Exploiting electronic health records for research on atrial fibrillation: risk factors, subtypes, and outcomes

Victoria Allan

University College London

PhD Health Informatics

I, Victoria Allan 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.

Foreword

This thesis is the product of three and a half years of research, exploring how electronic health records, collected on large populations as part of routine clinical care, can be used to investigate the common, yet incompletely understood, heart rhythm disorder atrial fibrillation.

Over the last hundred years, key advances have been made in understanding atrial fibrillation at both the cellular and individual level. Most notable is the invention of electrocardiography, which was awarded the Nobel Prize for medicine in 1924. Electrocardiography offers a non-invasive way of detecting atrial fibrillation in individuals and is a procedure firmly embedded in clinical practice today. But here in 2018, the era of big data, this thesis seeks to illustrate how population level approaches to the study of atrial fibrillation can also be used to yield novel disease insights.

I thank my supervisors Professor Harry Hemingway and Dr Amitava Banerjee for their guidance over the course of this PhD, as well as the below collaborators who have contributed to publications arising from this thesis:

1. **Allan V**, Honarbakhsh S, Casas JP, Wallace J, Hunter R, Schilling R, Perel P, Morley K, Banerjee A, Hemingway H. Are cardiovascular risk factors also associated with the incidence of atrial fibrillation? A systematic review and field synopsis of 23 factors in 32 population-based cohorts of 20 million participants.Thromb Haemost. 2017 May 3;117(5):837-850.

2. **Allan V**, Banerjee A, Shah AD, Patel R, Denaxas S, Casas JP, Hemingway H. Net clinical benefit of warfarin in individuals with atrial fibrillation across stroke risk and across primary and secondary care.Heart. 2017 Feb;103(3):210-218.

I am indebted to Giovanna Ceroni for her help in organising my PhD progress meetings, to Michail Katsoulis for his sound statistical advice and to the other doctoral students at the UCL Institute of Health Informatics with whom I have shared this journey and have established great friendships.

I am enormously grateful to my Mam, Sister and Luca Perletta. They are three pillars of strength who have supported me in this PhD and continue to support me in life in general.

I dedicate this work to my Dad.

Victoria Allan, April 2018

Short abstract

Background: Electronic health records (EHRs), collected on large populations in routine clinical care, may hold novel insights into the heart rhythm disorder atrial fibrillation (AF).

Aim: To exploit EHRs to investigate, validate and extend evidence for AF risk factors, subtypes, and outcomes.

Methods: The CALIBER dataset (1997–2010) linking primary care, secondary care, and mortality records for a representative subset of the UK population was used (i) to model associations between cardiovascular disease (CVD) risk factors and incident AF, including AF with (AF^+) and AF without (AF^-) intercurrent CVD, (ii) to create EHR definitions for eight AF subtypes (structural, focal, polygenic, postoperative, valvular, monogenic, respiratory and AF in athletes) and (iii) to investigate stroke outcomes by CHA_2DS_2-VASc, sex, and warfarin use.

Results: Among 1,949,052 individuals, 50,097 developed incident AF: 12,652 (25.3%) with AF^+ and 37,445 (74.7%) with AF^-. Smoking (HR [95%CI] for AF^+ vs. AF^-: 1.66 [1.56,1.77] vs. 1.21 [1.16,1.25]), hypertension (2.19 [2.11,2.27] vs. 1.65 [1.62,1.69]), and diabetes (2.03 [1.94,2.12] vs. 1.45 [1.41,1.49]) showed consistent direct associations with AF^+ and AF^-, while heavy drinking (1.17 [0.81,1.67] vs. 1.99 [1.68,2.34]) and total cholesterol levels (0.99 [0.96,1.02] vs. 0.85 [0.84,0.87]) showed inconsistent associations with AF^+ and AF^-. EHR definitions for AF subtypes were created by combining 2813 diagnosis, medication, and procedure codes. There were 12,751 individuals with AF and valvular heart disease. Prosthetic replacements, mitral stenosis and aortic stenosis showed higher HR [95%CI] for stroke, thromboembolism and mortality (1.13 [1.02,1.24], 1.20 [1.05,1.36], and 1.27 [1.19,1.37] respectively). The net-clinical benefit (NCB [95%CI] per 100 person-years) of warfarin was shown from CHA_2DS_2-VASc≥2 in men (0.5 [0.1,0.9]) and CHA_2DS_2-VASc≥3 in women (1.5 [1.1,1.9]).

Conclusion: AF is a heterogeneous condition associated with diverse disease mechanisms. EHRs can help refine understanding of risk factors, subtypes, and outcomes with relevance for clinical practice.

Extended abstract

Background: Electronic health records (EHRs), collected on large populations in routine clinical care, can be exploited for research to yield novel disease insights. This PhD thesis explores the use of EHRs for investigating the common, yet incompletely understood, heart rhythm disorder atrial fibrillation (AF). AF has captured recent clinical attention with advancements in treatments and new ideas on developmental mechanisms. EHRs offer a valuable data source in which to address outstanding clinical questions in relation to AF risk factors, subtypes and outcomes, however have been underutilised in AF research so far. A number of analytic complexities arise as EHRs are not collected for primary research purposes and therefore this thesis also describes and applies methods for overcoming these.

Setting: The ClinicAl research using LInked Bespoke studies and Electronic health Records (CALIBER) dataset was used. CALIBER links primary care, secondary care, and mortality records for a subset of the UK population that is representative of the overall population in terms of age, sex, ethnicity and mortality. The years 1997 to 2010 were studied, reflecting the time period when linked data sources were aligned.

Aim: To exploit EHRs in CALIBER to investigate, validate and extend evidence for AF risk factors, subtypes, and outcomes.

Specific objectives:

- To conduct a systematic review and field synopsis of the existing observational epidemiology on the associations of 23 cardiovascular risk factors with incident AF
- To use the CALIBER dataset to model the associations of 23 cardiovascular risk factors with two novel AF endpoints: AF with (AF^+) and without (AF^-) intercurrent cardiovascular disease (CVD)
- To create EHR definitions for eight AF subtypes relevant to the 2016 European Society of Cardiology guidelines, which are: (1) structural, (2) focal, (3) polygenic, (4) postoperative, (5) valvular, (6) AF in athletes, (7) monogenic and (8) respiratory AF.
- To use the CALIBER dataset to implement, improve and validate the EHR definition for valvular AF.
- To use the CALIBER dataset to investigate stroke outcomes by CHA_2DS_2-VASc, sex and warfarin use.

Methods: A systematic review (Pubmed to October 2015) and field synopsis of population-based, consented or EHR, cohorts that investigated the associations of one or more of 23 cardiovascular risk factors and incident AF was carried out. For each risk factor relative risks (RR) and 95% confidence intervals [95% CI] were extracted and visualised using forest plots. Associations between 23 cardiovascular risk factors and incident AF, AF^+ and AF^-, were investigated in CALIBER using Cox regression adjusted for age and sex and stratified on GP practice and the hazard ratios (HR) and 95% CI obtained were compared with estimates from prior literature.

In line with CALIBER guidelines, computable EHR definitions were created for eight AF subtypes by identifying relevant clinical codes from pre-existing code lists, new and updated code searches (to December 2016) and by matching synonymous codes between the Read (primary care diagnoses and procedures), ICD-10 (secondary care diagnoses) and OPCS-4 (secondary care procedures) classification systems. Where relevant codes were unavailable to describe a subtype of interest, algorithms for inferred cases were devised. The definition for valvular AF was implemented in EHRs and Cox regression was used to model the associations of different valvular heart diseases (VHDs) with a composite endpoint of incident stroke (ischaemic, haemorrhagic and unspecified), systemic embolism, and all-cause mortality. Incidence rates and 95% confidence intervals per 100 person-years (IR [95% CI] /100 PY) were calculated for ischaemic strokes (IS) and hemorrhagic strokes (HS) in men and women, with and without use of warfarin, and across levels of the CHA_2DS_2-VASc stroke risk score. The net clinical benefit (i.e. number of IS avoided vs. number of HS caused) per 100 person-years of warfarin use (NCB [95%CI] /100 PY) was estimated using the formula: (IS rate $_{\text{without warfarin}}$ - IS rate $_{\text{with warfarin}}$) – 1.5(HS rate $_{\text{with warfarin}}$ - HS rate $_{\text{without warfarin}}$).

Results: Overall, the systematic review and field synopsis included 73 out of 2777 publications (84 reports based upon 28 consented and 4 EHR cohorts from 10 countries), with 576,602 AF events in 20,420,175 participants. Hypertension (13/17 reports) and obesity (19/19 reports) showed direct associations, while 4 other factors showed associations with AF in the opposite direction of known associations with CHD. These were inverse associations for non-White ethnicity (5/5 reports, with RR from 0.35 to 0.84 [0.82,0.85]), total cholesterol (4/13 reports from 0.76 [0.59,0.98] to 0.94 [0.90,0.97]; 8/13 reports with non-significant inverse associations), and diastolic blood pressure (2/11 reports from 0.87 [0.78,0.96] to 0.92 [0.85,0.99]; 5/11 reports with non-significant inverse associations), and direct associations for taller height (7/10 reports from 1.03 [1.02,1.05] to 1.92 [1.38,2.67]). Among 1,949,052 initially healthy individuals within the CALIBER dataset, 50,097 developed incident AF: 12,652 (25.3%) with AF^+ and 37,445 (74.7%) with AF^-. Smoking (HR [95%CI] for AF^+ vs. AF^-: 1.66 [1.56,1.77] vs. 1.21 [1.16,1.25]), hypertension (HR [95%CI] for AF^+ vs. AF^-: 2.19 [2.11,2.27] vs. 1.65 [1.62,1.69]), and diabetes (HR [95%CI] for AF^+ vs. AF^-: 2.03 [1.94,2.12] vs. 1.45 [1.41,1.49]) showed consistent direct associations with AF^+ and AF^-, while heavy drinking (HR [95%CI] for AF^+ vs. AF^-: 1.17 [0.81,1.67] vs. 1.99 [1.68,2.34]) and total cholesterol levels (HR [95%CI] for AF^+ vs. AF^-: 0.99 [0.96,1.02] vs. 0.85 [0.84,0.87]) showed inconsistent associations with AF^+ and AF^-. EHR definitions were set out for all eight AF subtypes based on code combinations and plausible inferences. A total of 2813 applicable clinical codes were identified. In the absence of family relationships or genomics data, the definition for polygenic AF was derived based on inferences of an early age of AF onset and AF not explained by any of the other subtypes. Implementation, improvement and validation of the valvular definition identified 12,751 individuals with AF and VHD at baseline. Compared with individuals with AF and no VHD, individuals with prosthetic valves, mitral stenosis and aortic stenosis had higher HR [95% CI] for stroke, systemic embolism and mortality of 1.13 [1.02, 1.24], 1.20 [1.05, 1.36], and 1.27 [1.19, 1.37] respectively after adjustment for age, sex, warfarin and CHA_2DS_2-VASc risk factors, while individuals with bioprosthetic valve re-

placements had a lower adjusted hazard ratio of 0.78 [0.68, 0.88]. A significant positive net clinical benefit of warfarin was found from CHA_2DS_2-VASc≥2 in men (NCB [95% CI] /100 PY: 0.5 [0.1,0.9]) and from CHA_2DS_2-VASc≥3 in women (NCB [95% CI] /100 PY: 1.5 [1.1,1.9]).

Conclusion: AF is a heterogeneous condition associated with diverse disease mechanisms for onset and progression. A systematic evaluation of the available observational evidence suggests similarities as well as important differences in the risk factors for incidence of AF as compared with other CVDs, which has implications for the primary prevention strategies for AF. Primary preventions strategies will however only work given that the target of prevention is clearly defined. The development of computable definitions of eight AF subtypes demonstrates the viability of EHR data in validating and refining understanding about these more precise targets of prevention. In addition to prosthetic heart valves and mitral stenosis, EHR data suggest aortic stenosis may also be clinically relevant in the progression of AF. Using a highly representative population of individuals with AF from primary and secondary care settings, incidence rates of ischaemic stroke in men and women with one 1-point scoring risk factor from CHA_2DS_2-VASc (irrespective of sex) were lower than previously reported, which may change the decision to start anticoagulation with warfarin in these individuals. EHRs can thus help to advance understanding of AF risk factors, subtypes, and outcomes and should be increasingly utilised to inform future clinical trials, future clinical guidelines and future clinical practice.

Impact statement

The work of PhD thesis was carried out in order to advance understanding of the common, yet incompletely understood, heart rhythm disorder atrial fibrillation and, ultimately, to improve the clinical care and outcomes of diagnosed individuals.

In studying how electronic health records can be used in research on atrial fibrillation, the clinical code lists, algorithms and practical insights I derived are of value to the scientific community going forward in future investigations.

My work is also of direct benefit to public health with findings in relation to stroke risk in individuals with atrial fibrillation cited against treatment recommendations in recently updated international clinical practice guidelines (European Society of Cardiology, 2016).

Contents

List of tables

List of figures

Chapter 1

Introduction to overall aims, objectives and background motivating this research

1.1 Chapter overview and broader thesis outline

Electronic health records (EHRs), collected on large populations as part of routine clinical care, can be exploited for research in order to yield novel disease insights. This PhD thesis explores the use of EHRs for investigating the common, yet incompletely understood, heart rhythm disorder atrial fibrillation (AF). AF has captured recent clinical attention with advancements in treatments and new ideas on developmental mechanisms. However, EHRs have been underutilised in AF research so far.

This first chapter presents the background, which has motivated my research, including the clinical importance of AF,[1] current clinical uncertainties in AF risk factors,[2] subtypes[3] and outcomes[4] and the opportunity for studying these using EHRs. **Chapter 2** reviews the existing observational epidemiology on the link between a range of cardiovascular risk factors and incidence of AF and summarises the findings using a newer field synopsis methodology.[2] **Chapter 3** presents CALIBER (an acronym for ClinicAl research using LInked Bespoke studies and Electronic health Records),[5] which links EHRs in the United Kingdom (UK) and is the data source used within the analytic chapters of this thesis. **Chapter 4** addresses limitations in the existing observational epidemiology on risk factors for AF,[2] by using EHRs to examine the role of intercurrent cardiovascular diseases (CVDs) in the development of AF. **Chapter 5** shifts in focus from AF risk factors to AF subtypes and investigates whether EHR definitions can be derived for a range of AF clinical distinctions recently outlined in the 2016 European Society of Cardiology (ESC) guidelines for the management of AF.[3] **Chapter 6** takes forward the EHR definition derived for AF in the context of valvular heart diseases[6] for further examination and refinement. **Chapter 7** turns attention again from AF subtypes toward AF outcomes and uses EHRs to investigate stroke rates according to estimated stroke risk, sex and use of warfarin.[7] Finally, **chapter 8** discusses the overall implications of this thesis including 12 recommendations for future AF clinical practice and research.

It should be noted that the works presented in the chapters of this thesis were not necessarily completed in the order that they appear. Instead, I have chosen to present them in an order reflecting the natural transition between risk factors predisposing to AF (i.e. primary prevention), through to subtypes influencing AF treatment decisions (i.e. diagnosis and management) and then on to AF-related outcomes aimed at being prevented (i.e. secondary prevention; **figure 1.1**).

1.2 Atrial fibrillation

AF is the world's most common heart rhythm disorder and a leading cause of fatal and disabling strokes. In 2010, AF prevalence was estimated at 33.5 million people worldwide[8] with projec-

tions suggesting a doubling in the number of people affected over the next 50 years.[9 10] This includes the UK where recent estimates suggest an increase in prevalence from 700,000 people with AF in 2010 to between 1.3 and 1.8 million people with AF by 2060.[11] To convey the clinical importance of AF, in terms of pathophysiology and prognosis,[1] it helps to describe the electrophysiology behind normal heart rhythm.[12]

1.2.1 Normal heart rhythm: electrophysiology

Normal heart rhythm, as **figure 1.2** shows, is regulated by a wave of electrical activity passing through the upper and lower heart chambers known as the atria and the ventricles. Electric activity originates in the sinoatrial node (SA), located in the top right atrium. During each heart rhythm cycle (i.e. one heartbeat) the SA node emits an electrical impulse which passes through the atria stimulating contraction and transportation of blood into the ventricles. The electrical impulse is received by the atrioventricular (AV) node, located between the right atrium and right ventricle. At the AV node, the electrical impulse first pauses to ensure complete transportation of blood between the atria and the ventricles before passing through the ventricles stimulating contraction and transportation of blood to the lungs (i.e. the right ventricle for blood reoxygenation) or out towards the rest of the body (i.e. the left ventricle to supply body cells with new oxygen).[12] Conversely, while in AF the heart's co-ordinated electrical system malfunctions.[1]

1.2.2 Atrial fibrillation: pathophysiology and prognosis

AF arises when the heart's co-ordinated system of electrical activity malfunctions. Electrical impulses, which normally originate from the SA node, are instead emitted spontaneously from other parts of the atria. This stimulates the atria and ventricles to contract out of sequence and can result in incomplete transportation of blood between the heart's chambers.[1]

AF is a leading risk factor for stroke because irregular heart contractions mean that blood is transported through the heart less effectively and can accumulate inside the heart chambers forming blood clots. Blood clots may then be transported out of the heart, travel through the bloodstream and become trapped in blood vessels, most often in the brain.[1] AF therefore accounts for 1 in 4 strokes,[13 14] doubles the risk of death[15] and places a substantial economic burden on healthcare systems.[16] In the UK alone, AF was estimated to be the direct cause of 12,500 strokes in 2008, with associated treatment costs of £148 million (£11,900 per stroke).[17]

Early detection and diagnosis of AF is therefore crucial in order to manage the increased stroke risk.[3 18 19] Individuals with AF may experience a range of symptoms including lethargy, palpitations, dyspnoea, chest tightness, sleeping difficulties, and psychosocial distress,[3] although quite often AF is asymptomatic which makes detection and diagnosis difficult.

1.2.3 Atrial fibrillation: detection and diagnosis

Early detection and diagnosis of AF is crucial in order to manage the increased stroke risk, as well as any symptoms,[3 18 19] and this can be determined using electrocardiography (electrocardiogram; ECG). ECGs provide an image of the heart's electrical activity, which is obtained from

electrodes strategically placed upon the skin's surface. Each electrode captures the heart's electrical activity from a different vector which when combined produces a single waveform with characteristics to distinguish AF from normal heart rhythm.[12]

The ECG characteristics of normal heart rhythm, as **figure 1.3** shows, are the P, Q, R, S and T waves, which correspond to the co-ordinated sequence of electrical activity spreading through the heart's chambers. The P wave first captures the electrical impulses stimulating the atria to contract, the QRS waves then capture the electrical impulses stimulating the ventricles to contract, while the T wave then captures the ventricles relaxing before another heart rhythm cycle begins. Conversely, while in AF, the heart's electrical system is no longer co-ordinated and thus the consecutive P, QRS and T waves do not appear on an ECG. Instead, the ECG characteristics of AF, as **figure 1.4** shows, are an absence of P waves and QRS waves appearing at irregular intervals.[12]

AF may also be crudely detected using pulse palpation.[3 18 19] This involves placing the fingertips on an artery close to the skin's surface, usually at the neck or wrist, applying gentle pressure until the pulse is felt and then determining the number, and regularity, of beats within one minute. A normal pulse at rest is considered within the range 60 to 100 beats per minute with regular intervals between each beat. A pulse outside of the normal rate or appearing at irregular intervals may indicate AF. Improved detection of AF is being achieved through public awareness campaigns such as Know Your Pulse,[20] as well as the CATCH ME (Characterizing Atrial fibrillation by Translating its Causes into Health Modifiers in the Elderly) smart phone and tablet applications.[21]

1.2.4 Atrial fibrillation: clinical management

Once diagnosed, AF is predominantly managed using pharmacological interventions[3 19 22] with procedural approaches usually reserved for more troublesome cases. Anticoagulants, which limit the blood's ability to form blood clots, are the cornerstone of AF management, used to reduce the risk of stroke by a half.[23 24] Drugs to control heart rate and rhythm are also available however there are presently little data to suggest that these alone are protective against stroke and thus further clinical trials are underway and in the pipeline.[25 26] Procedural approaches for AF include direct current cardioversion, which shocks the heart back into normal heart rhythm, and catheter ablation, which destroys the source of abnormal electrical impulses in the atria.[27] However the comparative effectiveness of these procedures, including in comparison to pharmacological interventions, also remains under review.[28]

Though proven to be effective in preventing strokes, anticoagulants can however increase the risk of bleeding and therefore not all individuals with AF undergo treatment. The decision to treat individuals with anticoagulants is therefore made on balance of stroke and bleeding risk. This can be determined using appropriate risk prediction models such as the CHA_2DS_2-VASc (Congestive heart failure, Hypertension, Age ≥75 years, Diabetes mellitus, history of Stroke or thromboembolism, Vascular disease, Age 65–74 years, and Sex category)[29] and HAS-BLED

(Hypertension, Abnormal renal and liver function, Stroke, Bleeding, Labile International Normalised Ratio (INR), Elderly, Drugs or alcohol)[30] scores which are recommended in international guidelines governing the way that AF is managed in clinical practice.[3 19 22]

1.2.5 Atrial fibrillation: recent clinical advances

At the time of commencing this thesis, AF had captured recent clinical attention with excitement and promise around the newly available direct oral anticoagulants (DOACs) for stroke prevention in individuals with AF.[31-34] Previously the choice of anticoagulant had been limited to that of warfarin which has a key disadvantage of needing regular INR tests to ensure that the blood clots at a rate beneficial for stroke prevention but doesn't cause harmful bleeding. Yet, in DOAC trials of dabigatran,[31] rivaroxaban,[32] apixaban[33] and edoxaban,[34] it was shown that all four agents were as effective as warfarin in preventing ischaemic strokes and associated with fewer bleeding complications, negating the need for regular blood testing.

The recent introduction of DOACs reflects a major advancement for AF research and clinical practice, however as I now go on to describe, a number of clinical uncertainties remain, including in relation to AF risk factors,[2] subtypes[3] and outcomes.[4]

1.2.6 Atrial fibrillation: clinical uncertainty

In spite of recent advances in understanding AF aetiology many clinical uncertainties in relation to risk factors,[2] subtypes[3] and outcomes[4] remain.

Uncertainty surrounds the risk factors to target in primary prevention of AF[2] because the predominant focus of research to date has been on secondary prevention of stroke once AF has developed. So far there has been a lack of clinical trials of healthy individuals without pre-existing CVD and with AF as the primary outcome.[35] Community screening programmes for detection of AF have been proposed,[36] however these programmes are also designed to identify people at high risk of stroke and thromboembolism, and do not identify those who are at an initially high risk of later developing AF. Current clinical guidelines therefore make no recommendations for the primary prevention of AF itself, among people without pre-existing CVDs.[3 19 22]

Of course, for AF primary preventions strategies to work, the target of prevention must be clearly defined, but uncertainty also surrounds definitions for AF and its subtypes.[3] Over the past 15 years, AF has been classified according to internationally agreed definitions describing the frequency and duration of heart rhythm disorder, either: paroxysmal (i.e. self-terminating), persistent (i.e. not self-terminating), or permanent (i.e. persistent and resistant to treatment).[37] However, each type is understood to confer a similar stroke risk and therefore warrants the same management according current clinical guidelines.[3 19 22] More recently, the 2016 updates to ESC guidelines for the management of AF outlined seven mechanistically distinct subtypes of AF, which are: (1) AF secondary to structural heart disease, (2) focal AF, (3) polygenic AF, (4) postoperative AF, (5) AF in mitral stenosis or prosthetic heart valves (often referred to as 'valvular' AF), (6) AF in athletes and (7) monogenic AF.[3] These definitions reflect contemporary clini-

cal thinking but were derived based on expert consensus and remain unsupported by any quantitative evidence.

As far as AF outcomes are concerned, the link between AF and subsequent stroke risk is well-established however uncertainty on the level of stroke risk (as determined by the CHA_2D_S2-VASc risk score) at which individuals require preventative therapy with anticoagulants continues to be the topic of much debate.[4]

These uncertainties have not only shaped the direction of this thesis but have also gained recognition from some of the major funders of biomedical research in recent years.

1.2.7 Atrial fibrillation: clinical recognition

In recent years, two major funders of biomedical research have recognised that the clinical determinants of AF need better understood and, as a result, global consortia have formed in order to tackle challenges in AF research together.[38 39]

Firstly in 2008, the National Institutes of Health (NIH; a major funder of biomedical research in the United States (US)) set out to discover risk factors for AF by combining data from existing consented cohort studies.[38] The CHARGE-AF (Cohorts for Heart and Aging Research in Genomic Epidemiology – AF) consortium was formed bringing together data from three US cohorts (the Atherosclerosis Risk in Communities study, the Cardiovascular Health Study and the Framingham Heart Study) and two European cohorts (the Age, Gene and Environment-Reykjavik study and the Rotterdam Study) with a combined total of 1771 incident AF events.[40] However, the NIH neglected the potential role that population level EHR data can play in advancing AF prevention research.

In 2017, and while in the process of writing this thesis, the Innovative Medicines Initiative (IMI) BigData@Heart project was launched to understand how EHR data sources across Europe compare and can be combined in order to drive progress in cardiovascular research.[39] AF (along with myocardial infarction and heart failure) is one of three CVDs of particular interest to the IMI BigData@Heart project. In many ways, this thesis lays important foundations for future AF research using EHRs, including that of IMI BigData@Heart. As I now go on to describe, EHRs in the UK offer a number of unique advantages for studying the onset and progression of AF.

1.3 Electronic health records

EHRs concern the digital collection of individuals' health and health-related information.[5] They are collected on a population level in several countries (e.g. the UK,[41] Denmark[42] and Sweden[43]) as part of routine clinical care. EHRs offer unparalleled opportunities for research in terms of large samples sizes and breadth of clinical information. However, they are primarily collected for administrative and financial reasons and therefore challenges can arise in the re-purposing of EHR data for research.[44]

1.3.1 Electronic health records: overview of opportunities and challenges

EHRs contain invaluable insights into disease prevention, development, progression, detection, diagnosis and treatment. EHRs can contain information on symptoms, diagnoses, drug prescriptions, operations and procedures, results of pathological tests, anthropometric measurements, and health behaviours. EHRs contain mostly structured information that is coded using universal systems such as the International Statistical Classification of Diseases and Health-Related Problems (ICD)[45] but they may also contain unstructured data such as free-text descriptions[46] or images[39] (such as ECGs; although these are a lot less common).

In recent years, there has been a rapid expansion in the provision of EHR for use in research.[47] EHRs offer unparalleled opportunities for research, which include the potential for studying diseases at much larger scale (e.g. at a whole country level), and across a wide range of risk factors and disease endpoints (e.g. any condition defined in the 21 disease chapters of ICD). EHRs are also highly representative of patient populations are they consists of unselected individuals treated with usual clinical care.[39]

The challenges of using EHR data for research largely reflect the fact that data are not collected for primary research purposes.[48] A key challenge lies in the identification of diseases cases. Diseases are rarely explained by a single clinical code and even where relevant codes exist they may not be used in clinical practice. Researchers must therefore invest considerable time into compiling code lists and understanding how codes can be combined in order to identify cases as accurately as possible.[49] A second important challenge lies in the fact that EHRs collected in different healthcare settings (such as in primary care and in secondary care) are often unlinked.[50] Linking EHRs brings advantages for research by allowing the full pathway of care to be studied including, for example, any diagnoses made or medications prescribed exclusively in each setting. However, EHRs from different settings may be coded using diverse classification systems with varying levels of clinical detail and may also cover non-identical patient populations.[49]

In **chapter 3** I provide a full description of the CALIBER database, which is the EHR data source used within the subsequent analytic chapters of this thesis.[5] CALIBER links EHRs from primary care, secondary care, an acute coronary syndromes registry and the mortality registry in the UK creating unique opportunities for AF research. As part of a wider objective the CALIBER programme seeks to overcome challenges and extract the value from EHR data for the benefit of health research in the UK and beyond. The CALIBER portal (www.caliberresearch.org/portal) is an online repository for sharing knowledge, algorithms and code lists for identifying (currently around >600) risk factors and outcomes in EHRs.[51]

1.3.2 Electronic health records: opportunities for atrial fibrillation research

This thesis exploits EHRs to investigate three aspects of AF aetiology: AF risk factors, AF subtypes and AF outcomes. I focussed on these aspects because they relate to current limitations in clinical practice guidelines influencing the way individuals with, or at risk of, AF are managed[3]

[19 22] but also because the opportunity exists to study these using EHRs in the UK.[51] Indeed, the analyses of this thesis are made possible because of the structure of the UK healthcare system[52] and because AF can be accurately identified in EHRs.[49]

EHR data exist in vast quantities in the UK because the National Health Service (NHS) provides healthcare to 98% of the population.[53] Importantly, every healthcare consultation generates an EHR registered against an NHS number unique to each individual. Although EHRs from different NHS healthcare settings (e.g. across primary and secondary care) are not linked centrally in the UK at the present, subsequent data linkage (i.e. bringing together disconnected records belonging to the same individual) is achieved by matching on the basis of NHS numbers (as was done to create the CALIBER dataset, explained in **chapter 3**[5]).

As described, EHRs are not collected for primary research purposes and therefore, among other challenges, identifying disease cases can be complex.[48] This PhD therefore benefits from earlier work by Morley and colleagues using the CALIBER dataset to show that AF cases can be accurately identified in primary and secondary care linked EHRs. ECG images (i.e. the gold standard method for ascertaining AF cases) cannot currently be extracted from NHS systems and made available for population level research.[54] Therefore in the absence of these, Morley created an EHR definition for AF which, as **figure 1.4** shows, is composed of coded AF cases and inferred AF cases based on warfarin or digoxin prescriptions without thromboembolic disease or heart failure.[49] Using this definition, Morley found an AF prevalence estimate in CALIBER of 1.6% which is comparable to the 2.0% (95% CI: 1.6–2.4%) estimate from the UK-based ECHOES (Echocardiographic Heart Of England Screening) study with access to ECG data. [55]
The ability to accurately ascertain AF events using EHR represents an important research opportunity, which leads me on to the overall thesis aim, objectives and hypotheses I address.

1.4 Overall thesis aim, objectives and hypotheses

The overall aim and specific objectives of this thesis are the following:

1.4.1 Aim

To exploit EHRs to investigate, validate and extend evidence for AF risk factors, subtypes, and outcomes

1.4.2 Specific objectives and hypotheses

- **Objective:** To conduct a systematic review and field synopsis of the existing observational epidemiology on the associations of 23 cardiovascular risk factors with incident AF

  Hypothesis: I hypothesise that a range of demographic, behavioural and biological factors that are known to increase the risk of developing other CVDs like myocardial infarction and stroke will also increase the risk of developing AF.

- **Objective:** To use the CALIBER dataset to model the associations of 23 cardiovascular risk factors with two novel AF endpoints: AF with and without intercurrent CVD

  Hypothesis: I hypothesise that observational associations between cardiovascular risk factors and incident AF may be mediated by intercurrent diagnoses of CVD.

- **Objective:** To create EHR definitions for eight AF subtypes relevant to the 2016 ESC guidelines, which are: (1) structural, (2) focal, (3) polygenic, (4) postoperative, (5) valvular, (6) AF in athletes, (7) monogenic and (8) respiratory AF.

  Hypothesis: I hypothesise that EHR definitions can be developed in order to identify a range of diverse AF subtypes; however in the absence of genomic information and ECG images it may be infeasible to create definitions for polygenic and focal AF.

- **Objective:** To use the CALIBER dataset to implement, improve and validate the EHR definition for valvular AF.

  Hypothesis: I hypothesise that the EHR definition I created for valvular AF will show clinical validity in replicating known associations between prosthetic heart valves, mitral valve stenosis and an increased risk of stroke, systemic embolism and mortality and that there may be other valve diseases associated with poorer prognosis.

- **Objective:** To use the CALIBER dataset to investigate stroke outcomes by CHA_2DS_2-VASc, sex and warfarin use.

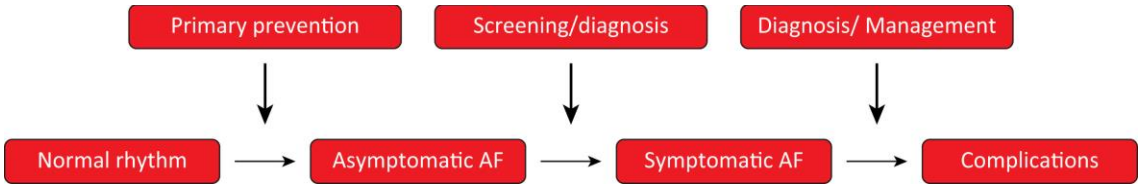
  Hypothesis: I hypothesise that the stroke rate estimates I obtain from a population of individuals with AF diagnosed in one of two clinical settings (i.e. in either primary care or secondary care) may be different from prior reports, which have focussed exclusively on individuals with AF diagnosed in secondary care.

1.5 Chapter summary

In summary this introduction chapter outlines the motivation and specific objectives of this PhD thesis including the clinical importance of AF, current clinical uncertainties in AF risk factors, subtypes and outcomes and the opportunity for studying these using EHRs. In **chapter 2**, which follows, I present the findings of a systematic review and field synopsis I conducted into the associations of 23 cardiovascular risk factors and incidence of AF. Given the lack of focus on AF in the primary prevention setting, this review helps to identify risk factors with preventive potential to prioritise in future research.

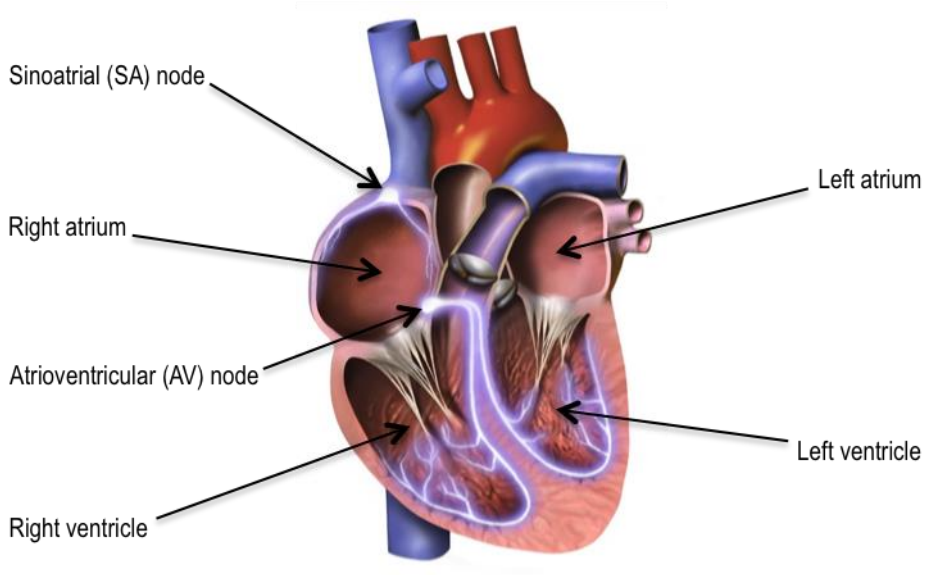
1.6 Chapter figures

Figure 1.1 Onset and progression of atrial fibrillation and opportunities for intervention.



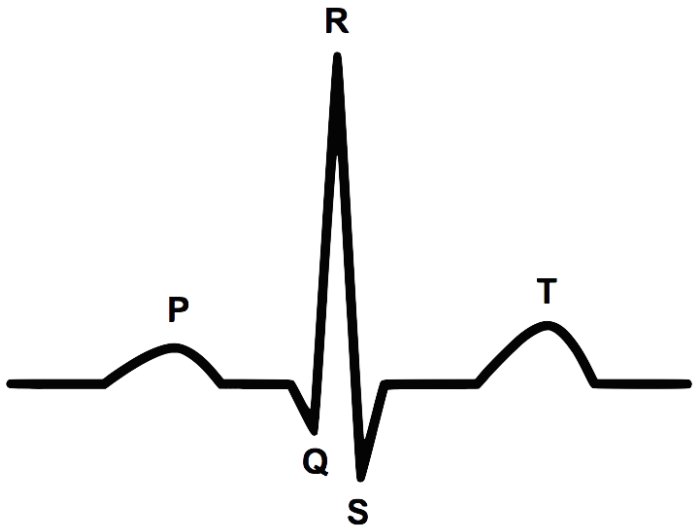
Notes: Figure from Murphy A, Banerjee A, Breithardt G, Camm AJ, Commerford P, Freedman B, Gonzalez-Hermosillo JA, Halperin JL, Lau CP, Perel P, Xavier D, Wood D, Jouven X, Morillo CA. The World Heart Federation Roadmap for Nonvalvular Atrial Fibrillation. Glob Heart. 2017 Dec;12(4):273-284. doi: 10.1016/j.gheart.2017.01.015. [Reused with permission]

Figure 1.2 Illustrative diagram of normal heart rhythm electrophysiology, which initiates in the sinoatrial node located in the right atrium



Notes: figure adapted from BruceBlaus [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)], via Wikimedia Commons

Figure 1.3 Electrocardiographic characteristics of normal heart rhythm: P, Q, R, S and T waves corresponding to electrical impulses stimulating atrial contract, ventricular contraction and ventricular relaxation

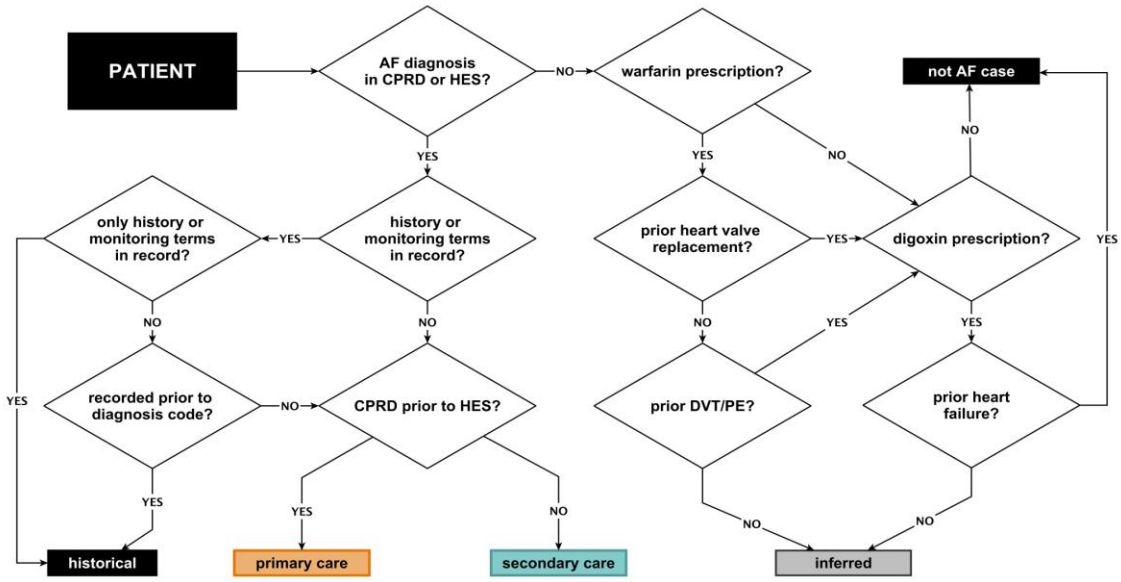


Notes: figure adapted from ECG-PQRST+popis.svg: *SinusRhythmLabels.svg: Created by Agateller (Anthony Atkielski), converted to svg by atom. derivative work: Kychot (talk) derivative work: Kychot (ECG-PQRST+popis.svg) [Copyrighted free use], via Wikimedia Commons

Figure 1.4 Electrocardiogram tracing comparing normal heart rhythm (top) with characteristic P, QRS and T waves and atrial fibrillation (bottom) with an absence of P waves and QRS waves appearing at irregular intervals

Notes: figure adapted from BruceBlaus [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)], via Wikimedia Commons

Figure 1.5 Electronic health record algorithm for atrial fibrillation as defined by Morley and colleagues, which incorporates coded diagnoses with inferred diagnoses based on warfarin prescriptions in the absence of thromboembolic disease



Notes: figure from Morley KI, Wallace J, Denaxas SC, Hunter RJ, Patel RS, Perel P, et al. (2014) Defining Disease Phenotypes Using National Linked Electronic Health Records: A Case Study of Atrial Fibrillation. PLoS ONE9(11): e110900. https://doi.org/10.1371/journal.pone.0110900

Chapter 2

Systematic review and field synopsis of the link between 23 cardiovascular risk factors and incidence of atrial fibrillation

2.1 Chapter outline

In this chapter, I describe the findings of a systematic review and field synopsis I conducted into the associations of 23 cardiovascular risk factors and incidence of atrial fibrillation (AF), which has been published in *Thromb Haemost* (with CC-BY-NC-ND 4.0 licence).[2] Given the lack of focus on AF in the primary prevention setting, this review helps to identify risk factors with preventive potential to prioritise in future research. I hypothesised that a range of demographic, behavioural and biological factors that are known to increase the risk of developing other CVDs like myocardial infarction and stroke will also increase the risk of developing AF. I used a novel field synopsis methodology to synthesise the results of the literature search. Field synopses, unlike meta-analyses, are concerned with bringing together all of the available evidence, regardless of diverse study design features, and evaluating (1) the overall amount of evidence, (2) the extent of replication and (3) the quality and likelihood of bias.[56] A key measure of study quality was whether previous reports had accounted for intercurrent cardiovascular disease (CVD) in the development of AF. This is important to understand whether risk factors lead directly to AF, or if risk factors lead to CVD, which in turn leads to AF. Collaborator contributions for this work are reflected in the text.

List of collaborators: Shohreh Honarbakhsh, Juan-Pablo Casas, Joshua Wallace, Ross Hunter, Richard Schilling, Pablo Perel, Katherine Morley, Amitava Banerjee, and Harry Hemingway.

2.2 Abstract

Background Established primary prevention strategies of CVDs are based on understanding of risk factors, but whether the same risk factors are associated with AF remains unclear.

Methods I conducted a systematic review (Pubmed to October 2015) and field synopsis of population-based, consented or electronic health record (EHR) cohorts that investigated the associations of one or more of 23 cardiovascular risk factors, and incident AF. For each risk factor I extracted relative risks (RR) and 95% confidence intervals [95% CI], and extent of risk factor adjustment. I used forest plots to visualise the number of reports with inverse (RR [95% CI] <1.00), or direct (RR [95% CI] >1.00) associations.

Results Overall 73 publications were included (84 reports based upon 28 consented and 4 EHR cohorts), with 576,602 AF events in 20,420,175 participants. The number of reports ranged from 3 to 19 (median 10) per risk factor, with 66 (78·6%) published in 2010–2015. I found substantial heterogeneity in AF event definition, and quality of reporting with age range not reported in 30 reports (35·7%), lack of adjustment for six standard CVD risk factors in 63

reports (75·0%), and lack of adjustment for intercurrent CVD in 69 reports (82·1%). For alcohol intake, I identified 10 reports, 10 disparate alcohol definitions, and only 3 reports which showed a direct association. Hypertension (13/17 reports) and obesity (19/19 reports) showed direct associations, while 4 other factors showed associations with AF in the opposite direction of known associations with CHD. These were inverse associations for non-White ethnicity (5/5 reports, with RR from 0.35 to 0.84 [0.82-0.85]), total cholesterol (4/13 reports from 0.76 [0.59-0.98] to 0.94 [0.90-0.97]; 8/13 reports with non-significant inverse associations), and diastolic blood pressure (2/11 reports from 0.87 [0.78-0.96] to 0.92 [0.85-0.99]; 5/11 reports with non-significant inverse associations), and direct associations for taller height (7/10 reports from 1.03 [1.02-1.05] to 1.92 [1.38-2.67]).

Conclusion A systematic evaluation of the available evidence suggests similarities as well as important differences in the risk factors for incidence of AF as compared with other cardiovascular diseases, which has implications for the primary prevention strategies for AF.

2.3 Introduction

AF is the world's most common heart rhythm disorder, affecting 33.5 million people globally in 2010.[8] AF accounts for 1 in 4 ischaemic strokes,[14] doubles the risk of death,[15] places an economic burden on healthcare systems,[57] and is projected to affect twice as many people over the next 50 years.[9 10] Yet to date, there have been no clinical trials of healthy participants without CVD, and with AF as the primary outcome.[35] The focus of trials has instead been on prevention of stroke and thromboembolism after diagnosis of AF. Community screening programmes for detection of AF,[58] are also designed to identify patients at high risk of stroke and thromboembolism, and do not identify those who are at an initially high risk of later developing AF. Thus, current clinical guidelines make no recommendations for the primary prevention of AF itself, among people without CVDs.[3 19 22]

Established primary prevention strategies of other CVDs, such as coronary heart disease (CHD),[59] and stroke,[60] are based on understanding of risk factors, but the extent to which the same risk factors are associated with the incidence of AF is not fully understood. Ultimately, it is not known whether existing CVD prevention strategies can also work in preventing AF, or whether there may be important clinical differences. In synthesising available evidence the conventional (near universal) approach is to examine risk factors one at a time. Single risk factor systematic reviews and meta-analyses have been carried out for alcohol,[61-63] C-reactive protein,[64] diabetes mellitus,[65] obesity,[66] physical activity,[67 68] and renal function[69] in relation to AF risk. Each of these reviews uses non-identical methods, for example varying in the extent to which incident AF is analysed among people free from pre-existing CVD. While there is an important ongoing role for the vertical approach of a single risk factor meta-analysis (particularly if methods can be aligned), there is also a complementary role for a horizontal 'field synopsis' approach across multiple potential risk factors. The term field synopsis is defined as a systematic evaluation of evidence in which the (i) overall amount, (ii) extent of replication, and (iii) protection from bias is considered across the whole field.[56 70] One advantage of a field synopsis in

multifactorial diseases is to provide an unbiased empirical basis for prioritising further research into risk factors with preventive potential.

I therefore conducted a systematic review and field synopsis of the associations of a wide range of demographic, behavioural, and biological CVD risk factors and incidence of AF among population based cohorts. Field synopses of cumulative evidence.[56 70] are common in genetics but have seldom been applied in the context of preventive medicine. My objectives were (i) to determine the amount of evidence for each risk factor, (ii) to evaluate the extent to which each risk factor shows concordant or discordant associations with AF incidence across independent study populations, and (iii) to systematically appraise the quality of the observational evidence across the field of AF prevention research.

2.4 Methods

My approach to the search, selection, data collection and analysis of reports was systematic, and guided by the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) checklist which is provided in **table S2.1** in **appendix**.[71]

2.4.1 Search strategy

I queried the PubMed database using the search terms listed in **table S2.2** in **appendix**, for original research reports that were published in English up to 1 October 2015; involving prospective, population based cohorts in which the proportion of people with diagnosed CVD at baseline was either zero (because of exclusions) or low reflecting prevalence in the general population (hereafter referred to as population based cohorts). Cohorts were of any age and without prior AF and investigated the association between "risk factors" and incident AF, over any follow–up period using Cox proportional hazards or Poisson regression models which were adjusted or stratified for age and sex as a minimum. The 23 cardiovascular risk factors listed in **table 2.1** were shortlisted for inclusion in the review based on clinical relevance as an established predictor or treatment target in the prevention of CVD,[59] on clinical opinion of an association with AF[22] and on the expert suggestions of collaborators. Reference lists of identified reports, existing reviews and meta-analyses (which were not restricted to prospective cohorts of individuals either free from or with general population levels of baseline CVD: for alcohol,[61-63] C-reactive protein,[64] diabetes mellitus,[65] obesity,[66] physical activity,[67 68] and renal function[69]) were hand-searched for additional reports. I, together with two other collaborators (Joshua Wallace and Shohreh Honarbakhsh), reviewed the inclusion of each report based on title, then abstract, then full–text. Disagreements were resolved by joint full-text review with a third independent reviewer (Ross Hunter).

2.4.2 Data extraction

From each included report I extracted the following information: design of cohort (consented participant cohort with research measures at baseline and follow up, or EHR cohort in which anonymised data collected as part of usual clinical care was used for baseline and follow–up measures), country, sample size (number of participants at baseline) and number of AF events

over follow–up (based on the highest figure reported), age range, proportion of female participants, mean or median follow–up, methods of AF ascertainment, risk factor definition, statistical model, and risk factors used in adjustment. I extracted data on whether cardiovascular events, prevalent at baseline and incident during follow–up and preceding AF were accounted for. For each risk factor, I extracted adjusted RR, and 95% CI. Where there were multiple RR reported within a publication or across multiple publications from the same cohort, I selected the most adjusted estimate, modelled with the highest number of AF cases.

2.4.3 Summary and visualisation of risk factor associations

I summarised the overall results of the field of cohort epidemiology of AF by plotting the number of reports with inverse (RR<1.00), null or mixed (RR=1.00 or shows opposite associations among subpopulations), or direct relationship (RR>1.00) with AF incidence. I regarded the association as significant if the 95% CI did not cross 1.00. Unless stated, RR are given as originally reported. For each factor, I then plotted the RR and 95% CI using statistical software R version 3.2.0.

2.4.4 Summary and visualisation of quality of reporting and analysis

I summarised the quality of reporting by completeness of the items listed in the above data extraction section (items not reported (NR) are clearly indicated in tables and figures). I summarised the quality of analysis by assessment of the number (%) of adjustment made for the 23 risk factors, and whether adjustment was made for 6 standard CVD risk factors (age, sex, smoking, blood pressure, lipids and diabetes mellitus), and for prevalent and incident CVD events. I visualised these as “Swiss cheese” plots.[72]

2.5 Results

2.5.1 Characteristics of included reports

Overall 73 out of 2777 publications were included with a total of 84 reports based upon 32 independent cohorts from 10 countries and 20,420,175 participants (flow diagram in **figure S2.1** in **appendix**).[40 63 73-143] As **table 2.2** shows 28 cohorts (87.5%) involved consented participants with 39,900 (6.9%) events and 4 cohorts (12.5%) were based on EHR populations with 536,702 (93.1%) events. AF events were ascertained from a research or healthcare electrocardiogram (12 reports (14.3%)), diagnosis codes from medical records (37 reports (44.1%)), or using a combination of both methods (35 reports (41.7%)). As **table S2.3** in **appendix** shows, 17 reports (20.2%) described using two out of four types of medical records (i.e. general practitioner, hospital care, prescriptions, or mortality records), but no report used three or all four types combined.

2.5.2 Quality of reporting

Age range was not reported in 30 reports (35.7%), mean or median follow-up in 18 reports (21.4%), and risk factor definition was not reported in 9 reports (10.7%). Information was consistently reported on country, sample size, female participants, and AF events.

2.5.3 Quality of analysis

Overall, 63 reports (75.0%) lacked adjustment for all six standard CVD risk factors (**table S2.4** in **appendix**). Age was adjusted for in 84 reports (100.0%), sex in 80 reports (95.2%), smoking in 49 reports (53.3%), blood pressure in 63 reports (75.9%), lipids in 32 reports (38.1%), and diabetes mellitus in 59 reports (70.2%). The total number of adjustment factors ranged from 2 to 14 factors, with a median of 8 factors. There was lack of adjustment for prevalent CVD in 30 reports (35.7%), and for incident CVD in 69 reports (82.1%).

2.5.4 Associations of 23 risk factors and incidence of AF

A summary of the heterogeneity of associations of 23 risk factors and incidence of AF are visualised in **figure 2.1**, and for each factor separately in **figures 2.2 to 2.6** and **figures S2.2 to S2.19** in **appendix**.

- **Demographic factors**

For age, all 15 reports showed significant direct associations, but these were heterogeneous. RR [95%CI] ranged from 1.02 [1.01–1.03] to 1.14 [1.10–1.18] for every 1–year, from 1.43 [1.29–1.59] to 1.65 [1.57–1.74] for every 5–year, from 1.09 [1.09–1.09] to 2.35 [2.03–2.72] for every 10–year, and from 1.36 [1.27–1.45] to 4.34 [3.72–5.07] for every standard deviation (NR) year increase in age (**figure S2.2**).[133,143,81,40,40,74,89,100,115,139,135,83,78,95,112] For men (compared to women), 1 report showed a significant inverse association (0.70 [0.50–0.90]),[124] 2 reports were inverse but non-significant (from 0.95 to 0.96),[143,133] and 8 reports showed significant direct associations (from 1.45 [1.29–1.63] to 1.90 [1.58–2.29]) (**figure S2.3**).[83,40,89,139,40,95,100,115] For African American, Asian, Chinese, Hispanic and Non-Hispanic Black (compared to White) ethnicities, all 5 reports showed significant inverse associations (from 0.35 [NR–NR] to 0.84 [0.82–0.85]).[130,90,90,74,137] Only 1 country reported estimates for the association of ethnicity and incidence of AF (**figure 2.2**). For socio-economic status, 2 reports showed significant inverse associations (from 0·91 [0·86-0·96] to 0·98 [0·98-0·99]),[74,139] 3 were inverse but non-significant (from 0·88 to 0·98),[101,93,84] and 1 showed a mixed association (**figure S2.4**).[81]

- **Health behaviours**

For current smoking, 1 report was inverse but non-significant (0.78),[81] 1 report showed a mixed association,[40] 5 reports were direct but non-significant (from 1.01 to 1.20),[128,133,101,115,99] and 6 reports showed significant direct associations (from 1.32 [1.19–1.46] to 2.00 [1.40–2.80]) (**figure S2.5**).[74,83,40,123,86,124] For physical activity, 3 reports showed significant inverse associations,[94,79,73] 4 reports were inverse but non-significant,[135,92,131,104] 2 reports showed null or mixed associations,[75,111] and 2 reports were direct but non-significant (**figure S2.6**).[101,82] For alcohol intake in drinks per day or week, in grams per day or week, or for current alcohol drinkers, 2 reports showed significant inverse associations (from 0.65 [0.45–0.94] to 0.96 [0.93–0.99]), [128,98] 1 report was inverse but non-significant (0.97),[92] 1 report showed a null association,[74] 3 reports were direct but non-significant (from 1.04 to 1.20),[81,115,124] and 3 reports showed significant direct associations (from 1.39 [1.22–1.58] to 2.90 [1.61–5.23]).[63,109,133] All 10 alcohol reports de-

fined alcohol intake differently, and as shown for the 3 direct alcohol associations, the increased risk of developing AF was only among the highest alcohol intake categories (**figure 2.3**).

- **Blood pressure**

For every 10–22mmHg increase in systolic blood pressure, or systolic blood pressure ≥160mmHg, 1 report showed a null association,[124] 5 reports were direct but non-significant (from 1.01 to 1.24),[81,100,128,40,129] and 8 reports showed significant direct associations (from 1.14 [1.05–1.25] to 2.63 [1.83–3.78]; **figure S2.7**).[95,40,135,114,136,101,110,92] For every 10–11mmHg increase in diastolic blood pressure, or diastolic blood pressure ≥95–100mmHg, 2 reports showed significant inverse associations (from 0.87 [0.78–0.96] to 0.92 [0.85–0.99]),[114,95] 5 reports were inverse but non-significant (from 0.82 to 0.99),[129,128,40,136,100] 2 reports were direct but non-significant (from 1.02 to 1.23),[40,110] and 2 reports showed significant direct associations (from 1.24 [1.10–1.40] to 2.02 [1.20–3.41]).[135,92] No EHR cohorts reported estimates for the association of diastolic blood pressure and incidence of AF (**figure 2.4**). For hypertension, 1 report was inverse but non-significant (0.93),[133] 3 reports were direct but non-significant (from 1.21 to 1.37),[100,124,81] and 13 reports showed significant direct associations (from 1.28 [1.08–1.51] to 2.60 [1.60–4.40]) (**figure S2.8**). [95,83,112,74,136,40,40,77,115,86,101,143,132]

- **Lipid profile**

For every 10–50mg/dl increase in total cholesterol, or total cholesterol ≥220–280mg/dl, 4 reports showed significant inverse associations (from 0.76 [0.59–0.98] to 0.94 [0.90–0.97]),[106,98,40,78] 8 reports were inverse but non-significant (from 0.57 to 0.99), [81,87,112,128,133,101,40,116] and 1 report was direct but non-significant (1.13).[116] Both inverse and direct associations were shown in the 3 total cholesterol reports that adjusted for prevalent and incident CVD events (**figure 2.5**). For every 10–40mg/dl increase in low–density lipoprotein cholesterol, or low–density lipoprotein cholesterol ≥150mg/dl, 2 reports showed significant inverse associations (from 0.72 [0.56–0.92] to 0.92 [0.88–0.96]),[106,78] 4 reports were inverse but non-significant (from 0.85 to 0.95),[87,100,128,116] and 1 report was direct but non-significant (1.15) (**figure S2.9**).[116] For every 15mg/dl increase in high–density lipoprotein cholesterol, or high–density lipoprotein cholesterol ≥60mg/dl, 5 reports were inverse but non-significant (from 0.85 to 0.98),[116,40,116,78,40] 2 reports showed null or mixed associations,[87,40] 2 reports were direct but non-significant (from 1.01 to 1.07),[128,106] and 1 report showed a significant direct association (1.16 [1.04–1.28]) (**figure S2.10**).[112] For triglycerides, 3 reports were inverse but non-significant (from 0·83 to 0·98),[106,128,78] 1 showed a mixed association,[40] 2 were direct but non-significant (from 1·02 to 1·02),[87,40] and 3 showed significant direct associations (from 1·09 [1·01-1·17] to 1·16 [1·02-1·33]) (**figure S2.11**).[40,116,116]

- **Diabetes mellitus, renal function**

For diabetes mellitus (type unspecified), 2 reports were inverse but non-significant (from 0.86 to 0.98),[143,128] 8 reports were direct but non-significant (from 1.02 to 1.49),[83,115,99,103,40,40,112,101] and 6 reports showed significant direct associations (from 1.17 [1.16–1.19] to 1.80 [1.30–2.60]) (**figure S2.12**).[140,95,133,86,74,124] For renal function, 3 reports were inverse but non-significant (from

0·66 to 0·93),[40,97,128] 5 were direct but non-significant (from 1·02 to 1·36),[76,143,40,40,107] and 3 showed significant direct associations (from 1·78 [1·49-2·13] to 3·41 [1·50-7.76]) (**figure S2.13**).[139,102,91]

- **Anthropometric factors**

For every 1–10cm increase in height, or height ≥173cm, 3 reports were direct but non-significant (from 1.14 to 1.17), [115,112,40] and 7 reports showed significant direct associations (from 1.03 [1.02–1.05] to 1.92 [1.38–2.67]) (**figure 2.6**).[98,101,80,124,40,134,92] For weight, all 8 reports showed significant direct associations (from 1·17 [1·04-1·31] to 1·55 [1·28-1·87]) (**figure S2.14**).[40,40,40,40,80,124,134,40] For every 1–10kg/m^2 increase in body mass index, or body mass index ≥25–30kg/m^2, all 19 reports showed significant direct associations (from 1.04 [1.02–1.05] to 2.24 [1.41–3.58]) (**figure S2.15**).[105,101,100,136,93,74,126,135,112,115,121,134,80,128,83,124,77,85,133]

- **Inflammatory biomarkers**

For C–reactive protein, 4 reports were direct but non-significant (from 1·01 to 1·07),[128,120,118,113] and 4 reports showed significant direct associations (from 1·11 [1·02-1·20] to 1·24 [1·11-1·40]) (**figure S2.16**).[108,120,100,95] For fibrinogen, 2 reports were inverse but non-significant (from 0·94 to 0·98),[118,113] 1 report was direct but non-significant (1·07),[119] and 3 reports showed significant direct associations (from 1·10 [1·02-1·20] to 2·30 [1·34-3·95]) (**figure S2.17**).[108,88,125]

- **Thyroid function, autoimmune disease**

For every 1.0mU/L decrease in thyroid stimulating hormone, or thyroid stimulating hormone <0.10–0.45mU/L, 1 report was inverse but non-significant (0.34),[127] 5 reports were direct but non-significant (from 1.06 to 2.85),[96,122,127,127,127] and 2 reports showed significant direct associations (from 1.41 [1.25–1.59] to 3.10 [1.70–5.50]) (**figure S2.18**).[141,117] For autoimmune diseases, 1 report showed a significant direct association for coeliac disease (1·33 [1·23-1·43]),[142] 1 showed a significant direct association for rheumatoid arthritis (1·43 [1·33-1·53]),[138] and 1 showed a significant direct association for mild (1·22 [1·14-1·30]) and severe psoriasis (1·53 [1·23-1·91]) (**figure S2.19**).[139]

2.6 Discussion

As far as I am aware, this is the first example of a field synopsis evaluating associations across multiple risk factors and disease incidence. I systematically evaluated 84 reports from 32 independent cohorts for the impact of 23 cardiovascular risk factors in the development of AF. Unlike previous reviews (e.g. for position papers[3] [144]), I focussed exclusively on primary prevention among populations initially free from diagnosed CVD or general populations in which baseline levels of CVD reflected prevalence in the general population. I found some evidence that ethnicity, height, diastolic blood pressure and serum cholesterol, are associated with AF incidence in opposite directions to their known associations with CHD and stroke. Furthermore I found only modest evidence for the widely held clinical opinion that excess alcohol is associated with increased risk of AF. Taken together these findings suggest that a primary prevention

strategy for AF may require some different elements from the current strategies used for other CVDs.

Concordant associations

For some risk factors, namely hypertension and higher body mass index, there were consistent, direct associations with incidence of AF, as there are for CHD. This could reflect a causal link with AF, or that the risk factor causes CHD, which in turn causes AF. Surprisingly, I found that only 3 (out of 14) reports investigating the association between systolic blood pressure and incident AF accounted for both prevalent and intercurrent incident cardiovascular events, and only 1 of which reported a significant direct association. Several post hoc analyses of trials have suggested a possible benefit of angiotensin-converting-enzyme inhibitors/angiotensin receptor blockers,[145] and other blood–pressure lowering medications,[146] in the prevention of AF. However, as I have shown here the available observational evidence on the associations of 23 cardiovascular risk factors and incidence of AF, does not fully consider a mechanism of confounding or mediation by other intercurrent CVDs.

Current clinical guidelines include alcohol in a list of potentially "reversible" causes of AF, but acknowledge that there is no evidence to suggest that addressing any of these causes is effective in preventing AF.[22] Three earlier reviews by Samokhvalov and colleagues to April 2009,[61] Kodama and colleagues to January 2009,[62] and Larsson and colleagues to January 2014[63]) have all reported a dose-response relationship between alcohol and AF. I, on the other hand, found only a small number of reports (3 out of 10) showing a statistically significant direct association but can offer at least six possible explanations as to why the findings I report here are different. First, I considered only prospective studies whereas Samokhvalov and Kodama included retrospective studies. Second, I considered only general population cohorts, while Larsson included one cohort with pre-existing CVD. Third, I considered only incident AF events, while Kodama included studies on AF recurrence. Fourth, I considered only estimates from Cox or Poisson regression models, whereas Samokhvalov, Kodama and Larsson all included estimates from cross-sectional logistic regression. Fifth, I considered only the most adjusted alcohol estimate per cohort while Samokhvalov included the study with the most comprehensive alcohol data and Larsson did not report an approach to selecting from multiple estimates per cohort. And sixth, by running an updated literature search to October 2015 and combining results using a more inclusive field synopsis method I included 8 additional reports that have not been involved in the previous reviews.[74 81 92 98 115 124 128 133] Based on the 3 statistically significant direct alcohol associations that I identified, the increased risk of developing AF was confined to the highest levels of alcohol intake, as opposed to there being a J–shaped or dose–response relationship. Overall, these findings indicate that at present, there is limited consistent evidence from observational studies on which recommended alcohol intake levels for the primary prevention of AF could be based.

Discordant associations

I found some evidence that white ethnicity, a taller height, lower total cholesterol and lower diastolic blood pressure might confer a higher risk of incident AF, which is in the opposite direction to their known associations with incident CHD.[59] The cholesterol finding suggests that reducing lipid levels may not be relevant for the primary prevention of AF, which is in line with an existing meta–analysis of trial evidence that did not support the role of statins for prevention of AF in participants with underlying CVD.[147] Previously, it has been shown that blood pressure has markedly different associations with the incidence of twelve different CVDs (not including AF).[148] And this review now provides some, albeit mixed, evidence that this may also be the case for AF. The direct and inverse associations shown for systolic and diastolic blood pressure respectively, may indicate high pulse pressure, which is a marker of arterial stiffness and is more prevalent in older populations.[149] Two prior studies have reported an association between pulse pressure and incidence of AF,[114 129] however pulse pressure was not considered in this review as its clinical utility is not well defined.[150]

Clinical implications

The observational cohort evidence summarised here suggests that programmes for the primary prevention of AF may need to differ slightly from those which have guided clinicians and public health practitioners in the primary prevention of other CVDs. Existing management strategies to tackle obesity, smoking, alcohol and hypertension may have a role but the current evidence is insufficient to design an AF specific primary prevention programme. The risk factors included in available risk prediction tools for development of AF are supported by my systematic review, and these tools should be used more frequently in clinical practice.[40 115] Such risk prediction tools may be used to identify high–risk individuals for inclusion in primary prevention trials in AF, where there is the largest knowledge gap.

Overall characteristics of the field

I systematically evaluated and visualised the field of cohort epidemiology of AF. Although systematic reviews across multiple risk factors have been used for the global burden of disease estimation, and in genetics,[56] the field synopsis approach has seldom been applied in the context of preventive medicine. Overall, I found a relatively "young" field, which has been rapidly expanding over the last five years (see **figure S2.20**). Although my review included 32 cohorts of 20 million participants and 600,000 AF events, I found a limited number of reports (between 3 and 19) per risk factor, and evidence to suggest that there is unpublished risk factor data. Whereas the majority of consented cohorts measure blood pressure, I found reports for less than half (only 14) of the included cohorts. Although I identified some efforts at pooling studies (e.g. the CHARGE–AF consortium of 5 cohorts, 3 countries, and 1771 AF events[40]), the amount of evidence available is markedly smaller than the scale of cohort evidence available on risk factors for CHD or stroke incidence. By way of comparison, the Emerging Risk Factor Collaboration consists of over 100 prospective, population–based cohorts.[151] I found that the AF field is dominated by North American and North European cohorts, which is consistent with other cardiovascular and non-communicable diseases.[152] However, in the interest of reducing

the global burden of AF, there is currently no global data on which to base a global primary prevention strategy.[153] Next, I found that the AF field is beginning to span both consented population and EHR studies, with all 7 EHR reports published in the years 2011 to 2015. In the era of big data research, EHRs offer the potential for studying associations at much larger scale, at population–level, in comparison with other risk factors, and across a wide range of diseases.[5] However none of the included EHR cohorts reported associations for numerical clinical data such as blood pressure, lipids, and body mass index and instead only reported coded risk factor data. Finally, I found considerable heterogeneity in study design and reporting, and a lack of consistent approach to adjustment for other risk factors (visualised as a "Swiss cheese"). Field synopses allow for differences in study designs, however in order to further inform primary preventive programmes and estimate the precise relative risk estimates in meta-analyses; there is a need for large-scale strategic co-ordination of the field of AF prevention research.

Strengths and Limitations

The work I present here is subject to the inherent limitations of systematic reviews and field synopses.[70] A principal strength – evaluation across a comprehensive range of risk factors – is also the principal weakness. As in order to evaluate the breadth of the field there is a necessary restriction in the depth of analysis of any one risk factor, or relations between them. Most notably as I only searched the PubMed database, it is possible that I may have missed relevant studies. I therefore conducted a sensitivity analysis comparing search results in PubMed with that of Embase for the year 2013, which is the median year between 2010 and 2015 when the majority (75%) of included reports were published. As **table S2.5** shows, I found no further eligible studies in Emabse, which is consistent with other reports showing limited additional value of searching biomedical databases beyond PubMed.[154 155] There are of course other publications in support of searching multiple databases to identify further studies.[156 157] However, as I did not perform meta-analysis, I have not introduced any computational bias in to the present work and therefore consider the results and conclusions unlikely to change. Field synopses provide a systematic foundation, unbiased by a particular interest in one or more risk factors,[158] for hypothesis generation and further research. One example of how this work could be taken forward would be to evaluate the extent to which findings in relation to ethnicity, height and lipids[159] are inter-related.

2.7 Conclusions

A systematic evaluation of the available evidence suggests similarities as well as important differences in the risk factors for AF as compared with other common CVDs like CHD and stroke. This has implications for the primary prevention of AF.

2.8 Chapter summary

To summarise, in this chapter I reviewed the current observational evidence for the associations of 23 cardiovascular risk factors in relation to incidence of AF highlighting similarities (e.g. hypertension and obesity) and differences (e.g. ethnicity and lipids) as compared to known asso-

ciations with other CVDs. Overall I found a relatively young field of research with a small number of reports for each risk factor, not to mention vast heterogeneity in the way risk factors and AF events were defined. Although most studies of AF risk factors to date have involved consented participant cohorts, cohorts formed from EHRs began to emerge in the last five years of the review, albeit without any numerical clinical data such as blood pressure, lipids, and body mass index. Thus in **chapter 3**, which follows, I present the CALIBER dataset linking EHRs from primary care, secondary care and mortality records in the UK. CALIBER data is used in the later analytic chapters of this thesis to derive novel insights into AF risk factors, subtypes and outcomes. As I go on to describe, CALIBER offers some unique advantages for studying AF.

