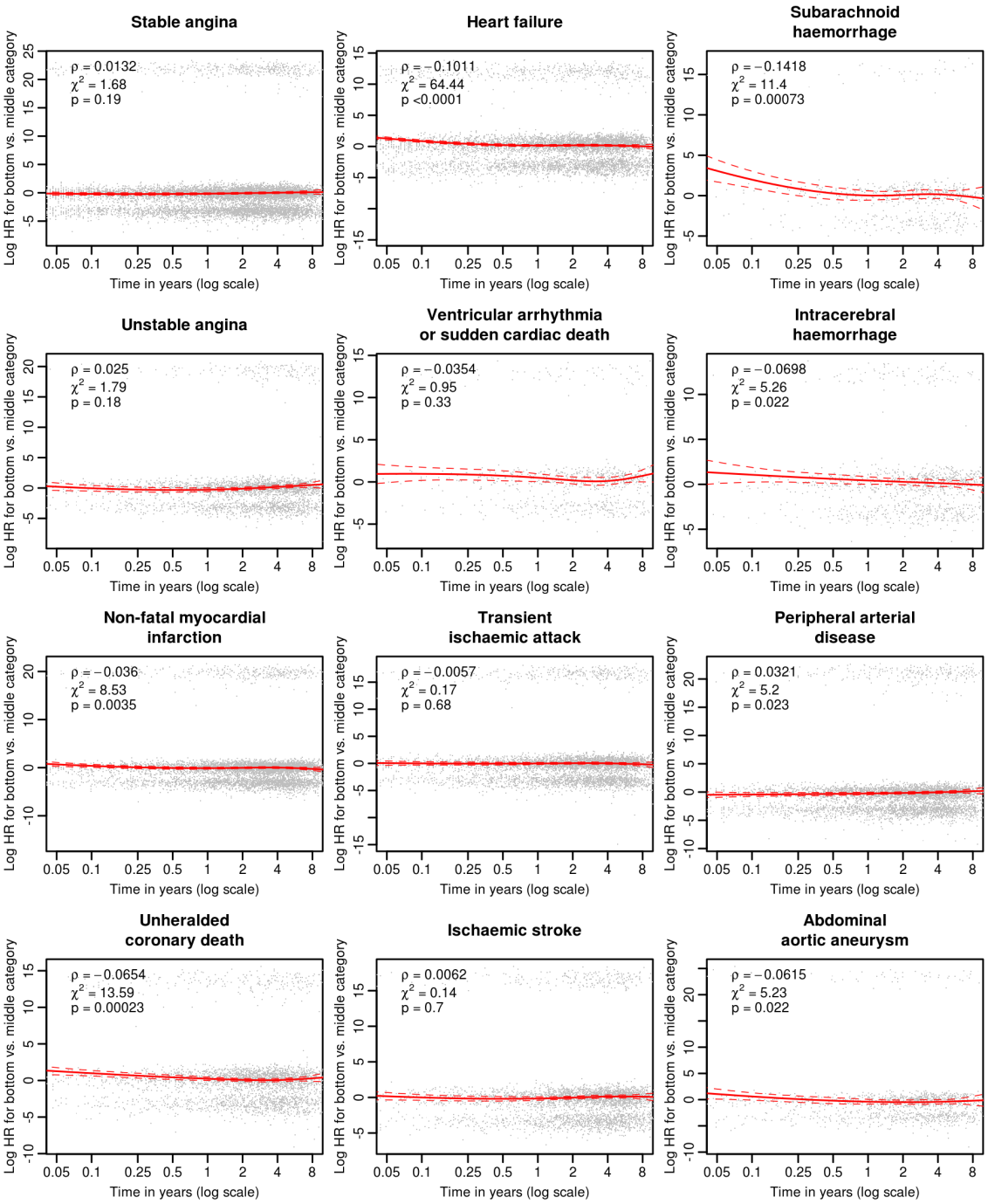


Figure 8.8: Association of eosinophil categories with different initial presentations of cardiovascular disease, by time after measurement

Hazard ratios are adjusted for age, sex, deprivation, ethnicity, smoking, diabetes, systolic blood pressure, body mass index, total cholesterol, HDL cholesterol, eGFR, atrial fibrillation, autoimmune conditions, inflammatory bowel disease, COPD, cancer, statin use, blood pressure medication and acute conditions at the time of blood testing. P values * < 0.05, ** < 0.01, *** < 0.001

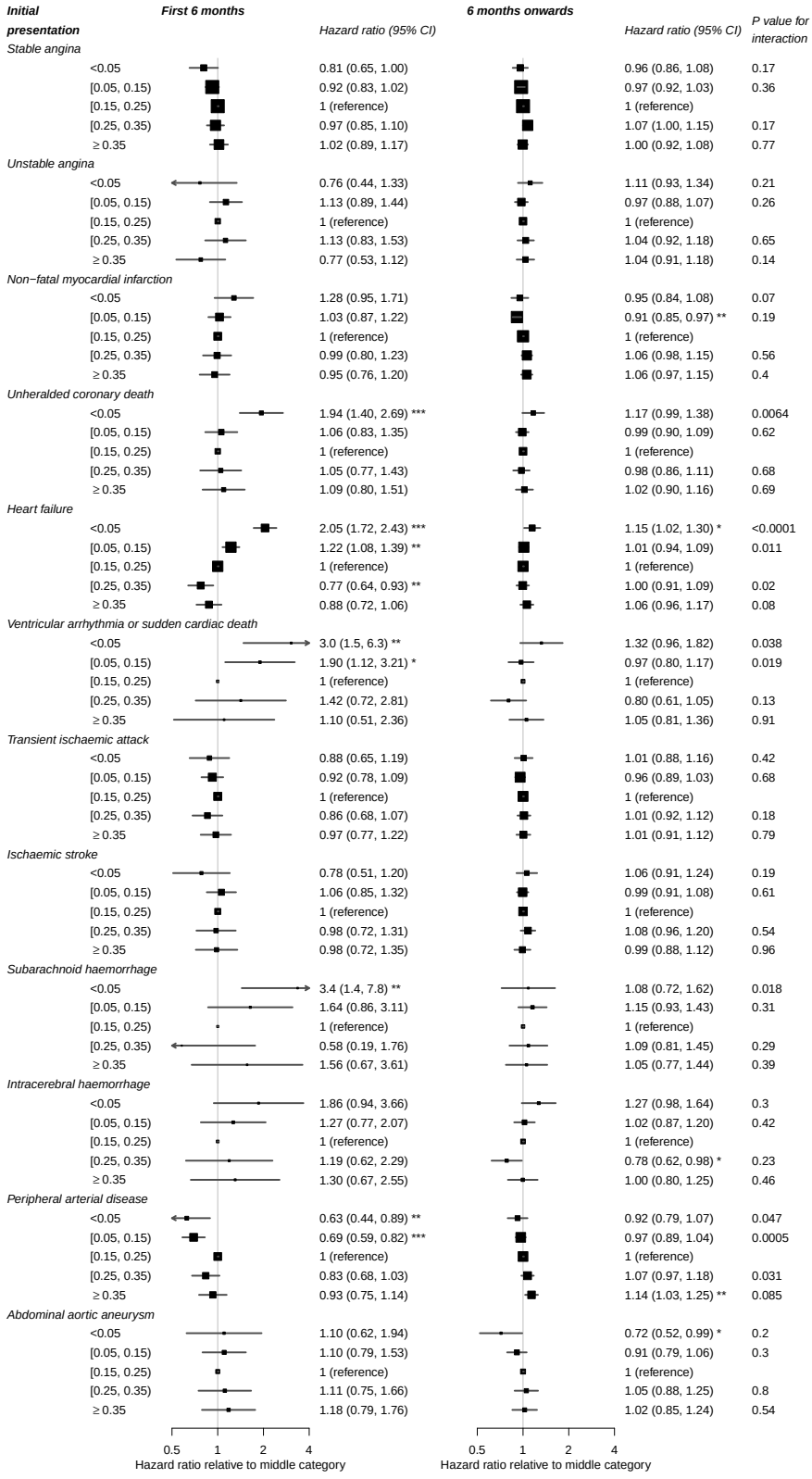


Figure 8.9: Linear association of eosinophil count with different initial presentations of cardiovascular disease, by acuity and time since measurement

Hazard ratios are adjusted for age, sex, deprivation, ethnicity, smoking, diabetes, systolic blood pressure, body mass index, total cholesterol, HDL cholesterol, eGFR, atrial fibrillation, autoimmune conditions, inflammatory bowel disease, COPD, cancer, statin use, blood pressure medication and acute conditions at the time of blood testing. P values * < 0.05, ** < 0.01, *** < 0.001

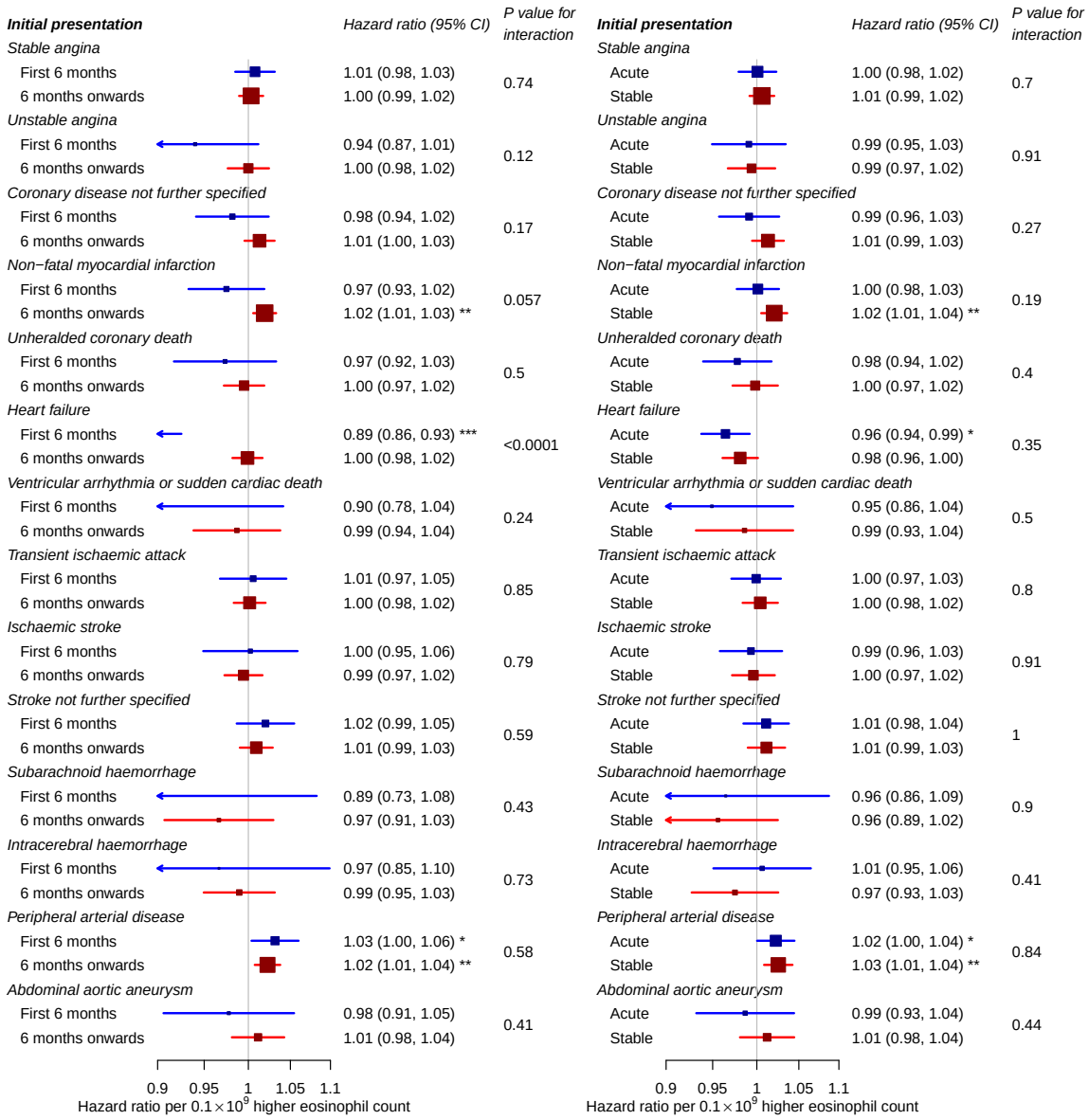


Figure 8.10: Association of categories of eosinophil count with different initial presentations of cardiovascular disease, by age group

Hazard ratios are adjusted for age, sex, deprivation, ethnicity, smoking, diabetes, systolic blood pressure, body mass index, total cholesterol, HDL cholesterol, eGFR, atrial fibrillation, autoimmune conditions, inflammatory bowel disease, COPD, cancer, statin use, blood pressure medication and acute conditions at the time of blood testing. P values * < 0.05, ** < 0.01, *** < 0.001

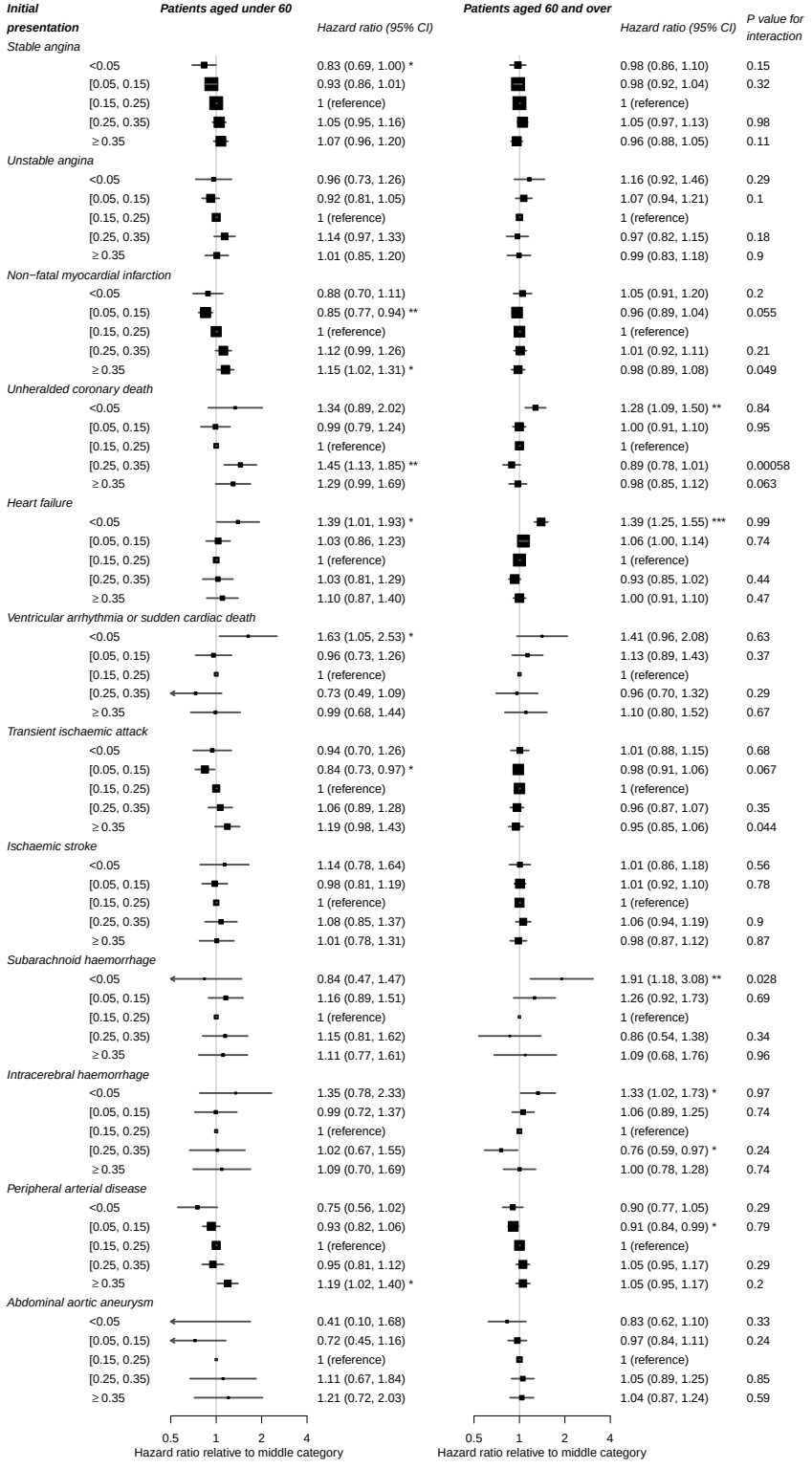


Figure 8.11: Association of categories of eosinophil count with different initial presentations of cardiovascular disease, by sex

Hazard ratios are adjusted for age, sex, deprivation, ethnicity, smoking, diabetes, systolic blood pressure, body mass index, total cholesterol, HDL cholesterol, eGFR, atrial fibrillation, autoimmune conditions, inflammatory bowel disease, COPD, cancer, statin use, blood pressure medication and acute conditions at the time of blood testing. P values * < 0.05, ** < 0.01, *** < 0.001

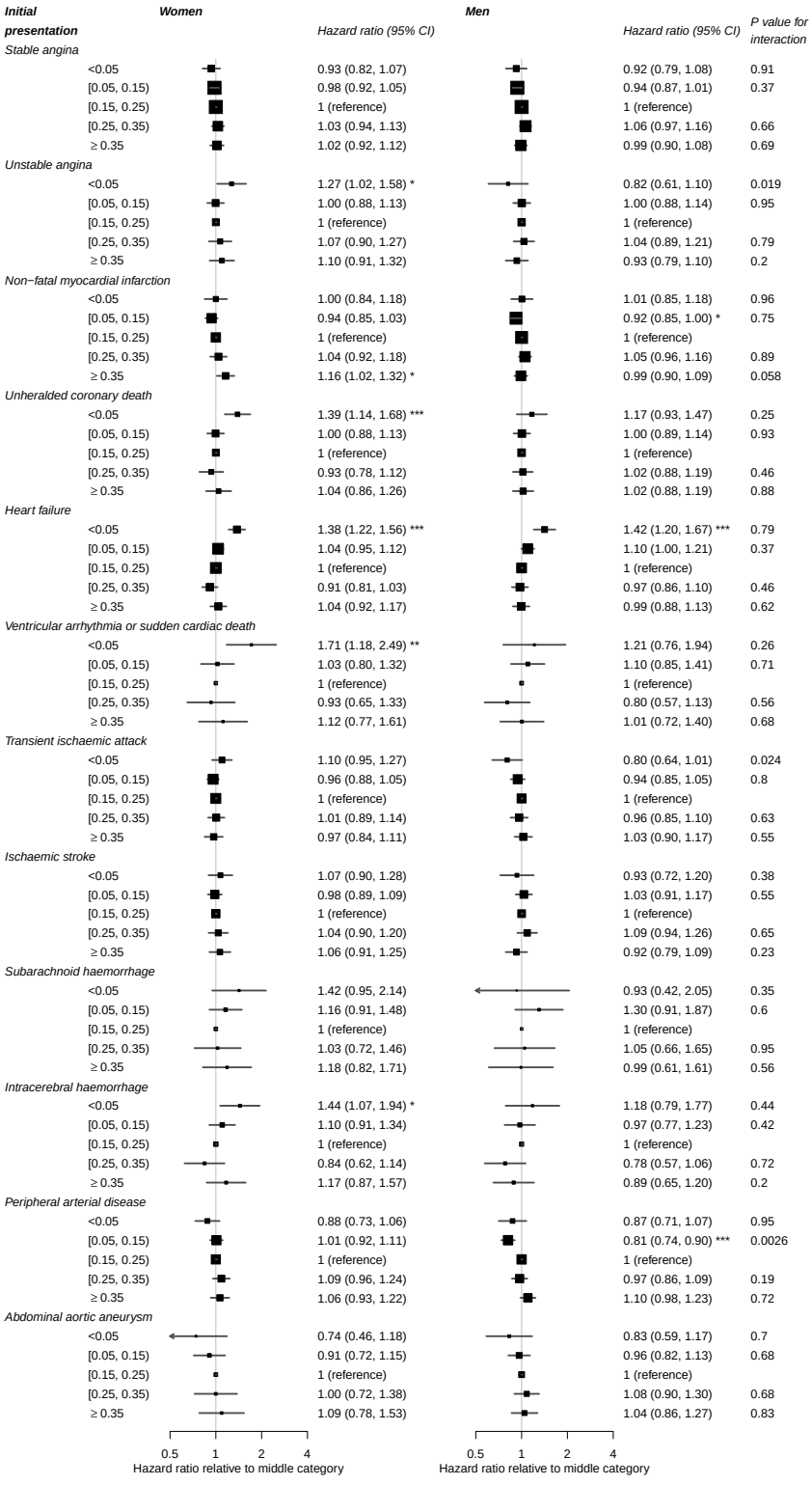
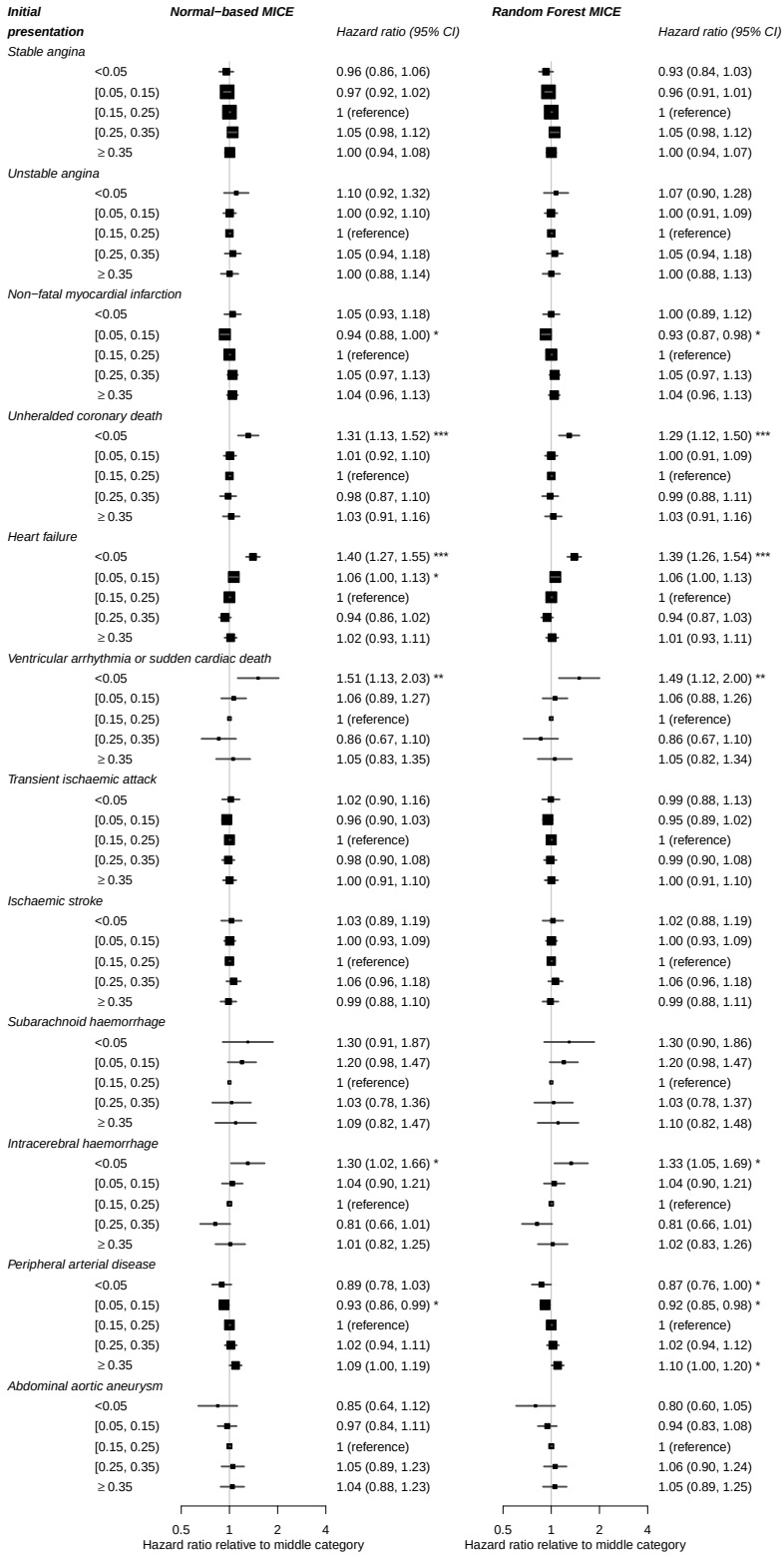


Figure 8.12: Association of eosinophil categories with different initial presentations of cardiovascular disease, using different methods of multiple imputation

Hazard ratios are adjusted for age, sex, deprivation, ethnicity, smoking, diabetes, systolic blood pressure, body mass index, total cholesterol, HDL cholesterol, eGFR, atrial fibrillation, autoimmune conditions, inflammatory bowel disease, COPD, cancer, statin use, blood pressure medication and acute conditions at the time of blood testing. P values * < 0.05, ** < 0.01, *** < 0.001



8.6 Discussion

Low eosinophil counts were associated with increased short-term incidence of heart failure, unheralded coronary death and ventricular arrhythmia or sudden cardiac death as the initial presentation of cardiovascular disease.