2.9 Chapter tables

Table 2.1 List of 23 cardiovascular risk factors investigated for associations with incident atrial fibrillation in population based cohorts

Demographic factors	
1	Age
2	Sex
3	Ethnicity
4	Socio–economic status
Health behaviors	
5	Smoking
6	Physical activity
7	Alcohol intake
Blood pressure	
8	Systolic blood pressure
9	Diastolic blood pressure
10	Hypertension
Cholesterol	
11	Total cholesterol
12	Low–density lipoprotein cholesterol
13	High–density lipoprotein cholesterol
14	Triglycerides
Metabolic	
15	Diabetes mellitus
16	Renal function
Anthropometry	
17	Height
18	Weight
19	Body Mass Index
Inflammation	
20	C–reactive protein
21	Fibrinogen
Thyroid/autoimmunity	
22	Thyroid function
23	Autoimmune diseases

Table 2.2 Characteristics of reports included in systematic review and field synopsis, sorted by cohort and number of atrial fibrillation events

Cohort	Country	Age range	Sample size	Women (%)	Mean / median follow–up	ECG	medical records	self–reports	AF events	Risk factors	References
Consented population cohorts:											
WHI–OS	United States	50–79	81317	100	11.5	○	●	○	9792	physical activity	73
		50–79	81892	100	9.8	○	●	○	8252	age ethnicity SES smoking alcohol hypertension diabetes BMI	74
COSM	Sweden	45–79	44410	0	12.0	○	●	○	4568	physical activity	75
		45–83	43841	0	10.9	○	●	○	4488	alcohol	63
NPMS	Japan	20–NR	223877	68	5.9	●	○	○	2974	renal	76
		20–NR	28449	66	4.5	●	○	○	265	hypertension BMI	77
		20–NR	28449	66	4.5	●	○	○	265	age total chol. LDL HDL triglycerides	78
SMC	Sweden	49–83	36513	100	12.0	○	●	○	2915	physical activity	79
		45–83	35178	100	10.9	○	●	○	2757	alcohol	62
DCHS	Denmark	50–64	55273	52	13.5	○	●	○	2581	height weight BMI	80
		50–64	47589	53	5.7	○	●	○	553	age SES smoking alcohol SBP hyper-tension total chol.	81
		50–64	38400	49	5.7	○	●	○	418	physical activity	82
MPP	Sweden	26–61	30865	32	23.3	○	●	○	2312	age sex smoking hypertension diabetes BMI	83
ARIC	United States	45–64	14352	55	20.6	●	●	○	1794	SES	84
		45–64	14219	55	18.2	●	●	○	1775	BMI	85
		45–64	14598	55	17.1	●	●	○	1520	smoking diabetes hypertension	86
		45–64	13969	55	18.7	●	●	○	1433	total chol LDL HDL triglycerides	87
		45–64	14858	55	16.8	●	●	○	1209	fibrinogen	88
		45–65	15407	55	14.8	●	●	○	1085	age sex	89
		45–64	14419	55	16.0	●	●	○	1068	ethnicity	90
		45–64	10328	57	10.1	○	●	○	788	renal	91
		45–64	14546	55	NR	●	●	○	515	physical activity alco-hol SBP DBP height	92
		46–94	10675	57	NR	●	●	○	419	weight	40
CHS	United States	65–89	5685	58	11.2	●	●	○	1585	SES BMI	93
		65–NR	5365	57	10.0	●	●	○	1172	ethnicity	89
		65–NR	5446	58	8.7	●	●	○	1061	physical activity	94
		65–NR	5491	45	6.9	●	●	●	897	age sex SBP DBP hypertension diabetes CRP	95
		65–NR	2673	56	NR	●	●	○	812	thyroid	96
		65–NR	5043	60	NR	●	●	○	624	smoking HDL triglycerides weight	40
		65–NR	4321	59	7.4	●	●	●	579	renal	97
		65–NR	4844	58	3.3	●	●	●	304	alcohol total chol. height	98

MDCS	Sweden	44–73	30441	60	11.2	○	●	○	1430	smoking diabetes	99
		41–71	5135	59	14.0	○	●	○	284	age sex SBP DBP hypertension LDL BMI CRP	100
GPPS	Sweden	47–56	6903	0	NR	○	●	○	1253	SES smoking physical activity SBP hypertension total chol. diabetes height BMI	101
IPHS	Japan	40–79	132250	69	13.8	●	○	○	1232	renal	102
WHS	United States	45–NR	33372	100	16.4	●	●	○	1027	diabetes	103
		45–NR	34759	100	14.4	●	●	○	968	physical activity	104
		NR–NR	34309	100	12.9	●	●	○	834	BMI	105
		45–NR	23738	100	16.4	●	●	○	795	total chol. LDL HDL triglycerides	106
		45–NR	24746	100	15.4	●	●	○	786	renal	107
		45–NR	24734	100	14.4	●	●	○	747	CRP fibrinogen	108
		45–NR	34715	100	12.4	●	●	○	653	alcohol	109
		45–NR	34221	100	12.4	●	●	○	644	SBP DBP	110
NorPD	Norway	40–45	309540	52	NR	○	●	○	863	physical activity	111
TS	Norway	25–NR	22815	52	11.1	○	●	○	822	age hypertension total chol. HDL diabetes height BMI	112
		25–84	6315	51	10.9	○	●	○	566	CRP fibrinogen	113
FHS	United States	35–91	5331	55	16.0	●	○	○	698	SBP DBP	114
		45–95	4764	55	NR	●	○	○	457	age sex smoking alcohol hypertension diabetes height BMI	115
		30–87	2608	56	11.9	●	○	○	259	total chol. LDL HDL triglycerides	116
		60–NR	2007	59	NR	●	○	○	192	thyroid	117
		NR–NR	2863	55	6.2	●	●	○	148	CRP fibrinogen	118
		46–94	2838	55	NR	●	○	○	143	renal weight	40
MCS	Sweden	26–61	6031	0	25.0	○	●	○	667	fibrinogen	119
AGES	Iceland	46–94	4469	60	NR	●	●	○	408	age sex smoking SBP DBP hypertension total chol. HDL triglycerides diabetes renal height weight	40
		45–95	4467	60	NR	●	●	○	408	CRP	120
		45–95	4238	63	4.2	○	●	○	226	BMI	121
RS	Netherlands	45–NR	9166	57	6.8	●	●	○	402	thyroid	122
		55–NR	5668	65	7.2	●	●	○	371	smoking	123
		55–NR	3203	59	NR	●	●	○	177	age sex SBP DBP hypertension total chol. HDL triglycerides diabetes renal height weight	40
		45–95	3203	59	NR	●	●	○	177	CRP	120
CCHS	Denmark	40–79	18167	56	NR	○	●	○	379	sex smoking alcohol SBP hypertension diabetes height weight BMI	124
		20–NR	8410	58	7.5	○	●	○	268	fibrinogen	125
HABC	United States	70–79	2717	52	NR	○	●	○	371	BMI	126
		70–79	1850	52	8.1	●	○	○	17	thyroid	127

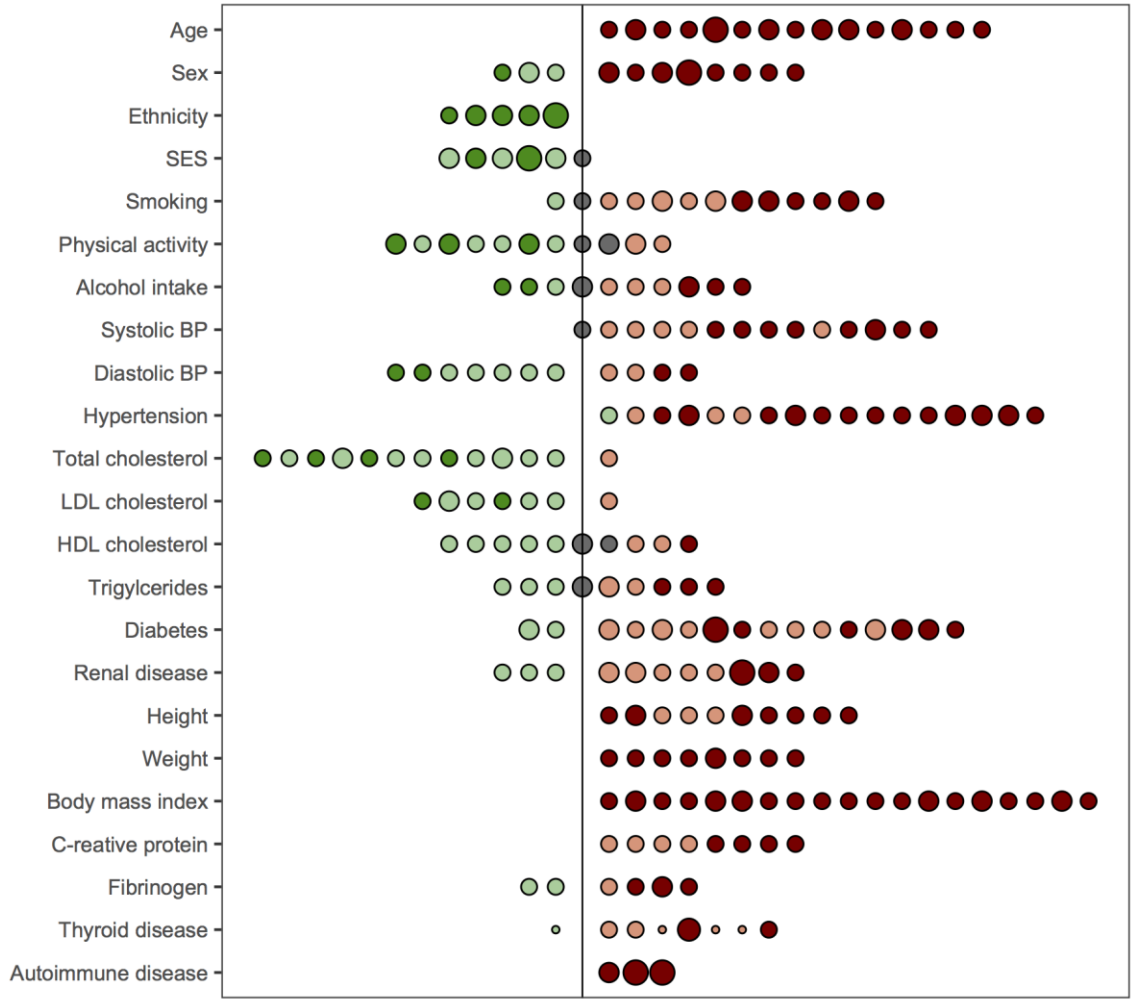
Cohort	Country	Age range	Sample size	Women (%)	Mean / median follow–up	ECG	medical records	self–reports	AF events	Risk factors	References
BHS	Australia	25–84	4267	56	NR	○	●	○	343	smoking alcohol SBP DBP total chol. LDL HDL triglycerides diabetes renal BMI CRP	128
		18–90	1048	48	20.0	○	●	○	14	thyroid	127
MESA	United States	45–84	6630	53	7.8	○	●	○	307	SBP DBP	129
		45–84	6721	53	7.0	○	●	○	305	ethnicity	130
		45–84	4534	52	8.2	○	●	○	221	total chol. LDL HDL triglycerides	116
		45–84	5793	53	7.7	○	●	○	199	physical activity	131
		45–84	5311	53	5.3	○	●	○	182	hypertension	132
CIRCS	Japan	30–80	7206	63	6.4	●	●	●	296	age sex smoking alcohol hypertension total chol. diabetes BMI	133
S–HS	Sweden	60–60	4021	52	13.6	○	●	○	285	height weight BMI	134
OCS	Norway	40–59	1997	0	30.0	●	●	○	270	age physical activity SBP DBP BMI	135
TSS	Japan	30–84	8360	53	12.8	●	●	●	253	SBP DBP hyperten-sion BMI	136
L85PS	Netherlands	85–85	420	64	5.2	●	○	○	39	thyroid	127
SHIP	Germany	20–81	2891	47	10.1	●	○	○	34	thyroid	127
	Participants:		1112394				AF events:		39900		
Electronic health record cohort:											
HCUP	United States	18–NR	13967949	57	3.2	○	●	○	375318	ethnicity	137
D–EHR	Denmark	16–NR	4182335	51	4.8	○	●	○	156484	autoimmune	138
		10–NR	4518484	49	9.2	○	●	○	126217	age sex SES renal autoimmune	139
		18–100	5081087	45	NR	○	●	○	115956	diabetes	140
		18–NR	586460	61	5.5	○	●	○	17154	thyroid	141
S–EHR	Sweden	00–95	170368	62	10.4	○	●	○	3859	autoimmune	142
T–NHIRD	Taiwan	18–NR	88377	61	NR	○	●	○	1041	age sex hyperten-sion diabetes renal	143
	Participants:		19307781				AF events:		536702		
Total participants:			**20420175**		**Total AF events:**				**576602**		

Abbreviations: AF – atrial fibrillation, SES - socioeconomic status, SBP – systolic blood pressure, DBP - diastolic blood pressure, total chol. – total cholesterol, LDL – low–density lipoprotein cholesterol, HDL – high–density lipoprotein cholesterol, BMI – body mass index, CRP – C-reactive protein, ● – yes, ○ – no.

Cohort abbreviations: WHI–OS – Women's Health Initiative Observational Study, COSM – Cohort of Swedish Men, NPMS – Niigata preventive medicine study, SMC – Swedish Mammography Cohort, DCHS – Diet Cancer and Health study, MPP – Malmö Preventive Project, ARIC – Atherosclerosis Risk in Communities, CHS – Cardiovascular Health Study, MDCS – Malmö Diet and Cancer study, GPPS – Göteborg Primary Prevention Study, IPHS – Ibaraki prefectural health study, WHS – Women's Health Study, NorPD – Norwegian Prescription Database, TS – Tromsø Study, FHS – Framingham Heart Study, MCS – Malmö Cardiovascular Screening, AGES – Age, Gene and Environment–Reykjavik study, RS – Rotterdam Study , CCHS – Copenhagen City Heart Study, HABC – Health, Aging, and Body Composition, BHS – Busselton Health Study, MESA – Multi–Ethnic Study of Atherosclerosis, CIRCS – Circulatory Risk in Communities Study, S–HS – Stockholm Health Screening cohort , OCS – Oslo Cardiovascular Survey, TSS – The Suita Study, L85PS – Leiden 85–Plus Study, SHIP – Study of Health in Pomerania, HCUP – Healthcare Cost and Utilization Project, D–EHR – Denmark Electronic Health Record cohort, S–EHR – Sweden Electronic Health Record cohort, T–NHIRD – Taiwan National Health Insurance Research Database.

2.10 Chapter figures

Figure 2.1 Direction of association reported for 23 risk factors and incidence of atrial fibrillation, ordered from most extreme inverse (green) to most extreme direct (red)



Notes: each dot represents one report in order of most extreme inverse to most extreme direct point estimate. Dots are colour-coded to indicate significant inverse (RR [95% CI]<1.00; dark green), non-significant inverse (RR< 1.00; light green), null or mixed (RR=1.00 or show opposite associations among subpopulations; grey), non-significant direct (RR>1.00; light red) and significant direct (RR [95% CI]>1.00; dark red) associations and scaled in size to indicate <100 (smallest dot), <1000, <10000, <100000 and ≥100000 (largest dot) atrial fibrillation events. For example, for age there were 15 reports and all showed significant direct associations. Risk factor and reference category definitions are detailed in individual risk factors plots (**figures 2.2–2.6** and **S2.2–S2.19**).

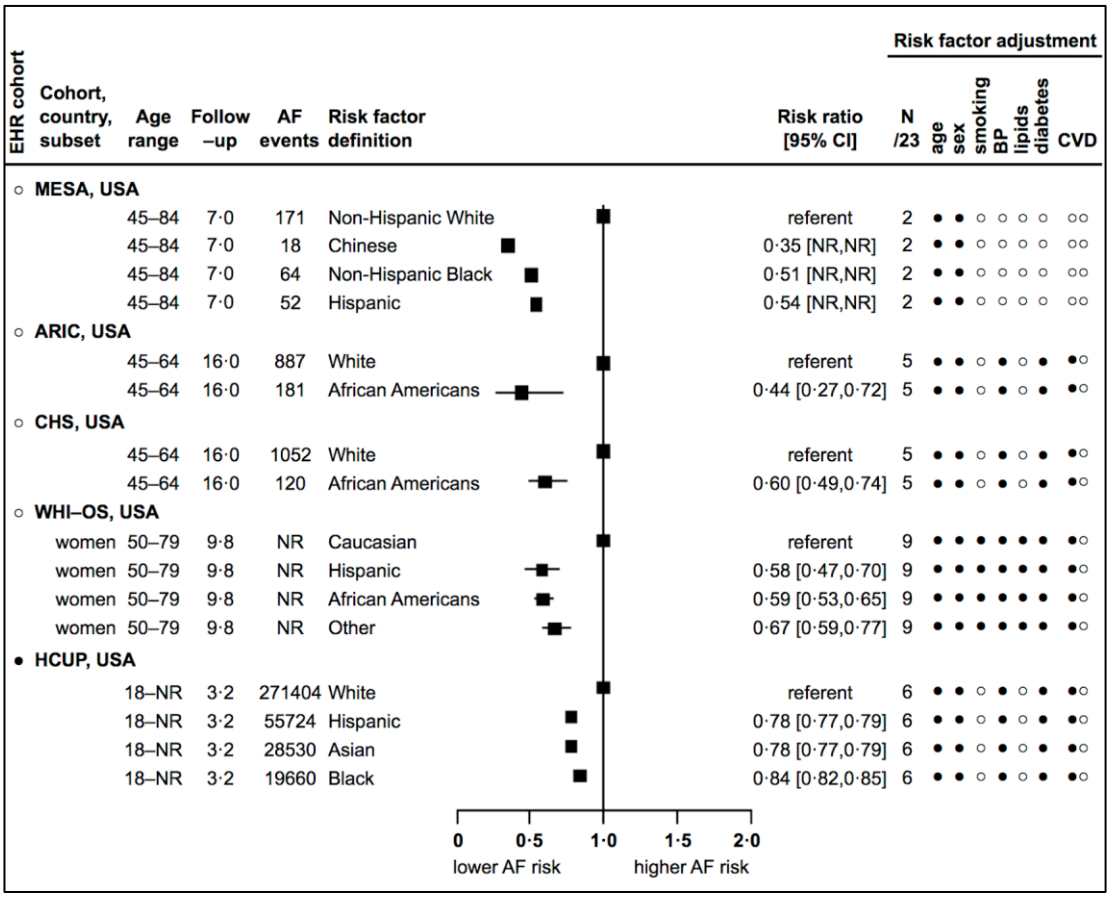
Abbreviations: SES – socioeconomic status, BP – blood pressure, LDL – low density lipoprotein, HDL – high density lipoprotein, RR – relative risk, 95% CI – 95% confidence interval.

References pertaining to each report: on following page.

References pertaining to each report from left to right sequence:

Age - 133,143**,**81**,**100**,**139**,**95,40,40,74,89,135,83,115,78,112
Sex - 124,143,133,83,40,89,139,40,95,100,115
Ethnicity - 130,90,90,74,137
SES - 101,74,93,139,84,81
Smoking - 81,40,128,133,101,115,99,74,83,40,123,86,124
Physical activity - 94,135,79,92,131,73,104,111,75,101,82
Alcohol - 128,98,92,74,81,115,124,63,109,133
Systolic BP - 124, 81,100,128,40,95,40,135,114,129,136,101,110,92
Diastolic BP - 114,95,129, 128,40,136,100,40,110,135,92
Hypertension - 133,100,95,83,124,81,112,74, 136,40,40,77,115,86,101,143,132
Total cholesterol - 106,81,98,87,40,112,128, 78,133,101,40,116,116
LDL cholesterol - 106, 87,100,78,128,116,116
HDL cholesterol -116, 40,116,78,40,87,40,128,106,112
Triglycerides - 106,128,78,40,87,40,40,116,116
Diabetes - 143,128,83,115,99,103,140,95,40,40,112,133,101,86,74,124
Renal disease - 40,97,128,76,143,40, 40,107,139,102,91
Height - 98,101,115,112,40,80,124,40,134,92
Weight - 40,40,40,40,80,124,134,40
Body mass index - 105,101,100,136,93,74,126,135,112,115,121,134,80,128,83,124,77,85,133
C-reative protein - 128,120,118,113,108, 120,100,95
Fibrinogen - 118,113,119,108,88,125
Thyroid disease - 127,96,122,127,141,127, 127,117
Autoimmune disease - 142,138,139

Figure 2.2 Association of ethnicity and incidence of atrial fibrillation: 5 reports from 1 country with 386 115 events

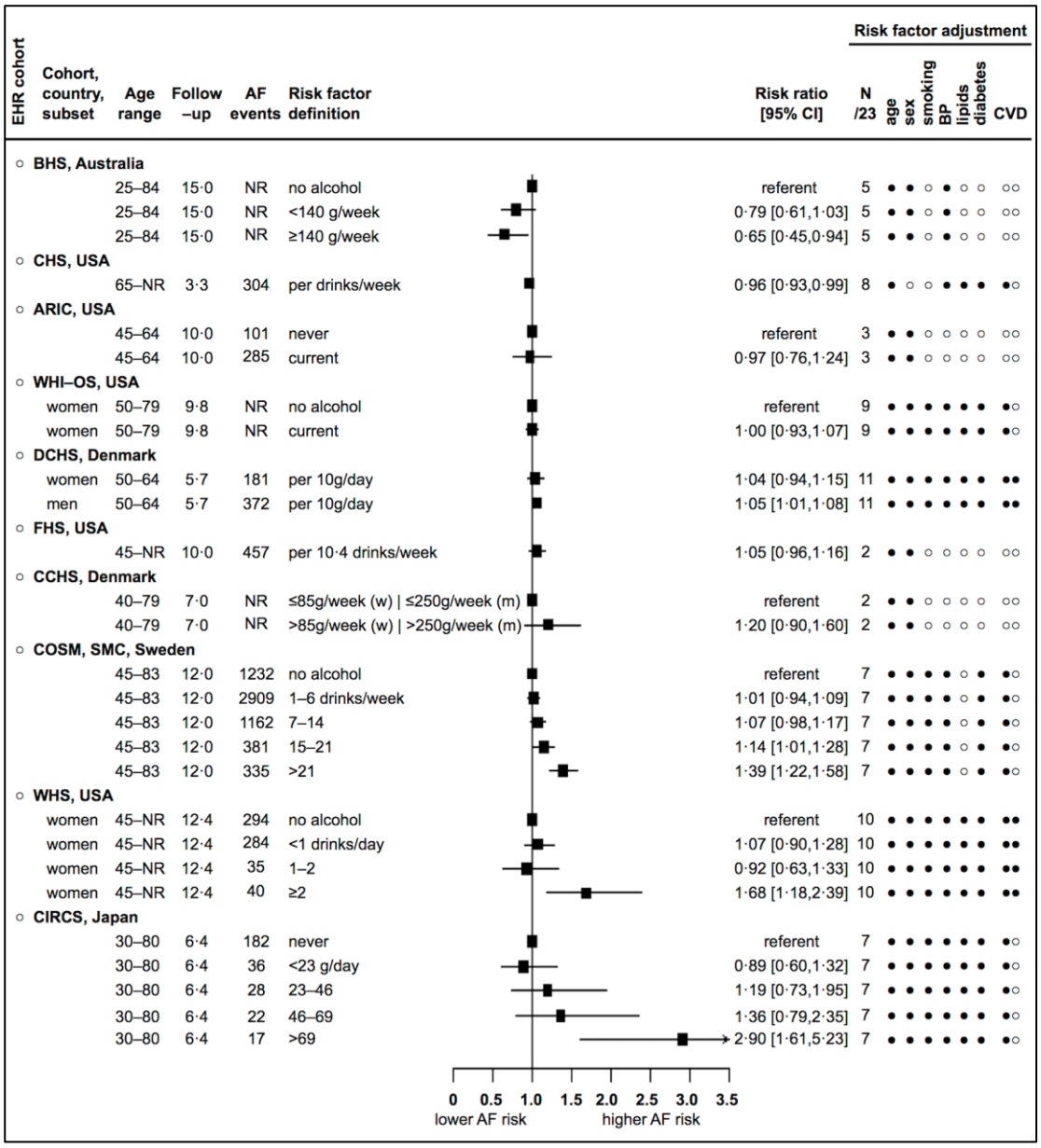


Notes: risk factor adjustment refers to whether adjustment was made for the 23 risk factors under review, 6 CVD risk factors, and prevalent and incident CVD events. Example: ARIC adjusted for 5/23 risk factors, age, sex, blood pressure (i.e. any of systolic blood pressure, diastolic blood pressure, hypertension, or blood pressure lowering medication), and diabetes mellitus, but not smoking or lipids (i.e. any of total cholesterol, low–density lipoprotein cholesterol, high–density lipoprotein cholesterol, triglycerides, hyperlipidaemia, or lipid lowering medication), and prevalent, but not incident CVD events.

Abbreviations: EHR – electronic health record, age range in years, follow–up in years (mean, median, or maximum), AF – atrial fibrillation, CI – confidence interval, N/23 – number (of factors) out of 23, CVD – cardiovascular disease, SD – standard deviation, NR – not reported, USA – United States of America, ● – yes, ○ – no, -- – not applicable. For cohort abbreviations see **table 2.2**.

References pertaining to each report: MESA,[130] ARIC,[90] CHS,[90] WHI-OS,[74] HCUP.[137]

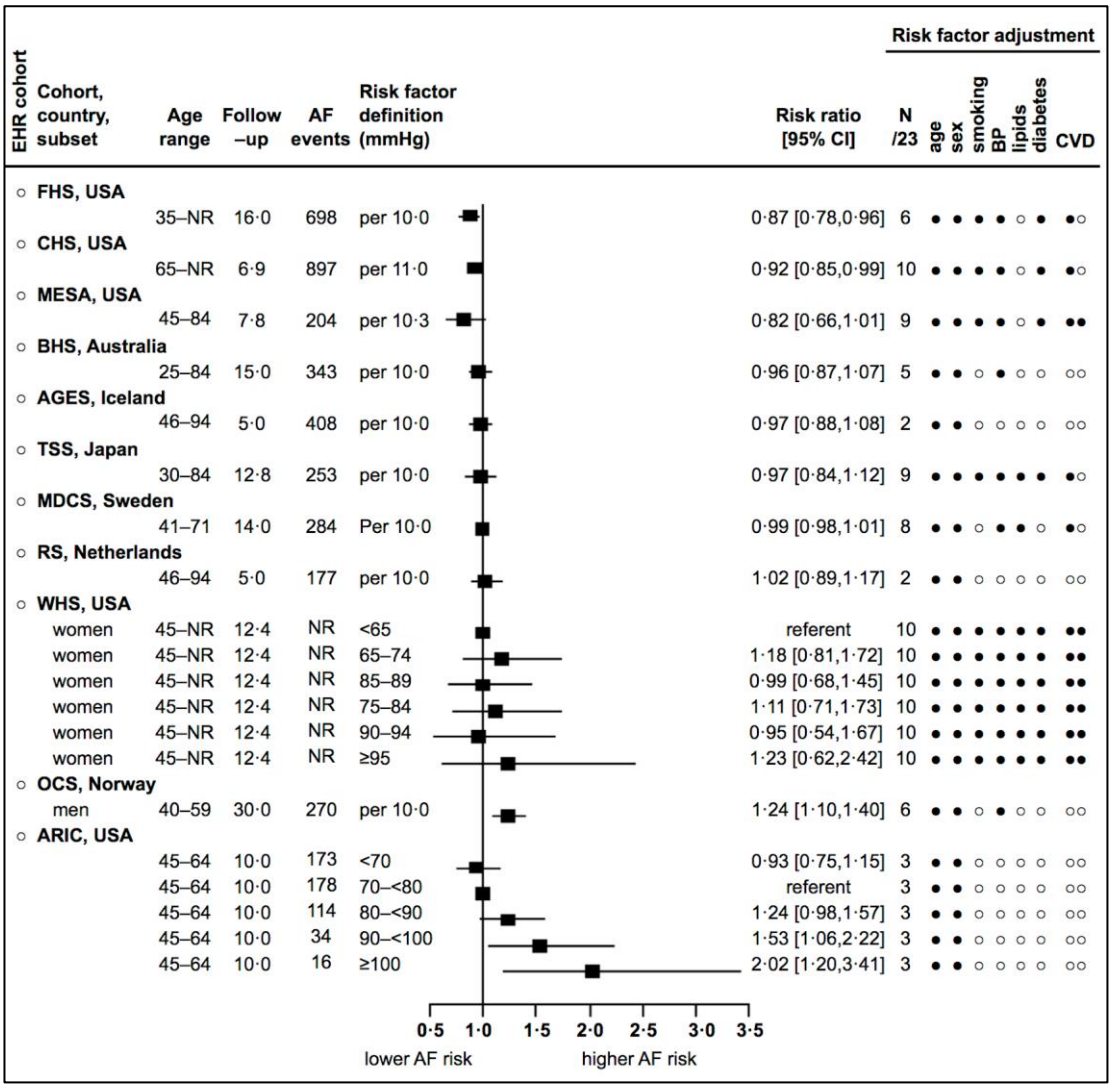
Figure 2.3 Association of alcohol intake and incidence of atrial fibrillation: 10 reports from 5 countries with 18 997 events



Abbreviations: see **figure 2.2** and g – grams, (w) – women, (m) – men.

References pertaining to each report: BHS,[128] CHS,[98] ARIC,[92] WHI-OS,[74] DCHS,[81] FHS,[115] CCHS,[124] COSM,[63] WHS,[109] CIRCS.[133]

Figure 2.4 Association of diastolic blood pressure and incidence of atrial fibrillation: 11 reports from 7 countries with 4796 events

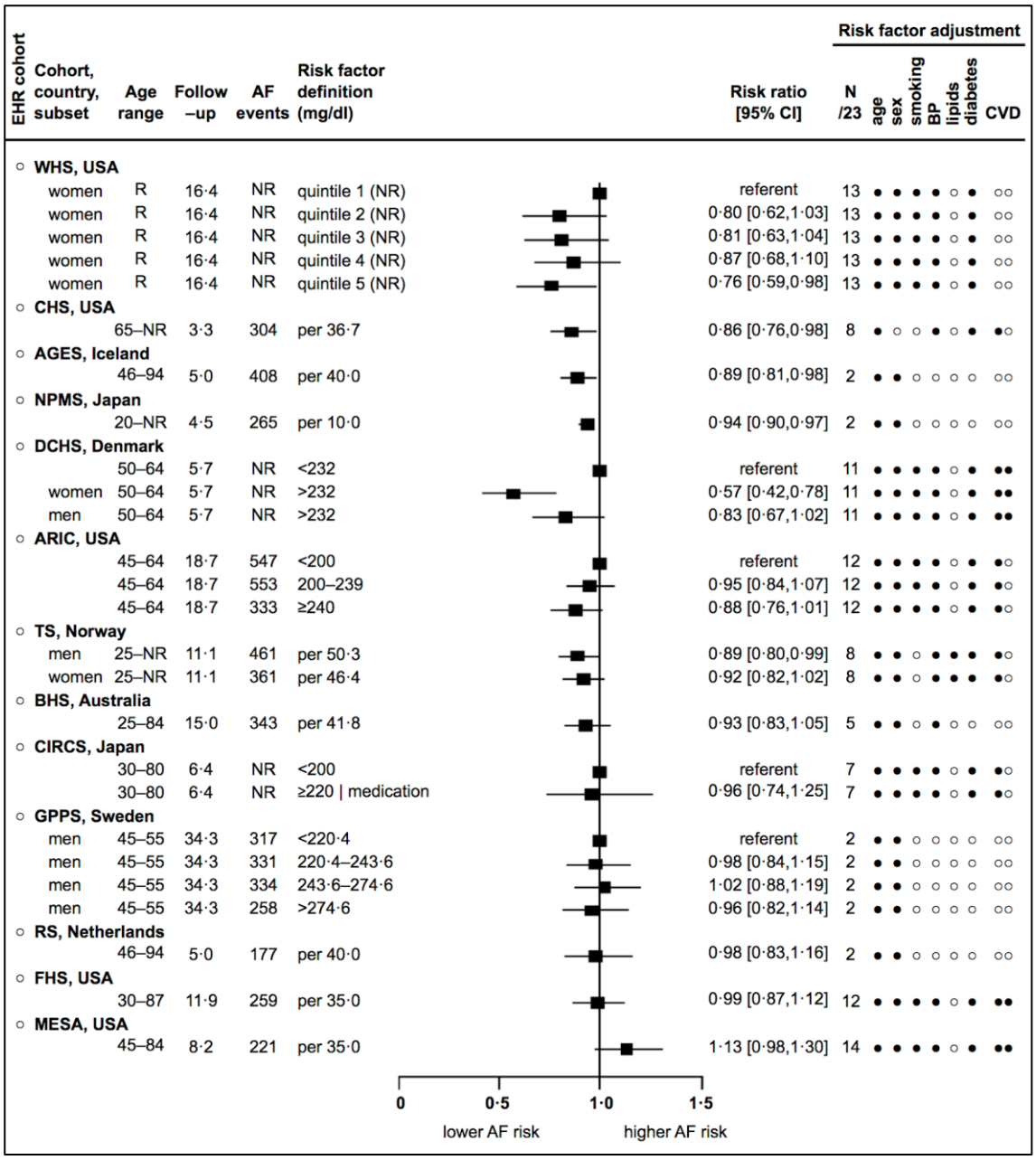


Notes: risk factor adjustment for BP in this instance refers to whether systolic blood pressure, hypertension, or blood pressure lowering medication were adjusted for.

Abbreviations: see **figure 2.2** and mmHg – millimetres of mercury.

References pertaining to each report: FHS,[114] CHS,[95] MESA,[129] BHS,[128] AGES,[40] TSS,[136] MDCS,[100] RS,[40] WHS,[110] OCS,[135] ARIC.[92]

Figure 2.5 Association of total cholesterol and incidence of atrial fibrillation: 13 reports from 8 countries with 7129 events

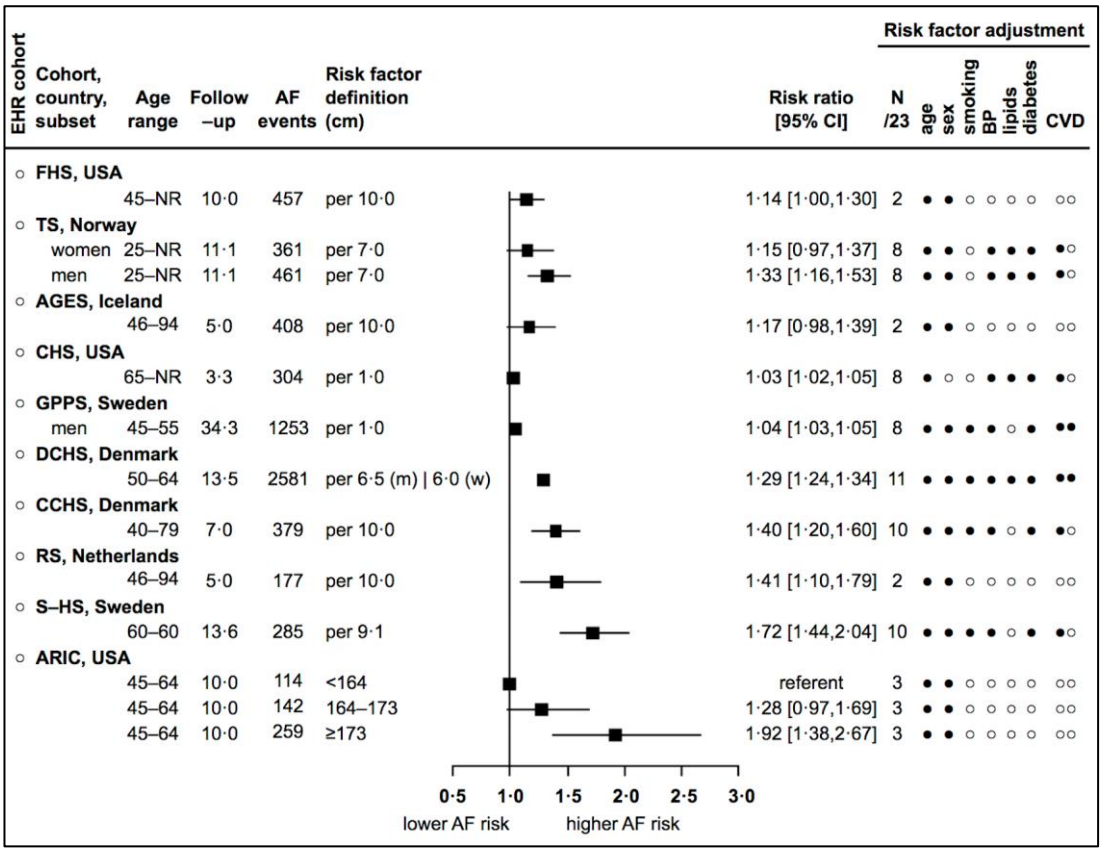


Notes: risk factor adjustment for lipids in this instance refers to whether low–density lipoprotein cholesterol, high–density lipoprotein cholesterol, triglycerides, hyperlipidaemia, or lipid lowering medication were adjusted for. Total cholesterol reported as mmol/l for CHS, GPPS, TS and BHS was converted to mg/dl using the conversion 1mmol/l = 38·66976 mg/dl.

Abbreviations: see **figure 2.2** and mg/dl – milligrams per decilitre, mmol/l – millimoles per litre.

References pertaining to each report: WHS,[106] CHS,[98] AGES,[40] NPMS,[78] DCHS,[81] ARIC,[87] TS,[112] BHS,[128] CIRCS,[133] GPPS,[101] RS,[40] FHS,[116] MESA.[116]

Figure 2.6 Association of height and incidence of atrial fibrillation: 10 reports from 6 countries with 7181 events



Abbreviations: see **figure 2.2** and cm – centimetres, (m) – men, (w) – women.

References pertaining to each report: FHS,[115] TS,[112] AGES,[40] CHS,[98] GPPS,[101] DCHS,[80] CCHS,[124] RS,[40] S-HS,[134] ARIC.[92]

Chapter 3

Clinical research using linked bespoke studies and electronic health records (CALIBER): description of data sources, motivation for use, and analytic considerations in research on atrial fibrillation

3.1 Chapter outline

CALIBER, an acronym for ClinicAl research using LInked Bespoke studies and Electronic health Records, is the name given to the programme of work initiated in 2010 to exploit the scientific opportunity in electronic health records (EHR) in the United Kingdom (UK). EHRs previously unlinked from primary care (general practitioners (GP)), secondary care (inpatient hospital admissions), an acute coronary syndromes registry, and the mortality registry were unified for the first time creating the CALIBER data platform; a new resource for conducting health-related research.[5]

In this chapter I present the motivation for using CALIBER data in research on atrial fibrillation (AF). I first give an overview of how CALIBER data are exploited in the successive chapters of this thesis to study AF risk factors (**chapter 4**), AF subtypes (**chapter 5** and **chapter 6**) and AF outcomes (**chapter 6** and **chapter 7**). I go on to describe the origins of each of the four CALIBER data sources, the process of linking them together to form an enhanced dataset for studying AF, and how CALIBER data, more generally, can be accessed for research. I then discuss analytic challenges in using CALIBER data for research, including ways in which these can be overcome, and I provide a short investigation into the extent to which the data are clinically valid for studying AF. To do so I model the associations of 23 cardiovascular risk factors and incident AF and compare with the findings of my earlier systematic review and field synopsis (**chapter 2**). Finally, I close this chapter by summarising the strengths and limitations of using CALIBER data in epidemiological research as compared to the 32 mostly consented cohort studies identified in my earlier systematic review**.**

3.2 CALIBER: motivation for use in this PhD thesis

EHR, especially when linked, offer the ability to study many different aspects of clinical care including across various time points in disease onset and progression. [5 39 44 51] In this PhD I exploit EHRs to investigate three aspects of AF aetiology: AF risk factors, AF subtypes and AF outcomes. These aspects were chosen, primarily, for two reasons. First, because they relate to current limitations in clinical practice guidelines[3 19 22] influencing the way individuals with, or at risk of, AF are managed (i.e. a lack of recommendations on risk factors to target in the primary prevention of AF,[2] a lack of understanding of clinically distinct subtypes of AF,[3] and a lack of clarity on the level of stroke risk at which individuals with AF require preventative therapy[4]). The second reason I chose to study AF risk factors, subtypes and outcomes refers to the opportunity to study these using data from CALIBER.

As **figure 3.1** shows, the constituent data sources of CALIBER capture complementary information on the onset and progression of AF.[5 51] Primary care records capture comprehensive risk factor data, potentially, predisposing to AF, including a range of demographic, behavioural and biological factors.[53] Linked primary care[53] and secondary care[41] records capture diagnoses of AF made in two different clinical settings, as well as associated comorbidities, medications and surgical interventions relevant for subtyping. Linked primary care,[53] secondary care[41] and mortality records[160] capture comprehensive outcomes information; secondary care records are particularly important for capturing strokes almost always treated in hospitals, and mortality records are important for capturing all deaths recorded in England and Wales.

Of course, it must be reiterated that the analyses of this PhD thesis are only possible because of the structure of the UK healthcare system[52 53] and because AF can be accurately identified in EHRs.[49] EHR data exist in vast quantities in the UK because the National Health Service (NHS) provides healthcare to 98% of the population.[53] Importantly, every healthcare consultation generates an EHR registered against an NHS number unique to each individual. At present, EHRs from different NHS healthcare sectors (e.g. across primary and secondary care) are not centrally connected, however subsequent data linkage (i.e. the bringing together disconnected records belonging to the same individual) is achieved by matching on the basis of NHS numbers. While increasingly used in research, EHRs are not collected for primary research purposes and therefore, among other challenges, identifying disease cases can be complex.[48] This PhD therefore benefits from earlier work by Morley and colleagues showing that, in primary and secondary care linked EHRs, AF cases can be identified based on diagnosis codes, as well as inferred AF cases based on warfarin or digoxin prescriptions in the absence of thromboembolic disease or heart failure.[49] Morley's research unveiled a number of important insights to consider when using EHRs in research on AF and I make reference to these throughout this chapter.

EHRs, undoubtedly, represent a new paradigm for research on AF, however, as I will now describe, understanding the origin of these data is fundamental to the design and conduct of meaningful analyses.

3.3 CALIBER: description of data sources

The CALIBER data platform connects multiple diverse sources of EHR, providing an enhanced view of the AF care pathway, however exploiting EHR data for research requires a firm understanding of where the data have originated.[5] As described below, each of CALIBER's data sources of are managed by separate data providers, consist of complementary data content relevant to AF care, are coded with diverse classifications systems, and cover non-identical populations and time periods with implications for data linkage. Using CALIBER data in research is also subject to data access, governance, and ethical approval.

3.3.1 Data providers

The CALIBER platform comprises linkages between four national sources of EHR, which are managed by separate data providers. These are (1) primary care data from the Clinical Practice

Research Datalink GP OnLine Database (CPRD GOLD), provided by CPRD,[53] (2) secondary care data from Hospital Episode Statistics (HES), provided by NHS Digital,[41] (3) data on admissions to hospital with an acute coronary syndrome from the Myocardial Ischaemia National Audit Project (MINAP), provided by the National Institute for Cardiovascular Outcomes Research (NICOR),[161] and (4) cause-specific mortality data from death certificates, provided by the Office for National Statistics (ONS).[160]

AF records in CPRD GOLD are entered by GPs during primary care consultations or by practice staff based on hospital discharge summaries.[53] AF records in HES are entered by non-clinical coders who are trained to digitise paper-based records maintained by hospital consultants and other hospital staff.[162] AF records in ONS are entered by non-clinical coders who are trained to digitise death registrations certified by doctors.[160] Records in MINAP are entered by trained cardiology nurses. AF is not recorded within MINAP, which focuses exclusively on the management of acute coronary syndromes (ACSs).[161] MINAP records are described here for reference but are not analysed as part of this PhD thesis.

3.3.2 Data content

Records in CPRD GOLD, relevant to AF, include coded data on AF history, diagnosis, and monitoring, AF-related medications (i.e. rate control, rhythm control, and anticoagulants), AF-related procedures (i.e. cardioversion and ablation), AF-related comorbidities (i.e. heart failure, hypertension, diabetes mellitus, stroke and vascular disease history relevant for calculating the guideline recommended CHA_2DS_2-VASc score for stroke risk assessment in individuals with AF), AF-related outcomes (i.e. subsequent ischaemic and hemorrhagic strokes) and numerical clinical values for risk and prognostic factors such as blood pressure and the International Normalised Ratio (INR). Records in HES, relevant to AF, include coded data on AF diagnosis, AF-related procedures, AF-related comorbidities, and AF-outcomes. Records in ONS, relevant to AF, include coded data on AF-related mortality and AF-related outcomes mortality.[7 49] Records in MINAP lack relevant AF data, however otherwise include coded data on clinical variables relating to patient characteristics (demographics and cardiovascular comorbidities), admission characterises (initial and final diagnoses of ACS) and cardiovascular care (interventions, drug therapy) and outcomes.[161]

3.3.3 Data coding

AF records in CALIBER are coded with four classifications systems. These are (1) Read (a subset of Systematised Nomenclature Of Medicine Clinical Terms (SNOMED CT);[163] for primary care diagnoses and procedures in CPRD GOLD with 22 codes for AF diagnosis, (2) British National Formulary (BNF;[164] for primary care prescriptions in CPRD GOLD) with 36 codes for warfarin prescriptions, (3) International Statistical Classification of Diseases and Health-Related Problems, Tenth Edition (ICD-10;[45] for secondary care diagnoses and cause-specific mortality in HES and ONS) with one main code for AF diagnosis, and (4) Office of Population Censuses and Surveys' Classification of Interventions and Procedures, version 4 (OPCS-4;[165] secondary care procedures in HES) with 11 codes for AF-related procedures (including cardioversion and

ablation). MINAP records (not analysed in this PhD because of lack of AF data but described here for reference), are coded using a bespoke classification system that adheres to Cardiology Audit and Registration Data Standards.[166]

3.3.4 Data coverage and linkage

CALIBER data sources cover non-identical time periods and populations with implications for data linkage.[5] CPRD GOLD records have been collected since 1987 and include over 11.3 million individuals from 674 (a subset of) GP practices in England, Wales, Scotland and Northern Ireland.[53] Paper-based records collected before the year 1987 have also been digitised in retrospect, and thus are included in CPRD GOLD, with the earliest records of AF, among the data I have analysed, dating from the 1920s. HES have been collected since 1996 and include inpatient admissions at all NHS hospitals in England and Wales.[41] MINAP records have been collected since 2001 and include ACS admissions at 230 NHS hospitals in England and Wales.[161] ONS mortality statistics have been collected since as early as the 1900s and include all registered deaths in the UK.[160]

CPRD GOLD, HES, MINAP and ONS data were linked in 2010 to form CALIBER, bringing together for the first time AF records from across multiple sectors of healthcare provision in the UK. Because not all GP practices in CPRD GOLD consent to data being linked with other EHR sources (i.e. to HES, ONS and MINAP), the denominator population of CALIBER is therefore based on patients registered at GP practices permitting record linkage.[5] The extract of CALIBER data analysed for this PhD thesis focusses on the years 1997 to 2010, reflecting the time period when all datasets are in alignment, and consists of approximately 2 million individuals initially free from a diagnosis of cardiovascular disease (CVD). As reported by Morley and colleagues, both CPRD GOLD and HES capture unique patient diagnoses of AF, with over one third (39.6%) of diagnoses captured in both data sources.[49] Linking data sources is therefore crucially important for complete ascertainment of AF cases as well as comorbidities for risk stratification, which I go on demonstrated in **chapter 7**.

Data linkage was carried out by a trusted third party, coordinated by the Medicines and Healthcare products Regulatory Agency, and using deterministic matching of patient identifiers (these were unique 10-digit NHS number, date of birth, sex and postcode). Over 95% of individuals with a valid NHS number were matched successfully.[5] Unfortunately no further information are available to CALIBER researchers on the quality of the data linkage or exact criteria for matching two pairs of records. The linkage was performed externally by trusted third party (rather internally than by myself) in order to protect patient confidentially and in respect of data governance laws.

3.3.5 Data access, governance and ethics

Access to CALIBER data for research depends upon approval of the intended research purpose, required standards in data governance and, potentially, ethical approval.[5] These require-

ments exist to safeguard against improper data use and to protect the confidentiality of the general public who contribute health data, often without any direct reciprocal gain. Principles in data access, governance and ethics, as well as expectations on behalf of both the research community and the general public, have evolved in recent years with impact on timelines and scope of this PhD thesis.[167-170]

Data access

Approval to access CALIBER data is obtained by submitting a study protocol for consideration to the Independent Scientific Advisory Committee (ISAC).[5] ISAC was set up by the Secretary of State for Health to provide expert advice on health-related research projects. Protocols are assessed for scientific rigour in terms of study quality, feasibility, and value to public health. Protocols must outline the background to the proposed study in terms of clinical importance, current unknowns and what the study will potentially add, methods in terms of use of linked data, definitions for the study population and clinical variables and analytical approach, as well as future plans for research dissemination.

The projects of this PhD thesis have approval under ISAC protocol no. 12_165 (AF risk factors and subtypes) and 15_028 (AF outcomes). Approval for protocol no. 12_165 was obtained by a former UCL colleague (Dr Katherine Morley) in January 2013, and I was subsequently added as a named investigator. Approval for protocol 15_028, which I sought, was obtained in March 2015. The approval process, in my case, took approximately two months and involved (1) an initial dialogue with ISAC to discuss research intentions, (2) submission of the full ISAC protocol, (3) revision and resubmission of the protocol based on ISAC suggestions, (4) protocol approval granting access to CALIBER dataset. Approvals for research projects involving new or bespoke data linkages (e.g. to the cancer registry), the recontact of patients or GPs (e.g. for clinical trial recruitment or validation studies), or less well validated methodologies (e.g. machine learning) can however take considerably longer to obtain. Further details of the ISAC process are available on the CPRD webpages: https://www.cprd.com/isac/.

Data governance

Once ISAC approval has been granted, CALIBER data is made available to researchers providing that required standards in data governance (also known as information governance) are met. Data governance refers to the safe and proper use and storage of sensitive information. In accordance with UK and European law, CALIBER data are pseudonymised, meaning all personal information (i.e. names, NHS numbers, addresses and full date of birth information) are removed and replaced with an artificial unique person identifier. CALIBER data must be held and analysed in restricted access drives and increasingly via remote servers known as 'data safe-havens'.[167] The move to data safe-havens offers greater data security however current technology can restrict research progress with working days lost due downed servers or in waiting for research outputs to be manually reviewed before exporting from the server. Access to CALIBER data is also restricted to what has been approved by ISAC to answer the research question and is time limited. Any analyses which fall outside of the original study remit must be

approved by ISAC in the form of a protocol revision. Data retention policies dictate the duration of agreed access to CALIBER data and all data must be destroyed after research objectives have been achieved.

The UCL School of Life and Medical Sciences (SLMS; under which the Institute of Health Informatics currently sits) maintain an Information Governance Framework for the promotion of safe working practices and researchers are provided with annual awareness training.[171]

Data ethics

Ethical approval was not required for the specific projects of this PhD as they involve only pseudonymised data. However, the CALIBER dataset overall has been granted both Ethics approval (ref: 09/H0810/16) and Ethics and Confidentiality Committee of the National Information Governance Board for Health and Social Care approval (ref: ECC 2-06(b)/2009 CALIBER dataset).

Evolution of data access, governance and ethics

Alongside the expansion in the use of EHRs in research over the past decade,[47] principles in data access, governance and ethics have been evolving in recent years.

In 2012, the UK government signalled support for EHR research in the Health and Social Care Information Act,[172] promising to unlock and make available new health-related datasets including, under the Care.data programme, a centralised database connecting records from primary and secondary care. However, the implementation of Care.data was poorly handled. Lack of public consultation led to mistrust about the reuse of EHR data, including whether health records could be exploited (e.g. by insurance companies) for commercial gain. Care.data was therefore suspended and subjected to a government inquiry.[169] The mishandling of Care.data had a direct impact on EHR research with requests for new and updated data extracts placed on indefinite hold. As a consequence, the extract of CALIBER data I have analysed is censored at the year 2010; almost a decade prior the submission of the PhD thesis.

The failures of Care.data have however helped to positively redefine expectations of both the research community and the general public with respect to EHR research.[169 170] One, if not the biggest, reason Care.data failed was the assumption that people would be willing to share their health records and have a moral obligation to do so, and furthermore the public health benefits derived far outweighed any potential risks (e.g. in re-identification of individuals). In a positive step, public engagement (i.e. the involvement of the general population in research activities), is becoming an increasingly integral part of research project design and conduct.[170] In a personal effort towards public outreach, I wrote a non-specialist summary of my PhD thesis (available in **appendix**) that was later shortlisted for the 2016 Max Perutz Science Writing Award.[173]

Understanding the origins, modes of access and public perceptions of EHR data are examples of the complexities associated with EHR research. In the next section I describe analytic complexities to consider when using EHR data, including ways in which these can be overcome,

and I investigate whether EHR data are clinically valid by way of replicating the findings of my systematic review and field synopsis into cardiovascular risk factors associations with incident AF.

3.4 CALIBER: analytical considerations

EHRs offer a valuable data source for the pursuit of clinical research; however there are a number of associated analytical complexities which arise from the fact that EHRs are not collected for primary research purposes.[48] In this section I describe analytic consideration of EHR data which relate mainly to data definitions, data quality and data validity.

3.4.1 Data definitions

A key analytical consideration for research involving EHRs lies in the management of millions of rows of unsorted data. Sorting and selecting relevant data for each research question requires rules and definitions for the study population, clinical variables and outcomes of interest, including how to combine data from diverse sources coded with differing levels of clinical detail and whether to use single or repeated data point over time. These rules then need to be translated into coding language and executed on powerful computing systems.[51]

Preparing a single clinical variable for analysis is time and resource intensive. Diseases, in particular, are rarely expressed by a single clinical code (e.g. as Morley and colleagues showed there are 22 primary care (Read) codes and one main secondary care (ICD-10) code for AF diagnosis[49]) and creative strategies (known an EHR algorithms) may be needed to identify disease cases missing a diagnosis code but with other highly relevant information suggesting disease presence (e.g. Morley's 'inferred AF' based on warfarin prescriptions in the absence of thromboembolic disease[49]). Robust definitions for identifying risk factors and outcomes are necessary in order to accurately estimate the associations between them. However, the process of developing EHR definitions and algorithms requires a thorough understanding of the data and its structures.

To advance progress and raise standards in EHR research, CALIBER researchers therefore created a framework for extracting the value out of linked EHR data, as well as an online repository (www.caliberresearch.org/portal) for sharing knowledge, algorithms and code lists for identifying (currently around >600) risk factors and outcomes in EHRs.[51] Future EHRs studies can therefore align to these definitions, and thus be more easily compared. This is of huge relevance to the field of AF risk factor research, given that, as I showed in my systematic review,[2] widely different definitions for risk factors and AF events have been used across the available studies to date. I fully describe and apply the CALIBER framework in **chapter 5** and **chapter 6** when I develop my own EHR definitions for identifying AF subtypes.

3.4.2 Data quality

The quality of CALIBER data depends upon how well, and whether, information has been captured and recorded at source.[48] Data quality is beyond researchers' direct control and therefore

detailed checks must be undertaken (e.g. correcting for duplicates, date inconsistencies, implausible values, assessing and addressing extent of missing data) before commencing data analysis.[53]

Temporal improvements in data quality have been driven by standards imposed by data providers, as well as national initiatives. CPRD GOLD assesses practice-level data quality based on proportion of absences or inconsistencies in the recording of age, gender, practice registration or deregistration information, pregnancy outcomes, and rates of prescriptions, referrals, and deaths, and provides the date at which practices are deemed of research standard.[53 174] For example, 75% of practices, among the data I have analysed, were considered up-to-standard by 8 September 1999. CPRD GOLD data is also further enhanced by the Quality and Outcomes Framework (QOF) introduced in 2004 by the UK Government to remunerate GPs for completeness of recording a number of specified care quality indicators (e.g. the proportion of patients over 40 years of age with a blood pressure measurement in the past five years, or the proportion of patients with AF on an anticoagulant).[175] AF was introduced as a QOF indicator in 2006, and as observed by Morley and colleagues, the recording of AF improved (i.e. fewer inferred cases) between 1998 and 2010. Increased recording of AF over a similar period (1995 to 2010) has also been reported for mortality records.[176] Mortality records are not subject to QOF incentives and therefore changes in AF recording practices may instead reflect increasing incidence of AF associated with longer living populations as well as greater clinical awareness around detection and diagnosis of cases. HES data quality is monitored by NHS Digital who publish periodic online reports including the automatic data cleaning rules applied to datasets,[177] however no quality indicators for individual hospitals are provided to researchers.

Missing data is an important aspect of data quality affecting EHR research. Data in epidemiological studies can be missing for many reasons, however the two most specific to EHRs are that (1) information capture is intermittent (i.e. individuals only have data recorded if they interact with healthcare services) and (2) the type of information captured depends on the nature of the healthcare interaction (e.g. an individual who has AF but isn't prescribed warfarin is unlikely to have any data on INR levels).[48]

Popular methods for handling missing data in EHR studies include using values recorded within a defined 'look back' period (e.g. one, two or five years) before study start date, defining study baseline dates based on when values have been recorded, and artificially creating plausible value using single or multiple imputation.[178] Single imputation involves computing a single plausible value for the missing value, whereas multiple imputation involves computing multiple plausible values for the missing value reflecting uncertainty about what the missing value is. Multiple imputation, while a more sophisticated method for handling missing data, is difficult to apply in EHRs because of complexities relating to data size and structure. New methods specific to the imputation of EHR data (e.g. the two-fold fully conditional specification algorithm[179]) are currently under development.

3.4.3 Data validity

A fundamental question about EHRs is whether the data are valid for clinical research. EHRs, as Hripcsak and Albers argue,[48] do not provide a direct window into clinical events, but rather the recording processes surrounding clinical events and are therefore subject to systematic biases. Testing the validity of EHR data is therefore an important analytical step.

The clinical validity of EHRs can be tested using three main methods, which are (1) re-contacting patients or GPs to first hand verify information (e.g. confirming whether an individual with an AF record really has AF)[180], (2) verifying measures against a trusted external data source (e.g. comparing EHR estimates of AF prevalence with estimates from a gold standard electrocardiography (ECG) study)[55], and (3) performing internal verification tests against existing clinical knowledge (e.g. testing whether EHRs show an association between AF and subsequent stroke)[7].

Adopting the first method, Ruigómez and colleagues re-contacted GPs to verify AF diagnosis records with 1540 out of 1606 individuals confirmed as AF cases (a positive predictive value of 96%).[180] Adopting the second method, Morley and colleagues found an AF prevalence estimate in CALIBER of 1.6% which is comparable to the 2.0% (95% CI: 1.6–2.4%) estimate from the UK-based ECHOES (Echocardiographic Heart of England Screening) study with access to ECG data. Adopting the third method of testing clinical validity, I investigated whether CALIBER records are consistent with the existing observational literature on the associations of 23 cardiovascular risk factors and incident AF, as found in my systematic review.[2]

The methods I used to create an analytic dataset for studying AF risk factors are described in full in the **appendix** (and further reflections are provided in **chapter 4** when I reuse this dataset to analyse novel risk factor associations with AF both with and without an intercurrent diagnosis of CVD). Briefly, I included individuals available for analysis between 1998 and 2010, of at least 30 years of age, without prior record of AF or ten other CVDs, and with a minimum of one year follow-up at a primary care practice with research quality data. I used existing EHR definitions for the 23 risk factors and AF cases as available on the CALIBER portal and in published papers (e.g. sex,[181] ethnicity,[182] socio-economic status,[183] smoking status,[184] alcohol,[185] blood pressure,[148] lipids,[186] diabetes mellitus,[187] height, weight and body mass index,[188] autoimmune diseases[189] and AF[49]).

To maximise the use of observed risk factor data, I selected values available at baseline (i.e. earliest date study inclusion criteria fulfilled) using a look back period of one year (unless otherwise specified in appendices), together with values available after baseline but before date of occurrence of any CVD or AF events. This available case strategy,[190] therefore means the number of individuals included in analyses varies from risk factor to risk factor and findings should be interpreted on a risk factor by risk factor basis, rather than for the cohort as a whole.
I used Cox regression models adjusted for age and sex and stratified on GP practice to model associations between risk factors and AF with the underlying timescale as the difference be-

tween date of censoring and date of risk factor measurement (i.e. baseline if measure available at baseline, or date of measurement if available after baseline). Individuals were censored in the event of an AF record, death from a cause other than AF, and for administrative reasons such as deregistration from GP practice and end of GP practice follow-up.

The resulting CALIBER estimates for the associations of 23 cardiovascular risk factors and incident AF, in terms of relative risks and 95% confidence intervals (RR [95% CI]), are detailed in **table 3.1.** I updated **figure 2.1**, which summarises the overall results of the systematic review and field synopsis in terms inverse (RR<1.00), null (RR=1.00) and direct (RR>1.00) risk factor associations, to include the CALIBER estimates (thus creating **figure 3.2**). As shown (in **figure 3.2**), CALIBER estimates were entirely consistent with the existing observational literature, which, as this analysis aimed to test, confirms the validity of using EHR data in research on AF.

3.5 CALIBER: strengths and limitations

In this final section of **chapter 3**, describing the CALIBER dataset and motivation for use in research on AF, I now summarise the strengths and limitations of EHR data. I draw comparisons with the 32 epidemiological cohorts used in studies of AF risk factors to date, as identified in my systematic review. I also consider the strengths and limitations of CALIBER in relation to other types of data collected at scale, within or outside of healthcare, and relevant for research on AF (**table 3.2**).

3.5.1 Data strengths

CALIBER contains diverse health datasets which when linked together offer a powerful resource for epidemiological research. Key strengths of using CALIBER in research on AF include:

- **Broad range of clinical variables**

  CALIBER contains a broad range of clinical variables, allowing studies of, but not limited to, AF risk factors, subtypes and outcomes. None (0) of the 32 cohorts identified in the review reported associations for all 23 cardiovascular risk factors,[2] whereas, as I have shown, this is possible using CALIBER data. Traditional epidemiological cohorts (as were the majority in my systematic review) are restricted to studies on a limited set of clinical measures, predefined and collected to meet research objectives. EHRs, on the other hand, are collected without any predefined research questions in mind allowing a broader range of clinical investigations. Four other EHR cohorts were identified in the review (including the nation-wide Danish[42] and Swedish[43] cohorts) however only coded risk factor data (e.g. a record of diagnosis of hypertension or diabetes mellitus) were reported. A major strength of CALIBER, as an EHR resource, is the availability of numerical risk factor data (e.g. blood pressure, lipids and body mass index), which can be used, for example, to study optimal risk factor levels to target in the primary prevention of AF.

- **Large scale population denominator**

  CALIBER data comprises a large scale population denominator. For example, in the above validity study, I analysed data on 1,949,052 individuals with 50,097 incident AF cases over 12 years of follow-up. By contrast, the Framingham heart study, an important cohort for uncovering the link between AF and subsequent stroke risk,[13] accrued only 1500 incident AF cases in over 50 years of follow-up.[191] Large-scale population denominators are necessary for reliably estimating disease incidence and prevalence (e.g. as shown by Morley[49] for AF) or for intricately investigating AF subtypes (e.g. as I go on to show in **chapter 6**).

- **Life course data**

  CALIBER records consist of longitudinal, life course data. In the UK, individuals begin generating EHRs linked their unique NHS number when they first come into contact with healthcare services, which is usually at birth. CALIBER's linkages between healthcare and mortality records can therefore be used for epidemiological studies over the entire life course, or so-called 'cradle-to-grave' analyses. The longitudinal nature of data collection, including repeated clinical measurements, offers the chance to more accurately study how risk factors develop and evolve over time, unlike traditional cohorts which capture only health snapshots at study baseline and periodic follow-up intervals.

- **Objectivity in data collection**

  CALIBER records are collected with a greater level of objectively than perhaps is found in traditional epidemiological cohort studies. Data are collected by health professionals and at the time of the clinical event, thus are not subject to biases in information recall or self-reporting of events. Moreover because EHRs are collected without a research hypothesis in mind, the process of recording information is not impacted by confirmatory bias[158] (i.e. the influence of pre-existing beliefs).

- **Contemporary real world clinical care**

  CALIBER data reflect contemporary real world clinical care, which is particularly important for understanding how newly implemented medical interventions perform in the longer-term and beyond the controlled setting of a clinical trial.[192] Studies of drug efficacy, safety, persistence, and adherence require longitudinal and accurate medical information. Traditional cohorts containing only cross-sectional snapshots of self-reported medications are therefore inadequate for these types of studies. EHR cohorts, such as CALIBER, are therefore in a unique position to study the impact of the four newly introduced direct oral anticoagulants (DOACs) for stroke prevention in AF.[193]

  Aside from an inability to study clinical effectiveness, traditional observational cohorts may also be limited in the extent to which they relate to contemporary health issues. Obesity and diabetes mellitus, for example, may not be so well reflected in the Framingham Heart study cohort with historic risk factor data collected at study inception in 1948.[191] The nature of collection of EHRs ensures modern day healthcare issues are continuously reflected.

- **International comparisons**

  CALIBER data is coded using standardised classification systems also used in other countries which facilitates international comparisons research.[194 195] International comparisons help to understand how similarities and differences in the health and health systems of diverse populations impact on outcomes. Like CALIBER, both the nation-wide Danish[42] and Swedish[43] cohorts contain diagnoses coded with ICD-10. Therefore, between-country differences in AF care and outcomes could be easily compared. The CHARGE–AF (Cohorts for Heart and Aging Research in Genomic Epidemiology - AF) consortium of 5 cohorts across 3 countries reflects efforts to make such comparisons in the traditional epidemiological cohort setting.[40] However, the process of harmonising non-standard data definitions can be challenging.

- **Data linkage**

  CALIBER's data linkages piece together patient journeys from across primary and secondary care, providing greater understanding of AF care and outcomes than if using a single data source in isolation.[49] In the UK, EHRs can be linked easily because of the existence of a unique patient identifier in the NHS number. The nation-wide Danish[42] and Swedish[43] cohorts also benefit from a unique personal identification number upon which health and health-related records can be linked. Traditional cohort studies are increasingly exploring linkages to EHR resources for the purpose of validating self-reported research measures and as an enhanced method of participant follow-up.[196 197] However, this can process can be challenging without the availability of a unique identifier (e.g. an NHS number) to link the datasets and can be restricted by the study's consent model (i.e. what participants have originally agreed to).

3.5.2 Data limitations

CALIBER data, while offering unique opportunities for clinical research, have a number of important limitations. Key limitations of using CALIBER in research on AF include:

- **Lack of imaging, free text and genomics data**

  CALIBER's broad range of clinical variables and validated algorithm for identifying individuals with AF unfortunately does not include imaging, free text or genomics data at present (**table 3.2**). Linkage to ECG images (i.e. the gold standard method for diagnosing AF) would help to improve AF case ascertainment and further strengthen the validity of cases identified through coded AF diagnoses. Linkage to free text descriptions, in the form of clinical notes, discharge summaries and letters, would provide greater context around AF events including any symptoms, vital signs, tests performed and medications prescribed. Linkage to genomics information would help to improve the development of risk prediction models by factoring in individuals' underlying genetic susceptibility of first developing AF and then subsequently responding to AF-related treatments. However, imaging, free text and genomics data are not extracted from NHS systems in a centralised manner which means they cannot currently be linked and made available for population level research.[51] Almost all of the non-

EHR cohort studies identified in my systematic review collected research ECGs at baseline and study follow-up,[2] and therefore have the advantage over CALIBER in this regard.

- **Data not in a research ready format**

  As described in the above analytic considerations section, CALIBER data do not exist in a format that is immediately amenable to analysis. Extracting the value out of EHRs therefore requires the development of robust disease definitions as well as powerful computing systems to sort and select through the millions of rows of data.[51] Data quality issues in EHR research, which depend on whether, and how well, information has been captured and recorded at source, can be overcome with methods such as multiple imputation, however these methods are complicated by data size and structure.[48 178] By contrast, the imputation of missing data in traditional cohort studies is easier because of smaller sample sizes, regular time points of data capture (i.e. at baseline and follow-up) and because research measures are predefined and, in principle, collected for all participants. Conversely in EHRs, data are only collected when individuals interact with healthcare and there are no records to indicate lack of disease. An individual without an AF record is therefore assumed to be free from AF. Moreover, even if relevant data exist for a particular disease endpoint, risk factor or characteristic, they may not always be useful for research. Morley, for example, examined whether the ascertainment of AF cases could be improved with information on pulse palpation (i.e. a method of screening for AF), however these data added limited value.[49] Challenges in the preparation of EHR data for research are helped by the knowledge shared on the CALIBER data portal: www.caliberresearch.org/portal

- **Individuals not captured**

  CALIBER data comprises a large-scale population denominator, however does not capture all individuals in the UK. Individuals not captured in CALIBER include those from primary care practices not included, or not permitting data linkage, in CPRD, as well as those who individually withdraw consent from the reuse of their data for research purposes.[169] Whereas withdrawals from traditional cohort studies can be easily managed and quantified, EHR data providers do not share information of the numbers of withdrawn individuals. CALIBER data have however been shown to be representative of the UK population in terms of demographics (age, sex and ethnicity[198]) and overall mortality.[53]

- **Lack of real-time or recent data**

  EHR data, while continuously captured, are not currently available for research in real-time. Datasets are extracted from health systems with some time lags and are made available to researchers at some time thereafter. The ability to conduct research and, potentially, influence clinical care in real time is an attractive prospect however is unlikely to be available in the immediate future. Delays in updating the contemporaneity of the CALIBER dataset due to the information governance challenges described earlier[168 169] meant that the most recent year of data I analysed was the year 2010. As such, I was unable to provide any novel in-

sights on the introduction of DOACs for stroke prevention in AF, which is a limitation of this thesis.

In describing CALIBER's strengths and limitations and comparing and contrasting these with the current observational literature on AF I have given a sense of the opportunities for research presented in EHRs. Of course, observational cohort studies in the traditional sense (e.g. Framingham) and in a more modern, 'big data' sense (e.g. CALIBER) are not the only study designs available. AF registries (e.g. the Euro Heart Survey[199] or the recent GARFIELD-AF registry[200]) and well-designed single centre studies (e.g. The Loire Valley Atrial Fibrillation Project[201]) contain far smaller sample sizes limiting population level understanding but instead benefit from higher resolution and complete information specific to AF care. These, and other, study designs should therefore be viewed as complementary. Over the course of the analytic chapters of this PhD thesis, I therefore endeavor to provide greater understanding on the role EHR data can play in advancing progress in AF research.

3.6 Chapter summary

In summary, this chapter served to describe how the data sources comprising the CALIBER dataset link together to offer a number of unique advantages for research on AF. In the successive chapters of the PhD thesis, I exploit CALIBER data to study three aspects of AF aetiology: risk factors, subtypes and outcomes. I begin in **chapter 4** with a novel investigation into AF risk factors, which looks at whether the 23 cardiovascular factors considered in my systematic review and earlier validation exercise show consistent associations with AF with (AF^+) and AF without (AF^-) an intercurrent diagnosis of CVD. This was an important question in my systematic review that could not be answered based on the current field of AF risk factor research.

3.7 Chapter tables

Table 3.1 Hazard ratios and 95% confidence intervals for the associations of 23 cardio-vascular risk factors with incidence of atrial fibrillation as estimated using CALIBER data

Demographics	N	AF	HR [95%CI]
Age per 15.3 years	1949052	50097	3.99 [3.95, 4.03]
Gender:			
Women	999020	25472	reference
Men	950032	24625	1.49 [1.47, 1.51]
Ethnicity:			
White	933202	39097	reference
Black	32370	193	0.46 [0.40, 0.54]
South Asian	30309	220	0.51 [0.45, 0.59]
Other and Mixed	35972	320	0.68 [0.61, 0.77]
Socio-economic status:			
First quintile (least deprived)	388102	9214	0.84 [0.81,0.88]
Second quintile	389244	10479	0.91 [0.88,0.94]
Third quintile	387307	10580	0.93 [0.89,0.95]
Fourth quintile	388527	10095	0.96 [0.93,0.99]
Fifth quintile (most deprived)	386921	9564	reference
Health behaviours			
Smoking status:			
Non smoker	978954	26104	reference
Former smoker	298538	9208	1.21 [1.18, 1.24]
Current smoker	365859	5131	1.27 [1.23, 1.31]
Alcohol drinker status:			
Non-drinker	201184	5526	reference
Former drinker	49320	1468	1.11 [1.04, 1.17]
Occasional drinker	198352	5432	0.96 [0.93, 1.00]
Moderate drinker	949745	21272	1.00 [0.97, 1.04]
Heavy drinker	6426	180	1.75 [1.50, 2.03]
Physical activity status:			
Inactive	84686	2639	reference
Gentle activity	224567	7349	0.90 [0.86, 0.94]
Moderate activity	409656	8401	0.78 [0.75, 0.82]
Vigorous activity	76248	818	0.71 [0.67, 0.77]
Blood pressure			
Systolic blood pressure per 20.2 mmHG	1574301	44098	1.13 [1.12, 1.14]
Diastolic blood pressure per 10.9 mmHG	1574301	44098	1.12 [1.11, 1.13]
Hypertension	1949052	50097	1.69 [1.66, 1.72]
Lipids			
Total cholesterol per 1.1 mmol/L	731521	22234	0.88 [0.87, 0.89]
LDL cholesterol per 1.0 mmol/L	528681	13254	0.88 [0.86, 0.89]
HDL cholesterol per 0.4 mmol/L	607883	16314	0.99 [0.97, 1.01]
Triglycerides per 1.2 mmol/L	618403	17308	0.95 [0.93, 0.97]
Metabolic factors			
Diabetes mellitus:			
Type I	8834	282	1.91 [1.70, 2.15]
Type II	93514	7359	1.51 [1.47, 1.54]
Uncertain	8681	563	1.75 [1.61, 1.90]
Renal failure	1949052	50097	1.55 [1.44, 1.66]
Anthropometry			
Height per 0.1 m	1604843	43434	1.32 [1.30, 1.34]
Weight per 17.1 Kg	1385286	33269	1.38 [1.37, 1.40]
Body mass index per 5.2 Kg/m^2	1347412	32257	1.24 [1.23, 1.25]
Inflammation			
C-reactive protein per 13.0 mg/L	284450	6851	1.12 [1.10, 1.14]
Fibrinogen per 1.2 g/L	16594	619	1.08 [0.99, 1.17]

Thyroid disease			
Thyroid disease:			
Hypothyroidism	35632	1660	1.15 [1.09, 1.21]
Hyperthyroidism	8361	394	1.30 [1.18, 1.44]
Uncertain type	4768	211	1.15 [1.00, 1.32]
Autoimmune disease			
Rheumatoid arthritis	1949052	50097	1.53 [1.42, 1.63]
Psoriasis	1949052	50097	1.11 [1.04, 1.17]

Notes: continuous variables are analysed for strength of association per one standard deviation increase in value.

Abbreviations: N – number, AF – atrial fibrillation, HR [95%CI] – hazard ratio and 95% confidence interval, L/HDL – low/high density lipoprotein cholesterol, mmHG - millimetres of mercury, mmol/L - millimoles per Liter, mg/L - milligrams per litre, g/L - grams per litre, m – metres, kg – kilograms.

Table 3.2 Examples of data collected at scale, within or outside of healthcare, and relevant for research on atrial fibrillation

Data type	Data description	Data example
Coded	Diagnoses of AF coded and recoded in electronic health records (e.g. CALIBER) using standardised classification systems such as Read and ICD-10	**Atrial fibrillation:** 14AN.00 H/O: atrial fibrillation 212R.00 Atrial fibrillation resolved 662S.00 Atrial fibrillation monitoring 6A9..00 Atrial fibrillation annual review G573200 Paroxysmal atrial fibrillation G573400 Permanent atrial fibrillation G573500 Persistent atrial fibrillation 3272.00 ECG: atrial fibrillation G573000 Atrial fibrillation G573300 Non-rheumatic atrial fibrillation G573.00 Atrial fibrillation and flutter G573z00 Atrial fibrillation and flutter NOS 3273.00 ECG: atrial flutter G573100 Atrial flutter I48 Atrial fibrillation and flutter (ICD-10) **No atrial fibrillation:** Inferred from absence of above clinical codes
Imaging	Electrocardiography (i.e. the gold standard method for diagnosing AF)	**Atrial fibrillation:**  **No atrial fibrillation:** 

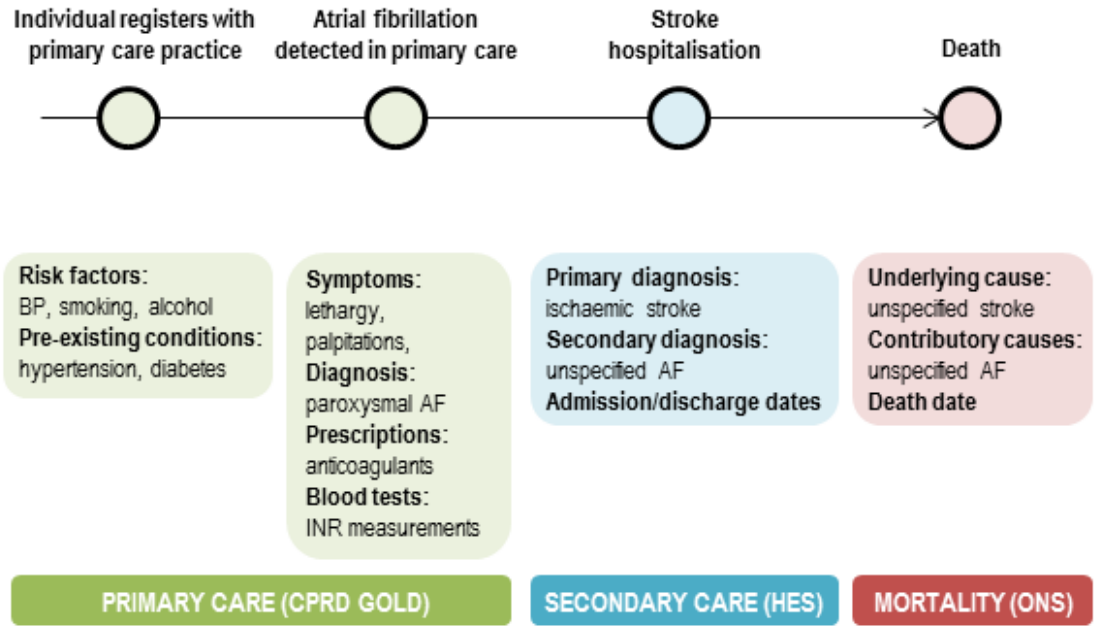
Free text	Clinical notes, discharge summaries and letters to GPs containing unstructured information relevant to diagnoses of AF[202]	**Atrial fibrillation:**
		`Patient 65 years male admitted with shortness of breath and palpitations. Known T2DM and HT. HR 124 irregular. ECG showed AF. CHADSVasc of 3. Initiated with warfarin.`
		
		No atrial fibrillation:
		`Patient 73 year old female admitted by ambulance with chest pain. HR 130 irregular, BP 145/73. ECG normal.`

Genomics	Genetic variants associated with AF[203]	**Atrial fibrillation [genetic basis]:** METTL11B-KIFAP3, ANXA4–GMCL1, CEP68, TTN–TTN-AS1, KCNN2, KLHL3–WNT8A–FAM13B, SLC35F1–PLN, ASAH1–PCM1, SH3PXD2A, KCNJ5, KCNN3, PRRX1, CAND2, PITX2, GJA1, CAV1/2, C9orf3, SYNPO2L, NEURL1, TBX5, SYNE2, HCN4, ZFHX3 **No atrial fibrillation [genetic basis]:** Inferred from absence of above genetic variants

Abbreviations: H/O – history of, NOS – not otherwise specified, T2DM – Type II diabetes mellitus, HT - hypertension, HR – heart rate, BP – blood pressure.