As far as I am aware, this is the first time an adverse association of low eosinophil counts on cardiovascular disease has been demonstrated in a healthy population. I also found that low eosinophil counts were associated with increased risk of non-cardiovascular mortality; this is consistent with studies in other settings, such as critical care [241], which showed that low eosinophil counts were predictive of short-term mortality. It may be eosinophils *per se*, or metabolic or immune features associated with low eosinophil counts, that are causally related to these outcomes. For example, endogenous mineralocorticoids suppress eosinophil count (eosinophilia is associated with adrenal insufficiency [242]), and may cause increased incidence of heart failure via sodium retention.

Eosinophil counts are raised in asthma [82], tropical pulmonary eosinophilia [243], helminth infestations [74] and hypereosinophilic syndrome [244], but their main physiological role is thought to be in restoring tissue homeostasis after inflammation [75, 244].

8.6.1 Interleukin-5 and cardiovascular disease

Interleukin-5 (IL-5) is the principal cytokine mediating the release of eosinophils into the bloodstream [80]. It is strongly associated with eosinophil counts in cross-sectional studies [81] and anti-IL5 therapy is being investigated as a therapy for asthma, which is characterised by an overactive immune response and high eosinophil counts [83]. However, IL-5 also has other roles in the immune system, such as the differentiation of B lymphocytes to immunoglobulin-secreting plasma cells [245]. These mechanisms may explain why higher IL-5 is associated with better outcomes in critical care patients [246], and IL-5 administration after the onset of sepsis improves survival in mice [246].

IL-5 may also be relevant to atherosclerosis; a cross-sectional study found that IL-5 was correlated with decreased subclinical atherosclerosis [84] and variants in the IL-5 gene are associated with coronary artery disease [85]. In mouse studies macrophage-specific overexpression of IL-5 led to increased secretion of IgM antibodies that inhibit uptake of oxidized low-density lipoprotein, with consequent inhibition of atherosclerosis [86].

8.6.2 Relation to previous studies

Previous studies have reported a long-term positive association of eosinophil count with coronary disease which was stronger than observed in this study [107, 116] (**Table 2.6 on page 51**); this may be because I adjusted more completely for other cardiovascular risk factors (higher eosinophil count was associated with higher body mass index,

smoking, diabetes and deprivation). Genome wide association studies found that a non-synonymous SNP rs3184504[T] in SH2B3 is associated both with higher eosinophil count and myocardial infarction [79], and is thought to contribute to atherosclerosis through reduced anti-inflammatory activity of SH2B3. However, given that my results did not show a strong association of eosinophil count with myocardial infarction, it seems likely that the causal pathway from rs3184504[T] is not mediated by eosinophils.

8.6.3 Limitations

The limitations of the neutrophil study described in **Section 7.6.3** also apply to this study. An additional limitation regarding the eosinophil count was that it was imprecise, with many results recorded to only one decimal place; this may attenuate associations because of random error (regression dilution bias).

8.6.4 Clinical and research implications

Eosinophil counts are commonly performed in clinical practice, but they tend to inform clinical management only if the results are outside the reference range. Low lymphocyte count, which is known to be associated with worse prognosis of heart failure [247], had a similar pattern of associations to eosinophils (**Figure 7.17 on page 181**). Clinicians should be aware that low eosinophil and lymphocyte counts are associated with increased risk of heart failure and coronary death, and studies should investigate the utility of these biomarkers in predicting adverse outcomes.

I recommend that these findings are replicated in bespoke cohort studies such as UK Biobank [169], with eosinophil counts measured under standardised conditions with concurrent measurement of IL-5. This would provide more precise results (many of the values in this study were measured to only one decimal place), and would also help to elucidate the mechanistic pathway. If the atheroprotective effect of IL-5 is found to be clinically significant, there is the theoretical potential that asthma patients treated with anti-IL-5 may be at increased cardiovascular risk, and this should be monitored in clinical trials and post-marketing surveillance.

8.6.5 Conclusion

Low eosinophil counts were strongly associated with increased short-term incidence of heart failure and coronary death in a healthy population. This may have implications for monitoring the cardiovascular safety of anti-IL-5 therapy for asthma.

8.7 Summary

This study harnesses the large CALIBER dataset to help to clarify the role of eosinophils in cardiovascular diseases, and found an association with low eosinophil counts. Repli-

cation of novel findings such as these in other cohorts is important; the next chapter, chapter 9, is an example of a replication study using CALIBER and the New Zealand PREDICT cohort.

9 Association of total leukocyte count with all-cause mortality: a study of cohorts in England (CALIBER) and New Zealand (PREDICT)

Cardiovascular diseases do not respect national boundaries, and it is important to validate study findings in other populations, preferably in different countries with different population compositions. This study compares the association of total leukocyte count with all-cause mortality in CALIBER and the New Zealand PREDICT cohort. It was a collaborative project with Simon Thornley in New Zealand. We performed analyses independently in the two countries following a similar protocol.

9.1 Abstract

Background: The total white blood cell count (WBC) is a commonly performed blood test in medical practice. Associations with all cause mortality have been suggested, but previous studies have been small or limited to selected populations.

Methods: We performed a cohort study in England and New Zealand using linked electronic health record databases. The English data source was CALIBER (linked primary care records, hospitalisations, mortality and acute coronary syndrome registry). The New Zealand data source was PREDICT (cardiovascular risk assessments in primary care linked to hospitalisations, mortality, dispensed medication and laboratory results). We included people aged 30–75 with no prior cardiovascular disease, and assessed WBC and cardiovascular risk factors at baseline. We followed patients until death, transfer out of the practice (in CALIBER) or study end.

Results: The study included 686 475 individuals in CALIBER and 194 513 individuals in PREDICT. The two highest quintiles of WBC were associated with increased mortality; the adjusted hazard ratio for highest compared to the middle quintile was 1.90 (95% confidence interval (CI) 1.82, 1.99) in CALIBER and 1.53 (95% CI 1.31, 1.79) in PREDICT. The strength of this association was strongest close to the time of WBC measurement (in CALIBER: adjusted hazard ratio 4.1 (95% CI 3.7, 4.6) for the first 6 months, 1.60 (95% CI 1.53, 1.68) thereafter, P for interaction < 0.0001).

Conclusions: Elevated WBC is associated with increased mortality even within the 'normal' range. Replication of this finding in two independent populations increases confidence in our results.

9.2 Introduction

Incomplete recording of clinical characteristics or outcomes may create potential biases. This makes it important to replicate epidemiological findings in other settings. The restricted ethnic diversity of single country studies may limit their generalisability, particularly when investigating traits that vary significantly between ethnic groups.

Reasons for performing international replication or comparison include:

- Replication in cohort with different selection criteria and confounding structure
- Increasing the overall sample size
- Allowing for greater diversity of genotype and ethnicity
- Allowing for differences in geography, development, lifestyle, healthcare system, education

The analyses in chapters 7 and 8 present novel findings about the associations of leukocyte counts with the initial presentation of cardiovascular diseases. It is vital that these findings are replicated in other cohorts in order to verify that they are generalisable and can be potentially used to inform changes to clinical practice. This study is a replication study in a New Zealand cohort which also offers the opportunity to extend the investigation to different ethnic groups, and increase the overall size of the combined study.

The English cohort was CALIBER, as used for the previous studies in this thesis and described in chapter 3. CALIBER patients are predominantly Caucasian, with small proportions of South Asian or Black ethnicities [197] whereas the New Zealand population contains a sizeable proportion of Māori, Pacific, and Chinese people [248]. The PREDICT programme in New Zealand links cardiovascular risk assessments from primary care with hospital admissions, mortality, dispensed medication and laboratory results [249]. The data sources and characteristics of the study populations are shown in **Table 9.1 on page 213**.

Differential leukocyte counts were not available at the time of this project so we compared associations with total leukocyte count, which is similar to neutrophil count (as neutrophils are the most numerous leukocytes in peripheral blood) and has similar associations with other patient characteristics (see **Table 7.1 on page 160** and **Table 7.2 on page 161**). As this was the first collaborative study using PREDICT and CALIBER we chose all-cause mortality as the endpoint. Investigation of individual cardiovascular diseases will require harmonisation of disease definitions in the two countries, and will be the subject of future work.

Previous studies have suggested that higher total white cell count is associated with higher all-cause mortality [250], but there have been few large-scale population-based studies (**Table 9.2 on page 214**). Some studies were limited to government employees and their dependants [109] or survey respondents [114].

This study aimed to replicate associations between total white cell count and all-cause mortality in two ethnically different populations from England and New Zealand.

Table 9.1: Comparison of study populations

	CALIBER (England)	PREDICT (New Zealand)
Data sources		
General practice data	All coded data and some anonymised free text available	Cardiovascular risk assessments
Cardiovascular risk factors	Opportunistic and incomplete recording	Complete data at time of cardiovascular risk assessment
Medication recording	Prescribed medication	Dispensed medication
Hospital admission data	Yes	Yes
Death certificate data	Coded cause of death	Coded and free text cause of death
Acute coronary syndrome registry	Yes, MINAP	Not currently; due to be linked to ANZACS
Coverage		
Coverage of country's population	5% of England	Cardiovascular risk assessment for large proportion of North Island, lab results for Auckland and Christchurch regions
Number of patients	4.7 million	2 million with lab results, 400,000 with cardiovascular risk assessment
Follow-up period	While registered with GP	Until end of recording (PREDICT cannot trace residency)
Time period	GP data from 1990s, hospital data from 2000, cause of death from 2001, MINAP from 2003	2002 onwards, but poor coverage of dispensing data before 2005, so researchers use data from 2005 onwards.
Coding		
Primary care diagnoses	Read version 2	N/A
Hospital diagnoses	WHO ICD-10	ICD-10 Australian modification
Cause of death	WHO ICD-10	ICD-10 Australian modification (6th edition) using WHO rules and guidelines for mortality coding
Procedures	OPCS-4	ICD-10 Australian modification

9.3 Methods

9.3.1 PREDICT study population (New Zealand)

In New Zealand, general practitioners use a standardised web-based tool to assess cardiovascular risk for primary prevention, and the PREDICT study captures this information centrally. PREDICT includes commonly measured risk factors for cardiovascular disease such as smoking status, diabetes, gender, age and systolic blood pressure [255]. Ethnicity was classified into Māori, Asian (e.g. Indian, Pakistani, Bangladeshi, Chinese, Vietnamese, Thai), Pacific (e.g. Fijian, Tongan, Samoan), European or 'Other', according to a standard protocol which prioritises Māori and minority ethnic groups if more than one ethnicity is recorded for a person [256]. These records are linked to national databases of hospital admission data and mortality using an encrypted New Zealand national health index (NHI). Records were also linked with the NZ Pharmaceutical Information database [8], a register of community dispensing, and Testsafe, a repository of laboratory test results for the Auckland and Northland region of the North Island of

Table 9.2: Association of total white cell count and all cause mortality in previous cohort studies with over 2000 participants

Author, year, study	**Population**	**N deaths**	**Adjusted measure of association**
Jee, 2005, Korean Cancer Prevention Study [109]	438 500 government employees, teachers, and their dependants insured with Korea Medical Insurance Corporation	48 757	HR (95% CI) for WBC ≥ 9 vs $< 5 \times 10^9$/L: male non-smokers 1.38 (1.22, 1.56) male smokers 1.14 (1.07, 1.21) female non-smokers 1.20 (0.94, 1.06) female smokers 1.26 (1.12, 1.41)
Margolis, 2005, Women's Health Initiative [112]	66 261 women in observational study	1919	HR 1.52 (95% 1.33, 1.74) for top vs bottom quartile (P for trend <0.001)
Kabagambe, 2011, REGARDS [251]	17 845 men and women in the Reasons for Geographic and Racial Differences in Stroke Study cohort	1062	HR 1.33 (95% CI 1.03, 1.72) for top vs bottom quartile in Whites (P for trend = 0.03) HR 1.13 (95% CI 0.86, 1.50) for top vs bottom quartile in Blacks (P for trend = 0.55)
Ruggiero, 2007, Baltimore Longitudinal Study of Aging [252]	2803 people recruited from Baltimore-Washington, DC	944	HR 1.62 (95% CI 0.92, 2.85) for WBC > 10 vs $3.5 - 6.0 \times 10^9$/L
Tamakoshi, 2007, NIPPON DATA90 [114]	6756 Japanese residents followed up from nationwide random survey	576	RR 1.61 (95% CI 1.07, 2.40) for WBC $9 - 10 \times 10^9$/L vs $4 - 4.9 \times 10^9$/L
Shankar, 2007, Blue Mountains Eye Study [115]	2904 people with no prior cardiovascular disease or cancer	575	RR 1.68 (95% CI 1.35, 2.09) for top vs bottom quartile
De Labry, 1990, Normative Aging Study [253]	2011 healthy men in Boston	183	HR 1.16 (95% CI 1.08, 1.25) per 10^9/L higher WBC
Kim, 2013, Korean elderly cohort [254]	9996 men and women aged 65 or over	118	HR 1.57 (95% 0.88, 2.80) for top vs bottom quartile (P for trend <0.073)

New Zealand [7]. White cell count results were linked to PREDICT information using the encrypted NHI.

This study includes patients assessed between 1 July 2005 and 24 July 2012, as full dispensing, laboratory and outcome data were available for this period. Testsafe contains community and hospital laboratory results from July 2006 onwards; prior to this date only hospital test results were available from this source, or community tests which were also copied to hospital services. The PREDICT project was approved by the national Multi Region Ethics Committee in 2007 (MEC/07/19/EXP).

9.3.2 Inclusion and exclusion criteria

The CALIBER study population was similar to that used for the studies of differential leukocyte count (section 6.4) but with slightly different inclusion criteria to align with the PREDICT study population.

This study included patients aged 30 to 75 years with no prior history of cardiovascular disease and no use of loop diuretics in the previous six months were eligible to enter the study. In PREDICT, prior cardiovascular disease was ascertained by the cardiovascular risk assessment questionnaire and hospitalisation records; in CALIBER prior cardiovascular disease was defined as a recorded diagnosis in primary care or hospitalisation data, or an entry in the acute coronary syndrome registry.

In CALIBER, patients entered the study on the date of the first white cell count measurement after study eligibility. In PREDICT, patients entered the study on the date of the cardiovascular risk assessment. The most recent measure of total white cell count up to five years before or two weeks after cardiovascular risk assessment was chosen. If no white cell count was available during this period, it was considered missing.

9.3.3 Outcomes

The primary outcome of the study was all-cause mortality. Patients who had died were identified by linkage to the relevant national death registry, with follow-up in New Zealand until 26 July 2012 and in CALIBER until 25 March 2010.

9.3.4 Other covariates

Cardiovascular risk factor information such as smoking status, blood pressure and total: high density lipoprotein (HDL) cholesterol ratio were collected in PREDICT as part of the cardiovascular risk assessment and were largely complete, as general practitioners are required to enter all the data to receive a cardiovascular risk prediction. In CALIBER smoking status (current smoker or non-smoker, to match the PREDICT categories) was derived using primary care records prior to study entry, and for continuous risk factors (blood pressure, total cholesterol and HDL cholesterol) the most recent value up to 1 year prior to study entry was used.

Diabetes status was assessed in CALIBER by a diagnosis of diabetes recorded prior to study entry in primary care or in a hospital admission (subsection 5.4.2). In PREDICT, patients were considered to have diabetes at baseline if their general practitioner recorded that they were diabetic in their cardiovascular risk assessment, or if they had been dispensed an oral hypoglycaemic agent or insulin in the six months before assessment, or if they had been hospitalised with a primary diagnosis of diabetes within the previous five years.

9.3.5 Statistical analysis

Analyses were performed on CALIBER and PREDICT data separately using a similar protocol, as we did not have permissions to combine data from the two countries in a single location and analyse them together. We shared R code to ensure that the analyses were as similar as possible in the two countries, although some differences were necessary because of differences in the available data.

The primary analysis used a Cox regression model with time-to-death from baseline enrolment as the outcome. Individuals were considered censored if they reached the end of the study period alive (PREDICT) or transferred out of the practice (CALIBER). The total white cell counts were grouped into quintiles, based on the observed distribution of total white cell count in the CALIBER data. As in the previous two chapters, all category intervals were closed at the lower end and open at the higher end.

The proportional hazards assumption was assessed by plotting the scaled Schoenfeld residuals against log time [257]. If the hazard ratio varied over time, the follow-up time was split into two intervals. Survival models included the following cardiovascular risk factors, chosen *a priori*: age, sex, total: HDL cholesterol ratio, systolic blood pressure, diabetes status, and current smoking status. All analyses were done using R software (version 3.0.2) [157], using the survival [258] and rms [259] packages for Cox regression. Missing covariate data was handled using multiple imputation, with 10 multiply imputed datasets as described below. Secondary analyses included assessment for interactions with age group, sex and ethnicity, which were carried out in CALIBER only as it was the larger of the two datasets.

The main analysis used a Cox proportional hazards approach to model the association between white cell count quintiles and time to death, adjusting for the following cardiovascular risk factors, chosen a priori: age, sex, total:HDL cholesterol ratio, systolic blood pressure, ethnicity, diabetes status, and current smoking status. Quintiles were based on the CALIBER dataset as it was larger than PREDICT. For CALIBER analyses I stratified the baseline hazard of the Cox model by general practice.

9.3.6 Multiple imputation

PREDICT

The baseline population in PREDICT consisted of all individuals undergoing cardiovascular risk assessment in primary care. Cardiovascular risk factors were almost completely observed but about 30% of people did not have a record of total white cell count. As individuals with a record of white cell count measured would be a (possibly biased) sample, we carried out analyses both limited to those with a white cell count measurement and among the entire baseline population, using multiple imputation to handle missing values.

Predictive mean matching for multiple imputation was implemented by aRegImpute in the Hmisc package [260]. This function uses predictive mean matching with flexible additive models, and transforms continuous predictors using restricted cubic splines with 3 knots. Imputation models included all the variables in the substantive model, event indicator and time in the form of the marginal Nelson-Aalen cumulative hazard [193].

Analyses of 10 imputed datasets were combined using Rubin's rules [261]. White cell count quintiles were imputed directly (instead of imputing actual white cell count) in order to avoid imposing linearity on the association between white cell count and any of the predictors.

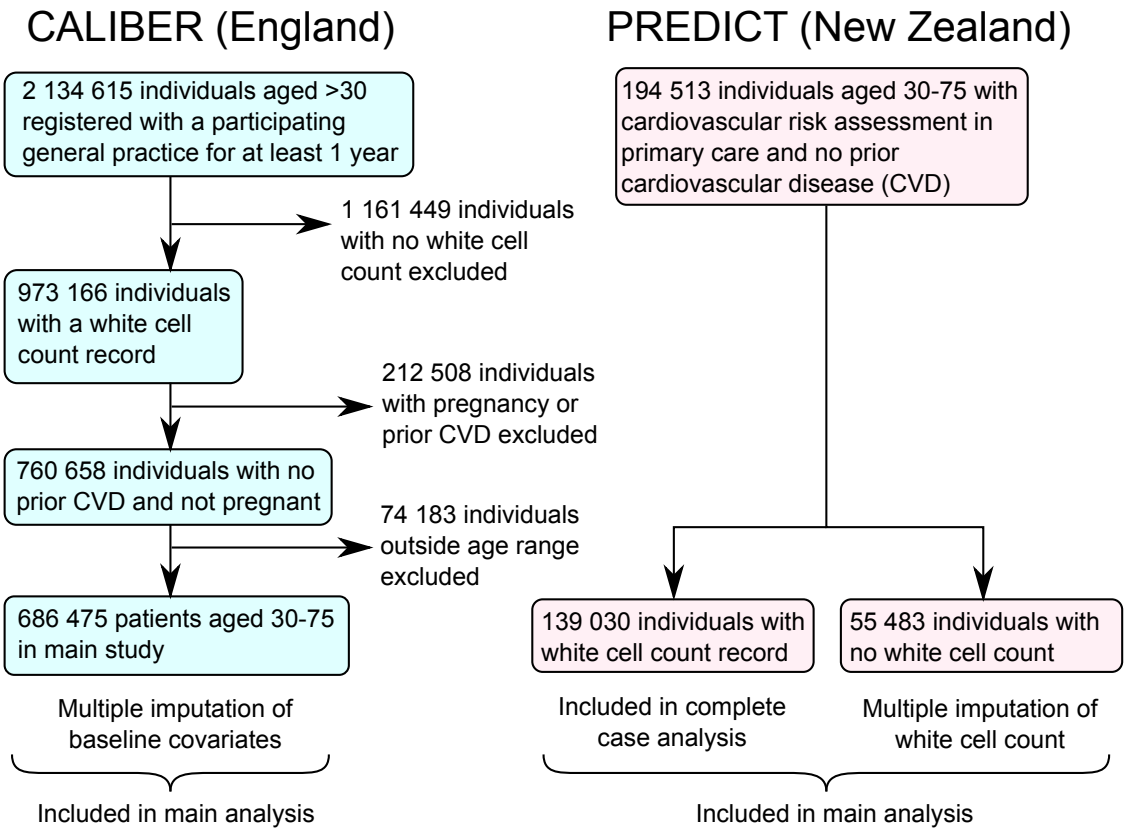
CALIBER

In CALIBER, the cohort consisted of individuals with a white cell count record and no prior history of cardiovascular disease, whether or not they had had a formal cardiovascular disease assessment. Total white cell count data was therefore completely observed but some people did not have records of other baseline covariates: total cholesterol, HDL cholesterol, blood pressure or smoking status. I wished to examine for interactions between total white cell count and predictor variables in their association with mortality, so I chose a method of multiple imputation which would account for these interactions. For the primary analysis I used multiple imputation using chained equations (MICE) [187] with Random Forest multiple imputation [147], as described in section 4.4.

As Random Forest multiple imputation can be slow on very large datasets, I carried out the imputation within general practice, thus also accounting for between-practice variability. I also performed a secondary analysis using normal-based multiple imputation in MICE. With normal-based MICE I included the general practice as a fixed effect (i.e. a categorical variable) and carried out the imputation separately by age group and sex, which accounted for interactions between these variables without requiring too many additional parameters in the imputation models.

Imputation models included all the variables in the substantive model, event indicator and time in the form of the marginal Nelson-Aalen cumulative hazard, as well as the following auxiliary variables: counts of white cell subtypes (neutrophils, basophils, eosinophils, monocytes and lymphocytes), renal function (estimated glomerular filtration

Figure 9.1: Patient flow diagrams for CALIBER and PREDICT studies



rate), and diastolic blood pressure. I used 10 iterations and generated 10 imputations. I verified that the number of iterations was sufficient by inspecting plots of chain means and variances. I combined the results of analysis using Rubin's rules.

9.4 Results

The number of individuals analysed was 686 475 in CALIBER and 194 513 in PREDICT (**Table 9.3 on page 219**). The median age was 50 years in CALIBER and 55 in PREDICT, and 58% were men. The majority of patients in CALIBER were white (92% of those with ethnicity recorded, 383 428 out of 416 828) but in PREDICT just over half were European (104 000 / 194 513, 53%), and significant proportions of individuals belonged to Asian, Indian, Pacific or Māori ethnic groups. In PREDICT 139 030 individuals (71%) had at least one white cell count recorded, and 77% (107 063 / 109 874) of these records were taken within one year before or two weeks after risk assessment. All patients in CALIBER had record of a white cell count (as it was one of the inclusion criteria) (**Table 9.3 on page 219**).

Lower white cell count was associated with Asian or European ethnicity in PREDICT and Black ethnicity in CALIBER, whereas higher white cell count was associated with

Table 9.3: Study populations in England and New Zealand

	CALIBER (England)			PREDICT (New Zealand)		
Characteristics	**Women**	**Men**	**Overall**	**Women**	**Men**	**Overall**
N patients	284 478	401 997	686 475	86 084	108 429	194 513
Age in years, median (IQR)	52 (42, 61)	49 (39, 60)	50 (40, 60)	57 (50, 63)	52 (46, 60)	55 (47, 62)
N (%) with white cell count record *	284 478 (100%)	401 997 (100%)	686 475 (100%)	63 880 (74.2%)	75 150 (69.3%)	139 030 (71.5%)
White cell count ($\times 10^9$/L), median(IQR)	6.6 (5.5, 8.0)	6.7 (5.5, 8.1)	6.6 (5.5, 8.1)	6.6 (5.4, 8.0)	6.7 (5.6, 8.1)	6.6 (5.5, 8.0)
Ethnicity						
N (%) with ethnicity recorded	160 102 (56.3%)	256 726 (63.9%)	416 828 (60.7%)	86 084 (100%)	108 429 (100%)	194 513 (100%)
European (PREDICT) / White (CALIBER)	148 288 (92.6%)	235 140 (91.6%)	383 428 (92.0%)	45 462 (52.8%)	58 538 (54.0%)	104 000 (53.5%)
Asian or Indian (PREDICT) / South Asian (CALIBER)	4810 (3.0%)	8140 (3.2%)	12 950 (3.1%)	14 117 (16.4%)	18 076 (16.7%)	32 193 (16.6%)
Pacific (PREDICT)	N/A	N/A	N/A	12 810 (14.9%)	15 754 (14.5%)	28 564 (14.7%)
Māori (PREDICT)	N/A	N/A	N/A	12 193 (14.2%)	13 777 (12.7%)	25 970 (13.4%)
Black (CALIBER)	3261 (2.0%)	6373 (2.5%)	9634 (2.3%)	N/A	N/A	N/A
Other	3743 (2.3%)	7073 (2.8%)	10 816 (2.6%)	1502 (1.7%)	2284 (2.1%)	3786 (2.0%)
N (%) with smoking status record **	268 766 (94.5%)	385 575 (95.9%)	654 341 (95.3%)	86 084 (100%)	108 429 (100%)	194 513 (100%)
Current smoker, n (%)	74 003 (27.5%)	87 540 (22.7%)	161 543 (24.7%)	12 275 (14.3)	19 518 (18.0)	31 793 (16.4)
N (%) with blood pressure record **	178 209 (62.6%)	259 876 (64.6%)	438 085 (63.8%)	86 083 (100%)	108 429 (100%)	194 512 (100%)
Systolic blood pressure in mmHg, median(IQR)	140 (128, 150)	130 (119, 144)	134 (120, 148)	130 (120, 140)	130 (120, 140)	130 (120, 140)
N (%) with HDL and total cholesterol record	116 184 (40.8%)	103 696 (25.8%)	219 880 (32.0%)	86 078 (100%)	108 423 (100%)	194 501 (100%)
HDL : total cholesterol ratio, median (IQR)	4.4 (3.5, 5.3)	3.6 (2.9, 4.5)	4.0 (3.2, 4.9)	3.6 (2.9, 4.4)	4.3 (3.5, 5.2)	4.0 (3.2, 4.9)
Diabetes at baseline, n (%)	16 219 (5.7%)	12 741 (3.2%)	28 960 (4.2%)	7764 (9.0%)	8882 (8.2%)	16 646 (8.6%)
Deaths during follow-up, n (%)	9961 (3.5%)	9636 (2.4%)	19 597 (2.9%)	892 (1.04%)	1338 (1.23%)	2230 (1.15%)
Follow-up time (years), median (IQR)	3.75 (1.73, 5.97)	4.21 (1.96, 6.42)	4.01 (1.86, 6.23)	2.23 (0.98, 3.78)	2.21 (0.99, 3.86)	2.22 (0.99, 3.83)

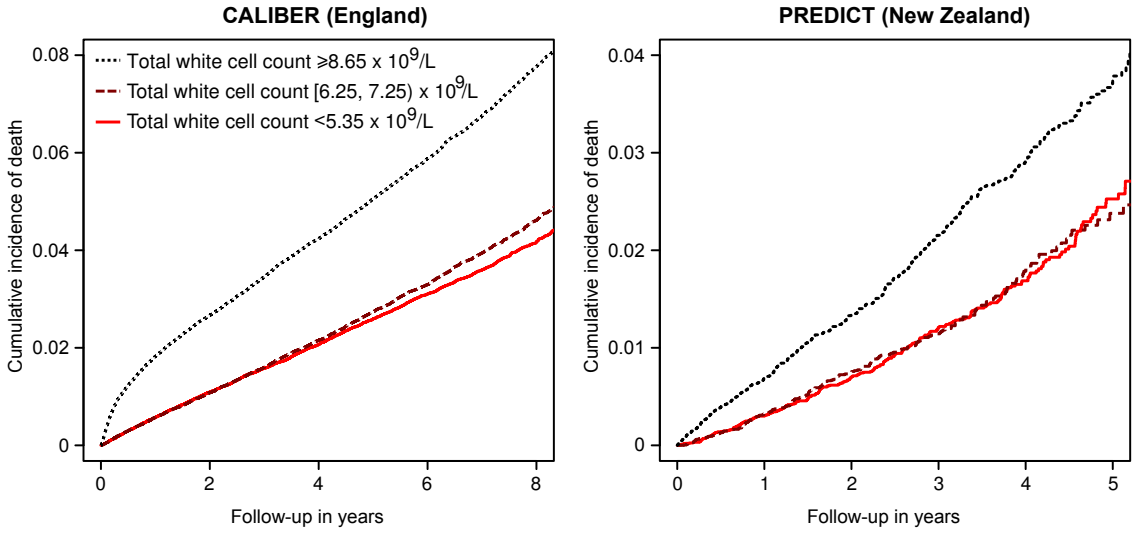
Abbreviations: HDL, high density lipoprotein; IQR, interquartile range; SD, standard deviation.
* In PREDICT the baseline total white cell count was the most recent within 5 years prior to 2 weeks after the cardiovascular risk assessment. In CALIBER the study start date was the date of the white cell count measurement, and patients without any white cell count measurement were excluded.
** The CALIBER analysis used the most recent blood pressure and cholesterol measurements within 1 year prior to study entry. In PREDICT these measurements were taken at the time of the cardiovascular risk assessment.

Table 9.4: Characteristics of study populations by quintiles of white cell count

		Total white cell count range ($\times$ 10^9/L)				
Characteristic	**Population**	**<5.35**	**[5.35, 6.25)**	**[6.25, 7.25)**	**[7.25, 8.65)**	**$\geq$ 8.65**
N patients	CALIBER	148 686	135 107	141 491	133 816	127 375
	PREDICT	29 156	28 312	29 136	27 466	24 960
Men, n (%)	CALIBER	59 148 (39.8%)	58 588 (43.4%)	60 565 (42.8%)	55 429 (41.4%)	50 748 (39.8%)
	PREDICT	14 047 (48.2%)	15 565 (55.0%)	16 267 (55.8%)	15 362 (55.9%)	13 909 (55.7%)
Age, median	CALIBER	51 (41, 60)	51 (41, 61)	51 (41, 61)	49 (40, 60)	47 (38, 59)
(IQR)	PREDICT	56 (50, 63)	56 (48, 63)	55 (47, 62)	53 (46, 62)	52 (45, 60)
Ethnicity, n (%):						
White	CALIBER	78 169 (90.2%)	72 833 (91.8%)	78 597 (92.1%)	76 628 (92.4%)	77 201 (93.4%)
South Asian	CALIBER	1823 (2.1%)	2321 (2.9%)	2910 (3.4%)	3117 (3.8%)	2779 (3.4%)
Black	CALIBER	4112 (4.7%)	1965 (2.5%)	1561 (1.8%)	1106 (1.3%)	890 (1.1%)
European	PREDICT	17 184 (58.9%)	15 849 (56.0%)	14 994 (51.5%)	12 934 (47.1%)	10 785 (43.2%)
Indian	PREDICT	2012 (6.9%)	2554 (9.0%)	3072 (10.5%)	3206 (11.7%)	2645 (10.6%)
Asian	PREDICT	4438 (15.2%)	3072 (10.9%)	2556 (8.8%)	1851 (6.7%)	1213 (4.9%)
Pacific	PREDICT	2780 (9.5%)	3669 (13.0%)	4681 (16.1%)	5178 (18.9%)	5581 (22.4%)
Māori	PREDICT	2056 (7.1%)	2531 (8.9%)	3171 (10.9%)	3725 (13.6%)	4279 (17.1%)
Current smoker,	CALIBER	16 009 (11.3%)	20 415 (15.9%)	29 769 (22.1%)	39 694 (31.1%)	55 656 (45.8%)
n (%)	PREDICT	1661 (5.7%)	2561 (9.1%)	3995 (13.7%)	5423 (19.7%)	7481 (30.0%)
Diabetes, n (%)	CALIBER	4382 (2.9%)	4840 (3.6%)	6071 (4.3%)	6710 (5.0%)	6957 (5.5%)
	PREDICT	1227 (4.2%)	1705 (6.0%)	2470 (8.5%)	3068 (11.2%)	3615 (14.5%)
Systolic blood pressure,	CALIBER	132 (120, 146)	135 (120, 148)	135 (121, 149)	135 (120, 149)	132 (120, 147)
median (IQR)	PREDICT	130 (120, 140)	130 (120, 140)	130 (120, 140)	130 (120, 140)	130 (120, 140)
Total cholesterol: HDL	CALIBER	3.7 (3.0, 4.5)	3.9 (3.2, 4.8)	4.1 (3.3, 5.0)	4.2 (3.4, 5.1)	4.3 (3.5, 5.3)
ratio, median (IQR)	PREDICT	3.7 (3.0, 4.5)	3.9 (3.2, 4.8)	4.0 (3.3, 4.9)	4.1 (3.3, 5.0)	4.1 (3.4, 5.1)

Figure 9.2: Unadjusted Kaplan Meier curves for all cause mortality by total white cell count, in CALIBER and PREDICT

Curves shown for top, middle and bottom quintiles.



South Asian ethnicity in CALIBER, and Māori or Pacific ethnicity in PREDICT. Higher white cell count was associated with current smoking, diabetes and higher total: HDL cholesterol ratio (**Table 9.4 on page 220**).

Median follow-up was 2.2 years in PREDICT and 4.0 years in CALIBER. The total number of deaths was 2230 in PREDICT and 19 597 in CALIBER. Crude Kaplan Meier curves showed a greater cumulative incidence of mortality in the highest white cell count quintile ($\geq$ 8.65 $\times 10^9$/L) in both populations (**Figure 9.2 on page 221**). In CALIBER, the gradient of this curve was steepest immediately after study entry, suggesting that the association between white cell count and mortality was particularly strong in the first few months.

Hazard ratios estimated from Cox models showed a J-shaped association between white cell count and mortality, with the second quintile (5.35 to 6.25 $\times 10^9$/L) associated with the lowest risk. The strength of the association was slightly attenuated by adjustment for cardiovascular risk factors (**Figure 9.3 on page 222**). High white cell count was associated with particularly increased risk of mortality: the multiply adjusted hazard ratio comparing the highest and middle quintiles ($\geq$ 8.65 $\times 10^9$/L vs. 6.25 to 7.25 $\times 10^9$/L) was 1.90 (95% CI 1.82, 1.99) in CALIBER and 1.53 (95% CI 1.31, 1.79) in PREDICT (**Figure 9.4 on page 222**). However, the hazard ratio decreased with time as shown by the plot of scaled Schoenfeld residuals (**Figure 9.5 on page 223**, $p < 0.0001$ for correlation with log time). Splitting the follow-up time by 6 months after the full blood count measurement in CALIBER produced an adjusted hazard ratio of 4.12 (95% CI 3.69, 4.60) for the first 6 months and 1.60 (95% CI 1.53, 1.68) for the subsequent time period (**Figure 9.4 on page 222**), with no evidence of statistically significant non-proportionality of hazards after 6 months. A similar pattern was observed in PREDICT, but with a less extreme difference in hazard ratios between the time periods.

Figure 9.3: Hazard ratios for all cause mortality by quintile of total white cell count in CALIBER and PREDICT

'Multiple adjustment' comprises adjustment for age, sex, total:HDL cholesterol ratio, systolic blood pressure, ethnicity, diabetes and smoking. P values *** <0.001, ** <0.01, * <0.05

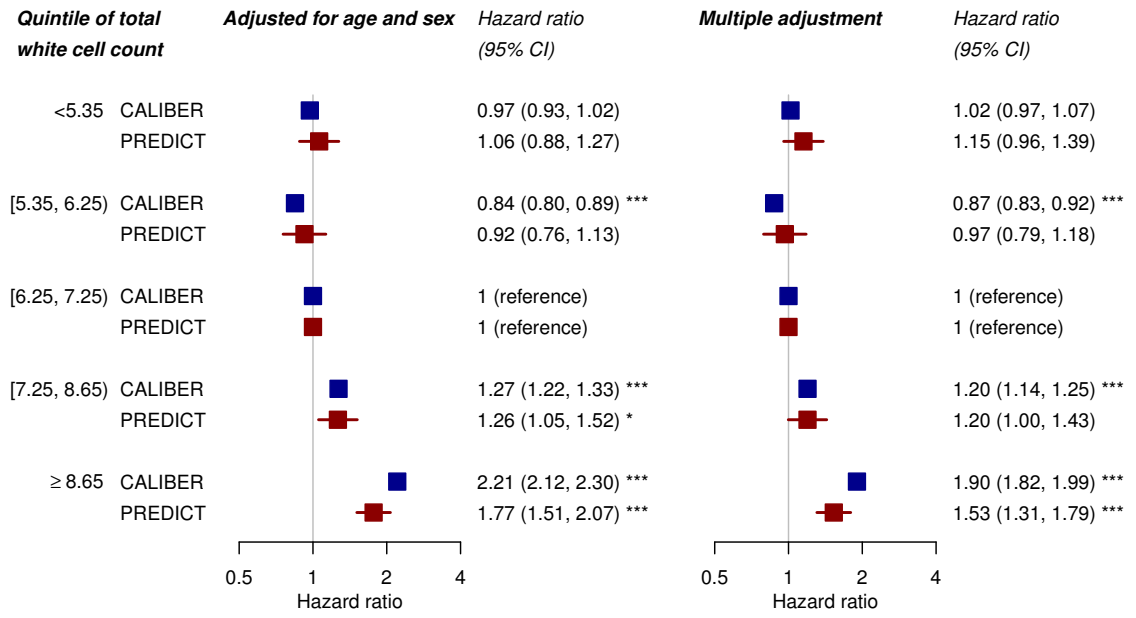


Figure 9.4: Hazard ratios for all cause mortality by quintile of total white cell count, by time period in CALIBER and PREDICT

Hazard ratios were adjusted for age, sex, total:HDL cholesterol ratio, systolic blood pressure, ethnicity, diabetes and smoking. P values *** <0.001, ** <0.01, * <0.05

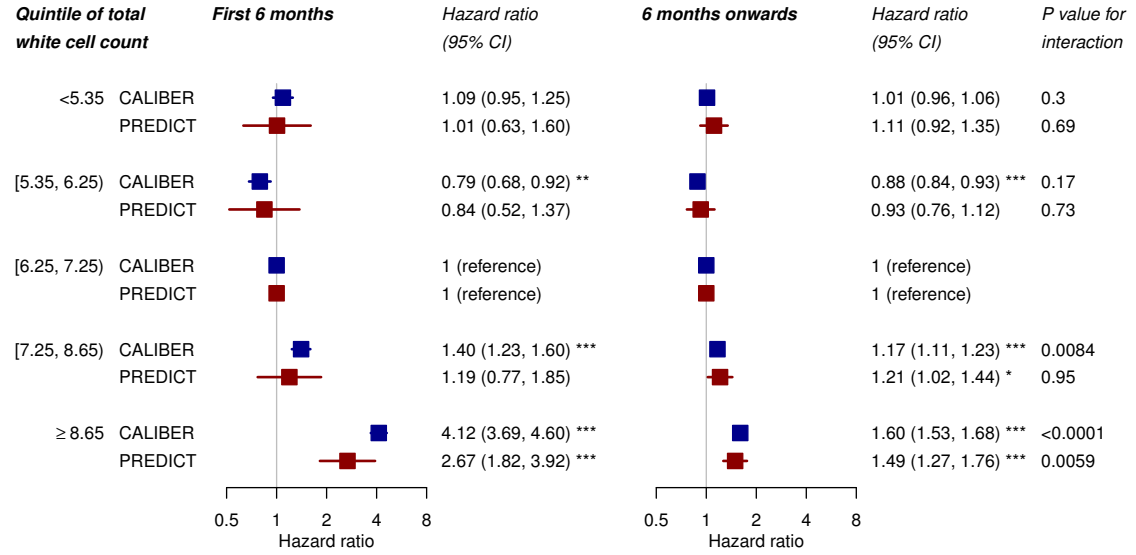


Figure 9.5: Scaled Schoenfeld residuals for multiply adjusted hazard ratio for mortality comparing quintiles of total white cell count in CALIBER

Hazard ratios were adjusted for age, sex, total:HDL cholesterol ratio, systolic blood pressure, ethnicity, diabetes and smoking, and use the middle quintile as the reference category. The lines show the estimate and 95% confidence interval for the beta coefficient. The top left corner of each graph contains ρ, χ^2 and p value for the correlation between log survival time and scaled Schoenfeld residuals. For clarity, points plotted on the graph are for a random sample of 20 000 patients. The middle quintile was the reference category.