3.8 Chapter figures

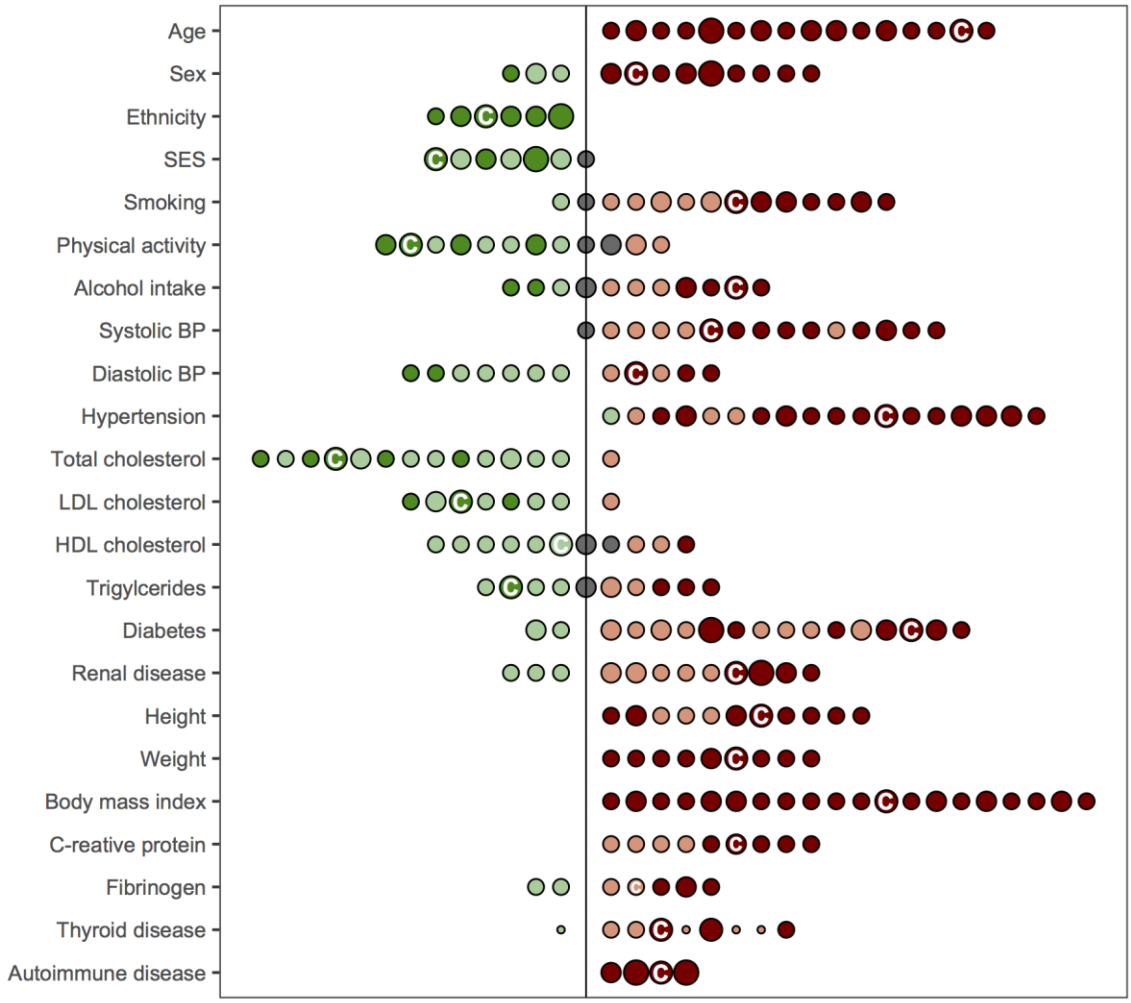
Figure 3.1 Illustrative diagram of CALIBER linked primary care, secondary care and mortality records with complementary information captured on the onset and progression of atrial fibrillation



Notes: Example of longitudinal clinical information captured in primary care, secondary care and mortality records for a theoretical individual with atrial fibrillation. Figure is adapted from Denaxas SC, George J, Herrett E, Shah AD, Kalra D, Hingorani AD, Kivimaki M, Timmis AD, Smeeth L, Hemingway H. Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). Int J Epidemiol. 2012 Dec;41(6):1625-38. doi: 10.1093/ije/dys188. Epub 2012 Dec 5. Original figure shows example patient journey for an individual with coronary disease and has been modified with elements relevant to the onset and progression of atrial fibrillation.

Abbreviations: BP – blood pressure, AF – atrial fibrillation, INR – International Normalised Ratio, CPRD GOLD - Clinical Practice Research Datalink GP OnLine Database (source of primary care data), HES - Hospital Episode Statistics (source of secondary care records), ONS – Office For National Statistic (source of mortality records).

Figure 3.2 [updated figure 2.1] Consistency of CALIBER with systematic review and field synopsis findings on the associations of 23 cardiovascular risk factors and incidence of atrial fibrillation.



Notes: each dot represents one report in order of most extreme inverse to most extreme direct point estimate. Dots are colour-coded to indicate significant inverse (RR [95% CI]<1.00; dark green), non-significant inverse (RR< 1.00; light green), null or mixed (RR=1.00 or show opposite associations among subpopulations; grey), non-significant direct (RR>1.00; light red) and significant direct (RR [95% CI]>1.00; dark red) associations and scaled in size to indicate <100 (smallest dot), <1000, <10000, <100000 and ≥100000 (largest dot) atrial fibrillation events. Dots labeled with the letter C indicated CALIBER estimates.

Abbreviations: SES – socioeconomic status, BP – blood pressure, LDL – low density lipoprotein, HDL – high density lipoprotein, RR – relative risk, 95% CI – 95% confidence interval.

Chapter 4

Novel associations between 23 cardiovascular risk factors and incident atrial fibrillation with and without intercurrent cardiovascular disease

4.1 Chapter outline

In the previous two chapters I presented current understanding on atrial fibrillation (AF) risk factors as suggested by the observational literature to date (**chapter 2**), I also described how the CALIBER dataset offers some unique advantages to advance the study of AF (**chapter 3**) including how these data, though not collected for primary research purposes, hold clinically valid disease insights. Here in **chapter 4**, I go on to exploit the CALIBER dataset to investigate the role of intercurrent cardiovascular disease (CVD) in the development of AF. I hypothesised that observational associations between cardiovascular risk factors and incident AF may be influenced by intercurrent diagnoses of CVD. This was an important question under evaluation as part of my systematic review and field synopsis however it could not be answered from the results of existing studies.

4.2 Abstract

Background The existing observational literature on the associations of a range of demographic, behavioural and biological cardiovascular risk factors in relation to incidence of AF suggest similarities (e.g. obesity and hypertension), as well as important differences (e.g. lipids and ethnicity), in the risk factors for AF, as compared to other CVDs like myocardial infarction and stroke. However, the role of intercurrent CVD in the development of AF remains uncertain.

Methods I used cardiovascular risk factor data on up to 1,949,052 individuals initially free from AF and other pre-existing CVD as available in CALIBER linked primary care, secondary care and mortality records between 1998 and 2010. I defined and investigated the risk factors for two diverse AF endpoints: AF with (AF^+) and AF without (AF^-) an intercurrent diagnosis of CVD. I used Cox regression adjusted for age and sex and stratified on GP practice to model risk factor associations with follow-up defined as the difference between date of risk factor measurement and date of AF^+ or AF^- events or censoring due to administration reasons as the underlying time scale. I used forest plots to visualise differences in risk factor associations (i.e. hazard ratios and 95% confidence intervals; HR [95% CI]) with AF^+ or AF^-.

Results Over a mean (standard deviation; SD) follow-up of 6.1 (4.2) years, a total of 50,097 incident AF events occurred: 12,652 (25.3%) with AF^+ and 37,445 (74.7%) with AF^-. Availability of risk factor data ranged from 16,594 (0.9%) available measures for fibrinogen (a non-routinely collected biomarker) to 1,949,052 (100%) available measures for age and gender and diagnoses of hypertension, diabetes mellitus, renal failure, thyroid disease, rheumatoid arthritis and psoriasis which are defined based on presence or absence of a clinical code. Current smoking, hypertension, and diabetes showed consistent direct associations with AF^+ and AF^- albeit with

stronger risk estimates for AF^+. While for heavy drinking and lower lipids levels inconsistent associations were shown with AF^+ and AF^-.

Conclusion AF has diverse clinical presentations including AF with and AF without intercurrent CVD. A better understanding of these clinical distinctions will help to direct research into risk factors. EHRs can help in this regard.

4.3 Introduction

Electronic health record (EHR) resources, such as CALIBER,[5] consist of large-scale population denominators (e.g. whole countries or regions) and a broad range of clinical variables (i.e. are not restricted to pre-defined measures collected for narrow study objectives),[204] which provides a unique advantage to ask new questions about poorly understood clinical conditions such as AF. The systematic review and field synopsis I conducted on the associations of 23 cardiovascular risk factors and incident AF involving 32 population based cohorts of 20 million participants was helpful in understanding the current field of AF risk factor research.[2] The review suggested similarities (e.g. obesity and hypertension) as well as, potentially, some important differences (e.g. lipids and ethnicity) in the risk factors for AF as compared to other CVDs (e.g. myocardial infarction and stroke), warranting further investigation.

A key question under evaluation in my systematic review was the extent to which existing studies had investigated the impact of intercurrent CVDs in the development of AF. This question is important to understand whether risk factors lead directly to AF, or rather lead to CVD, which in turn leads to AF. Unfortunately the question of intercurrent CVD could not be answered based on the current observational literature.[2] Traditional epidemiological cohort studies (e.g. the Framingham Heart Study with 1544 incident cases AF accrued over 50 years of follow-up[191]), as were most in my systematic review, are limited in ability to answer deeper mechanistic questions about how diseases develop for reasons relating to sample size and type and frequency of clinical information capture. EHRs, on the other hand, consist of large numbers of individuals and continuous capture of clinical information means the way that risk factors progress into disease can be accurately tracked.[204] EHRs therefore have an important role in addressing the question of intercurrent CVD in the development of AF.

In order to untangle the relationship between risk factors, AF and the influence of other CVDs, as well as to identify opportunities for AF primary prevention, the cohorts included in my systematic review were either populations initially free from pre-existing CVD or general populations in which the level of baseline CVD reflected prevalence in the general population. Secondary prevention populations (e.g. all with pre-existing CVDs such as myocardial infarction or heart failure) were excluded in order to minimise any biases in risk factor values due to established structural heart damage. In analysing the relationship between risk factors and AF, over 80% of included reports did not account for intercurrent incident CVD events thus rendering it impossible to understand, from the current observational literature, the direct influence that these risk factors have on AF. Among the few reports that did account for intercurrent incident

CVD events, the methods used included censoring individuals in the event of a CVD diagnosis[109 110] and making model adjustments using time-varying covariates.[80 81 84 91 101 103 104 116 123 125 129 140] Time-varying covariates, account for changes in risk factors over time and how this impacts upon outcomes, however are difficult to implement in EHRs because data are captured at irregular time points.[205]

I therefore decided to define and investigate the risk factors for two diverse AF trajectories: AF with (hereafter referred to as AF^+) and AF without (hereafter referred to as AF^-) an intercurrent diagnosis of CVD. Analysing AF^+ and AF^- side by side allows for similarities, as well as any, potentially important, differences between them to be more easily compared.

4.4 Methods

4.4.1 Analysis dataset

I used the same analysis dataset as was used in **chapter 3** to investigate the clinical validity and consistency of CALIBER records in estimating the associations of 23 cardiovascular risk factors with an all-encompassing AF endpoint (i.e. AF^+ and AF^- modelled together as in existing literature). The full methods for building this dataset, which links primary care, secondary care and mortality records are provided in section 4 of the **appendix**. However, to recap and further reflect on these methods, the key analytic points are as follows:

Individuals included in the dataset

Individuals were included based on the following criteria:

1. at least 30 years of age
2. without prior record of diagnosis of AF
3. without prior record of diagnosis of ten CVDs (heart failure, myocardial infarction, unstable angina, stable angina, coronary heart disease, abdominal aortic aneurysm, peripheral arterial disease, haemorrhagic stroke, ischaemic stroke, and transient ischaemic attack)
4. with a minimum of one year follow-up at a GP practice with research quality data between 1998 and 2010

Minimum age criteria was applied because of low incidence of AF in individuals younger than 30 years of age, and because early onset AF could be the result of a congenital heart malformation rather than the result of a CVD risk factor.[206] Individuals with prior AF and prior CVD diagnoses were excluded in order to focus on risk factor associations in initially healthily individuals and limit any biases in risk factor values due to established structural heart damage (i.e. reverse causation). I focussed on the exclusion of the ten cardiovascular disorders listed above as they reflect the most prevalent CVDs in Europe,[207] are commonly used in composite CVD endpoints in clinical trials,[208] and have robust EHR definitions as used in prior CALIBER studies.[148 181-185 187 189 209-214] One year of follow-up at a GP practice with research quality data[174] was required to accurately ascertain baseline disease status.[215] The study period 1998 to 2010 refers to when all data sources are in alignment. Baseline was defined as the earliest date that the above criteria were satisfied.

Data definitions for 23 cardiovascular risk factors and AF outcomes

I used existing EHR definitions for the 23 risk factors and AF cases as available on the CALIBER portal and in published papers (e.g. sex,[181] ethnicity,[182] socio-economic status,[183] smoking status,[184] alcohol,[185] blood pressure,[148] lipids,[186] diabetes mellitus,[187], height, weight and body mass index,[188], autoimmune diseases[189] and AF[49]). Full details are provided in section 4 of the **appendix**.

To maximise the use of observed risk factor data, I selected values available at baseline (i.e. earliest date study inclusion criteria fulfilled) using a look-back period of one year (unless otherwise specified in the appendix), together with values available after baseline but before date of occurrence of any CVD or AF events. This available case strategy,[190] therefore means the number of individuals included in analyses varies from risk factor to risk factor, as do the related age, sex, follow-up and outcomes characteristics, and findings should be interpreted on a risk factor by risk factor basis, rather than for the cohort as a whole. **Table S4.2** of the **appendix** summarises the characteristics of individuals with data available for each risk factor. Most notably, lipids and inflammatory factors were recorded in older individuals (e.g. mean (SD) age was 55.2 (13.7) years for total cholesterol and 54.6 (15.3) years for C-reactive protein vs. 46.9 (15.3) years in the overall cohort). Blood pressure, anthropometric and inflammatory factors were less often recorded in men (e.g. 44.0% had measured systolic blood pressure, 43.9% had measured weight, and 37.2% had measured C-reactive protein vs. 48.7% men in the overall cohort), and AF outcomes were higher in individuals with recorded ethnicity and measured fibrinogen (3.6% and 3.7% had AF respectively vs. 2.6% with AF in the overall cohort).

In addition to using Morley's definition for AF in linked primary and secondary care records,[49] I also included AF recorded as a cause of death in ONS using ICD 9 codes 42731 (atrial fibrillation) and 42732 (atrial flutter) for deaths registered before the year 2000 and ICD 10 code I48 (atrial fibrillation and flutter) for deaths registered after the year 2000. Individuals with incident AF were subcategorised into AF^+ and AF^- based on whether an intercurrent diagnosis of CVD (based on the ten CVDs excluded at baseline) preceded the diagnosis of AF.

4.4.2 Statistical analysis

I assessed differences in risk factors distributions in individuals without incident AF, with AF^+ and with AF^- using means (SDs) for continuous variables and numbers and percentages (%) for categorical variables. I then used Cox regression models adjusted for age and sex and stratified on GP practice (in order to account for local level differences in the management, recording and coding of risk factors and disease endpoints) to estimate associations between risk factors and AF^+ and AF^-. The underlying timescale was follow-up, calculated as the difference between date of censoring and date of risk factor measurement (i.e. baseline if measure available at baseline, or date of measurement if available after baseline). Individuals were censored in the event of occurrence of AF^+ or AF^-, death from a cause other than AF, and for administrative reasons such as deregistration from GP practice and end of GP practice follow-up. The Cox model assumption of proportionality was assessed with plots (**appendix**). Continuous variables

are analysed for strength of association per one SD increase in value. Analyses were performed using statistical software Stata/SE 13.1 and figures produced using R 3.2.0.

4.5 Results

4.5.1 Availability of data on 23 cardiovascular risk factors

As **figure 4.1** (flow diagram) depicts, the final analysis dataset comprised a cohort of 1,949,052 individuals. Availability of risk factor data ranged from 16,594 (0.9%) measures for fibrinogen (a non-routinely collected biomarker) to 1,949,052 (100%) measures for age, gender, hypertension, diabetes mellitus, renal failure, thyroid disease, rheumatoid arthritis and psoriasis. The percentage of individuals with AF^+ and AF^- was comparable across all 23 cardiovascular risk factors.

4.5.2 Distribution of 23 cardiovascular risk factors in individuals without AF or with AF^+ or AF^-

Table 4.1 summarises the risk factor distributions in individuals without incident AF, with AF^+ and with AF^-. As shown individuals without AF had lowest mean (SD) age, had a lower percentage of White ethnicity, were more likely to be a current smoker, moderate drinker and engage in moderate levels of physical activity, had lower mean (SD) body mass index and were less likely to have hypertension and type II diabetes than individuals with AF^+ or AF^-. In general, individuals with AF^+ had higher average values and percentages of cardiovascular risk factors than individuals with AF^-. Individuals with AF^+ had the highest mean (SD) age, were more likely to be men, were least likely to be within the least deprived socio-economic status quintile, and had the highest levels of physical inactivity, hypertension and type II diabetes.

4.5.3 Associations of 23 cardiovascular risk factors with AF^+ and AF^-

Risk factors associations in terms of HR [95% CI] for AF^+ and AF^- are detailed in **table 4.2,** visualised in **figures 4.2 to 4.6** and described as follows:

- **Demographics**

Age, per 15 years, showed direct associations with AF^+ and AF^-, but with a higher risk estimate for AF^+ (HR [95%CI] for AF^+ vs. AF^-: 4.68 [4.58, 4.77] vs. 4.02 [3.98, 4.07]). For women compared to Men, inverse associations were shown with AF^+ and AF^- with a lower risk estimate for AF+ (HR [95%CI] for AF^+ vs. AF^-: 0.56 [0.54, 0.58] vs. 0.71 [0.70, 0.73]). Compared to White ethnicity, Black ethnicity showed inverse associations with AF^+ and AF^- with similar risk estimates (HR [95%CI] for AF^+ vs. AF^-: 0.51 [0.38, 0.68] vs. 0.45 [0.38, 0.53]), South Asian ethnicity showed inverse associations with AF^-, but not with AF^+ (HR [95%CI] for AF^+ vs. AF^-: 0.91 [0.74, 1.13] vs. 0.39 [0.32, 0.47]), and Other and Mixed ethnicities showed inverse associations with AF^+ and AF^- with similar risk estimates (HR [95%CI] for AF^+ vs. AF^-: 0.76 [0.60, 0.94] vs. 0.66 [0.58, 0.75]). For most (compared to least) deprived socio-economic status, direct associations were shown with AF^+ and AF^- with a higher risk estimate for AF+ (HR [95%CI] for AF^+ vs. AF^-: 1.39 [1.29, 1.50] vs. 1.12 [1.08, 1.18]).

- **Health behaviours**

For current smokers (compared non-smokers) direct associations were shown with AF^+ and AF^- with a higher risk estimate for AF+ (HR [95%CI] for AF^+ vs. AF^-: 1.66 [1.56, 1.77] vs. 1.21 [1.16,1.25]). For heavy alcohol drinkers (compared to non-drinkers) a direct association was shown with AF^- but not with AF^+ (HR [95%CI] for AF^+ vs. AF^-: 1.17 [0.81, 1.67] vs. 1.99 [1.68, 2.34]). For vigorous physical activity (compared to inactivity) inverse associations were shown with AF^+ and AF^- with similar risk estimates (HR [95%CI] for AF^+ vs. AF^-: 0.59 [0.53, 0.70] vs. 0.72 [0.65, 0.79]).

- **Blood pressure**

Systolic blood pressure, per 20 mmHG, showed direct associations with AF^+ and AF^- with a higher risk estimate for AF^+ (HR [95%CI] for AF^+ vs. AF^-: 1.22 [1.20, 1.24] vs. 1.12 [1.11, 1.13]). Diastolic blood pressure, per 10 mmHG, showed direct associations with AF^+ and AF^- with similar risk estimates (HR [95%CI] for AF^+ vs. AF^-: 1.15 [1.13, 1.17] vs. 1.12 [1.11, 1.13]). Hypertension (compared to no record of hypertension) showed direct associations with AF^+ and AF^- with a higher risk estimate for AF^+ (HR [95%CI] for AF^+ vs. AF^-: 2.19 [2.11, 2.27] vs. 1.65 [1.62, 1.69]).

- **Lipids**

Total cholesterol, per 1.1 mmol/L, showed an inverse association with AF^- but not with AF^+ (HR [95%CI] for AF^+ vs. AF^-: 0.99 [0.96, 1.02] vs. 0.85 [0.84, 0.87]). Low density lipoprotein cholesterol, per 1.0 mmol/L, showed an inverse association with AF^- but not with AF^+ (HR [95%CI] for AF^+ vs. AF^-: 0.97 [0.93, 1.01] vs. 0.86 [0.84, 0.88]). High density lipoprotein cholesterol, per 0.4 mmol/L showed an inverse association with AF^+ but not with AF^- (HR [95%CI] for AF^+ vs. AF^-: 0.88 [0.84, 0.91] vs. 1.01 [0.99, 1.03]). Triglycerides, per 1.2 mmol/L, showed an inverse association with AF^- but not with AF^+ HR [95%CI] for AF^+ vs. AF^-: 1.08 [1.05, 1.11] vs. 0.92 [0.90, 0.94]).

- **Metabolic factors**

Type II diabetes mellitus (compare to no record of type II diabetes mellitus) showed direct associations with AF^+ and AF^- with a higher risk estimate for AF^+ (HR [95%CI] for AF^+ vs. AF^-: 2.03 [1.94, 2.12] vs. 1.45 [1.41, 1.49]). Renal failure (compare to no record of renal failure) showed direct associations with AF^+ and AF^- with similar risk estimates (HR [95%CI] for AF^+ vs. AF^-: 1.78 [1.54, 2.05] vs. 1.53 [1.41, 1.66]).

- **Anthropometrics factors**

Height , per 10 cm, showed direct associations with AF^+ and AF^- with a higher risk estimate for AF^- (HR [95%CI] for AF^+ vs. AF^-: 1.15 [1.11, 1.18] vs. 1.38 [1.36, 1.40]). Weight, per 17 kg, showed direct associations with AF^+ and AF^- with similar risk estimates (HR [95%CI] for AF^+ vs. AF^-: 1.40 [1.36, 1.43] vs. 1.39 [1.37, 1.41]). Body mass index, per 5 kg/m^2, showed direct associations with AF^+ and AF^- with a higher risk estimate for AF^+ (HR [95%CI] for AF^+ vs. AF^-: 1.31 [1.28, 1.33] vs. 1.23 [1.22, 1.25]).

- **Inflammatory factors**

C-reactive protein, per 13.0 mg/L, showed direct associations with AF^+ and AF^- with similar risk estimates (HR [95%CI] for AF^+ vs. AF^-: 1.16 [1.12, 1.21] vs. 1.11 [1.09, 1.13]). Fibrinogen, per 1.2 g/L, showed similar null associations with AF^+ and AF^- (HR [95%CI] for AF^+ vs. AF^-: 1.13 [0.92, 1.37] vs. 1.07 [0.97, 1.17]).

- **Thyroid disease**

Hypothyroidism (compared to no record of hypothyroidism) showed direct associations with AF^+ and AF^- with a higher risk estimate for AF^+ (HR [95%CI] for AF^+ vs. AF^-: 1.33 [1.21, 1.47] vs. 1.13 [1.06, 1.20]). Hyperthyroidism (compared to no record of hyperthyroidism) showed direct associations with AF^+ and AF^- with similar risk estimates (HR [95%CI] for AF^+ vs. AF^-: 1.35 [1.10, 1.65] vs. 1.30 [1.16, 1.46]).

- **Autoimmune disease**

Rheumatoid arthritis (compared to no record of rheumatoid arthritis) showed direct associations with AF^+ and AF^- with a higher risk estimate for AF^+ (HR [95%CI] for AF^+ vs. AF^-: 1.87 [1.65, 2.12] vs. 1.48 [1.36, 1.60]). Psoriasis (compared to no record of psoriasis) showed direct associations with AF^+ and AF^- with similar risk estimates (HR [95%CI] for AF^+ vs. AF^-: 1.17 [1.05, 1.31] vs. 1.12 [1.05, 1.19]).

4.6 Discussion

Overview of key findings

This study draws upon the breadth and depth of clinical information contained within the CALIBER resource[5] to provide novel insights into the associations of 23 cardiovascular risk factors in relation to two diverse AF trajectories: AF with (AF^+) and AF without (AF^-) an intercurrent CVD. Whereas the existing observational literature did not fully account for the influence of intercurrent CVDs,[2] by doing so I found that standard cardiovascular risk factors, age, sex, smoking, blood pressure and diabetes mellitus, have stronger direct associations with AF^+ than AF^-, while the prior reported discordant association between higher lipids levels and lower risk of AF, was shown for AF^- but not for AF^+. These findings could imply that existing CVD prevention programmes[59 60] may work for some, but not all, AF and a firmer understanding of the different clinical presentations of AF is therefore needed in order to more accurately identify risk factors upon which to intervene. The ability to detect diverse AF trajectories in CALIBER strengthens the support for the use of EHR data in these investigations. The work presented here was, however, largely exploratory; designed to provide an initial screening of the risk data to identify possible signals with AF^+/AF^- risk. More detailed work, studying each individual risk factor in turn, is needed to understand how sources of bias in the collection of EHRs data (e.g. missed AF cases and comorbidities, over/under-represented populations and limited information on disease severity) may impact the interpretation of findings (further reflections provided below).

Diverse AF trajectories

As already described the existing literature on AF risk factors, as identified in my systematic review,[2] failed to address the question of the role of intercurrent CVDs in the development of AF. This makes it difficult to directly compare with the results of the present analysis. Two relevant reports recently published using the Atherosclerosis Risk in Communities (ARIC) and Framingham Heart Study (FHS) cohorts investigated timing and sequence of acquiring CVD risk factors and diagnoses in relation to future AF risk.[216 217] In ARIC, a rapid rise in prevalence of stroke, myocardial infarction and heart failure was found near the time of AF diagnosis.[216] While in FHS, using a form of latent class analysis (a data driven method of identifying underlying patterns and sub-groupings of risk factors),[218] five distinct systolic blood pressure trajectories were found to be associated with 15 year risk of AF.[217] Overall this confirms that diverse trajectories leading to AF exist and future studies need to more adequately account for these, rather than focussing on all-encompassing AF endpoints. Looking beyond the current observational evidence, a number of consensus documents,[219 220] and more recently, changes in European clinical practice guidelines[3] have emerged offering new ideas on AF mechanisms. The updated 2016 European Society of Cardiology (ESC) guidelines for the management of AF outlines seven clinical distinctions of AF (which I go on to consider in greater detail in **chapter 5**) including AF secondary to structural heart disease, defined as:

"AF in patients with left ventricular systolic or diastolic dysfunction, long-standing hypertension with left ventricular hypertrophy, and/or other structural heart disease".

Although not an exact match in definition, parallels can be drawn with the two AF endpoints investigated here with AF^+ reflecting more severe structural heart damage than AF^- and hence the stronger associations with CVD risk factors make more plausible biological sense. The new AF definitions suggested by the ESC are largely based upon expert consensus rather than any large-scale quantitative evidence and therefore EHR resources, such as CALIBER, provide a viable setting in which understanding about these can be refined and validated.

Clinical implications

From a clinical stand-point, the results of this study suggest that AF will be prevented in some individuals based on existing management strategies to tackle obesity, smoking, and hypertension.[59 60] However, there are other groups of individuals who develop AF via diverse biological mechanisms which are not fully understood at present. A firmer understanding of the different clinical presentations of AF can help to look back 'upstream' for risk factors and opportunities upon which to intervene.[35] It is unclear, at this stage, how current risk models for prediction of AF[40 115] (which have been derived based on all-encompassing definitions of AF) perform in relation to higher resolution AF endpoints (e.g. AF^+ and AF^-). The clinical utility of these risk models should be more robustly tested before being adopted in clinical practice.

Research implications

The results of this study give rise to three key implications for future research. First, and above all else, differences in the magnitude of risk factor associations shown for AF^+ and AF^- highlight the importance of adequately accounting for these in observational analyses. Future studies, hoping to identify novel risk factors for AF, therefore need to more carefully consider AF as, not just a single entity, but a reflection of multiple diverse biological mechanisms leading to the same clinical manifestation. Second, the results of this study strengthen the assertion that AF mechanisms research, which is usually reserved for the basic science setting,[221] is also possible using EHRs. In response to the new ESC definitions for seven clinical AF distinctions,[3] a relevant next line of inquiry would be to explore the extent to which these can be operationalised in EHRs (as I proceed to do in **chapter 5** and **chapter 6**). Operationalising these will allow the robustness of definitions to be tested in terms of level upon which they are clinically distinct (or conversely overlapping), relative contributions to AF populations, as well as helping to refine understanding about how different AF subtypes develop and progress. The benefit of conducting this research in the EHR setting with standardised ways of classifying disease (e.g. ICD-10) means definitions can also be implemented in international systems (e.g. the Danish[42] and Swedish[43] nation-wide cohorts). This will bring together even larger populations of individuals with AF and generate even more opportunities for research. Finally, the results of this study are also useful for research in the clinical trial setting. With the exception of several post hoc analyses in relation to statins and blood pressure lowering medications,[145-147] trials of healthy participants (i.e. without pre-existing CVD) with AF as the primary endpoint have been lacking.[35] However, as this study shows, the strongest associations between CVD risk factors and incident AF are mediated by intercurrent CVDs such as myocardial infarction and coronary heart disease. This suggests there is little justification to include AF as a primary endpoint in future CVD prevention trials.

Strengths and limitations

The strengths and limitations of this study revolve around what is feasibly possible within the CALIBER dataset. Population size and breadth of clinical variables allowed more intricate investigations than before on how 23 demographic, behavioural and biological cardiovascular risk factors relate to two higher resolution AF endpoints (i.e. AF^+ and AF^- intercurrent CVD), and all of this was possible within a single dataset. Traditional epidemiological cohort studies lacking large scale population denominators are, on the other hand, often limited to studies on all-encompassing AF definitions in relation to a narrower set of research measures.[204] Limitations of the present study go back to what is not currently captured in CALIBER as well as sources of bias in the data that is captured within CALIBER. Robust definitions for risk factors and AF endpoints are necessary in order to accurately estimate the associations between them. Researchers using EHR data therefore invest considerable time and resources into developing disease definitions, which hold clinical validity (e.g. Morley's EHR definition for AF). However, unless each and every disease diagnosis and risk factor is individually verified it is impossible to definitively quantify the extent or direction of bias in EHR data. Lack of ECG tracings (i.e. the gold standard method for diagnosing AF) means that some AF cases will have inevitably been

missed. AF events may also have failed to be recorded, if, for example, incidentally detected during a hospital admission for a predominant cardiac disease (e.g. an acute myocardial infarction). Systematic under-recording of AF events is likely to attenuate associations with risk factors towards the null. Over- and under-represented populations in EHR data are another potential source of bias. In the UK, the vast majority of the general population are registered with a National Health Service (NHS) general practitioner and baseline levels of disease risk (including health behaviours, anthropometric measures and blood pressure) are captured in the form of new patient questionnaires and NHS health check schemes. Other risk markers, such as inflammatory factors CRP and fibrinogen, are not routinely collected measures and are thus more likely to be recorded in sicker populations as part of a clinical investigation. An overrepresentation of sicker patients with CRP and fibrinogen measures is therefore likely to have amplified the direct risk factor associations shown for AF^+ and AF^-. CRP and fibrinogen are also examples of where reverse causation could be at play; in that the presence of AF may induce an inflammatory response, rather contributing to its development. So-called 'worried well' populations are also likely to over-represented in EHRs, while 'doctor avoiders' and those genuinely healthy are likely be under-represented. Individuals who avoid doctors are less likely to have comorbidities recorded. This means that some AF^+ cases will have been misclassified as AF^- and therefore the difference between these two groups may actually be even more pronounced than was observed. The resolution of clinical coding in EHR data can also be limited, particularly in relation to disease severity. In analysing renal failure, for example, I used a definition that did not differentiate acute, mild, moderate and severe forms of renal failure. Under the assumption that more severe disease cases are likely to be recorded in EHRs, again this is likely to have amplified the direct risk factor associations shown for AF^+ and AF^-. One method EHR researchers use to assess potential for bias in EHR data is to compare findings against estimates from an external non-EHR population. Therefore given that the risk factor associations estimated in my earlier validation study (**chapter 3**) were consistent with existing clinical evidence generated from consented cohorts, I am reassured that even if a smaller number of cases have been missed, it has not majorly impacted the study quality. The question of intercurrent CVD in the development of AF, as in whether AF^+ reflects more severe structural heart damage than AF^-, would be interesting to address using whole heart imaging data (e.g. magnetic resonance imaging; MRI). However, as with ECG tracing data, the infrastructure to extract these from hospital systems does not exist in the UK at present.

The methods I used to select and analyse risk factor data also have positive and negative aspects. Rather than focussing on a single cardiovascular risk factor in fine granular detail, I instead, like in my systematic review and field synopsis, took a horizontal view across a broad range of factors. I took this approach for the purpose of prioritisation in future research (e.g. the discordant lipids findings), as well as to highlight the scope of data opportunities in CALIBER. This approach, however, came at the expense of not being able to fully address missing data. The available case strategy[190] I employed to maximise the use of observed risk factor data meant that the number of individuals and their associated characteristics varied for each risk factor and therefore results should be interpreted in isolation and not for the cohort as a whole.

Fixing baseline dates based on when EHR data are available may again introduce bias because of health seeking behaviours due to sickness and ill-health. Again, linking back to my earlier validation study showing results closely matching existing literature, it is unlikely that the methods for selecting risk factor data have impacted study quality. Of course, more sophisticated methods for handling missing data (e.g. multiple imputation) are available, however these can be challenging to implement in EHRs because of data size and structure.[178] Multiple imputation was not a suitable option for this analysis because each individual analysis model requires a separate strategy for imputing covariate data, thus with 23 different risk factors under investigation it was not scalable to impute. Likewise it was not scalable to provide any adjusted analyses. The discordant observational association shown for higher lipids and lower risk of AF as found in my systematic review and now shown for AF^- but not for AF^+ could potentially be explained by statin use. It could also be an example of survivor bias in that only those individuals who survive other CVDs like myocardial infarction and stroke or other competing risks will go on to develop AF. Single risk factors investigations with imputed covariate data, fully adjusted models and further considerations of the potential biases of EHR data will be the subject of future analyses in CALIBER. Finally, it should be remarked that the associations presented here are observational and, as always, may also be influenced by unmeasured confounding. The gold-standard study design to determine the risk factors upon which to intervene to prevent AF would be a randomised clinical trial.

4.7 Conclusion

AF has diverse clinical presentations including AF with and AF without intercurrent CVD. A better understanding of these clinical distinctions will help to direct research into risk factors. EHRs can help in this regard.

4.8 Chapter summary

To summarise, this chapter presented a novel investigation into the associations of 23 cardiovascular risk factors and AF both with and without an intercurrent CVD with differences in the magnitude of risk factor associations. Standard cardiovascular risk factors, age, sex, smoking, blood pressure and diabetes mellitus, had stronger direct associations with AF^+ than AF^-, while higher lipids levels were inversely associated with AF^- but not with AF^+ and heavy alcohol consumption was directly associated with AF^- but not with AF^+. These differences suggest that existing CVD prevention programmes[59 60] may work for some, but not all, AF. A firmer understanding of the different clinical presentations of AF can help to look back 'upstream' for risk factors and opportunities upon which to intervene.[35] The importance of looking at AF as a collection of diverse biological mechanisms leading to a similar clinical manifestation, rather than a single entity, was reflected in recent updates to ESC guidelines for the management of AF in which new ideas on AF mechanisms were put forward.[3] The ability to detect diverse AF trajectories in CALIBER, as I have demonstrated in this chapter, exemplifies the value of using these data in research on AF. In the next chapter, **chapter 5,** I go on to evaluate whether the new ESC definitions for AF can be operationalised in CALIBER records, as well as describing more generally the methods for creating and testing EHR definitions and algorithms for robustly identifying disease cases.

4.9 Chapter tables

Table 4.1 Risk factor distributions in individuals without AF, with AF^{+} and with AF^{-}

Demographics	Individuals without AF	Individuals with AF^{-}	Individuals with AF^{+}	Individuals overall
Age (years)	46.3 (15.0)	68.1 (13.6)	69.6 (11.6)	46.9 (15.3)
Gender:				
Men	925407 (48.7%)	17942 (47.9%)	6683 (52.8%)	17942 (48.7%)
Women	973548 (51.3%)	19503 (52.1%)	5969 (47.2%)	19503 (51.3%)
Ethnicity:				
White	894105 (90.1%)	28778 (98.3%)	10319 (97.9%)	933202 (90.4%)
South Asian	32177 (3.2%)	142 (0.5%)	51 (0.5%)	32370 (3.1%)
Black	30089 (3.0%)	127 (0.4%)	93 (0.9%)	30309 (2.9%)
Other and Mixed	35652 (3.6%)	239 (0.8%)	81 (0.8%)	35972 (3.5%)
Socioeconomic status:				
First quintile (least deprived)	378888 (20.0%)	7118 (19.1%)	2096 (16.6%)	388102 (20.0%)
Second quintile	378765 (20.0%)	7842 (21.0%)	2637 (20.9%)	389244 (20.1%)
Third quintile	376727 (19.9%)	7904 (21.2%)	2676 (21.2%)	387307 (20.0%)
Fourth quintile	378432 (20.0%)	7486 (20.1%)	2609 (20.7%)	388527 (20.0%)
Fifth quintile (most deprived)	377357 (20.0%)	6980 (18.7%)	2584 (20.5%)	386921 (19.9%)
Health behaviours				
Smoking status:				
Non smoker	952850 (59.4%)	20036 (64.9%)	6068 (63.3%)	978954 (59.6%)
Former smoker	289330 (18.1%)	7018 (22.7%)	2190 (22.8%)	298538 (18.2%)
Current smoker	360728 (22.5%)	3800 (12.3%)	1331 (13.9%)	365859 (22.3%)
Alcohol drinker status:				
Non-drinker	195658 (14.3%)	4117 (16.0%)	1409 (17.2%)	201184 (14.3%)
Former drinker	47852 (3.5%)	1113 (4.3%)	355 (4.3%)	49320 (3.5%)
Occasional drinker	192920 (14.1%)	4073 (15.9%)	1359 (16.6%)	198352 (14.1%)
Moderate drinker	928473 (67.7%)	16221 (63.2%)	5051 (61.6%)	949745 (67.6%)
Heavy drinker	6246 (0.5%)	149 (0.6%)	31 (0.4%)	6426 (0.5%)
Physical activity status:				
Inactive	82047 (10.6%)	1979 (13.5%)	660 (14.4%)	84686 (10.7%)
Gentle activity	217218 (28.0%)	5530 (37.8%)	1819 (39.8%)	224567 (28.2%)
Moderate activity	401255 (51.7%)	6477 (44.3%)	1924 (42.1%)	409656 (51.5%)
Vigorous activity	75430 (9.7%)	646 (4.4%)	172 (3.8%)	76248 (9.6%)
Blood pressure (mmHG)				
Systolic blood pressure	130.9 (19.9)	147.5 (22.2)	151.2 (21.6)	131.4 (20.2)
Diastolic blood pressure	79.2 (10.8)	83.4 (10.9)	84.1 (10.7)	79.3 (10.9)
Hypertension	420846 (22.2%)	20004 (53.4%)	7800 (61.7%)	448650 (23.0%)
Lipids (mmol/L)				
Total cholesterol	5.6 (1.1)	5.6 (1.1)	5.8 (1.2)	5.6 (1.1)
LDL cholesterol	3.4 (1.0)	3.4 (1.0)	3.5 (1.0)	3.4 (1.0)
HDL cholesterol	1.5 (0.4)	1.5 (0.5)	1.4 (0.4)	1.5 (0.4)
Triglycerides	1.6 (1.2)	1.6 (1.0)	1.8 (1.2)	1.6 (1.2)
Metabolic factors				
Diabetes mellitus:				
Type I	8552 (0.5%)	181 (0.5%)	101 (0.8%)	8834 (0.5%)
Type II	86155 (4.5%)	4908 (13.1%)	2451 (19.4%)	93514 (4.8%)
Uncertain	8118 (0.4%)	383 (1.0%)	180 (1.4%)	8681 (0.4%)
Renal failure	21525 (1.1%)	606 (1.6%)	204 (1.6%)	22335 (1.1%)
Anthropometry				
Height (m)	1.69 (0.1)	1.69 (0.1)	1.68 (0.1)	1.69 (0.1)
Weight (kg)	75.3 (17.1)	77.3 (18.3)	77.6 (17.0)	75.4 (17.1)
Body mass index (kg/m^2)	26.4 (5.2)	27.2 (5.5)	27.5 (5.2)	26.4 (5.2)
Inflammation				
C-reactive protein (mg/L)	8.1 (12.8)	12.4 (17.4)	13.7 (19.0)	8.2 (13.0)
Fibrinogen (g/L)	3.7 (1.2)	4.0 (1.2)	4.0 (1.2)	3.7 (1.2)

Thyroid disease	Individuals without AF	Individuals with AF^-	Individuals with AF^+	Individuals overall
Thyroid disease:				
Hypothyroidism	33972 (1.8%)	1213 (3.2%)	447 (3.5%)	35632 (1.8%)
Hyperthyroidism	7967 (0.4%)	297 (0.8%)	97 (0.8%)	8361 (0.4%)
Uncertain type	4557 (0.2%)	151 (0.4%)	60 (0.5%)	4768 (0.2%)
Autoimmune disease				
Rheumatoid arthritis	11178 (0.6%)	601 (1.6%)	253 (2.0%)	12032 (0.6%)
Psoriasis	39861 (2.1%)	895 (2.4%)	329 (2.6%)	41085 (2.1%)

Notes: continuous values are presented as means and (standard deviations) and categorical variables are presented as percentages.

Abbreviations: AF^+/ AF^- - atrial fibrillation with/without intercurrent cardiovascular disease, mmHG - millimetres of mercury, mmol/L - millimoles per Liter, L/HDL – low/high density lipoprotein cholesterol, m – metres, kg – kilograms, mg/L - milligrams per litre, g/L - grams per litre.

Table 4.2 Age and sex adjusted hazard ratios and 95% confidence intervals for the associations of 23 cardiovascular risk factors with incident AF+ and AF⁻ and all AF combined

Demographics	AF⁻ HR [95%CI]	AF⁺ HR [95%CI]	AF HR [95%CI]
Age per 15.3 years	4.02 [3.98, 4.07]	4.68 [4.58, 4.77]	3.99 [3.95, 4.03]
Gender:			
Men	reference	reference	reference
Women	0.71 [0.70, 0.73]	0.56 [0.54, 0.58]	0.67 [0.66, 0.68]
Ethnicity:			
White	reference	reference	reference
Black	0.45 [0.38, 0.53]	0.51 [0.38, 0.68]	0.46 [0.40, 0.54]
South Asian	0.39 [0.32, 0.47]	0.91 [0.74, 1.13]	0.51 [0.45, 0.59]
Other and Mixed	0.66 [0.58, 0.75]	0.76 [0.60, 0.94]	0.68 [0.61, 0.77]
Socio-economic status:			
First quintile (least deprived)	reference	reference	reference
Second quintile	1.01 [0.97, 1.05]	1.17 [1.09, 1.24]	1.04 [1.01, 1.08]
Third quintile	1.06 [1.02, 1.10]	1.18 [1.10, 1.26]	1.08 [1.05, 1.12]
Fourth quintile	1.06 [1.02, 1.10]	1.24 [1.16, 1.33]	1.10 [1.06, 1.14]
Fifth quintile (most deprived)	1.12 [1.08, 1.18]	1.39 [1.29, 1.50]	1.19 [1.14, 1.23]
Health behaviours			
Smoking status:			
Non smoker	reference	reference	reference
Former smoker	1.20 [1.16, 1.23]	1.30 [1.23, 1.37]	1.21 [1.18, 1.24]
Current smoker	1.21 [1.16, 1.25]	1.66 [1.56, 1.77]	1.27 [1.23, 1.31]
Alcohol drinker status:			
Non-drinker	reference	reference	reference
Former drinker	1.12 [1.05, 1.20]	1.07 [0.95, 1.21]	1.11 [1.04, 1.17]
Occasional drinker	0.97 [0.93, 1.02]	0.92 [0.85, 0.99]	0.96 [0.93, 1.00]
Moderate drinker	1.02 [0.98, 1.06]	0.93 [0.88, 0.99]	1.00 [0.97, 1.04]
Heavy drinker	1.99 [1.68, 2.34]	1.17 [0.81, 1.67]	1.75 [1.50, 2.03]
Physical activity status:			
Inactive	reference	reference	reference
Gentle activity	0.90 [0.85, 0.95]	0.85 [0.78, 0.94]	0.90 [0.86, 0.94]
Moderate activity	0.79 [0.75, 0.83]	0.68 [0.62, 0.75]	0.78 [0.75, 0.82]
Vigorous activity	0.72 [0.65, 0.79]	0.59 [0.53, 0.70]	0.71 [0.67, 0.77]
Blood pressure			
Systolic blood pressure per 20.2 mmHG	1.12 [1.11, 1.13]	1.22 [1.20, 1.24]	1.13 [1.12, 1.14]
Diastolic blood pressure per 10.9 mmHG	1.12 [1.11, 1.13]	1.15 [1.13, 1.17]	1.12 [1.11, 1.13]
Hypertension	1.65 [1.62, 1.69]	2.19 [2.11, 2.27]	1.69 [1.66, 1.72]
Lipids			
Total cholesterol per 1.1 mmol/L	0.85 [0.84, 0.87]	0.99 [0.96, 1.02]	0.88 [0.87, 0.89]
LDL cholesterol per 1.0 mmol/L	0.86 [0.84, 0.88]	0.97 [0.93, 1.01]	0.88 [0.86, 0.89]
HDL cholesterol per 0.4 mmol/L	1.01 [0.99, 1.03]	0.88 [0.84, 0.91]	0.99 [0.97, 1.01]
Triglycerides per 1.2 mmol/L	0.92 [0.90, 0.94]	1.08 [1.05, 1.11]	0.95 [0.93, 0.97]
Metabolic factors			
Diabetes mellitus:			
Type I	1.76 [1.52, 2.04]	3.04 [2.49, 3.70]	1.91 [1.70, 2.15]
Type II	1.45 [1.41, 1.49]	2.03 [1.94, 2.12]	1.51 [1.47, 1.54]
Uncertain	1.66 [1.50, 1.84]	2.31 [1.99, 2.68]	1.75 [1.61, 1.90]
Renal failure	1.53 [1.41, 1.66]	1.78 [1.54, 2.05]	1.55 [1.44, 1.66]
Anthropometry			
Height per 0.1 m	1.38 [1.36, 1.40]	1.15 [1.11, 1.18]	1.32 [1.30, 1.34]
Weight per 17.1 Kg	1.39 [1.37, 1.41]	1.40 [1.36, 1.43]	1.38 [1.37, 1.40]
Body mass index per 5.2 Kg/m²	1.23 [1.22, 1.25]	1.31 [1.28, 1.33]	1.24 [1.23, 1.25]
Inflammation			
C-reactive protein per 13.0 mg/L	1.11 [1.09, 1.13]	1.16 [1.12, 1.21]	1.12 [1.10, 1.14]
Fibrinogen per 1.2 g/L	1.07 [0.97, 1.17]	1.13 [0.92, 1.37]	1.08 [0.99, 1.17]

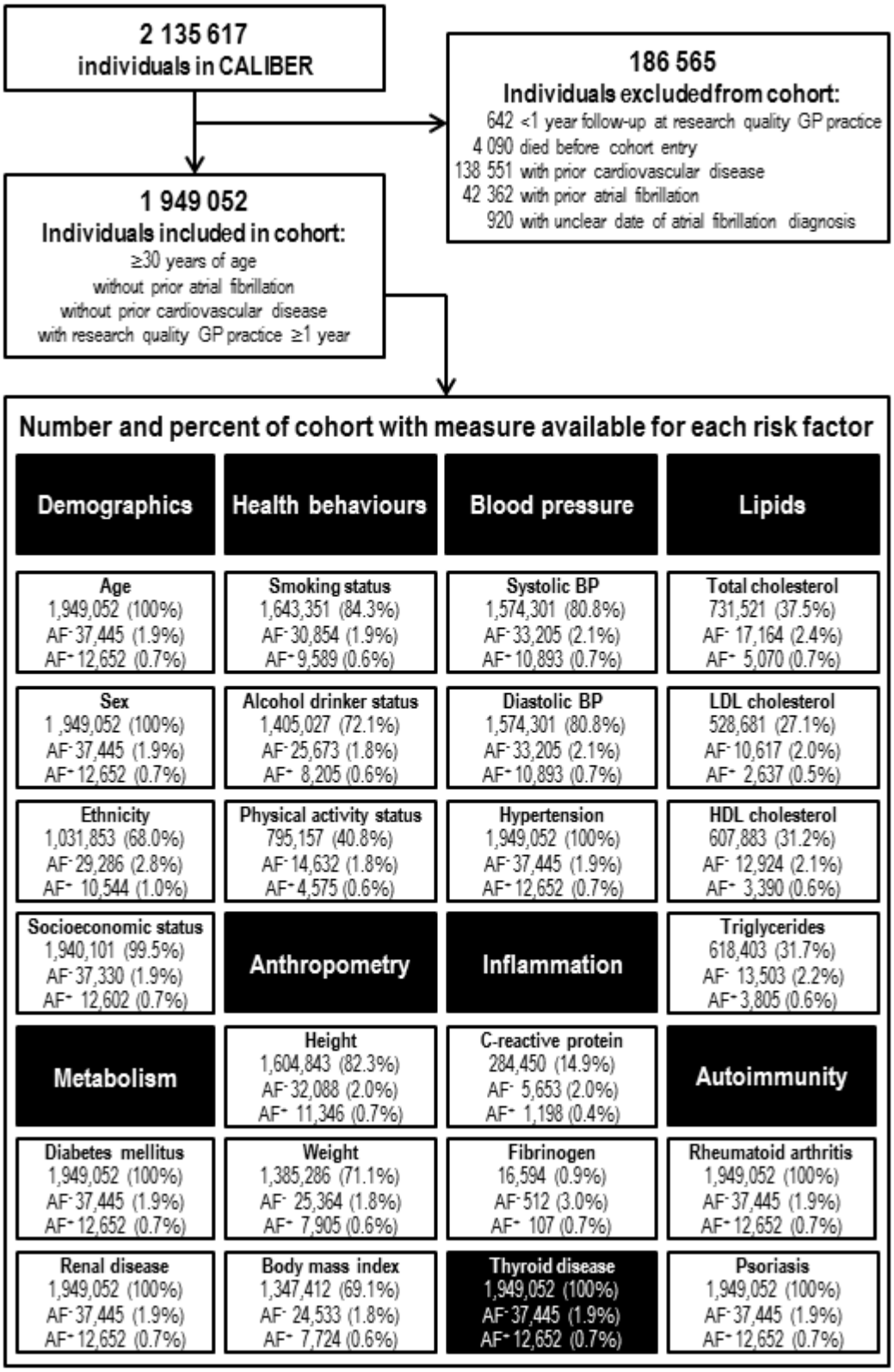
	AF⁻ HR [95%CI]	AF⁺ HR [95%CI]	AF HR [95%CI]
Thyroid disease			
Thyroid disease:			
Hypothyroidism	1.13 [1.06, 1.20]	1.33 [1.21, 1.47]	1.15 [1.09, 1.21]
Hyperthyroidism	1.30 [1.16, 1.46]	1.35 [1.10, 1.65]	1.30 [1.18, 1.44]
Uncertain type	1.11 [0.94, 1.30]	1.38 [1.07, 1.79]	1.15 [1.00, 1.32]
Autoimmune disease			
Rheumatoid arthritis	1.48 [1.36, 1.60]	1.87 [1.65, 2.12]	1.53 [1.42, 1.63]
Psoriasis	1.12 [1.05, 1.19]	1.17 [1.05, 1.31]	1.11 [1.04, 1.17]

Notes: continuous variables are analysed for strength of association per one standard deviation increase in value.

Abbreviations: AF^{+}/ AF^{-} - atrial fibrillation with/without intercurrent cardiovascular disease, HR [95%CI] – hazard ratio and 95% confidence interval, L/HDL – low/high density lipoprotein cholesterol, mmHG - millimetres of mercury, mmol/L - millimoles per Liter, mg/L - milligrams per litre, g/L - grams per litre, m – metres, kg – kilograms.

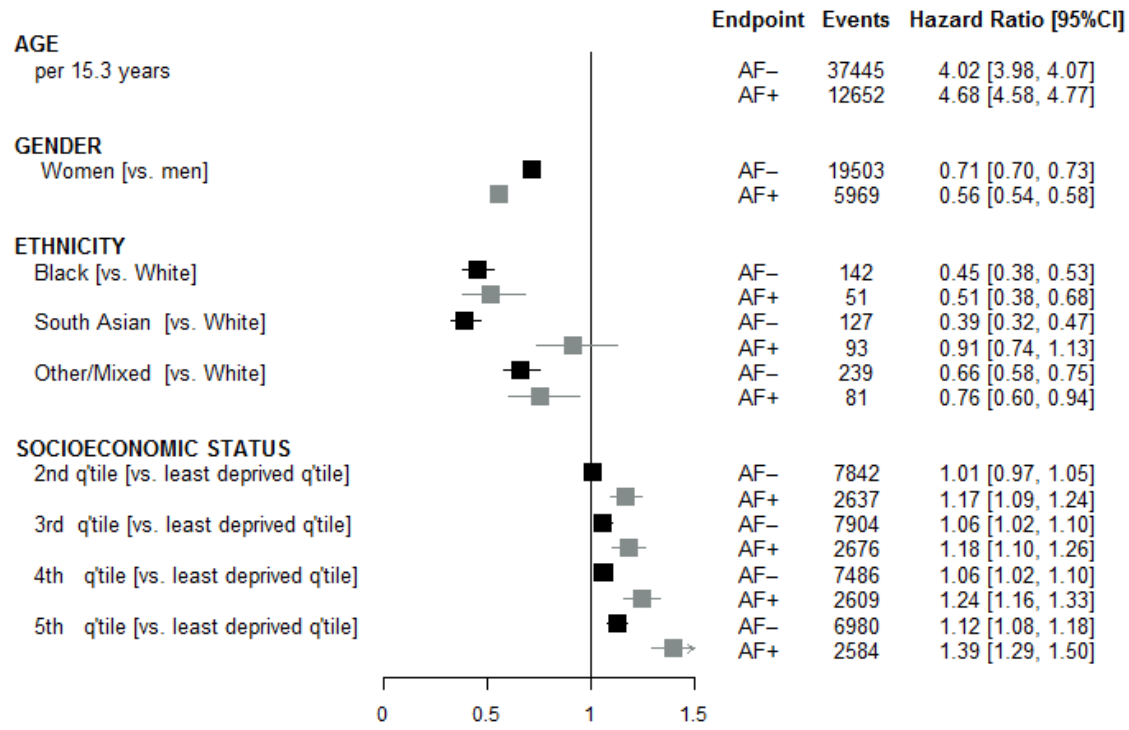
4.10 Chapter figures

Figure 4.1 Cohort flow diagram showing number and percentage of individuals with available data for 23 cardiovascular risk factors



Abbreviations: GP – general practitioner, BP – blood pressure, LDL – low density lipoprotein cholesterol, HDL – high density lipoprotein cholesterol, AF^+/ AF^- - atrial fibrillation with/without intercurrent cardiovascular disease.

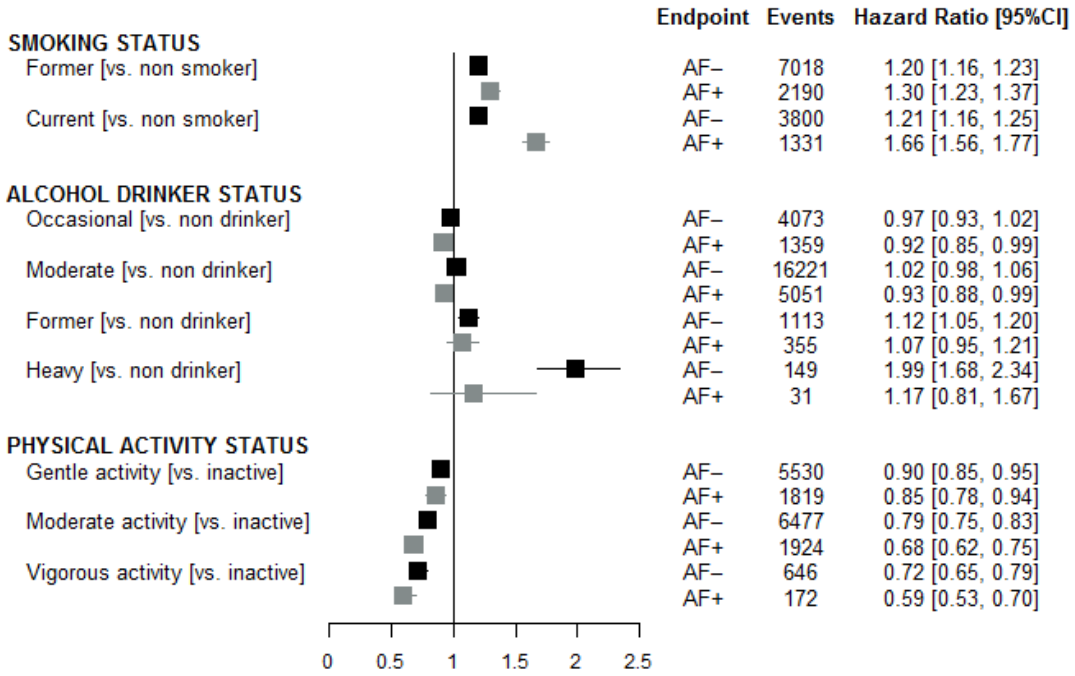
Figure 4.2 Age and sex adjusted hazard ratios and 95% confidence intervals for associations with incident atrial fibrillation with and without incurrent cardiovascular disease: age, sex, ethnicity and socio-economic status



Notes: plot for age is not shown as the extreme direct risk estimates fall beyond the scale of the x axis. Continuous variables are analysed for strength of association per one standard deviation increase in value.

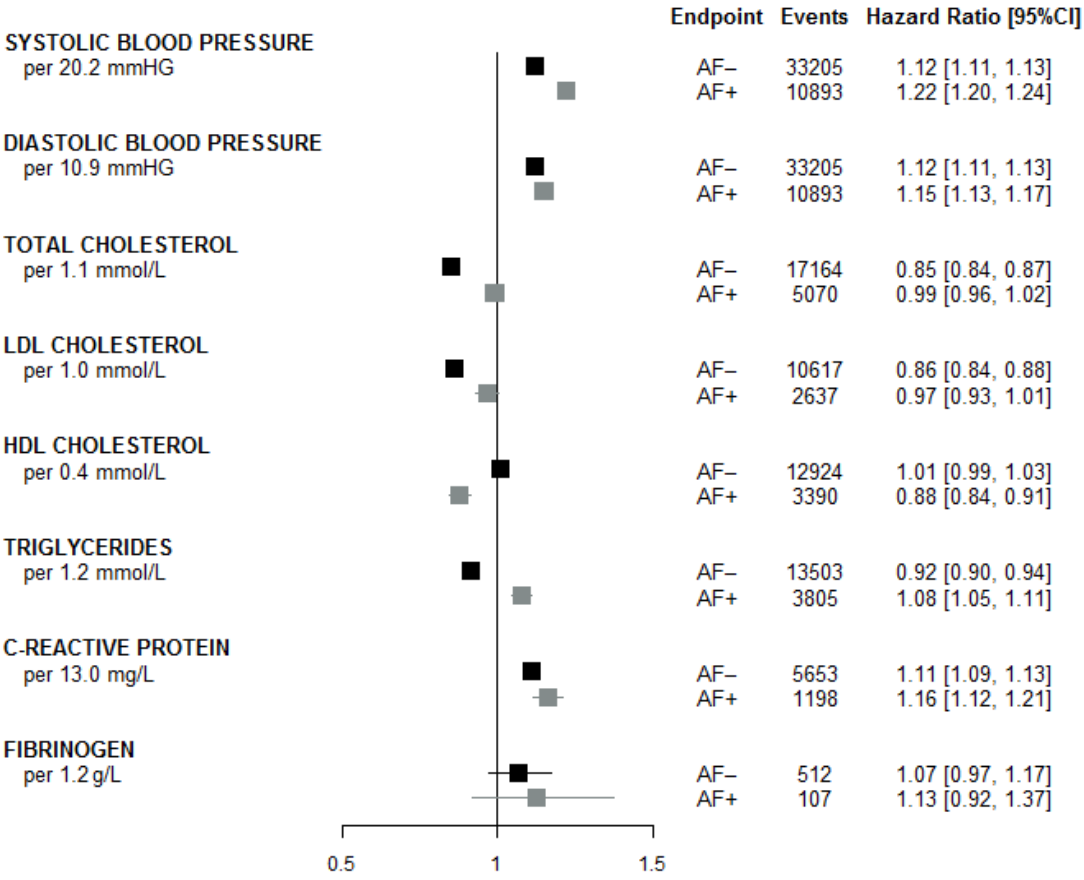
Abbreviations: AF^+/ AF^- - atrial fibrillation with/without intercurrent cardiovascular disease, [95%CI] – 95% confidence interval, q'tile – quintile.

Figure 4.3 Age and sex adjusted hazard ratios and 95% confidence intervals for associations with incident atrial fibrillation with and without incurrent cardiovascular disease: smoking, alcohol and physical activity



Abbreviations: AF^{+}/ AF^{-} - atrial fibrillation with/without intercurrent cardiovascular disease, [95%CI] – 95% confidence interval.

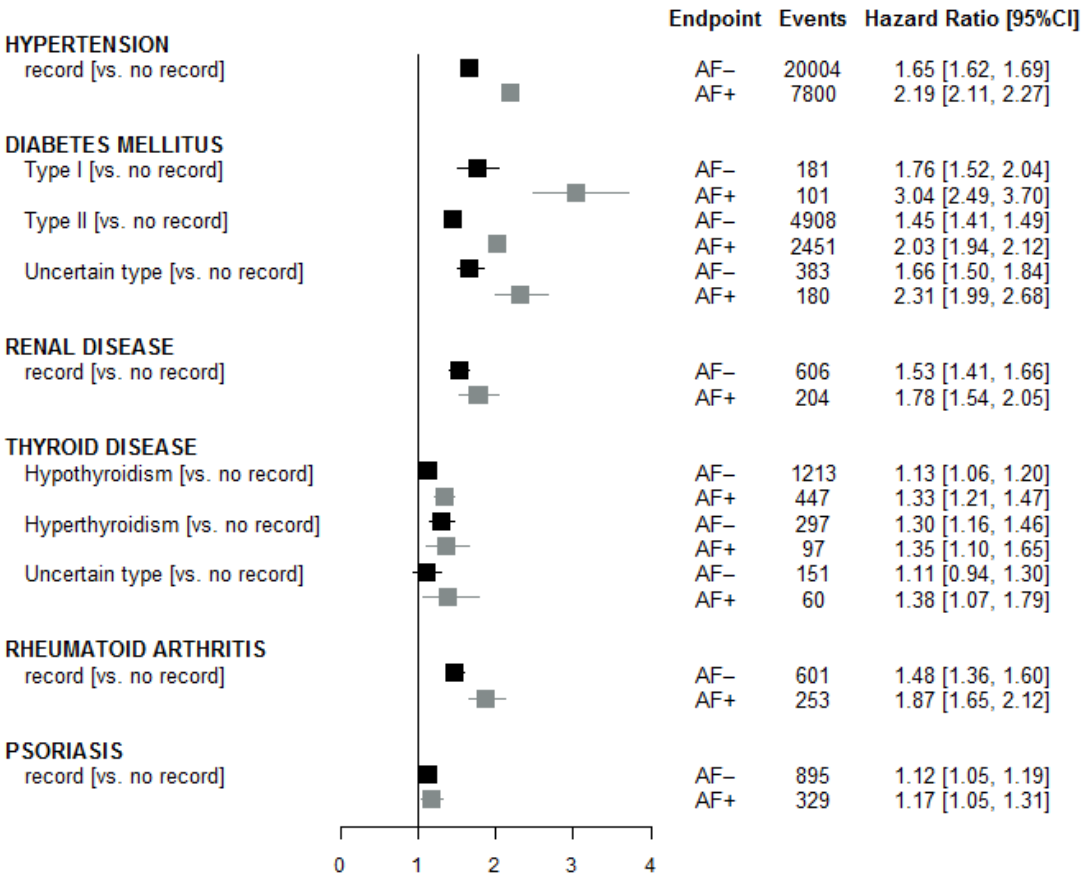
Figure 4.4 Age and sex adjusted hazard ratios and 95% confidence intervals for associations with incident atrial fibrillation with and without incurrent cardiovascular disease: systolic blood pressure, diastolic blood pressure, total cholesterol, low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, triglycerides, C-reactive protein and fibrinogen



Notes: continuous variables are analysed for strength of association per one standard deviation increase in value.

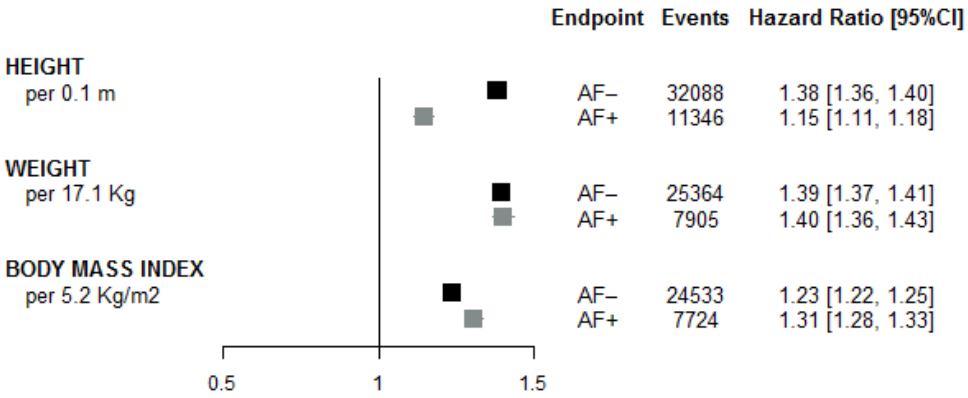
Abbreviations: AF^+/ AF^- - atrial fibrillation with/without intercurrent cardiovascular disease, [95%CI] – 95% confidence interval, L/HDL – low/high density lipoprotein cholesterol, mmHG - millimetres of mercury, mmol/L - millimoles per Liter, mg/L - milligrams per litre, g/L - grams per litre.

Figure 4.5 Age and sex adjusted hazard ratios and 95% confidence intervals for associations with incident atrial fibrillation with and without incurrent cardiovascular disease: hypertension, diabetes mellitus, renal disease, thyroid disease, rheumatoid arthritis and psoriasis diagnoses



Abbreviations: AF^{+}/ AF^{-} - atrial fibrillation with/without intercurrent cardiovascular disease, [95%CI] – 95% confidence interval.

Figure 4.6 Age and sex adjusted hazard ratios and 95% confidence intervals for associations with incident atrial fibrillation with and without incurrent cardiovascular disease: height, weight and body mass index



Notes: continuous variables are analysed for strength of association per one standard deviation increase in value.

Abbreviations: AF^{+}/ AF^{-} - atrial fibrillation with/without intercurrent cardiovascular disease, [95%CI] – 95% confidence interval, m – metres, kg – kilograms.

Chapter 5

Development of electronic health record definitions for AF subtypes relevant to the 2016 European Society of Cardiology guidelines

5.1 Chapter outline

This chapter shifts in focus from using electronic health records (EHRs) to obtain novel insights into atrial fibrillation (AF) risk factors (**chapter 4**) towards how EHRs can be used to investigate AF subtypes. Research into AF risk factors and AF subtypes is intimately related because primary preventions strategies for AF will only work given that the target of prevention is clearly defined. In a step forward, the 2016 updates to European Society of Cardiology (ESC) guidelines for the management of AF outlined a framework of new ideas on mechanisms and clinical distinctions of AF; albeit without any large-scale quantitative evidence to support its use.[3] Therefore, as I describe in this chapter, I investigated whether the new ESC definitions for AF can be operationalised in CALIBER records.[5] In addition to the seven distinctions suggested in the ESC guidelines, I also investigated the creation an initial case definition for 'AF secondary to respiratory disease' because of a growing evidence base in support of this as a potential AF mechanism.[222] I hypothesised that EHR definitions can be developed in order to identify a range of diverse AF subtypes; however in the absence of genomic information and ECG images it may be infeasible to create definitions for polygenic and focal AF. The ability to identify AF subtypes in CALIBER offers a potential setting in which to validate and refine understanding about them. While presenting the methods for this work, I also describe the general CALIBER approach which I followed in order to create robust definitions and algorithms for identifying disease cases in EHR.[49] Note that this chapter does not involve the analysis of any data, but rather reports the results of systematic searches and reviews of thousands of clinical codes in order to arrive at computable definitions for each AF subtype.

5.2 Abstract

Background Progress in identifying risk factors and designing primary prevention programmes for AF is held back by a limited understanding of the different ways that AF presents clinically. I investigated whether EHR compatible definitions could be derived for the AF subtypes described in recently updated ESC guidelines for the management of AF.

Methods In line with CALIBER guidance, I aimed to develop EHR definitions for eight AF subtypes: (1) structural, (2) focal, (3) polygenic, (4) postoperative, (5) valvular, (6) AF in athletes, (7) monogenic and (8) respiratory AF. I translated the ESC subtypes descriptions into computable definitions as accurately as possible by identifying applicable clinical codes and other information relevant to the diagnosis of cases**.** Applicable codes were identified from pre-existing code lists, new and updated code searches and by matching synonymous codes between the Read (primary care diagnoses and procedures), ICD-10 (secondary care diagnoses) and OPCS-4 (secondary care procedures) classification systems. Where applicable codes were unavailable I instead created strategies for inferred cases.

Results Overall a total of 2813 applicable clinical codes were identified. EHR definitions were set out for all eight AF subtypes based on code combinations and plausible inferences. With the exception of post-operative AF (based exclusively on OPCS-4 codes) and AF in athletes (based exclusively on Read codes), all other AF subtypes definitions can be derived, fully or at least in part, using solely ICD-10 codes, and are thus compatible with international EHR resources also using ICD-10 (e.g. in Denmark and Sweden). In the absence of family relationships or genomics data, the definition for polygenic AF was derived based on inferences of an early age of AF onset and AF not explained by any of the other subtypes.

Conclusions I created potentially workable EHR definitions for eight mechanistically diverse subtypes of AF. Further work is now needed to implement, improve and confirm the validity of these definitions.

5.3 Introduction

Progress in identifying risk factors and designing primary prevention programmes for AF is held back by a limited understanding of the different ways that AF presents clinically.[39] In a step forward, the 2016 updates to ESC guidelines for management of AF, outlined seven newly defined clinical distinctions: (1) AF secondary to structural heart disease, (2) focal AF, (3) polygenic AF, (4) postoperative AF, (5) AF in mitral stenosis or prosthetic heart valves (often referred to as 'valvular' AF), (6) AF in athletes, and (7) monogenic AF (**table 5.1**).[3] However, these definitions were derived based on expert consensus and remain unsupported by quantitative evidence.

EHR resources like CALIBER,[5] which are collected on large populations as part of routine clinical care, are a viable data source in which to systematically validate and refine understanding of AF subtypes.[39] Traditional epidemiological cohort studies (e.g. the Framingham Heart Study with 1544 incident cases AF accrued over 50 years of follow-up[191]) and snap-shot AF registries (e.g. the EuroHeart Survey of 5,333 AF patients enrolled from one of 182 hospitals in Europe in 2003 to 2004[199]) have been instrumental in studying the link between AF and subsequent stroke risk,[13 29] however often lack the large-scale population denominator needed to intricately investigate AF subtypes.[204] EHRs, as already described, are increasingly used in research[47] but as they are not collected for primary research purposes it can be challenging to identify disease cases which are rarely explained by a single clinical code.[223] Identifying AF cases in UK primary and secondary care linked EHRs, as studied by Morley and colleges, requires an algorithm incorporating 286 codes from four different coding systems as well as inferred diagnoses based on warfarin prescriptions in the absence of thromboembolic disease.[49]

No study to date has used EHRs to investigate validity of AF subtypes definitions, including whether computable definitions for the new ESC distinctions can feasibly be derived. The novel associations I found between cardiovascular risk factors and two separate AF endpoints (AF with and without intercurrent cardiovascular disease; as presented in **chapter 4**) nevertheless show that there is the potential to detect diverse clinical presentations of AF within EHRs. The aim of this study was therefore to create computable definitions (i.e. definitions that can be

computerised and coded to work within EHRs) for the newly defined clinical distinctions of AF as indicated in the latest ESC guidelines for management of AF.[3] Operationalising these can help progress AF research by offering transparent and reproducible EHR definitions on which future studies, both in the UK and internationally, can align which in turn may lead to new insights about risk factors and outcomes specific to each subtype.[39]

5.4 Methods

To create computable definitions for AF subtypes relevant to the 2016 ESC guidelines,[3] I followed the CALIBER approach to EHR algorithm development.[49] I will first describe the general approach before applying in relation to AF subtypes.