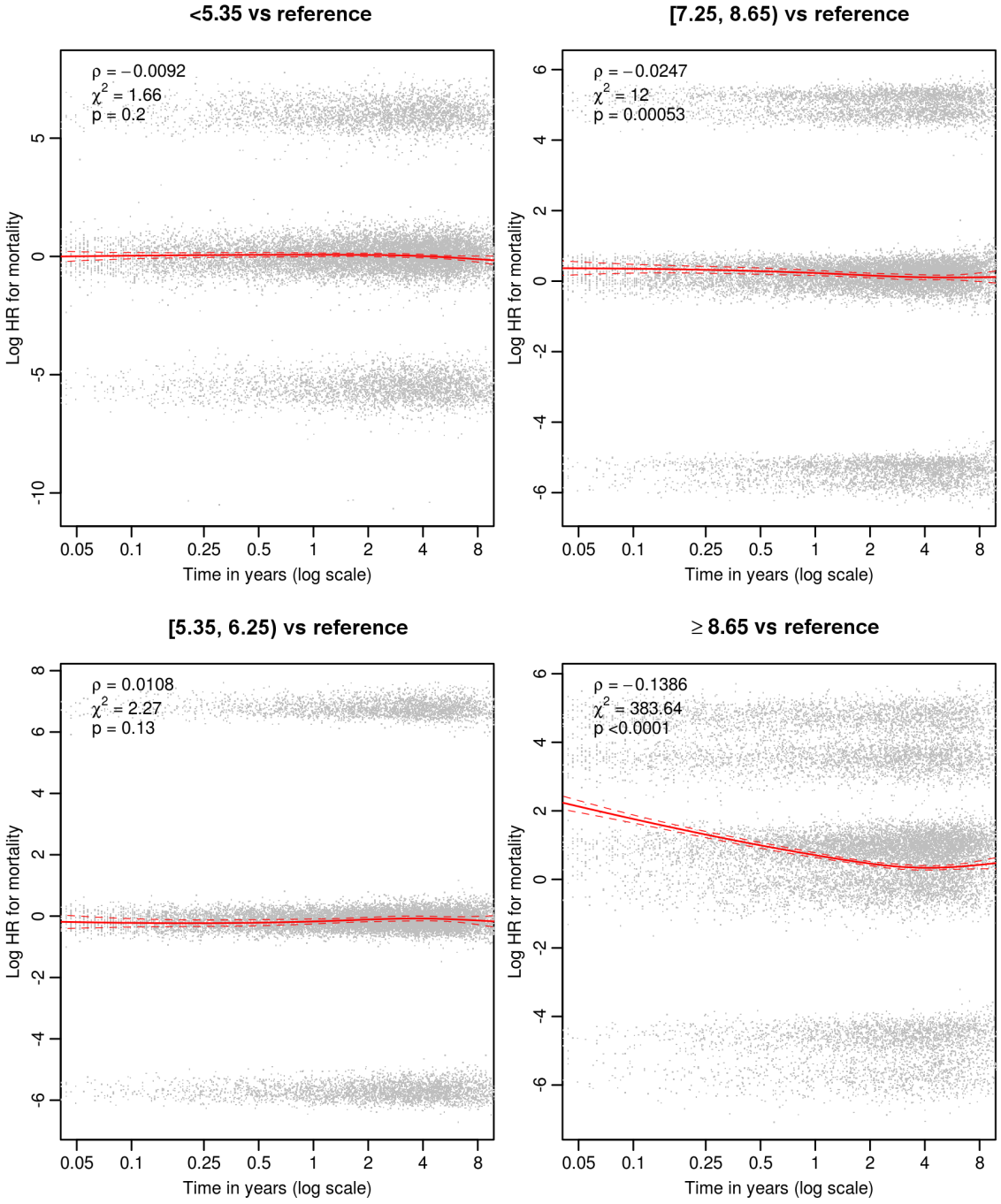
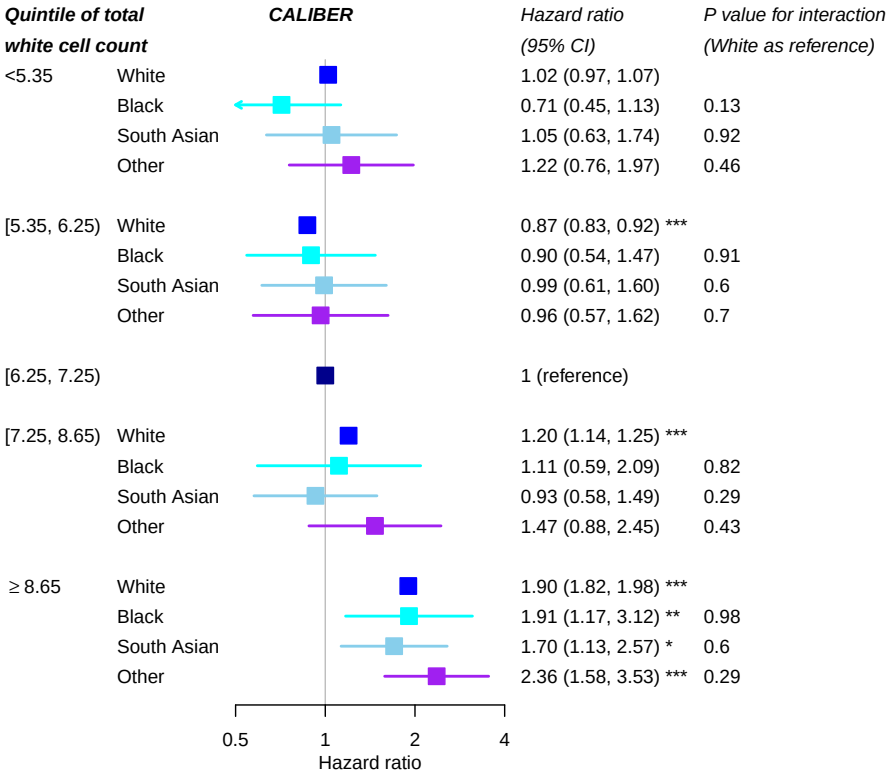


Figure 9.6: Hazard ratios for all cause mortality by quintile of total white cell count, by ethnicity in CALIBER

Hazard ratios were adjusted for age, sex, total:HDL cholesterol ratio, systolic blood pressure, ethnicity, diabetes and smoking. P values *** <0.001, ** <0.01, * <0.05



There were no statistically significant interactions with ethnicity (**Figure 9.6 on page 224**) or sex (**Figure 9.7 on page 225**), although the proportion of ethnic minorities was low, which limits the power of the interaction test with ethnicity. There were statistically significant interactions with age; in the youngest age group (individuals aged 30–45), the lowest quintile of white cell count was associated with increased risk of mortality (adjusted hazard ratio 1.33 for comparison with middle quintile, 95% CI 1.14, 1.54) but there was no significant association for older age groups. Conversely, the association between the highest quintile of white cell count and mortality was weaker in the youngest age group (p = 0.0015 for comparison with 45–60) (**Figure 9.8 on page 225**).

Complete case analysis of PREDICT data yielded similar estimates to the main imputed analysis (**Figure 9.9 on page 226**). The two methods of multiple imputation used in CALIBER data (normal-based and Random Forest MICE) yielded almost identical estimates (**Figure 9.10 on page 226**).

Figure 9.7: Hazard ratios for all cause mortality by quintile of total white cell count, by sex in CALIBER

Hazard ratios were adjusted for age, sex, total:HDL cholesterol ratio, systolic blood pressure, ethnicity, diabetes and smoking. P values *** <0.001, ** <0.01, * <0.05

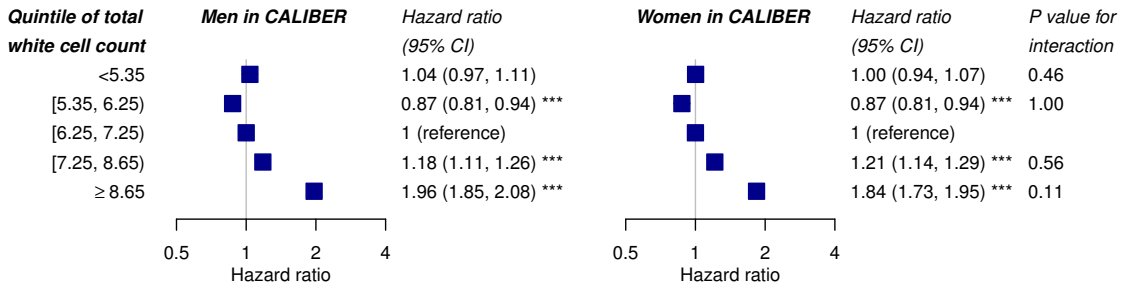


Figure 9.8: Hazard ratios for all cause mortality by quintile of total white cell count, by age group in CALIBER

Hazard ratios were adjusted for age, sex, total:HDL cholesterol ratio, systolic blood pressure, ethnicity, diabetes and smoking. P values *** <0.001, ** <0.01, * <0.05

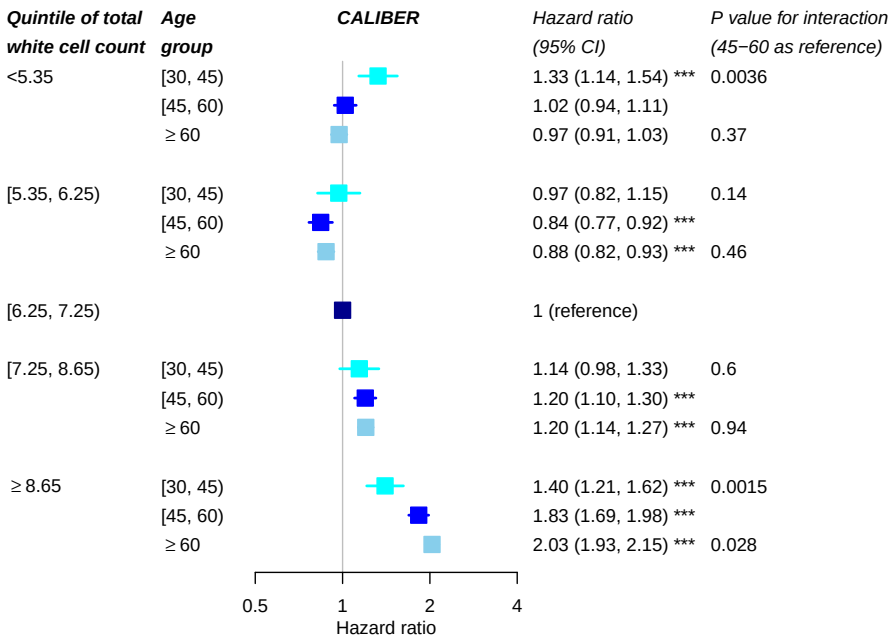


Figure 9.9: Hazard ratios for all cause mortality by quintile of total white cell count for complete case analysis in PREDICT

Hazard ratios were adjusted for age, sex, total:HDL cholesterol ratio, systolic blood pressure, ethnicity, diabetes and smoking. P values *** <0.001, ** <0.01, * <0.05

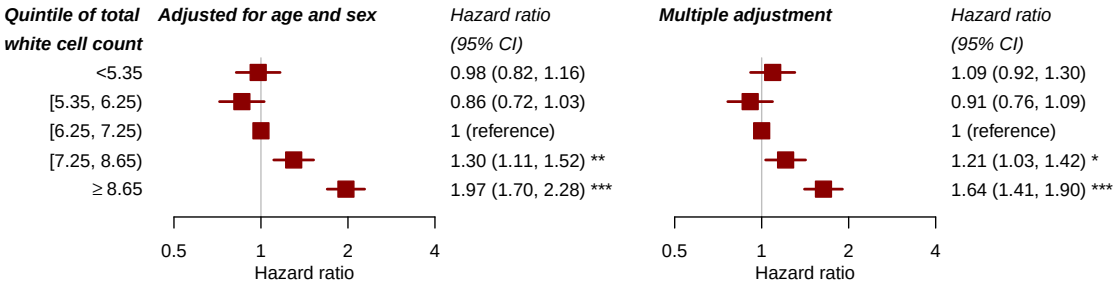
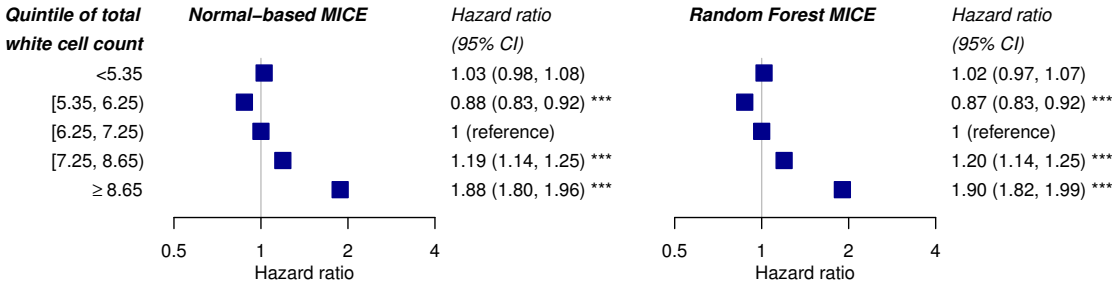


Figure 9.10: Multiply adjusted hazard ratios for all cause mortality by quintile of total white cell count, by method of multiple imputation in CALIBER

Hazard ratios were adjusted for age, sex, total:HDL cholesterol ratio, systolic blood pressure, ethnicity, diabetes and smoking. P values *** <0.001, ** <0.01, * <0.05



9.5 Discussion

Total leukocyte count had a J-shaped relationship with all-cause mortality in two cohorts in different countries. Individuals in the second quintile of white cell count (5.35 to 6.25 $\times 10^9$/L) had the lowest mortality risk, and those in the highest quintile had the highest risk (**Figure 9.3 on page 222**). Mortality risk was increased at 'high normal' levels of total white cell count below the upper limit of laboratory reference ranges. The validity of these findings are increased by the fact that they were observed in two populations with different ethnic compositions looked after by different healthcare systems. This was the first collaborative study between the CALIBER programme in England and PREDICT in New Zealand, and paves the way for future studies to improve electronic health record research by international collaboration.

The strength of the association between white cell count and mortality was strongest close to the time of white cell count measurement, suggesting that it reflects the patient's acute medical state. However, there was a significant but weaker association which persisted for several years after measurement, suggesting that the white cell count is also a useful long-term prognostic marker (**Figure 9.4 on page 222**).

These findings corroborate previous studies which have found an association between total white cell count and mortality or cardiovascular events [253, 262, 263]. Total white cell count is a marker of inflammation, and there are a number of putative biological mechanisms which may explain the inflammation. One potential source is periodontitis [264], which is linked to higher white cell count [265,266]. Cigarette smoking and diabetes are both associated with higher white cell count and dental disease [267, 268]. The increased risk of mortality observed in the lowest quintile of may be because of frailty or malnutrition, which were not recorded in our dataset.

The main strengths of this study were the fact that two cohorts were available in different countries. Both cohorts were large and population based, avoiding the selection bias inherent in many bespoke cohorts with low response rates. The studies differed in the method of patient selection and timing of white cell count measurement relative to study entry. In CALIBER the study population consisted of all patients undergoing a full blood count; this meant there was no missing white cell count data but some covariate data were missing. In PREDICT the study population consisted of all individuals undergoing cardiovascular risk assessment in primary care, with complete recording of cardiovascular risk factors.

9.5.1 Limitations

A limitation of both studies is that white cell count was measured only when it was thought to be clinically necessary, so the probability that a patient has a white cell count depends on a range of factors such as the patient's health-seeking behaviour and the doctor's tolerance of uncertainty or propensity to investigate a patient, as well as the patient's

clinical condition. Replication of this study in a bespoke investigator-led cohort such as UK Biobank [169] would address these issues.

Other weaknesses of the study relate to the variation in time from recording of white cell count to cardiovascular assessment (in PREDICT), and the amount of missing data. Complete case analysis assumes there is no selection bias caused by the removal of patients with incomplete records. Multiple imputation relies on the missingness mechanism being Missing at Random, i.e. whether a value is missing is independent of the value itself conditional on all observed covariates [187, 193]. However, analyses in the two cohorts (with different types of information missing) and analyses using different methods for handling missing data yielded similar results.

Finally, as this is an observational study it is can be used to demonstrate association but not causation. It is possible, though unlikely, that residual confounding may account for the association observed between total white cell count and mortality.

9.5.2 Conclusions

This collaborative study yielded similar results in cohorts based on routine electronic health records in England and New Zealand. In both countries total white cell count had a 'J'-shaped relationship with overall mortality, with similar hazard ratios. Replication of studies in different countries can be challenging but also extremely useful for generalising study results to a wider population.

10 Discussion

The Discussion of this thesis is divided into two sections. Firstly, I discuss the use of electronic health record databases for research, and how national initiatives such as the Farr Institute can facilitate their use for patient benefit. Secondly, I discuss the clinical and research implications of the association of differential leukocyte counts with incidence of cardiovascular diseases, and what lessons can be learned for similar studies investigating biomarkers in electronic health records.

10.1 Using large electronic health record databases for research

10.1.1 Using linked general practice records for research

The work in this thesis illustrates the process of preparing raw structured clinical data for statistical analysis. Patient data are increasingly being collected electronically worldwide [269], creating a rich resource for research, but there is much work involved in creating a research-ready dataset from clinical data.

A primary hurdle is the gathering of data from multiple sources. For the work in this thesis I benefited from the resource built up by the Clinical Practice Research Datalink (CPRD), which has been collecting pseudonymised research data from general practices for many years [4]. Individual hospitals can have hundreds of incompatible information systems, but over 90% of general practices in England use one of 3 clinical software systems [132]. All these general practice systems share a terminology system for recording medical information, the Read codes [136]. The completeness and quality of information recorded in GP systems has been improving over time, and most practices are now paperless. A key advantage of general practice databases over other clinical data sources for research is that healthy people are registered with a general practice, providing a meaningful follow-up period and denominator for the calculation of incidence and prevalence (although prison, army and homeless populations are not included).

There are a number of limitations to the general practice data in CPRD which I encountered, some of which may be overcome with new initiatives.

Population coverage The current coverage of CPRD is 6.9% of the population of the UK [129]. While this has been adequate for many studies on disease epidemiology when used alone or linked to national datasets (such as CALIBER), it is of limited use when linking to detailed datasets which do not cover the entire population, such as UK Biobank [169] or local hospital datasets. In such cases the overlap between CPRD and the second data source may be too small to be useful. The incomplete coverage of CPRD also means

it is impossible to use it to calculate estimates of morbidity for small areas, which would be immensely useful for local healthcare planning and service evaluation.

There are other projects which collect primary care data from all general practices, including the National Diabetes Audit [11], but only a limited subset of information is extracted. There are other research databases in the UK (QResearch, `www.qresearch.org` and ResearchOne, `www.researchone.org`) which collect data from other GP systems, and if used in combination with CPRD they could significantly increase coverage. For an individual researcher it is prohibitively expensive and time-consuming to seek access to multiple primary care datasets when a single dataset suffices for most projects, and there is no unified portal or access scheme. Studies that have used more than one general practice database have generally done so only for the purpose of validation of risk prediction algorithms [270]. Wider collaborations may need higher level facilitation, such as Public Health England's Burden of Disease project or the Farr Institute. In the longer term, national projects such as the care.data programme `www.hscic.gov.uk/gpes/caredata` will compile a single uniform dataset from all general practices for audit and research.

Recommendation:

- Improve the population coverage of general practice databases.

The Read terminology. Read codes have the advantage that they are used throughout general practice and convey much of the meaning of clinical entries without requiring a detailed information model. However, the Read terminology does have limitations which need to be taken into account in interpreting the data [136]. Read codes were developed for clinical use to enable information to be stored more efficiently; therefore they encode clinical statements that a general practitioner might want to record, but unlike ICD-10 they are not strictly hierarchical (there are broad chapters but many terms are in unordered lists) and there are many synonymous terms giving a wide variety of ways that the same diagnosis can be coded. There is no consistent way to record relationships between terms, or qualifiers such as laterality or negation. Read will be superseded by SNOMED-CT (a combination of Read Clinical Terms Version 3 and SNOMED) but this change in terminology system will not, on its own, solve this problem [271]. SNOMED-CT contains far too many terms to be used for data entry as is; it is often better to use information models to store contextual information (as described in section 10.1.2) rather than construct complicated multi-part codes.

A suggested solution is to develop and curate subsets of the terminology in use (e.g. SNOMED-CT in the NHS, possibly others such as ICD-11 or ICD-10-CM elsewhere). These subsets should be developed for particular clinical use cases in different settings, such as in general practice or a cancer multidisciplinary team meeting, and curated by the relevant professional group. This will facilitate data entry at the point of care. In parallel, information models (archetypes), as described later (section 10.1.2), should be developed by clinical groups in order to standardise data collection for clinical care, research

and audit across healthcare sites and clinical systems.

Recommendations:

- Develop terminology subsets for data entry in different situations.
- Develop better user interfaces (possibly with real-time natural language processing) to encourage appropriate coding.

Lack of secondary care linkages. Secondary care records (e.g. hospital inpatient, outpatient, and mental health records) contain rich information on diseases and investigations which is not captured in administrative sources or disease registries, but this has not yet been linked to primary care for research purposes; even for clinical care the integration between GP and hospital data is operational in only a few sites. Lack of standardisation of data formats in different hospitals is one of the problems, and can only be addressed in the long term by using open information models such as openEHR archetypes [272].

Recommendations:

- Develop hospital research databases linked to primary care.
- Encourage standardisation of secondary care information systems.

Information models used in general practice systems. CPRD primarily records clinical 'events' such as diagnoses, prescriptions or laboratory results. Although this is suited to events related to a single time point, such as tests or prescriptions, it is unclear how chronic conditions should be recorded; there is no structured way for representing onset, duration or exacerbations. A problem-orientated medical record, in which problems or diagnoses are recorded once in a specific part of the record, may help, as long as doctors use the system in the way it is intended. Event-based records in CPRD are sometimes misused by repeatedly recording a diagnosis for a consultation related to that diagnosis (e.g. myocardial infarction).

Although clinical terminologies are standardised, the information models (archetypes) which describe how data elements are organised to describe clinical entities, are created *ad hoc* by system manufacturers. The openEHR programme seeks to standardise information models used in electronic health record systems in order to facilitate interoperability and ensure the integrity of patient data [272]. Clinical professional groups should be closely involved in specifying the information structures for storing clinical information in their domain (see section 10.1.2).

Recommendation:

- Clinical professional groups should be involved in specifying information structures for storing clinical information in their domain.

Lack of free text in primary care research databases. Much of the electronic information in healthcare records is stored as free text rather than in a structured form [273], and is difficult to access for research. There has been intense interest in developing computer algorithms to extract coded information from free text [274]. Natural language processing can be used to classify patients [275] or assign diagnostic codes [276], but the results are not reliable enough to be used in clinical practice. Ideally, real-time natural language processing could be used to assist data entry by clinicians, as recommended in the NHS Common User Interface guidelines [277]. This would allow clinicians to select appropriate diagnosis codes without having to be intimately familiar with the terminology. The codes can be verified by the clinician at the point of data entry, improving the accuracy of the records. There are a few examples of such systems (e.g. `http://clinithink.com/`), but they are not yet in general use.

There is a potential risk of identifying patients in research databases if names and addresses are included in the free text. Free text is collected by CPRD but not routinely made available to researchers (except in limited quantities after anonymisation), and is not collected by other primary care research databases. Natural language processing tools may be able to tap into this resource, extracting useful information for research from large databases without compromising patient confidentiality [276].

However, it is possible that CPRD may not be permitted to collect free text at all in the future. If this occurs it will limit the potential for using free text for case validation or extracting further information about clinical events in general practice databases.

Potential solutions:

- Individual patient consent for free text data extraction – this may be possible for disease cohorts or with concerted effort in a small set of practices.
- Encode as much data as possible, using software that facilitates coding at the point of data entry.
- Provide a mechanism for natural language processing to take place on the raw patient records with only anonymised data (extracted codes) being extracted for the research database. This would require liaison with the GP system supplier.

Patients moving between practices. As patient data in CPRD were pseudonymised at the practice level, individuals would not be recognised as being the same person if they subsequently re-registered at another practice. At the current level of coverage of CPRD (6.9%) this is not too much of a problem, but if research is performed using much larger

proportions of the population, it will be important to ensure that patients are not double counted.

Recommendation:

- Link patients between healthcare providers to compile a longitudinal view in research databases.

Documentation of the process of data generation. Conversion of raw clinical data to a research-ready format requires knowledge of the clinical domain as well as the health system and health informatics. It is essential to understand what information is unlikely to be captured as well as what is recorded. The phenotyping process used by the CALIBER group [179] brings these disciplines together along with an exploration of the data. Publication of phenotyping algorithms will help to make research more efficient, by reducing duplication of effort, and will hopefully improve the quality of research. If researchers are able to explore the user interface of electronic health record systems used in practice, it may help them understand why information is captured in a certain way. Although CALIBER has its own data portal for publication of phenotyping algorithms (`https://caliberresearch.org/portal`), a more general repository of variable definitions and methods for a range of healthcare datasets would be desirable.

Recommendation:

- The Farr should set up a repository of variable definitions and methods for reproducible EHR research.

Problems with data linkage. When performing research using identifiable clinical data (without explicit patient consent), specific approval from the Confidentiality Advisory Group of the Health Research Authority [144] is required, and researchers must show that it is not practical to seek consent. Identifiable data should be used for the specific parts of the project for which pseudonymised data would not be sufficient. This commonly means that identifiers are used for linkage but are then removed and replaced by pseudo-identifiers for analysis.

A potential solution which does not involve extraction of identifiers from the original data location is to run pseudonymisation software separately in the two datasets to be linked, using an encryption algorithm with a study-specific key ('salt') which is accessible to a minimum number of people or stored within an encryption service, and is not provided to researchers using the linked data. Examples of pseudonymisation software include OpenPseudonymiser (`www.openpseudonymiser.org`). This approach may be suitable if both datasets have strong identifiers (e.g. different databases within a single hospital trust) but there is no possibility to detect or quantify the quality of links or linkage error.

The linkage in CALIBER was deterministic with CPRD as the denominator population and other datasets providing additional information (for example, deaths are recorded

in CPRD but the death registry provides additional information such as a more precise date of death and the cause of death; see subsection 3.4.1). When using linked datasets without a denominator, and existence of a patient in a particular dataset is itself of interest, mismatches and missed matches can be more of a problem. If the probability of a match varies by the factor under investigation this can bias the results; for example in a study comparing ethnic groups, Hispanic people were less likely to be matched by name than White people [17].

The datasets used in CALIBER have a strong identifier, the NHS number, but linkage to other health and social care datasets may have to use probabilistic methods because a unique identifier is not available. This is particularly the case for research involving marginalised or migrant groups.

Recommendation:

- The Farr Institute should develop best practice guidance for data linkage and handling uncertainty in the linkage process.

10.1.2 Improving health record design to facilitate research

Developments in electronic health record systems should seek to improve the quality of information entered (by having an intuitive user interface which makes it easy to enter data correctly) and make it easy to share and extract information (using standardised information models).

Although the terminologies in clinical use (e.g. ICD-10, Read) are standardised, they can be difficult to use in practice without being narrowed down to a small subset for a clinical use case, or without intelligent user interfaces which suggest the appropriate term to the user. Terms require additional data elements to convey clinical information in full, such as whether a diagnosis is confirmed or suspected. No structured information system can represent all the nuances of meaning so there will always be a need for unstructured free text to convey clinical narrative faithfully.

Even for information that is intrinsically structured, such as clinical measurements or scores, the format of the information can vary between information systems. For example, a blood pressure can be recorded as text (e.g. '120/76'), a sequence of two numbers, or two numbers with labels such as 'Systolic' and 'Diastolic', or 'SYS' and 'DIA'. The myriad of ways of recording even a simple clinical measurement such as blood pressure makes it difficult to share health records or to extract usable research data from multiple incompatible EHR systems. The costs of data migration are huge, and loss of information or incorrect translation can be harmful to patient care.

This situation has arisen because the way clinical information is recorded in EHR systems is determined largely by the system vendor. These information models are not developed in an open and transparent way and may have little clinical input. In the current situation it may be too expensive, difficult and disruptive for a healthcare provider to

switch from an inadequate EHR system ('vendor lock-in').

There are a number of initiatives to try to resolve this problem [278]. One key area of work is in healthcare information standards; there are now commonly used standards for the structure of electronic clinical documents and inter-system messages such as Health Level 7 (HL7) versions 2 and 3, HL7 Clinical Document Architecture and Fast Healthcare Interoperable Resources (`www.hl7.org.uk`). However, although these standards dictate the format of the document (e.g. how a laboratory request message should be organised), they do not specify how the clinical information within the document is structured. Standardised documents containing non-standardised clinical information are a slight improvement but still present significant barriers to interoperability.

Therefore clinically-led standardised information models for EHR are essential. OpenEHR 'archetypes' fulfil this role, and are described in the next section [279].

Archetypes

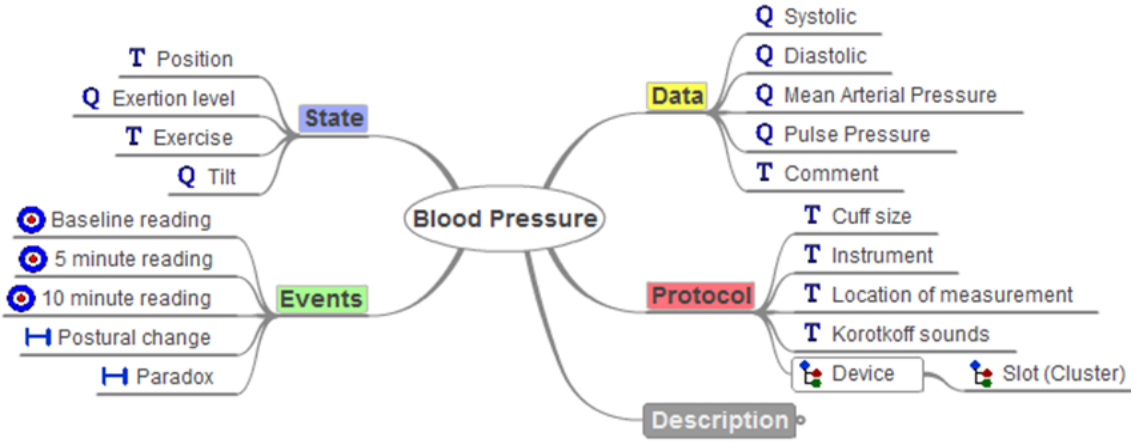
Archetypes are 're-usable formal models of a domain concept' [278], or formal ways of representing an information model, which can make it easier for clinicians and system developers to understand the information how the data need to be organised [279] (**Figure 10.2 on page 237**). For example, an archetype for blood pressure contains data elements for systolic and diastolic pressures (with units and allowable minima and maxima), body site, patient position and device details (see **Figure 10.1 on page 236**). Not all of these data elements need to be used in a particular instance, but because all these extra elements exist in the archetype, they will be clearly defined if they are used in the future. OpenEHR archetypes contain certain administrative data elements by default, such as the time and date, and identity of the healthcare provider entering the data. This means that archetype designers can focus on the clinical information rather than the administrative overhead.

The advantage of recording blood pressure using an archetype is that blood pressure readings can easily be retrieved in the future, whether from ward observations, admission clerkings or clinic notes.

The CPRD uses information models called 'entity types'. These fulfil the function of archetypes but are specific to the GP system Vision, and would have to be mapped individually if records are to be combined between clinical system. If all GP systems use archetypes, the information can be easily retrieved and combined. It also becomes much easier to compile a research database.

By separating the user application from the information model and the physical database, the openEHR archetype model permits EHR systems to be developed in a more flexible way and reduces vendor lock-in. This should improve the quality of healthcare information systems and enable more efficient care, as well as providing better data for answering research questions [279]. It should also give more power to clinicians to shape the data structures of EHR systems so that they faithfully represent cinical prac-

Figure 10.1: Mindmap of the openEHR blood pressure archetype



OpenEHR blood pressure archetype, openEHR-EHR-OBSERVATION.blood_pressure.v1
`http://openehr.org/ckm/`

tice and facilitate audit and research, as suggested in the schematic **Figure 10.2 on page 237**.

10.1.3 Data science

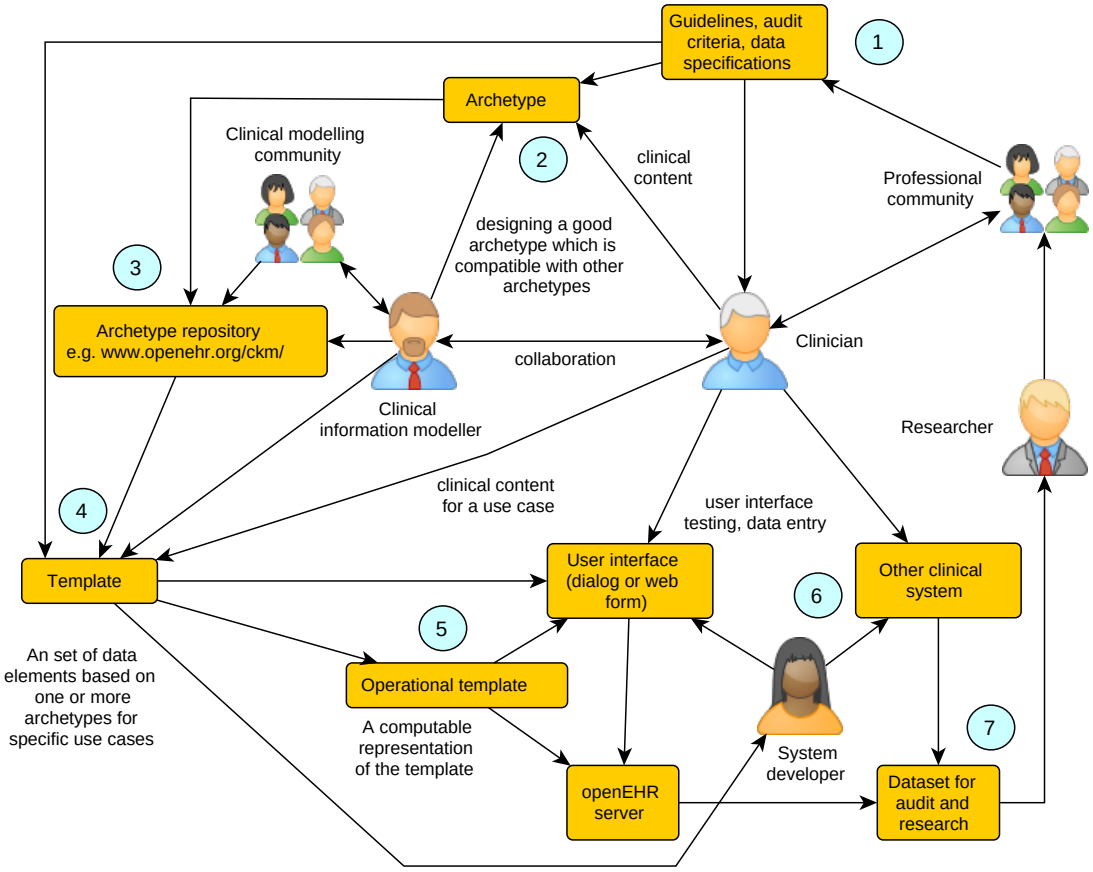
The work in this thesis involved the development of 'big data' approaches to EHR. The open source statistical package R has a well-documented application programming interface and packaging system, so it is possible for individuals to contribute packages to perform functions that are not well supported in base R. Packages that fulfil strict requirements (such as my CALIBERrfimpute package [147]) can be published on the Comprehensive R Archive Network `http://cran.r-project.org/`.

EHR research should seek to emulate the wide range of tools and methods published in bioinformatics. The Farr can use its position as the institute leading EHR research in the UK to set up a methods repository for EHR research. This will enable researchers to share their methods in a way that ensures they receive appropriate credit, and that their methods can be readily reused.

Throughout the work described in this thesis I sought to apply good software engineering practices to data management and analysis programs, to yield re-usable methods and ensure that analyses could be replicated. I used the Git version control system [280] to require me to document changes to my code, and allow me to revert to a previous version at any time, even when the program code is distributed among multiple files. For functions that were frequently used, I formally documented and packaged them into R packages, as described in section 4.3. This required an initial investment of time in making the methods generally applicable, and in testing them with data in different formats. However, it made subsequent work much quicker, and particularly facilitated rapid exploration of new datasets.

Another initiative in reproducible research is to combine documentation and code into

Figure 10.2: Schematic showing how openEHR archetypes can bridge the gap between clinical practice and big data



1. A need to capture certain types of clinical information is established among a clinical community. For example, this might relate to a new type of investigation being introduced, and it is important to evaluate how the outcome of the investigation correlates with outcomes.
2. Clinicians and information modellers develop an archetype for the data elements that could be recorded in this new investigation.
3. The archetype can be reviewed more widely among clinicians and the clinical modelling community, to make sure that it is unambiguous, well structured and well documented.
4. Templates are limited versions of archetype which include only the data elements required in a particular setting, and are designed to be easy to convert into data entry forms.
5. Operational templates can be used to create forms and data structures automatically on openEHR systems.
6. For non-openEHR systems, the template can provide a guide for developers implementing a new data entry form.
7. A common information model between hospitals allows standardised datasets to be extracted for audit and research.

a hybrid document that can carry out an analysis and produce a report. I applied this to the simulation study in the CALIBERrfimpute package [147]. The vignette contains both documentation and code, and when it is re-run by the user, a formatted report is generated with tables and graphs generated from the new data. However, for large analysis it was essential to split the tasks into several steps in order to save time and be able to selectively re-run specific parts of the analysis.

Some computing tasks could be very time-consuming, and I tried to take advantage of the parallel processing capabilities of the CALIBER high performance cluster to perform some tasks at the same time, such as multiple imputation of different subsets of the cohort. Tasks that require the dataset to be considered as a whole, such as fitting statistical models, have generally not been optimised for parallel processing in most statistical software packages. There is much scope for software development to optimise statistical methods for large datasets.

10.1.4 Machine learning

'Machine learning' covers a broad range of techniques which involve a computer program learning a set of rules from a dataset, and then applying it to new data. This can take the form of a prediction model, as an alternative to a conventional statistical model. The Random Forest algorithm, which I used for my novel multiple imputation method, is a machine learning method (section 4.4); some investigators have used it instead of Cox models for modelling survival [281, 282], but this is still uncommon. Machine learning methods can be good at building predictive models from large complex healthcare datasets with many data values per patient, whereas conventional statistical models can only include a limited number of predictor variables.

Machine learning approaches have the disadvantage that the models can be complex and difficult for humans to understand. Statisticians may be unfamiliar with these methods, and there is no mathematical theory that can prove how these models will behave, instead simulation studies are required to learn how the method will perform in different situations. Machine learning methods may also be computationally intensive, but this disadvantage is becoming less important as computing power increases.

10.1.5 Validation and replication of studies

It is good practice to replicate epidemiological studies to show that the conclusions apply more generally, to different populations looked after by different healthcare systems. There have been few examples of replication of studies in electronic health records from different countries. This may arise because of the lack of comparable datasets in different countries and the requirement for international collaborations. However, international replication of studies increases the quality and utility of the research.

As it is generally not possible for entire electronic health record datasets to be transferred between countries, international replication generates additional challenges of en-

suring that analyses are comparable. Sharing analysis code is essential, and meta-analysis techniques may be required to combine results from individual analyses. Publication of phenotypes, dataset descriptions and analysis scripts will facilitate replication of studies.

In this thesis I was able to carry out a replication study in New Zealand data looking at a simple outcome (all cause mortality) (chapter 9). Further work will be required to harmonise the definitions of cardiovascular diseases in England and New Zealand for more detailed studies (for example the coding systems are different; England uses the original WHO version of ICD-10 but New Zealand uses the Australian Modification).

10.2 Differential leukocyte count and onset of cardiovascular diseases

I found that counts of different leukocyte subtypes in peripheral blood had widely differing associations with specific initial presentations of cardiovascular disease. Neutrophil counts within the normal range had strong monotonic positive associations with myocardial infarction, unheralded coronary death, heart failure, and ventricular arrhythmia / sudden cardiac death, but no association with stable or unstable angina. Neutrophil counts were moderately associated with cerebral infarction but not with intracerebral haemorrhage, stable angina or unstable angina. Monocytes showed similar associations with cardiovascular diseases to neutrophils, but less strong. Low counts of lymphocytes were associated with increased heart failure, unheralded coronary death and other death (section 7.6). Low eosinophil counts were associated with increased short-term incidence of heart failure, unheralded coronary death and ventricular arrhythmia or sudden cardiac death (section 8.6).

A recommendation arising from this thesis is that the components of the leukocyte differential should be tested in predictive models for cardiovascular diseases [30, 47]. The total leukocyte count is included in a prognostic model for patients with stable coronary disease [283], but incorporating components of the differential count may enable more accurate predictions. I investigated leukocyte counts in isolation, but the power of predictive models may be increased by combining them in various ways. For example, a combination of high neutrophil count and low lymphocyte count has been found to be a particularly strong predictor of poor prognosis in many diseases, including cancers [284–286], pulmonary embolism [287] and heart failure [288].

10.2.1 Utility of biomarkers in electronic health records for research

The differential leukocyte count is one of many commonly performed clinical measurements which can be researched at scale in electronic health record datasets. Leukocyte counts can be performed for a range of diverse indications, and a strength of electronic health record datasets over bespoke cohort data collections is the longitudinal record of all the important clinical events. This allowed the patient state to be classified, addressing

potential concerns that the predictive value of the leukocyte differential may be affected by the indication for the test.

Statistical methods used for analysis should allow for variation in the timing of covariate recording; in this study I used multiple imputation to fill in missing values using measurements outside the desired time window as auxiliary variables, as described in subsection 7.4.7. This would incorporate the within-person correlation of repeat measurements but also model changes over time and associations with other variables, and make more complete use of the available data.

However, studies do need to take into account the population in which the test is performed, and the likely frequency with which it is performed. Selection bias can limit the generalisability of results from investigating a test that is performed in a limited group of patients. This would be the case for glucose and HbA1c tests – they are generally only performed in patients that the GP suspects of having diabetes, and would be unrepresentative of these measurements in the general population. This is less of a problem when studying a population with pre-existing disease, where the test is routinely recommended in patients with that condition.

Bespoke cohort studies can overcome this type of selection bias if tests are performed in all participants. Such studies may also incorporate information that is not collected clinically, such as genotype information or a comprehensive biomarker assay [169].

However, bespoke studies or surveys with low response rates can be subject to another source of selection bias – independent factors that each increase the probability of a person's agreeing to participate will be over-represented in the cohort, and the combination of such factors will also be over-represented [20]. For example, if diabetic individuals and older individuals are particularly enthusiastic about entering a research cohort, there may be a relative deficit of young non-diabetic people in the cohort. If the cohort is used to make inferences about the general population, it could give a misleading impression that diabetes is common in young people. This is why, unlike population-based cohorts such as CALIBER, consented cohorts with low response rates cannot be used to obtain unbiased estimates of population disease incidence or prevalence [20].

10.2.2 Combining phenotype and genotype information

Large cohorts with detailed genotype and phenotype information will allow new insights into the causal relevance of epidemiological findings. Mendelian randomisation studies take advantage of the random association of chromosomes during meiosis, so presence of an allele associated with a higher level of a biomarker is independent of potential confounders. The allele can therefore be used as an instrumental variable to assess whether higher level of the biomarker is causally related to an outcome, or merely a marker of an underlying condition [289]. This may be helpful for validation of drug targets. For example, Mendelian randomisation studies found that CRP is not causally related to cardiovascular disease [52], so drugs acting on CRP itself may not be clinically useful. CRP is increased

by interleukin-6, which is thought to be causally relevant [290]. Mendelian randomisation studies using SNPs for genes associated with neutrophil count [56, 237] may help to determine whether the neutrophil count itself is causal in atherosclerotic disease.

Genetic studies have typically required large consortia to obtain sufficient statistical power by combining the results of many studies, but large cohorts with genetic information are now being gathered in the UK, and will hopefully accelerate these studies in the future. The UK Biobank (500 000 participants with 100 000 SNP assay) and Genomics England (100 000 whole genome sequences, `www.genomicsengland.co.uk`) have been set up for genetic studies from the outset, and there are other large cohorts such as 50 000 blood donors in the INTERVAL study (a clinical trial of reducing the interval between blood donations) with long term storage of blood samples and participant consent for genetic analyses and linkage to EHR [291].

10.2.3 Clinical and research implications

The strong and diverse associations of leukocyte subtypes with initial presentation of cardiovascular diseases suggests that clinicians should be aware of the significance of these 'routine' tests, and should treat measurements such as the neutrophil count as a continuous measurement. In the same way as doctors inform patients that reducing salt in their diet will reduce their blood pressure, they can also say that stopping smoking and exercising regularly will reduce the inflammation in their body. In clinical trials, smoking cessation reduced the mean neutrophil count by 1.0×10^9/L [239], and moderate regular exercise (8 kcal per kg body weight per week) reduced the mean neutrophil count by 0.1×10^9/L in overweight postmenopausal women [240].

Differential leukocyte counts should be investigated for incorporation into better risk prediction models and clinical decision support algorithms. If these tools run in the electronic health record itself, they can make use of rich clinical information in the system, including repeated measures (for example, I found that the mean of two neutrophil counts was more strongly associated with cardiovascular diseases, see **Figure 7.9 on page 173**). It will likely be more cost-effective to make better use of existing information in the patient record than to carry out additional testing for a novel biomarker.

This thesis provides evidence for the importance of inflammation in the development of atherosclerotic diseases, and therefore the importance of public health measures to combat the causes of this inflammation (such as obesity [292], smoking [239], physical inactivity [240] and air pollution [229]). Inflammatory pathways are under investigation as potential therapeutic targets for cardiovascular diseases; the results of the trials of anti-inflammatory agents such as methotrexate (CIRT) [233] and canakinumab, a human monoclonal antibody that selectively neutralizes interleukin-1β (CANTOS) [234], are awaited.

Finally, this thesis has brought to light the limitations of electronic health record data and the need for better systems that enable information to be recorded easily, accurately,

and contemporaneously. Clinicians will be more willing to enter high quality information if they see a benefit in making it easier to look after their patients. Efforts to improve user interfaces and standardise the clinical content of electronic health records are extremely important [273], both for clinical care and research utility of the data.