5.4.1 The CALIBER approach to EHR algorithm development

The CALIBER approach to EHR algorithm development is depicted in **figure 5.1** and involves the following five steps:

1. **Creating an initial case definition**
   i.e. setting out an initial approach to identifying disease cases in records

2. **Translating initial case definition into a computable definition**
   i.e. identifying applicable clinical codes and other information relevant to disease diagnosis

3. **Implementing computable definition in electronic health records**
   i.e. executing clinical codes and examining ability to detect disease cases

4. **Improving computable definition based on electronic health record insights**
   i.e. investigating whether case detection can be refined with code additions or omissions and incorporation of relevant supporting information

   Steps three and four are iterated until a final definition is reached

5. **Validating final definition against established clinical knowledge:**
   i.e. testing whether case definitions hold true against known clinical associations

5.4.2 Application of the CALIBER approach to EHR algorithm development for AF subtypes

Creating initial case definitions for AF subtypes

Initial case definitions were created with the view to map the ESC definitions as closely as possible to the available data. However as some of the subtype descriptions lack specificity (e.g. structural AF with limited detail on the structural heart diseases comprising the definition) or are less feasible to operationalise (e.g. polygenic AF without access to genomics data), I carried out

further reading of the literature (e.g. consensus documents, reviews and expert opinions)[6 222 224-227] and consulted local clinical expertise (Dr Amitava Banerjee) in order to attain greater clarity on how to target these in the records, making reasonable modifications where necessary. In addition to the seven distinctions suggested in the ESC guidelines, I also created an initial case definition for 'AF secondary to respiratory disease' because of a growing evidence base in support of this as a potential AF mechanism.[222] Further considerations for including 'AF secondary to respiratory disease' in this work are provided in **this chapter's discussion.**

Translating initial case definitions into computable definitions for AF subtypes

Initial case definitions were then translated into computable definitions using three methods to identify relevant clinical codes and supporting information. First of all, I sourced applicable pre-existing code lists used in prior CALIBER and CPRD studies. Second, I ran new and updated code searches up to and including the year 2016 using the CPRD code browser software which returns Read codes based on a key word search (as shown in **figure 5.2**). And third, I performed code matching using the CALIBERcodelists package available in statistical software package R, which matches synonymous codes between the Read, ICD 10 and OPCS4 classification systems (as shown in **figure 5.3**). All identified codes were then reviewed for inclusion in the subtype definitions and independently verified by a clinical expert (Dr Amitava Banerjee).

Implementing, improving, and validating

In **chapter 6**, which follows, I describe and apply the next steps of EHR algorithm development which involves iteratively implementing and improving the computable definition before validating the final definition against established clinical knowledge. These three steps are more time and resource intensive than creating initial case definitions and translating into code lists. I therefore selected only one AF subtype, valvular AF, to take forward for further development. The rationale for selecting valvular AF over other subtypes reflects uncertainties in the valvular heart diseases comprising the definition,[228] which can be helped with insights drawn from EHRs, as well as there being established clinical knowledge of a higher associated risk of subsequent stroke and thromboembolism,[6] which can be used as a point of validation.

5.5 Results

Table 5.2 shows the resultant initial case definitions for eight subtypes of AF (i.e. seven ESC definitions plus AF secondary to respiratory disease), which map to the guidelines as far as feasibility possible and include reasonable modifications where necessary. The process of translating initial case definitions into computable definitions for AF yielded a combined total of 2813 clinical codes. These are summarised for each subtype below with full code lists available in **tables S5.1** to **S5.7** of **appendix**.

- **AF secondary to structural heart disease**

  Based on pre-existing code lists for heart failure,[209] hypertension[148], congenital heart malformations,[229] cardiomyopathies,[230] valvular heart diseases,[7] and post-myocardial infarction ruptures/defects[213] and new and updated code searches and matches, I iden-

tified a total of 912 relevant codes: 745 (81.7%) from Read, 78 (8.6%) from ICD-10 and 89 (9.8%) from OPCS-4 (**table S5.1** of **appendix**). Example coding:

Read codes

G5yy900	Left ventricular systolic dysfunction
G5yyA00	Left ventricular diastolic dysfunction
G580.00	Congestive heart failure
G20..00	Essential hypertension
G5y3411	Left ventricular hypertrophy
P54..00	Ventricular septal defect [congenital]
P55..00	Ostium secundum atrial septal defect [congenital]
G551.00	Hypertrophic obstructive cardiomyopathy
790G200	Percutaneous occlusion of left atrial appendage
G341.00	Aneurysm of heart
G30..13	Cardiac rupture following myocardial infarction (MI)
G361.00	Atrial septal defect/curr comp folow acut myocardal infarct

ICD-10 codes

I50	Heart failure
I10	Essential (primary) hypertension
Q21	Congenital malformations of cardiac septa
I421	Obstructive hypertrophic cardiomyopathy
I253	Aneurysm of heart
I231	Atrial septal defect as current complication following acute myocardial infarction
I232	Ventricular septal defect as current complication following acute myocardial infarction

OPCS-4 codes

K22.3	Exclusion of left atrial appendage NEC
K243	Repair of right ventricular aneurysm
K244	Repair of left ventricular aneurysm

- **Focal AF**

Based on pre-existing code lists for paroxysmal AF,[49] AF symptoms (chest tightness,[231] sleeping difficulties,[232] and psychosocial distress[212]) and new and updated code searches and matches, I identified a total of 284 relevant codes: 268 (94.4%) from Read and 16 (5.6%) from ICD-10 (**table S5.2 of appendix**). Example coding:

Read codes

G573200	Paroxysmal atrial fibrillation
168..12	Lethargy - symptom
181..00	Palpitations
173..12	Dyspnoea - symptom
R065800	[D]Chest tightness
1B1B.11	C/O - insomnia

E200.00	Anxiety states
3264.00	ECG: atrial ectopics
G570000	Paroxysmal atrial tachycardia

ICD-10 codes

R53	Malaise and fatigue
R002	Palpitations
R060	Dyspnoea
R074	Chest pain, unspecified
F510	Nonorganic insomnia
F41	Other anxiety disorders
I491	Atrial premature depolarization
I471	Supraventricular tachycardia

- **Polygenic AF**

  No relevant codes were identified for polygenic AF because the initial case definition is based wholly upon inferences (i.e. polygenic AF potentially can be inferred in individuals with early onset AF,[233] but in whom AF cannot be explained by any of the other clinical distinctions).

- **Post-operative AF**

  Based on the entire OPCS-4 classification system for operations and procedures, I identified a total of 1402 relevant (top level) codes (**table S5.3 of appendix**).

- **'Valvular AF'** *(modified from AF in patients with mitral stenosis or prosthetic valves)*

  Based on pre-existing code lists for valvular heart diseases,[7] and AF-related procedures (available on the CALIBER portal) and new and updated code searches and matches, I identified a total of 370 relevant codes: 235 (63.5%) from Read, 49 (13.2%) from ICD-10 and 86 (23.42%) from OPCS-4 (**table S5.4 of appendix**). Example coding:

Read codes

G540.16	Mitral regurgitation
G110.00	Mitral stenosis
G11..11	Rheumatic mitral valve disease
G11z.00	Mitral valve disease NOS
7910200	Prosthetic replacement of mitral valve
7910000	Allograft replacement of mitral valve
7910100	Xenograft replacement of mitral valve
7910.12	Replacement of mitral valve
7910.00	Plastic repair of mitral valve
7916000	Open mitral valvotomy
7917000	Closed mitral valvotomy
7918000	Annuloplasty of mitral valve

ICD-10 codes

I05	Rheumatic mitral valve diseases
I050	Mitral stenosis
I051	Rheumatic mitral insufficiency
I052	Mitral stenosis with insufficiency
I058	Other mitral valve diseases
I059	Mitral valve disease, unspecified
I34	Nonrheumatic mitral valve disorders
I340	Mitral (valve) insufficiency
I341	Mitral (valve) prolapse
I342	Nonrheumatic mitral (valve) stenosis
I348	Other nonrheumatic mitral valve disorders
I349	Nonrheumatic mitral valve disorder, unspecified

OPCS-4 codes

K253	Prosthetic replacement of mitral valve
K251	Allograft replacement of mitral valve
K252	Xenograft replacement of mitral valve
K254	Replacement of mitral valve NEC
K25	Plastic repair of mitral valve
K311	Open mitral valvotomy
K321	Closed mitral valvotomy
K341	Annuloplasty of mitral valve

- **AF in athletes**

  Based on new code searches focussed on the occupational codes available within the Read classification system, I identified a total of 27 relevant codes (**table S5.5 of appendix**). Example coding:

Read codes

1386.00	Competitive athlete
04A3.00	Jockey
68L1.11	Keen sportsman
04A2.00	Prof. association footballer
04AB.00	Professional boxer
04A6.11	Professional runner
04AZ.00	Professional sport occup. NOS
04A..00	Professional sport occupations
04A6.00	Professional sportsman
04A..13	Professional sportsmen
04A4.00	Racing car driver
04A5.00	Racing motor cyclist
04AA.00	Sport trainee

68L1.00 Sportsman

- **'AF secondary to inherited rhythm disorders'** *(modified from monogenic AF)*
  Based on new code searches and matches, I identified a total of 8 (75%) relevant codes: 6 from Read and 2 (25%) from ICD-10 (**table S5.6 of appendix**). Example coding:

  Read codes

  G56y500 Long Q-T syndrome
  G57y200 Brugada syndrome
  G567400 Wolff-Parkinson-White syndrome
  32K3.00 ECG: Q-T interval prolonged
  32K2.00 ECG: Q-T interval abnormal
  32K4.00 ECG: Q-T interval shortened

  ICD-10 codes

  I498 Other specified cardiac arrhythmias: Brugada syndrome, Long QT syndrome
  I456 Pre-excitation syndrome: Wolff-Parkinson-White syndrome

- **'AF secondary to respiratory disease'** *(included in addition to ESC subtypes)*
  Based on pre-existing code lists for chronic obstructive pulmonary disease,[234] and new and updated code searches and matches, I identified a total of 180 relevant codes: 167 from Read and from 13 (%) ICD-10 (**table S5.7 of appendix**). Example coding:

  Read codes

  H3...00 Chronic obstructive pulmonary disease
  H5B0.00 Obstructive sleep apnoea
  G410.00 Primary pulmonary hypertension
  G41y000 Secondary pulmonary hypertension
  G411.00 Kyphoscoliotic heart disease

  ICD-10 codes

  J440 Chronic obstructive pulmonary disease with acute lower respiratory infection
  J441 Chronic obstructive pulmonary disease with acute exacerbation, unspecified
  G473 Sleep apnoea
  I270 Primary pulmonary hypertension
  I271 Kyphoscoliotic heart disease
  I272 Other secondary pulmonary hypertension

A summary of how the computable AF subtypes definitions utilise data across the linked EHR sources is provided in **table 5.3**.

5.6 Discussion

Overview of key findings

In this study I investigated the feasibility of using EHRs to identify a range of AF subtypes which were recently defined in the 2016 updates to ESC guidelines for the management of AF.[3] I

found a total of 2813 relevant clinical codes from the Read, ICD-10 and OPCS-4 classification systems and defined rules for combining them, together with plausible inferences, in order detect eight mechanistically diverse AF types. These initial definitions confirm that EHR resources such as CALIBER[5] hold important insights into disease processes and offer a foundation for future research. More detailed work is now needed to implement, improve and confirm the strength and clinical validity of these AF subtype definitions before they can be reliably used in research.

Code use and clinical validity

Although I have shown it is feasibly possible to identify eight mechanistically diverse subtypes of AF in EHRs, a crucial next step is to confirm whether these definitions are usable (i.e. do actually detect cases) and are clinically valid (i.e. make clinical sense). This is important because even though clinical codes exist it may not mean they are used in clinical practice. As an example, Morley and colleagues identified available codes for pulse palpation (a guideline recommended method of screening for AF [3 19 22]), hypothesising that these codes could be used to improve AF case ascertainment. While it was shown that over 70% of individuals with a record of pulse palpation prior to AF diagnosis had an irregular pulse (as would be expected); pulse palpation was recorded in less than 2% of all cases and thus of little value to case ascertainment.[49] Robust definitions are also necessary in order to accurately calculate disease risk or rate estimates. Newly defined EHR definitions therefore require validation against established clinical knowledge in order to gain reassurance about their ability to derive novel clinical insights going forward. Implementing, improving and confirming the clinical validity of EHR definitions is an involved process as I will demonstrate in the next chapter (**chapter 6**) in relation to 'valvular AF'. The rationale for selecting valvular AF over other subtypes reflects uncertainties in the valvular heart diseases comprising the definition,[228] which can be helped with insights drawn from EHRs, as well as there being established clinical knowledge of a higher associated risk of subsequent stroke and thromboembolism,[6] which can be used as a point of validation. As described in this chapter's introduction the clinical distinctions of AF set out by the ESC reflect contemporary expert opinion but are currently unsupported by any large-scale evidence. Although not explicitly classified as an AF distinction, the ESC guidelines acknowledge there is a growing evidence base for a possible respiratory mechanism,[222] suggesting that chronic obstructive pulmonary disease, sleep apnoea and other respiratory diseases should be treated in order to improve AF outcomes.[3] It was for this reason, as well as the fact that an earlier CALIBER code list for chronic obstructive pulmonary disease was available,[234] that I also developed an EHR definition for AF secondary to respiratory disorders. It may be that AF secondary to respiratory disorders is not sufficiently distinct from other AF subtypes to stand alone. Individuals classified as with AF secondary to respiratory disorders (e.g. COPD) are likely to overlap with individuals classified as with AF secondary to structural heart disease as these conditions frequently coexist. The extent to which the eight AF subtypes considered here are clinically distinct will be more greatly understood on implementation of the EHR code lists I have developed.

Potential for internationalisation

Looking beyond CALIBER, the strength of the EHR definitions provided here will be further judged on compatibility and external validity in international systems. Particularly for AF subtypes representing very small proportions of the overall AF population (e.g. AF secondary to inherited rhythm disorders estimated to account for less than 5% of all AF cases[3]) combining data from the UK together with data collected in other countries will be necessary in order to amass large enough numbers for meaningful investigations. Denmark[42] and Sweden[43] represent important targets for replication of this work, given that both countries have nation-wide systems for the collection and linkage of EHRs as well as elements in common with CALIBER (i.e. secondary care diagnoses coded with ICD-10 and linkage to prescriptions and mortality data). As **table 5.3** (which summarises how Read, ICD-10 and OPCS-4 codes combine to form the basis of the EHR definitions) shows at least six out of eight subtypes are compatible with the Danish and Swedish records given that they can be implemented (fully or at least partially) based on ICD-10 codes alone. However, the lack of large-scale linkage to primary care records in Denmark and Sweden, which in the UK offers advantages in terms of data on numerical clinical values (rather than just coded data) and information on professional occupations, means that it may be more difficult to detect some subtype distinctions such as AF in athletes. The Danish records, on the other hand, have a particular advantage for studying polygenic AF in that complete information on parent and sibling relationships have been collected since 1930 and 1942 respectively.[233] These data were recently used to show that familial AF presents at a significantly younger age of onset as compared to non-familial AF (e.g. median [interquartile range] of 50 [43, 54] years vs. 77 [67, 84] years) although with no differences in the subsequent risk of death and thromboembolism. Of course, other international systems may not have such compatible data and a data harmonisation process (i.e. understanding where similarities and differences lie and ultimately finding a common denominator) will be needed to translate CALIBER definitions into definitions that fit with local coding classifications. Understanding how EHR data sources across Europe compare and can be combined in order to drive progress in cardiovascular research is the subject the recently launched Innovative Medicines Initiative BigData@Heart project.[39] The present analysis therefore represents important groundwork in relation to big data approaches to the study of AF.

Strengths and limitations

The strengths of the CALIBER resource, in terms of breadth of clinical variables and opportunities for clinical research, is further demonstrated here by showing that eight mechanistically diverse subtypes of AF can feasibly be identified in records. Limitations relate to the fact that these initial EHR definitions now need thorough testing in terms of actual ability to detect AF cases and determining whether detected cases hold clinically valid insights both internally in the CALIBER dataset and externally in other EHR data sources. Missing/misclassified/ and limited resolution of clinical codes in EHRs may mean that some subtypes may not be reliably identifiable in the data. Clearly, some of the definitions set out here are more robust than others. As always, lack of linkage to ECG imaging and genomics data limits the extent to which CALIBER can be used to characterise heart rhythm patterns (as would be ideal for identifying focal AF) or

inherited conditions (as would be ideal for identifying polygenic AF). Creative solutions are therefore needed to overcome shortcomings such as these. Combining the knowledge that familial AF presents at considerably earlier age of onset[233] with the knowledge that AF is not explained by any of the other subtype distinctions, is an example of a solution I created for inferring polygenic AF in the absence of genomics data. The validity of this definition however needs to be cautiously explored, and, ideally, validated in an EHR data source with available genomics data such as the Kaiser Permanente Research Bank combining biological samples data with health insurance claims in the United States.[235] Another limitation to reflect upon is that EHR definitions are time-limited and require updating as new clinical codes come into existence. While the ICD coding system, now in its tenth revision, updates only periodically (i.e. ICD-10 introduced in 1992 and ICD-11 expected in 2018), the Read classification system[163] used in UK primary care offers the flexibility for primary care providers to add new codes to the classification. Therefore, even though pre-existing code lists (e.g. for heart failure[209] and chronic obstructive pulmonary disease[234]) were available, I updated the searches up to and including the year 2016 in order to ensure code lists were as up to date as possible and can be applied in more recent data sets than the 2010 extract I have had access to within the timeframe of this PhD. The CALIBER portal, providing existing EHR disease definitions and algorithms, details the dates upon which code lists were compiled, as well as code list authors which helps to decide whether definitions require updating and facilitates the process. Future research and clinical practice will benefit from higher resolution definitions for the coding and recording of diseases and greater interoperability between healthcare systems. Lastly, the study design I selected to develop EHR definitions for AF subtypes reflects traditional hypothesis driven research, starting with theory about what the subtype looks like and manually sorting and selecting through thousands of clinical codes. While code lists were compiled using comprehensive and systematic methods (i.e. sourcing pre-existing code lists used in prior studies, running new and updated code searches and matches and independent verification from a practicing cardiologist), it is possible that some codes may have been missed completely or may not be relevant. An alternative study design would have been to use a novel machine learning approach, whereby the discovery of disease subtypes is guided by correlations within the data. This more automated approach may remove potential biases due to human error, however could also suggest correlations that are not clinically meaningful.

5.7 Conclusion

I created potentially workable EHR definitions for eight mechanistically diverse subtypes of AF. Further work is now needed to implement, improve and confirm the validity of these definitions.

5.8 Chapter summary

To summarise, in this chapter I explored the feasibility of using EHR data to investigate the new AF subtype distinctions suggested in recent 2016 updates to ESC guidelines for management of AF.[3] I sourced applicable clinical codes to potentially identify eight AF subtypes in the records, these were: (1) structural, (2) focal, (3) polygenic, (4) postoperative, (5) valvular, (6) AF in athletes, (7) monogenic and (8) respiratory AF. Where applicable clinical codes were unavaila-

ble I instead created strategies for inferring cases, although the validity of these inferences requires rigorous testing. In **chapter 6**, which follows, I take forward the EHR definition for valvular AF set out here for further development. The rationale for selecting valvular AF over other subtypes reflects uncertainties in the valvular heart diseases comprising the definition,[228] which can be helped with insights drawn from EHRs, as well as there being established clinical knowledge of a higher associated risk of subsequent stroke and thromboembolism,[6] which can be used as a point of validation.

5.9 Chapter tables

Table 5.1 Seven clinical subtypes of atrial fibrillation newly outlined in 2016 European Society of Cardiology guidelines for the management of atrial fibrillation.

AF type	Clinical presentation
AF secondary to structural heart disease	AF in patients with left ventricular systolic or diastolic dysfunction, long-standing hypertension with left ventricular hypertrophy, and/or other structural heart disease. The onset of AF in these patients is a common cause of hospitalization and a predictor of poor outcome.
Focal AF	Patients with repetitive atrial runs and frequent, short episodes of paroxysmal atrial fibrillation. Often highly symptomatic, younger patients with distinguishable atrial waves (coarse AF), atrial ectopy, and/ or atrial tachycardia deteriorating in AF.
Polygenic AF	AF in carriers of common gene variants that have been associated with early onset AF.
Post-operative AF	New onset of AF (usually self-terminating) after major (typically cardiac) surgery in patients who were in sinus rhythm before surgery and had no prior history of AF.
AF in patients with mitral stenosis or prosthetic heart valves	AF in patients with mitral stenosis, after mitral valve surgery and in some cases other valvular disease.
AF in athletes	Usually paroxysmal, related to duration and intensity of training.
Monogenic AF	AF in patients with inherited cardiomyopathies, including channelopathies.

Notes: table recreated from (table 6) Kirchhof P, Benussi S, Kotecha D, Ahlsson A, Atar D, Casadei B, Castella M, Diener HC, Heidbuchel H, Hendriks J, Hindricks G, Manolis AS, Oldgren J, Popescu BA, Schotten U, Van Putte B, Vardas P, Agewall S, Camm J, Baron Esquivias G, Budts W, Carerj S, Casselman F, Coca A, De Caterina R, Deftereos S, Dobrev D, Ferro JM, Filippatos G, Fitzsimons D, Gorenek B, Guenoun M, Hohnloser SH, Kolh P, Lip GY, Manolis A, McMurray J, Ponikowski P, Rosenhek R, Ruschitzka F, Savelieva I, Sharma S, Suwalski P, Tamargo JL, Taylor CJ, Van Gelder IC, Voors AA, Windecker S, Zamorano JL, Zeppenfeld K. 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS. Eur Heart J. 2016 Oct 7;37(38):2893-2962. Epub 2016 Aug 27.

Abbreviations: AF – atrial fibrillation

Table 5.2 Initial case definitions for eight subtypes of atrial fibrillation relevant to the 2016 European Society of cardiology guidelines

ESC definitions	Initial case definitions for identification in EHRs
AF secondary to structural heart disease	
AF in patients with left ventricular systolic or diastolic dysfunction, long-standing hypertension with left ventricular hypertrophy, and/or other structural heart disease. The onset of AF in these patients is a common cause of hospitalization and a predictor of poor outcome.	As the description provided lacks clarity on the specific conditions constituting 'structural heart diseases', I unpacked and expanded the definition into the following six subcategories, identified as priority areas based on other relevant literature: [224-226] (1) LV systolic/diastolic dysfunction/ heart failure (2) long-standing hypertension with LV hypertrophy (3) congenital heart malformations (4) cardiomyopathies (5) valvular heart diseases (6) other structural heart diseases
Focal AF	
Patients with repetitive atrial runs and frequent, short episodes of paroxysmal atrial fibrillation. Often highly symptomatic, younger patients with distinguishable atrial waves (coarse AF), atrial ectopy, and/ or atrial tachycardia deteriorating in AF.	In the absence of ECG imaging data to characterise heart rhythm patterns, I instead created an initial case definition for focal AF based on the following code combinations and inferences: (1) early onset AF (2) codes for paroxysmal AF, AF symptoms (e.g. lethargy, palpitations, dyspnoea, chest tightness, sleeping difficulties, psychosocial distress), atrial ectopy, and/or atrial tachycardia (3) AF not explained by any other subtypes
Polygenic AF	
AF in carriers of common gene variants that have been associated with early onset AF.	In the absence of family relationships or genomics data to determine if AF was inherited, I instead created an initial case definition for polygenic AF based on the following combined inferences: (1) early onset AF; as familial AF is known to present at a younger age of onset[233] (2) AF not explained by any other subtypes
Post-operative AF	
New onset of AF (usually self-terminating) after major (typically cardiac) surgery in patients who were in sinus rhythm before surgery and had no prior history of AF.	As the description provided lacks clarity on the specific conditions constituting post-operative AF and the qualifying time frame for diagnosis, I unpacked and expanded the definition into two subcategories, focusing on AF diagnoses made within 30 days[227 236] after surgery: (1) cardiac procedures (2) non-cardiac procedures
AF in patients with mitral stenosis or prosthetic valves	***'valvular AF'***
AF in patients with mitral stenosis, after mitral valve surgery and in some cases other valvular disease.	The separation of individuals with AF and mitral stenosis or prosthetic heart valves is due to a higher subsequent stroke risk, but uncertainty remains as to whether other valvular heart diseases should also be included in this subtype distinction.[6] I therefore cast a broad initial case definition to capture all adult valvular heart conditions with the view to examine which ones warrant being kept clinically distinct and which ones should fall in the AF secondary to structural heart disease category above. Congenital malformations were excluded in order focus on acquired disease: (1) stenosis, regurgitation and other and unspecified disorders of mitral, aortic, tricuspid, pulmonary and unspecified heart valves (2) prosthetic, bioprosthetic and unspecified heart valve replacements (3) heart valve repairs and operations

AF in athletes	
Usually paroxysmal, related to duration and intensity of training.	I created an initial case definition for AF in athletes based on the following code combinations and inferences: (1) codes for competitive athletes (2) codes for sports professions (inferring high level or long duration of sports participation)
Monogenic AF	***'AF secondary to inherited rhythm disorders'***
AF in patients with inherited cardiomyopathies, including channelopathies.	As cardiomyopathies is captured in the AF secondary to structural heart disease category above, I refocussed this subtype distinction towards 'AF secondary to inherited rhythm disorders', and in particular the following conditions: (1) Long QT syndrome (2) Brugada syndrome (3) Short QT syndrome (4) Wolff-Parkingson-White syndrome
	'AF secondary to respiratory disease'
	Although not technically a clinical subtype of AF put forward by the ESC, there is growing evidence and mechanistic relevance to support investigation of AF secondary to respiratory disease. I therefore created an initial case definition based on the following disorders: (1) Chronic obstructive pulmonary disorder (2) Sleep apnoea (3) Pulmonary hypertension

Abbreviations: ESC – European Society of Cardiology, EHR – electronic health records, AF – atrial fibrillation, LV – left ventricular, ECG – electrocardiogram, QT – refers to QT interval measured on an ECG (see **section 1.2.3** for explanation).

Table 5.3 Summary of how codes from the Read (primary care diagnoses and procedures), ICD-10 (secondary care diagnoses) and OCPS-4 (secondary care procedures) classification systems combine to form electronic health record definitions for eight atrial fibrillation subtypes

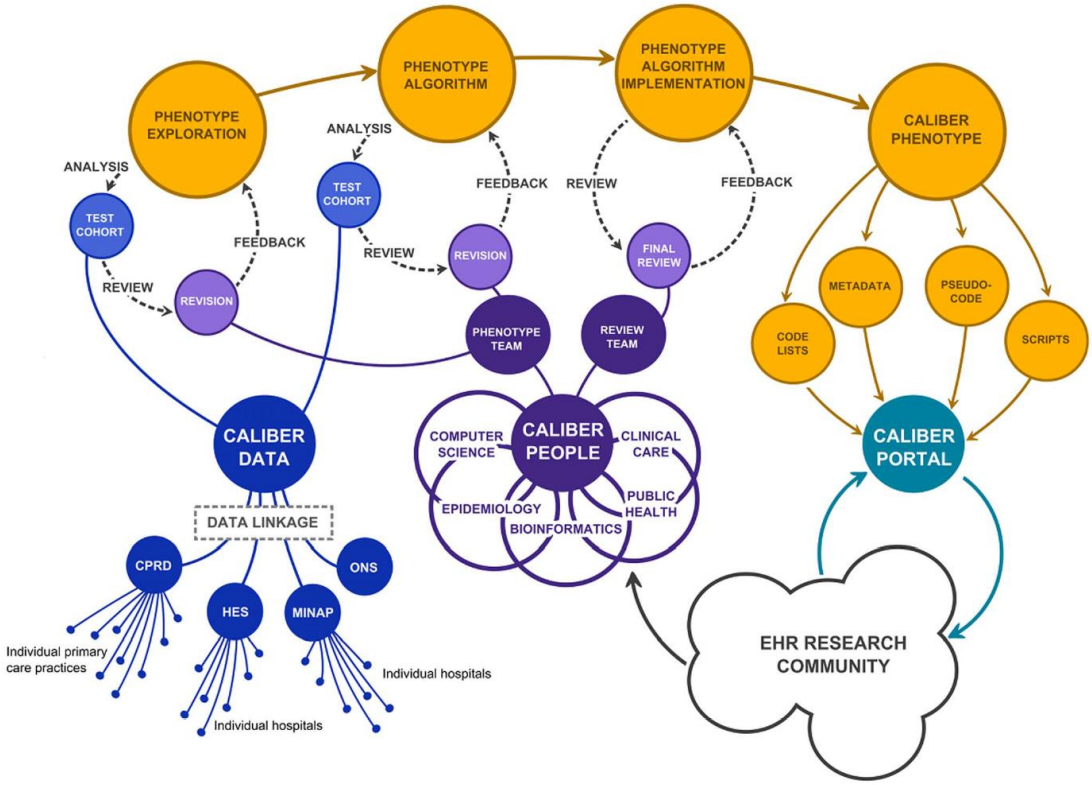
	Coding classification		
AF secondary to structural heart disease	**Read**	**ICD-10**	**OPCS-4**
LV systolic/diastolic dysfunction	✓		
Heart failure	✓	✓	
Hypertension	✓	✓	
LV hypertrophy	✓		
Congenital heart malformations	✓	✓	✓
Cardiomyopathies	✓	✓	
Valvular heart diseases	✓	✓	
Other: left atrial appendage (occlusion)	✓		✓
Other: heart aneurysm	✓	✓	✓
Other: post myocardial infarction ruptures/defects	✓	✓	
Focal AF			
Paroxysmal AF	✓	✓	
AF symptoms: lethargy	✓	✓	
AF symptoms: palpitations	✓	✓	
AF symptoms: dyspnoea	✓	✓	
AF symptoms: chest tightness	✓	✓	
AF symptoms: sleeping difficulties	✓	✓	
AF symptoms: psychosocial distress	✓	✓	
Atrial ectopy	✓	✓	
Atrial tachycardia	✓	✓	
Polygenic AF			
	NA	NA	NA
Post-operative AF			
cardiac procedures			✓
non-cardiac procedures			✓
Valvular AF			
valvular heart diseases	✓	✓	
valve replacements	✓		✓
valve repairs and operations	✓		✓
AF in athletes			
Competitive athletes and sports professionals	✓		
Other sports-related occupations	✓		
AF secondary to inherited rhythm disorders			
Long QT syndrome	✓	✓	
Brugada syndrome	✓	✓	
Short QT syndrome	✓		
Wolff-Parkingson-White syndrome	✓	✓	

AF secondary to respiratory disease	✓	✓	
Chronic obstructive pulmonary disorder	✓	✓	
Sleep apnoea	✓	✓	
Pulmonary hypertension	✓	✓	

Abbreviations: ICD-10 – International Statistical Classification of Diseases and Health-Related Problems, Tenth Edition, OPCS-4 – Office of Population Censuses and Surveys' Classification of Interventions and Procedures version 4, LV – left ventricular, AF – atrial fibrillation, NA – none available, QT – refers to QT interval measured on an electrocardiogram (see **section 1.2.3** for explanation).

5.10 Chapter figures

Figure 5.1 Illustrative diagram of the CALIBER approach to electronic health record algorithm development showing iterative cycles between development, implementation and validation



Notes: figure from Morley KI, Wallace J, Denaxas SC, Hunter RJ, Patel RS, Perel P, et al. (2014) Defining Disease Phenotypes Using National Linked Electronic Health Records: A Case Study of Atrial Fibrillation. PLoS ONE9(11): e110900. https://doi.org/10.1371/journal.pone.0110900

Figure 5.2 Screenshot of CPRD code browser software with example key word search for codes relating to "atrial fibrillation"

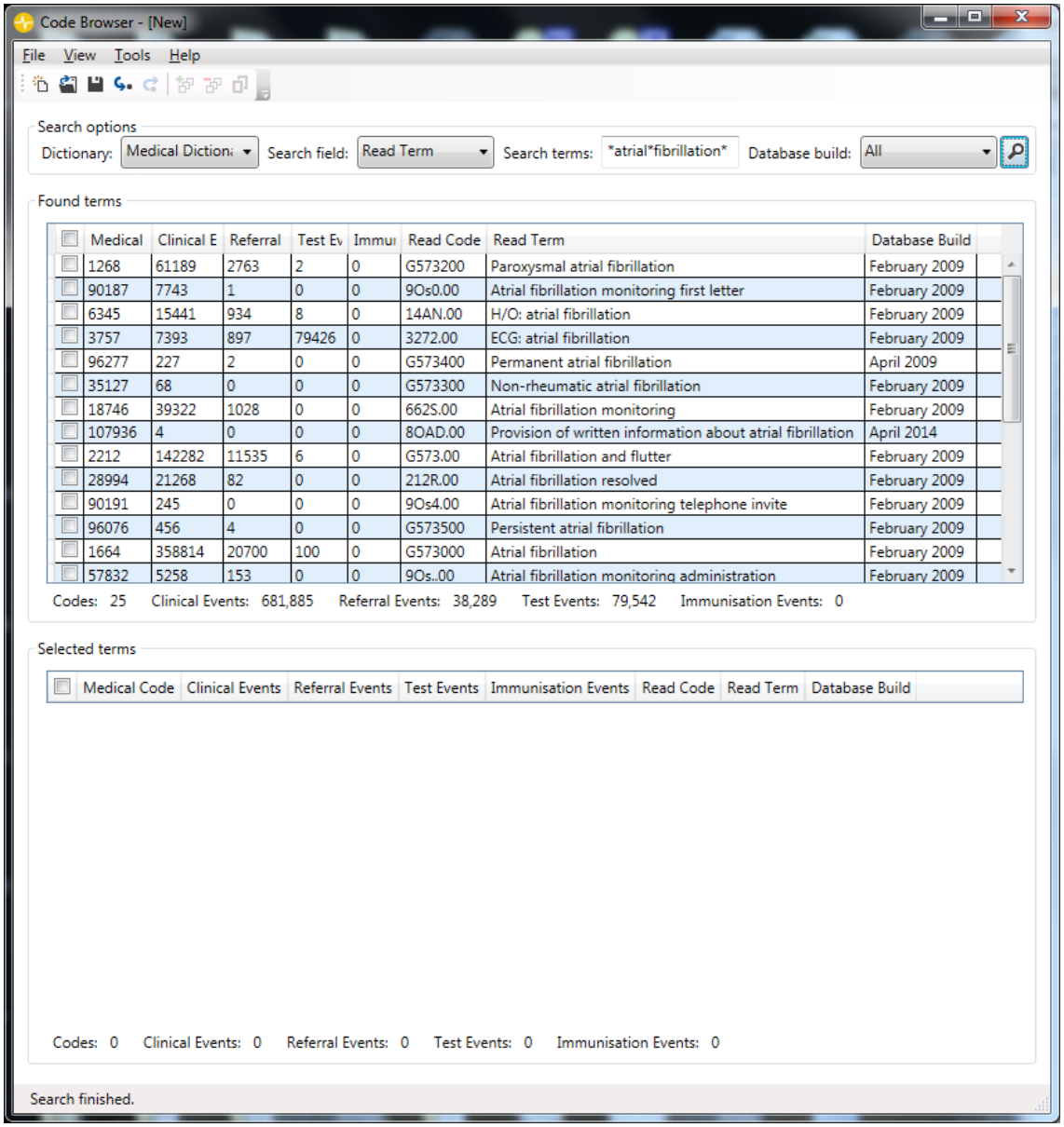
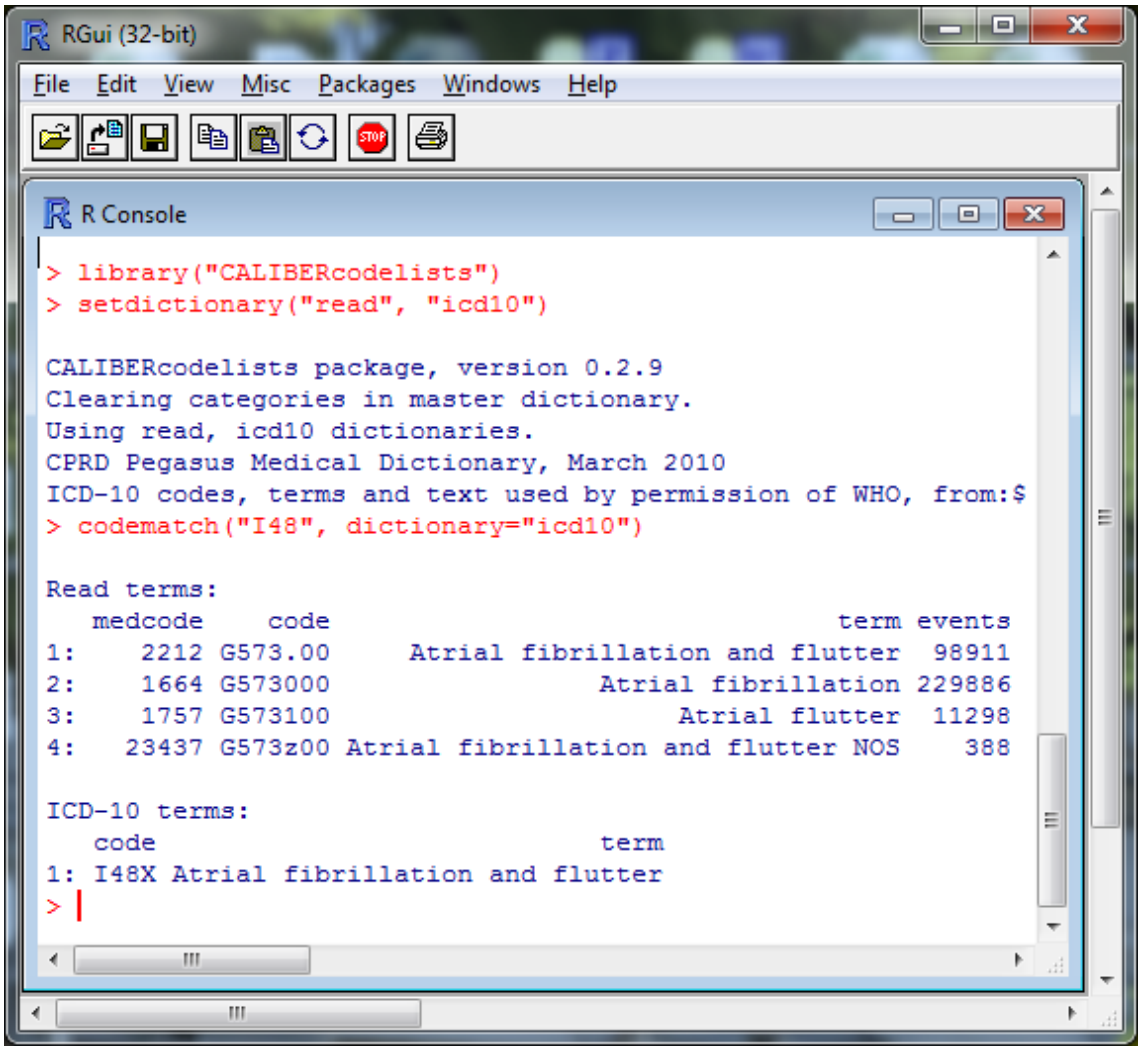


Figure 5.3 Screenshot of CALIBERcodelists package available in R with example of matching synonymous codes for "I48 atrial fibrillation" between the Read and ICD 10 classifications



Chapter 6

What is 'valvular' atrial fibrillation'? A reappraisal exploiting electronic health records

6.1 Chapter outline

Progress in atrial fibrillation (AF) research is hampered by a lack of standardised disease definitions and limited understanding of subtypes. In **chapter 5** I created computable definitions (i.e. definitions that can be computerised and coded to work within electronic health records; EHRs) for eight AF subtypes with relevance to the 2016 updates to European Society of Cardiology (ESC) guidelines for management of AF.[3] However, these definitions now require rigorous testing to ensure that they are usable (i.e. do actually detect cases) and are clinically valid (i.e. make clinical sense). Thus in this chapter I take forward the definition for valvular AF for implementation, improvement and validation. I hypothesised that the EHR definition I created for valvular AF will show clinical validity in replicating known associations between prosthetic heart valves, mitral valve stenosis and an increased risk of stroke, systemic embolism and mortality and that there may be other valve diseases associated with poorer prognosis.

6.2 Abstract

Background Progress in AF research is hampered by a lack of standardised disease definitions and limited understanding of subtypes. I investigated the viability of EHRs for validating and refining AF subtype definitions. I used 'valvular' AF as an example because of known aetiological differences in onset and progression as well as uncertainty in the types of valvular heart disease (VHD) comprising the definition.

Methods I used data on 76,019 individuals with AF recorded in primary or secondary care EHRs in England in 1998 to 2010 as available in CALIBER. I created an algorithm to identify VHD based on diagnosis, procedure and prescription codes from four classification systems: Read (primary care diagnoses and procedures), British National Formulary (BNF; primary care prescriptions), ICD-10 (secondary care diagnoses) and OPCS-4 (secondary care procedures). I improved the algorithm based on EHR insights and used Cox proportional hazards regression to model the associations of VHDs with a composite endpoint of incident stroke (ischaemic, haemorrhagic and unspecified), systemic embolism, and all-cause mortality.

Results The final algorithm combined 165 diagnosis codes, 205 procedure codes and 36 prescription codes (406 codes in total). Algorithm implementation identified 8,623 (11.3%) individuals with AF and VHD at baseline, and a further 4,128 (5.4%) who developed VHD after baseline. Over a median (interquartile range) follow-up of 2.2 (4.2) years, 31,934 endpoint events (9.2% ischaemic stroke, 13.8% unspecified stroke, 1.7% systemic embolism, 2.3% haemorrhagic stroke, 73.1% mortality) occurred with 3,764 (11.8%) in individuals with VHD. Compared with individuals with AF and no VHD, individuals with prosthetic valves, mitral stenosis and aortic stenosis had higher hazard ratios [95% confidence intervals] for stroke, systemic embolism and

mortality of 1.13 [1.02, 1.24], 1.20 [1.05, 1.36], and 1.27 [1.19, 1.37] respectively after adjustment for age, sex, warfarin and CHA_2DS_2-VASc risk factors, while individuals with bioprosthetic valve replacements had a lower adjusted hazard ratio of 0.78 [0.68, 0.88].

Conclusion EHRs are a valuable data source for investigating AF subtypes. In addition to prosthetic heart valves and mitral stenosis, EHR data suggest aortic stenosis is also clinically relevant in the progression of AF.

6.3 Introduction

Progress in AF research is hampered by a lack of standardised disease definitions and limited understanding of subtypes. In recent 2016 updates to the ESC guidelines for management of AF, a range of newly defined clinical distinctions were outlined.[3] However, these AF subtypes were derived based on expert consensus and remain unsupported by large scale quantitative evidence.

EHRs resources like CALIBER,[5] collected on large populations in routine clinical care, offer a viable data source in which to systematically validate and refine understanding of AF subtypes. In the previous chapter (**chapter 5**) I explored whether it is feasibly possible to operationalise the newly defined ESC definitions for AF subtypes in EHRs, creating computable definitions for: (1) AF secondary to structural heart disease, (2) focal AF, (3) polygenic AF, (4) postoperative AF, (5) AF in mitral stenosis or prosthetic heart valves (often referred to as 'valvular' AF), (6) AF in athletes, (7) monogenic AF and (8) AF secondary to respiratory disease. However, these theoretical EHR definitions require rigorous testing to ensure they are usable (i.e. do actually detect cases) and are clinically valid (i.e. make clinical sense).[49] The process of implementing, improving and validating EHR definitions against established clinical knowledge is time and resource intensive and therefore I selected only one AF subtype, valvular AF, to take forward for further development.

According the ESC, the valvular AF distinction exists because of aetiological differences in onset and progression, including a higher risk of subsequent stroke and thromboembolism.[3] However, uncertainty surrounds the types of VHDs comprising the definition as underscored by the following three lines of evidence. First, across international clinical guidelines for the management of AF (i.e. in Europe and the United States), definitions for valvular AF differ and have changed over time (**table 6.1**).[3 22 237-239] Second, recent clinical trials testing the newly introduced direct oral anticoagulants (DOACs) for stroke prophylaxis in AF excluded individuals with valvular AF due to the higher thromboembolic risk, however employed non-identical exclusion criteria (**table 6.2**).[31-34] And third, a recent survey of AF-treating clinicians in Europe found that among 157 cardiologists only 57.1% responded that existing guideline distinctions between valvular and non-valvular AF *"are sufficiently clear"*.[228] In a literature review titled *"What is 'valvular' atrial fibrillation'? A reappraisal",* AF experts, Professor Raffaele De Caterina and Professor John Camm, highlighted the importance of prosthetic heart valves and mitral valve stenosis, however reported a dearth of evidence for thromboembolic risk in other forms of VHD.[6]

Stroke risk assessment and treatment decisions in AF hinge upon the ability to accurately distinguish between valvular and non-valvular subtypes, and thus the current lack of clear-cut definitions could lead to mismanagement and potential harms. Clinicians require clearer guidelines on valvular AF and therefore in this study I had the following three aims: (1) to investigate how EHR data can be used to identify valvular AF cases, (2) to estimate the prevalence of VHDs in the context of AF in the UK between 1998 and 2010, and (3) to quantify the impact of different VHDs on AF prognosis in terms of incident stroke, systemic embolism and all-cause mortality with the potential to offer new evidence on which to base the valvular AF definition.

6.4 Methods

6.4.1 Analysis dataset

Data sources

I used data from the Clinical research using Linked Bespoke studies and Electronic health Records (CALIBER) database[5] which, as a reminder, connects primary care, secondary care and mortality records for a subset of the UK population that is representative of the overall population in terms of age, sex, ethnicity[198] and mortality.[53] Data are coded using four classification systems: Read (primary care diagnoses and procedures),[163] British National Formulary (BNF; primary care prescriptions),[164] ICD-10 (secondary care diagnoses and cause-specific mortality),[45] and OPCS-4 (secondary care procedures).[165]

Study population

Data on 76,019 individuals were available based on the following study entry criteria: age greater than 18 years, an AF diagnosis code (both prevalent and incident cases)[49] recorded in primary or secondary care during the study period of 1998 to 2010, and at least one year of follow-up at a primary care practice with research quality data[174] before analysis. Individuals entered the study on the earliest date these criteria were met and were followed up until the earliest date of occurrence of one of the following study exit criteria: incident stroke (ischaemic, haemorrhagic and unspecified), systemic embolism or all-cause death, transfer out of primary care practice, or end of primary care practice follow-up.

6.4.2 Electronic health record algorithm development: valvular atrial fibrillation

To investigate how EHR data can be used to identify valvular AF cases, I followed the CALIBER guidelines on EHR algorithm development, which I have previously described in **chapter 5** and includes five steps: (1) creating an initial case definition, (2) translating initial case definition into a computable definition based on available clinical codes and supporting information, (3) implementing computable definition in EHRs, (4) improving computable definition based on EHR insights (steps three and four are iterated until a final definition is reached), (5) validating final definition against established clinical knowledge.

(1) creating an initial case definition

The initial case definition for valvular AF was broadly set to capture all adult VHDs with the view to examine the evidence upon which VHDs warrant inclusion in the subtype distinction. I excluded congenital heart valve malformations in order to focus on acquired disease.

(2) translating initial case definition into a computable definition

I translated the initial case definition into a computable definition by identifying 165 diagnosis and 205 procedure codes (370 codes in total) relevant to VHD within the Read (n=235 (63.5%)), ICD-10 (n=49 (13.2%)) and OPCS-4 (n=86 (23.2%)) classification systems (full code list in **Table S5.4** of **appendix**). Diagnosis codes were available to describe stenosis, regurgitation and other and unspecified disorders of the mitral, aortic, tricuspid, pulmonary, and unspecified heart valves with a limited number of codes specifying if the disorder was of rheumatic or non-rheumatic origin. Procedure codes were available to describe prosthetic heart valve replacements (i.e. codes referring to prosthetic or artificial valves, Starr prosthesis, and Bjork-Shiley prosthesis), bioprosthetic heart valve replacements (i.e. codes relating to allografts, xenografts, Carpentier and Edwards prosthesis) and unspecified heart valve replacements, and valvuloplasty (valvotomy), annuloplasty and other and unspecified valve repairs and operations specific to each of the four, and unspecified, heart valves.

In identifying relevant codes, I employed three methods (as described in greater detail in **chapter 5**): (1) sourcing pre-existing code lists, (2) running new and updated code searches and (3) matching synonymous codes across the different classification systems. All identified codes were then reviewed for inclusion and independently verified by a clinical expert (Dr Amitava Banerjee).

(3) implementing computable definition in EHRs

I implemented the computable valvular AF definition in the records identifying individuals with AF and VHD both prevalent at baseline (i.e. VHD first recorded before study entry) and incident over follow up (i.e. VHD first recorded after study entry and before study exit). I examined the quality of recording incident VHD diagnoses and procedures in EHRs in 1998 to 2010. For each individual, I took the earliest VHD diagnosis or procedure date and combined all relating information captured within a subsequent 30-day window (i.e. further VHD diagnosis or procedure codes captured within and across data sources). I then examined code usage and resolution to distinguish between the specific heart valve(s) affected by VHD, between prosthetic and bioprosthetic valve replacements, and between disorders of rheumatic and non-rheumatic origin.

(4) improving computable definition based on EHR insights

Based on insights obtained from the implementation stage, I investigated whether the computable valvular AF definition could be improved to reclassify individuals with un-

specified valve replacements as having a prosthetic or bioprosthetic heart valve. I did this by combining information on age at valve replacement and subsequent warfarin prescriptions, which are key criteria influencing choice of valve replacement.[240]

(5) *validating final definition against established clinical knowledge*

To validate the final valvular AF algorithm, I assigned each individual, based on baseline VHD disease status, to one of nineteen mutually exclusive categories ranked in order of disease severity and importance in AF aetiology (i.e. prosthetic valve replacements, bioprosthetic valve replacements, valve repairs, and then stenosis, regurgitation, and other and unspecified disorders of each of the four, and unspecified, heart valves; **table 6.3**). I then modelled the associations of these baseline VHDs (using no record of VHD as the reference category) with a commonly used AF clinical trial endpoint composed of incident stroke (ischaemic, haemorrhagic and unspecified), systemic embolism, and all-cause mortality.[31-34] As a point of validation, I expected a higher thromboembolic risk in individuals with prosthetic heart valve replacements and mitral valve stenosis as reported in existing literature.[6]

6.4.3 Statistical analysis

Baseline characteristics of individuals with AF with and without VHDs were analysed using numbers and percentages (%) for categorical variables and medians and interquartile ranges and intervals (IQR; IQI) for continuous variables. VHD prevalence percentages were calculated at monthly and yearly intervals over the study period 1998 to 2010 (i.e. 144 months; 12 years) by dividing the total number of individuals with prevalent VHD by the total number of individuals at risk during each interval. LOESS (LOcally wEighted Scatterplot Smoothing) lines, which make no assumption about the shape of the data distribution, were fitted to identify any temporal trends in VHD prevalence.[241] Associations of baseline VHDs with incident stroke, systemic embolism, and mortality were modelled using incrementally adjusted Cox regression with model assumptions and goodness of fit assessed graphically (see **figure S6.1** in **appendix**). Model 1 was adjusted for age and sex, model 2 for age, sex and baseline warfarin prescriptions, and model 3 for age, sex, warfarin and risk factors from the CHA_2DS_2-VASc risk score: congestive heart failure, hypertension, diabetes mellitus, stroke, transient ischaemic attack or SE, and vascular disease. Interaction testing was carried out between baseline VHDs and key confounders of interest: age, sex, warfarin and prior stroke, systemic embolism or transient ischaemic attack (see **table S6.1** in **appendix**). All models were stratified on primary care practice to account for potential local differences in the way clinical codes were applied or conditions managed. Data analyses were performed using statistical software Stata/SE 13.1 and figures produced using R 3.2.0.

6.5 Results

6.5.1 Implementation and improvement of valvular atrial fibrillation case definition

Identification of individuals with atrial fibrillation and valvular heart disease

Based on the 370 applicable codes relating to VHD diagnoses and procedures, 12,751 (16.8%) individuals out of the total AF population of 76,019 had a record of VHD, of which 8,623 (11.3%) had prevalent VHD (i.e. VHD first recorded before study entry), and 4,128 (5.4%) had incident VHD (i.e. VHD first recorded after study entry). A total of 2578 (3.3%) individuals had a record of heart valve replacement including 1902 (2.5%) prevalent at baseline and 676 (0.9%) incident over follow-up.

Quality of recording of valvular heart diseases in electronic health records

Quality of recording: specific heart valves

Among the 4,128 individuals with an incident record of VHD in 1998 to 2010, 2700 (65.4%) had a record of mitral valve disorders, 1288 (31.2%) had a record of aortic valve disorders, 398 (9.6%) had a record of tricuspid valve disorders, and 48 (1.2%) had a record of pulmonary valve disorders. For 63 (1.5%) individuals, only unspecified heart valve disorder codes were recorded (see **figure S6.2** in **appendix**).

Quality of recording: prosthetic vs. bioprosthetic valves

Among the 676 individuals with an incident record of heart valve replacements in 1998 to 2010, a majority of 570 (84.3%) individuals had a record specifying that the replacement was prosthetic or bioprosthetic while 106 (15.7%) individuals only had a record indicating an unspecified heart valve replacement. Unspecified valve replacement codes were used in combination with both prosthetic and bioprosthetic replacements, however prosthetic and bioprosthetic replacement codes were rarely (i.e. in only 8 (1.2%) individuals) used in combination (**figure S6.3** in **appendix**).

Quality of recording: rheumatic valve disease

Among the 4,128 individuals with an incident record of VHD in 1998 to 2010, 316 (7.7%) had a record of rheumatic heart valve disorders, 359 (8.7%) had a record of non-rheumatic disorders, while for the majority of 3637 (88.1%) individuals, the rheumatic basis of the disorder was unspecified (**figure S6.4** in **appendix**). Because of limited recording of rheumatic VHD and no supporting data to confirm cases, I did not include this in the final algorithm.

Improving identification of prosthetic vs. bioprosthetic heart valve replacements

For individuals with incident record of prosthetic, bioprosthetic and unspecified heart valve replacements in 1998 to 2010, there were differences in the distributions for age at heart valve replacement (**figure S6.5** in **appendix**) and in the percentages with warfarin prescriptions in the 6 months after heart valve replacement surgery. The median (IQI) ages for prosthetic, bioprosthetic and unspecified heart valve replacements were 70.1 (62.2, 75.8), 75.8 (72.2, 80.8), and 75.1 (65.7, 81.1) years respectively. The percentages with warfarin prescriptions for prosthetic, bioprosthetic and unspecified heart valve replacements were 74.3%, 52.1% and 57.6% respectively. Based on these insights it could be inferred that individuals younger than 70 years are likely to have a prosthetic valve replacement and individuals greater than 75 years are likely to have a bioprosthetic valve replacement. For individuals 70 to 75 years, where the age distribu-

tion for prosthetic and bioprosthetic replacements overlap, I used warfarin prescriptions (63 codes) to infer a prosthetic valve replacement and no warfarin prescription to infer a bioprosthetic valve replacement.

6.5.2 Application of the final algorithm

The final algorithm combining 406 diagnosis, procedure and prescription codes and ages at, and warfarin following, heart valve replacement surgery is depicted in **figure 6.1**.

Baseline valvular heart diseases

Implementation of the final algorithm at baseline classified 67,396 (88.7%) individuals as with no VHD, 1207 (1.6%) with a prosthetic heart valve replacement, 695 (0.9%) with a bioprosthetic valve replacement, 434 (0.6%) with a heart valve repair, 527 (0.7%) with mitral valve stenosis, 2374 (3.1%) with mitral valve regurgitation, 974 (1.3%) with other and unspecified mitral valve disorders, 1494 (2.0%) with aortic valve stenosis, 444 (0.6%) with aortic valve regurgitation, and 197 (0.3%) with other and unspecified aortic valve disorders. The algorithm identified low numbers of individuals with tricuspid (n=187 (0.2%)), pulmonary (n=39 (0.1%)) and unspecified valve disorders (n=51 (0.1%)), and therefore these individuals were not included in any further analyses.

Baseline characteristics

An overview of characteristics of the overall study population and in individuals with different forms of VHD is provided in **table 6.4**. The overall study population had median (IQR) age at baseline of 77.7 (15.0) years, were 49.1% women, were followed-up for a median (IQR) of 2.2 (4.2) years, and consisted of 26.1% with heart failure, 82.7% with hypertension, 14.2% with diabetes mellitus, 18.2% with stroke, transient ischaemic attack or systemic embolism, 19.8% with vascular disease, 3.4% with CHA_2DS_2-VASc=0, 8.0% with CHA_2DS_2-VASc=1, 88.6% with CHA_2DS_2-VASc≥2, and 21.5% with a warfarin prescription in the six months before study entry.

Compared to individuals without VHD, individuals with VHD had higher percentages of heart failure and hypertension: 23.6% vs. ≥37.6% and 81.8% vs. ≥86.1% respectively. Individuals with prosthetic heart valve replacements had the lowest median (IQR) baseline age of 70.5 (13.4) years and highest percentage of warfarin prescription of 68.7%. Individuals with mitral valve stenosis were 75.3% women, had the highest percentage of prior stroke, transient ischaemic attack or systemic embolism of 24.5%, and had the second highest percentage of warfarin prescriptions of 59.6%. Individuals with aortic stenosis had the highest median (IQR) baseline age of 82.2 (11.3) years, the highest percentage of diabetes mellitus of 17.1%, and highest percentage of vascular disease of 28.9%.

Prevalence and incidence of valvular heart diseases in 1998 to 2010

The monthly prevalence of VHDs over the study period 1998 to 2010 is visualised in **figure 6.2**. Increases in prevalence were shown for bioprosthetic valve replacements (from 0.7% in 1998-9 to 2.1% in 2009-10), heart valve repairs (from 0.9% to 2.2%), mitral valve regurgitation (from

4.9% to 8.1%), aortic stenosis (from 2.3% to 5.0%) and aortic regurgitation (from 1.2% to 2.3%). The prevalence of prosthetic heart valve replacements increased from 2.1% in 1998-9 to 2.6% in 2002-3, and then plateaued over the remaining years of follow-up. The prevalence of mitral valve stenosis decreased from 2.0% in 1998-9 to 1.4% in 2009-10.

Associations of valvular heart diseases with incident stroke, systematic embolism and death

Results of the incrementally adjusted cox models are provided in **table 6.5** and **figure 6.3**. Overall 31,934 endpoint events (9.2% ischaemic stroke, 13.8% unspecified stroke, 1.7% systemic embolism, 2.3% haemorrhagic stroke, 73.1% mortality) occurred with 3,764 (11.8%) in individuals with VHD. Compared with individuals with AF and no VHD, individuals with prosthetic valve replacements, mitral stenosis and aortic stenosis had higher hazard ratios [95% confidence intervals] for stroke, systemic embolism and mortality of 1.13 [1.02, 1.24], 1.20 [1.05, 1.36], and 1.27 [1.19, 1.37] respectively after adjustment for age, sex, warfarin prescription and CHA_2DS_2-VASc risk factors, while individuals with bioprosthetic valve replacements had a lower adjusted hazard ratio of 0.78 [0.68, 0.88].

6.6 Discussion

Overview of key findings

This study reports a transparent and reproducible algorithm to identify VHDs in individuals with AF in linked EHRs with new evidence on which to base the 'valvular' AF distinction and key implications for future research. The algorithm, underpinned by 406 clinical codes from four classification systems, compatible in primary and secondary care records and tested on 76,019 individuals with AF, was shown to be clinically valid in replicating known associations of a higher thromboembolic risk in individuals with prosthetic heart valve replacements and mitral valve stenosis.[6] In addition, individuals with aortic stenosis were also shown to have worse outcomes (than individuals with no VHD), suggesting that aortic stenosis is also likely to be clinically relevant in the progression of AF. Overall these findings strengthen support for the use of EHR data in validating and refining understanding of AF subtypes.

Risks of thromboembolism and mortality

Consistent with the often variable definitions for valvular AF reported in clinical practice guidelines,[3 22] DOAC trials,[31-34] and De Caterina and Camm's literature review from 2014,[6] I found a higher risk of stroke, systemic embolism and mortality in individuals with prosthetic heart valve replacements and mitral valve stenosis, which reinforces the evidence for including these in the valvular AF subtype distinction. As highlighted by De Caterina and Camm, a dearth of evidence currently exists for prognosis in individuals with AF and other forms of VHD.[6] Thus a novel finding from this analysis is that individuals with AF and aortic valve stenosis fared significantly worse than individuals with AF and no VHD, and in particular had higher levels of mortality. Traditional epidemiological cohort studies (e.g. the Framingham Heart Study with 1544 incident cases of AF accrued over 50 years of follow-up[191]) and snap-shot AF registries (e.g. the Euro-Heart Survey of 5,333 AF patients enrolled from one of 182 hospitals in Europe in 2003 to 2004[199]), have in the past provided important insights into AF epidemiology, however are unlike-

ly to have large enough population denominators to detect associations with AF subtypes.[204] Therefore, assuming the association between aortic stenosis and poorer prognosis is true, the smaller sample sizes of prior studies is one possible explanation as to why it has not been reported before in the observational literature. My finding regarding aortic stenosis is however supported by evidence from the interventional setting. A recent reanalysis of data from the ROCKET-AF clinical trial of the DOAC rivoroxaban found that compared to individuals with mitral regurgitation (MR), aortic regurgitation (AR) or no significant valve disease (SVD), those with aortic stenosis (AS) had higher rates of the composite endpoint of stroke, systemic embolism or vascular death (AS: 10.84, MR or AR: 4.54 and no SVD: 4.31 events per 100 patient-years, p=0.0001) and had higher all-cause death (AS: 11.22, MR or AR 4.90: and no SVD: 4.39 events per 100 patient-years, p=0.0003).[242] Another confirmatory finding of this study is that a negligible number of individuals had AF and disorders involving the tricuspid or pulmonary valves, which suggests that these conditions are unlikely to play a substantive role in the progression of AF.

Value of electronic health records in investigating AF subtypes

As well as offering a large population denominator giving rise to novel associations between VHDs and AF prognosis, the value of EHR data was also demonstrated in the quality of recording incident VHDs over the study period 1998 to 2010. As shown, the 98.5% of records had detailed resolution to distinguish between the specific heart valve(s) affected by VHD and 84.3% of records had detailed resolution to distinguish between prosthetic and bioprosthetic heart valve replacements. Furthermore, using insights obtained from the EHRs I showed that it was possible to infer whether an individual with an unspecified heart valve replacement code had a prosthetic or bioprosthetic valve based on differences in age at, and warfarin prescriptions following, valve replacement operations. While a small number of clinical codes exist to differentiate between VHDs of rheumatic and non-rheumatic origin, I found that they were seldom used in practice. While this most likely reflects a systematic limitation in the way the EHRs are captured and recorded at source,[48] another contributing factor may lie in declining prevalence rates of rheumatic heart diseases in industrialised countries, such as the UK.[243] In keeping with prior reports,[244 245] mitral regurgitation and aortic stenosis, which are disorders associated with ageing, were the most common VHDs in this study with year-on-year increases in prevalence. Mitral valve stenosis, which is almost always on a rheumatic basis, was the only VHD shown to decrease in prevalence over the study period.

Clinical implications

A lack of clarity in the definition for valvular AF has previously been reported among clinicians surveyed in Europe with the potential for harm through inappropriate treatment decisions.[228] This is particularly pertinent to AF because most of the epidemiology and trial literature focuses on non-valvular AF, which has been variably defined.[246 247] The findings of this analysis – the largest systematic attempt to investigate valvular AF to my knowledge to date – can therefore give clinicians reassurance that individuals with AF and prosthetic heart valve replacements or mitral stenosis require separate clinical attention. According to current guidelines these individ-

uals require lifelong stroke prophylaxis by way of warfarin; DOACs are contraindicated due to higher complication rates (prosthetic valves) and lack of randomised trial evidence (mitral stenosis).[247] Clinicians in industrialised countries are likely to see fewer and fewer individuals with mitral valve stenosis in the future because of improved living environments and prevention and treatment of acute rheumatic fever with antibiotics.[248] However, rheumatic heart disease remains the predominant cause of valvular disease in developing countries, and therefore the results of this study, reinforcing the link between AF, mitral stenosis and poorer prognosis, have important implications in terms of global health provision. While this study also suggests aortic stenosis may be clinically relevant in the progression of AF, one study in isolation is insufficient to change practice at this stage.

Research implications

This study has four key implications for future research. First, given that DOAC trials used disparate criteria to exclude individuals with valvular AF, [31-34] the confirmation that prosthetic heart valve replacements and mitral stenosis confer a significantly higher thromboembolic risk, provides a stronger basis on which to inform the design of future clinical trials. Second, the VHD code list generated and shared through this study helps progress AF research by providing a transparent and reproducible definition on which future EHR analyses can align and results can be more easily compared. The code list generated is useful, not only in EHR studies of AF, but for EHR studies of valvular disease in general, which is timely considering that the new 2017 European guidelines for management of VHDs pinpoint extensive evidence gaps in risk stratification and comparative effectiveness of different surgical interventions.[249] The third research implication refers to the quality of coding of VHDs in EHRs in which I uncovered a potential underuse of available codes for rheumatic and non-rheumatic heart valve diseases. Whereas for unspecified heart valve replacements, I showed that prosthetic and bioprosthetic replacements could be inferred by combing information on age at valve replacement and warfarin prescriptions, I unfortunately did not have access to any relevant supporting data to create an inference for rheumatic heart disease. Future research could therefore examine whether the rheumatic basis of VHD can be inferred in EHRs by incorporating antibiotics data. The final research implication involves internationalisation. As shown the algorithm is valid for primary and secondary care linked EHRs in the UK, however external validity remains untested. An important next step for this research would be to examine the extent to which the algorithm is compatible with comparable international datasets. Both Denmark[42] and Sweden[43] have nation-wide secondary care records with data elements in common with CALIBER (e.g. diagnoses coded with ICD-10 and linkage to prescriptions and mortality data). Successful implementation of the algorithm in these international datasets would offer the large sample size needed to re-examine the aortic stenosis finding.

Strengths and limitations

This study's prevailing strengths lie in the vast availability and proven validity of EHRs in the CALIBER resource.[5] The availability of a large sample of 76,019 individuals with AF from across both primary and secondary care allowed meaningful analyses of the valvular AF subtype,

which would have likely been underpowered if using, for example, data from traditional consented cohort studies.[204] Given that the primary purpose of collection of EHRs does not include research, the validity of data is often questioned,[48] however this study provides further evidence of the validity of EHR data by way of replicating known clinical associations between prosthetic valves, mitral stenosis and poorer prognosis in AF.[6] A limitation already reflected upon is that the quality of EHR data depends upon how well information has been captured and recorded at source; in particular whether specific codes exist to describe a condition of interest and are actually used in practice.[48] I showed that while higher resolution VHD codes were available (e.g. indicating the type of heart valve replacement) they were not always used, but by combining diagnosis codes with supplementary information, these issues can in some cases be overcome. One area where EHR data can be weak is in the coding of disease severity. The definitions I used to identify VHD therefore did not differentiate between mild, moderate and severe forms. Assuming that in general more severe disease cases will likely enter EHRs, this may have exaggerated the impact that, for example, aortic stenosis has on prognosis, and future investigations stratified on valve disease severity should be carried out in order to clarify the robustness of my findings. Similarly as I focussed on a composite endpoint of stroke, systemic embolism and mortality this assumes that each individual component has equal clinical importance and can mask whether there is any heterogeneity in how each one relates to the exposure. Individuals with aortic stenosis experienced higher levels of mortality, for example, and therefore further investigations are needed into how this may have impacted associations with the composite endpoint. In order to separate out the impact of different forms of VHD I created a hierarchy for classifying individuals into one category at baseline based on perceived clinical importance (**table 6.3)**. I first considered valve replacements, valve repairs, followed by valve diseases: mitral stenosis, aortic stenosis, mitral regurgitation and aortic regurgitation (diseases of tricuspid and pulmonary valves were excluded due to limited numbers of events). The order for classifying valve diseases is supported by evidence from the subgroup analysis of the ROCKET-AF trial (showing worse outcomes in AS vs. MR and AR).[242] However, a limitation of analysing individuals based on the most clinically relevant VHD does not take into account prognosis in individuals who are impacted by more than one disorder. Likewise this may have given a distorted impression of the distribution of VHDs within the cohort. For example, the most prevalent VHD at baseline was mitral regurgitation recorded in 3312 (4.4%) individuals but execution of the hierarchy reduced this down to 2374 (3.1%) individuals. A distortion in the proportion of individuals with mitral regurgitation (relative to those with mitral stenosis) may also have occurred due to recording practices relating to disease severity. Given the poor prognosis associated with mitral valve stenosis, it's likely that mild, moderate and severe forms are frequently captured whereas for mitral regurgitation only severe forms are entered into EHRs. This potential systematic bias in recording should therefore be taken into account when interpreting the results of this study and ideally quantified in future EHR validation exercises (e.g. GP/patient re-contact studies to first hand verify the information). Further limitations of this study reflect current challenges in data access and data retention.[169] Data access challenges meant the most recent year of follow-up for this study was the year 2010. A more recent data extract clearly would have been desirable, however, with the exception of lack of information on individuals treated

with DOACs, EHRs collected within the last ten years, as I have shown, still provide contemporary insights into the clinical management of AF. Furthermore, because I conducted code searches up until the year 2016, it means the algorithm can be implemented in more recent data extracts. Due to restrictions in data retention I could not request any additional clinical data points in relation to this study. I was therefore unable to investigate whether antibiotic prescriptions data could enhance the capture of rheumatic heart disease cases. Finally, while linkages between primary care, secondary care and mortality records in CALIBER provide an enhanced data source in which to intricately investigate healthcare journeys, linkages to imaging datasets are not currently available at scale within the UK. Therefore without electrocardiogram and echocardiogram images, it is possible that some AF and VHD cases may have been missed or misclassified. Although a prior validation study suggests 96% accuracy of coded AF diagnoses in UK primary care data.[180] Lastly, the associations presented here are observational and, as always, may be influenced by unmeasured confounding. Ideally, individuals with AF and valvular heart disease should be recruited into randomised clinical trials testing, for example, whether individuals with AF and aortic valve stenosis derive a prognostic benefit from oral anticoagulants.

6.7 Conclusion

EHRs are a valuable data source for investigating AF subtypes. In addition to prosthetic heart valves and mitral stenosis, EHR data suggest aortic stenosis is also clinically relevant in the progression of AF.

6.8 Chapter summary

To summarise, in this chapter I took forward the computable definition I previously developed for AF in the context of VHDs for further refinement and validation. As well as using relevant clinical codes to identify cases, I showed that individuals with prosthetic and bioprosthetic heart valves could be better detected in EHRs when combining information on warfarin prescriptions and age at valve replacement surgery, which are key criteria influencing the choice of heart valve replacement.[240] In line with current understanding,[6] individuals with prosthetic heart valves and mitral valve stenosis were found to have a higher relative risk of stroke, systemic embolism and mortality as compared to individuals with no record of VHD. In addition, individuals with aortic valve stenosis were also shown to have a poorer prognosis, which is a novel contribution of this work. Moving on to **chapter 7**, I take a closer look at AF outcomes by investigating rates of ischaemic stroke by sex, estimated stroke risk and use of warfarin.

6.9 Chapter tables

Table 6.1 Comparison of differences and changes over time in definitions for valvular atrial fibrillation across international clinical guidelines

	Current clinical guideline Definitions for valvular AF	Previous clinical guideline Definitions for valvular AF
Guidelines	**Year: description**	**Year: description**
European Society of Cardiology	**2016:** Mitral stenosis or prosthetic heart valves. [3]	**2012:** Rheumatic valvular disease (predominantly mitral stenosis) or prosthetic heart valves.[238]
American Heart Association	**2014**: Rheumatic mitral stenosis, a mechanical or bioprosthetic heart valve, or mitral valve repair.[22]	**2006:** Rheumatic mitral valve disease, a prosthetic heart valve, or mitral valve repair.[239]

Abbreviations: AF – atrial fibrillation

Table 6.2 Comparison of non-identical criteria used to exclude individuals with valvular atrial fibrillation in recent clinical trials testing efficacy and safety of direct oral anticoagulants (DOACs) for stroke prevention

Clinical trial definitions used to exclude individuals with valvular AF	
Trial: drug	**Description**
RE-LY: Dabigatran etexilate	History of heart valve disorder (i.e., prosthetic valve or hemodynamically relevant valve disease)[31]
ROCKET AF: Rivaroxaban	Hemodynamically significant mitral valve stenosis Prosthetic heart valve (annuloplasty with or without prosthetic ring, commissurotomy and/or valvuloplasty are permitted)[32]
ARISTOTLE: Apixaban	Moderate or severe mitral stenosis, conditions other than atrial fibrillation that required anticoagulation (e.g., a prosthetic heart valve)[33]
ENGAGE AF-TIMI 48: Edoxaban	Moderate or severe mitral stenosis, unresected atrial myxoma, or a mechanical heart valve (subjects with bioprosthetic heart valves and/or valve repair can be included)[34]

Abbreviations: AF – atrial fibrillation, RE-LY - Randomized Evaluation of Long-Term Anticoagulation Therapy, ROCKET AF - Rivaroxaban Once Daily Oral Direct Factor Xa Inhibition Compared with Vitamin K Antagonism for Prevention of Stroke and Embolism Trial in Atrial Fibrillation, ARISTOTLE - Apixaban for Reduction in Stroke and Other Thromboembolic Events in Atrial Fibrillation, ENGAGE AF-TIMI 48 - Effective Anticoagulation with Factor Xa Next Generation in Atrial Fibrillation–Thrombolysis in Myocardial Infarction 48.