10.3 Conclusion

A set of tools were developed to assist the process of converting raw electronic health records into a research-ready format. Counts of leukocyte subtypes are differentially associated with the short and long term risk of different initial presentations of cardiovascular diseases. Leukocyte counts should be investigated for use in cardiovascular risk prediction models, and inflammatory pathways should be further developed as therapeutic targets for cardiovascular diseases, both directly (using medication) and indirectly (by eliminating factors that cause or contribute to inflammation).

Bibliography

[1] 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 Dec;41(6):1625–1638. http://dx.doi.org/10.1093/ije/dys188.

[2] Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012 May;13(6):395–405. http://dx.doi.org/10.1038/nrg3208.

[3] Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med. 2015 Feb;372(9):793–795. http://dx.doi.org/10.1056/NEJMp1500523.

[4] Wood L, Martinez C. The general practice research database: role in pharmacovigilance. Drug Saf. 2004;27(12):871–881.

[5] Sheikh A, Cornford T, Barber N, Avery A, Takian A, Lichtner V, et al. Implementation and adoption of nationwide electronic health records in secondary care in England: final qualitative results from prospective national evaluation in "early adopter" hospitals. BMJ. 2011;343:d6054.

[6] Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med. 2013 Oct;15(10):761–771. http://dx.doi.org/10.1038/gim.2013.72.

[7] TestSafe. New Zealand. http://www.testsafe.co.nz/.

[8] Pharmaceutical Information Database (PHARMS); 2015. http://www.health.govt.nz/publication/pharmaceutical-information-database-data-guide.

[9] Herrett E, Smeeth L, Walker L, Weston C, MINAP Academic Group. The Myocardial Ischaemia National Audit Project (MINAP). Heart. 2010 Aug;96(16):1264–1267. http://dx.doi.org/10.1136/hrt.2009.192328.

[10] Jernberg T, Attebring MF, Hambraeus K, Ivert T, James S, Jeppsson A, et al. The Swedish Web-system for enhancement and development of evidence-based care in heart disease evaluated according to recommended therapies (SWEDEHEART). Heart. 2010 Oct;96(20):1617–1621. http://dx.doi.org/10.1136/hrt.2010.198804.

[11] Gadsby R, Young B. Diabetes care in England and Wales: information from the 2010-2011 National Diabetes Audit. Diabet Med. 2013 Jul;30(7):799–802. http://dx.doi.org/10.1111/dme.12182.

[12] Spencer SA, Davies MP. Hospital episode statistics: improving the quality and value of hospital data: a national internet e-survey of hospital consultants. BMJ Open. 2012;2(6). http://dx.doi.org/10.1136/bmjopen-2012-001651.

[13] Dregan A, Moller H, Murray-Thomas T, Gulliford MC. Validity of cancer diagnosis in a primary care database compared with linked cancer registrations in England. Population-based cohort study. Cancer Epidemiol. 2012 Oct;36(5):425–429. http://dx.doi.org/10.1016/j.canep.2012.05.013.

[14] Harron K, Goldstein H, Wade A, Muller-Pebody B, Parslow R, Gilbert R. Linkage, evaluation and analysis of national electronic healthcare data: application to providing enhanced blood-stream infection surveillance in paediatric intensive care. PLoS One. 2013;8(12):e85278. http://dx.doi.org/10.1371/journal.pone.0085278.

[15] Harron K, Wade A, Gilbert R, Muller-Pebody B, Goldstein H. Evaluating bias due to data linkage error in electronic healthcare records. BMC Med Res Methodol. 2014;14:36. http://dx.doi.org/10.1186/1471-2288-14-36.

[16] Harron K, Wade A, Muller-Pebody B, Goldstein H, Gilbert R. Opening the black box of record linkage. J Epidemiol Community Health. 2012 Dec;66(12):1198. http://dx.doi.org/10.1136/jech-2012-201376.

[17] Lariscy JT. Differential record linkage by Hispanic ethnicity and age in linked mortality studies: implications for the epidemiologic paradox. J Aging Health. 2011;23(8):1263–1284.

[18] 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 5;346.

[19] Goldstein H, Harron K, Wade A. The analysis of record-linked data using multiple imputation with data value priors. Stat Med. 2012 Dec;31(28):3481–3493. http://dx.doi.org/10.1002/sim.5508.

[20] Swanson JM. The UK Biobank and selection bias. Lancet. 2012;380:110.

[21] Dawber TR, Meadors GF, Moore FE. Epidemiological approaches to heart disease: the Framingham Study. Am J Public Health Nations Health. 1951 Mar;41(3):279–281. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1525365/.

[22] Roland M, Guthrie B, Thomé DC. Primary medical care in the United kingdom. J Am Board Fam Med. 2012 Mar;25 Suppl 1:S6–11. http://dx.doi.org/10.3122/jabfm.2012.02.110200.

[23] Gulliford MC, van Staa T, McDermott L, Dregan A, McCann G, Ashworth M, et al. Cluster randomised trial in the General Practice Research Database: 1. Electronic decision support to reduce antibiotic prescribing in primary care (eCRT study). Trials. 2011;12:115. http://dx.doi.org/10.1186/1745-6215-12-115.

[24] Vlug AE, van der Lei J, Mosseveld BM, van Wijk MA, van der Linden PD, Sturkenboom MC, et al. Postmarketing surveillance based on electronic patient records: the IPCI project. Methods Inf Med. 1999 Dec;38(4-5):339–344. http://dx.doi.org/10.1267/METH99040339.

[25] Pont LG, Sturkenboom MC, van Gilst WH, Denig P, Haaijer-Ruskamp FM. Trends in prescribing for heart failure in Dutch primary care from 1996 to 2000. Pharmacoepidemiol Drug Saf. 2003 Jun;12(4):327–334. http://dx.doi.org/10.1002/pds.809.

[26] Vermeer-de Bondt PE, Schoffelen T, Vanrolleghem AM, Isken LD, van Deuren M, Sturkenboom MCJM, et al. Coverage of the 2011 Q fever vaccination campaign in the Netherlands, using retrospective population-based prevalence estimation of cardiovascular risk-conditions for chronic Q fever. PLoS One. 2015;10(4):e0123570. http://dx.doi.org/10.1371/journal.pone.0123570.

[27] de Lusignan S, Metsemakers JF, Houwink P, Gunnarsdottir V, van der Lei J. Routinely collected general practice data: goldmines for research? A report of the European Federation for Medical Informatics Primary Care Informatics Working Group (EFMI PCIWG) from MIE2006, Maastricht, The Netherlands. Inform Prim Care. 2006;14(3):203–209.

[28] Andersen CL, Siersma VD, Karlslund W, Hasselbalch HC, Felding P, Bjerrum OW, et al. The Copenhagen Primary Care Differential Count (CopDiff) database. Clin Epidemiol. 2014;6:199–211. http://dx.doi.org/10.2147/CLEP.S60991.

[29] Andersen CL, Siersma VD, Hasselbalch HC, Lindegaard H, Vestergaard H, Felding P, et al. Eosinophilia in routine blood samples and the subsequent risk of hematological malignancies and death. Am J Hematol. 2013 Oct;88(10):843–847. http://dx.doi.org/10.1002/ajh.23515.

[30] David C Goff J, Lloyd-Jones DM, Bennett G, Coady S, Ralph B D'Agostino S, Gibbons R, et al. 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk: A Report of the American College of Cardiology / American Heart Association Task Force on Practice Guidelines. Circulation. 2014 Jun;129(25 Suppl 2):S49–S73.

[31] Luengo-Fernández R, Leal J, Gray A, Petersen S, Rayner M. Cost of cardiovascular diseases in the United Kingdom. Heart. 2006;92:1384–1389.

[32] 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.

[33] Feigin VL, Lawes CM, Bennett DA, Barker-Collo SL, Parag V. Worldwide stroke incidence and early case fatality reported in 56 population-based studies: a systematic review. Lancet Neurol. 2009 Apr;8(4):355–369.

[34] Bhatnagar P, Wickramasinghe K, Williams J, Rayner M, Townsend N. The epidemiology of cardiovascular disease in the UK 2014. Heart. 2015.

[35] Pujades-Rodriguez M, George J, Shah AD, Rapsomaniki E, Denaxas S, West R, et al. Heterogeneous associations between smoking and a wide range of initial presentations of cardiovascular disease in 1 937 360 people in England: lifetime risks and implications for risk prediction. Int J Epidemiol. 2015 Feb;44(1):129–141.

[36] Rapsomaniki E, Timmis A, George J, Pujades-Rodriguez M, Shah AD, Denaxas S, et al. Blood pressure and incidence of twelve cardiovascular diseases: lifetime risks, healthy life-years lost, and age-specific associations in 1·25 million people. Lancet. 2014 May;383(9932):1899–1911. http://dx.doi.org/10.1016/S0140-6736(14)60685-1.

[37] Swirski FK, Nahrendorf M. Leukocyte behavior in atherosclerosis, myocardial infarction, and heart failure. Science (New York, NY). 2013 Jan;339(6116):161–166. http://www.ncbi.nlm.nih.gov/pubmed/23307733.

[38] Cecelja M, Chowienczyk P. Role of arterial stiffness in cardiovascular disease. J R Soc Med Cardiovasc Dis. 2012 Jul;1:11.

[39] Mayet J, Hughes A. Cardiac and vascular pathophysiology in hypertension. Heart. 2003 Sep;89:1104–1109. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1767863/.

[40] Söderholm M, Zia E, Hedblad B, Engström G. Leukocyte count and incidence of subarachnoid haemorrhage: a prospective cohort study. BMC Neurol. 2014;14(1471-2377 (Linking)):71. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4234394/.

[41] Kosierkiewicz TA, Factor SM, Dickson DW. Immunocytochemical studies of atherosclerotic lesions of cerebral berry aneurysms. J Neuropathol Exp Neurol. 1994 Jul;53(4):399–406.

[42] Cohen JR, Parikh S, Grella L, Sarfati I, Corbie G, Danna D, et al. Role of the neutrophil in abdominal aortic aneurysm development. Cardiovasc Surg. 1993 Aug;1(4):373–376.

[43] Hannawa KK, Eliason JL, Upchurch GR. Gender differences in abdominal aortic aneurysms. Vascular. 2009;17 Suppl 1:S30–S39.

[44] Nordon IM, Hinchliffe RJ, Loftus IM, Thompson MM. Pathophysiology and epidemiology of abdominal aortic aneurysms. Nat Rev Cardiol. 2011 Feb;8(2):92–102. http://dx.doi.org/10.1038/nrcardio.2010.180.

[45] Twig G, Afek A, Shamiss A, Derazne E, Tzur D, Gordon B, et al. White blood cell count and the risk for coronary artery disease in young adults. PLoS ONE. 2012;7(10):e47183.

[46] Barron HV, Harr SD, Radford MJ, Wang Y, Krumholz HM. The association between white blood cell count and acute myocardial infarction mortality in patients $\geq$65 years of age: findings from the Cooperative Cardiovascular Project. J Am Coll Cardiol. 2001;38(6):1654–1661.

[47] Madjid M, Fatemi O. Components of the complete blood count as risk predictors for coronary heart disease: in-depth review and update. Tex Heart Inst J. 2013;40(1):17–29. www.ncbi.nlm.nih.gov/pmc/articles/PMC3568280/.

[48] Hansson GK, Hermansson A. The immune system in atherosclerosis. Nat Immunol. 2011;12(3):204–212.

[49] Wheeler JG, Mussolino ME, Gillum RF, Danesh J. Associations between differential leucocyte count and incident coronary heart disease: 1764 incident cases from seven prospective studies of 30,374 individuals. Eur Heart J. 2004 Aug;25(15):1287–1292. http://dx.doi.org/10.1016/j.ehj.2004.05.002.

[50] Emerging Risk Factors Collaboration, Kaptoge S, Angelantonio ED, Pennells L, Wood AM, White IR, et al. C-reactive protein, fibrinogen, and cardiovascular disease prediction. N Engl J Med. 2012 Oct;367(14):1310–1320. http://dx.doi.org/10.1056/NEJMoa1107477.

[51] Danesh J, Kaptoge S, Mann AG, Sarwar N, Wood A, Angleman SB, et al. Long-Term Interleukin-6 Levels and Subsequent Risk of Coronary Heart Disease: Two New Prospective Studies and a Systematic Review. PLoS Medicine. 2008 Feb;5(4):e78.

[52] C Reactive Protein Coronary Heart Disease Genetics Collaboration (CCGC). Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. BMJ. 2011;342:d548.

[53] Ridker PM. Moving beyond JUPITER: will inhibiting inflammation reduce vascular event rates? Curr Atheroscler Rep. 2013 Jan;15(1):295. http://dx.doi.org/10.1007/s11883-012-0295-3.

[54] Travlos GS. Normal structure, function, and histology of the bone marrow. Toxicol Pathol. 2006;34(5):548–565. http://dx.doi.org/10.1080/01926230600939856.

[55] eMERGE Network Supplemental Genotyping Project – Phenotype Description & Specifications for White Blood Cell Count. eMERGE Network Supplemental Genotyping Project; 2013. http://phenotype.mc.vanderbilt.edu/group/emerge-phenotype-wg.

[56] Crosslin DR, McDavid A, Weston N, Nelson SC, Zheng X, Hart E, et al. Genetic variants associated with the white blood cell count in 13,923 subjects in the eMERGE Network. Hum Genet. 2012 Apr;131(4):639–652. http://dx.doi.org/10.1007/s00439-011-1103-9.

[57] Siminiak T, Flores N, Sheridan D. Neutrophil interactions with endothelium and platelets: possible role in the development of cardiovascular injury. Eur Heart J. 1995;16:160–170.

[58] Borissoff JI, Cate HT. From neutrophil extracellular traps release to thrombosis: an overshooting host-defense mechanism? J Thromb Haemost. 2011 Sep;9(9):1791–1794.

[59] Hansson GK, Libby P. The immune response in atherosclerosis: a double-edged sword. Nat Rev Immunol. 2006 Jul;6(7):508–519. http://dx.doi.org/10.1038/nri1882.

[60] Coller BS. Leukocytosis and Ischemic Vascular Disease Morbidity and Mortality: Is It Time to Intervene? Arterioscler Thromb Vasc Biol. 2005;25(4):658–670.

[61] Eliason JL, Hannawa KK, Ailawadi G, Sinha I, Ford JW, Deogracias MP, et al. Neutrophil depletion inhibits experimental abdominal aortic aneurysm formation. Circulation. 2005 Jul;112(2):232–240. http://dx.doi.org/10.1161/CIRCULATIONAHA.104.517391.

[62] Smith MR, Kinmonth AL, Luben RN, Bingham S, Day NE, Wareham NJ, et al. Smoking status and differential white cell count in men and women in the EPIC-Norfolk population. Atherosclerosis. 2003 Aug;169(2):331–337.

[63] Gkrania-Klotsas E, Ye Z, Cooper AJ, Sharp SJ, Luben R, Biggs ML, et al. Differential white blood cell count and type 2 diabetes: systematic review and meta-analysis of cross-sectional and prospective studies. PLoS One. 2010;5(10):e13405. http://dx.doi.org/10.1371/journal.pone.0013405.

[64] Morris-Rosenfeld S, Lipinski MJ, McNamara CA. Understanding the role of B cells in atherosclerosis: potential clinical implications. Expert Rev Clin Immunol. 2014;10(1):77–89. http://dx.doi.org/10.1586/1744666X.2014.857602.

[65] Ammirati E, Moroni F, Magnoni M, Camici PG. The role of T and B cells in human atherosclerosis and atherothrombosis. Clin Exp Immunol. 2015;179(2):173–187. http://dx.doi.org/10.1111/cei.12477.

[66] Moretta L, Moretta A. Unravelling natural killer cell function: triggering and inhibitory human NK receptors. EMBO J. 2004 Jan;23(2):255–259. http://dx.doi.org/10.1038/sj.emboj.7600019.

[67] Bobryshev YV, Lord RSA. Identification of natural killer cells in human atherosclerotic plaque. Atherosclerosis. 2005 Jun;180(2):423–427. http://dx.doi.org/10.1016/j.atherosclerosis.2005.01.046.

[68] Seltzer MH, Bastidas JA, Cooper DM, Engler P, Slocum B, Fletcher HS. Instant nutritional assessment. J Parenter Enteral Nutr. 1979;3:157–159. http://pen.sagepub.com/content/3/3/157.long.

[69] Vaduganathan M, Ambrosy AP, Greene SJ, Mentz RJ, Subacius HP, Maggioni AP, et al. The Predictive Value of Low Relative Lymphocyte Count in Patients Hospitalized for Heart Failure with Reduced Ejection Fraction: Insights from the EVEREST Trial. Circ Heart Fail. 2012;5:750–758.

[70] Núñez J, Núñez E, Bodí V, Sanchis J, Mainar L, Miñana G, et al. Low lymphocyte count in acute phase of ST-segment elevation myocardial infarction predicts long-term recurrent myocardial infarction. Coron Artery Dis. 2010 Jan;21(1):1–7.

[71] Tacke F, Alvarez D, Kaplan TJ, Jakubzick C, Spanbroek R, Llodra J, et al. Monocyte subsets differentially employ CCR2, CCR5, and CX3CR1 to accumulate within atherosclerotic plaques. J Clin Invest. 2007 Jan;117(1):185–194. http://dx.doi.org/10.1172/JCI28549.

[72] Rogacev KS, Cremers B, Zawada AM, Seiler S, Binder N, Ege P, et al. CD14++CD16+ monocytes independently predict cardiovascular events: A cohort study of 951 patients referred for elective coronary angiography. J Am Coll Cardiol. 2012;60(16):1512–1520.

[73] Swirski FK, Libby P, Aikawa E, Alcaide P, Luscinskas FW, Weissleder R, et al. Ly-6Chi monocytes dominate hypercholesterolemia-associated monocytosis and give rise to macrophages in atheromata. J Clin Invest. 2007 Jan;117(1):195–205. http://dx.doi.org/10.1172/JCI29950.

[74] Nutman TB. Evaluation and differential diagnosis of marked, persistent eosinophilia. Immunol Allergy Clin North Am. 2007 Aug;27(3):529–549. http://dx.doi.org/10.1016/j.iac.2007.07.008.

[75] Lee JJ, Jacobsen EA, McGarry MP, Schleimer RP, Lee NA. Eosinophils in health and disease: the LIAR hypothesis. Clin Exp Allergy. 2010 Apr;40(4).

[76] Kirkeby K, Paudal B. The eosinophil count in the diagnosis and prognosis of myocardial infarction. Acta Med Scand. 1960;168:21–24.

[77] Sakai T, Inoue S, Matsuyama TA, Takei M, Ota H, Katagiri T, et al. Eosinophils may be involved in thrombus growth in acute coronary syndrome. Int Heart J. 2009 May;50(3):267–277.

[78] Atkinson JB, Robinowitz M, McAllister HA, Virmani R. Association of eosinophils with cardiac rupture. Hum Pathol. 1985;16(6):562–8.

[79] Gudbjartsson DF, Bjornsdottir US, Halapi E, Helgadottir A, Sulem P, Jonsdottir GM, et al. Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction. Nat Genet. 2009 mar;41(3):342–347. http://dx.doi.org/10.1038/ng.323.

[80] Khoury P, Grayson PC, Kilon AD. Eosinophils in vasculitis: characteristics and roles in pathogenesis. Nat Rev Rheumatol. 2014 Aug;10(8):474–483.

[81] Silveira A, McLeod O, Strawbridge RJ, Gertow K, Sennblad B, Baldassarre D, et al. Plasma IL-5 concentration and subclinical carotid atherosclerosis. Atherosclerosis. 2015 Mar;239(1):125–130. http://dx.doi.org/10.1016/j.atherosclerosis.2014.12.046.

[82] Tran TN, Khatry DB, Ke X, Ward CK, Gossage D. High blood eosinophil count is associated with more frequent asthma attacks in asthma patients. Ann Allergy Asthma Immunol. 2014 Jul;113(1):19–24. http://dx.doi.org/10.1016/j.anai.2014.04.011.

[83] Garcia G, Taillé C, Laveneziana P, Bourdin A, Chanez P, Humbert M. Anti-interleukin-5 therapy in severe asthma. Eur Respir Rev. 2013 Sep;22(129):251–257. http://dx.doi.org/10.1183/09059180.00004013.

[84] Sämpi M, Ukkola O, Päivänsalo M, Kesäniemi YA, Binder CJ, Hörkkö S. Plasma interleukin-5 levels are related to antibodies binding to oxidized low-density lipoprotein and to decreased subclinical atherosclerosis. J Am Coll Cardiol. 2008 Oct;52(17):1370–1378. http://dx.doi.org/10.1016/j.jacc.2008.06.047.

[85] IBC 50K CAD Consortium. Large-scale gene-centric analysis identifies novel variants for coronary artery disease. PLoS Genet. 2011 Sep;7(9):e1002260. http://dx.doi.org/10.1371/journal.pgen.1002260.

[86] Zhao W, Lei T, Li H, Sun D, Mo X, Wang Z, et al. Macrophage-specific overexpression of interleukin-5 attenuates atherosclerosis in LDL receptor-deficient mice. Gene Ther. 2015 Apr;http://dx.doi.org/10.1038/gt.2015.33.

[87] Majlesi Y, Samorapoompichit P, Hauswirth AW, Schernthaner GH, Ghannadan M, Baghestanian M, et al. Cerivastatin and atorvastatin inhibit IL-3-dependent differentiation and IgE-mediated histamine release in human basophils and downmodulate expression of the basophil-activation antigen CD203c/E-NPP3. J Leukoc Biol. 2003 Jan;73(1):107–117.

[88] Hansson GK, Robertson AKL, Söderberg-Nauclér C. Inflammation and atherosclerosis. Annu Rev Pathol. 2006;1:297–329. http://dx.doi.org/10.1146/annurev.pathol.1.110304.100100.

[89] Ambrose J, Barua R. The pathophysiology of cigarette smoking and cardiovascular disease: An update. J Am Coll Cardiol. 2004;43:1731–1737.

[90] Madjid M, Awan I, Willerson JT, Casscells SW. Leukocyte count and coronary heart disease: Implications for risk assessment. J Am Coll Cardiol. 2004;44(10):1945–1956. http://dx.doi.org/10.1016/j.jacc.2004.07.056.

[91] Stewart RAH, White HD, Kirby AC, Heritier SR, Simes RJ, Nestel PJ, et al. White blood cell count predicts reduction in coronary heart disease mortality with pravastatin. Circulation. 2005;111(14):1756–1762.

[92] 3-hydroxy-3-methylglutaryl-CoA reductase. Online Mendelian Inheritance in Man, John Hopkins University; 2010. http://omim.org/entry/142910.

[93] Mach F. Statins as novel immunomodulators: from cell to potential clinical benefit. Thromb Haemost. 2003 Oct;90(4):607–610. http://dx.doi.org/10.1267/THRO03040607.

[94] Maher BM, Dhonnchu TN, Burke JP, Soo A, Wood AE, Watson RWG. Statins alter neutrophil migration by modulating cellular Rho activity–a potential mechanism for statins-mediated pleotropic effects? J Leukoc Biol. 2009 Jan;85(1):186–193. http://dx.doi.org/10.1189/jlb.0608382.

[95] Guasti L, Marino F, Cosentino M, Maio RC, Rasini E, Ferrari M, et al. Prolonged statin-associated reduction in neutrophil reactive oxygen species and angiotensin II type 1 receptor expression: 1-year follow-up. Eur Heart J. 2008 May;29(9):1118–1126. http://dx.doi.org/10.1093/eurheartj/ehn138.

[96] Hillyard DZ, Cameron AJM, McDonald KJ, Thomson J, MacIntyre A, Shiels PG, et al. Simvastatin inhibits lymphocyte function in normal subjects and patients with cardiovascular disease. Atherosclerosis. 2004 Aug;175(2):305–313. http://dx.doi.org/10.1016/j.atherosclerosis.2004.03.018.

[97] Ferro D, Parrotto S, Basili S, Alessandri C, Violi F. Simvastatin inhibits the monocyte expression of proinflammatory cytokines in patients with hypercholesterolemia. J Am Coll Cardiol. 2000 Aug;36(2):427–431.

[98] McCarey DW, McInnes IB, Madhok R, Hampson R, Scherbakova O, Ford I, et al. Trial of Atorvastatin in Rheumatoid Arthritis (TARA): double-blind, randomised placebo-controlled trial. Lancet. 2004;363(9426):2015 – 2021.

[99] Weber MS, Prod'homme T, Steinman L, Zamvil SS. Drug Insight: using statins to treat neuroinflammatory disease. Nat Clin Pract Neurol. 2005 Dec;1(2):106–112. http://dx.doi.org/10.1038/ncpneuro0047.

[100] Ibáñez L, Jaramillo AM, Ferrer A, de Zegher F. High neutrophil count in girls and women with hyperinsulinaemic hyperandrogenism: normalization with metformin and flutamide overcomes the aggravation by oral contraception. Hum Reprod. 2005 Sep;20(9):2457–2462. http://dx.doi.org/10.1093/humrep/dei072.

[101] Krysiak R, Okopien B. Lymphocyte-suppressing and systemic anti-inflammatory effects of high-dose metformin in simvastatin-treated patients with impaired fasting glucose. Atherosclerosis. 2012 Dec;225(2):403–407. http://dx.doi.org/10.1016/j.atherosclerosis.2012.09.034.

[102] Crittenden DB, Lehmann RA, Schneck L, Keenan RT, Shah B, Greenberg JD, et al. Colchicine use is associated with decreased prevalence of myocardial infarction in patients with gout. J Rheumatol. 2012;39(7):1458–1464.

[103] Nidorf SM, Eikelboom JW, Budgeon CA, Thompson PL. Low-dose colchicine for secondary prevention of cardiovascular disease. J Am Coll Cardiol. 2013;61(4):404–406.

[104] Taglieri N, Bacchi RML, Palmerini T, Cinti L, Saia F, Guastaroba P, et al. Baseline white blood cell count is an independent predictor of long-term cardiovascular mortality in patients with non-ST-segment elevation acute coronary syndrome, but it does not improve the risk classification of the GRACE score. Cardiology. 2013;124(2):97–104.

[105] Prospective Studies Collaboration. Blood cholesterol and vascular mortality by age, sex, and blood pressure: a meta-analysis of individual data from 61 prospective studies with 55 000 vascular deaths. Lancet. 2007;370(9602):1829–1839.

[106] Emerging Risk Factors Collaboration, Kaptoge S, Angelantonio ED, Lowe G, Pepys MB, Thompson SG, et al. C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: an individual participant meta-analysis. Lancet. 2010 Jan;375(9709):132–140. http://dx.doi.org/10.1016/S0140-6736(09)61717-7.

[107] Prentice RL, Szatrowski TP, Fujikura T, Kato H, Mason MW, Hamilton HH. Leukocyte counts and coronary heart disease in a Japanese cohort. Am J Epidemiol. 1982;116(3):496–509. http://aje.oxfordjournals.org/cgi/content/abstract/116/3/4.

[108] Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, et al. Meta-analysis of Observational Studies in Epidemiology: A Proposal for Reporting. JAMA. 2000 apr;283(15):2008. http://dx.doi.org/10.1001/jama.283.15.2008.

[109] Jee SH, Park JY, Kim HS, Lee TY, Samet JM. White blood cell count and risk for all-cause, cardiovascular, and cancer mortality in a cohort of Koreans. Am J Epidemiol. 2005 Dec;162(11):1062–9.

[110] Prentice RL, Szatrowski TP, Kato H, Mason MW. Leukocyte counts and cerebrovascular disease. J Chronic Dis. 1982;35(9):703–714.

[111] Li C, Engstrom G, Hedblad B. Leukocyte count is associated with incidence of coronary events, but not with stroke: A prospective cohort study. Atherosclerosis. 2010;209(2):545–550.

[112] Margolis KL, Manson JE, Greenland P, Rodabough RJ, Bray PF, Safford M, et al. Leukocyte count as a predictor of cardiovascular events and mortality in postmenopausal women: the Women's Health Initiative observational study. Arch Intern Med. 2005;165(5):500–508. http://dx.doi.org/10.1001/archinte.165.5.500.

[113] Pfister R, Sharp SJ, Luben R, Wareham NJ, Khaw KT. Differential white blood cell count and incident heart failure in men and women in the EPIC-Norfolk study. Eur Heart J. 2012;33(4):523–30.

[114] Tamakoshi K, Toyoshima H, Yatsuya H, Matsushita K, Okamura T, Hayakawa T, et al. White blood cell count and risk of all-cause and cardiovascular mortality in nationwide sample of Japanese–results from the NIPPON DATA90. Circ J. 2007 Apr;71(4):479–85.

[115] Shankar A, Mitchell P, Rochtchina E, Wang JJ. The association between circulating white blood cell count, triglyceride level and cardiovascular and all-cause mortality: Population-based cohort study. Atherosclerosis. 2007;192(1):177 – 183.

[116] Sweetnam PM, Thomas HF, Yarnell JW, Baker IA, Elwood PC. Total and differential leukocyte counts as predictors of ischemic heart disease: the Caerphilly and Speedwell studies. Am J Epidemiol. 1997;145(5):416–21.

[117] Lee CD, Folsom AR, Nieto FJ, Chambless LE, Shahar E, Wolfe DA. White blood cell count and incidence of coronary heart disease and ischemic stroke and mortality from cardiovascular disease in African-American and White men and women: Atherosclerosis Risk in Communities Study. Am J Epidemiol. 2001;154(8):758–764.

[118] Engstrom G, Melander O, Hedblad B. Leukocyte count and incidence of hospitalizations due to heart failure. Circ Heart Fail. 2009;2(3):217–222.

[119] Kavousi M, Elias-Smale S, Rutten JHW, Leening MJG, Vliegenthart R, Verwoert GC, et al. Evaluation of newer risk markers for coronary heart disease risk classification: a cohort study. Ann Intern Med. 2012;156(6):438–44.

[120] Olivares R, Ducimetière P, Claude JR. Monocyte Count: A Risk Factor for Coronary Heart Disease? Am J Epidemiol. 1993;137(1):49–53.

[121] Adamsson Eryd S, Smith JG, Melander O, Hedblad B, Engström G. Incidence of coronary events and case fatality rate in relation to blood lymphocyte and neutrophil counts. Arterioscler Thromb Vasc Biol. 2012;32(2):533–9. http://atvb.ahajournals.org/content/32/2/533.long.

[122] Zia E, Melander O, Bjorkbacka H, Hedblad B, Engstrom G. Total and differential leucocyte counts in relation to incidence of stroke subtypes and mortality: A prospective cohort study. J Intern Med. 2012;272(3):298–304.

[123] Shah N, Parikh V, Patel N, Badheka A, Deshmukh A, Rathod A, et al. Neutrophil lymphocyte ratio significantly improves the Framingham risk score in prediction of coronary heart disease mortality: Insights from the National Health and Nutrition Examination Survey-III. Int J Cardiol. 2014;171(3):390–397.

[124] Gillum RF, Mussolino ME, Madans JH. Counts of neutrophils, lymphocytes, and monocytes, cause-specific mortality and coronary heart disease: The NHANES-I epidemiologic follow-up study. Ann Epidemiol. 2005;15(4):266–271.

[125] Karino S, Willcox BJ, Fong K, Lo S, Abbott R, Masaki KH. Total and differential white blood cell counts predict eight-year incident coronary heart disease in elderly Japanese-American men: the Honolulu Heart Program. Atherosclerosis. 2015 Feb;238(2):153–158. http://dx.doi.org/10.1016/j.atherosclerosis.2014.12.003.

[126] Bekwelem W, Lutsey PL, Loehr LR, Agarwal SK, Astor BC, Guild C, et al. White Blood Cell Count, C-Reactive Protein, and Incident Heart Failure in the Atherosclerosis Risk in Communities (ARIC) Study. Ann Epidemiol. 2011;21(10):739–748.

[127] Toor IS, Jaumdally R, Lip GYH, Millane T, Varma C. Eosinophil count predicts mortality following percutaneous coronary intervention. Thromb Res. 2012;130(4):607–11.

[128] Cikrikcioglu MA, Soysal P, Dikerdem D, Cakirca M, Kazancioglu R, Yolbas S, et al. Absolute blood eosinophil count and 1-year mortality risk following hospitalization with acute heart failure. Eur J Emerg Med. 2012;19(4):257–263.

[129] Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol. 2015 Jun;http://dx.doi.org/10.1093/ije/dyv098.

[130] Hospital Episode Statistics. http://www.hscic.gov.uk/hes.

[131] Department for Communities and Local Government. English indices of deprivation 2010: technical report; 2010. https://www.gov.uk/government/publications/english-indices-of-deprivation-2010-technical-report.

[132] Kontopantelis E, Buchan I, Reeves D, Checkland K, Doran T. Relationship between quality of care and choice of clinical computing system: retrospective analysis of family practice performance under the UK's quality and outcomes framework. BMJ Open. 2013;3(8). http://dx.doi.org/10.1136/bmjopen-2013-003190.

[133] Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. Br J Clin Pharmacol. 2010 Jan;69(1):4–14. http://dx.doi.org/10.1111/j.1365-2125.2009.03537.x.

[134] Mathur R, Bhaskaran K, Chaturvedi N, Leon DA, Vanstaa T, Grundy E, et al. Completeness and usability of ethnicity data in UK-based primary care and hospital databases. J Public Health (Oxf). 2013 Dec.

[135] Campbell J, Dedman DJ, Eaton SC, Gallagher AM, Williams TJ. Is the GPRD GOLD population comparable to the UK population? Pharmacoepidemiol Drug Saf. 2013;22(Suppl 1):280.

[136] Smith N, Wilson A, Weekes T. Use of Read codes in development of a standard data set. BMJ. 1995 Jul;311(7000):313–315. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2550372/.

[137] Taylor B, Jick H, Maclaughlin D. Prevalence and incidence rates of autism in the UK: time trend from 2004-2010 in children aged 8 years. BMJ Open. 2013;3(10):e003219. http://dx.doi.org/10.1136/bmjopen-2013-003219.

[138] Mackenzie IS, Morant SV, Bloomfield GA, MacDonald TM, O'Riordan J. Incidence and prevalence of multiple sclerosis in the UK 1990-2010: a descriptive study in the General Practice Research Database. J Neurol Neurosurg Psychiatry. 2014 Jan;85(1):76–84. http://dx.doi.org/10.1136/jnnp-2013-305450.

[139] World Health Organization. International statistical classification of diseases and health related problems: 10th revision, Volume 2. Geneva. http://www.who.int/classifications/icd/en/index.html.

[140] Methodology for creation of the HES Patient ID (HESID). Health and Social Care Information Centre; 2014. http://www.hscic.gov.uk/article/1825/The-processing-cycle-and-HES-data-quality.

[141] Wright FL, Green J, Canoy D, Cairns BJ, Balkwill A, Beral V, et al. Vascular disease in women: comparison of diagnoses in hospital episode statistics and general practice records in England. BMC Med Res Methodol. 2012;12:161. http://dx.doi.org/10.1186/1471-2288-12-161.

[142] Wood DM, Conran P, Dargan PI. ICD-10 coding: poor identification of recreational drug presentations to a large emergency department. Emerg Med J. 2011 May;28(5):387–389. http://dx.doi.org/10.1136/emj.2009.088344.

[143] Rooney C, Griffiths C, Cook L. The implementation of ICD-10 for cause of death coding – some preliminary results from the bridge coding study. Health Stat Q. 2002;13:31–41.

[144] Section 251 and the Confidentiality Advisory Group (CAG). http://www.hra.nhs.uk/about-the-hra/our-committees/section-251/.

[145] Shah AD. CALIBERcodelists: Generate ICD10, Read and OPCS codelists; 2013. R package version 0.2-5. http://caliberanalysis.r-forge.r-project.org/.

[146] Shah AD. CALIBERdatamanage: Data management tools for CALIBER datasets; 2013. R package version 0.1-6. http://caliberanalysis.r-forge.r-project.org/.

[147] Shah AD. CALIBERrfimpute: Imputation in MICE using Random Forest [R package]; 2013. R package version 0.1-2. http://caliberanalysis.r-forge.r-project.org/.

[148] Shah AD, Bartlett JW, Carpenter JR, Nicholas O, Hemingway H. Comparison of Random Forest and parametric imputation models when imputing missing data using MICE: a CALIBER study. Am J Epidemiol. 2014;179:764–774.

[149] 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 Feb;13(2):133–138.

[150] Hammar N, Alfredsson L, Rosén M, Spetz CL, Kahan T, Ysberg AS. A national record linkage to study acute myocardial infarction incidence and case fatality in Sweden. Int J Epidemiol. 2001 Oct;30 Suppl 1:S30–S34.

[151] Barchielli A, Balzi D, Naldoni P, Roberts AT, Profili F, Dima F, et al. Hospital discharge data for assessing myocardial infarction events and trends, and effects of diagnosis validation according to MONICA and AHA criteria. J Epidemiol Community Health. 2012 May;66(5):462–467. http://dx.doi.org/10.1136/jech.2010.110908.

[152] The West of Scotland Coronary Prevention Study Group. Computerised record linkage: compared with traditional patient follow-up methods in clinical trials and illustrated in a prospective epidemiological study. J Clin Epidemiol. 1995 Dec;48(12):1441–1452.

[153] Pajunen P, Koukkunen H, Ketonen M, Jerkkola T, Immonen-Räihä P, Kärjä-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 Apr;12(2):132–137.

[154] 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(5):237–247. http://dx.doi.org/10.1007/s10654-009-9335-x.

[155] Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis. 1987;40(5):373–383.

[156] 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 May;285(19):2486–2497.

[157] R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2010. ISBN 3-900051-07-0. http://www.R-project.org/.

[158] 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 Dec;14(4):338–342. http://dx.doi.org/10.1093/eurpub/14.4.338.

[159] Payne RA, Abel GA, Simpson CR. A retrospective cohort study assessing patient characteristics and the incidence of cardiovascular disease using linked routine primary and secondary care data. BMJ Open. 2012;2(2):e000723. http://dx.doi.org/10.1136/bmjopen-2011-000723.

[160] Boyle CA, Dobson AJ. The accuracy of hospital records and death certificates for acute myocardial infarction. Aust N Z J Med. 1995 Aug;25(4):316–323.

[161] Madsen M, Davidsen M, Rasmussen S, Abildstrom SZ, Osler M. The validity of the diagnosis of acute myocardial infarction in routine statistics: a comparison of mortality and hospital discharge data with the Danish MONICA registry. J Clin Epidemiol. 2003 Feb;56(2):124–130.

[162] Hammad TA, Graham DJ, Staffa JA, Kornegay CJ, Pan GJD. Onset of acute myocardial infarction after use of non-steroidal anti-inflammatory drugs. Pharmacoepidemiol Drug Saf. 2008 Apr;17(4):315–321. http://dx.doi.org/10.1002/pds.1560.

[163] Hammad TA, McAdams MA, Feight A, Iyasu S, Pan GJD. Determining the predictive value of Read/OXMIS codes to identify incident acute myocardial infarction in the General Practice Research Database. Pharmacoepidemiol Drug Saf. 2008 Dec;17(12):1197–1201. http://dx.doi.org/10.1002/pds.1672.

[164] Jick H, Jick S, Myers MW, Vasilakis C. Risk of acute myocardial infarction and low-dose combined oral contraceptives. Lancet. 1996 Mar;347(9001):627–628.

[165] Meier CR, Jick SS, Derby LE, Vasilakis C, Jick H. Acute respiratory-tract infections and risk of first-time acute myocardial infarction. Lancet. 1998 May;351(9114):1467–1471.

[166] Hall GC, Brown MM, Mo J, MacRae KD. Triptans in migraine: the risks of stroke, cardiovascular disease, and death in practice. Neurology. 2004 Feb;62(4):563–568.

[167] 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 Jun;109(24):3000–3006. http://dx.doi.org/10.1161/01.CIR.0000132491.96623.04.

[168] Gulliford MC, Charlton J, Ashworth M, Rudd AG, Toschke AM, eC R T Research Team. Selection of medical diagnostic codes for analysis of electronic patient records. Application to stroke in a primary care database. PLoS One. 2009;4(9):e7168. http://dx.doi.org/10.1371/journal.pone.0007168.

[169] Collins R. What makes UK Biobank special? Lancet. 2012 Mar;379(9822):1173–1174. http://dx.doi.org/10.1016/S0140-6736(12)60404-8.

[170] Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH). http://www.dor.kaiser.org/external/DORExternal/rpgeh/.

[171] Million Veterans Program. www.research.va.gov/mvp.

[172] Johannesdottir SA, Horváth-Puhó E, Ehrenstein V, Schmidt M, Pedersen L, Sørensen HT. Existing data sources for clinical epidemiology: The Danish National Database of Reimbursed Prescriptions. Clin Epidemiol. 2012;4:303–313. http://dx.doi.org/10.2147/CLEP.S37587.

[173] van Staa TP, Dyson L, McCann G, Padmanabhan S, Belatri R, Goldacre B, et al. The opportunities and challenges of pragmatic point-of-care randomised trials us-

ing routinely collected electronic records: evaluations of two exemplar trials. Health Technol Assess. 2014 Jul;18(43):1–146. http://dx.doi.org/10.3310/hta18430.

[174] Boggon R, van Staa TP, Chapman M, Gallagher AM, Hammad TA, Richards MA. Cancer recording and mortality in the General Practice Research Database and linked cancer registries. Pharmacoepidemiol Drug Saf. 2013 Feb;22(2):168–175. http://dx.doi.org/10.1002/pds.3374.

[175] Bhaskaran K, Forbes HJ, Douglas I, Leon DA, Smeeth L. Representativeness and optimal use of body mass index (BMI) in the UK Clinical Practice Research Datalink (CPRD). BMJ Open. 2013;3(9):e003389. http://dx.doi.org/10.1136/bmjopen-2013-003389.

[176] Taxonomies & Controlled Vocabularies Special Interest Group. http://www.taxonomies-sig.org/about.htm.

[177] de Lusignan S, van Weel C. The use of routinely collected computer data for research in primary care: opportunities and challenges. Fam Pract. 2006 Apr;23(2):253–263. http://dx.doi.org/10.1093/fampra/cmi106.

[178] Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 2012 Sep;20(1):117–121. http://dx.doi.org/10.1136/amiajnl-2012-001145.

[179] Morley KI, Wallace J, Denaxas SC, Hunter RJ, Patel RS, Perel P, et al. Defining disease phenotypes using national linked electronic health records: a case study of atrial fibrillation. PLoS One. 2014;9(11):e110900. http://dx.doi.org/10.1371/journal.pone.0110900.

[180] Diabetes UK. Diabetes in the UK 2012: Key statistics on diabetes. Diabetes UK; 2010. https://www.diabetes.org.uk/About_us/What-we-say/Statistics/Diabetes-in-the-UK-2012/.

[181] de Lusignan S, Sadek N, Mulnier H, Tahir A, Russell-Jones D, Khunti K. Miscoding, misclassification and misdiagnosis of diabetes in primary care. Diabet Med. 2012 Feb;29(2):181–189. http://dx.doi.org/10.1111/j.1464-5491.2011.03419.x.

[182] Shelton N. Chapter 8: Diabetes. In: Health Survey for England 2003, Volume 2: Risk factors for cardiovascular disease. Department of Health; 2004. .

[183] McEvoy PM, Erceg-Hurn DM, Saulsman LM, Thibodeau MA. Imagery enhancements increase the effectiveness of cognitive behavioural group therapy for social anxiety disorder: A benchmarking study. Behav Res Ther. 2015;65(0):42–51.

[184] Raju D, Su X, Patrician PA, Loan LA, McCarthy MS. Exploring factors associated with pressure ulcers: a data mining approach. Int J Nurs Stud. 2015 Jan;52(1):102–111. http://dx.doi.org/10.1016/j.ijnurstu.2014.08.002.

[185] 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(6):618–626. http://dx.doi.org/10.1002/pds.1934.

[186] Schafer JL. Analysis of Incomplete Multivariate Data. No. 72 in Monographs on Statistics and Applied Probability. Chapman & Hall; 1997.

[187] van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw. 2011 12;45(3):1–67. http://www.jstatsoft.org/v45/i03.

[188] Seaman SR, Bartlett JW, White IR. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. BMC Med Res Methodol. 2012 Apr;12(1):46. http://dx.doi.org/10.1186/1471-2288-12-46.

[189] Hardt J, Herke M, Leonhart R. Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research. BMC Med Res Methodol. 2012;12:184.

[190] Burgette LF, Reiter JP. Multiple Imputation for Missing Data via Sequential Regression Trees. Am J Epidemiol. 2010;172(9):1070–1076.

[191] Breiman L. Random Forests. Mach Learn. 2001;45:5–32.

[192] Stekhoven DJ, Buhlmann P. MissForest–non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012 Jan;28(1):112–118. http://dx.doi.org/10.1093/bioinformatics/btr597.

[193] White IR, Royston P. Imputing missing covariate values for the Cox model. Stat Med. 2009;28(15):1982–1998.

[194] Barnard J, Rubin D. Miscellanea. Small-sample degrees of freedom with multiple imputation. Biometrika. 1999;86(4):948–955.

[195] Bartlett JW, Seaman SR, White IR, Carpenter JR. Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. arXiv. 2013;p. 1210.6799v3. http://arxiv.org/abs/1210.6799v3.