Table 6.3 Hierarchy of nineteen mutually exclusive valvular heart disease categories used to classify individuals in terms of baseline status

Rank	Valvular heart disease description
1	Prosthetic valve replacement
2	Bioprosthetic valve replacement
3	Valve repair
4	Mitral valve stenosis
5	Aortic valve stenosis
6	Tricuspid valve stenosis
7	Pulmonary valve stenosis
8	Unspecified heart valve stenosis
9	Mitral valve regurgitation
10	Aortic valve regurgitation
11	Tricuspid valve regurgitation
12	Pulmonary valve regurgitation
13	Unspecified heart valve regurgitation
14	Other and unspecified mitral valve disorders
15	Other and unspecified aortic valve disorders
16	Other and unspecified tricuspid valve disorders
17	Other and unspecified pulmonary valve disorders
18	Other and unspecified heart valve disorders
19	No valvular heart disease

Table 6.4 Characteristics in individuals with atrial fibrillation with and without prevalent valvular heart diseases at baseline

		Valve re-placement			Mitral valve			Aortic valve			
	No VHD	Prosthetic	Bioprosthetic	Valve repair	Stenosis	Regurgitation	Other	Stenosis	Regurgitation	Other	Overall cohort
Individuals, N	67396	1207	695	434	527	2374	974	1494	444	197	76019
	(88.7)	(1.6)	(0.9)	(0.6)	(0.7)	(3.1)	(1.3)	(2.0)	(0.6)	(0.3)	(100.0)
Age, M (IQR)	77.7 (15.2)	70.5 (13.4)	78.7 (10.2)	71.8 (13.8)	75.5 (15.1)	77.8 (13.3)	78.2 (14.0)	82.2 (11.3)	78.9 (13.4)	80.3 (11.8)	77.7 (15.0)
Sex, N (%)											
Men	34587 (51.3)	670 (55.5)	397 (57.1)	211 (48.6)	130 (24.7)	1134 (47.8)	427 (43.8)	700 (46.9)	236 (53.2)	89 (45.2)	38720 (50.9)
Women	32809 (48.7)	537 (44.5)	298 (42.9)	223 (51.4)	397 (75.3)	1240 (52.2)	547 (56.2)	794 (53.1)	208 (46.8)	108 (54.8)	37299 (49.1)
Heart failure, N (%)	15896 (23.6)	539 (44.7)	287 (41.3)	187 (43.1)	237 (45)	1164 (49)	432 (44.4)	684 (45.8)	196 (44.1)	74 (37.6)	19840 (26.1)
Hypertension, N (%)	55110 (81.8)	1083 (89.7)	647 (93.1)	388 (89.4)	454 (86.1)	2129 (89.7)	845 (86.8)	1377 (92.2)	405 (91.2)	181 (91.9)	62849 (82.7)
Diabetes mellitus, N (%)	9568 (14.2)	152 (12.6)	101 (14.5)	49 (11.3)	81 (15.4)	327 (13.8)	139 (14.3)	255 (17.1)	51 (11.5)	26 (13.2)	10798 (14.2)
Stroke/ TIA/ SE, N (%)	12201 (18.1)	224 (18.6)	111 (16.0)	62 (14.3)	129 (24.5)	426 (17.9)	193 (19.8)	326 (21.8)	89 (20.0)	46 (23.4)	13862 (18.2)
Vascular disease, N (%)	13019 (19.3)	202 (16.7)	146 (21.0)	75 (17.3)	84 (15.9)	608 (25.6)	256 (26.3)	432 (28.9)	97 (21.8)	46 (23.4)	15036 (19.8)
CHA_2DS_2-VASc, N (%)											
0	2474 (3.7)	34 (2.8)	8 (1.2)	14 (3.2)	6 (1.1)	27 (1.1)	25 (2.6)	5 (0.3)	10 (2.3)	2 (1.0)	2609 (3.4)
1	5578 (8.3)	120 (9.9)	30 (4.3)	43 (9.9)	28 (5.3)	126 (5.3)	48 (4.9)	29 (1.9)	15 (3.4)	6 (3.0)	6047 (8.0)
2	9168 (13.6)	212 (17.6)	78 (11.2)	75 (17.3)	65 (12.3)	236 (9.9)	111 (11.4)	91 (6.1)	55 (12.4)	13 (6.6)	10143 (13.3)
3	13399 (19.9)	259 (21.5)	118 (17.0)	82 (18.9)	88 (16.7)	400 (16.8)	143 (14.7)	253 (16.9)	81 (18.2)	42 (21.3)	14896 (19.6)
4	16106 (23.9)	263 (21.8)	209 (30.1)	98 (22.6)	113 (21.4)	627 (26.4)	238 (24.4)	384 (25.7)	112 (25.2)	47 (23.9)	18260 (24.0)
5	10562 (15.7)	171 (14.2)	143 (20.6)	71 (16.4)	103 (19.5)	508 (21.4)	194 (19.9)	342 (22.9)	83 (18.7)	42 (21.3)	12273 (16.1)
6	6386 (9.5)	91 (7.5)	64 (9.2)	29 (6.7)	67 (12.7)	252 (10.6)	129 (13.2)	218 (14.6)	51 (11.5)	24 (12.2)	7344 (9.7)
7	2804 (4.2)	44 (3.6)	34 (4.9)	17 (3.9)	41 (7.8)	142 (6.0)	54 (5.5)	130 (8.7)	30 (6.8)	16 (8.1)	3337 (4.4)
8	789 (1.2)	11 (0.9)	11 (1.6)	3 (0.7)	14 (2.7)	52 (2.2)	27 (2.8)	37 (2.5)	7 (1.6)	4 (2.0)	959 (1.3)
9	130 (0.2)	2 (0.2)	0 (0.0)	2 (0.5)	2 (0.4)	4 (0.2)	5 (0.5)	5 (0.3)	0 (0.0)	1 (0.5)	151 (0.2)
Follow-up, M (IQR)											
	2.3 (4.3)	3.1 (5.0)	2.0 (3.3)	2.7 (4.5)	2.5 (4.7)	2.1 (3.8)	1.8 (3.6)	1.2 (2.7)	1.9 (4.3)	1.9 (3.7)	2.2 (4.2)
Endpoints, N (%)											
Ischaemic stroke	2606 (3.9)	44 (3.6)	33 (4.7)	8 (1.8)	32 (6.1)	88 (3.7)	32 (3.3)	52 (3.5)	21 (4.7)	9 (4.6)	2930 (3.9)
Unspecified stroke	4030 (6.0)	43 (3.6)	22 (3.2)	12 (2.8)	33 (6.3)	103 (4.3)	50 (5.1)	77 (5.2)	18 (4.1)	9 (4.6)	4405 (5.8)
Systemic embolism	479	6	4	1	6	14	12	10	5	0	540

	(0.7)	(0.5)	(0.6)	(0.2)	(1.1)	(0.6)	(1.2)	(0.7)	(1.1)	(0)	(0.7)
Haemorrhagic stroke	626	19	10	6	5	20	8	8	7	1	711
	(0.9)	(1.6)	(1.4)	(1.4)	(0.9)	(0.8)	(0.8)	(0.5)	(1.6)	(0.5)	(0.9)
Death	20428	353	165	95	174	825	319	647	164	75	23348
	(30.3)	(29.2)	(23.7)	(21.9)	(33)	(34.8)	(32.8)	(43.3)	(36.9)	(38.1)	(30.7)
Warfarin, N (%)	13098	829	190	236	314	839	310	306	91	47	16339
	(19.4)	(68.7)	(27.3)	(54.4)	(59.6)	(35.3)	(31.8)	(20.5)	(20.5)	(23.9)	(21.5)

Notes: continuous values are presented as median and interquartile ranges and categorical variables are presented as percentages.

Abbreviations: VHD – valvular heart disease, N – number, M (IQR) – median and interquartile range, TIA – transient ischaemic attack, SE – systemic embolism, CHA_2DS_2-VASc – an acronym for the congestive heart failure, hypertension, age (≥75 years), diabetes mellitus, stroke/TIA/SE, vascular disease, age (65 to 75) and sex (female) risk score.

Table 6.5 Incrementally adjusted hazard ratios for associations of baseline valvular heart diseases with incident stroke, systemic embolism, and mortality

	Individuals	Events	HR [95% CI][1]	HR [95% CI][2]	HR [95% CI][3]
No heart valve disease	67396	28169	referent	referent	referent
Prosthetic valve replacement	1207	465	1.17 [1.07, 1.28]	1.26 [1.14, 1.38]	1.13 [1.02, 1.24]
Bioprosthetic valve replacement	695	234	0.82 [0.72, 0.94]	0.84 [0.74, 0.95]	0.78 [0.68, 0.88]
Valve repair	434	122	0.88 [0.74, 1.06]	0.93 [0.78, 1.12]	0.84 [0.70, 1.01]
Mitral stenosis	527	250	1.25 [1.10, 1.42]	1.34 [1.18, 1.52]	1.20 [1.05, 1.36]
Mitral regurgitation	2374	1050	1.14 [1.07, 1.21]	1.17 [1.10, 1.25]	1.05 [0.99, 1.12]
Other mitral disorder	974	421	1.19 [1.08, 1.32]	1.22 [1.11, 1.35]	1.09 [0.99, 1.20]
Aortic stenosis	1494	794	1.41 [1.32, 1.52]	1.42 [1.32, 1.53]	1.27 [1.19, 1.37]
Aortic regurgitation	444	215	1.23 [1.07, 1.41]	1.23 [1.08, 1.41]	1.13 [0.98, 1.29]
Other aortic disorder	197	94	1.14 [0.93, 1.39]	1.15 [0.94, 1.41]	1.10 [0.90, 1.35]

Notes:

[1] model adjusted for age, sex and stratified on primary care practice

[2] model 1 plus adjustment for baseline warfarin prescription

[3] model 2 plus adjustment for heart failure, hypertension, diabetes mellitus, stroke, transient ischaemic attack or system embolism, and vascular disease.

[4] model 3 plus adjustment for covariate interactions: warfarin * heart failure, warfarin * hypertension, heart failure * vascular disease, heart failure * stroke

Abbreviations: HR [95%CI] – hazard ratio and 95% confidence interval.

6.10 Chapter figures

Figure 6.1 Electronic health record algorithm to classify valvular heart diseases in individuals with atrial fibrillation

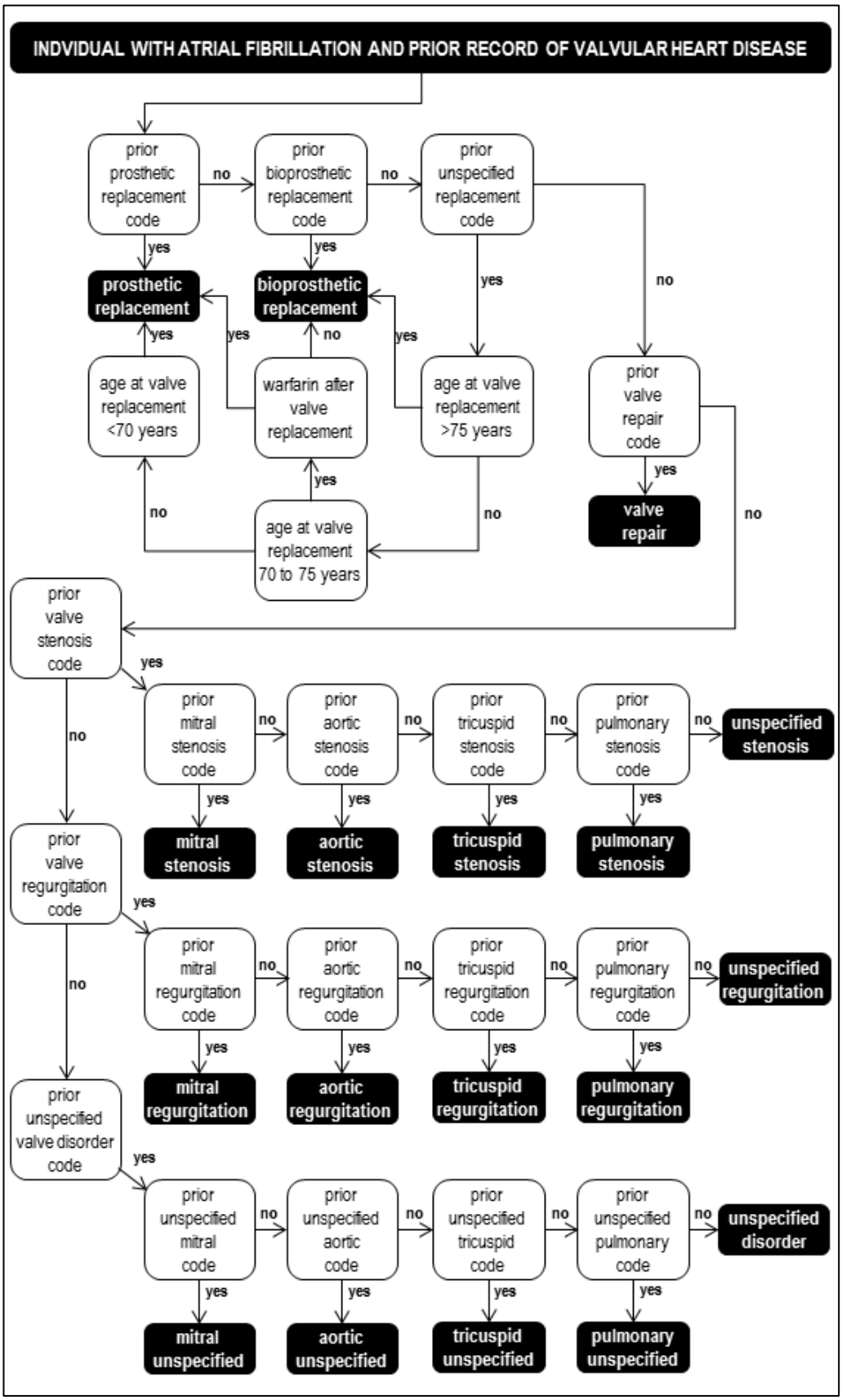


Figure 6.2 Prevalence of valvular heart diseases in individuals with atrial fibrillation between 1998 and 2010

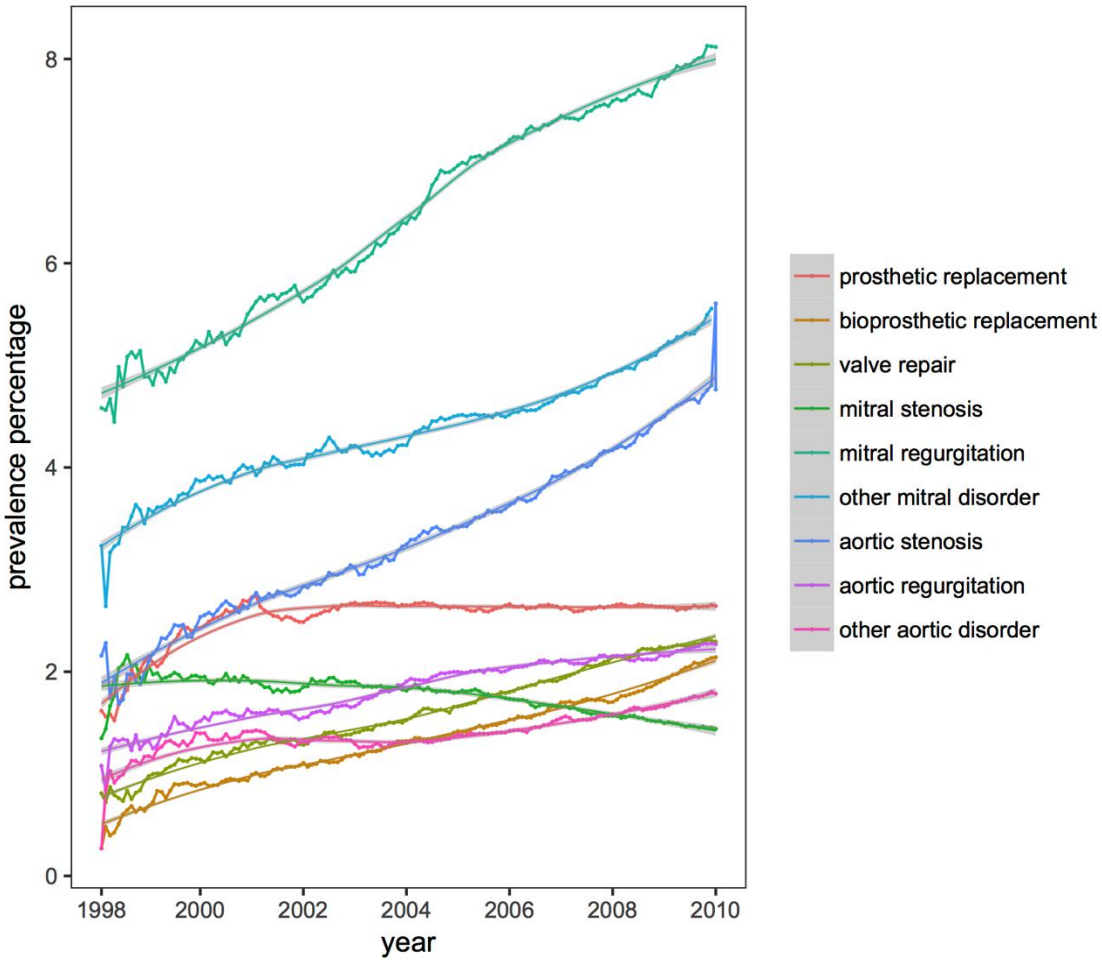
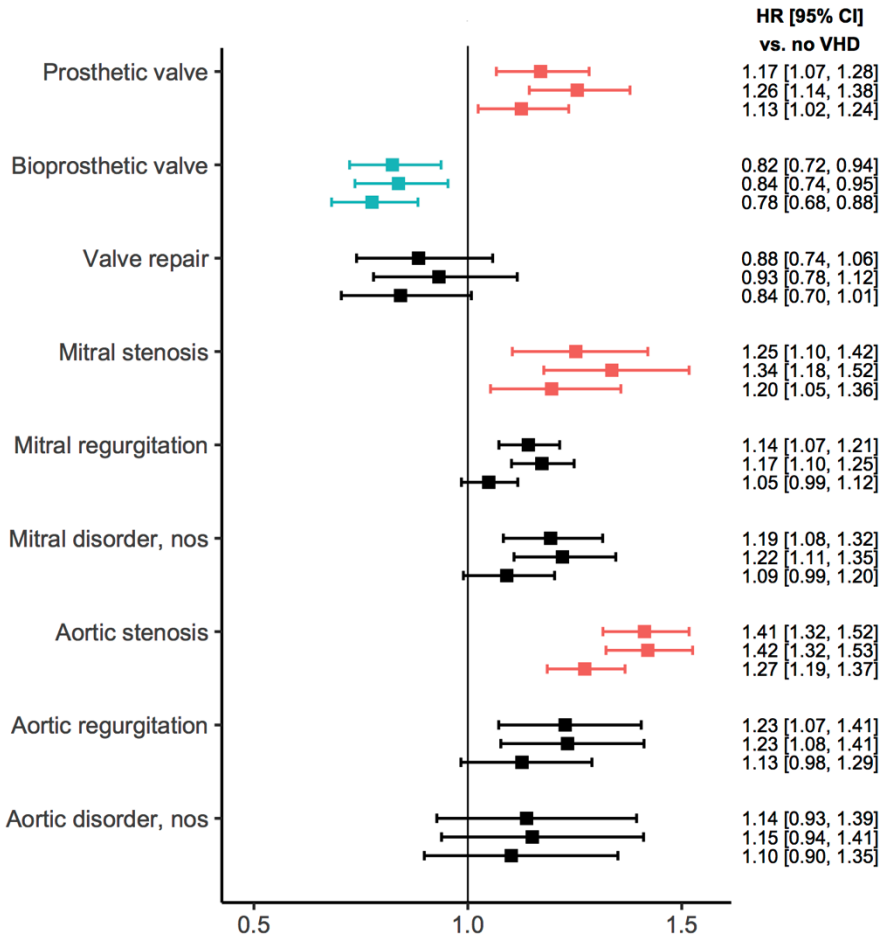


Figure 6.3 Plot of incrementally adjusted hazard ratios for associations of baseline valvular heart diseases with incident stroke, systemic embolism, and mortality



Notes: No record of valvular heart disease is always used as the reference category.

Abbreviations: HR [95%CI] – hazard ratio and 95% confidence interval, VHD – valvular heart disease, NOS – not otherwise specified.

Chapter 7

Net clinical benefit of warfarin in individuals with atrial fibrillation across stroke risk and across primary and secondary care

7.1 Chapter outline

In this final analysis chapter, I turn attention now to atrial fibrillation (AF) outcomes and present the results of an investigation using linked primary and secondary care electronic health records (EHRs) to estimate rates of ischaemic stroke by CHA_2DS_2-VASc scores, sex and use of warfarin. This work has been published in *Heart* (with CC-BY 4.0 license)[7] and was cited in the 2016 updates to European Society of Cardiology (ESC) guidelines for the management of AF.[3] Although one of the earlier completed projects of this PhD thesis, I have presented it last in order to reflect the natural transition between risk factors predisposing to AF, through to subtypes influencing AF treatment decisions and then on to AF-related outcomes aimed at being prevented. This project arose due the fact that, while it is clear there is a link between AF and stroke, the level of stroke risk upon which individuals should be recommended for preventative therapy with anticoagulants has been somewhat less clear. Whether individuals with just one factor from the CHA_2DS_2-VASc score for stroke risk stratification (which includes congestive heart failure, hypertension, age (≥75 years), diabetes mellitus, prior stroke, transient ischaemic attack or systemic embolism, vascular disease, age (65 to 75 years) and sex (female)[29]) is sufficient to warrant anticoagulants remains the topic of much debate.[4] I hypothesised that the stroke rate estimates I obtain from a population of individuals with AF diagnosed in one of two clinical settings (i.e. in either primary care or secondary care) may be different from prior reports, which have focussed exclusively on individuals with AF diagnosed in secondary care. Collaborator contributions for this work are reflected in the text.

List of collaborators: Amitava Banerjee, Anoop Dinesh Shah, Riyaz Patel, Spiros Denaxas, Juan-Pablo Casas, and Harry Hemingway

7.2 Abstract

Background The link between AF and subsequent stroke is well established but it remains unclear whether individuals with AF and one additional risk factor for stroke, defined as a CHA_2DS_2-VASc score equal to 1, benefit from stroke prevention with warfarin. Previous large-scale, population-based studies have reported incidence rates of stroke across levels of the CHA_2DS_2-VASc score however have neglected the full patient pathway by focussing exclusively on secondary care records. I therefore investigated the net clinical benefit of warfarin in individuals with AF across the levels of the CHA_2DS_2-VASc risk score using primary and secondary care records.

Methods I used CALIBER linked EHRs on 70,206 individuals with an initial record of diagnosis of AF in primary (n=29,568) or secondary care (n=40,638) in the UK between 1998 and 2010. For each individual I calculated baseline CHA_2DS_2-VASc scores, and followed them over a me-

dian 2.2 years for 7005 ischaemic strokes (IS) and for 906 haemorrhagic strokes (HS). I calculated incidence rates and 95% confidence intervals per 100 person-years (IR [95% CI] /100 PY) of IS and HS, with and without use of warfarin, and the net clinical benefit (i.e. number of IS avoided) per 100 person-years of warfarin use (NCB [95%CI] /100 PY).

Results Compared to individuals with initial record of diagnosis in secondary care, those in primary care had lower scores of IS risk (CHA_2DS_2-VASc≤2: 30.8% vs. 20.6%), and lower overall incidence of IS (IR [95% CI] /100 PY: 2.3 [2.2,2.4] vs. 4.3 [4.2,4.4]); however among individuals with CHA_2DS_2-VASc=0, 1 or 2 there were no differences in IS rate between those with initial record of diagnosis in primary care or secondary care (IR [95% CI] /100 PY: 0.2 [0.1,0.3] vs. 0.3 [0.2,0.5]), (IR [95% CI] /100 PY: 0.6 [0.4,0.7] vs. 0.7 [0.6,0.9]), and (IR [95% CI] /100 PY: 1.1 [1.00,1.3] vs. 1.4 [1.2,1.6]) respectively. For CHA_2DS_2-VASc=0, 1, and 2, incidence rates of IS with vs. without warfarin were (IR [95% CI] /100 PY: 0.4 [0.2,0.8] vs. 0.2 [0.1,0.3]), (IR [95% CI] /100 PY: 0.4 [0.3,0.7] vs. 0.7 [0.6,0.8]), and (IR [95% CI] /100 PY: 0.8 [0.7,1.0] vs. 1.4 [1.3,1.6]) respectively. I found a significant positive net clinical benefit of warfarin from CHA_2DS_2-VASc≥2 in men (NCB [95% CI] /100 PY: 0.5 [0.1,0.9]) and from CHA_2DS_2-VASc≥3 in women (NCB [95% CI] /100 PY: 1.5 [1.1,1.9]).

Conclusions CHA_2DS_2-VASc accurately stratifies ischaemic stroke risk in individuals with AF across both primary and secondary care. However as incidence rates of ischaemic stroke at CHA_2DS_2-VASc=1 are lower than previously reported this may change the decision to start anticoagulation with warfarin in these individuals.

7.3 Introduction

CHA_2DS_2-VASc (congestive heart failure, hypertension, age ≥75 years, diabetes mellitus, history of stroke or thromboembolism, vascular disease, age 65–74 years, and female sex) is the most widely used and validated clinical prediction score for assessment of ischaemic stroke risk in individuals with AF. The score ranges from 0–9, and assigns 1 or 2 points for each stroke risk factor.[29] CHA_2DS_2-VASc aims to identify individuals with the lowest stroke risk (CHA_2DS_2-VASc=0) in whom prevention with anticoagulants is not advised. The advice for individuals with one stroke risk factor (CHA_2DS_2-VASc=1) varies across United Kingdom (UK), European and US clinical practice guidelines, as it is debated whether the benefits of anticoagulants outweigh the harms. [3 19 22]

A recent systematic review of incidence rates of ischaemic stroke in individuals with CHA_2DS_2-VASc=1 reported highly heterogeneous annual rates ranging from 0.1% to 6.6% across studies, with wide uncertainty (0% to 3.23%) in the pooled estimate of 1.6%.[4] Among the 10 included studies, there were 0 studies that involved individuals with AF from across both primary and secondary care. Primary care, as Morley and colleagues showed, accounts for over 40% of initial AF diagnoses.[49] Therefore without the inclusion of these individuals it is unclear whether previously reported stroke rates are representative of the full patient pathway. In many countries (including the UK, Denmark, and Sweden) clinical information coded and recorded electronically

in secondary care is not integrated with that of primary care.[5] Thus, without linked information on risk factors diagnosed in primary care, previous studies based exclusively on secondary care data may have inaccurately calculated individuals' CHA_2DS_2-VASc scores (**figure 7.1**).[250] Moreover, without conducting net clinical benefit analyses, which weigh up benefits and harms of anticoagulants, the review was limited in the extent to which it could inform recommendations for clinical practice guidelines.

I therefore implemented the CHA_2DS_2-VASc score in linked CALIBER records on 70,206 individuals with an initial record of diagnosis of non-valvular AF in primary or secondary care in the UK in 1998 to 2010. The observation period is prior to the introduction of direct oral anticoagulants (DOACs), allowing outcomes with warfarin to be more accurately studied. My two main objectives were: (1) to investigate incidence of ischaemic stroke in individuals with AF across stroke risk and across primary and secondary care, and (2) to consider benefits and harms of warfarin in net clinical benefit analyses.

7.4 Methods

7.4.1 Analysis dataset

Data sources

I used CALIBER data connecting primary care, secondary care and mortality records for a subset of the UK population[5] that is representative of the overall population in terms of age, sex, ethnicity[198] and mortality.[53] Data are coded using four classification systems: Read (primary care diagnoses and procedures),[163] British National Formulary (BNF; primary care prescriptions),[164] ICD-10 (secondary care diagnoses and cause-specific mortality),[45] and OPCS-4 (secondary care procedures).[165]

Selection of individuals with non-valvular atrial fibrillation

I used Morley's definition for identifying AF in primary and secondary care linked EHRs to select all individuals with a diagnosis in primary or secondary care between 1998 and 2010. [49] To focus on non-valvular AF, as in recent DOAC trials,[31-34] I excluded individuals with codes for mitral valve diseases and prosthetic valve replacements which is based on current consensus definitions.[6]

CHA_2DS_2-VASc

I generated baseline CHA_2DS_2-VASc scores for each individual by assigning 1 or 2 points for each stroke risk factor.[29] Risk factors were defined according to existing CALIBER code lists (as available in published papers or at caliberresearch.org), which utilise all available clinical information across the linked primary and secondary care records. Briefly, age and sex[181] were obtained from general practice registration information; heart failure,[209] diabetes mellitus (type I, II and unclassified),[184] history of stroke or thromboembolism (stroke, transient ischaemic attack, or systemic embolism), and vascular disease (myocardial infarction[213] or peripheral artery disease) from clinical diagnosis codes; and hypertension[148] from diagnosis codes, two or more values of systolic or diastolic blood pressure measurements above UK diagnostic thresholds of

140/90mmHg,[150] or through repeat prescriptions for blood pressure lowering medications. The code list for systemic embolism, which I, together with Dr Anoop Shah, newly derived for this analysis, is provided in **table S7.1** in **appendix.**

Warfarin

I considered warfarin use throughout the entire study period and extracted information on prescriptions and International Normalised Ratio (INR) tests from primary care records. Prescriptions data includes drug type and date of administration but does not have information on whether medications were collected. Individuals were considered to be using warfarin continuously during follow-up if a prescription or INR test was administered every 30 days. This allowed each individual's follow-up time to be divided into periods with and without use of warfarin.[251]

Follow-up and endpoints

Individuals entered the study at their earliest coded diagnosis of AF during January 1998 and March 2010, provided they were aged 18 years and over and with a minimum of 1 year of continuous registration at a general practice with research quality data. I followed individuals for clinical diagnoses of ischaemic and unclassified strokes, and for haemorrhagic strokes (intracerebral and subarachnoid haemorrhage) as recorded in primary or secondary care and mortality registry records. Ischaemic and unclassified strokes were combined, as it has previously been shown that up to 90% of all strokes are ischaemic.[252] A clearance period of two weeks was imposed from date of recorded AF diagnosis, such that any stroke occurring during this time was attributed to baseline risk and not counted as an endpoint. The rationale is that AF is commonly first detected when an individual presents with a complication, such as stroke.[253] Clearance periods are important when analysing linked EHRs to avoid double-counting, as the same event can be recorded more than once, and in multiple data sources.[213] Total follow-up time for each individual was calculated as the number of days from the end of the clearance period to the point of censoring. Individuals were censored in the event of an ischaemic or haemorrhagic stroke endpoint, death (from a cause other than a stroke endpoint), transfer out of general practice, or last date of data collection.

7.4.2 Statistical analysis

I compared baseline CHA_2DS_2-VASc risk factors in individuals with initial record of diagnosis in primary and secondary care using number and percentages (%) for categorical variables, and means (standard deviations (SD)) and medians (range, interquartile range (IQR)) for continuous variables, as appropriate. I assessed the completeness of recording CHA_2DS_2-VASc risk factors in each data source using absolute proportions, i.e. number (%) of total cases captured in primary care records, and in secondary care records, compared to both data sources linked. I calculated incidence rates and 95% confidence intervals per 100 person-years (IR [95% CI] /100 PY]) for ischaemic and haemorrhagic stroke by dividing the number of endpoints by the accrued number of person-years. I assessed whether incidence rates were robust by comparing with estimates adjusted for propensity score quintiles.[254] This accounts for whether individuals were more or less likely to receive treatment with warfarin. I conducted net clinical benefit analyses

comparing number of ischaemic strokes (IS) avoided, against number of haemorrhagic strokes (HS) experienced per 100 person-years of warfarin use (NCB [95%CI] /100 PY). I used the formula: (IS rate $_{\text{without warfarin}}$ - IS rate $_{\text{with warfarin}}$) – 1.5(HS rate $_{\text{with warfarin}}$ - HS rate $_{\text{without warfarin}}$), whereby a positive estimate indicates a treatment benefit, and a negative estimate indicates treatment harm.[255] I regarded the net clinical benefit as significant if the 95% CI did not span both the positive and negative scale. All analyses were conducted in Stata/SE 13 and figures generated in R (version 3.2.0).

Sensitivity analyses

To facilitate comparisons between the ischaemic stroke rates I report and that of prior studies, I conducted a series of sensitivity analyses. These were predominantly in relation to men with a CHA_2DS_2-VASc=1. I tested the impact on stroke rates when (1) using CHA_2DS_2-VASc scores derived from clinical information captured exclusively in primary care vs. clinical information captured exclusively in secondary care, (2) including wider thromboembolic events (i.e. systemic embolism, pulmonary embolism and transient ischemic attack) in the ischaemic stroke endpoint definition, and (3) considering each 1 point scoring risk factor individually (i.e. heart failure, hypertension, diabetes mellitus, vascular disease, and age 65 to 74) as opposed to analysing them as one group. I also considered the impact of each 1 point scoring risk factor in women with CHA_2DS_2-VASc=2.

7.5 Results

Population characteristics

Population overall

The overall study population comprised 70,206 individuals with non-valvular AF with median age of 77.9 years (range: 18.0–108.7, IQR: 15.1), and median follow-up of 2.20 years (range: 0.03–12.2, IQR: 4.2). 34,286 (48.8%) individuals were women, and 2486 (3.5%) had CHA_2DS_2-VASc=0, 5637 (8.0%) had CHA_2DS_2-VASc=1, and 9339 (13.3%) had CHA_2DS_2-VASc=2. The mean (SD) CHA_2DS_2-VASc score of the population overall was 3.7 (1.8).

Individuals with initial record of diagnosis in primary vs. secondary care

29,568 (42.1%) individuals had initial record of diagnosis of AF in primary care, and 40,638 (57.9%) had initial record of diagnosis in secondary care. Individuals with initial record of diagnosis in secondary care were older (median (IQR) age: 79.1 (15.0) vs. 76.5 (14.8) years), more likely to be female (50.1% vs. 47.1%), and were less likely to have CHA_2DS_2-VASc=0 (2.9% vs. 4.4%), CHA_2DS_2-VASc=1 (6.6% vs. 10.1%), or CHA_2DS_2-VASc=2 (11.1% vs. 6.6%) than those with initial record of diagnosis in primary care. The mean (SD) CHA_2DS_2-VASc score among individuals with initial record of diagnosis of AF in primary compared to secondary care was 3.3 (1.7) vs. 4.0 (1.8). As **table 7.1** shows individuals with initial record of diagnosis in secondary care had a higher proportion of all CHA_2DS_2-VASc risk factors.

Individuals with vs. without use of warfarin

30,067 (42.8%) individuals underwent at least one period of warfarin use during follow-up; 50.1% of these had initial record of diagnosis in primary care (n=15,077), and 49.9% had initial record of diagnosis in secondary care (n=14,990). Individuals without use of warfarin were older (median (IQR) age: 80.7 (15.2) vs. 74.9 (13.4) years) and had a higher proportion of heart failure, but there was no difference in diagnosed hypertension, vascular disease (MI or PAD), diabetes, and previous strokes, when compared to those with at least one period of warfarin use (**table S7.2** of **appendix**).

Men vs. women

Women were older (median (IQR) age: 80.8 (13.3) vs. 74.9 (15.9) years) and had a higher proportion of heart failure, hypertension and previous strokes, while vascular disease (myocardial infarction and peripheral artery disease) and diabetes were more common in men (**table S7.3** of **appendix**).

Completeness of risk factors and reclassification of CHA_2DS_2-VASc scores

The completeness of recording CHA_2DS_2-VASc risk factors in primary and secondary care records ranged from 40.4% to 73.7% complete in secondary care records, and from 69.1% to 99.0% in primary care records (**table S7.4** and **figure S7.1** of **appendix**). Among individuals with initial record of diagnosis in secondary care, 975 (45.2%) were reclassified from CHA_2DS_2-VASc=0 to CHA_2DS_2-VASc≥1, and 2172 (53.7%) from CHA_2DS_2-VASc=1 to CHA_2DS_2-VASc≥2 when scores were calculated using linked primary–secondary care records. Only 15 (1.1%) individuals with initial record of diagnosis in primary care were reclassified from CHA_2DS_2-VASc=0 to CHA_2DS_2-VASc≥1, and 81 (2.7%) from CHA_2DS_2-VASc=1 to CHA_2DS_2-VASc≥2. For CHA_2DS_2-VASc scores calculated based on primary and secondary care records compared to both data sources linked, see **table S7.5** of **appendix**.

Stroke incidence

Ischaemic stroke rates in individuals with initial record of diagnosis in primary or secondary care

7005 ischaemic strokes occurred over 216,446 person-years, with IR [95% CI] /100 PY of 3.2 [3.2,3.3]. Compared to individuals with initial record of diagnosis in secondary care, those in primary care had lower overall ischaemic stroke incidence (IR [95% CI] /100 PY: 2.3 [2.2,2.4] vs. 4.3 [4.2,4.4]), however as **figure 7.2** shows there were no differences in incidence at CHA_2DS_2-VASc=0 (IR [95% CI] /100 PY: 0.2 [0.1,0.3] vs. 0.3 [0.2,0.5]), CHA_2DS_2-VASc=1 (IR [95% CI] /100 PY: 0.6 [0.4,0.7] vs. 0.7 [0.6,0.9]), or CHA_2DS_2-VASc=2 (IR [95% CI] /100 PY: 1.1 [1.00,1.3] vs. 1.4 [1.2,1.6]). Incidence rates in individuals with initial record of diagnosis in primary or in secondary care across all CHA_2DS_2-VASc scores are provided in **table 7.2**.

Ischaemic stroke rates by warfarin use

1015 (14.5%) ischaemic strokes occurred over 59,006 PY of warfarin use and 5990 (85.5%) over 157,439 PY of no warfarin use. Incidence rates were lower with warfarin use (IR [95% CI] /100 PY: 1.7 [1.6,1.8] vs. 3.8 [3.7,3.9]), with an incidence rate ratio [95% CI] of 0.5 [0.4,0.5]. For

CHA_2DS_2-VASc=0, CHA_2DS_2-VASc=1 and CHA_2DS_2-VASc=2, incidence rates with vs. without use of warfarin were (IR [95% CI] /100 PY: 0.4 [0.2,0.8] vs. 0.2 [0.1,0.3]), (IR [95% CI] /100 PY: 0.4 [0.3,0.7] vs. 0.7 [0.6,0.8]), and (IR [95% CI] /100 PY: 0.8 [0.7,1.0] vs. 1.4 [1.3,1.6]). As **figure 7.3** and **figure 7.4** show incidence rates were lower with use of warfarin from CHA_2DS_2-VASc≥2 in men (IR [95% CI] /100 PY: 0.9 [0.7,1.1] vs. 1.7 [1.5,1.9]), and from CHA_2DS_2-VASc≥3 in women (IR [95% CI] /100 PY: 0.7 [0.5,1.0] vs. 2.3 [2.0,2.5]). Incidence rates by sex and use of warfarin across all CHA_2DS_2-VASc scores are provided in **table 7.3.** Incidence rates adjusted for propensity score quintiles were consistent with the unadjusted estimates (see **table S7.6** of **appendix**).

Net clinical benefit of warfarin

906 haemorrhagic strokes occurred over 224,777 PY. The overall net clinical benefit of warfarin was 1.9 [1.8,2.1] ischaemic strokes avoided per 100 PY. For CHA_2DS_2-VASc=0, CHA_2DS_2-VASc=1 and CHA_2DS_2-VASc=2, net clinical benefit was (NCB [95% CI] /100 PY: -0.3 [-0.8,0.1]), (NCB [95% CI] /100 PY: 0.1 [-0.2,0.4]), and (NCB [95% CI] /100 PY: 0.2 [-0.1,0.6]), respectively. A significant positive net clinical benefit was observed from CHA_2DS_2-VASc≥2 in men (NCB [95% CI] /100 PY: 0.5 [0.1,0.9]) and from CHA_2DS_2-VASc≥3 in women (NCB [95% CI] /100 PY: 1.5 [1.1,1.9]). Net clinical benefit estimates across all CHA_2DS_2-VASc scores, are provided in **table 7.4.**

Sensitivity analyses

Table 7.5 provides the results of sensitivity analyses conducted on stroke rate estimates for men with a CHA_2DS_2-VASc=1. Compared to the estimate of 0.7 [0.6,0.8] obtained in the main analysis, stroke rates were higher when CHA_2DS_2-VASc scores were derived from clinical information captured in secondary care records (IR [95% CI] /100 PY: 1.4 [1.2,1.6]) and when systemic embolism, pulmonary embolism and transient ischaemic attack were added to the endpoint definition (IR [95% CI] /100 PY: 1.4 [1.2,1.5]). Stroke rates, considering each 1 point scoring risk factor individually, also differed with age 65 to 74 conferring the highest estimate (IR [95% CI] /100 PY: 1.2 [0.9,1.5]) in men with a CHA_2DS_2-VASc=1. Stroke rates, considering each 1 point scoring risk factor individually, in women with CHA_2DS_2-VASc=2 were: 0.8 [0.6,1.2] for age 65 to 75, 1.4 [0.4,5.7] for heart failure and 0.6 [0.4,0.8] for hypertension; there were no events for diabetes mellitus or vascular disease.

7.6 Discussion

I conducted a large-scale linked EHR study of the potential benefits and harms of warfarin in individuals with AF across stroke risk and across primary and secondary care with two major findings. First, I confirmed that CHA_2DS_2-VASc accurately stratifies stroke risk in individuals with an initial record of diagnosis in primary and secondary care, however clinical information recorded in both primary and secondary care must be considered in order to correctly assign CHA_2DS_2-VASc scores. Second, in individuals who were truly CHA_2DS_2-VASc=1, the absolute risk of ischaemic stroke (0.4 [0.3,0.7] with warfarin, and 0.7 [0.6,0.8] without warfarin) was relatively low and similar to the original derivation cohort of the CHA_2DS_2-VASc score,[29] and the net

clinical benefit of warfarin was positive but non-significant (0.1 [-0.2,0.4]). I therefore found insufficient evidence to support anticoagulation with warfarin in individuals with CHA_2DS_2-VASc=1.

Findings in context

The ischaemic stroke incidence rate I found of 0.7 [0.6,0.8] for CHA_2DS_2-VASc=1 without warfarin is low compared to previous reports (which are summarised in **table S7.7** of **appendix**)[7 29 253 256-264] although consistent with the levels of uncertainty in a recent systematic review which reported an annual rate of 1.6% but with wide confidence intervals of 0% to 3.23%.[4] To investigate the robustness of my estimates, and offer three explanations as to why they differ from previous studies, I conducted a series of sensitivity analyses in relation to men with CHA_2DS_2-VASc=1 (which are detailed in **table 7.5**). First, as demonstrated in the main analysis, secondary care records underestimate stroke risk in half of individuals, and therefore previous studies which have predominantly focused on secondary care populations may be biased by misclassification of CHA_2DS_2-VASc scores. In a sensitivity analysis of men estimated to have CHA_2DS_2-VASc=1 according to secondary care records (but which included 53% who were truly CHA_2DS_2-VASc≥2) I obtained an incidence rate of 1.4 [1.2,1.5], which is twice as high as the rate found for men who were truly CHA_2DS_2-VASc=1 (0.7 [0.6,0.8]) and similar to the meta-analysed rate found in the recent systematic review (1.6 [0,3.23]).[4] This confirms that if CHA_2DS_2-VASc risk is underestimated then stroke rates at the lower end of the CHA_2DS_2-VASc scale are overestimated, which has implications on treatment decisions. Second, Friberg and colleagues reported that variation in the literature also exists because of differences in the way stroke is defined, and that a 44% higher incidence rate is observed when including wider thromboembolic endpoints.[253] I also confirmed this in a sensitivity analysis, and found that for men with CHA_2DS_2-VASc=1, the incidence rate doubled from 0.7 [0.6,0.8] to 1.4 [1.2,1.5] when systemic embolism, pulmonary embolism and transient ischaemic attacks were included as a composite endpoint. Third, differences in previously reported stroke rates at CHA_2DS_2-VASc=1 may exist because not all 1 point scoring risk factors (heart failure, hypertension, diabetes mellitus, vascular disease, age 65 to 74, female sex) confer exactly the same stroke risk.[257] In sensitivity analyses considering each 1 point scoring risk factor separately, I found that age 65 to 74 years conferred the highest stroke risk with incidence rate of 1.2 [0.9,1.5]. Among women with CHA_2DS_2-VASc=2 (i.e. with 1 point scoring risk factor irrespective of sex), I found that heart failure conferred the higher stroke risk with incidence rate of 1.4 [0.4,5.7]; there were however very low numbers of events and therefore this finding should be interpreted as exploratory and with caution. Ultimately, population-based incidence rates of CHA_2DS_2-VASc=1 depend upon the distribution of 1 point scoring risk factors within the population.

Unlike the recent systematic review,[4] my findings with regard to CHA_2DS_2-VASc=1 are supported by net clinical benefit analyses.[255] To my knowledge, this is the first net clinical benefit analysis of warfarin to date that includes both individuals with an initial record of diagnosis of AF in primary and secondary care. I found a positive but non-significant treatment benefit of warfarin in individuals with CHA_2DS_2-VASc=1 (0.1 [-0.2,0.4]), and therefore insufficient evidence to sup-

port anticoagulation with warfarin in these individuals. Existing net clinical benefit analyses of warfarin have predominantly focussed on individuals with initial record of diagnosis in secondary care, but these have also shown an unclear benefit of treatment at CHA_2DS_2-VASc=1, including in both the nationwide Danish (-0.02 [-0.15,0.11]),[260] and Swedish (0.00 [not reported])[265] cohorts.

Direct oral anticoagulants

A question that remains is whether the newer DOACs (dabigatran, rivaroxaban, apixaban, and edoxaban) have a role in the treatment of lower risk individuals. In DOAC trials (e.g. RE-LY,[31] ROCKET-AF,[32] ARISTOTLE,[33] and ENGAGE AF-TIMI 48[34]) all four agents were shown to be as effective in preventing ischaemic strokes as warfarin, and associated with fewer haemorrhagic strokes (**table S7.8** of **appendix**). In the absence of DOAC data, I therefore applied the trial reported relative risks of ischaemic and haemorrhagic stroke to the available data to consider the net clinical benefit of DOACs compared to no treatment. I found a significant positive net clinical benefit at CHA_2DS_2-VASc=1 across all agents (**table S7.9** of **appendix**), and therefore some, albeit extrapolated evidence that DOACs may be a more suitable treatment option for those at lower stroke risk. I caution against interpreting this finding too strongly however, as the extrapolation takes on multiple assumptions. First, it assumes that the relative risk reduction is constant over the entire period of follow-up, and across all levels of the CHA_2DS_2-VASc score. Second, it assumes that the trial population is representative of the general AF population. And third, it assumes that the benefit and harm endpoint definitions are equivalent. All four DOAC trials were approximately 2 years in duration, which is comparable to the 2.2 median years of follow-up in this analysis. However based on trial eligibility criteria, patients at CHA_2DS_2-VASc=1 were largely excluded from DOAC trials, and as previously demonstrated these trials represented only a half to two thirds of the AF population in the UK.[192] Lastly, benefit and harm endpoint definitions did vary across trials (**table S7.8** of **appendix**).

Clinical implications

I confirmed that CHA_2DS_2-VASc is valid for estimating stroke risk in individuals with an initial record of diagnosis in primary and secondary care, and therefore advocate use of the score in both of these clinical settings. While I showed that lack of integration of primary and secondary care information may lead to inaccurate CHA_2DS_2-VASc scores, I do not regard this as an issue at point-of-care as clinicians are able to verify stroke risk factors with patients directly. However with the advent of clinical decision support systems,[266] greater integration of primary and secondary care, and better recording of risk factors is urgently required in order to avoid undue patient harm through underestimation of stroke risk. Finally, I found insufficient evidence to support stroke prevention with warfarin at CHA_2DS_2-VASc=1, which has relevance for treatment guidelines.

Research implications

The findings of this work highlight the value of linked EHRs in investigating individuals with AF across the full clinical pathway of primary and secondary care, and crucially in identifying indi-

viduals with CHA_2DS_2-VASc=1 in whom treatment guidelines have so far been unclear.[4] As shown, primary care records were instrumental in identifying individuals with CHA_2DS_2-VASc=1 and I therefore propose wider usage of existing and discovery of new primary care data sources for studying these individuals in future research, and in particular in 'real-world' comparative effectiveness studies of DOACs. This is something that will likely be achieved in the recently launched Innovative Medicines Initiative BigData@Heart programme which aims to unlock large-scale clinical and cardiovascular-related datasets collected across Europe.[39]

Strengths and limitations

The study's principal strength was the inclusion of individuals, risk factors and endpoints recorded across both primary and secondary care. Unlike prior studies, which have predominantly focussed on secondary care,[4] the inclusion of primary care records in this analysis meant that I ascertained a more representative sample of the overall AF population and was able to more accurately calculate CHA_2DS_2-VASc stroke risks by factoring in clinical information recorded exclusively in each setting. The completeness of recording clinical information in electronic heart records remains a wider ongoing issue[48] and thus as well as analysing data from multiple sources to minimise the impact of this, I also used established CALIBER definitions for determining AF cases, CHA_2DS_2-VASc risk factors and stroke endpoints.[51] An example of the value CALIBER definitions, which combine relevant diagnosis codes together with plausible inferences, comes from hypertension. As shown the proportion of individuals with baseline hypertension rose from 59.7% to 82.3% when recorded diagnoses were supplemented with inferred diagnoses from repeated blood pressure measurements and prescriptions for blood pressure lowering medications. It is however still possible that some individuals, risk factors and endpoints may have been overlooked, but this number is likely to be minimal and certainly less than previous studies. Another limitation of this work is the lack of DOAC data. This was an unfortunate consequence of recent challenges in accessing EHRs for research[169] and thus beyond my direct control. One advantage of using data prior to the introduction of DOACs means that outcomes with warfarin can be more independently studied and used as a benchmark for studies post-approval of DOACs. Warfarin is still widely used in clinical practice today and is particularly important in the treatment of individuals with valvular forms of AF.[247] It should be remarked that, while I had access to a large sample size of 70,000 individuals, I found that only a quarter of the overall population had CHA_2DS_2-VASc scores 0, 1, and 2. I therefore cannot rule out the possibility that a positive net clinical benefit of warfarin may be observed at CHA_2DS_2-VASc=1 given a larger study population. Finally, as with all observational epidemiology unmeasured confounder limits the interpretation of study findings. Whether individuals with CHA_2DS_2-VASc=1 require anticoagulation could be determined in the randomised control setting, however given that the net-clinical benefit from observational data suggests otherwise, it would be unethical to conduct such a study.

7.7 Conclusion

CHA_2DS_2-VASc accurately stratifies ischaemic stroke risk in individuals with AF across both primary and secondary care. However incidence rates of ischaemic stroke at CHA_2DS_2-VASc=1

are lower than previously reported, which may change the decision to start anticoagulation with warfarin in these individuals.

7.8 Chapter summary

To summarise, in this chapter I used primary and secondary care EHRs to estimate rates of ischaemic stroke by CHA_2DS_2-VASc scores, sex and use of warfarin highlighting the value of linked clinical information in ascertaining a more representative sample of the overall AF population and accurately calculating stroke risks. The work of this chapter has been published in *Heart*[7] journal and was cited the 2016 updates to ESC guidelines for the management of AF in relation to treatment recommendations for individuals at a low stroke risk (i.e. CHA_2DS_2-VASc=1).[3] This shows that EHR data sources can not only respond to changes in clinical guidelines (as I did in developing EHR definitions relevant to the new AF subtypes distinctions in **chapter 5** and **chapter 6**) but also have an important role in influencing guidelines as well. As this was the final analytical chapter, in **chapter 8**, which follows, I recap on the objectives of chapters one to seven before presenting an overall discussion of novel contributions, strengths, limitations and conclusions of this thesis using EHRs to study AF risk factors, subtypes and outcomes.

7.9 Chapter tables

Table 7.1 Comparison of baseline CHA_2DS_2-VASc risk factors in individuals with atrial fibrillation and initial record of diagnosis in primary or secondary care

	Initial record of diagnosis					
	Primary care		Secondary care		Population overall	
Number of individuals	**29 568**		**40 638**		**70 206**	
	N	**%**	**N**	**%**	**N**	**%**
Congestive heart failure	4768	16.1	12664	31.2	17432	24.8
Hypertension	23946	81.0	33817	83.2	57763	82.3
Diagnosis	16616	56.2	25273	62.2	41889	59.7
Blood pressure medication	20979	71.0	30164	74.2	51143	72.9
Blood pressure measures	17281	58.4	22329	55.0	39610	56.4
Age ≥ 75 [**2**]	16318	55.2	25872	63.7	42190	60.1
Diabetes	3316	11.2	6673	16.4	9989	14.2
Stroke/TIA/systemic embolism [**2**]	3887	13.2	8938	22.0	12825	18.3
Vascular disease	4049	13.7	9783	24.1	13832	19.7
Myocardial infarction	2635	8.9	6950	17.1	9585	13.7
Peripheral vascular disease	1776	6.0	3927	9.7	5703	8.1
Age 65–74	7744	26.2	8552	21.0	16296	23.2
Sex **C**ategory [female]	13930	47.1	20356	50.1	34286	48.8
CHA_2DS_2-VASc scores						
0	1305	4.4	1181	2.9	2486	3.5
1	2972	10.1	2665	6.6	5637	8.0
2	4820	16.3	4519	11.1	9339	13.3
3	6663	22.5	7107	17.5	13770	19.6
4	7332	24.8	9578	23.6	16910	24.1
5	3712	12.6	7514	18.5	11226	16.0
6	1866	6.3	4906	12.1	6772	9.7
7	724	2.5	2341	5.8	3065	4.4
8	156	0.5	707	1.7	863	1.2
9	18	0.1	120	0.3	138	0.2

Abbreviations: N – number.

Table 7.2 Incidence rates [95% confidence intervals] per 100 person-years of ischaemic stroke in individuals with atrial fibrillation by CHA_2DS_2-VASC scores and initial record of diagnosis in primary or secondary care

	Initial record of diagnosis					
	Primary care		Secondary care		Population overall	
CHA_2DS_2-VASc	Events	Rate	Events	Rate	Events	Rate
0	12	0.2 [0.1,0.3]	16	0.3 [0.2,0.5]	28	0.2 [0.2,0.4]
1	77	0.6 [0.4,0.7]	76	0.7 [0.6,0.9]	153	0.6 [0.5,0.7]
2	244	1.1 [1.0,1.3]	209	1.4 [1.2,1.6]	453	1.2 [1.1,1.3]
3	528	2.0 [1.8,2.2]	485	2.4 [2.2,2.6]	1013	2.2 [2.0,2.3]
4	766	2.9 [2.7,3.2]	907	3.9 [3.7,4.2]	1673	3.4 [3.2,3.6]
5	450	3.9 [3.5,4.2]	966	6.5 [6.1,6.9]	1416	5.3 [5.1,5.6]
6	332	6.4 [5.7,7.1]	1028	12.0 [11.3,12.8]	1360	9.9 [9.4,10.4]
7	146	7.7 [6.6,9.1]	546	14.8 [13.6,16.1]	692	12.4 [11.5,13.3]
8	34	9.3 [6.6,12.9]	151	15.8 [13.5,18.5]	185	13.9 [12.1,16.1]
9	5	13.4 [5.6,32.3]	27	22.1 [15.2,32.3]	32	20.0 [14.2,28.4]
0-9	**2594**	**2.3 [2.2,2.4]**	**4411**	**4.3 [4.2,4.42]**	**7005**	**3.2 [3.2,3.3]**

Table 7.3 Incidence rates [95% confidence intervals] per 100 person-years of ischaemic stroke in individuals with atrial fibrillation by CHA_2DS_2-VASC scores, sex, and warfarin.

CHA_2DS_2-VASc	With warfarin		Without warfarin		Population overall	
Overall population	Events	Rate	Events	Rate	Events	Rate
0	7	0.4 [0.2,0.8]	21	0.2 [0.1,0.3]	28	0.2 [0.2,0.4]
1	27	0.4 [0.3,0.7]	126	0.7 [0.6,0.8]	153	0.6 [0.5,0.7]
2	87	0.8 [0.7,1.0]	366	1.4 [1.3,1.6]	453	1.2 [1.1,1.3]
3	144	1.0 [0.9,1.2]	869	2.6 [2.5,2.8]	1013	2.2 [2.0,2.3]
4	226	1.7 [1.5,2.0]	1447	4.0 [3.8,4.2]	1673	3.4 [3.2,3.6]
5	233	3.2 [2.8,3.6]	1183	6.2 [5.8,6.5]	1416	5.3 [5.1,5.6]
6	159	4.2 [3.6,4.8]	1201	12.1 [11.4,12.8]	1360	9.9 [9.4,10.4]
7	111	7.1 [5.9,8.6]	581	14.5 [13.4,15.7]	692	12.4 [11.5,13.4]
8	18	4.8 [3.0,7.6]	167	17.6 [15.1,20.5]	185	14.0 [12.1,16.2]
9	3	7.5 [2.4,23.3]	29	24.3 [16.9,35.0]	32	20.1 [14.2,28.4]
0-9	**1015**	**1.7 [1.6,1.8]**	**5990**	**3.8 [3.7,3.9]**	**7005**	**3.2 [3.2,3.3]**
Men						
0	7	0.4 [0.2,0.8]	21	0.2 [0.1,0.3]	28	0.2 [0.2,0.4]
1	25	0.5 [0.3,0.7]	112	0.8 [0.6,0.9]	137	0.7 [0.6,0.8]
2	75	0.9 [0.7,1.1]	306	1.7 [1.5,1.9]	381	1.5 [1.3,1.6]
3	106	1.3 [1.1,1.5]	550	2.9 [2.7,3.2]	656	2.4 [2.2,2.6]
4	119	2.2 [1.8,2.6]	489	4.3 [3.9,4.7]	608	3.6 [3.3,3.9]
5	122	4.1 [3.4,4.9]	491	7.8 [7.1,8.5]	613	6.6 [6.1,7.1]
6	51	4.1 [3.1,5.4]	312	11.5 [10.3,12.9]	363	9.2 [8.3,10.1]
7	27	6.1 [4.3,9.2]	129	14.5 [12.2,17.3]	156	11.9 [10.1,13.9]
8	1	1.9 [0.3,13.3]	20	14.5 [9.4,22.5]	21	11.0 [7.2,16.8]
0-8	**533**	**1.6 [1.4,1.7]**	**2430**	**2.9 [2.8,3.1]**	**2963**	**2.5 [2.4,2.6]**
Women						
1	2	0.4 [0.1,1.5]	14	0.4 [0.2,0.7]	16	0.4 [0.3,0.7]
2	12	0.5 [0.3,0.9]	60	0.7 [0.6,0.9]	72	0.7 [0.5,0.8]
3	38	0.7 [0.5,1.0]	319	2.3 [2.0,2.5]	357	1.8 [1.6,2.0]
4	107	1.4 [1.1,1.7]	958	3.9 [3.7,4.1]	1065	3.3 [3.1,3.5]
5	111	2.5 [2.1,3.0]	692	5.4 [5.0,5.8]	803	4.7 [4.4,5.0]
6	108	4.2 [3.5,5.0]	889	12.3 [11.5,13.1]	997	10.2 [9.5,10.8]
7	84	7.4 [6.0,9.2]	452	14.5 [13.2,15.9]	536	12.6 [11.6,13.7]
8	17	5.3 [3.3,8.5]	147	18.1 [15.4,21.3]	164	14.5 [12.4,16.9]
9	3	7.5 [2.4,23.3]	29	24.3 [16.9,35.0]	32	20.1 [14.2,28.4]
1-9	**482**	**2.0 [1.8,2.1]**	**3560**	**4.8 [4.6,4.9]**	**4042**	**4.1 [3.9,4.2]**

Table 7.4 Net clinical benefit [95% confidence intervals] per 100 person-years with warfarin in individuals with atrial fibrillation by CHA_2DS_2-VASC scores and sex.

CHA_2DS_2-VASc scores	Total stroke events		Net clinical benefit
Overall population	**Ischaemic stroke**	**Haemorrhagic stroke**	
0	28	8	-0.3 [-0.8,0.1]
1	153	54	0.1 [-0.2,0.4]
2	453	95	0.2 [-0.1,0.6]
3	1013	165	1.5 [1.2,1.8]
4	1673	237	2.2 [1.8,2.6]
5	1416	180	3.2 [2.6,3.8]
6	1360	104	7.7 [6.7,8.8]
7	692	45	7.2 [5.2,9.1]
8	185	18	12.8 [8.9,16.9]
9	32	0	16.8 [1.8,31.5]
0-9	7005	906	1.9 [1.8,2.1]
Men			
0	28	8	-0.3 [-0.8,0.1]
1	137	48	0.1 [-0.2,0.4]
2	381	79	0.5 [0.1,0.9]
3	656	103	1.5 [1.1,1.9]
4	608	93	2.0 [1.3,2.7]
5	613	88	3.9 [2.6,4.9]
6	363	26	7.1 [5.2,9.1]
7	156	15	8.6 [4.7,13.0]
8	21	3	15.9 [7.0,25.7]
0-8	2963	463	1.2 [1.0,1.4]
Women			
1	16	6	0.3 [-0.4,0.8]
2	72	16	-0.1 [-0.6,0.3]
3	357	62	1.5 [1.1,1.9]
4	1065	144	2.4 [2.0,2.8]
5	803	92	3.1 [2.3,3.8]
6	997	78	8.0 [6.6,9.3]
7	536	30	6.8 [4.4,9.1]
8	164	15	12.4 [7.9,17.3]
9	32	0	16.8 [3.1,30.5]
1-9	4042	443	2.7 [2.4,3.0]

Table 7.5 Sensitivity analyses on ischaemic stroke incidence rates in men with CHA_2DS_2-VASc=1.

	Total patients	Initial AF diag.		CHA_2DS_2-VASc		Endpoint			Endpoint Definition					Total events	IR [95% CI] per 100 PY
		PC	SC	PC	SC	PC	SC	DR	IS	US	SE	PE	TIA		
By data source used for CHA_2DS_2-VASc:															
Secondary care	3015[a]	○	●	○	●	●	●	●	●	●	○	○	○	156	**1.4 [1.2,1.6]**
Primary care	2563[a]	●	○	●	○	●	●	●	●	●	○	○	○	72	0.6 [0.5,0.8]
Total	4690	●	●	●	●	●	●	●	●	●	○	○	○	137	0.7 [0.6,0.8]
By data source used for initial AF diagnosis:															
Secondary care	2185	○	●	●	●	●	●	●	●	●	○	○	○	68	0.8 [0.6,1.0]
Primary care	2505	●	○	●	●	●	●	●	●	●	○	○	○	69	0.6 [0.5,0.7]
Total	4690	●	●	●	●	●	●	●	●	●	○	○	○	137	0.7 [0.6,0.8]
By data source for endpoint:															
Death register	760[b]	●	●	●	●	○	○	●	●	●	○	○	○	14	0.6 [0.4,1.1]
Primary care	4690	●	●	●	●	○	●	○	●	●	○	○	○	106	0.5 [0.4,0.6]
Secondary care	4690	●	●	●	●	●	○	○	●	●	○	○	○	90	0.4 [0.4,0.5]
Total	4690	●	●	●	●	●	●	●	●	●	○	○	○	137	0.7 [0.6,0.8]
By endpoint definition:															
Composite endpoint	4690	●	●	●	●	●	●	●	●	●	●	●	●	274	**1.4 [1.2,1.5]**
Primary endpoint	4690	●	●	●	●	●	●	●	●	●	○	○	○	137	0.7 [0.6,0.8]
By use of anticoagulants:															
Without anticoagulants	4629[c]	●	●	●	●	●	●	●	●	●	○	○	○	112	0.8 [0.6,0.9]
With anticoagulants	2549[c]	●	●	●	●	●	●	●	●	●	○	○	○	25	0.4 [0.3,0.7]
Total	4690	●	●	●	●	●	●	●	●	●	○	○	○	137	0.7 [0.6,0.8]
By risk score component:															
Age 65 to 74	1336	●	●	●	●	●	●	●	●	●	○	○	○	68	**1.2 [0.9,1.5]**
Hypertension	3053	●	●	●	●	●	●	●	●	●	○	○	○	66	0.5 [0.4,0.6]
Heart failure	126	●	●	●	●	●	●	●	●	●	○	○	○	2	0.5 [0.1,1.8]
Vascular disease	110	●	●	●	●	●	●	●	●	●	○	○	○	1	0.2 [0.0,1.5]
Diabetes mellitus	65	●	●	●	●	●	●	●	●	●	○	○	○	0	No events
Total	4690	●	●	●	●	●	●	●	●	●	○	○	○	137	0.7 [0.6,0.8]

Notes:

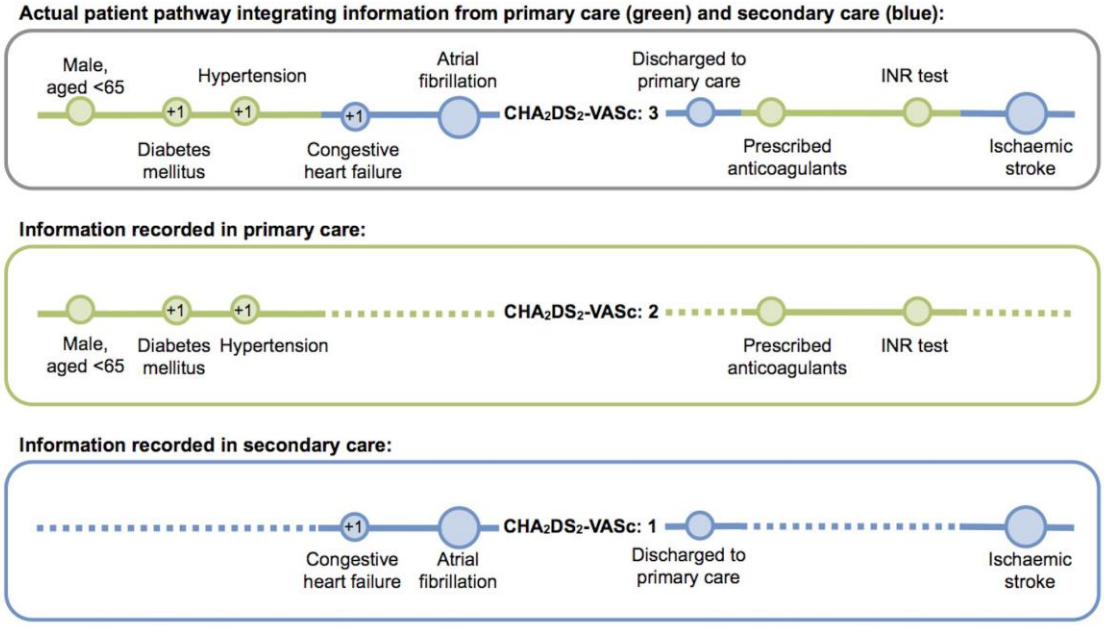
[a] Includes patients reclassified to CHA_2DS_2-VASc ≥2 when stroke risk calculated from linked records.

[b] restricted to 760 all-cause deaths during follow-up, [c] patients could contribute follow-up time to periods with and without anticoagulants.

Abbreviations: CI – confidence interval, PY – person-years, PC –primary care, SC – secondary care, DR – death registry, IS - ischaemic stroke, US – unclassified stroke, SE – systemic embolism, PE – pulmonary embolism, TIA - transient ischaemic attack, ○ – no, ● – yes. Initial AF diag. – refers to whether initial record of diagnosis of atrial fibrillation was in primary or secondary care.

7.10 Chapter figures

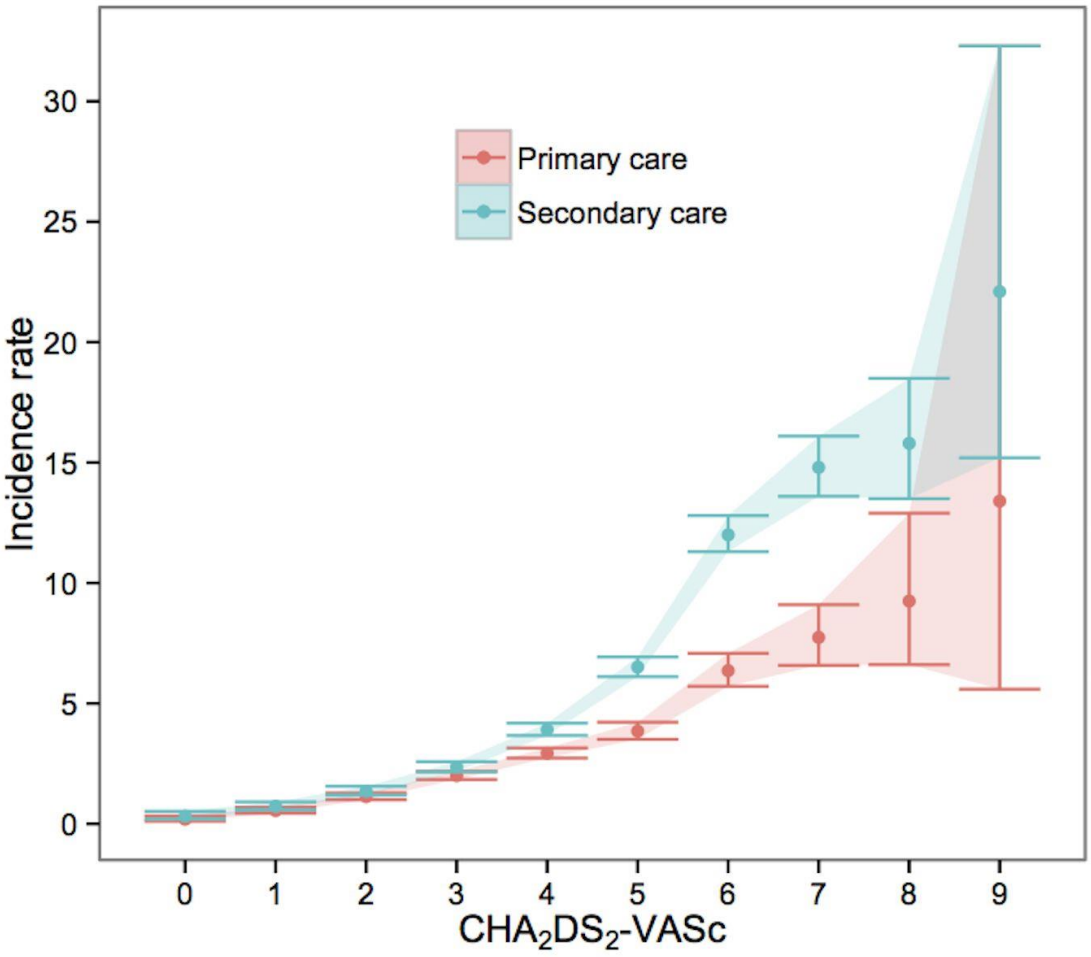
Figure 7.1 Hypothetical example of one patient's clinical pathway and how information could be captured exclusively in primary care or secondary care records. This illustrates how lack of integration of primary and secondary care information may lead to underestimation of CHA_2DS_2-VASc scores.



Notes: Patient interactions occurring in primary and secondary care are colour-coded green and blue respectively. +1 indicates scoring of one CHA_2DS_2-VASc risk factor point. As shown, this patient has CHA_2DS_2-VASc=1 based on secondary care records only, CHA_2DS_2-VASc=2 based on primary care records only, but is actually CHA_2DS_2-VASc=3 based on integrated primary and secondary care records.

Abbreviations: CHA_2DS_2-VASc – an acronym for the congestive heart failure, hypertension, age (≥75 years), diabetes mellitus, history of stroke, transient ischaemic attack or systemic embolism, vascular disease, age (65 to 75 year) and sex (female) risk score, INR - International Normalised Ratio.

Figure 7.2 Incidence rates [95% confidence intervals] per 100 person-years of ischaemic stroke in individuals with atrial fibrillation by CHA_2DS_2-VASC scores and initial record of diagnosis in primary or secondary care.



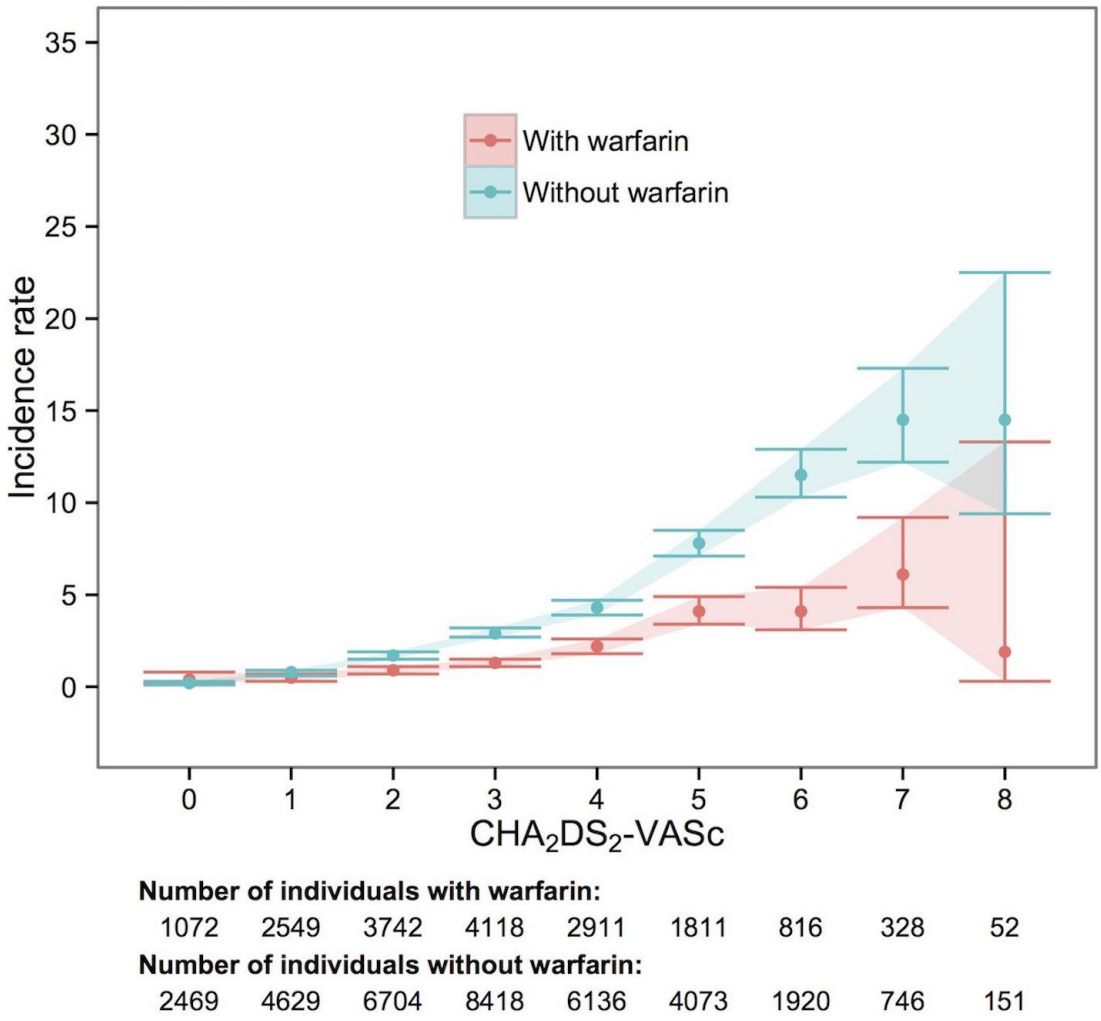
Number of individuals with initial record of diagnosis in primary care:

1305 2972 4820 6663 7332 3712 1866 724 156 18

Number of individuals with initial record of diagnosis in secondary care:

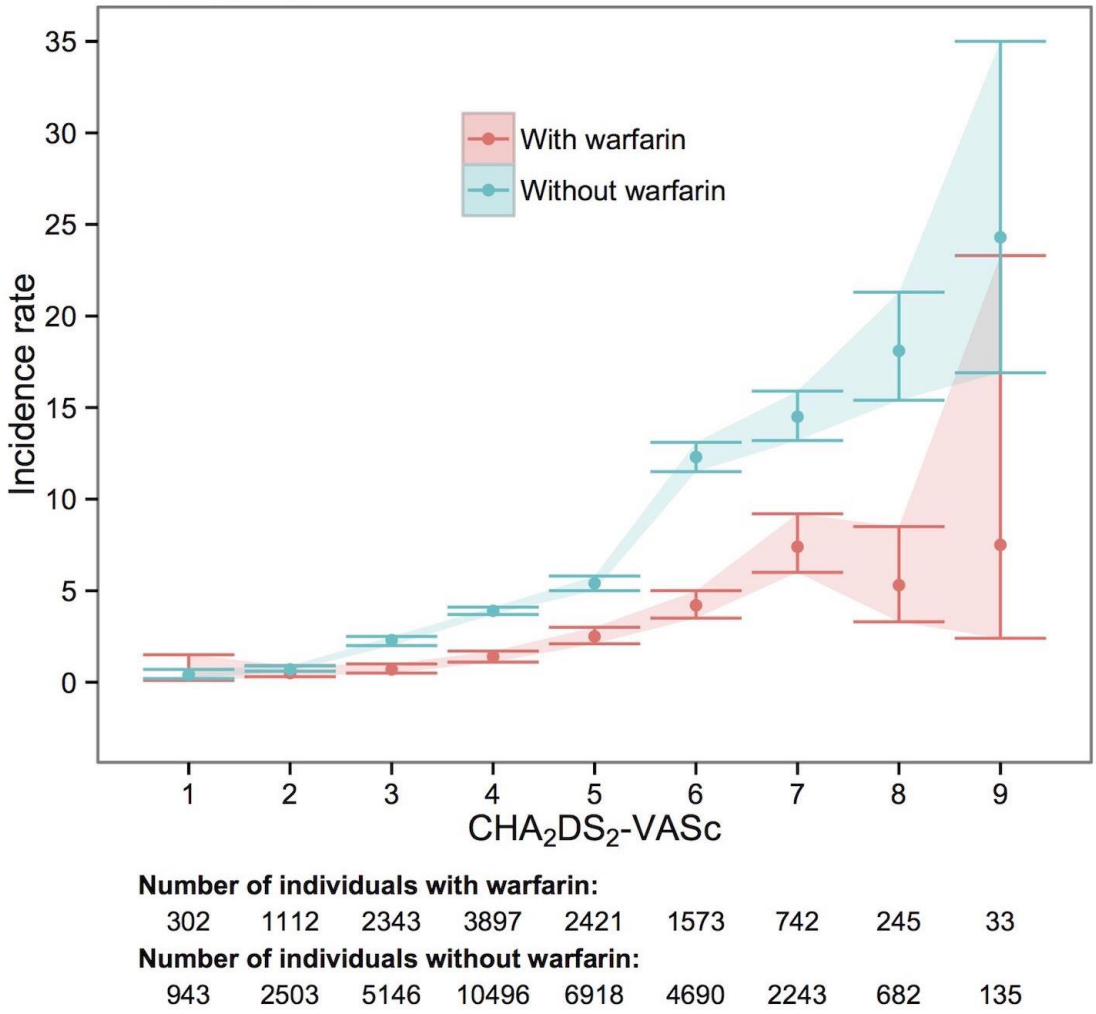
1181 2665 4519 7107 9578 7514 4906 2341 707 120

Figure 7.3 Incidence rates [95% confidence intervals] per 100 person-years of ischaemic stroke in men with atrial fibrillation by CHA_2DS_2-VASC scores, and use of warfarin.



Notes: Individuals could contribute follow-up time to periods with and without warfarin.

Figure 7.4 Incidence rates [95% confidence intervals] per 100 person-years of ischaemic stroke in women with atrial fibrillation by CHA_2DS_2-VASC scores, and use of warfarin.



Notes: Individuals could contribute follow-up time to periods with and without warfarin.

Chapter 8

Overall discussion of novel contributions, strengths, limitations and conclusion

8.1 Chapter overview

Over the previous seven chapters, I have both described and demonstrated the value of electronic health records (EHRs) in driving forward research on atrial fibrillation (AF) risk factors, subtypes and outcomes. In this overall discussion chapter, I start by recapping on the thesis objectives of chapters one to seven, I go on to reflect upon novel contributions including 12 recommendations for AF clinical practice and research, and I then describe overall strengths and limitations of the data and methods used, as well as emerging techniques which have not been used in the thesis but are of increasing relevance to the field of AF research. Finally, I present my conclusions.

8.2 Recap of thesis objectives

Ahead of presenting the novel contributions of this PhD thesis it helps to first recap on what was achieved in chapters one to seven.

Firstly, in **chapter 1** I set out how AF has captured recent clinical attention, including excitement and promise around the newly available direct oral anticoagulants (DOACs) for stroke prevention in AF,[31-34] recognition from major funders of biomedical research that the determinants of AF need better understood,[38] and how, in a positive step, global consortium are forming in order to tackle challenges in AF research together.[39 40] I described how EHR resources have been underutilised in studies of AF to date,[2] and therefore why I choose to exploit EHRs to investigate three areas of AF aetiology (i.e. risk factors,[2] subtypes,[3] and outcomes[4]) that were lacking clarity in international guidelines governing the way that AF is managed in clinical practice.[3 19 22]

In **chapter 2** I reviewed the existing observational epidemiology on the link between a range of cardiovascular risk factors and incident AF[2] and summarised the findings using a newer field synopsis methodology,[56] which, unlike a meta-analysis, brings together all of the available evidence despite heterogeneity in the way that risk factors and AF outcomes may have been defined. The review highlighted similarities (e.g. for obesity and hypertension) as well as differences (e.g. for lipids and ethnicity) in the associations of 23 cardiovascular risk factors with incident AF as compared to their known associations with other cardiovascular diseases (CVDs) like myocardial infarction and stroke. In addition, it was shown that there is a substantial disregard for the role that an intercurrent diagnosis of CVD may play in the development of AF and the associated impact this may have on estimating risk factor associations. However, as most of the included cohorts involved comparatively smaller samples sizes than are attainable when using EHRs,[204] it is likely that these cohorts may have been limited in the extent to which they could fully investigate whether risk factors lead directly to AF, or if risk factors lead to CVD, which in turn leads to AF.

In **chapter 3** I presented the CALIBER data resource,[5] describing each of the constituent data sources in detail [41 53 160 161] and how when linked together they provide higher resolution information about the onset and progression of AF.[51] I described and demonstrated analytic challenges which arise due to the fact data are not collected for primary research purposes including the importance of verifying data are clinically valid.[48] In a validation exercise of my own, I confirmed that CALIBER data are consistent with the results of my systematic review and field synopsis in showing similarities and differences in the associations of 23 cardiovascular risk factors with incident AF.[2] I summarised the strengths and limitations of using CALIBER records in research on AF, drawing comparisons with the epidemiological cohorts identified in my systematic review, and giving the balanced reasons why I ultimately choose to use CALIBER to study AF risk factors, subtype and outcomes as part of this PhD thesis.

Moving on to **chapter 4**, I resumed my examination of the question of the role of intercurrent cardiovascular diseases in the development of AF. I defined two separate AF endpoints, AF with (AF^+) and AF without (AF^-) an intercurrent diagnosis of CVD, and then modelled and compared the magnitude of associations with the 23 cardiovascular risk factors examined in my earlier systematic review[2] and validation exercise. I found that most standard CVD risk factors (e.g. age, sex, smoking, blood pressure, and diabetes mellitus) have stronger direct associations with AF^+ than AF^-, while the prior reported discordant association between higher lipids levels and lower risk of AF, was shown for AF^- but not for AF^+. These results suggest that limited progress in identifying risk factors for AF could in part be due to a lack of clarity in the definitions for AF and its subtypes.

The focus of **chapter 5** shifted away from AF risk factors and towards AF subtypes. Clearly, AF risk factors and AF subtypes are closely related as for AF primary prevention strategies to work the target of prevention must be clearly defined. However, this chapter arose largely in response to the 2016 updates to European Society of Cardiology (ESC) guidelines for the management of AF[3] in which a range of newly defined clinical distinctions of AF were outlined but without any large-scale systematic evidence to support their use. Seizing upon this important research gap, I therefore investigated whether these definitions could be operationalised in EHRs. By searching for applicable clinical codes and defining rules for combining them, together with plausible inferences, I set out how all seven subtypes (plus an additional eighth based on further reading[222]) could be potentially identified in records. I reflected upon the potential for these computable definitions to also be compatible with international systems,[42 43] in view of relevant upcoming cross-country projects such as the Innovative Medicines Initiative (IMI) BigData@Heart programme which seeks to drive progress in CVD research (AF included) through the bringing together of large-scale clinical datasets and collaborations.[39]

The goal of **chapter 6** was to take forward one of the EHR definitions derived in **chapter 5** for further implementation, improvement and validation. I selected 'valvular' AF for further development because of uncertainties in the valvular heart diseases comprising the definition,[228] which can be helped with insights drawn from EHRs, as well as there being established clinical

knowledge of a higher associated risk of subsequent stroke and thromboembolism,[6] which can be used as a point of validation. I found that not only do individuals with AF and prosthetic heart valves or mitral valve stenosis have a higher risk of stroke, systemic embolism and mortality than individuals with AF and no record of valvular heart disease (as expected),[6 247] but that individuals with AF and aortic stenosis may also have a poorer prognosis. From a methodological perspective, I showed that an underusage of available codes indicating whether heart valve replacements involved prosthesis or bioprosthesis could be overcome based on inferences about age at heart valve replacement and associated warfarin prescriptions (i.e. key criteria influencing valve replacement choice[240]). However, I could not similarly rectify an underusage of available codes indicating whether valvular heart diseases were of a rheumatic or non-rheumatic basis.

In **chapter 7** I turned the attention, for the last time, from AF subtypes toward AF outcomes; invoking the help of EHRs to investigate ischaemic stroke rates according to CHA_2DS_2-VASc scores, sex, and use of warfarin. Although one of the earlier completed projects of this PhD thesis, it is presented last in order to reflect the natural transition between risk factors predisposing to AF, through to subtypes influencing AF treatment decisions and then on to AF-related outcomes aimed at being prevented. This project was also completed prior to the 2016 updates to European Society of Cardiology guidelines for the management of AF and the results went on to be cited within the guidelines in relation to treatment recommendations for individuals at a low stroke risk.[3] Clinical guideline committees have, in the past, focussed on the results of clinical trials and meta-analyses in order to make recommendations and therefore this new model of using routinely collected clinical data to inform the preparation of guidelines, as adopted by the ESC, represents a very positive outlook for EHR research.

8.3 Novel contributions to atrial fibrillation research

This PhD thesis, exploiting EHRs for research, makes a number of novel contributions in relation to understanding of AF risk factors subtypes and outcomes. I describe these and suggest 12 recommendations for AF clinical practice and research.

Field synopsis methodology

In reviewing the existing field of AF risk factor research I employed a novel field synopsis methodology to summarise the findings.[2 56] Field synopses are commonly used to synthesise the results from genome-wide association studies (which look to find links between genetic variants and disease traits), but have seldom been applied in the context of preventative medicine. Field synopses, unlike single risk factor meta-analyses, take a horizontal view across a broad range of factors. Moreover, by applying the same study inclusion criteria (e.g. in my case prospective cohorts initially free from diagnosed CVD or with general population levels of CVD) it allows the findings yielded for each risk factors to be more easily compared. Whereas the calculation of a pooled risk estimate as part of a meta-analysis requires homogeneously defined risk factor and outcome variables and can thus lead to the exclusion of relevant reports, field synopsis allow for heterogeneity and instead are concerned with bringing together all the of the available evidence and appraising the overall amount, extent of replication and likelihood of bias.[56] Clearly, meta-

analyses have an important role in estimating precise risk factor associations, however field synopsis have an equally complementary role in providing a systematic foundation, unbiased by a particular interest in one or more risk factors,[158] for hypothesis generation and further research.

Recommendation 1:
Wider adoption of field synopsis methodology to summarise current understanding in AF research

AF endpoint definitions

As first described in my field synopsis,[2] and further illustrated in my investigation into the associations of risk factors with two diverse AF endpoints (AF^+ and AF^-), a lack of a regard for the role of intercurrent CVDs in the development of AF can result in potentially misleading findings. Traditional epidemiological cohort studies (e.g. the Framingham Heart Study with 1544 incident cases AF accrued over 50 years[191]) are limited in terms of sample size and frequency of follow-up information and are thus less able to answer deeper mechanistic questions about how diseases develop.[204] Limitations around sample sizes and follow-up information are starting to be overcome by combing data from multiple cohorts (e.g. the CHARGE-AF consortium of five cohorts[40]) and ad-hoc linkage of cohorts to EHRs resources. UK Biobank,[197] a newer cohort study set up in 2006, is set apart from traditional cohorts (e.g. Framingham), with 500,000 recruited participants (i.e. larger than usual), consent for linkage to EHR resources from the outset, and available genomics and whole body imaging data (ECGs included); making it a promising resource to study AF once sufficient cases amass. While the UK Biobank catalogue of data is impressive it has however been costly to generate (e.g. initial set up costs of £61 million[267]). EHRs, by comparison, offer big data approaches to cohort epidemiology with negligible costs as data are generated as routine part of clinical care.[204] Future investigations into risk factors potentially predisposing to AF, regardless of whether in the context of traditional cohorts or in emerging datasets like UK Biobank[197] or CALIBER,[5] all need to move away from using all-encompassing AF definitions and should instead be guided by the new recommendations for AF subtypes provided by the ESC.[3]

Recommendation 2:
Wider use of higher resolution AF endpoints in studies of AF risk factors

Cardiovascular disease prevention programmes

As supported by evidence from my systematic review,[2] validation exercise of risk factor data in CALIBER, and novel investigation into the link between risk factors and AF both with (AF^+) and without (AF^-) an intercurrent diagnosis of CVD, there are clear differences in how a range of demographic, behavioural and biological cardiovascular risk factors impact the development of AF as compared to how they impact the development of other CVDs such as myocardial infarction and stroke. I found that most standard CVD risk factors (e.g. age, sex, smoking, blood pressure, and diabetes mellitus) have consistent direct associations with AF^+ and AF^-; albeit with far stronger risk estimates for AF^+. While for heavy drinking and lower lipids levels the findings were

inconsistent. Heavy drinking increased the risk of AF^-, but not AF^+, and higher lipids levels were protective against AF^- (which is discordant with knowledge about other CVDs) but showed no relationship with AF^+. Aggressively lowering blood pressure, cholesterol, glucose levels beyond current targets has been the subject of research in recent years,[268] and often showing favourable results**.** However, what exactly aggressive CVD risk factor management means for AF has been inadequately explored. I therefore propose two recommendations:

Recommendation 3:
Further research into understanding the role of the lipids in the development of AF

Recommendation 4:
The redesign of CVD prevention programmes so that they work for all CVDs

Global AF data

In systematically evaluating the existing cohort epidemiology on 23 risk factors in relation to AF,[2] I found, overall, a relatively "young" field of research. Although the review included 32 cohorts of 20 million participants and 600,000 AF events, there were a limited number of reports (between 3 and 19) per risk factor and evidence to suggest unpublished risk factor data (e.g. most consented cohorts measure blood pressure but thus far not all have reported on it). I found that the AF field is dominated by North American and North European cohorts, with no cohorts included in my review from Central or South America, Eastern or Southern Europe or South Asia. This is not just unique to AF, but consistent with other cardiovascular and non-communicable diseases.[152] However, in the interest of reducing the global burden of AF, there is no global data on which to base global prevention strategies. Although I identified, efforts underway at pooling cohorts (e.g. CHARGE–AF [40]), the amount of evidence for AF risk factors is still markedly smaller than what is available for coronary heart disease or stroke (e.g. the Emerging Risk Factor Collaboration of over 100 cohorts[151]). The IMI BigData@Heart programme,[39] launched in 2017, therefore represents a positive step towards the discovery and integration of large-scale routinely collected clinical and cardiovascular-related datasets across Europe, and will be particularly useful in studying some of the rarer AF subtypes (e.g. AF secondary to inherited rhythm disorders estimated to account for less than 5% of all AF cases[3]). But equally, data from beyond Europe must also be leveraged, and especially data from developing countries for research into the rheumatic basis of valvular AF.[243]

Recommendation 5:
Discovery of global AF data, not just from Europe and the US, but in particular from developing countries

EHR definitions for eight AF subtypes

I successfully created EHR definitions for eight mechanistically diverse subtypes of atrial fibrillation with relevance to the recent 2016 updates to ESC guidelines for the management of AF.[3] With the exception of valvular AF, which I took forward for further refinement and validation, the

initial definitions for structural, focal, polygenic, postoperative, AF in athletes, monogenic, and respiratory AF require further testing, however they demonstrate a feasible approach to targeting a broad range of AF subtypes in the records and are a foundation for future research. In evaluating the potential for these EHR definitions to work beyond the CALIBER dataset,[5] I showed that six out of eight definitions are immediately amenable to analyses in full or at least in part (i.e. the definitions which can work with ICD-10 codes alone) in other international systems (e.g. in Denmark[42] and Sweden[43]). These definitions are likely to be of particular interest to the co-ordinators of IMI BigData@Heart working to align EHR resources in Europe for the progression of cardiovascular research with a special focus on AF.[39]

Recommendation 6:
Adoption, refinement and validation of the eight EHR definitions I developed for research into AF subtypes

Valvular AF

Given differences in definitions across international guidelines[3 22] and clinical trials,[31-34] and reports of clinician uncertainty and treatment confusion,[228] I used EHRs to investigate the widely debated subtype of 'valvular AF'.[6] In the largest systematic attempt (as far as I am aware) to answer the valvular AF question, I found, as expected, a higher risk of subsequent stroke, systematic embolism and mortality in individuals with prosthetic valve replacements or mitral stenosis as compared to individuals with AF and no valvular heart disease,[6] and these associations remained significant after adjustment for age, sex, stroke risk (CHA_2DS_2-VASc factors) and warfarin use. This finding can give reassurance to clinicians that these individuals require separate clinical attention, which, according to current guidelines, means lifelong stroke prophylaxis by way of warfarin; DOACs are contraindicated due to higher complication rates.[247] In a novel finding of this study, it was shown that individuals with AF and aortic valve stenosis also had significantly worse outcomes than individuals with no valvular heart disease. However, as this association has not been reported before in the observational literature (possibly due to limited sample sizes of prior studies[204]), further validation studies are required to confirm the clinical relevance of aortic stenosis in the progression of AF and ideally using comparable large-scale resources. Recommendations arising from this work are the following:

Recommendation 7:
Greater clinician confidence in treating valvular AF i.e. AF in the context of mitral valve stenosis and prosthetic heart valves

Recommendation 8:
Further research, ideally using large-scale EHR resources, to confirm the link between aortic stenosis and AF prognosis

Stroke risk

In using the CALIBER dataset[5] to investigate stroke rates in individuals with AF, by sex, CHA_2DS_2-VASC and use of warfarin I highlighted the value of linking records from primary and secondary care in order to capture representative patient populations and complete information on risk factors and outcome variables, and, ultimately, in order to provide accurate estimations.[7] Prior to the conduction of this study, the level of risk at which individuals should initiate stroke prevention with anticoagulants, as determined by CHA_2DS_2-VASC score, was unclear (namely whether just one stroke risk factor is sufficient to warrant treatment). A systematic review on this topic indicated vast heterogeneity in 10 prior studies.[4] Although other real-world EHR datasets had been used to investigate the question, including the Danish[260] and Swedish[265] nation-wide cohorts, these had predominantly focussed on AF patients in secondary care. Thus there was a lack of understanding on how individuals with AF managed in primary care and secondary care differ. Capitalising on the available linkages between primary and secondary care records, I therefore aimed to study these uncertainties. I found that the benefit of treatment was observed from a CHA_2DS_2-VASc score of 2 in men and of a CHA_2DS_2-VASc score of 3 in women. I obtained an ischaemic stroke rate of 0.7 [0.6 to 0.8] for CHA_2D_S2-VASc=1 without warfarin, which was low compared to previous reports although consistent with levels of uncertainty in the annualised rate of 1.6% [0% to 3.23%] estimated in the recent systematic review and meta-analysis.[4] In a series of sensitivities analyses I sought to find possible explanations as to why my estimate was lower. I found that hospital records, alone, were inadequate for estimating CHA_2DS_2-VASc risk with more than half of hospitalised AF patients reclassified to higher scores when factoring in clinical information recorded exclusively in primary care. Thus prior studies have potentially overestimated stroke rates in individual with a seemingly low stroke risk, as they were likely under classified. Furthermore, hospitalised patients were shown to reflect a riskier patient population, with higher proportions all of CHA_2DS_2-VASc risk factors. Evidence obtained in analyses of secondary care data may therefore give rise to misleading inferences about the overall AF population. Future studies should link wherever possible, and, if not possible, more firmly address the limitations of not being able to link. From a health systems perspective, improved recording of clinical events and greater interoperability between records collected in different healthcare sectors will benefit both future research and clinical practice. Based on this work I have the following recommendations:

Recommendation 9:

Individuals with AF should have warfarin from a CHA_2DS_2-VASc of 2 in men and 3 in women

Recommendation 10:

Linkage of data sources wherever possible to accurately ascertain baseline disease risk as well as improved recording and interoperability of clinical events collected across healthcare sectors

Clinical trials

Two of the studies I conducted had findings with implications for clinical trials. Firstly, I found that standard cardiovascular risk factors, age, sex, smoking, blood pressure and diabetes mellitus, were more strongly associated with AF with an intercurrent diagnosis of CVD, as compared

to AF without an intercurrent diagnosis of CVD. Trials of initially health individuals without pre-existing CVD and with AF as the primary endpoint have been lacking thus far.[35] However, this finding does not support the inclusion of AF as a primary endpoint in future CVD prevention trials (e.g. on blood pressure lowering medications), and instead the focus should remain on the initial prevention of other CVDs such as myocardial infarction and stroke. In my analysis using EHRs to refine understanding about valvular AF I found evidence to support the inclusion of mitral valve stenosis and prosthetic heart valves in the valvular AF distinction[6] as well as, potentially, the addition of aortic valve stenosis. Recent trials of DOACs for stroke prevention in AF used definitions to exclude valvular AF which ranged from both mitral valve stenosis and prosthetic heart valves to the exclusion of all valvular heart diseases combined.[31-34] Whereas, EHR resources like CALIBER have an already established role in examining the subsequent real-world implications of clinical trials (e.g. testing how DOACs perform in the usual clinical setting), here I show how EHRs can also be useful in informing the upfront trial design. The RE-LY,[31] ARISTOTLE[32] and ROCKET-AF[33] trials have previously been shown to be 74%, 72% and 56% representative of individuals with AF in the UK respectively.[192] Therefore, using EHRs to identify groups of individuals for inclusion in clinical trials can help to bring trial representativeness closer to 100%.

Recommendation 11:
Wider use of EHR data sources to inform the design of future clinical trials, including on AF primary prevention and outcomes in valvular AF where there is a larger knowledge gap

Guidelines

The work of this PhD thesis was able to both influence and directly respond to changes in major international guidelines governing the way that AF in managed in clinical practice. CALBER data was used to clarify uncertainties around the level of stroke risk at which individuals with AF should be recommended to take preventative anticoagulants and was cited in the 2016 updates to ESC guidelines for AF.[3] On the flip side, new ideas around different AF mechanisms and subtypes distinctions were suggested in the 2016 ESC guidelines but without any large-scale supporting evidence.[3] I therefore set out how these definitions could be operationalised in EHRs, providing a viable setting in which understanding of about AF subtypes could be refined and validated. Clinical guidelines are compiled by disease experts and based on the evidence available at the time, however rather than pinpointing knowledge gaps and caveats, guideline writers can proactively turn to large-scale systematic evidence available in EHRs in an attempt to answer outstanding clinical questions.

Recommendation 12:
Wider use of EHR data to enhance the preparation of AF clinical guidelines

8.4 Overall strengths and limitations

The overall strengths and limitations I wish you address fall into two categories: (1) data and (2) methods. Strengths and limitations are described alongside in order to ease comparison of the associated trade-offs:

Data

Throughout this PhD thesis I have showcased the breadth and depth of CALIBER dataset.[5] CALIBER's data linkages, large-scale population denominator and proven data validity are among some of the major advantages for its use in research on AF.[51] Data not currently captured in CALIBER and a lack of control over data quality are among the disadvantages.

CALIBER's data linkages capture complementary information on AF care, which have afforded me the ability to make multiple clinical and research focussed recommendations about three areas of AF aetiology (risk factors, subtypes and outcomes). Whereas secondary care[41] and mortality records[160] have been particularly important for capturing AF-related outcomes (i.e. fatal and non-fatal strokes), primary care records have a number of unique features,[53] which have majorly enhanced the opportunities for and quality of my research. The availability of numerical clinical values (e.g. blood pressure and lipids), which are rare in other EHR resources, [5 51] meant that all 23 cardiovascular risk factors could be studied in relation to incident AF using a single dataset. The coding of primary care records with the more extensive Read classification system[163] meant that EHR definitions could be created for all new AF subtype distinctions suggested by the ESC, including the occupational-related 'AF in athletes'. Primary care records were also instrumental in capturing a more representative AF patient sample (i.e. including AF patients never admitted to hospital) as well as more complete information on CHA_2DS_2-VASc components to more accurately risk stratify and calculate associated stroke outcomes.[7] In an ideal situation, CALIBER would of course be linked to ECG images, which are the gold standard technique for diagnosing AF. [3 19 22] Coded AF diagnoses in CALIBER have been verified in multiple validation exercises including in a GP re-contact study,[180] Morley's investigation showing similar AF prevalence to a UK-based ECG study,[49] and in my own investigation showing similar AF risk factor associations as in existing literature. However, in the absence ECGs, some AF cases will have been undoubtedly missed or misclassified. CALIBER's large scale population denominator was shown to be of particular value for studying AF subtypes. Out of 76,019 individuals with prevalent AF, I found 12,751 (16.8%) with a concomitant diagnosis of valvular heart disease which allowed the 'valvular' AF subtype definition to be meaningfully explored. The culmination of large amounts of data from diverse clinical settings does however come at the expense of data quality. Robustly identifying disease cases in records is a challenge that largely depends upon the quality of data capture at source.[48] The analyses of this PhD benefitted from established EHR definitions for disease risk factors and outcomes.[51] However, as I showed for valvular heart diseases of rheumatic origin, not all phenotypes can be easily abstracted. Finally, a key limitation of this work was the lack of up-to-date data which came as consequence of recent challenges in information governance (e.g. Care.data).[169] I was thus unable to use EHRs to generate novel insights about the new DOACs for stroke prevention in AF.

Data strengths

- Data linkages bring enhanced understanding of AF risk factors, subtypes and outcomes
- Large scale population denominator necessary for investigating AF subtypes
- Proven validity of data for AF research

Data limitations

- Lack of ECG imaging to refine understanding about AF cases
- Lack of control over data quality
- Lack of real time/recent data (e.g. currently no DOACs data)

Methods

In this PhD I have applied a range of methods including field synopsis,[2] EHR algorithm development[49] and statistical modelling in order to make novel contributions towards the understanding of AF onset and progression. The decision to study AF risk factors, subtypes and outcomes was based around shortcomings in clinical guidelines for AF as well as vast data opportunities in CALIBER and as a result this thesis provides a solid foundation for future EHR research in these three areas. Clearly, in considering several aspects of AF aetiology it has restricted the depth to which I have evaluated the evidence in relation to any one area. In reflecting in more detail upon the overall strengths and limitations of the methods applied, I also comment upon emerging techniques (i.e. machine learning[269] and mendelian randomisation[270]) which I have not applied but consider to be of increasing relevance to the field of AF research.

As already considered above, I employed a novel field synopsis methodology[56] to scalably and systematically summarise the existing field of AF risk factor research. Field synopses welcome in heterogeneity in the way that risk factors and outcome variables are defined which has the advantage of ensuring that all relevant reports are likely included. The trade-off of allowing for differences in study design is that I did not meta-analyse any factors and therefore my results are limited in the extent to which I can advise on optimal risk factor levels or precise threshold effects. A challenge when analysing EHR data lies in how to robustly define risk factors and disease endpoints.[48] This PhD benefitted from validated EHR definitions,[51] for example Morley's AF based on 286 clinical codes and inferences around warfarin prescriptions in the absence of thromboembolic disease.[49] In developing eight EHR definitions for AF subtypes, I, too, have helped to drive forward research into AF using EHRs. EHR definitions were derived in accordance with the CALIBER framework,[51] which involves translating disease descriptions into clinical codes and iteratively testing and improving definitions against data insights and established clinical knowledge. In many respects this reflects a traditional hypothesis driven approach towards research into disease subtypes. Though not explicitly explored within this PhD, machine learning[269] reflects a more contemporary data driven approach to the discovery of disease subtypes. Rather than starting with a theory about what the subtype looks like the idea is instead to be guided by correlations within the data. The application of machine learning in medicine remains controversial because the resultant models are often difficult to interpret or explain.[269] However one area where machine learning is beginning to make an impact is in the updating of already

existing risk prediction tools (e.g. CHA_2DS_2-VASc) to improve model performance.[271] Lastly, in analysing 23 cardiovascular risk factors in relation to AF risk I used an available case data strategy[190] fixing study baseline dates based on when risk factors were measured. This strategy meant that I could maximise the use observed risk factor data however the number of individuals and their associated characteristics in each risk model were not identical. Multiple imputation reflects a more sophisticated method for the handling of missing covariate data however this was not a scalable option due to the fact that multiple risk models were being considered and each model requires a separate imputation strategy.[178] Biases due to missing data may therefore be possible but are likely to be small given the consistency of risk factors associations with prior literature. All epidemiological studies are subject to inherent biases in the way that risk factors have been measured for reasons that includes reverse causation. Mendelian randomisation[270] therefore reflects an emerging approach for the unbiased estimation of casual relationships between risk factors and disease outcomes. Mendelian randomisation works by modelling the genetic variants associated with risk factors and disease outcomes and is considered unimpacted by confounders because genes are randomly determined at the point of conception. Moving forward mendelian randomisation techniques should be more widely applied in relation to AF risk factors for example in ascertaining whether there is a link between lipids, height and ethnicity.[159]

Methods strengths

- Evaluation across a broad range of factors to inform three aspects of AF aetiology
- Foundation for future EHR research into AF risk factors, subtypes and outcomes
- EHR definitions for AF subtypes developed in line with CALIBER guidance

Methods limitations

- Restricted in depth to which results can inform three aspects of AF aetiology
- Imputation of missing data impractical
- No machine learning (yet)
- No mendelian randomisation (yet)

As a final thought, it should be stated that the central focus of this PhD has been on the application of EHRs in observational epidemiology. Of course, other study designs in clinical medicine are available but this PhD serves to illustrate how the untapped potential in routinely collected EHR data can be leveraged for public health benefit, and in particular in advancing the understanding and prevention of AF and its consequences.

8.5 Conclusion

Atrial fibrillation is a heterogeneous condition associated with diverse mechanisms for onset and progression. Electronic health records, collected on large populations as part of routine clinical care, can help to refine understanding about atrial fibrillation risk factors, subtypes, and outcomes and should be increasingly utilised to inform future clinical trials, future clinical guidelines and future clinical practice.

Appendix of supplementary methods, tables and figures

Chapter 1

Introduction to overall aims, objectives and background motivating this research

No supplementary material.

Chapter 2

Systematic review and field synopsis of the link between 23 cardiovascular risk factors and incidence of atrial fibrillation

Table S2.1 Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	p24
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	p24
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	p25-6
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	p25-6
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	N/A
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	p26
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	p26
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	p153
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	p26-7
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	p26-7
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	p26
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	p27
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	P27
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	P27
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	P27

Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	N/A
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	p27 p154
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	p27-8
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	p27-8
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	p28-30
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	N/A
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	p27-8 p155-9
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	N/A
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	p30-3
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	p33
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	P33
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	N/A

Table S2.2 Search terms used to identify relevant reports

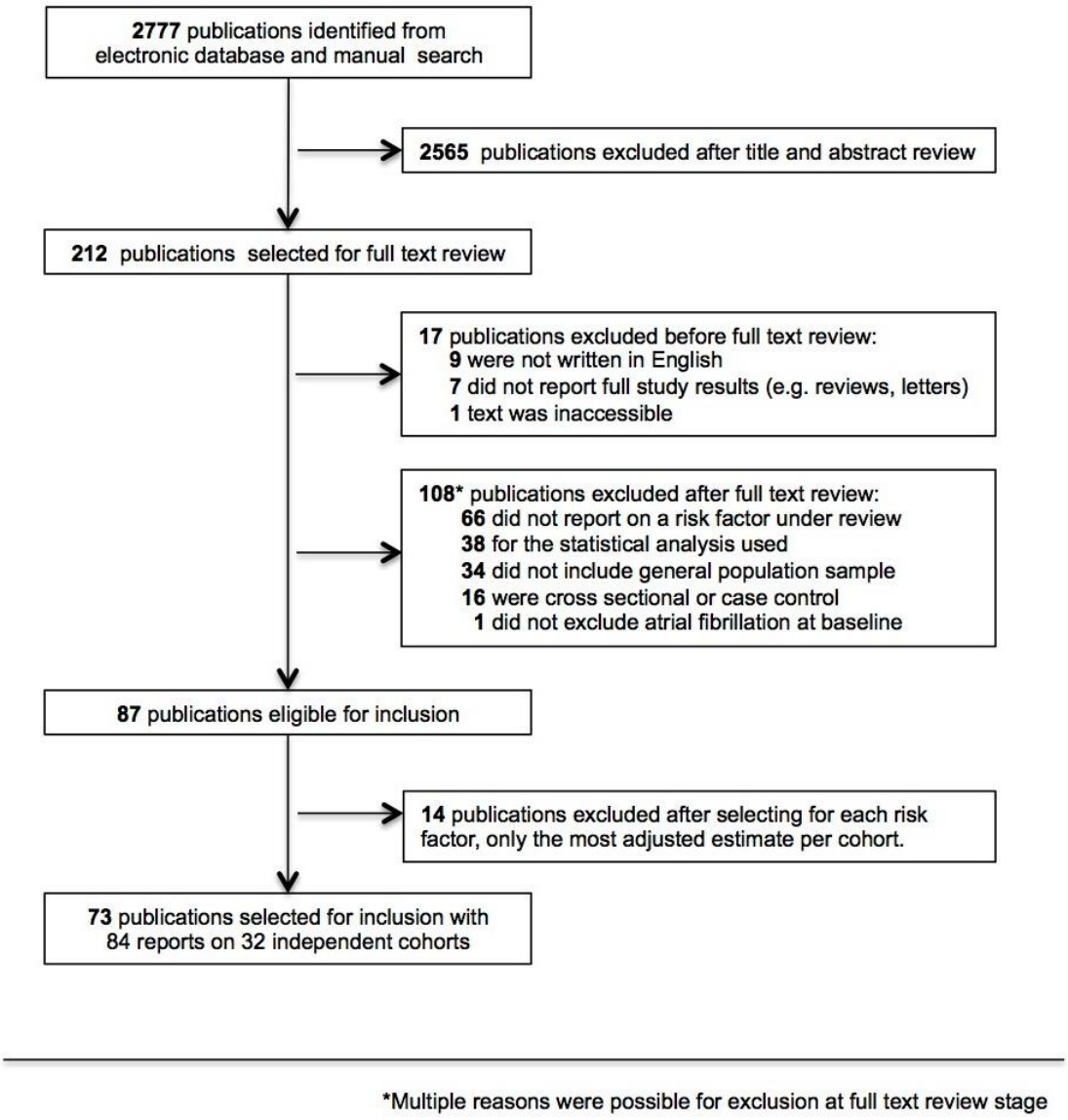
(“atrial fibrillation"[title] OR “atrial flutter"[title] OR “auricular fibrillation"[title] OR “auricular flutter"[title] OR "cardiac arrhythmia"[title])

AND (“epidemiology” OR “incidence” OR “risk factors” OR “risk score” OR “risk assessment”)

AND (“prospective studies” OR “cohort” OR “observational”)

AND ("1900–01–01"[Date – Publication] : "2015–10–01"[Date – Publication])

Figure S2.1 Systematic review flow diagram



As shown, 2777 publications were identified through electronic database and hand searching, 2565 were excluded on title and abstract, 212 were reviewed in full, 87 were deemed eligible for inclusion, and 73 publications with 84 reports on 32 independent cohorts remained after selecting for each risk factor, only the most adjusted risk ratio and 95 % confidence interval per cohort.

Table S2.3 Methods, data sources and clinical codes used in ascertaining incident atrial fibrillation events

					Medical records															
					Record types								Clinical codes used							
Cohort	Total AF events	AF from research ECG	AF from self-reports	AF from medical records:	General practice	Hospital care	Prescriptions	Mortality	Total record types used	Surveillance	Linkage	AF from healthcare ECG	AF from coded data	ICD7	ICD8	IC9	ICD9CM	ICD10	ATC	Reference
ARIC	1794	●	○	●	○	●	○	●	2	●	○	○	●				427.3 427.31 427.32	I48		84
ARIC	1775	●	○	●	○	●	○	●	2	●	○	○	●				427.31 427.32			85
ARIC	1520	●	○	●	○	●	○	●	2	●	○	○	●				427.31 427.32			86
ARIC	1433	●	○	●	○	●	○	●	2	●	○	○	●				427.3 427.31 427.32	I48		87
ARIC	1209	●	○	●	○	●	○	●	2	●	○	○	●				427.31 427.32	I48		88
ARIC	1085	●	○	●	○	●	○	●	2	●	○	○	●				427.3 427.31	I48		89
ARIC	1068	●	○	●	○	●	○	●	2	●	○	○	●				NR	NR		90
ARIC	788	○	○	●	○	●	○	●	2	●	○	○	●				427.3 427.31 427.32	I48		91
ARIC	515	●	○	●	○	●	○	●	2	●	○	○	●				427.3 427.31 427.32	I48		92
ARIC	419	●	○	●	○	●	○	●	2	●	○	○	●				427.3 427.31 427.32	I48		40
TS	822	○	○	●	○	●	○	●	2	○	●	●	●				427.0- 427.99	I47-I48		112
TS	566	○	○	●	○	●	○	●	2	○	●	●	●				427.0- 427.99	I47-I48		113
RS	402	●	○	●	●	●	○	○	2	●	●	○	●				NR	NR		122
RS	371	●	○	●	●	●	○	○	2	●	●	○	●				NR	NR		123
RS	177	●	○	●	●	●	○	○	2	●	●	○	●				NR	NR		40
RS	177	●	○	●	○	○	○	○	NR	●	●	○	●				NR	NR		120
TSS	253	●	●	●	○	●	○	●	2	●	○	○	NR							136
S-EHR	3859	○	○	●	○	●	○	●	2	○	●	○	●	433.12	427.92	427D		I48		142
HCUP	375318	○	○	●	○	●	○	○	1	○	●	○	●				427.31			137
D-EHR	156484	○	○	●	○	●	○	○	1	○	●	○	●					I48		138
D-EHR	126217	○	○	●	○	●	○	○	1	○	●	○	●		427.94 427.95			I48		139
D-EHR	115956	○	○	●	○	●	○	○	1	○	●	○	●					I48		140
D-EHR	17154	○	○	●	○	●	○	○	1	○	●	○	●					I48		141
T-NHIRD	1041	○	○	●	○	●	○	○	1	○	●	●	●				427.31			143
WHI-OS	9792	○	○	●	○	●	○	○	1	●	●	○	●				427.31			73
WHI-OS	8252	○	○	●	○	●	○	○	1	●	●	○	●				427.31			74

				Medical records																
					Record types															
Cohort	Total AF events	AF from research ECG	AF from self-reports	AF from medical records:	General practice	Hospital care	Prescriptions	Mortality	Total record types used	Surveillance	Linkage	AF from healthcare ECG	AF from coded data	Clinical codes used: ICD7	ICD8	IC9	ICD9CM	ICD10	ATC	Reference
COSM	4568	○	○	●	○	●	○	○	1	○	●	○	●		427.92	427D		I48		75
COSM	4488	○	○	●	○	●	○	○	1	○	●	○	●					I48		63
SMC	2915	○	○	●	○	●	○	○	1	○	●	○	●		427.92	427D		I48		79
SMC	2757	○	○	●	○	●	○	○	1	○	●	○	●					I48		63
DCHS	2581	○	○	●	○	●	○	○	1	○	●	○	●		427.93 427.94			I48		80
DCHS	553	○	○	●	○	●	○	○	1	○	●	○	●		427.93 427.94			I48		81
DCHS	418	○	○	●	○	●	○	○	1	○	●	○	●		427.93 427.94			I48		82
MPP	2312	○	○	●	○	●	○	○	1	○	●	○	●				427.3	I48		83
CHS	1585	●	○	●	○	●	○	○	1	○	●	○	●				427.3 427.31 427.32			93
CHS	1172	●	○	●	○	●	○	○	1	○	●	○	●				NR	NR		90
CHS	1061	●	○	●	○	●	○	○	1	○	●	○	●				427.3 427.31 427.32			94
CHS	897	●	●	●	○	●	○	○	1	○	●	○	●				427.3 427.31 427.32			95
CHS	812	●	○	●	○	●	○	○	1	○	●	○	●				427.3 427.31 427.32			96
CHS	624	●	○	●	○	●	○	○	1	○	●	○	●				427.31 427.32			40
CHS	579	●	●	●	○	●	○	○	1	○	●	○	●				427.3 427.31 427.32			97
CHS	304	●	●	●	○	●	○	○	1	○	●	○	●				427.3 427.31 427.32			98
MDCS	1430	○	○	●	○	●	○	○	1	○	●	○	●		427.92	427D		I48		99
MDCS	284	○	○	●	○	●	○	○	1	○	●	○	●		427.92	427D		I48		100
GPPS	1253	○	○	●	○	●	○	○	1	○	●	○	●		427.92	427D		I48		101
NorPD	863	○	○	●	○	○	●	○	1	○	●	○	●						C01BC04 C07AA07	111
MCS	667	○	○	●	○	●	○	○	1	○	●	○	●		427.92	427D		I48		119
AGES	408	●	○	●	○	●	○	○	1	○	●	○	●				427.3	I48		40
AGES	408	●	○	●	○	●	○	○	1	○	●	○	●				427.3	I48		120
AGES	226	○	○	●	○	●	○	○	1	○	●	○	●				NR	NR		121
CCHS	379	○	○	●	○	●	○	○	1	○	●	○	●		427.93 427.94			I48.9		124
CCHS	268	○	○	●	○	●	○	○	1	○	●	○	●		427.93 427.94			I48.9		125
HABC	371	○	○	●	○	●	○	○	1	○	●	○	●				427.31 427.32			126
HABC	17	●	○	NR	○	○	○	○	NR	○	○	○	NR							127

BHS	343	○	○	●	○	●	○	○	1	○	●	○	●				427.3 427.31 427.32	I48		128
BHS	14	●	○	NR	○	○	○	○	NR	○	○	○	NR							127
MESA	307	○	○	●	○	●	○	○	1	●	●	○	●				427.31 427.32			129
MESA	305	○	○	●	○	●	○	○	1	●	●	○	●				427.31 427.32			130
MESA	221	○	○	●	○	●	○	○	1	●	●	○	●				427.31 427.32			116
MESA	199	○	○	●	○	●	○	○	1	●	●	○	●				427.3			131
MESA	182	○	○	●	○	●	○	○	1	●	●	○	●				427.31 427.32			132
CIRCS	296	●	●	●	○	●	○	○	1	●	○	○	NR							133
S-HS	285	○	○	●	○	●	○	○	1	○	●	○	●					I48.0 I48.1 I48.2 I48.3 I48.4 I48.9		134
OCS	270	●	○	●	○	●	○	○	1	○	●	○	●							135
FHS	698	●	○	●	○	●	○	○	1	●	○	●	NR							114
FHS	457	●	○	●	○	●	○	○	1	●	○	●	NR							115
FHS	259	●	○	●	○	●	○	○	1	●	○	●	NR							116
FHS	192	●	○	●	○	●	○	○	1	●	○	●	NR							117
FHS	148	●	○	●	○	●	○	○	1	●	○	●	NR							118
FHS	143	●	○	●	○	●	○	○	1	●	○	●	NR							40
L85-PS	39	●	○	NR	○	○	○	○	NR	○	○	○	NR							127
SHIP	34	●	○	NR	○	○	○	○	NR	○	○	○	NR							127
NPMS	2974	●	○	NR	○	○	○	○	NR	○	○	○	NR							76
NPMS	265	●	○	NR	○	○	○	○	NR	○	○	○	NR							77
NPMS	265	●	○	NR	○	○	○	○	NR	○	○	○	NR							78
IPHS	1232	●	○	NR	○	○	○	○	NR	○	○	○	NR							102
WHS	1027	○	○	●	○	○	○	○	NR	●	○	●	NR							103
WHS	968	○	○	●	○	○	○	○	NR	●	○	●	NR							104
WHS	834	○	○	●	○	○	○	○	NR	●	○	●	NR							105
WHS	795	○	○	●	○	○	○	○	NR	●	○	●	NR							106
WHS	786	○	○	●	○	○	○	○	NR	●	○	●	NR							107
WHS	747	○	○	●	○	○	○	○	NR	●	○	●	NR							108
WHS	653	○	○	●	○	○	○	○	NR	●	○	●	NR							109
WHS	644	○	○	●	○	○	○	○	NR	●	○	●	NR							110

Table abbreviations: AF - atrial fibrillation, ECG - electrocardiogram, research ECG - refers to whether an ECG was performed as part of the study, healthcare ECG - refers to whether an ECG was performed as part of routine healthcare, surveillance - refers to whether medical records were accessed through active surveillance by study researchers, linkage - refers to whether medical records were accessed through linkage to electronic health records databases, NR - not reported, ICD-n - international classification of diseases and version number, CM - clinical modification, ATC - Anatomical Therapeutic Chemical Classification System.

Cohort abbreviations: AGES - Age, Gene and Environment-Reykjavik study, ARIC -Atherosclerosis Risk in Communities, BHS - Busselton Health Study, CCHS - Copenhagen City Heart Study, CHS - Cardiovascular Health Study, CIRCS - Circulatory Risk in Communities Study, COSM - Cohort of Swedish Men, DCHS - Diet Cancer and Health study, D-EHR - Denmark Electronic Health Record cohort, FHS - Framingham Heart Study, GPPS - Göteborg Primary Prevention Study, HABC - Health, Aging, and Body Composition, HCUP - Healthcare Cost and Utilization Project, IPHS - Ibaraki prefectural health study, L85-PS - Leiden 85-Plus Study, MCS - Malmö Cardiovascular Screening, MDCS - Malmö Diet and Cancer study, MESA - Multi-Ethnic Study of Atherosclerosis, MPP - Malmö Preventive Project, NorPD - Norwegian Prescription Database, NPMS - Niigata preventive medicine study, OCS - Oslo Cardiovascular Survey, RS - Rotterdam Study, S-EHR - Sweden Electronic Health Record cohort, SHIP - Study of Health in Pomerania, S-HS - Stockholm Health Screening cohort, SMC - Swedish Mammography Cohort, T-NHIRD - Taiwan National Health Insurance Research Database TS - Tromsø Study, TSS - The Suita Study, WHI-OS - Women's Health Initiative Observational Study, WHS - Women's Health Study.

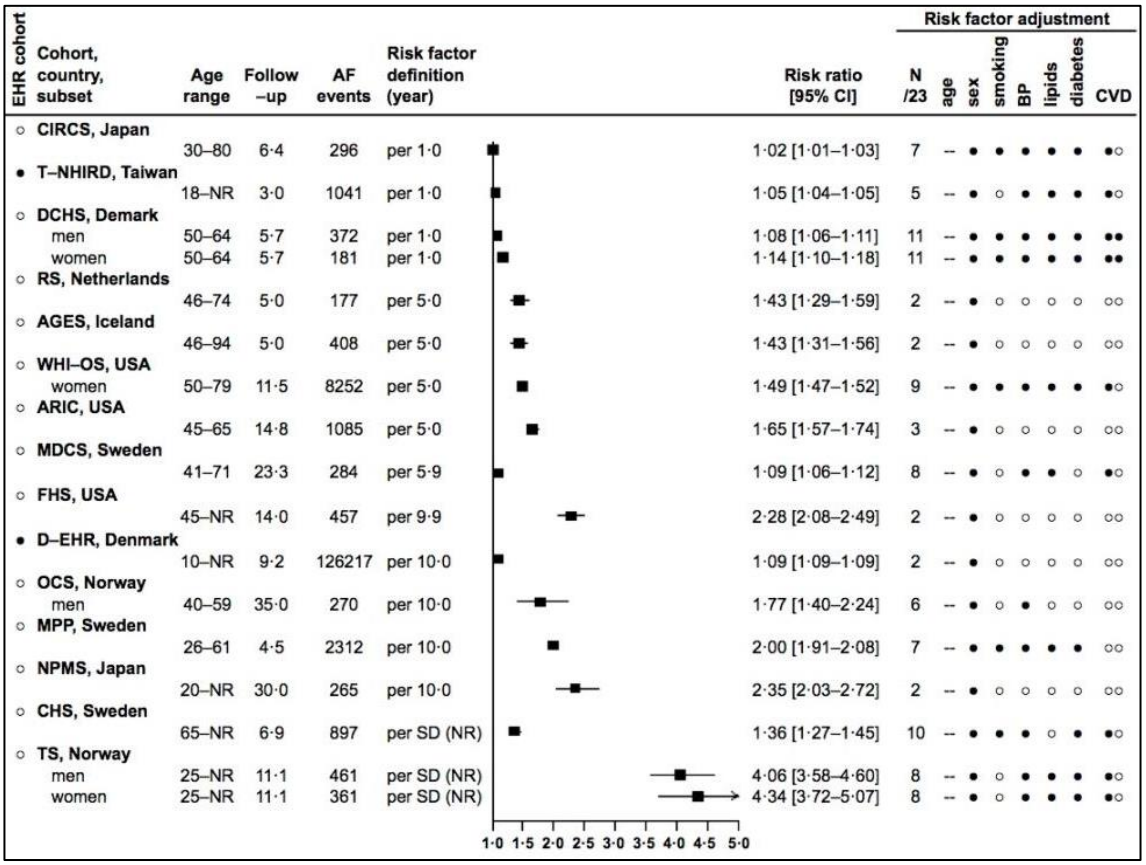
Table S2.4 Adjustment for 23 risk factors under review, 6 cardiovascular risk factors, and prevalent and incident cardiovascular disease

Reference	Cohort	Adjustment factors		Adjustment for six CVD risk factors								Adjustment for CVD	
		N /23	%	age	sex	smoking	BP	lipids	diabetes	N /6	%	Prevalent	Incident
88	ARIC	14	61	●	●	●	●	●	●	6	100	●	○
116	MESA	14	61	●	●	●	●	○	●	5	83	●	●
116	FHS	13	57	●	●	●	●	○	●	5	83	●	●
106	WHS	13	57	●	●	●	●	○	●	5	83	○	○
125	CCHS	12	52	●	●	●	●	●	●	6	100	●	●
87	ARIC	12	52	●	●	●	●	○	●	5	83	●	○
91	ARIC	12	52	●	●	●	●	○	●	5	83	●	●
80	DCHS	11	48	●	●	●	●	●	●	6	100	●	●
102	IPHS	11	48	●	●	●	●	●	●	6	100	○	○
131	MESA	11	48	●	●	●	●	●	●	6	100	●	○
81	DCHS	11	48	●	●	●	●	●	●	6	100	●	●
107	WHS	11	48	●	●	●	●	●	●	6	100	○	○
108	WHS	11	48	●	●	●	●	●	●	6	100	○	○
95	CHS	10	43	●	●	●	●	○	●	5	83	●	○
120	RS	10	43	●	○	●	●	○	●	4	67	●	○
120	AGES	10	43	●	○	●	●	○	●	4	67	●	○
126	HABC	10	43	●	●	●	●	●	●	6	100	●	○
124	CCHS	10	43	●	●	●	●	○	●	5	83	●	○
97	CHS	10	43	●	●	○	●	●	●	5	83	●	○
103	WHS	10	43	●	●	●	●	●	○	5	83	●	●
110	WHS	10	43	●	●	●	●	●	●	6	100	●	●
105	WHS	10	43	●	●	●	●	●	●	6	100	○	○
109	WHS	10	43	●	●	●	●	●	●	6	100	●	●
119	MCS	10	43	●	●	●	●	●	●	6	100	●	○
113	TS	10	43	●	●	●	●	●	●	6	100	●	○
133	S-HS	10	43	●	●	●	●	○	●	5	83	●	○
82	DCHS	10	43	●	●	○	●	●	○	4	67	○	○
136	TSS	9	39	●	●	●	●	●	●	6	100	●	○
132	MESA	9	39	●	●	●	○	●	●	5	83	●	○
86	ARIC	9	39	●	●	●	●	○	●	5	83	●	○
84	ARIC	9	39	●	●	●	●	○	●	5	83	●	●
129	MESA	9	39	●	●	●	●	○	●	5	83	●	●
74	WHI-OS	9	39	●	●	●	●	●	●	6	100	●	○
73	WHI-OS	9	39	●	●	●	●	●	●	6	100	●	○
104	WHS	9	39	●	●	●	●	●	●	6	100	●	●
96	CHS	8	35	●	●	●	●	○	●	5	83	●	○
118	FHS	8	35	●	●	●	●	○	●	5	83	●	○
85	ARIC	8	35	●	●	●	○	○	○	3	50	●	○
94	CHS	8	35	●	●	●	●	○	●	5	83	●	○
100	MDCS	8	35	●	●	○	●	●	○	4	67	●	○
98	CHS	8	35	●	○	○	●	●	●	4	67	●	○
101	GPPS	8	35	●	●	●	●	○	●	5	83	●	●
75	COSM	8	35	●	●	●	●	○	●	5	83	●	○
79	SMC	8	35	●	●	●	●	○	●	5	83	●	○
112	TS	8	35	●	●	○	●	●	●	5	83	●	○
139	D-EHR	7	30	●	●	○	●	●	●	5	83	●	○
76	CIRCS	7	30	●	●	●	●	●	●	6	100	●	○
134	NPMS	7	30	●	●	○	●	○	●	4	67	○	○
63	SMC	7	30	●	●	●	●	○	●	5	83	●	○

[63]	COSM	7	30	●	●	●	●	○	●	5	83	●	○
[122]	RS	7	30	●	●	●	●	●	●	6	100	○	○
[83]	MPP	7	30	●	●	●	●	●	●	6	100	○	○
[123]	RS	7	30	●	●	○	●	●	●	5	83	●	●
[138]	D-EHR	6	26	●	●	○	●	●	●	5	83	●	○
[78]	NPMS	6	26	●	●	○	●	○	●	4	67	○	○
[114]	FHS	6	26	●	●	●	●	○	●	5	83	●	○
[137]	HCUP	6	26	●	●	○	●	○	●	4	67	●	○
[111]	NorPD	6	26	●	●	●	○	○	○	3	50	●	○
[99]	MDCS	6	26	●	●	●	●	○	●	5	83	●	○
[135]	OCS	6	26	●	●	○	●	○	○	3	50	○	○
[143]	T-EHR	5	22	●	●	○	●	●	●	5	83	●	○
[90]	CHS	5	22	●	●	○	●	○	●	4	67	●	○
[90]	ARIC	5	22	●	●	○	●	○	●	4	67	●	○
[142]	S-EHR	5	22	●	●	○	○	○	●	3	50	○	○
[128]	BHS	5	22	●	●	○	●	○	○	3	50	○	○
[117]	FHS	5	22	●	●	●	●	○	●	5	83	●	○
[140]	D-EHR	4	17	●	○	○	●	○	○	2	33	●	●
[89]	ARIC	3	13	●	●	○	○	○	○	2	33	○	○
[141]	D-EHR	3	13	●	●	○	○	○	○	2	33	○	○
[92]	ARIC	3	13	●	●	○	○	○	○	2	33	○	○
[40]	AGES	2	9	●	●	○	○	○	○	2	33	○	○
[40]	RS	2	9	●	●	○	○	○	○	2	33	○	○
[40]	CHS	2	9	●	●	○	○	○	○	2	33	○	○
[40]	ARIC	2	9	●	●	○	○	○	○	2	33	○	○
[40]	FHS	2	9	●	●	○	○	○	○	2	33	○	○
[93]	CHS	2	9	●	●	○	○	○	○	2	33	○	○
[115]	FHS	2	9	●	●	○	○	○	○	2	33	○	○
[121]	AGES	2	9	●	●	○	○	○	○	2	33	○	○
[127]	L85PS	2	9	●	●	○	○	○	○	2	33	○	○
[127]	SHIP	2	9	●	●	○	○	○	○	2	33	○	○
[127]	BHS	2	9	●	●	○	○	○	○	2	33	○	○
[127]	HABC	2	9	●	●	○	○	○	○	2	33	○	○
[77]	NPMS	2	9	●	●	○	○	○	○	2	33	○	○
[130]	MESA	2	9	●	●	○	○	○	○	2	33	○	○
Total:				**84**	**80**	**49**	**63**	**32**	**59**			**54**	**15**
%				100	95	58	75	38	70			64	18

Abbreviations: LDL - low-density lipoprotein cholesterol, HDL - high density lipoprotein cholesterol, CVD - cardiovascular disease, ● - yes, ○ - no. Risk factor adjustment refers to whether adjustment was made for the 23 risk factors under review, 6 CVD risk factors (age, sex, smoking, blood pressure (i.e. any of systolic blood pressure, diastolic blood pressure, hypertension, or blood pressure lowering medication), lipids (i.e. any of total cholesterol, low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, triglycerides, hyperlipidaemia, or lipid lowering medication), and diabetes mellitus), and prevalent, and incident CVD events. For cohorts abbreviations see table S2.1.

Figure S2.2 Association of age and incidence of atrial fibrillation: 15 reports from 8 countries with 143 336 events

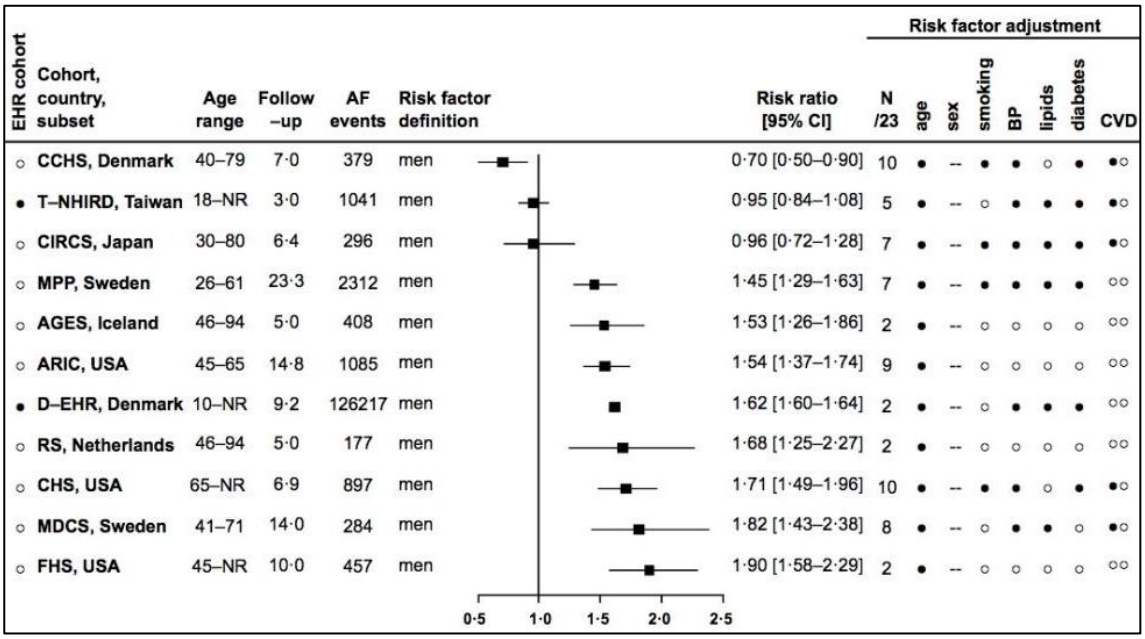


Abbreviations: EHR - electronic health record, age range in years, follow-up in years (mean, median, or maximum), AF - atrial fibrillation, CI - confidence interval, N/23 - number (of factors) out of 23, CVD - cardiovascular disease, SD - standard deviation, NR - not reported, USA - United States of America, ● - yes, ○ - no, -- - not applicable.

Risk factor adjustment refers to whether adjustment was made for the 23 risk factors under review, 6 CVD risk factors, and prevalent and incident CVD events. Example: CIRCS adjusted for 7/23 factors, age was not applicable as main effect, sex, smoking, blood pressure (i.e. any of systolic blood pressure, diastolic blood pressure, hypertension, or blood pressure lowering medication), lipids (i.e. any of total cholesterol, low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, triglycerides, hyperlipidaemia, or lipid lowering medication) and diabetes mellitus, and prevalent, but not incident CVD events. For cohort abbreviations see **table S2.1**.

References pertaining to each report: CIRCS,[133] T-NHIRD,[143] DCHS,[81] RS,[40] AGES,[40] WHI-OS,[74] ARIC,[89] MDCS,[100] FHS,[115] D-EHR,[139] OCS,[135] MPP,[83] NPMS,[78] CHS,[95] TS.[112]

Figure S2.3 Association of sex and incidence of atrial fibrillation: 11 reports from 7 countries with 133 553 events

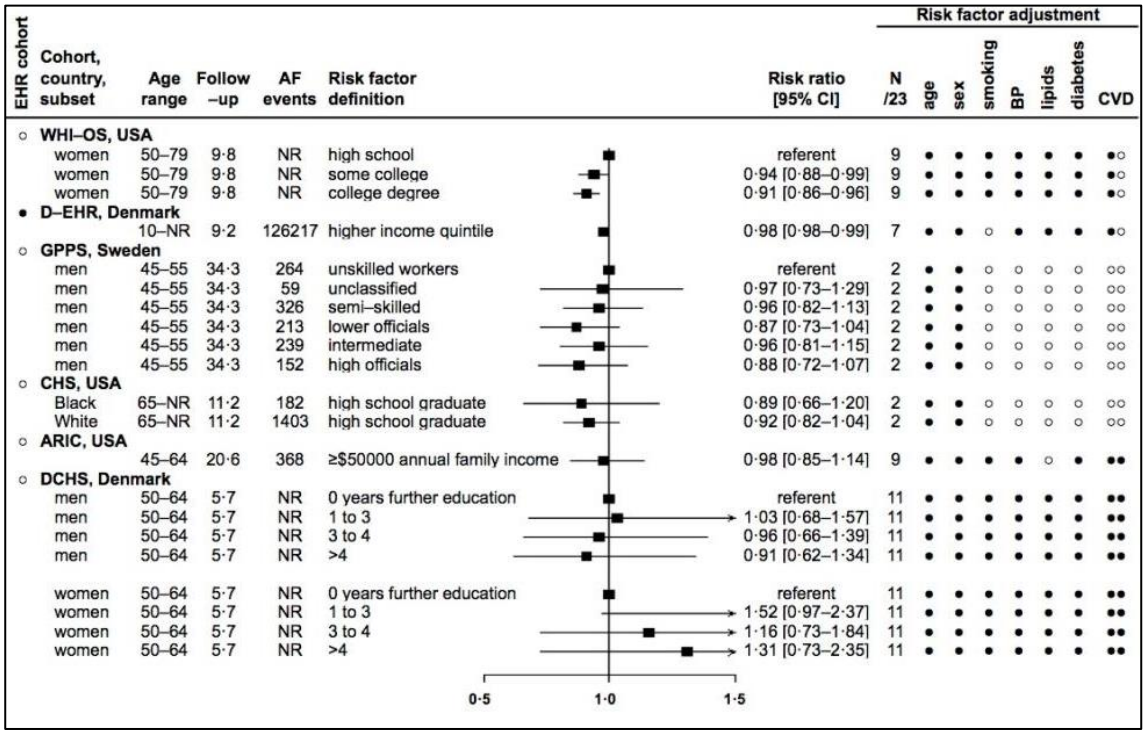


Abbreviations: see **figure S2.2**.

Notes: estimate for MDCS was inverted using formula exp{-ln(estimate)}, reported estimate for female sex: 0·55 [0·42-0·70].

References pertaining to each report: CCHS,[124] T-NHIRD,[143] CIRCS,[133] MPP,[83] AGES,[40] ARIC,[89] D-EHR,[139] RS,[40] CHS,[95] MDCS,[100] FHS.[115]

Figure S2.4 Association of socio-economic status and incidence of atrial fibrillation: 6 reports from 3 countries with 139 654 events

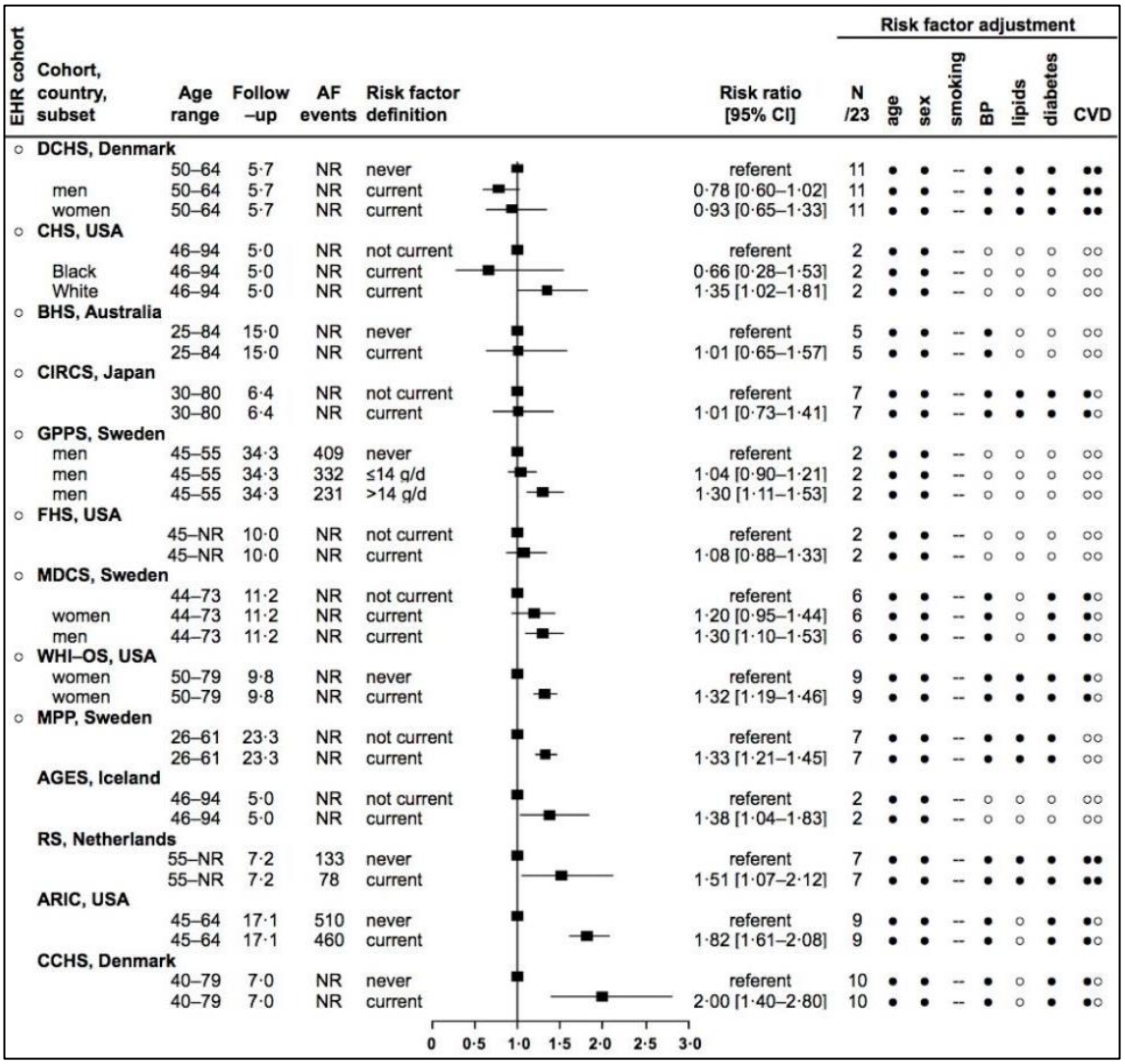


Notes: estimate for ARIC was inverted using formula exp{-ln(estimate)}, reported estimate for <$25000 (vs. ≥50000): 1·02 [0·88-1·17].

Abbreviations: see **figure S2.2**.

References pertaining to each report: WHI-OS,[74] D-EHR,[139] GPPS,[101] CHS,[93] ARIC,[84] DCHS.[81]

Figure S2.5 Association of current smoking and incidence of atrial fibrillation: 13 reports from 7 countries with 18 198 events

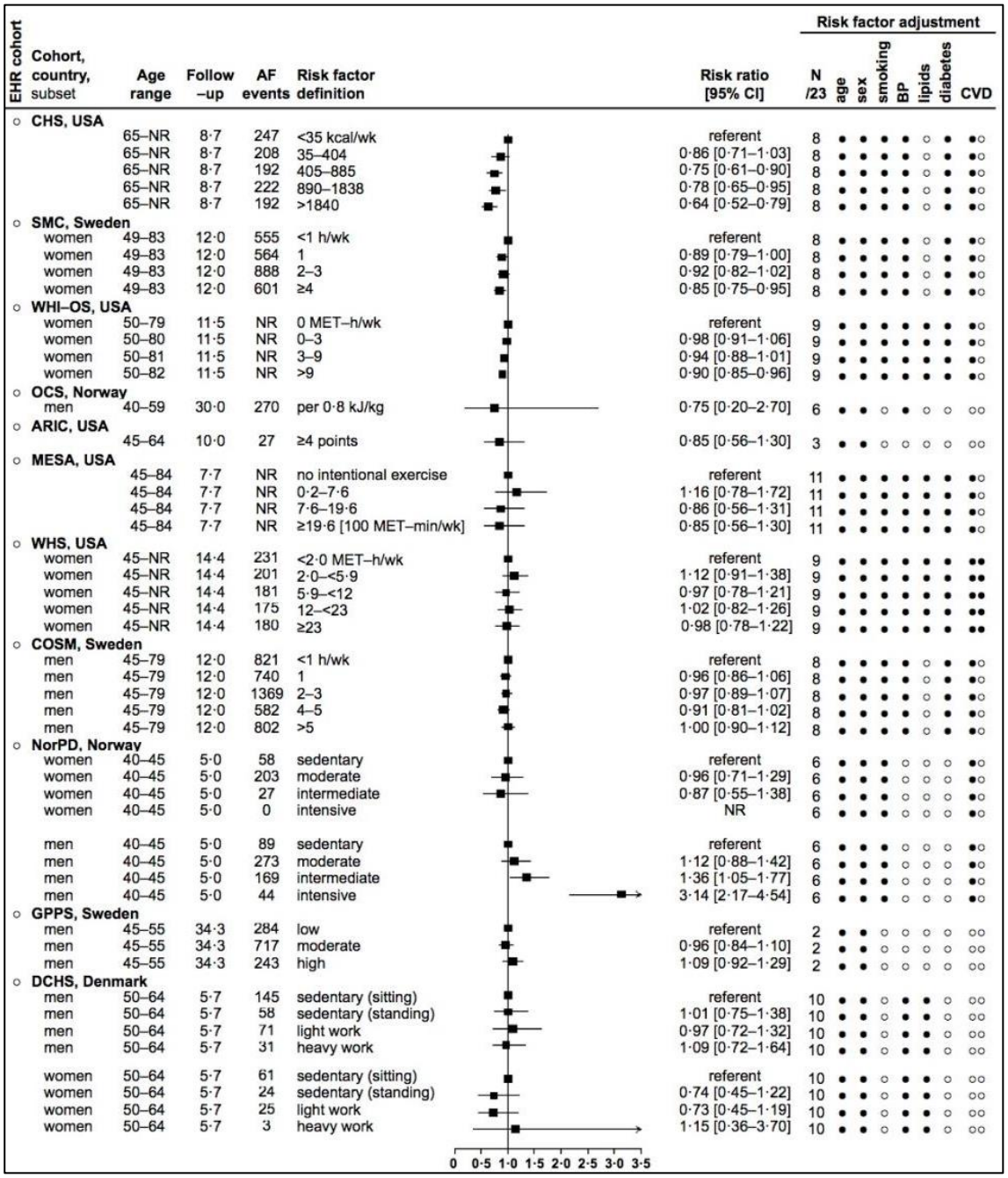


Notes: estimate for ARIC was inverted using formula exp{-ln(estimate)}, reported estimate for never (vs. current) smoking: 0·55 [0·48-0·62].

Abbreviations: see **figure S2.2**.

References pertaining to each report: DCHS,[81] CHS,[40] BHS,[128] CIRCS,[133] GPPS,[101] FHS,[115] MDCS,[99] WHI-OS,[74] MPP,[83] AGES,[40] RS,[123] ARIC,[86] CCHS.[124]

Figure S2.6 Association of physical activity and incidence of atrial fibrillation: 11 reports from 5 countries with 22 822 events

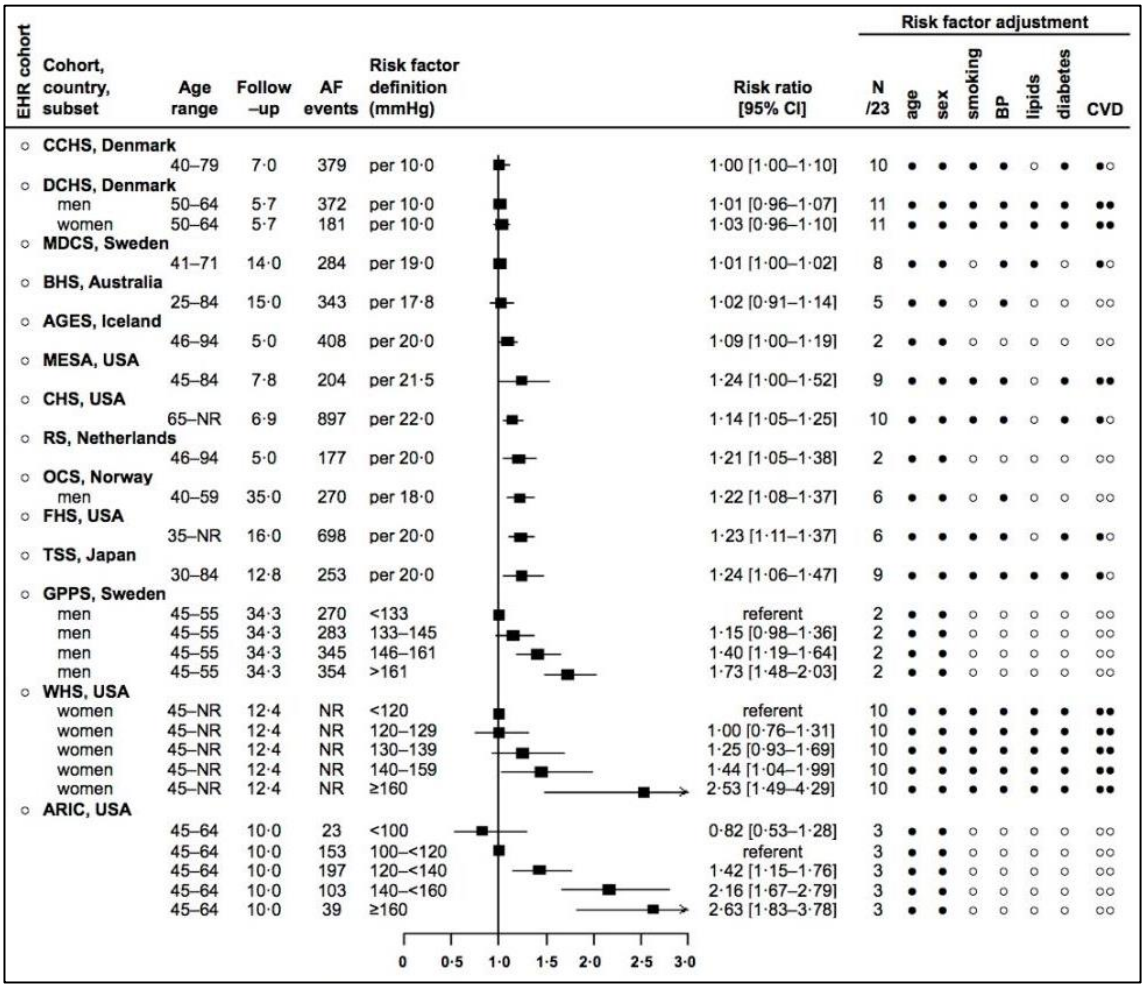


Notes: estimate for ARIC was inverted using formula exp{-ln(estimate)}, reported estimate for <2.0 points on sports index (vs. ≥4·0): 1·17 [0·77-1·78].