[196] Carpenter JR, Goldstein H, Kenward MG. REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. J Stat Softw. 2011;45(5):1–14.

[197] Shah AD, Langenberg C, Rapsomaniki E, Denaxas S, Pujades-Rodriguez M, Gale CP, et al. Type 2 diabetes and incidence of cardiovascular diseases: a cohort study in 1·9 million people. Lancet Diabetes Endocrinol. 2015 Feb;3(2):105–113. http://dx.doi.org/10.1016/S2213-8587(14)70219-0.

[198] van Dieren S, Beulens JWJ, van der Schouw YT, Grobbee DE, Neal B. The global burden of diabetes and its complications: an emerging pandemic. Eur J Cardiovasc Prev Rehabil. 2010 May;17 Suppl 1:S3–S8. http://dx.doi.org/10.1097/01.hjr.0000368191.86614.5a.

[199] Rydén L, Grant PJ, Anker SD, Berne C, Cosentino F, Danchin N, et al. ESC guidelines on diabetes, pre-diabetes, and cardiovascular diseases developed in collaboration with the EASD - summary. Diab Vasc Dis Res. 2014 May;11(3):133–173. http://dx.doi.org/10.1177/1479164114525548.

[200] Emerging Risk Factors Collaboration, Sarwar N, Gao P, Seshasai SRK, Gobin R, Kaptoge S, et al. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. Lancet. 2010 Jun;375(9733):2215–2222. http://dx.doi.org/10.1016/S0140-6736(10)60484-9.

[201] Peters SAE, Huxley RR, Woodward M. Diabetes as risk factor for incident coronary heart disease in women compared with men: a systematic review and meta-analysis of 64 cohorts including 858,507 individuals and 28,203 coronary events. Diabetologia. 2014 Aug;57(8):1542–1551. http://dx.doi.org/10.1007/s00125-014-3260-6.

[202] Gerds TA. prodlim: Product-limit estimation. Kaplan-Meier and Aalen-Johansson method for censored event history (survival) analysis; 2014. R package version 1.4.5. http://CRAN.R-project.org/package=prodlim.

[203] American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Care. 2014;37(Suppl 1):S81–S90.

[204] National Collaborating Centre for Chronic Conditions. Type 2 diabetes: national clinical guideline for management in primary and secondary care (update). London; 2008.

[205] Choudhury S, Hussain S, Yao G, Hill J, Malik W, Taheri S. The impact of a diabetes local enhanced service on quality outcome framework diabetes outcomes. PLoS One. 2013;8(12):e83738. http://dx.doi.org/10.1371/journal.pone.0083738.

[206] Fowkes FGR, Rudan D, Rudan I, Aboyans V, Denenberg JO, McDermott MM, et al. Comparison of global estimates of prevalence and risk factors for peripheral artery disease in 2000 and 2010: a systematic review and analysis. Lancet. 2013 Oct;382(9901):1329–1340. http://dx.doi.org/10.1016/S0140-6736(13)61249-0.

[207] 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 Sep;52(3):539–548. http://dx.doi.org/10.1016/j.jvs.2010.05.090.

[208] Rango PD, Farchioni L, Fiorucci B, Lenti M. Diabetes and abdominal aortic aneurysms. Eur J Vasc Endovasc Surg. 2014 Mar;47(3):243–261. http://dx.doi.org/10.1016/j.ejvs.2013.12.007.

[209] Ohrlander T, Merlo J, Ohlsson H, Sonesson B, Acosta S. Socioeconomic position, comorbidity, and mortality in aortic aneurysms: a 13-year prospective cohort study. Ann Vasc Surg. 2012 Apr;26(3):312–321. http://dx.doi.org/10.1016/j.avsg.2011.08.003.

[210] Iribarren C, Darbinian JA, Go AS, Fireman BH, Lee CD, Grey DP. Traditional and novel risk factors for clinically diagnosed abdominal aortic aneurysm: the Kaiser multiphasic health checkup cohort study. Ann Epidemiol. 2007 Sep;17(9):669–678. http://dx.doi.org/10.1016/j.annepidem.2007.02.004.

[211] Feigin VL, Rinkel GJE, Lawes CMM, Algra A, Bennett DA, van Gijn J, et al. Risk Factors for Subarachnoid Hemorrhage: An Updated Systematic Review of Epidemiological Studies. Stroke. 2005;36(12):2773–2780.

[212] Shantikumar S, Ajjan R, Porter KE, Scott DJA. Diabetes and the Abdominal Aortic Aneurysm. European Journal of Vascular and Endovascular Surgery. 2010;39(2):200–207.

[213] Siscovick D, Sotoodehnia N, Rea T, Raghunathan T, Jouven X, Lemaitre R. Type 2 diabetes mellitus and the risk of sudden cardiac arrest in the community. Rev Endocr Metab Disord. 2010;11(1):53–59. http://dx.doi.org/10.1007/s11154-010-9133-5.

[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(7):962–968.

[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(15):1737–1745.

[216] Peters SAE, Huxley RR, Woodward M. Diabetes as a risk factor for stroke in women compared with men: a systematic review and meta-analysis of 64 cohorts, including 775 385 individuals and 12 539 strokes. Lancet. 2014;383(9933):1973–1980.

[217] Mulnier HE, Seaman HE, Raleigh VS, Soedamah-Muthu SS, Colhoun HM, Lawrenson RA, et al. Risk of stroke in people with type 2 diabetes in the UK: a study using the General Practice Research Database. Diabetologia. 2006;49(12):2859–2865.

[218] Rivellese AA, Riccardi G, Vaccaro O. Cardiovascular risk in women with diabetes. Nutr Metab Cardiovasc Dis. 2010;20(6):474–480.

[219] de Rooij NK, Linn FHH, van der Plas JA, Algra A, Rinkel GJE. Incidence of subarachnoid haemorrhage: a systematic review with emphasis on region, age, gender and time trends. J Neurol Neurosurg Psychiatry. 2007 Dec;78(12):1365–1372. http://dx.doi.org/10.1136/jnnp.2007.117655.

[220] Pujades-Rodriguez M, Timmis A, Stogiannis D, Rapsomaniki E, Denaxas S, Shah A, et al. Socioeconomic Deprivation and the Incidence of 12 Cardiovascular Diseases in 1.9 Million Women and Men: Implications for Risk Prediction and Prevention. PLoS ONE. 2014 08;9(8):e104671. http://dx.doi.org/10.1371%2Fjournal.pone.0104671.

[221] Crosslin DR, McDavid A, Weston N, Zheng X, Hart E, de Andrade M, et al. Genetic variation associated with circulating monocyte count in the eMERGE Network. Hum Mol Genet. 2013;22(10):2119–2127.

[222] Bain BJ. Ethnic and sex differences in the total and differential white cell count and platelet count. J Clin Pathol. 1996 Aug;49(8):664–666.

[223] Lim EM, Cembrowski G, Cembrowski M, Clarke G. Race-specific WBC and neutrophil count reference intervals. Int J Lab Haematol. 2010 Dec;32(6 Pt 2):590–7.

[224] Rubin DB. Multiple imputation after 18+ years. J Am Stat Assoc. 1996 Jun;91:473–489.

[225] Yudkin JS, Stehouwer CDA, Emeis JJ, Coppack SW. C-Reactive Protein in Healthy Subjects: Associations With Obesity, Insulin Resistance, and Endothelial Dysfunction. Arterioscler Thromb Vasc Biol. 1999;19:972–978.

[226] Nakamura K, Fuster JJ, Walsh K. Adipokines: a link between obesity and cardiovascular disease. J Cardiol. 2014 Apr;63(4):250–259. http://dx.doi.org/10.1016/j.jjcc.2013.11.006.

[227] Schenkein HA, Loos BG. Inflammatory mechanisms linking periodontal diseases to cardiovascular diseases. J Clin Periodontol. 2013 Apr;40 Suppl 14:S51–S69. http://dx.doi.org/10.1111/jcpe.12060.

[228] Chen YW, Apostolakis S, Lip GYH. Exercise-induced changes in inflammatory processes: Implications for thrombogenesis in cardiovascular disease. Ann Med. 2014 Nov;46(7):439–55.

[229] Steenhof M, Janssen NAH, Strak M, Hoek G, Gosens I, Mudway IS, et al. Air pollution exposure affects circulating white blood cell counts in healthy subjects: the role of particle composition, oxidative potential and gaseous pollutants - the

RAPTES project. Inhal Toxicol. 2014 Feb;26(3):141–165. http://dx.doi.org/10.3109/08958378.2013.861884.

[230] Shah ASV, Langrish JP, Nair H, McAllister DA, Hunter AL, Donaldson K, et al. Global association of air pollution and heart failure: a systematic review and meta-analysis. Lancet. 2013 Sep;382(9897):1039–1048. http://dx.doi.org/10.1016/S0140-6736(13)60898-3.

[231] Newby DE, Mannucci PM, Tell GS, Baccarelli AA, Brook RD, Donaldson K, et al. Expert position paper on air pollution and cardiovascular disease. Eur Heart J. 2015 Jan;36(2):83–93.

[232] Ben-Chetrit E, Bergmann S, Sood R. Mechanism of the anti-inflammatory effect of colchicine in rheumatic diseases: a possible new outlook through microarray analysis. Rheumatology (Oxford). 2006 Mar;45(3):274–282. http://dx.doi.org/10.1093/rheumatology/kei140.

[233] Everett BM, Pradhan AD, Solomon DH, Paynter N, Macfadyen J, Zaharris E, et al. Rationale and design of the Cardiovascular Inflammation Reduction Trial: a test of the inflammatory hypothesis of atherothrombosis. Am Heart J. 2013 Aug;166(2):199–207.e15. http://dx.doi.org/10.1016/j.ahj.2013.03.018.

[234] Ridker PM, Thuren T, Zalewski A, Libby P. Interleukin-1β inhibition and the prevention of recurrent cardiovascular events: rationale and design of the Canakinumab Anti-inflammatory Thrombosis Outcomes Study (CANTOS). Am Heart J. 2011 Oct;162(4):597–605. http://dx.doi.org/10.1016/j.ahj.2011.06.012.

[235] Freedman DS, Flanders WD, Barboriak JJ, Malarcher AM, Gates L. Cigarette smoking and leukocyte subpopulations in men. Ann Epidemiol. 1996 Jul;6(4):299–306.

[236] von Vietinghoff S, Ley K. Homeostatic Regulation of Blood Neutrophil Counts. J Immunol. 2008;181(8):5183–5188.

[237] Okada Y, Kamatani Y, Takahashi A, Matsuda K, Hosono N, Ohmiya H, et al. Common variations in PSMD3-CSF3 and PLCB4 are associated with neutrophil count. Human molecular genetics. 2010 may;19(10):2079–2085. http://dx.doi.org/10.1093/hmg/ddq080.

[238] Nishida T, Sakakibara H. Association between underweight and low lymphocyte count as an indicator of malnutrition in Japanese women. J Womens Health (Larchmt). 2010 Jul;19(7):1377–1383. http://dx.doi.org/10.1089/jwh.2009.1857.

[239] Jensen EJ, Pedersen B, Frederiksen R, Dahl R. Prospective study on the effect of smoking and nicotine substitution on leucocyte blood counts and relation between

blood leucocytes and lung function. Thorax. 1998 Sep;53(9):784–789. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1745328/.

[240] Johannsen NM, Swift DL, Johnson WD, Dixit VD, Earnest CP, Blair SN, et al. Effect of different doses of aerobic exercise on total white blood cell (WBC) and WBC subfraction number in postmenopausal women: results from DREW. PLoS One. 2012;7(2):e31319. http://dx.doi.org/10.1371/journal.pone.0031319.

[241] Abidi K, Belayachi J, Derras Y, Khayari ME, Dendane T, Madani N, et al. Eosinopenia, an early marker of increased mortality in critically ill medical patients. Intensive Care Medicine. 2011;37(7):1136 – 1142. http://icmjournal.esicm.org.

[242] Oboni JB, Marques-Vidal P, Pralong F, Waeber G. Predictive factors of adrenal insufficiency in patients admitted to acute medical wards: a case control study. BMC Endocr Disord. 2013;13:3. http://dx.doi.org/10.1186/1472-6823-13-3.

[243] Mullerpattan JB, Udwadia ZF, Udwadia FE. Tropical pulmonary eosinophilia - A review. Indian J Med Res. 2013 Sep;138(3):295–302.

[244] Furuta GT, Atkins FD, Lee NA, Lee JJ. Changing roles of eosinophils in health and disease. Ann Allergy Asthma Immunol. 2014;113:3–8.

[245] Takatsu K. Interleukin-5 and IL-5 receptor in health and diseases. Proc Jpn Acad Ser B Phys Biol Sci. 2011;87(8):463–485.

[246] Linch SN, Danielson ET, Kelly AM, Tamakawa RA, Lee JJ, Gold JA. Interleukin 5 is protective during sepsis in an eosinophil-independent manner. Am J Respir Crit Care Med. 2012 Aug;186(3):246–254. http://dx.doi.org/10.1164/rccm.201201-0134OC.

[247] Vaduganathan M, Greene SJ, Butler J, Sabbah HN, Shantsila E, Lip GYH, et al. The immunological axis in heart failure: importance of the leukocyte differential. Heart Fail Rev. 2013 Nov;18:835–845.

[248] Infoshare; 2015. http://www.stats.govt.nz/infoshare/.

[249] Bannink L, Wells S, Broad J, Riddell T, Jackson R. Web-based assessment of cardiovascular disease risk in routine primary care practice in New Zealand: the first 18,000 patients (PREDICT CVD-1). N Z Med J. 2006;119(1245):U2313.

[250] Barron E, Lara J, White M, Mathers JC. Blood-Borne Biomarkers of Mortality Risk: Systematic Review of Cohort Studies. PLoS ONE. 2015 jun;10(6):e0127550.

[251] Kabagambe EK, Judd SE, Howard VJ, Zakai NA, Jenny NS, Hsieh M, et al. Inflammation Biomarkers and Risk of All-Cause Mortality in the Reasons for Geographic and Racial Differences in Stroke Cohort. Am J Epidemiol. 2011 jun;174(3):284–292. http://dx.doi.org/10.1093/aje/kwr085.

[252] Ruggiero C, Metter EJ, Cherubini A, Maggio M, Sen R, Najjar SS, et al. White Blood Cell Count and Mortality in the Baltimore Longitudinal Study of Aging. Journal of the American College of Cardiology. 2007;49(18):1841 – 1850. http://www.sciencedirect.com/science/article/pii/S0735109707006870.

[253] de Labry LO, Campion EW, Glynn RJ, Vokonas PS. White blood cell count as a predictor of mortality: results over 18 years from the Normative Aging Study. J Clin Epidemiol. 1990;43(2):153–157.

[254] Kim KI, Lee J, Heo NJ, Kim S, Chin HJ, Na KY, et al. Differential white blood cell count and all-cause mortality in the Korean elderly. Exp Gerontol. 2013 feb;48(2):103–108. http://dx.doi.org/10.1016/j.exger.2012.11.016.

[255] Riddell T, Wells S, Jackson R, Lee AW, Crengle S, Bramley D, et al. Performance of Framingham cardiovascular risk scores by ethnic groups in New Zealand: PREDICT CVD-10. N Z Med J. 2010 Feb;123(1309):50–61.

[256] Ministry of Health. Ethnicity data protocols for the health and disability sector. Wellington, New Zealand; 2004. https://www.health.govt.nz/system/files/documents/publications/ethnicitydataprotocols.pdf.

[257] Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika. 1994;81(3):515–526. http://biomet.oxfordjournals.org/content/81/3/515.abstract.

[258] Therneau T, original Splus->R port by Thomas Lumley. survival: Survival analysis, including penalised likelihood.; 2010. R package version 2.36-2. http://CRAN.R-project.org/package=survival.

[259] Harrell Jr FE. rms: Regression Modeling Strategies; 2015. R package version 4.3-0. http://CRAN.R-project.org/package=rms.

[260] Harrell Jr FE, with contributions from Charles Dupont and many others. Hmisc: Harrell Miscellaneous; 2015. R package version 3.15-0. http://CRAN.R-project.org/package=Hmisc.

[261] White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine. 2011;30(4):377–399. http://dx.doi.org/10.1002/sim.4067.

[262] Danesh J, Collins R, Appleby P, Peto R. Association of fibrinogen, C-reactive protein, albumin, or leukocyte count with coronary heart disease: meta-analyses of prospective studies. JAMA. 1998 May;279(18):1477–1482. http://jama.jamanetwork.com/article.aspx?articleid=187513.

[263] Grimm RH, Neaton JD, Ludwig W. Prognostic importance of the white blood cell count for coronary, cancer, and all-cause mortality. JAMA. 1985 Oct;254(14):1932–1937.

[264] Eke PI, Dye BA, Wei L, Thornton-Evans GO, Genco RJ, CDC Periodontal Disease Surveillance workgroup: James Beck (University of North Carolina CHSADPPAoP-PUoW. Prevalence of periodontitis in adults in the United States: 2009 and 2010. J Dent Res. 2012 Oct;91(10):914–920.

[265] Loos BG. Systemic markers of inflammation in periodontitis. J Periodontol. 2005 Nov;76(11 Suppl):2106–2115. http://dx.doi.org/10.1902/jop.2005.76.11-S.2106.

[266] Kweider M, Lowe GD, Murray GD, Kinane DF, McGowan DA. Dental disease, fibrinogen and white cell count; links with myocardial infarction? Scott Med J. 1993 Jun;38(3):73–74.

[267] Bergström J. Cigarette smoking as risk factor in chronic periodontal disease. Community Dent Oral Epidemiol. 1989 Oct;17(5):245–247.

[268] Löe H. Periodontal disease. The sixth complication of diabetes mellitus. Diabetes Care. 1993 Jan;16(1):329–334.

[269] McConnell H. International efforts in implementing national health information infrastructure and electronic health records. World Hosp Health Serv. 2004;40(1):33–40.

[270] 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 Jan;94(1):34–39. http://dx.doi.org/10.1136/hrt.2007.134890.

[271] Mullins HC, Scanland PM, Collins D, Treece L, Petruzzi P, Goodson A, et al. The efficacy of SNOMED, Read Codes, and UMLS in coding ambulatory family practice clinical records. Proc AMIA Annu Fall Symp. 1996;p. 135–139.

[272] openEHR: An open domain-driven platform for developing flexible e-health systems. http://www.openehr.org/.

[273] Kalra D, Fernando B. Approaches to enhancing the validity of coded data in electronic medical records. Prim Care Respir J. 2011 Mar;20(1):4–5. http://dx.doi.org/10.4104/pcrj.2010.00078.

[274] Meystre S, Savova G, Kipper-Schuler K, Hurdle J. Extracting Information from Textual Documents in the Electronic Health Record: a Review of Recent Research. Yearb Med Inform. 2008;p. 128–144.

[275] 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(1):e30412. http://dx.doi.org/10.1371/journal.pone.0030412.

[276] 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 Aug;12(1):88. http://dx.doi.org/10.1186/1472-6947-12-88.

[277] Microsoft. Design Guidance: Terminology – Matching. NHS Common User Interface Programme; 2007. https://www.uktcregistration.nss.cfh.nhs.uk/trud3/user/guest/group/0/home.

[278] Hwang KH, Chung KI, Chung MA, Choi D. Review of Semantically Interoperable Electronic Health Records for Ubiquitous Healthcare. Healthc Inform Res. 2010 Mar;16:1–5.

[279] Garde S, Chen R, Leslie H, Beale T, McNicoll I, Heard S. Archetype-based knowledge management for semantic interoperability of electronic health records. Stud Health Technol Inform. 2009;150:1007–1011.

[280] Torvalds L. Git: the stupid content tracker. http://git-scm.com/.

[281] Ishwaran H, Blackstone EH, Pothier CE, Lauer MS. Relative Risk Forests for Exercise Heart Rate Recovery as a Predictor of Mortality. J Am Stat Assoc. 2004;99(467):pp. 591–600. http://www.jstor.org/stable/27590433.

[282] Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random Survival Forests. Ann Appl Stat. 2008;2(3):pp. 841–860. http://www.jstor.org/stable/30245111.

[283] Rapsomaniki E, Shah A, Perel P, Denaxas S, George J, Udumyan R, et al. Prognostic models for stable coronary artery disease based on electronic health record cohort of 102,023 patients. Eur Heart J. 2014 Apr;35:844–852.

[284] Atila K, Arslan NC, Derici S, Canda AE, Sagol O, Oztop I, et al. Neutrophil-to-lymphocyte ratio: could it be used in the clinic as prognostic marker for gastrointestinal stromal tumor? Hepatogastroenterology. 2014 Sep;61(134):1649–53.

[285] Koh YW, Lee HJ, Ahn JH, Lee JW, Gong G. Prognostic significance of the ratio of absolute neutrophil to lymphocyte counts for breast cancer patients with ER/PR-positivity and HER2-negativity in neoadjuvant setting. Tumour Biol. 2014 Oct;35(10):9823–30.

[286] Kelkitli E, Atay H, Cilingir F, Güler N, Terzi Y, Ozatlı D, et al. Predicting survival for multiple myeloma patients using baseline neutrophil/lymphocyte ratio. Ann Hematol. 2014 May;93(5):841–846. http://dx.doi.org/10.1007/s00277-013-1978-8.

[287] Cavus UY, Yildirim S, Sonmez E, Ertan C, Ozeke O. Prognostic value of neutrophil/lymphocyte ratio in patients with pulmonary embolism. Turk J Med Sci. 2014;44(1):50–5.

[288] Benites-Zapata VA, Hernandez AV, Nagarajan V, Cauthen CA, Starling RC, Tang WHW. Usefulness of neutrophil-to-lymphocyte ratio in risk stratification of patients with advanced heart failure. Am J Cardiol. 2015 Jan;115(1):57–61.

[289] Mokry LE, Ahmad O, Forgetta V, Thanassoulis G, Richards JB. Mendelian randomisation applied to drug development in cardiovascular disease: a review. J Med Genet. 2015 Feb;52(2):71–79. http://dx.doi.org/10.1136/jmedgenet-2014-102438.

[290] Interleukin-6 Receptor Mendelian Randomisation Analysis (IL6R MR) Consortium. The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis. Lancet. 2012 Mar;379(9822):1214–1224. http://dx.doi.org/10.1016/S0140-6736(12)60110-X.

[291] Moore C, Sambrook J, Walker M, Tolkien Z, Kaptoge S, Allen D, et al. The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. Trials. 2014;15:363. http://dx.doi.org/10.1186/1745-6215-15-363.

[292] Yudkin JS, Eringa E, Stehouwer CDA. “Vasocrine" signalling from perivascular fat: a mechanism linking insulin resistance to vascular disease. Lancet. 2005;365:1817–1820.