Abbreviations: see **figure S2.2** and kJ/kg - kilojoules per kilogram, kcal/wk - kilocalories per week, h-wk - hours per week, MET-h/wk - metabolic equivalent task hours per week.

References pertaining to each report: CHS,[94] SMC,[79] WHI-OS,[73] OCS,[135] ARIC,[92] MESA,[131] WHS,[104] COSM,[75] NorPD,[111] GPPS,[101] DCHS.[82]

Figure S2.7 Association of systolic blood pressure and incidence of atrial fibrillation: 14 reports from 8 countries with 6981 events

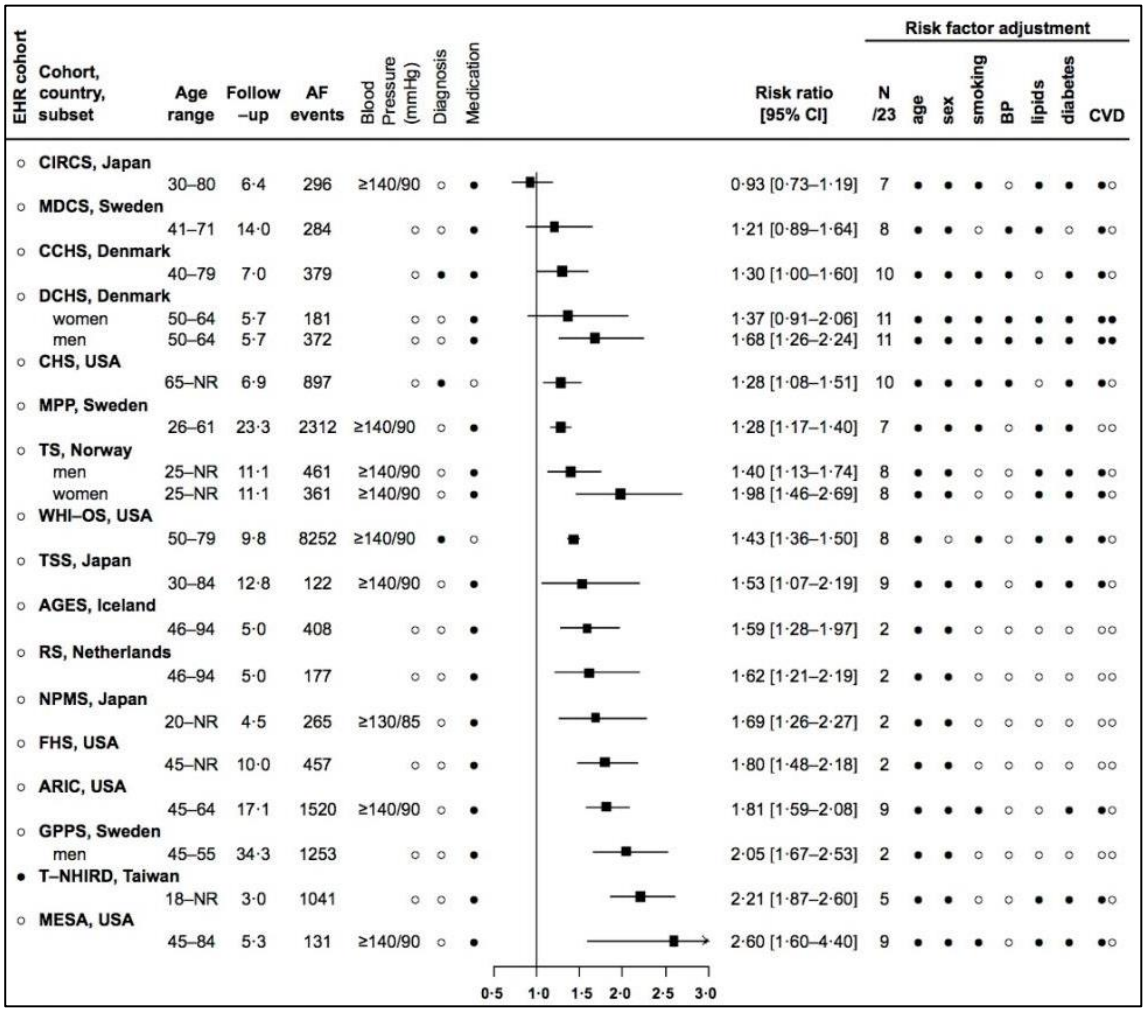


Notes: risk factor adjustment for BP in this instance refers to whether diastolic blood pressure, hypertension, or blood pressure lowering medication were adjusted for in addition to the risk factor definition used for systolic blood pressure.

Abbreviations: see figure S2.2 and mmHg - millimetres of mercury.

References pertaining to each report: CCHS,[124] DCHS,[81] MDCS,[100] BHS,[128] AGES,[40] MESA,[129] CHS,[95] RS,[40] OCS,[135] FHS,[114], TSS,[136] GPPS,[101], WHS,[110], ARIC.[92]

Figure S2.8 Association of hypertension and incidence of atrial fibrillation: 17 reports from 8 countries with 19 169 events

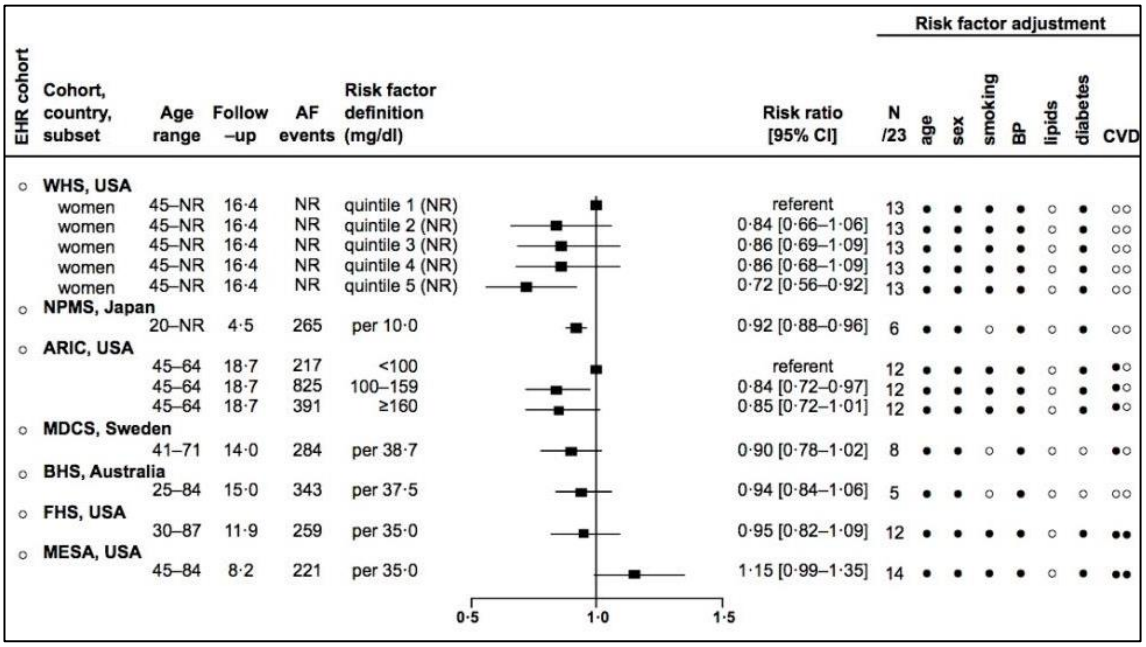


Notes: risk factor adjustment for BP in this instance refers to whether systolic blood pressure, diastolic blood pressure, or blood pressure lowering medication were adjusted for in addition to the risk factor definition used for hypertension. Estimate from ARIC was inverted using formula exp{-ln(estimate)}, reported estimate for blood pressure <120/80 (vs. ≥140/90): 0·55 [0·48-0·63].

Abbreviations: see **figure S2.2** and mmHg - millimetres of mercury.

References pertaining to each report: CIRCS,[133] MDCS,[100] CCHS,[124] DCHS,[81] CHS,[95] MPP,[83] TS,[112] WHI-OS,[74] TSS,[136] AGES,[40] RS,[40] NPMS,[77] FHS,[115] ARIC,[86] GPPS,[101] T-NHIRD,[143] MESA.[132]

Figure S2.9 Association of low-density lipoprotein cholesterol and incidence of atrial fibrillation: 7 reports from 4 countries with 3600 events

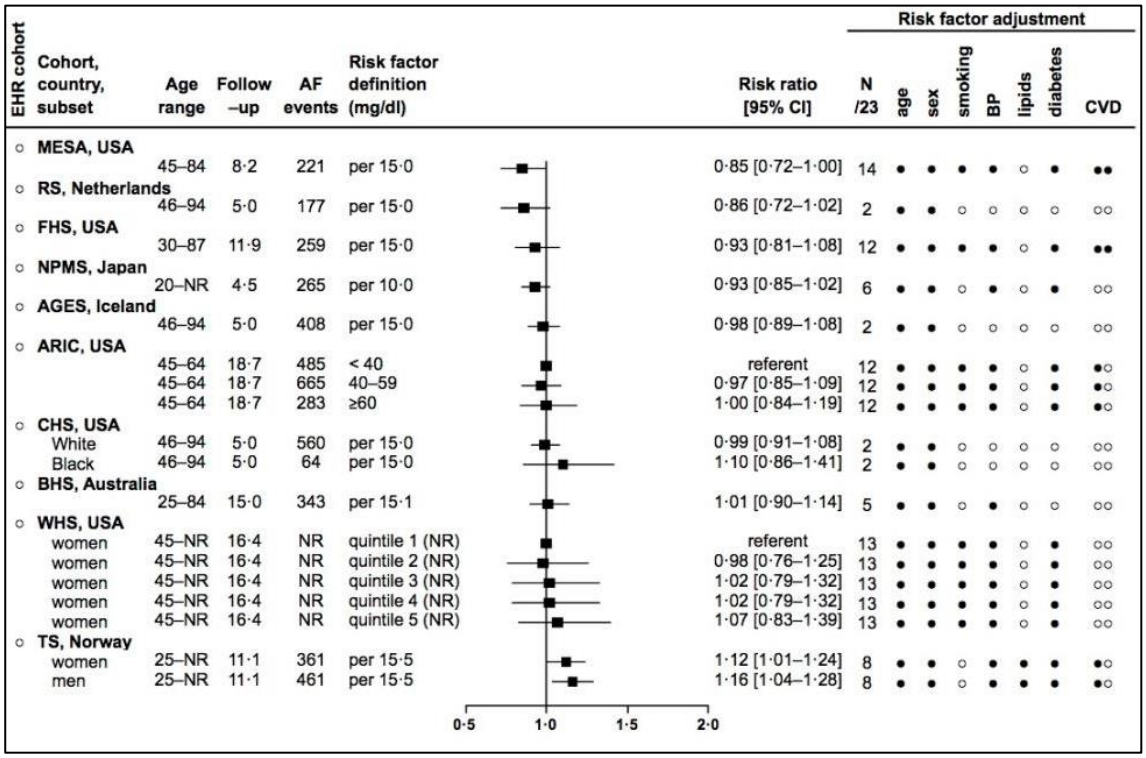


Notes: risk factor adjustment for lipids in this instance refers to whether total cholesterol, high-density lipoprotein cholesterol, triglycerides, hyperlipidaemia, or lipid lowering medication were adjusted for. Low-density lipoprotein reported as mmol/l for MDCS and BHS were converted to mg/dl using the conversion 1mmol/l = 38·66976 mg/dl.

Abbreviations: see **figure S2.2** and mg/dl - milligrams per decilitre, mmol/l - millimoles per litre.

References pertaining to each report: WHS,[106] NPMS,[78] ARIC,[87] MDCS,[100] BHS,[128] FHS,[116] MESA.[116]

Figure S2.10 Association of high-density lipoprotein cholesterol and incidence of atrial fibrillation: 10 reports from 6 countries with 5347 events

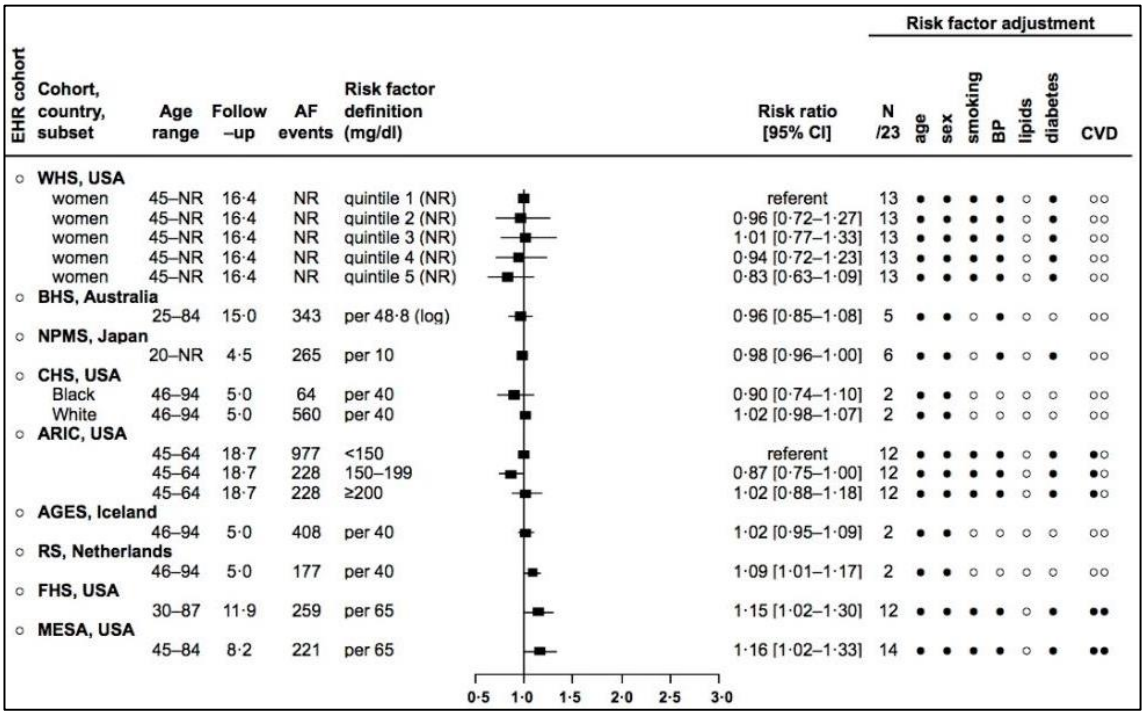


Notes: risk factor adjustment for lipids in this instance refers to whether total cholesterol, low-density lipoprotein cholesterol, triglycerides, hyperlipidaemia, or lipid lowering medication were adjusted for. High-density lipoprotein reported as mmol/l for BHS and TS were converted to mg/dl using the conversion 1mmol/l = 38·66976 mg/dl. Estimate for NPMS was converted from a unit decrease to a unit increase using the formula 1/[23], reported estimate for unit increase in high-density lipoprotein: 1·08 [0·98-1·18].

Abbreviations: see **figure S2.2** and mg/dl - milligrams per decilitre, mmol/l - millimoles per litre.

References pertaining to each report: MESA, [116] RS,[40] FHS,[116] NPMS,[78] AGES,[40] ARIC,[87] CHS,[40] BHS,[128] WHS,[106] TS.[112]

Figure S2.11 Association of triglycerides and incidence of atrial fibrillation: 9 reports from 5 countries with 4525 events

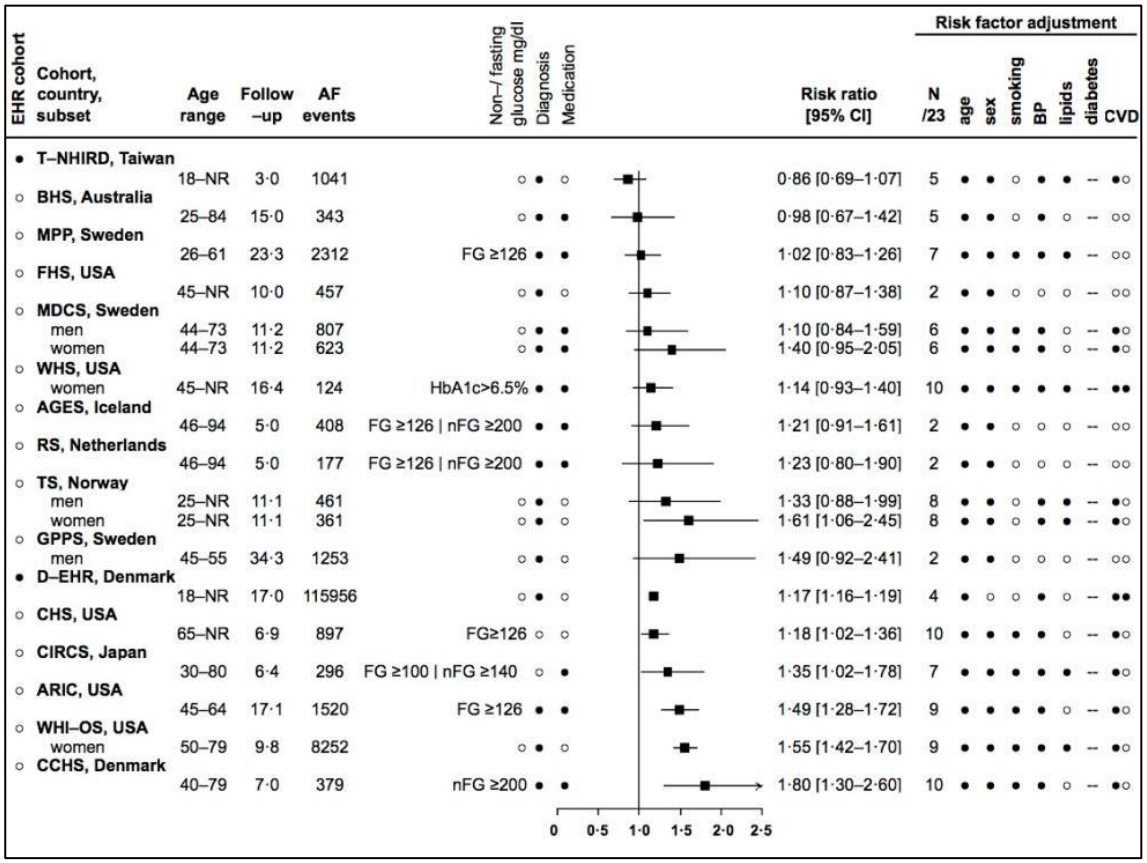


Notes: risk factor adjustment for lipids in this instance refers to whether total cholesterol, low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, hyperlipidaemia, or lipid lowering medication were adjusted for. Triglyceride levels reported as mmol/l for BHS were converted to mg/dl using the conversion 1mmol/l = 88·57396 mg/dl.

Abbreviations: see **figure S2.2** and mg/dl - milligrams per decilitre, mmol/l - millimoles per litre, (log) - log transformed.

References pertaining to each report: WHS,[106] BHS,[128] NPMS,[78] CHS,[40] ARIC,[87] AGES,[40] RS,[40] FHS,[116] MESA.[116]

Figure S2.12 Association of diabetes mellitus (type unspecified) and incidence of atrial fibrillation: 16 reports from 8 countries with 135 667 events

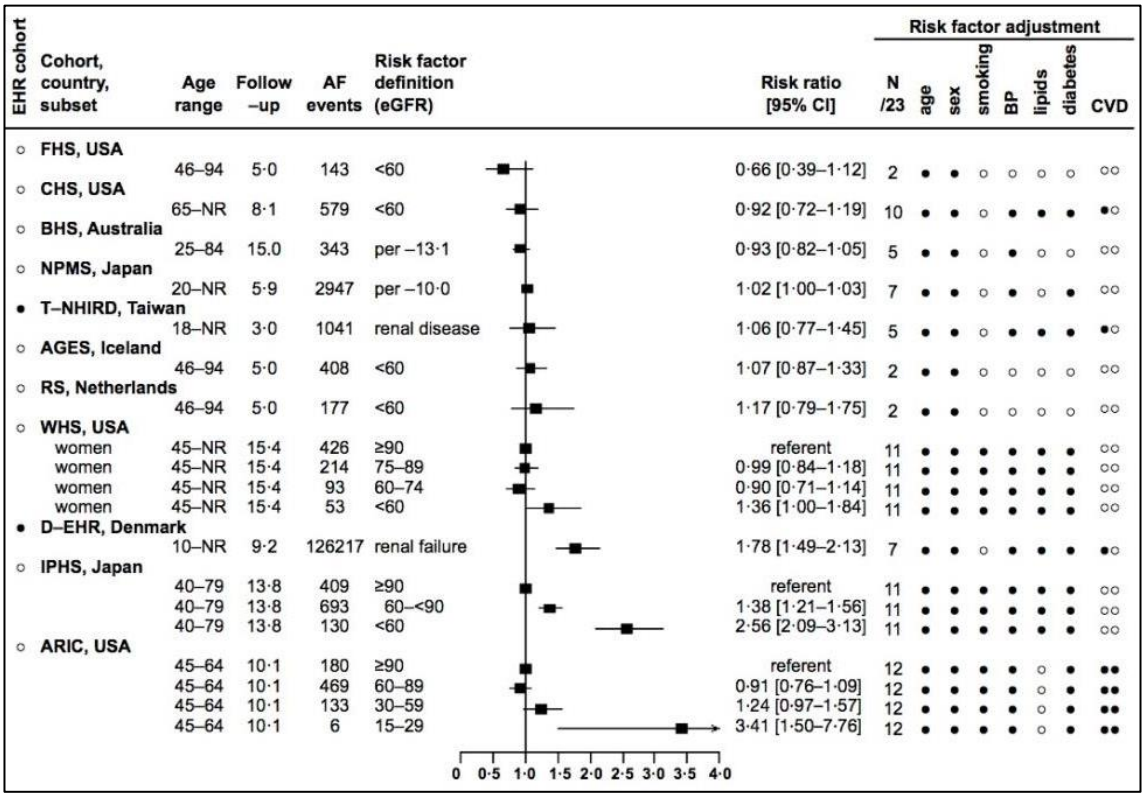


Notes: estimate for ARIC was inverted using the formula exp{-ln(estimate)}, reported estimate for FG<100 (vs. ≥126): 0·67 [0·58-0·78].

Abbreviations: see **figure S2.2** and mg/dl - milligrams per decilitre, FG - fasting serum glucose, nFG - non-fasting serum glucose, HbA1c - hemoglobin A1c.

References pertaining to each report: T-NHIRD,[143] BHS,[128] MPP,[83] FHS,[115] MDCS,[99] WHS,[103] AGES,[40] RS,[40] TS,[112] GPPS,[101] D-EHR,[140] CHS,[95] CIRCS,[133] ARIC,[86] WHI-OS,[74] CCHS.[124]

Figure S2.13 Association of impaired renal function and incidence of atrial fibrillation: 11 reports from 7 countries with 134 661 events

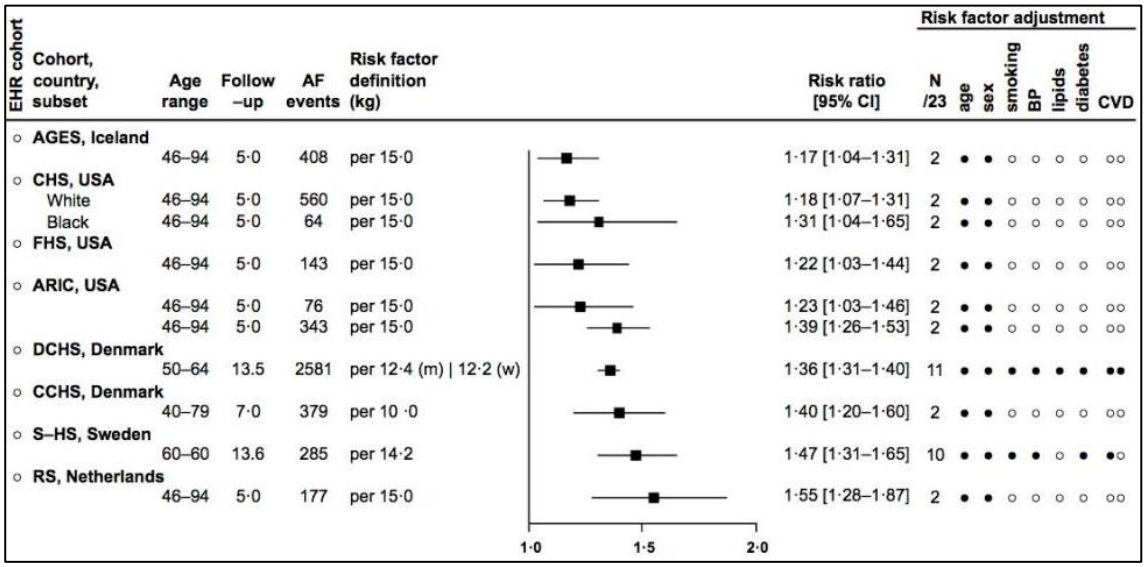


Notes: estimate for BHS was inverted from unit increase to unit decrease using the formula 1/{estimate}, reported estimate for unit increase in eGFR: 1·08 [0·95-1·22].

Abbreviations: see **figure S2.2** and eGFR - estimated glomerular filtration rate.

References pertaining to each report: FHS,[40] CHS,[97] BHS,[128] NPMS,[76] T-NHIRD,[143] AGES,[40] RS,[40] WHS,[107] D-EHR,[139] IPHS,[102] ARIC.[91]

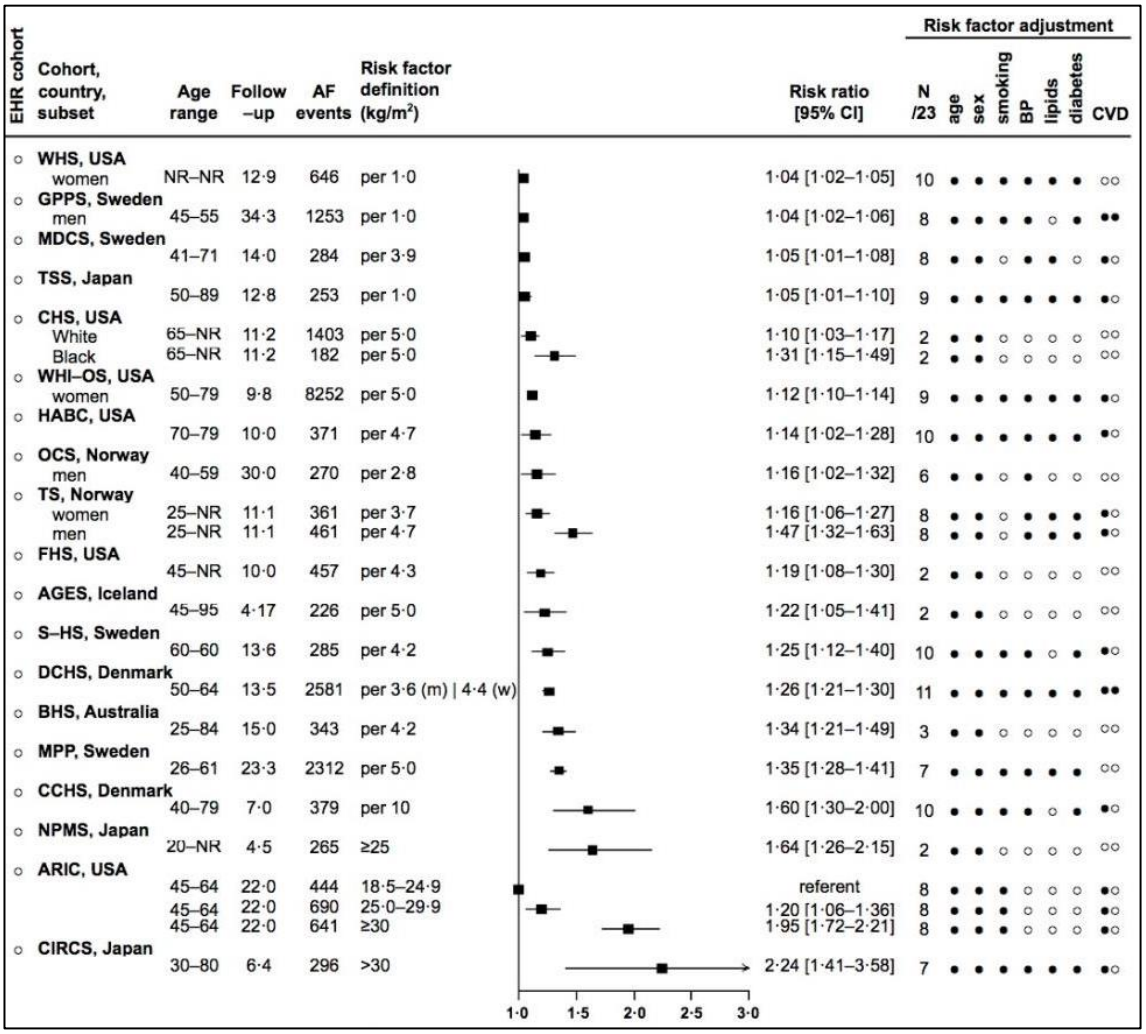
Figure S2.14 Association of weight and incidence of atrial fibrillation: 8 reports from 5 countries with 5016 events



Abbreviations: see **figure S2.2** and kg - kilograms, (m) - men, (w) - women.

References pertaining to each report: AGES,[40] CHS,[40] FHS,[40] ARIC,[40] DCHS,[80] CCHS,[124] S-HS,[134] RS.[40]

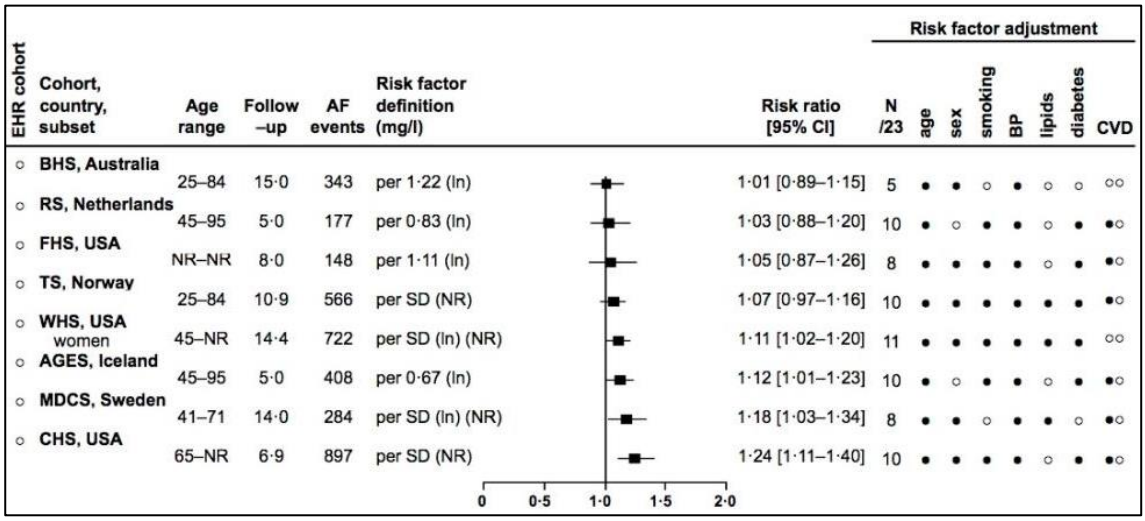
Figure S2.15 Association of body mass index and incidence of atrial fibrillation: 19 reports from 7 countries with 22 655 events



Abbreviations: see **figure S2.2** and kg/m2 - kilograms per metre squared, (w) - women, (m) - men.

References pertaining to each report: WHS,[105] GPPS,[101] MDCS,[100] TSS,[136] CHS,[93] WHI-OS,[74] HABC,[126] OCS,[135] TS,[112] FHS,[115] AGES,[121] S-HS,[134] DCHS,[80] BHS,[128] MPP,[83] CCHS,[124] NPMS,[77] ARIC,[85] CIRCS.[133]

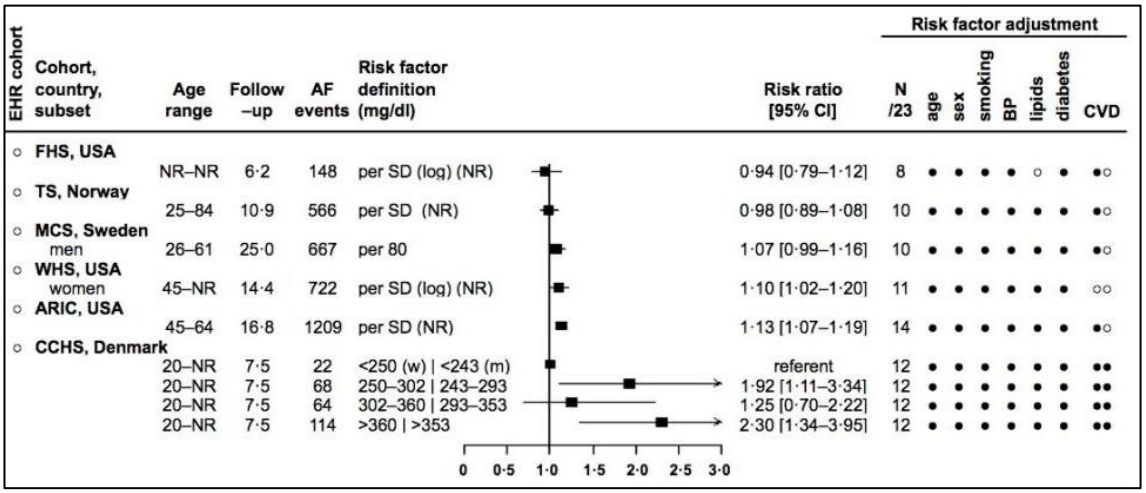
Figure S2.16 Association of C-reactive protein and incidence of atrial fibrillation: 8 reports from 6 countries with 3545 events



Abbreviations: see **figure S2.2** and mg/l, milligrams per litre, (ln) - log transformed.

References pertaining to each report: BHS,[128] RS,[120] FHS,[118] TS,[113] WHS,[108] AGES,[120] MDCS,[100] CHS.[95]

Figure S2.17 Association of fibrinogen and incidence of atrial fibrillation: 6 reports from 4 countries with 3580 events

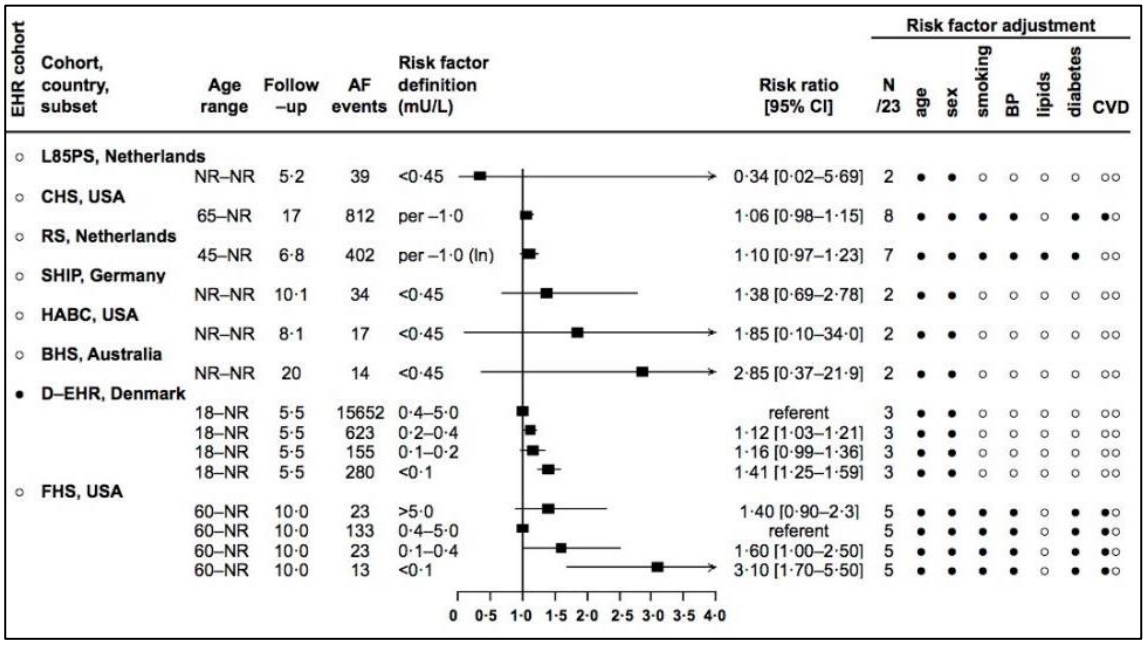


Notes: fibrinogen levels reported as g/l were converted to mg/dl using the conversion 1g/l = 100 mg/dl.

Abbreviations: see **figure S2.2** and mg/l, milligrams per litre, (log) - log-transformed, (w) - women, (m) - men.

References pertaining to each report: FHS, [118] TS,[113] MCS,[119] WHS,[108] ARIC,[88] CCHS.[125]

Figure S2.18 Association of impaired thyroid function and incidence of atrial fibrillation: 8 reports from 5 countries with 18 220 events

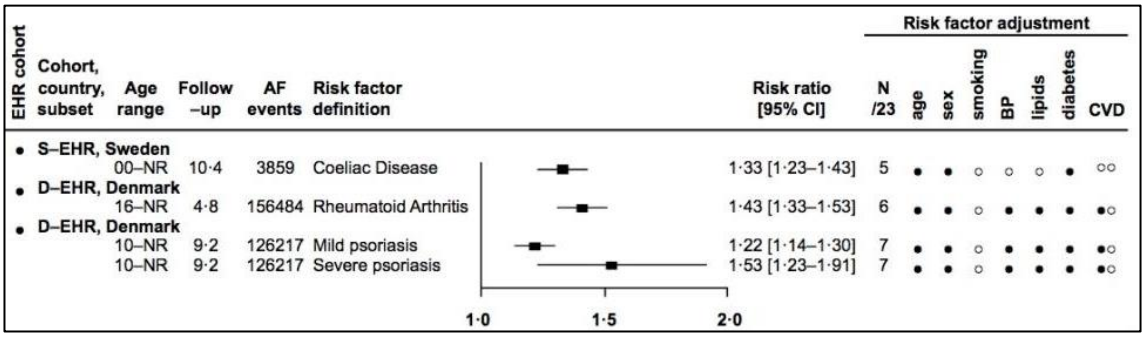


Notes: estimates for CHS and RS were inverted from a unit increase to a unit decrease, reported estimates for unit increase in TSH were 0·94 [0·87-1·02], and 0·91 [0·81-1·03] respectively.

Abbreviations: see **figure S2.2** and TSH - thyroid stimulating hormone, mU/L - milliunits per litre.

References pertaining to each report: L85PS,[127] CHS,[96] RS,[122] SHIP,[127] HABC,[127] BHS,[127] D-EHR,[141] FHS.[117]

Figure S2.19 Association of autoimmune diseases and incidence of atrial fibrillation: 3 reports from 2 countries with 286 560 events



Abbreviations: see **figure S2.2**.

References pertaining to each report: S-EHR,[142] D-EHR,[138] D-EHR.[139]

Figure S2.20 Summary of included reports by risk factor and publication year

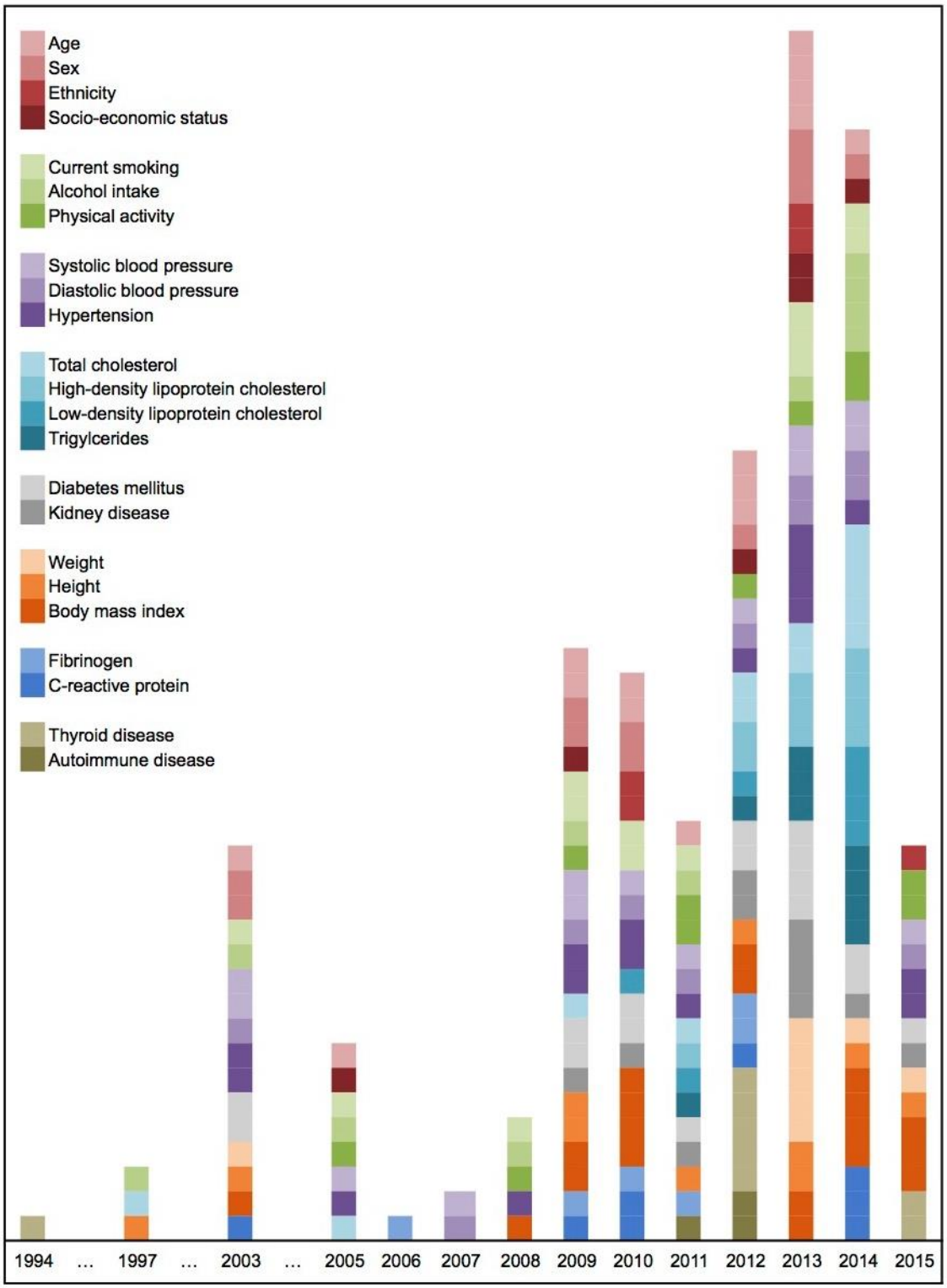


Table S2.5 Systematic review sensitivity analysis of single vs multiple database search

	PubMed	EMBASE
Search terms	("atrial fibrillation"[title] OR "atrial flutter"[title] OR "auricular fibrillation"[title] OR "auricular flutter"[title] OR "cardiac arrhythmia"[title]) AND ("epidemiology" OR "incidence" OR "risk factors" OR "risk score" OR "risk assessment") AND ("prospective studies" OR "cohort" OR "observational") AND ("2013-01-01"[Date - Publication] : "2013-12-31"[Date - Publication])	(atrial fibrillation or atrial flutter or auricular fibrillation or auricular flutter or cardiac arrhythmia).ti. and (epidemiology or incidence or risk factors or risk score or risk assessment).af. and (prospective studies or cohort or observational).af. and "2013".yr.
Overall publications retrieved	408	331
Exclusions based on duplicate titles and 'online first' articles published in 2014	-81	-10
Title review	327	321
Exclusions based on title	-319	-304
Abstract review	8	17
Exclusions based on abstract	-2 2 not general population cohorts	-11 2 not general population cohorts 9 conference abstracts only
Full text review	6	6
Final selection	6 + 1 from hand search 1. Alonso, et al. Journal of the American Heart Association 2013; 2(2): e000102 2. Chiang, et al. International Journal of Cardiology 164(2): 201-204. 3. Dewland, et al. Circulation 128(23): 2470-2477. 4. Nyrnes, et al. European Journal of Preventive Cardiology 20(5): 729-736. 5. Perez, et al. Heart 99(16): 1173-1178. 6. Thelle, et al. Heart 99(23): 1755-1760. Hand search: 7. Jensen et al. Journal of the American Geriatrics Society 2013; 61(2): 276-80.	6 + 1 from hand search 1. Alonso, et al. Journal of the American Heart Association 2013; 2(2): e000102 2. Chiang, et al. International Journal of Cardiology 164(2): 201-204. 3. Dewland, et al. Circulation 128(23): 2470-2477. 4. Nyrnes, et al. European Journal of Preventive Cardiology 20(5): 729-736. 5. Perez, et al. Heart 99(16): 1173-1178. 6. Thelle, et al. Heart 99(23): 1755-1760. Hand search: 7. Jensen et al. Journal of the American Geriatrics Society 2013; 61(2): 276-80.

Chapter 3

Clinical research using linked bespoke studies and electronic health records (CALIBER): description of data sources, motivation for use, and analytic considerations in research on atrial fibrillation

2016 Max Perutz Science Writing Award shortlisted entry:

Preventing a heart that goes ba-boom, ba-, ba-, ba- ,-boom, ba-boom

Your heart is a mighty engine. Sitting in the centre left of your upper chest, it beats tirelessly to ensure that your brain, kidney, liver and lungs are all adequately fuelled. Size-wise, it's about as small as two clenched fists. Structurally, it consists of pumping chambers, values, and pipework, and is powered by a series of electrical impulses. Each component of the heart's system works together in an orderly sequence: ba-boom, ba-boom, ba-boom, baboom.

Unfortunately, as with all great feats of engineering, the heart malfunctions sometimes. A blocked pipe: heart attack. A weakened pump: heart failure. An electrical fault causing the heart to beat in a rapid or disorganised manner … baboom, ba-, ba-, ba- , -boom, ba-boom: Atrial Fibrillation.

Yet despite being the world's most common heartbeat disorder, atrial fibrillation is less well known to people than other cardiovascular diseases such as heart attack and heart failure. Indeed, the first time you are likely to hear of atrial fibrillation is when either you or your family member are diagnosed with it. Atrial fibrillation is so common that the risk of developing it in your lifetime is one in four.

Sadly, living with atrial fibrillation doesn't just mean having a heart that goes baboom, ba-, ba-, ba- , -boom, ba-boom. Living with atrial fibrillation means living with debilitating symptoms like chronic fatigue, shortness of breath, and heart palpitations. Living with atrial fibrillation means living with a leading risk factor for suffering a stroke. Living with atrial fibrillation means living with lifelong medications.

Atrial fibrillation can devastate lives and therefore preventing people from developing it is a public health priority. Public health campaigns aimed at quitting smoking, improving diet, and increasing exercise have been hugely successful in cutting the number of people who have heart attacks. But where are the public health campaigns for atrial fibrillation? How can you or I reduce our own personal risks of developing atrial fibrillation?

The truth is there aren't any public health campaigns for atrial fibrillation. The risk factors for developing atrial fibrillation are not very well defined. There have been no clinical trials testing prevention strategies for atrial fibrillation, as researchers do not know which interventions to test, nor the types of people to recruit to take part.

Current understanding about risk factor factors for atrial fibrillation is based on findings from cohort studies. Typically, these studies involve several thousand people who volunteer information about their health behaviours, environment, and medical history; they might be screened for a range of physical and biological measures and then they are followed forward through time to see who develops the disease of interest.

My research takes a different approach to investigating risk factors for atrial fibrillation, and involves the use of electronic health records. Electronic health records concern the digital collection of people's health and health-related information. They are collected routinely each time you or I visit our doctor or attend a hospital appointment. Electronic health records can contain symptoms, diagnoses, drug prescriptions, operations, procedures, results of pathological tests, anthropometric measurements, and health behaviours.

Instead of recruiting several thousand people to take part in a cohort study, electronic health records allow a whole country to become a cohort. Instead of being limited to the information collected as part of a cohort study, electronic health records allow a much wider array of factors to be studied, as well as how these factors may develop and progress over time, and how multiple factors may be interrelated.

Traditionally, researchers have formed study hypotheses based on what they think they know and what they expect to find. However, it could be that the risk factors for atrial fibrillation remain unclear because researchers have been looking in the wrong place entirely.

My research therefore aims to use electronic health records to refine understanding about existing risk factors for atrial fibrillation, as well as to discover novel factors that researchers hadn't thought to consider previously. In this way, I hope to stimulate greater awareness about atrial fibrillation, contribute knowledge to help shape atrial fibrillation prevention strategies, and ultimately lead to a future where fewer people suffer from atrial fibrillation, and more people's hearts keep on beating healthily: ba-boom, ba-boom, ba-boom, ba-boom.

Chapter 4

Novel associations between 23 cardiovascular risk factors and incident atrial fibrillation with and without intercurrent cardiovascular disease

Methods used to create an analytic dataset for studying 23 cardiovascular risk factors in relation to incidence of atrial fibrillation

Study inclusion criteria

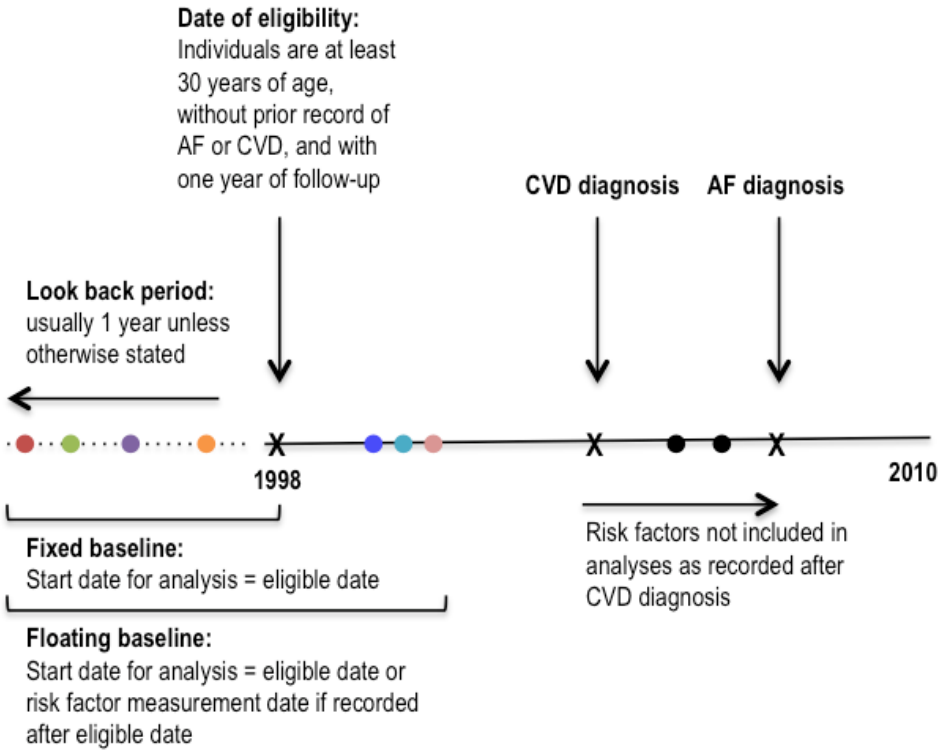
1. at least 30 years of age
2. without prior record of diagnosis of atrial fibrillation
3. without prior record of diagnosis of ten cardiovascular diseases (heart failure, myocardial infarction, unstable angina, stable angina, coronary heart disease, abdominal aortic aneurysm, peripheral arterial disease, haemorrhagic stroke, ischaemic stroke, and transient ischaemic attack)
4. with a minimum of one year follow-up at a GP practice with research quality data* between 1998 and 2010

* GP practice-level data quality is determined by CPRD based on proportion of absences or inconsistencies in the recording of age, gender, practice registration or deregistration information, pregnancy outcomes, and rates of prescriptions, referrals, and deaths, and provides the date at which practices are deemed of research standard.[174]

Approach to selecting risk factor data

Figure S4.1 illustrates my approach to selecting risk factor data. To maximise the use of observed values, I selected data available at baseline (i.e. earliest date study inclusion criteria fulfilled) using a lookback period usually of one year (unless otherwise specified in below definitions), together with data available after baseline but before date of occurrence of atrial fibrillation or any of the 10 cardiovascular diseases listed in the above study inclusion criteria.

Figure S4.1 Illustration of approach used to selecting risk factor data



Risk factor definitions

I used existing EHR definitions for the 23 risk factors and AF cases as available on the CALIBER portal and in published papers as indicated as follows:

Table S4.1 Data definitions for 23 cardiovascular risk factors investigated for associations with incident atrial fibrillation

Risk factor	Description
Age	Age was derived from the difference between date of birth as recorded in GPRD GOLD and date of eligibility, scaled in years [(date index – date birth)/365.25]. Due to data confidentiality, GPRD GOLD provides only birth year information from an individual's full date of birth (DDMMYYYY). Therefore for the purpose of generating an age at entry all individuals were given the birth day 1 January, which individuals are known to be as old as at most. More details are provide on the CALIBER portal: https://www.caliberresearch.org/portal/show/birthyear
Sex[181]	Gender, available from GPRD GOLD, was defined as a categorical variable with two levels: 1 - Men 2 - Women https://www.caliberresearch.org/portal/show/sex_gprd
Ethnicity[182]	Ethnicity, available from GPRD GOLD and HES, was defined as a categorical variable with four levels: 1 - White 2 - South Asian 3 - Black 4 - Other and Mixed Conflicts within and between GPRD GOLD and HES records are dealt with using the CAL-

	IBER algorithm developed by George and colleagues.[181] The recording of self-identifying ethnic group has been a mandatory requirement in UK primary and secondary care since 1991, and therefore data are increasingly complete,[198] and shown to be representative of the UK population, when compared to census data.[181] https://www.caliberresearch.org/portal/show/ethnic_gprd https://www.caliberresearch.org/portal/show/ethnic_hes
Socioeconomic status[183]	Socioeconomic status, available from ONS, was estimated for individuals with a valid post-code in GPRD GOLD using the Index of Multiple Deprivation (IMD). All postcodes were removed before receipt of the data. IMD ranks geographic areas across England from least to most deprived based on indicators of income, employment, education, health, crime, barriers to housing and services, and living environment.[183] Because of non-linearity (IMD is a relative measure), I analysed IMD according to score quintiles: 1 - IMD quintile 1 (least deprived 20%) 2 - IMD quintile 2 3 - IMD quintile 3 4 - IMD quintile 4 5 - IMD quintile 5 (most deprived 20%) https://www.caliberresearch.org/portal/show/imd_score_gprd
Smoking status[184]	Smoking status, recorded in GPRD GOLD, was defined as a categorical variable with three levels: 1 - Non-smoker 2 - Former smoker 3 - Current smoker Smoking status has been validated against smoking prevalence estimates reported in the Health survey for England (an annual survey of health and health-related behaviours on UK households carried out since 1991).[272] Because smoking status is subject to temporal change, I recategorised non-smokers to former smokers if a prior record indicated a history of smoking. For non- and former smokers, I used smoking status records with no restriction on the look back period (i.e. taken at any time point prior to eligible date) because adult non-current smoking tends to be a stable measure, only updated in the event of change. [178] For current smokers, I used smoking status records up to a maximum of three years prior to eligible date.[184] https://www.caliberresearch.org/portal/show/smoking_status_composite
Alcohol drinker status[185]	**Alcohol drinker status**, recorded in GPRD GOLD, was defined as a categorical variable with five levels: 1 - Non-drinker 2 - Former drinker 3 - Occasional drinker (less than once per week) 4 - Moderate drinker 5 - Heavy drinker (exceeds recommend alcohol intake levels) As in a prior CALIBER study of alcohol consumption,[185] I used a look back period of five years to capture alcohol status, and recategorised non-drinkers as former drinkers if a prior record indicated a history of drinking alcohol. https://www.caliberresearch.org/portal/show/alcohol_drinker_composite
Physical activity	**Physical activity**, recorded in GPRD GOLD, was defined as a categorical variable with four levels: 1 - Inactive 2 - Gentle activity 3 - Moderate activity 4 - Vigorous activity Health behaviours, such as physical activity status, are self-reported by individuals, usually collected as part as new patient registrations and NHS health checks. As for alcohol drinker status, I used a look back period for physical activity status of up to five years. https://www.caliberresearch.org/portal/show/physact_gprd
Systolic blood pressure[148]	Systolic blood pressure (SBP), recorded in GPRD GOLD, was based on measurements taken during primary care consultations and defined as plausible if in the range 20 to 350 millimetres of mercury (mmHg). As in a prior CALIBER,[148] and CPRD study,[273] I used a look back period of two years to capture baseline measurements. https://www.caliberresearch.org/portal/show/bp_gprd

Diastolic blood pressure[148]	Diastolic blood pressure (DBP), recorded in GPRD GOLD, was based on measurements taken during primary care consultations and defined as plausible if in the range 20 to 200 mmHg. As in a prior CALIBER,[148] and CPRD study,[273] I used a look back period of two years to capture baseline measurements. https://www.caliberresearch.org/portal/show/bp_gprd
Hypertension[148]	Hypertension, captured in GPRD GOLD and HES, was based on the CALIBER definition combining: **1 -** coded diagnoses From both the Read and ICD 10 classifications **2 -** repeat high blood pressure measurements At least two measurements within a six month period or three measurements within a one year period of systolic or diastolic blood pressure greater than 140/90 mmHg. **3 -** repeat prescriptions of blood pressure lowering medication At least two prescriptions within a six month period of blood pressure lowering medications referenced in the British National Formulary (pharmacy guidelines published by the British Medical Association and the Royal Pharmaceutical Society). https://www.caliberresearch.org/portal/show/phenotype_ht
Total cholesterol	Total cholesterol, recorded in GPRD GOLD, was based on serum and plasma measurements taken during primary care consultations and defined as plausible if in the range 1-15 millimoles per litre (mmol/L).[186] https://www.caliberresearch.org/portal/show/tot_chol_comp
Low density lipoprotein cholesterol	Low density lipoprotein (LDL) cholesterol, recorded in GPRD GOLD, was based on serum and plasma measurements taken during primary care consultations and defined as plausible if in the range 0.1-10 millimoles per litre (mmol/L). [186] https://www.caliberresearch.org/portal/show/LDL
High density lipoprotein cholesterol	High density lipoprotein (HDL) cholesterol, recorded in GPRD GOLD, was based on serum and plasma measurements taken during primary care consultations and defined as plausible if in the range 0.1-5 millimoles per litre (mmol/L). [186] https://www.caliberresearch.org/portal/show/HDL
Triglycerides	Triglycerides, recorded in GPRD GOLD, was based on serum and plasma measurements taken during primary care consultations and defined as plausible if in the range 0.01-20 millimoles per litre (mmol/L).[186] https://www.caliberresearch.org/portal/show/TRI_gprd
Diabetes mellitus[187]	Diabetes mellitus, recorded in GPRD GOLD and HES, was defined using the CALIBER definition developed by Shah and colleagues,[187] which is a categorical variable with four levels: 1 - No record of diabetes mellitus 2 - Type I diabetes mellitus 3 - Type II diabetes mellitus 4 - Diabetes of uncertain type https://www.caliberresearch.org/portal/show/phenotype_diabetes
Renal disease	Renal failure, recorded in GPRD GOLD and HES, was defined according to presence or absence of Read or ICD 10 codes relating to acute, mild, moderate, and severe failure, as well as dialysis and transplantations. https://www.caliberresearch.org/portal/show/renal_gprd https://www.caliberresearch.org/portal/show/renal_hes
Height[188]	Height, recorded in GPRD GOLD, was based on measurements taken during primary care consultations and defined as plausible if in the range 0.80 to 2.50 metres (m). Because adult height is subject to negligible change, I used the height measurement recorded closest to eligible date but with no restriction on the look back period.[188]
Weight[188]	Weight, recorded in GPRD GOLD, was based on measurements taken during primary care consultations and defined as plausible if in the range 10 to 50 kilograms (Kg). I used the weight measurement recorded closest to eligible date but with a look back period of three years, which has been validated again Health Survey for England data.[188] https://www.caliberresearch.org/portal/show/weight_gprd
Body mass index[188]	Body mass index (BMI), was derived from height and weight measurements (as above), using the universal equation weight/height2 and defined as plausible if in the range 10 to 80

	Kg/m2. I used height and weight measurements recorded on the same date, where available.[188] https://www.caliberresearch.org/portal/show/bmi
C-reactive protein	C-reactive protein (CRP), recorded in GPRD GOLD, was based on serum and plasma measurements taken during primary care consultations and defined as plausible if in the range 0-2000 milligrams per litre (mg/L). https://www.caliberresearch.org/portal/show/c_reactive_protein_gprd
Fibrinogen	Fibrinogen, recorded in GPRD GOLD, was based on serum and plasma measurements taken during primary care consultations and defined as plausible if in the range 0-50 milligrams per litre (mg/L). https://www.caliberresearch.org/portal/show/fibrinogen_gprd
Thyroid disease	Thyroid disease, recorded in GPRD GOLD and HES, was defined according to presence or absence of Read or ICD 10 codes relating to diagnosis of hyperthyroidism and or hypothyroidism. I created a categorical variable with four levels to describe thyroid disease status: 1 - No record of thyroid disease 2 - Hypothyroidism 3 - Hyperthyroidism 4 - Thyroid disease of uncertain type (if records suggest both hypo- and hyperthyroidism) https://www.caliberresearch.org/portal/show/hyperthyroid_gprd, https://www.caliberresearch.org/portal/show/hyperthyroid_hes, https://www.caliberresearch.org/portal/show/hypothyroid_gprd https://www.caliberresearch.org/portal/show/hypothyroid_hes
Rheumatoid arthritis	Rheumatoid arthritis, recorded in GPRD GOLD and HES, was defined according to presence or absence of Read or ICD 10 codes for diagnosis as compiled by Pujades-Rodriguez and colleagues.[189] https://www.caliberresearch.org/portal/show/phenotype_ra
Psoriasis	Psoriasis, recorded in GPRD GOLD and HES, was defined according to presence or absence of Read or ICD 10 codes for diagnosis. https://www.caliberresearch.org/portal/show/phenotype_psoriasis

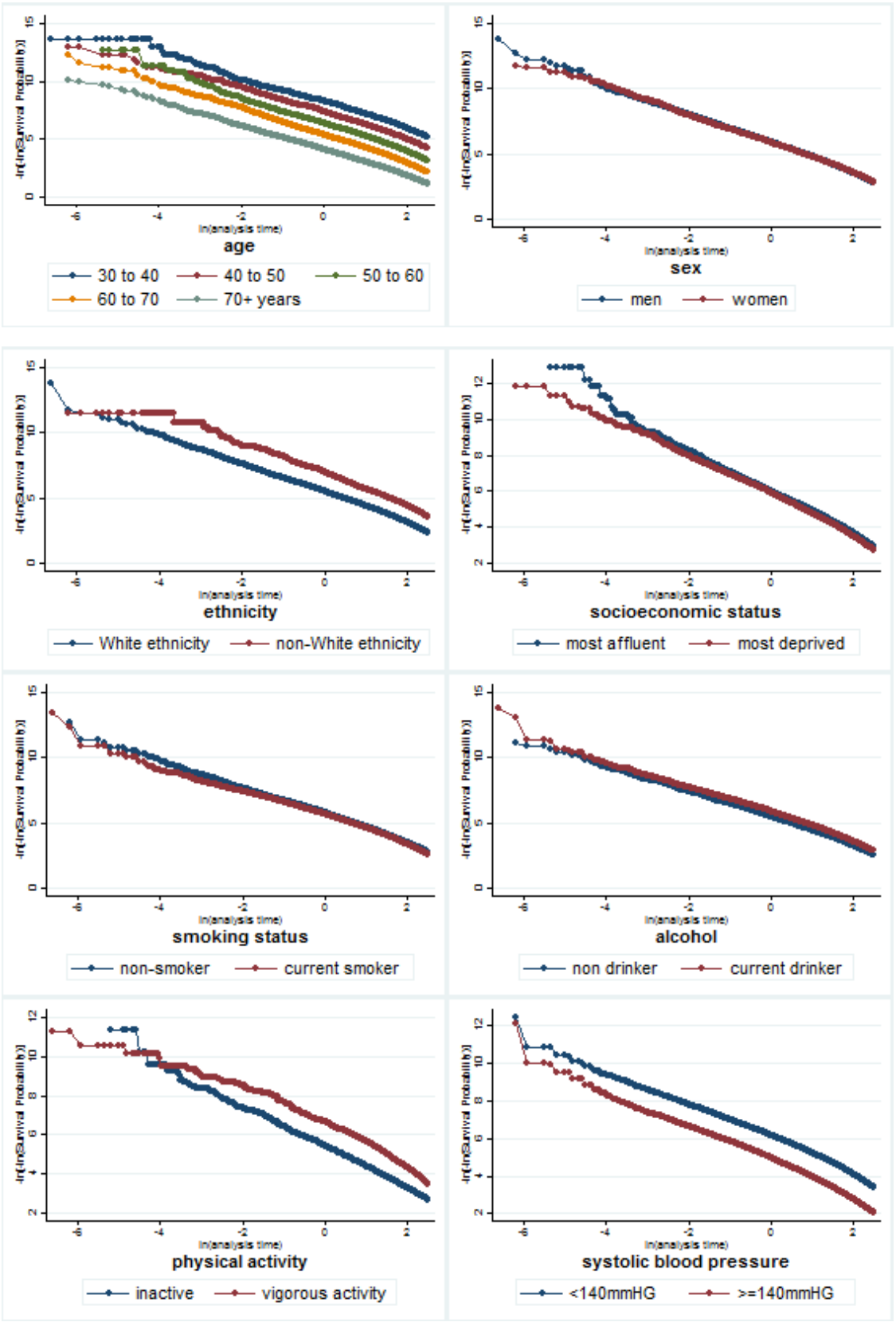
Table S4.2 Characteristics of individuals with data available for 23 cardiovascular risk factors investigated for associations with incident atrial fibrillation

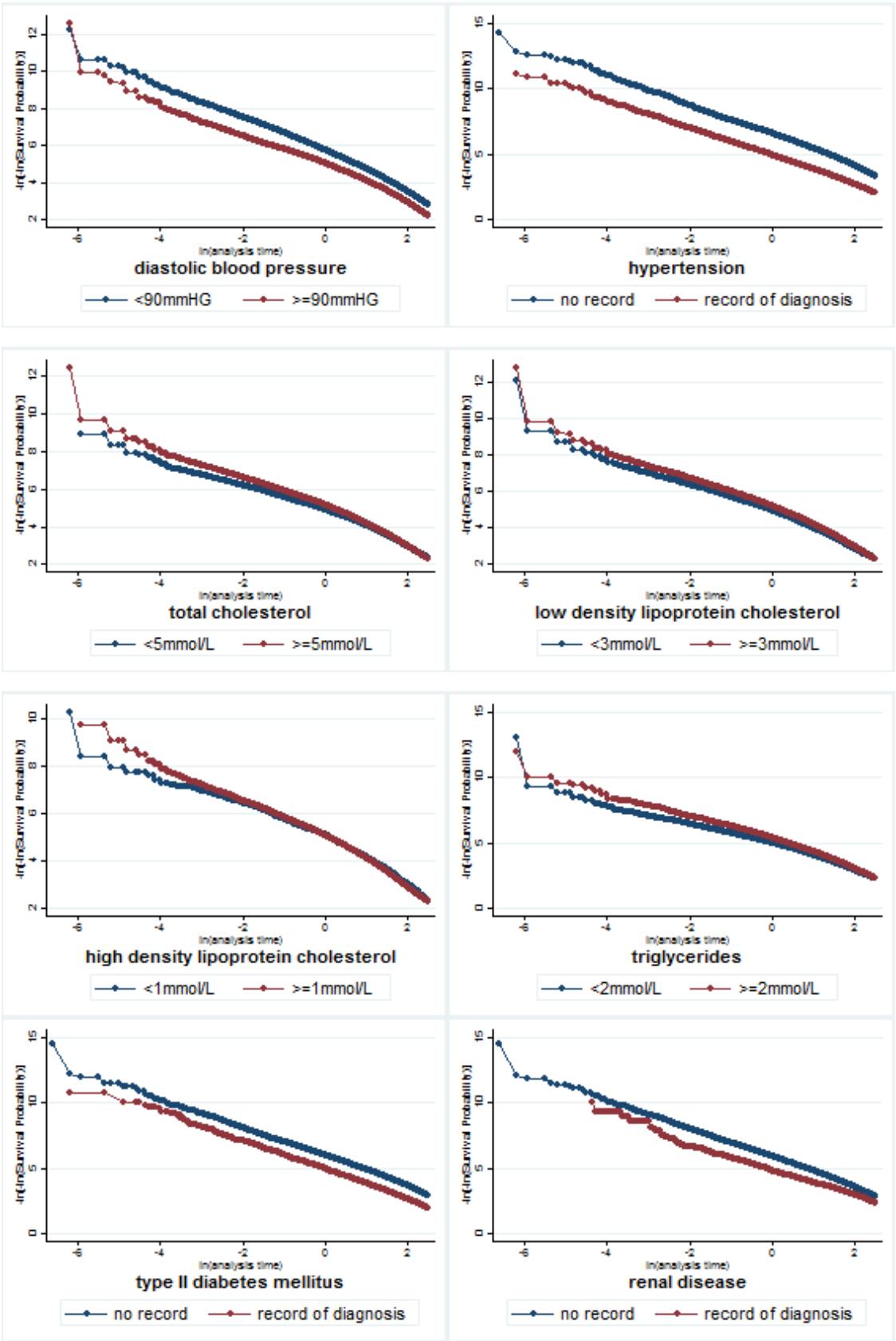
	Individuals N (%)	Age mean (SD)	Male %	Follow-up mean (SD)	AF N (%)
Demographics					
Age	1949052 (100)	46.9 (15.3)	48.7	6.1 (4.2)	50097 (2.6)
Gender	1949052 (100)	46.9 (15.3)	48.7	6.1 (4.2)	50097 (2.6)
Ethnicity	1324575 (68.0)	48.2 (15.9)	43.9	6.2 (4.3)	47481 (3.6)
SES	1940101 (99.5)	46.9 (15.3)	48.7	6.2 (4.2)	49932 (2.6)
Health behaviours					
Smoking status	1643351 (84.3)	47.9 (15.1)	46.4	5.5 (3.9)	40443 (2.5)
Alcohol drinker status	1405027 (72.1)	47.6 (14.9)	46.2	5.6 (4.0)	33878 (2.4)
Physical activity status	795157 (40.8)	48.3 (15.1)	43.5	5.5 (4.0)	19207 (2.4)
Blood pressure					
Systolic blood pressure	1574301 (80.8)	48.8 (15.3)	44.0	5.7 (3.8)	44098 (2.8)
Diastolic blood pressure	1574301 (80.8)	48.8 (15.3)	44.0	5.7 (3.8)	44098 (2.8)
Hypertension	1949052 (100)	46.9 (15.3)	48.7	6.1 (4.2)	50097 (2.6)
Lipids					
Total cholesterol	731521 (37.5)	55.2 (13.7)	47.4	4.5 (3.1)	22234 (3.0)
Low density lipoprotein cholesterol	528681 (27.1)	56.2 (13.4)	47.2	3.7 (2.6)	13254 (2.5)
High density lipoprotein cholesterol	607883 (31.2)	55.9 (13.5)	47.4	4.0 (2.7)	16314 (2.7)
Triglycerides	618403 (31.7)	55.7 (13.4)	47.6	4.2 (2.9)	17308 (2.8)
Metabolic factors					
Diabetes mellitus	1949052 (100)	46.9 (15.3)	48.7	6.1 (4.2)	50097 (2.6)
Renal failure	1949052 (100)	46.9 (15.3)	48.7	6.1 (4.2)	50097 (2.6)
Anthropometry					
Height	1604843 (82.3)	46.8 (14.8)	46.0	6.4 (4.2)	43434 (2.7)
Weight	1385286 (71.1)	48.2 (15.1)	43.9	5.3 (3.8)	33269 (2.4)
Body mass index	1347412 (69.1)	48.1 (14.9)	43.9	5.4 (3.8)	32257 (2.4)
Inflammation					
C-reactive protein	284450 (14.6)	54.6 (15.3)	37.2	3.4 (2.6)	6851 (2.4)
Fibrinogen	16594 (0.9)	54.4 (15.9)	30.6	3.7 (2.8)	619 (3.7)
Thyroid disease					
Thyroid disease	1949052 (100)	46.9 (15.3)	48.7	6.1 (4.2)	50097 (2.6)
Autoimmune disease					
Rheumatoid arthritis	1949052 (100)	46.9 (15.3)	48.7	6.1 (4.2)	50097 (2.6)
Psoriasis	1949052 (100)	46.9 (15.3)	48.7	6.1 (4.2)	50097 (2.6)

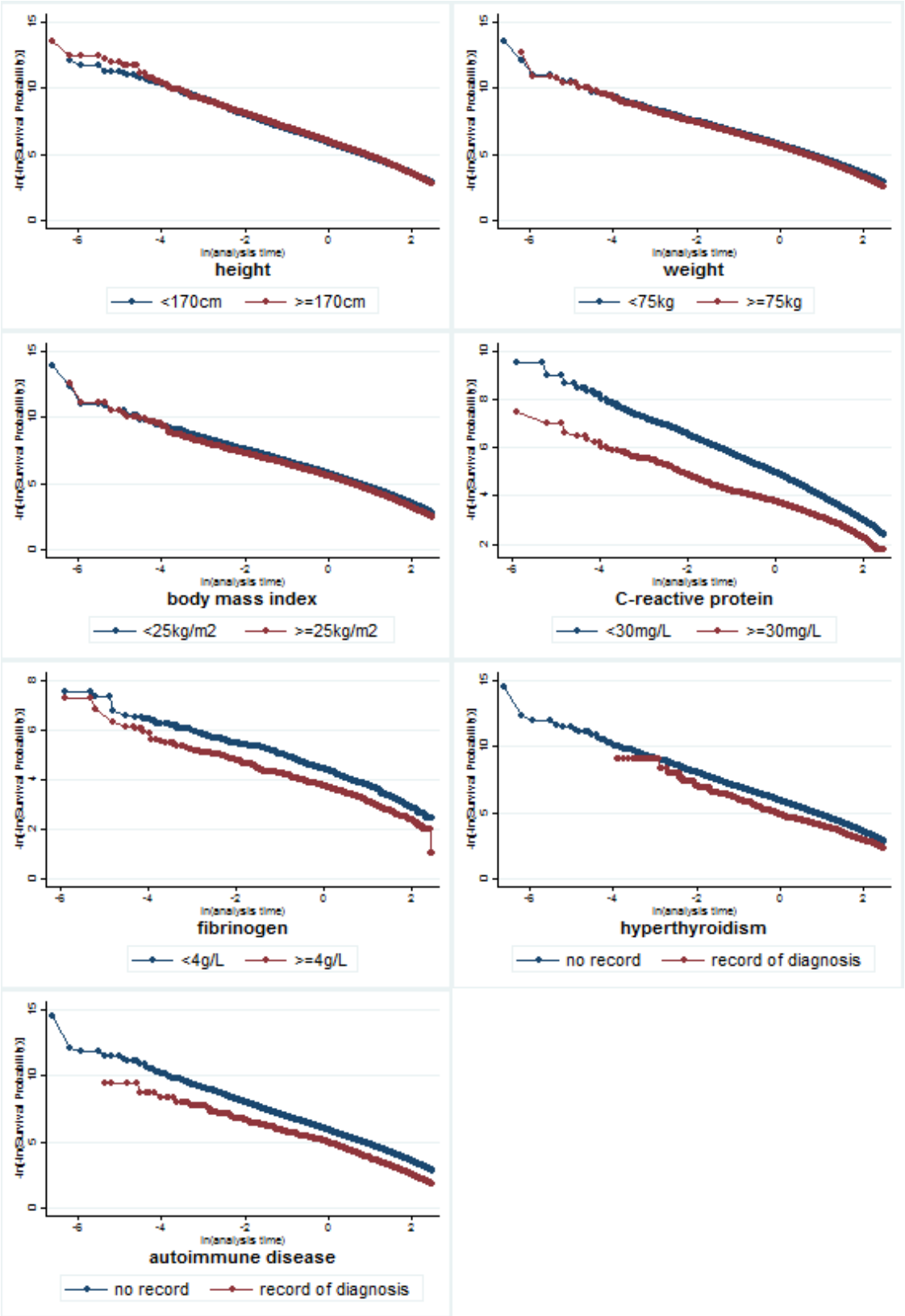
Table abbreviations: N – number, SD – standard deviation, SES – socioeconomic status.

Figure S4.1 Proportional hazard assumptions for 23 cardiovascular risk factors investigated for associations with incident atrial fibrillation

The very large sample size of the cohort (n=1,949,052) means that any test of proportionality is likely to be violated, but, as the below plots show, hazard lines are broadly parallel for all risk factors. Furthermore, as risk estimates derived are consistent with previous observational literature this confirms the Cox model is a reasonable analysis method.




ln[-ln(Survival Probability)]
ln(analysis time)
diastolic blood pressure
<90mmHG
>=90mmHG
hypertension
no record
record of diagnosis
total cholesterol
<5mmol/L
>=5mmol/L
low density lipoprotein cholesterol
<3mmol/L
>=3mmol/L
high density lipoprotein cholesterol
<1mmol/L
>=1mmol/L
triglycerides
<2mmol/L
>=2mmol/L
type II diabetes mellitus
no record
record of diagnosis
renal disease
no record
record of diagnosis


-ln(-ln(Survival Probability))
ln(analysis time)
height
<170cm
>=170cm
weight
<75kg
>=75kg
body mass index
<25kg/m2
>=25kg/m2
C-reactive protein
<30mg/L
>=30mg/L
fibrinogen
<4g/L
>=4g/L
hyperthyroidism
no record
record of diagnosis
autoimmune disease
no record
record of diagnosis

Chapter 5

Development of electronic health record definitions for AF subtypes relevant to the 2016 European Society of Cardiology guidelines

Table S5.1 Code lists for atrial fibrillation secondary to structural heart disease: 745 Read, 78 ICD-10 and 89 OPCS-4 codes

Atrial fibrillation secondary to structural heart disease										
LV systolic/diastolic dysfunction										
Read:	32BA.00	585f.00	585g.00	9h1..00	9h11.00	9h12.00	9On..00	9On0.00	9On1.00	9
9On2.00	9On3.00	9On4.00	G581.13	G5yy900	G5yyA00	G5yyD00				16
Heart failure										
Read:	2126400	14A6.00	14AM.00	1J60.00	1O1..00	388D.00	661M500	662f.00	662g.00	25
662h.00	662i.00	662p.00	662T.00	662W.00	679W100	679X.00	67D4.00	8B29.00	8CeC.00	35
8CL3.00	8CMK.00	8CMW800	8H2S.00	8HBE.00	8Hg8.00	8HgD.00	8HHb.00	8HHz.00	8Hk0.00	45
8HTL.00	8HTL000	8I98.00	8IB8.00	8IE0.00	8IE1.00	9hH..00	9hH0.00	9hH1.00	9m5..00	55
9N0k.00	9N2p.00	9N4s.00	9N4w.00	9N6T.00	9Or..00	9Or0.00	9Or1.00	9Or2.00	9Or3.00	65
9Or4.00	9Or5.00	G1yz100	G210.00	G210000	G210100	G211100	G21z100	G230.00	G232.00	75
G234.00	G400.00	G41z.11	G554000	G554011	G58..00	G58..11	G580.00	G580.11	G580.12	85
G580.13	G580.14	G580000	G580100	G580200	G580300	G580400	G581.00	G581.11	G581000	95
G582.00	G583.00	G583.11	G583.12	G58z.00	G58z.12	Q48y100	R2y1000	ZRad.00		104
ICD10:	I50									105
Hypertension										
Read:	6624.00	6627.00	6628.00	6629.00	2126100	6146200	14A2.00	1JD..00	212K.00	114
246M.00	661M600	661N600	662..12	662b.00	662c.00	662d.00	662F.00	662G.00	662H.00	124
662O.00	662P.00	662P000	662P100	662q.00	662r.00	7Q01.00	7Q01y00	8B26.00	8BL0.00	134
8CR4.00	8I3N.00	8IA5.00	9h3..00	9h31.00	9h32.00	9N03.00	9N1y200	9N4L.00	9OI..00	144
9OI..11	9OI1.00	9OI2.00	9OI3.00	9OI4.00	9OI5.00	9OI6.00	9OI7.00	9OI8.00	9OI9.00	154
9OIA.00	9OIA.11	9OIB.00	9OIZ.00	F404200	F421300	G2...00	G2...11	G20..00	G20..12	164
G200.00	G201.00	G202.00	G203.00	G20z.00	G20z.11	G21..00	G210.00	G210000	G210100	174
G210z00	G211.00	G211000	G211100	G211z00	G21z.00	G21z000	G21z011	G21z100	G21zz00	184
G22..00	G220.00	G221.00	G222.00	G22z.00	G22z.11	G23..00	G230.00	G231.00	G232.00	194
G233.00	G234.00	G23z.00	G24..00	G240.00	G240000	G240z00	G241.00	G241000	G241z00	204
G244.00	G24z.00	G24z000	G24z100	G24zz00	G25..00	G25..11	G26..00	G26..11	G27..00	214
G28..00	G2y..00	G2z..00	G41y100	G672.00	G672.11	Gyu2.00	Gyu2000	Gyu2100	L121200	224
L122.00	L122000	L122100	L122300	L122z00	L127.00	L127400	L127z00	L128.00	L128000	234
L128200	L12z400	TJC7.00	TJC7z00	U60C500	U60C511	U60C51A				241
ICD10:	I10	I11	I12	I13	I15					246
Left ventricular hypertrophy										
Read:	G5y3411	324..00	2492.00	3242.00	324Z.00					251
Congenital heart malformations										
Read:	66g..00	9RD0.00	L185.11	P5...12	P5...13	P50..12	P500.00	P500.11	P500.12	260
P502.00	P502.11	P50z.00	P51..00	P510.00	P511.00	P511100	P511200	P511300	P511z00	270
P512.00	P51y.00	P51y.11	P51z.00	P51z.11	P52..00	P520.00	P520.11	P520.12	P521.00	280
P52z.00	P53..00	P54..00	P540.00	P541.00	P542.00	P544.00	P545.00	P54y.00	P54z.00	290
P55..00	P550.00	P550.11	P550.12	P550.13	P551.00	P552.00	P552.11	P553.00	P55y.00	300
P55y.11	P55z.00	P56..00	P561.00	P561.11	P56y.00	P56z.00	P56z100	P56z200	P56zz00	310
P58..00	P59..00	P5y..00	P5z..00	P6...00	P60..00	P600.00	P601.00	P601000	P601z00	320
P602.00	P602100	P602z00	P603.00	P603.11	P60z.00	P60z000	P60z100	P60zz00	P61..00	330
P610.00	P611.00	P61z.00	P62..00	P63..00	P64..00	P640.00	P641.00	P64z.00	P65..00	340
P65..11	P650.00	P651.00	P652.00	P65z.00	P66..00	P67..00	P68..00	P69..00	P6X..00	350
P6y..00	P6y0.00	P6y1.00	P6y2.00	P6y3.00	P6y3000	P6y3100	P6y3z00	P6y4.00	P6y4000	360
P6y4100	P6y4300	P6y4500	P6y4600	P6y4z00	P6y5.00	P6y5000	P6y5100	P6y5200	P6y5z00	370

P6y6.00	P6y6.11	P6y6000	P6y6100	P6y6111	P6y6200	P6y6300	P6y6400	P6y6z00	P6y8.00	380
P6yy.00	P6yy.11	P6yy.12	P6yy000	P6yy100	P6yy200	P6yy300	P6yy400	P6yy411	P6yy500	390
P6yy600	P6yy700	P6yy900	P6yyA00	P6yyC00	P6yyz00	P6z..00	P6z..11	P6z0.00	P6z1.00	400
P6z1000	P6z1100	P6z2.00	P6z3.00	P6z3.11	P6zz.00	P70..00	P70..11	P71..00	P710.00	410
P711.13	P712.12	P713.00	P713.11	P71z.00	P72..00	P72..11	P720.00	P721.00	P721000	420
P721111	P721200	P721211	P721300	P721500	P721600	P721700	P721z00	P722.00	P722100	430
P722200	P722300	P722400	P722411	P722500	P722z00	P72z.00	P72z000	P72z100	P72z111	440
P73..00	P730.00	P731.00	P731.11	P732.00	P733.00	P734.00	P735.00	P737.00	P737.11	450
P738.00	P73y.00	P73z.00	P74..00	P740.00	P740000	P740100	P741.00	P741000	P741100	460
P741z00	P742.00	P742.11	P743.00	P744.00	P74z.00	P74z.11	P74z000	P74z100	P74z200	470
P74z300	P74z500	P74z600	P74z700	P74z800	P74zz00	P7X..00	PK33.00			478
ICD10:	Q20	Q21	Q22	Q23	Q24	Q25	Q26			485
Cardiomyopathies										
Read:	F391B00	G551.00	G554300	G554400	G554500	G554511	G555.00	G557.00	G558.00	494
G558000	G558100	G558400	G558z00	G559.00	G55A.00	G55A.11	G55y.11	G55y000	G55z.00	504
Gyu5M00	Gyu5R00	L186500								507
ICD10:	I420	I421	I422	I423	I425	I426	I427	I43	O903	516
O994										517
Valvular heart diseases										
See table S5.5										887
Other: left atrial appendage										
Read:	790G300	790G200								889
OPCS4:	K223									890
Other: heart aneurysm										
Read:	G341.00	G341.11	G341000	G341100	G341z00	790N200	790N300			897
ICD10:	I253									898
OPCS4:	K243	K244								900
Other: post-myocardial infarction ruptures/defects										
Read:	G30..13	G361.00	G362.00	G363.00	G364.00	G365.00	G366.00			907
ICD10:	I231	I232	I233	I234	I235					912

Note: final table column gives the cumulative total number of identified codes.

Table S5.2 Code lists for focal atrial fibrillation: 268 Read and 16 ICD-10 codes

Focal atrial fibrillation										
Paroxysmal atrial fibrillation										
Read:	G573200									1
Symptomatic atrial fibrillation: lethargy										
Read:	168..11	168..12	1682.00	1684.00	Eu46011	R007.00	R007100	R007300	R007200	10
R007500	R007z00	R007z11	168..00	1683.00	1683.11	168Z.00	E205.12			18
ICD10:	R53									19
Symptomatic atrial fibrillation: palpitations										
Read:	1812.00	R051.00	181..00	181Z.00	R051000	R051z00	181..11			26
ICD10:	R00.2									27
Symptomatic atrial fibrillation: dyspnoea										
Read:	1736.00	2322.00	173D.00	R060800	R060A00	173..12	173..13	1735.11	1739.00	36
173C.11	2323.00	388H.00	R060200	ZR3Q.00	R060D00	173..00	173..11	1738.00	173Z.00	46
173H.00	173I.00	173J.00	173K.00	173L.00	173N.00	173P.00	173Q.00	173R.00	173S.00	56
173T.00	173V.00	173W.00	173X.00	173Y.00	173a.00	1735.00	1734.00	1733.00	1732.00	66
173G.00	173C.12	173C.00	173F.00	173b.00	173g.00	173f.00	38Gb.00			74
ICD10:	R06.0									75
Symptomatic atrial fibrillation: chest tightness										
Read:	182..00	1822.00	1823.00	1824.00	1826.00	1828.00	1829.00	182B000	182C.00	84

182Z.00	8HTG.00	8HTJ.00	9N0f.00	R065.00	R065000	R065011	R065100	R065200	R065600	94
R065700	R065800	R065900	R065C00	R065D00	R065z00	Ryu0400	182A.00			102
ICD10:	R072	R073	R074							105
Symptomatic atrial fibrillation: sleeping difficulties										
Read:	R005200	1B1B.11	E274111	1B1B.00	R005.11	E274100	1B1B200	1B1B000	Eu51000	114
E274200	1B1B100	E274.12	38D1.00	Fy0..00	Fy00.00	E274.00	E274D11	Eu51.00	Eu51z11	124
E274z00	E274000	Fyu5800	E274y00	Eu51z00	Eu51y00	R005000	R005.00	R005z00	Z1M1.00	134
1B1Q.00	1BX0.00									136
ICD10:	F510	F518	F519							139
Symptomatic atrial fibrillation: psychosocial distress										
Read:	E112100	E112200	Eu32000	Eu32100	Eu32400	Eu32500	Eu32600	E112300	Eu32700	148
E11..12	E112400	E130.00	E130.11	Eu32311	Eu32312	Eu32313	Eu32314	Eu32800	Eu33311	158
E112.00	E112000	E112500	E112z00	E11y200	E11z200	E204.00	E2B..00	Eu32.00	Eu32.11	168
Eu32.12	Eu32.13	Eu32212	Eu32213	Eu32y00	Eu32y11	Eu32y12	Eu32z00	Eu32z11	Eu32z12	178
Eu32z13	Eu32z14	Eu33z11	Eu34111	Eu34113	E112.12	E112.13	E112.14	Eu33211	E112.11	188
E135.00	Eu32211	E211200	Eu34100	Eu34112	E2B1.00	E113100	E113200	Eu33000	Eu33100	198
E113300	Eu33200	E113400	Eu33300	Eu33313	Eu33314	Eu33315	Eu33316	E113.00	E113.11	208
E113000	E113500	E113600	E113700	E113z00	E118.00	Eu33.00	Eu33.11	Eu33.12	Eu33.13	218
Eu33.14	Eu33.15	Eu33212	Eu33214	Eu33400	Eu33y00	Eu33z00	Eu3y111	62T1.00	E204.11	228
E2B0.00	Eu53011	Eu53012	R007z13	E200200	E200400	E200500	Eu41100	Eu41111	E200300	238
Eu34114	Eu41200	Eu41211	E200100	E200111	E280.00	Eu41000	Eu41011	Eu41012	1Bb1.00	248
E202B00	E262000	Eu45311	Eu45313	285..00	286..00	E200.00	E200000	E200z00	Eu4..00	258
Eu41.00	Eu41112	Eu41113	Eu41300	Eu41y00	Eu41y11	Eu41z00	Eu41z11	Z4I7200		267
ICD10:	F32	F33	F341	F381	F41					272
Atrial Ectopy										
Read:	3264.00	G576300	G576100							275
ICD10:	I49.1									276
Atrial tachycardia										
Read:	G570000	G570.00	G570z00	G57y900	G570100	G570200	G570300			283
ICD10:	I471									284

Note: final table column gives the cumulative total number of identified codes.

Table S5.3 Code lists for post-operative atrial fibrillation: 1402 (top level) OPCS-4 codes

Post-operative atrial fibrillation		
Cardiac procedures		
OPCS4:	Chapter K	71
Non-cardiac procedures		
OPCS4:	Chapter A Nervous System	139
OPCS4:	Chapter B Endocrine System and Breast	169
OPCS4:	Chapter C Eye	243
OPCS4:	Chapter D Ear	265
OPCS4:	Chapter E Respiratory Tract	331
OPCS4:	Chapter F Mouth	376
OPCS4:	Chapter G Upper Digestive System	448
OPCS4:	Chapter H Lower Digestive System	507
OPCS4:	Chapter J Other Abdominal Organs, Principally Digestive	580
OPCS4:	Chapter L Arteries and Veins	666
OPCS4:	Chapter M Urinary	738
OPCS4:	Chapter N Male Genital Organs	763
OPCS4:	Chapter O Overflow codes	793
OPCS4:	Chapter P Lower Female Genital Tract	817
OPCS4:	Chapter Q Upper Female Genital Tract	866
OPCS4:	Chapter R Female Genital Tract Associated with Pregnancy, Childbirth and the Puerperium	898

OPCS4:	Chapter S Skin	958
OPCS4:	Chapter T Soft Tissue	1036
OPCS4:	Chapter U Diagnostic Imaging, Testing and Rehabilitation	1077
OPCS4:	Chapter V Bones and Joints of Skull and Spine	1138
OPCS4:	Chapter W Other Bones and Joints	1237
OPCS4:	Chapter X Miscellaneous Operations	1319
OPCS4:	Chapter Y Subsidiary Classification of Methods of Operation	1402

Note: final table column gives the cumulative total number of identified codes.

Table S5.4 Code lists for valvular atrial fibrillation: 235 Read, 49 ICD10 and 86 OPCS-4 codes

Valvular atrial fibrillation										
Valvular heart disease										
Read:	G540.16	G541100	G541212	G541200	G11..00	G540.15	G542000	G110.00	G541300	9
G543300	G542.00	G540.00	G541.00	G540000	G54z500	G543215	G121.12	G130.00	G140413	19
G120.00	G540.14	G541500	G13..00	G54z100	G541012	G541400	G133.12	G543.00	G543100	29
G541000	G543012	G543011	G140.00	G11..11	G131.14	G12..00	G544200	G541z00	G544100	39
G543z00	G111.11	G140111	G54z013	G111.12	G543000	G540z00	G113.00	G54..11	G13z.00	49
G11z.00	G541600	G140000	G133.00	G132.12	G540200	G121.00	G110.11	G132.00	G132.13	59
G540100	G140412	G543400	G542012	G542100	G141z00	G543213	G54z000	G540300	G544.00	69
G544000	G540.12	G140400	G140112	G121.11	G542z00	G141.00	G112.13	G112.00	G543200	79
G541011	Gyu5600	G131.00	G140514	G12z.00	G112.12	G111.00	G542200	Gyu1000	Gyu5800	89
G141100	G140300	G544X00	G114.00	G541211	G140100	G131.13	G543311	G14021Y	G141000	99
G122.00	A932.11	G13y.00	Gyu5A00	G140500	G140z00	G54z014	G14021X	G140200	Gyu1100	109
G133.11	G542011	G542X00	Gyu5500	Gyu1200	Gyu5D00	Gyu5B00				116
ICD10:	I05	I050	I051	I052	I058	I059	I06	I060	I061	125
I062	I068	I069	I07	I070	I071	I072	I078	I079	I08	135
I080	I081	I082	I083	I088	I089	I34	I340	I341	I342	145
I348	I349	I35	I350	I351	I352	I358	I359	I36	I360	155
I361	I362	I368	I369	I37	I370	I371	I372	I378	I379	165
Prosthetic valve replacements										
Read:	7914200	14T3.00	7910200	ZV43300	7910211	7911200	7914212	ZV45H00	7910212	174
7913200	7912200									176
OPCS4	K253	K263	K273	K283	K293					181
Bioprosthetic valve replacements										
Read:	7911100	7911000	7913000	7914211	7910213	7910100	7914100	7914000	7910214	190
7913100	7910000	7912000	7912100							194
OPCS4	K251	K252	K261	K262	K271	K272	K281	K282	K291	203
K292										204
Unspecified valve replacements										
Read:	7911.12	7910300	7910.12	7914300	14S4.00	7911300	ZV42200	7914.11	7912300	213
7913.12	7912.11	7913300	7911500							217
OPCS4	K254	K264	K274	K284	K294	K357				223
Valve repairs										
Read:	791z.00	7918.00	7N40000	7910.00	7N40.00	7N40100	7918000	7911.00	7N40300	232
7N40200	7914z00	7916z11	7916.11	7916100	7915000	7916000	7917000	7917z00	7910400	242
7918100	7912.00	7917.11	7916300	7913.11	791..00	7918200	7911.11	7910.11	7919300	252
7917300	7919100	7915300	7919.11	7911y00	7910z00	7915.00	7917100	7918y00	7915100	262
7913.00	791y.00	7918z00	7916.00	7914400	7913400	7917311	7911400	7918111	7919000	272
7911z00	7919400	7910y00	7914.00	7915200	7918300	7916z00	7916y00	7915y00	7917.00	282
7916200	7912z00	7912y00	7913y00	7910411	791D100	7911411	791D000	7912.12	791D.00	292
7913z00	7918500	7913411	7914y00	7912511	7918400	7912500	7917200	7917y00	7915z00	302
7914411	7919200	791Dy00								305

OPCS4	K25	K255	K258	K259	K26	K265	K268	K269	K27	314
K275	K276	K278	K279	K28	K285	K288	K289	K29	K295	324
K298	K299	K30	K301	K302	K303	K304	K308	K309	K31	334
K311	K312	K313	K314	K318	K319	K32	K321	K322	K323	344
K324	K328	K329	K34	K341	K342	K343	K344	K345	K346	354
K348	K349	K35	K351	K352	K353	K354	K355	K356	K358	364
K359	K36	K361	K362	K368	K369					370

Note: final table column gives the cumulative total number of identified codes.

Table S5.5 Code lists for atrial fibrillation in athletes: 27 Read codes

Atrial fibrillation in athletes										
Competitive athletes and sports professional (e.g. professional runner, footballer, boxer)										
Read:	1386.00	04A3.00	68L1.11	04A2.00	04AB.00	04A6.11	04AZ.00	04A..00	04A6.00	9
04A..13	04A4.00	04A5.00	04AA.00	68L1.00						14
Sports-related occupations (e.g. coaches, managers, officials)										
Read:	04A..11	04A8.00	04A7.00	04A1.00	06C..13	04A..12	06C..00	04A9.00	04A..14	23
06C4.00	06C5.00	04Z..00	04...00							27

Note: final table column gives the cumulative total number of identified codes.

Table S5.6 Code lists for atrial fibrillation secondary to inherited rhythm disorders: 6 Read and 2 ICD-10 codes

Atrial fibrillation secondary to inherited rhythm disorders										
Long QT syndrome, Brugada syndrome, Short QT syndrome and Wolff-Parkingson-White syndrome										
Read:	G56y500	G57y200	G567400	32K3.00	32K2.00	32K4.00				6
ICD10:	I49.8	I45.6								8

Note: final table column gives the cumulative total number of identified codes.

Table S5.7 Code lists for atrial fibrillation secondary to respiratory disease: 167 Read and 13 ICD-10 codes

Atrial fibrillation secondary to respiratory disease										
Chronic obstructive pulmonary disorder										
Read:	14B3.11	14OJ.00	1J71.00	66Yg.00	679V.00	8CE6.00	H060000	H060200	H060v00	9
H06z200	H30..12	H300.00	H301.00	H302.00	H30z.00	H310.00	H310000	H310z00	H311100	19
H311z00	H312.00	H312000	H312011	H313.00	H31y.00	H31y100	H31yz00	H32y000	H32y100	29
H32y111	H32y200	H3y..00	H581.00	H582.00	66YB.00	66YD.00	66YL.00	66YM.00	66YS.00	39
66YT.00	9Oi..00	9Oi0.00	9Oi1.00	9Oi2.00	H3...00	H3...11	H31..00	H312100	H312z00	49
H31z.00	H32..00	H320.00	H320000	H320100	H320200	H320z00	H321.00	H322.00	H32y.00	59
H32yz00	H32z.00	H36..00	H37..00	H38..00	H39..00	H3y..11	H3z..00	H3z..11	Hyu3000	69
Hyu3100	66Yf.00	H06..00	H060.00	H060.11	H060300	H060400	H060500	H060600	H060700	79
H060800	H060900	H060A00	H060B00	H060C00	H060D00	H060E00	H060F00	H060w00	H060x00	89
H060z00	H06z.00	H06z000	H06z011	H20..11	H21..11	H22..11	H23..11	H24..11	H25..11	99
H26..11	H270.11	H30..00	H30..11	H311.00	H311000	Hyu1000	H312200	H3y1.00	8CR1.00	109
H26..11	H270.11	H30..00	H30..11	H311.00	H311000	Hyu1000	H312200	H3y1.00	8CR1.00	119
9Oi3.00	9Oi4.00	14OX.00	H583200	H060100	AB63600	H320300	H3A..00	SK07.11	38Dg.00	129
8IEZ.00	9Nk7000	8CMW500	9NgP.00	8CMV.00	8BMW.00	9NgP.11	661N300	9e03.00	661M300	139
9kf0.11	9kf2.11	8Hkw.00	8CeD.00	9kf1.11	9kf0.00	66YB200	66YB100	66YB000	38Dd.00	149
14B3.12										150

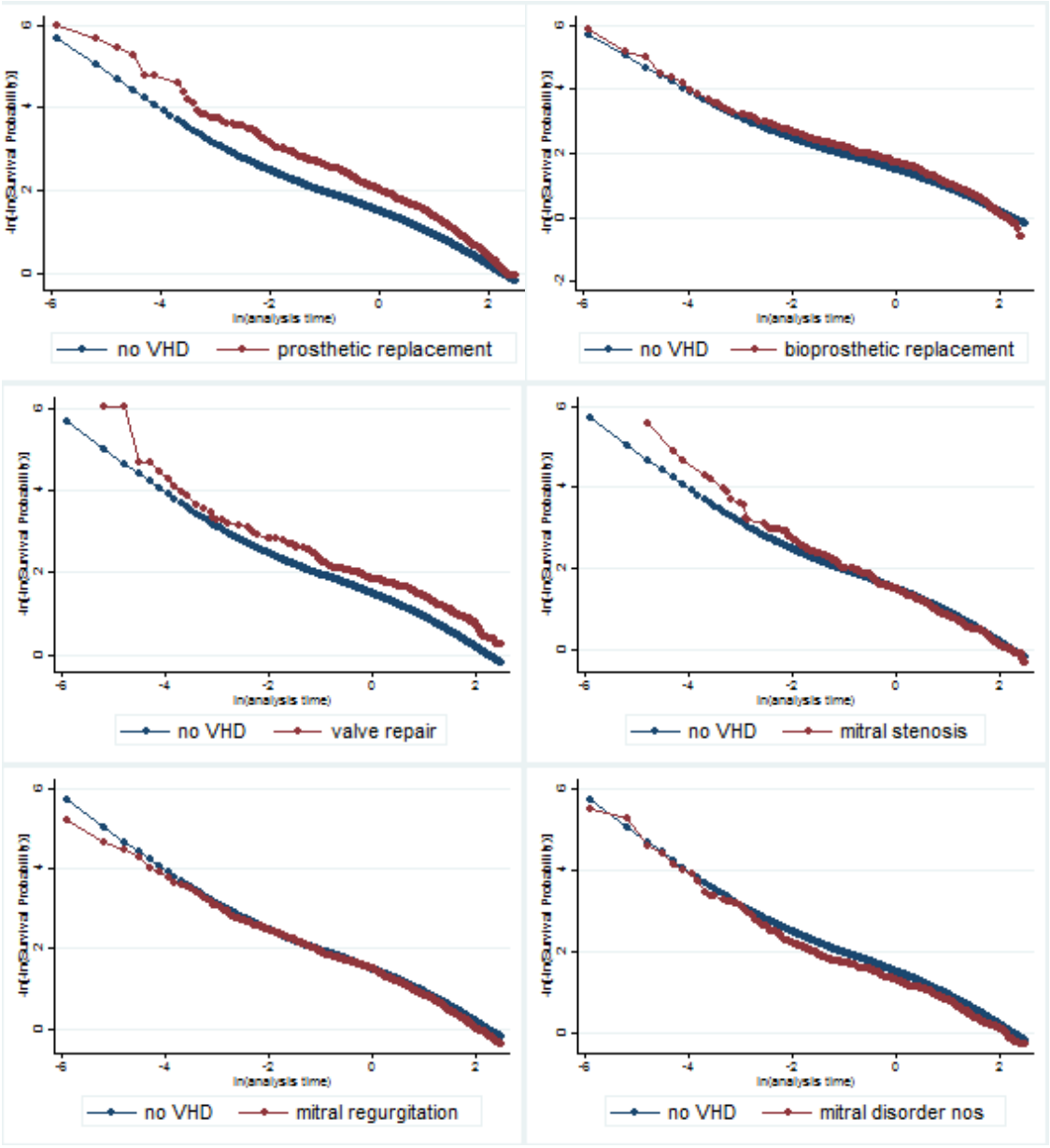
ICD10:	J40	J41	J42	J43	J448	J449	J20	J440	J441	159
Sleep apnoea										
Read:	H5B0.00	R005100	R005312	R005311	R005300	Fy03.11	38Da.00	H5B..00	Fy03.00	168
ICD10:	G47.3									169
Pulmonary hypertension										
Read:	G410.00	7Q01000	7Q01100	7Q01200	7Q01300	G41y100	G41y000	G411.00	G410.00	177
ICD10:	I27.0	I27.1	I27.2						I27.0	180

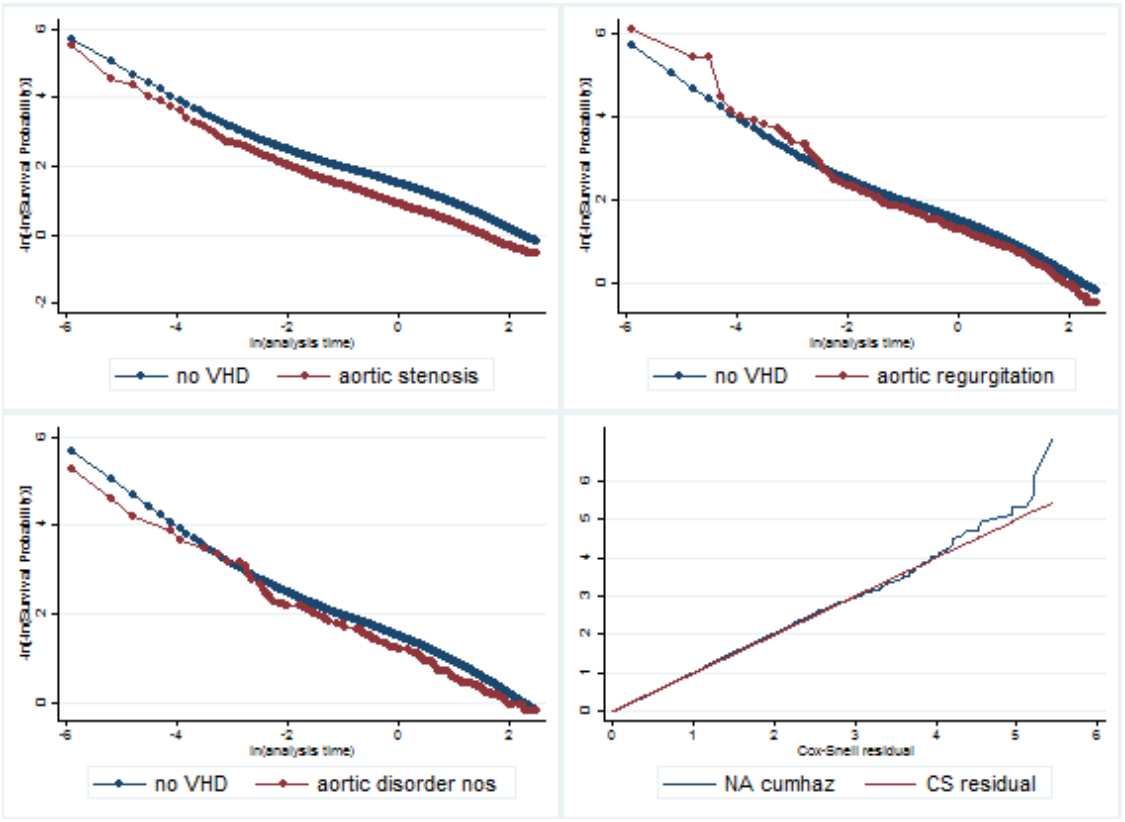
Note: final table column gives the cumulative total number of identified codes.

Chapter 6

What is ‘valvular’ atrial fibrillation’? A reappraisal exploiting electronic health records

Figure S6.1 Proportional hazard assumptions (first nine plots) and goodness of model fit assessment (final plot) for valvular heart diseases investigated for associations with incident stroke, systemic embolism and mortality





Notes: Due to the large sample size any test of proportionality is likely to be violated, but, as the above plots show, hazard lines are broadly parallel for all valvular heart disease. The final model adjusted for adjusted for age and sex, warfarin prescriptions, congestive heart failure, hypertension, diabetes mellitus, stroke, transient ischaemic attack or systemic embolism, and vascular disease is shown to fit the data well.

Abbreviations: VHD – valvular heart disease.

Table S6.1 Interaction testing between baseline valvular heart diseases and key confounders of interest: age, sex, warfarin and prior stroke

	Age		Sex		Warfarin		Prior stroke	
	β	p-value	β	p-value	β	p-value	β	p-value
Prosthetic valve replacement	-0.026	0.000	0.068	0.482	0.384	0.000	-0.112	0.320
Bioprosthetic valve replacement	0.013	0.188	0.068	0.593	0.416	0.001	0.189	0.195
Valve repair	0.003	0.774	0.207	0.261	0.403	0.029	0.371	0.080
Mitral stenosis	-0.005	0.454	-0.101	0.477	0.109	0.403	-0.022	0.872
Mitral regurgitation	-0.003	0.455	-0.093	0.143	0.064	0.341	-0.068	0.361
Other mitral disorder	-0.005	0.406	0.126	0.207	0.080	0.457	0.071	0.521
Aortic stenosis	0.002	0.605	0.096	0.185	0.013	0.890	-0.216	0.010
Aortic regurgitation	-0.014	0.059	-0.063	0.647	0.012	0.942	-0.194	0.234
Other aortic disorder	0.013	0.330	0.276	0.189	-0.221	0.385	0.065	0.779

Figure S6.1 Quality of recording valvular heart diseases in electronic health records in 1998 to 2010: specified heart valves

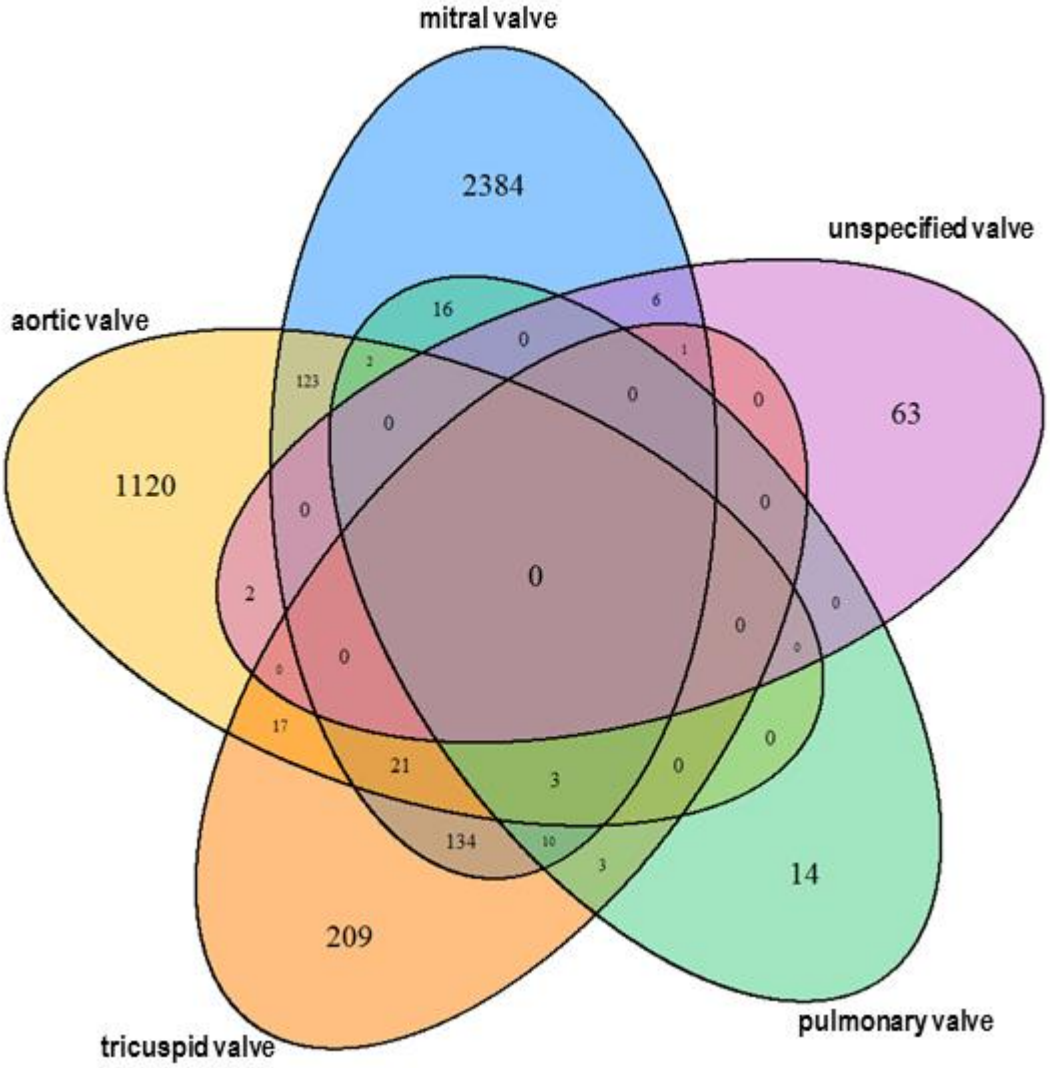


Figure S6.2 Quality of recording valvular heart diseases in electronic health records in 1998 to 2010: prosthetic vs. bioprosthetic valve replacements

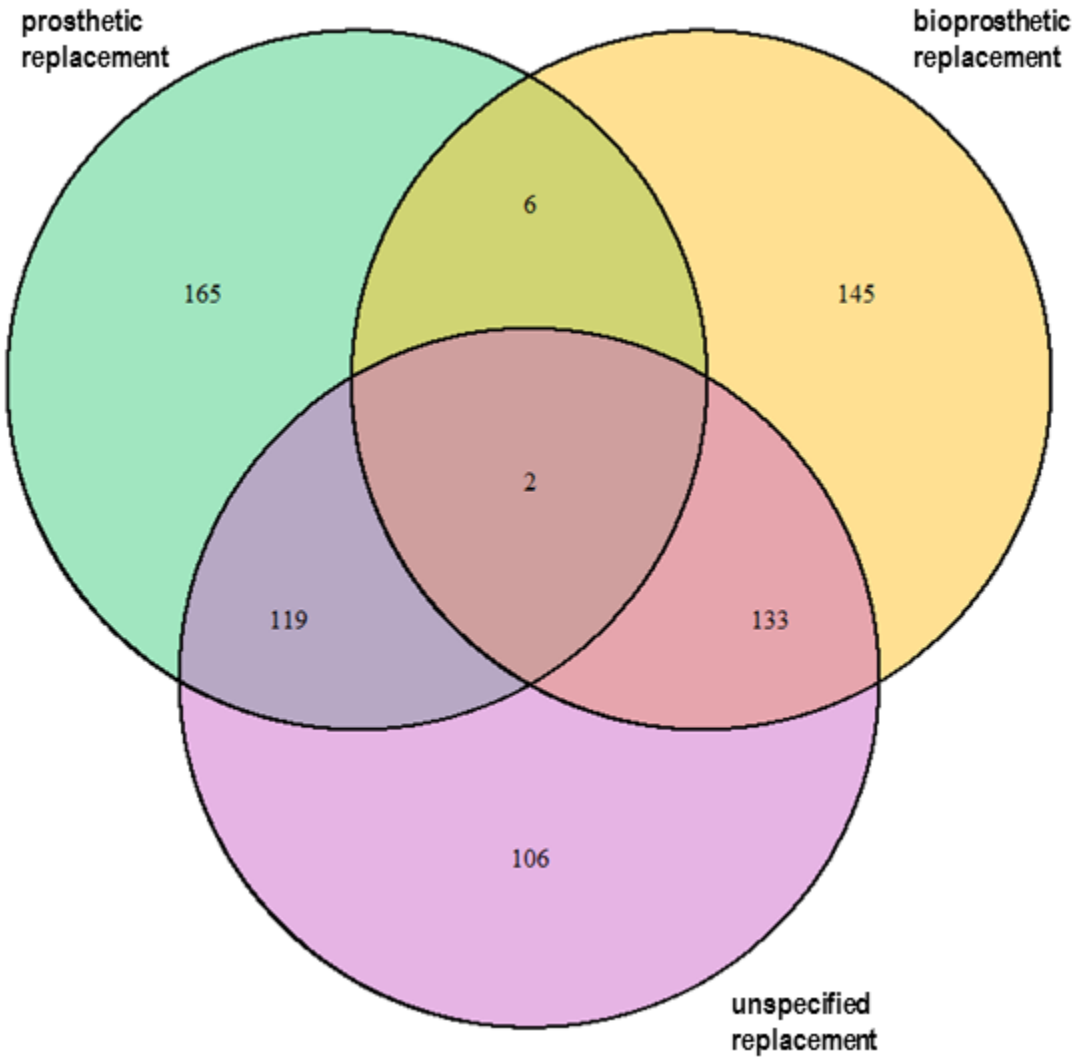


Figure S6.3 Quality of recording valvular heart diseases in electronic health records in 1998 to 2010: rheumatic vs. non-rheumatic valve disorders

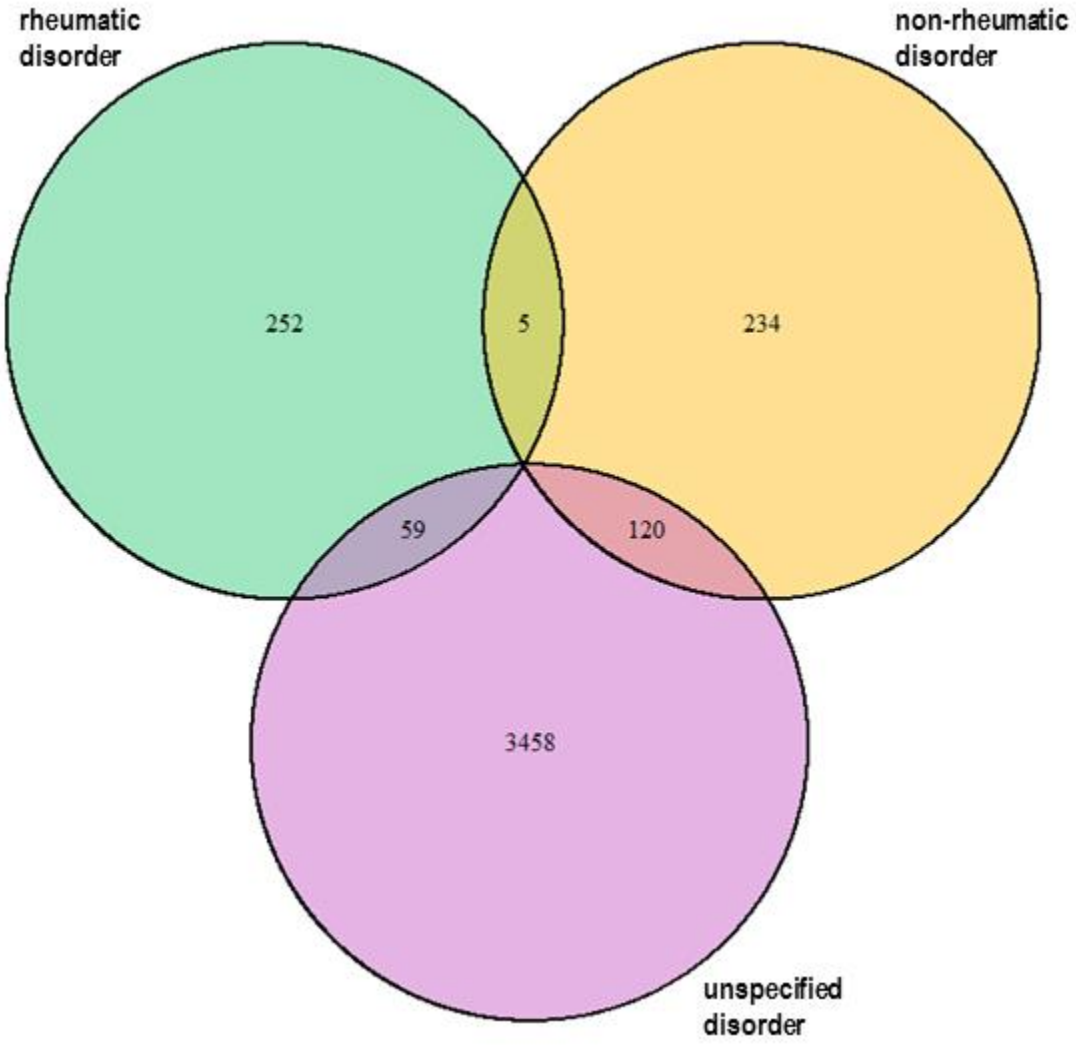
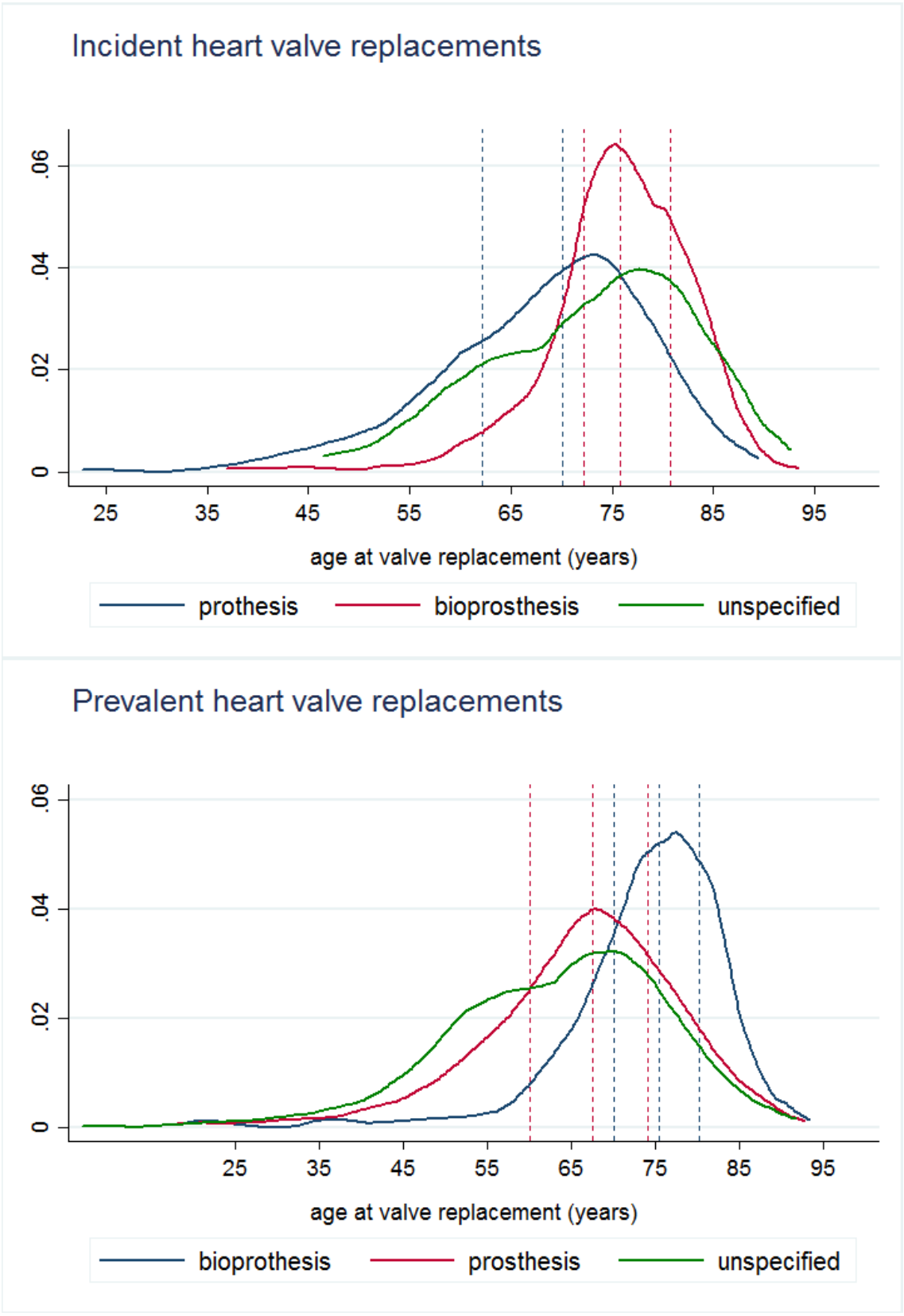


Figure S6.4 Differences in age distributions in individuals with prosthetic, bioprosthetic and unspecified valve codes



Notes: vertical dashed lines represent median and interquartile interval values showing overlap in the age distributions for individuals with prosthetic and bioprosthetic valve replacements between the ages 70 to 75 years.

Chapter 7

Net clinical benefit of warfarin in individuals with atrial fibrillation across stroke risk and across primary and secondary care

Table S7.1 Code lists for systemic embolism: 31 Read codes and 1 ICD-10 code

Systemic embolism										
Read:	K138000	K138011	G74..00	G74..11	G74..13	G740.00	G740.14	G741.00	G742.00	9
G742000	G742100	G742200	G742300	G742400	G742500	G742600	G742700	G742900	G742z00	19
G74y.00	G74y000	G74y100	G74y200	G74y300	G74y500	G74y600	G74y700	G74y800	G74y900	29
G74yz00	G74z.00									31
ICD10:	I74									32

Note: final table column gives the cumulative total number of identified codes.

Table S7.2 Comparison of baseline CHA_2DS_2-VASc risk factors in individuals with atrial fibrillation with and without use of warfarin.

	With warfarin		Without warfarin		Overall	
Number of individuals	**30 067**		**40 139**		**70 206**	
	N	**%**	**N**	**%**	**N**	**%**
Congestive heart failure	10400	25.9	7032	23.4	17432	24.8
Hypertension	32641	81.3	25122	83.6	57763	82.3
Diagnosis	23973	59.7	17916	59.6	41889	59.7
Blood pressure medication	28657	71.4	22486	74.8	51143	72.9
Blood pressure measures	22060	55.0	17550	58.4	39610	56.4
Age ≥ 75 [**2**]	27235	67.9	14955	49.7	42190	60.1
Diabetes	5746	14.3	4243	14.1	9989	14.2
Stroke/TIA/systemic embolism [**2**]	7506	18.7	5319	17.7	12825	18.3
Vascular disease	7935	19.8	5897	19.6	13832	19.7
Myocardial infarction	5439	13.6	4146	13.8	9585	13.7
Peripheral vascular disease	3341	8.3	2362	7.9	5703	8.1
Age 65–74	6975	17.4	9321	31.0	16296	23.2
Sex **C**ategory [female]	21618	63.0	12668	37.0	34286	48.8
CHA_2DS_2-VASc scores						
0	1414	3.5	1072	3.6	2486	3.5
1	2786	6.9	2851	9.5	5637	8.0
2	4485	11.2	4854	16.1	9339	13.3
3	7309	18.2	6461	21.5	13770	19.6
4	10102	25.2	6808	22.6	16910	24.1
5	6994	17.4	4232	14.1	11226	16.0
6	4383	10.9	2389	8.0	6772	9.7
7	1995	5.0	1070	3.6	3065	4.4
8	566	1.4	297	1.0	863	1.2
9	105	0.3	33	0.1	138	0.2

Table S7.3 Comparison of baseline CHA_2DS_2-VASc risk factors in men and women with atrial fibrillation

	Men		Women		Overall	
Number of individuals	**35 920**		**34 286**		**70 206**	
	N	**%**	**N**	**%**	**N**	**%**
Congestive heart failure	8458	23.6	8974	26.2	17432	24.8
Hypertension	28350	78.9	29413	85.8	57763	82.3
Diagnosis	19702	54.9	22187	64.7	41889	59.7
Blood pressure medication	24784	69	26359	76.9	51143	72.9
Blood pressure measures	18794	52.3	20816	60.7	39610	56.4
Age ≥ 75 [**2**]	17828	49.6	24362	71.1	42190	60.1
Diabetes	5545	15.4	4444	13.0	9989	14.2
Stroke/TIA/systemic embolism [**2**]	6136	17.1	6689	19.5	12825	18.3
Vascular disease	8435	23.5	5397	15.7	13832	19.7
Myocardial infarction	6120	17.0	3465	10.1	9585	13.7
Peripheral vascular disease	3292	9.2	2411	7.0	5703	8.1
Age 65–74	9861	27.5	6435	18.8	16296	23.2
Sex **C**ategory [female]	0	0.0	34286	100	34286	48.8
CHA_2DS_2-VASc scores						
0	2486	6.9	0	0.0	2486	3.5
1	4690	13.1	947	2.8	5637	8.0
2	6813	19.0	2526	7.4	9339	13.3
3	8562	23.8	5208	15.2	13770	19.6
4	6279	17.5	10631	31.0	16910	24.1
5	4168	11.6	7058	20.6	11226	16.0
6	1991	5.5	4781	13.9	6772	9.7
7	775	2.2	2290	6.7	3065	4.4
8	156	0.4	707	2.1	863	1.2
9	0	0.0	138	0.4	138	0.2

Table S7.4 Completeness of recording CHA_2DS_2-VASc risk factors in primary care records, and secondary care records

	Total individuals with risk factor	Individuals with risk factor recorded in primary care		Individuals with risk factor recorded in secondary care	
	N	N	%	N	%
Congestive heart failure	17432	12043	69.1	11145	63.9
Hypertension	57763	57197	99.0	23310	40.4
Diabetes mellitus	9989	9339	93.5	7366	73.7
Stroke/ TIA / systemic embolism	12825	10671	83.2	6171	48.1
Vascular disease	13832	11956	86.4	6419	46.4

Notes: total individuals with risk factor refers to total number of individuals with recorded diagnosis of the risk factor, as recorded in either primary or secondary care records. Individuals with risk factor recorded in primary care refers to the 'completeness' of recording the risk factor in primary records, where 100% indicates absolute completeness. For example 99% of total individuals with hypertension, had a record of hypertension in primary care records. Individuals with risk factor recorded in secondary care refers to the 'completeness' of recording the risk factor in secondary records.

Figure S7.1 Venn diagrams comparing numbers of CHA_2DS_2-VASc risk factors captured in primary care, secondary care and in both sources linked. Venn circles are scaled according to the proportion of individuals that they represent

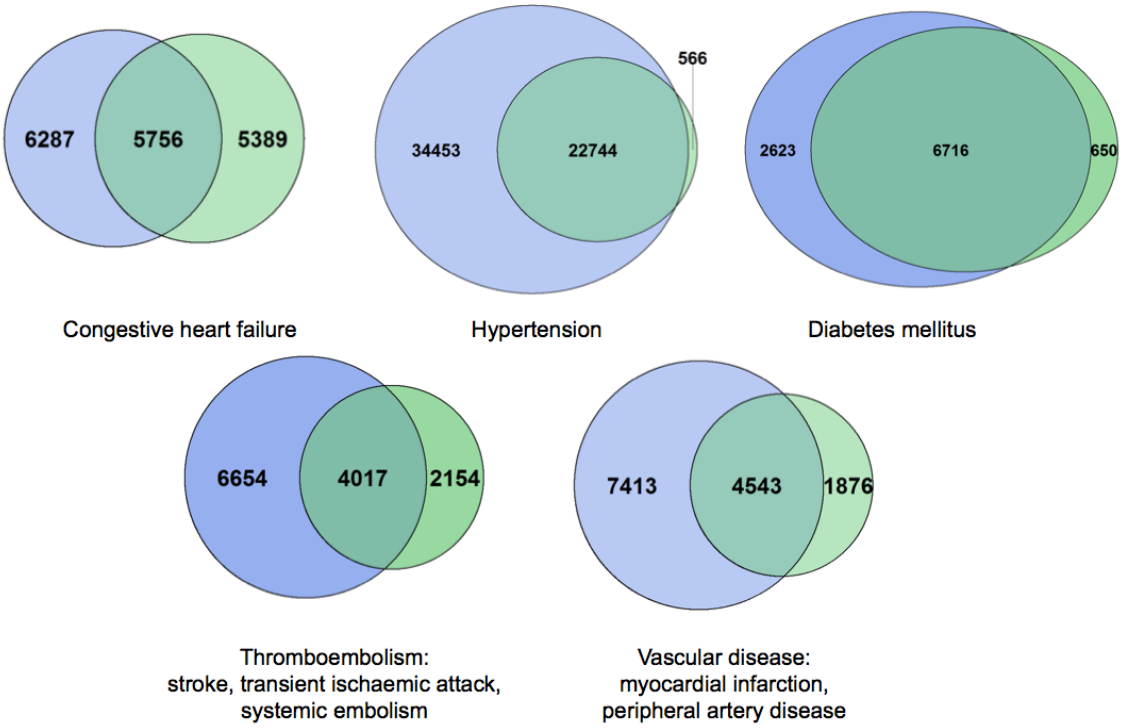


Table S7.5 CHA_2DS_2-VASc scores based on primary and secondary care records, compared to both data sources linked

	Secondary care records only		Primary–secondary care records	
Individuals initially in secondary care	**40 638**		**40 638**	
CHA_2DS_2-VASc scores	**N**	**%**	**N**	**%**
0	2156	5.3	1181	2.9
1	4042	10.0	2665	6.6
2	7196	17.7	4519	11.1
3	10309	25.4	7107	17.5
4	8,768	21.6	9578	23.6
5	4,794	11.8	7514	18.5
6	2,361	5.8	4906	12.1
7	776	1.9	2341	5.8
8	201	0.5	707	1.7
9	35	0.1	120	0.3
	Primary care records only		**Primary–secondary care records**	
Individuals initially in primary care	**29 568**		**29 568**	
CHA_2DS_2-VASc scores	**N**	**%**	**N**	**%**
0	1320	4.5	1305	4.4
1	3041	10.3	2972	10.1
2	4889	16.5	4820	16.3
3	6838	23.1	6663	22.5
4	7450	25.2	7332	24.8
5	3537	12.0	3712	12.6
6	1750	5.9	1866	6.3
7	611	2.1	724	2.5
8	121	0.4	156	0.5
9	11	0.0	18	0.1

Table S7.6 Propensity score adjusted incidence rates [95% confidence intervals] per 100 person-years of ischaemic stroke by CHA_2DS_2-VASC scores, sex, and use of warfarin

CHA_2DS_2-VASc scores	With warfarin		Without warfarin	
Overall population	**Predicted events**	**Adjusted rate**	**Predicted events**	**Adjusted rate**
0	7	0.4 [0.1,0.7]	23	0.2 [0.1,0.3]
1	27	0.4 [0.3,0.6]	127	0.7 [0.6,0.8]
2	89	0.8 [0.6,1.0]	355	1.4 [1.2,1.5]
3	152	1.1 [0.9,1.3]	816	2.5 [2.3,2.6]
4	236	1.8 [1.6,2.0]	1366	3.8 [3.6,4.0]
5	236	3.2 [2.8,3.6]	1129	5.9 [5.6,6.2]
6	170	4.4 [3.7,5.1]	1154	11.6 [11.0,12.2]
7	116	7.4 [6.1,8.8]	553	13.8 [12.7,14.8]
8	19	5.0 [2.7,7.3]	167	17.6 [15.2,20.0]
9	3	8.7 [0.4,16.9]	30	24.9 [17.0,32.7]
Men				
0	7	0.4 [0.1,0.7]	23	0.2 [0.1,0.3]
1	25	0.4 [0.3,0.6]	110	0.7 [0.6,0.9]
2	78	0.9 [0.7,1.1]	306	1.7 [1.5,1.9]
3	114	1.4 [1.1,1.6]	542	2.9 [2.6,3.1]
4	171	3.2 [2.3,4.1]	552	4.8 [4.3,5.3]
5	153	5.1 [3.8,6.5]	580	9.2 [8.2,10.1]
6	74	5.9 [1.8,10.0]	425	15.7 [13.5,17.9]
7	39	9.1 [1.4,16.7]	177	19.9 [15.5,24.3]
8	1	1.1 [0.0,3.0]	28	20.6 [9.7,31.4]
Women				
1	1	0.1 [0.0,0.4]	13	0.4 [0.1,0.7]
2	7	0.3 [0.1,0.4]	42	0.5 [0.3,0.6]
3	38	0.7 [0.5,0.9]	271	1.9 [1.7,2.2]
4	102	1.3 [1.1,1.6]	815	3.3 [3.1,3.5]
5	109	2.5 [2.0,2.9]	610	4.8 [4.4,5.2]
6	107	4.1 [3.3,4.9]	791	10.9 [10.2,11.7]
7	82	7.3 [5.7,8.8]	397	12.7 [11.5,13.9]
8	17	5.3 [2.8,7.7]	140	17.3 [14.8,19.9]
9	3	8.7 [0.4,16.9]	30	24.9 [17.0,32.7]

Notes: Incidence rates were adjusted for propensity score quintiles. Propensity score was generated using a logistic regression model, which predicted probability of warfarin use (yes, no), based on CHA2DS2-VASC risk factors, age at initial record of diagnosis of atrial fibrillation, and source of initial record of atrial fibrillation (primary, or secondary care).

Table S7.7 Supplementary evidence table comparing studies that reported incidence rates for moderate risk atrial fibrillation patients, sorted by estimate size

Author year	Country	Data source		Total individuals	Maximum follow-up	Endpoints					Total events	Risk group		Incidence Rate [95% CI] per 100 PY	
		PC	SC			IS	US	SE	PE	TIA		Score	Sex	With anticoagulants	Without anticoagulants
Huang 2014	China	○	●	358	4.5	●	○	○	○	○	70	1	men	NR [NR]	6.60 [NR]
Chao 2015	Taiwan	○	●	12935	6.0	●	○	○	○	○	1858	1	men	NR [NR]	2.75 [NR]
Chao 2015	Taiwan	○	●	7900	6.0	●	○	○	○	○	1174	2	women	NR [NR]	2.55 [NR]
Olesen 2011	Denmark	○	●	8203	1.0	●	○	●	●	○	NR	1	both	NR [NR]	2.01 [1.70,2.36]
Olesen 2012	Denmark	○	●	10062	12.0	●	○	●	○	●	159	1	both	NR [NR]	1.79 [1.53,2.09]
Olesen 2011	Denmark	○	●	14515	1.0	●	○	●	○	●	256	1	both	1.28 [1.02,1.61]	1.62 [1.37,1.92]
Lip 2015	Denmark	○	●	15860	1.0	●	○	●	○	○	188	1 2	men women	1.06 [NR]	1.55 [NR]
Olesen 2011	Denmark	○	●	8203	5.0	●	○	●	●	○	NR	1	both	NR [NR]	1.51 [1.37,1.67]
Lip 2015	Denmark	○	●	15860	1.0	●	○	○	○	○	182	1 2	men women	1.02 [NR]	1.50 [NR]
Olesen 2011	Denmark	○	●	8203	10.0	●	○	●	●	○	NR	1	both	NR [NR]	1.45 [1.32,1.58]
Olesen 2012	Denmark	○	●	10062	12.0	●	○	●	○	●	662	1	both	NR [NR]	1.44 [1.34,1.56]
Lip 2015	Denmark	○	●	15860	4.5	●	○	●	○	○	987	1 2	men women	1.08 [NR]	1.24 [NR]
Lip 2015	Denmark	○	●	15860	4.5	●	○	○	○	○	936	1 2	men women	1.02 [NR]	1.18 [NR]
Friberg 2012	Sweden	○	●	6770	3.5	●	●	●	○	●	NR	1	both	NR [NR]	0.90 [NR]
Guo 2012	China	○	●	114	3.0	●	○	●	●	○	NR	1	both	NR [NR]	0.90 [NR]
Allan 2016	**England**	●	●	**5637**	**12.0**	●	●	●	●	●	**153**	**1**	**both**	**0.4 [0.3,0.7]**	**0.7 [0.6,0.8]**
Lip 2010	Europe	○	●	162	1.0	●	○	●	●	○	1	1	both	NR [NR]	0.60 [0.00,3.40]
Friberg 2012	Sweden	○	●	6770	3.5	●	○	○	○	○	NR	1	both	NR [NR]	0.60 [NR]
Friberg 2015	Sweden	○	●	NR	5.0	●	○	○	○	○	NR	1	men	NR [NR]	[0.50,0.70]
Forslund 2014	Sweden	●	●	6682	1.0	●	○	○	○	○	NR	≤1	both	[0.00,0.30]	[0.30,0.50]
Friberg 2015	Sweden	○	●	NR	5.0	●	○	○	○	○	NR	1	women	NR [NR]	[0.10,0.20]

Abbreviations: CI – confidence interval, PY – person-years, PC –primary care, SC – secondary care, IS - ischaemic stroke, US – unclassified stroke, SE – systemic embolism, PE – pulmonary embolism, TIA - transient ischaemic attack, NR – not reported, ○ – no, ● – yes. Example: Huang (2014) used secondary care data for identifying patients, risk factors and endpoints, and included ischaemic stroke in the endpoint definition.

Reference pertaining to each report: Huang 2014,[256] Chao 2015,[257] Chao 2015,[257] Olesen 2011,[258] Olesen 2012,[259] Olesen 2011,[260] Lip 2015,[261] Olesen 2011,[258] Lip 2015,[261] Olesen 2011,[258] Olesen 2012,[259] Lip 2015,[261] Lip 2015,[261] Friberg 2012,[262] Guo 2012,[263] Allan 2016,[7] Lip 2010,[29] Friberg 2012,[262] Friberg 2015,[253] Forslund 2014,[264] Friberg 2015.[253]

Table S7.8 Relative risks of ischaemic and haemorrhagic stroke as reported in clinical trials of direct oral anticoagulants

Trial	Follow-up in years	Ischaemic stroke		Haemorrhagic stroke	
		Definition	Relative risk	Definition	Relative risk
RE-LY					
Dabigatran 110mg	2.0	Ischaemic or unspecified stroke	1.11 [0.89,1.40]	Haemorrhagic stroke	0.31 [0.17,0.56]
Dabigatran 150mg	2.0	Ischaemic or unspecified stroke	0.76 [0.60,0.98]	Haemorrhagic stroke	0.26 [0.14,0.49]
ROCKET AF					
Rivaroxaban as treated	1.9	Stroke or systemic embolism	0.79 [0.66,0.96]	Intracranial haemorrhage	0.67 [0.47,0.93]
Rivaroxaban intention to treat	1.9	Stroke or systemic embolism	0.88 [0.75,1.03]	Intracranial haemorrhage	0.67 [0.47,0.93]
ARISTOTLE					
Apixaban	1.8	Ischaemic or uncertain type of stroke	0.92 [0.74,1.13]	Haemorrhagic stroke	0.51 [0.35,0.75]
ENGAGE AF-TIMI 48					
Edoxaban 30mg	2.8	Ischaemic stroke	1.41 [1.19,1.67]	Haemorrhagic stroke	0.33 [0.22,0.50]
Edoxaban 60mg	2.8	Ischaemic stroke	1.00 [0.83,1.19]	Haemorrhagic stroke	0.54 [0.38,0.77]

References pertaining to each report: RE-LY,[31] ROCKET-AF,[32] ARISTOTLE,[33] and ENGAGE AF-TIMI 48.[34]

Table S7.9 Net clinical benefit [95% confidence intervals] per 100 person-years of warfarin or direct oral anticoagulants compared to no treatment, by CHA_2DS_2-VASC scores, and sex

Score	Events		Net clinical benefit of oral anticoagulants vs. no treatment							
			Warfarin	Dabigatran		Rivaroxaban		Apixaban	Edoxaban	
	IS	HS		110mg	150mg	AT	ITT		30mg	60mg
Overall population										
0	28	8	-0.3 [-0.8,0.1]	-0.2 [-0.6,0.1]	0.0 [-0.3,0.2]	-0.2 [-0.5,0.1]	-0.2 [-0.6,0.1]	-0.2 [-0.5,0.1]	-0.3 [-0.8,0.1]	-0.2 [-0.6,0.1]
1	153	54	0.1 [-0.2,0.4]	0.4 [0.1,0.6]	0.5 [0.3,0.7]	0.3 [0.1,0.6]	0.3 [0.1,0.5]	0.4 [0.1,0.6]	0.2 [-0.1,0.5]	0.3 [0.1,0.5]
2	453	95	0.2 [-0.1,0.6]	0.6 [0.3,0.9]	0.9 [0.7,1.1]	0.6 [0.4,0.9]	0.5 [0.3,0.8]	0.6 [0.4,0.9]	0.3 [0.0,0.6]	0.5 [0.3,0.8]
3	1013	165	1.5 [1.2,1.8]	1.8 [1.5,2.0]	2.2 [1.9,2.4]	1.9 [1.6,2.1]	1.8 [1.5,2.1]	1.8 [1.6,2.1]	1.4 [1.1,1.7]	1.7 [1.5,2.0]
4	1673	237	2.2 [1.8,2.6]	2.5 [2.2,2.9]	3.2 [2.9,3.4]	2.8 [2.5,3.1]	2.7 [2.3,3.0]	2.7 [2.4,3.0]	2.0 [1.6,2.4]	2.6 [2.2,2.9]
5	1416	180	3.2 [2.6,3.8]	3.4 [2.8,4.0]	4.6 [4.1,5.1]	4.1 [3.6,4.7]	3.8 [3.3,4.4]	3.9 [3.3,4.4]	2.5 [1.8,3.1]	3.6 [3.0,4.2]
6	1360	104	7.7 [6.7,8.8]	8.1 [7.0,9.1]	9.6 [8.7,10.5]	9.0 [8.0,9.9]	8.6 [7.6,9.6]	8.6 [7.6,9.6]	6.8 [5.6,8.0]	8.3 [7.2,9.3]
7	692	45	7.2 [5.2,9.1]	7.3 [5.1,9.3]	9.8 [8.1,11.6]	9.1 [7.4,10.8]	8.5 [6.7,10.3]	8.4 [6.5,10.2]	5.1 [2.6,7.5]	7.8 [5.8,9.7]
8	185	18	12.8 [8.9,16.9]	13.6 [9.7,17.5]	15.4 [12.0,19.0]	14.5 [10.8,18.2]	14.0 [10.3,17.9]	14.1 [10.5,18.0]	12.1 [7.8,16.4]	13.7 [10.0,17.6]
9	32	0	16.8 [1.8,31.5]	16.0 [0.3,31.3]	18.6 [5.8,32.5]	18.4 [5.2,32.4]	17.7 [3.6,32.0]	17.4 [3.0,31.7]	13.7 [-4.3,30.0]	16.8 [1.8,31.5]
	7005	**906**	**1.9 [1.8,2.1]**	**2.2 [2.1,2.4]**	**2.9 [2.7,3.0]**	**2.5 [2.4,2.7]**	**2.4 [2.2,2.5]**	**2.4 [2.3,2.6]**	**1.7 [1.5,1.9]**	**2.3 [2.1,2.4]**
Men										
0	28	8	-0.3 [-0.8,0.1]	-0.1 [-0.6,0.1]	0.0 [-0.3,0.2]	-0.2 [-0.5,0.1]	-0.2 [-0.6,0.1]	-0.2 [-0.5,0.1]	-0.3 [-0.8,0.1]	-0.2 [-0.6,0.1]
1	137	48	0.1 [-0.2,0.4]	0.5 [0.1,0.7]	0.6 [0.4,0.8]	0.4 [0.1,0.6]	0.3 [0.1,0.6]	0.4 [0.1,0.6]	0.3 [-0.1,0.6]	0.3 [0.1,0.6]
2	381	79	0.5 [0.1,0.9]	1.0 [0.6,1.2]	1.2 [0.9,1.5]	0.9 [0.6,1.2]	0.8 [0.5,1.1]	0.9 [0.6,1.2]	0.6 [0.2,1.0]	0.8 [0.5,1.1]
3	656	103	1.5 [1.1,1.9]	2.0 [1.4,2.2]	2.3 [1.9,2.6]	2.0 [1.6,2.3]	1.8 [1.5,2.2]	1.9 [1.5,2.3]	1.4 [0.9,1.8]	1.8 [1.4,2.2]
4	608	93	2.0 [1.3,2.7]	2.8 [1.7,3.0]	3.2 [2.6,3.7]	2.8 [2.2,3.3]	2.6 [1.9,3.2]	2.6 [2.0,3.2]	1.7 [0.9,2.4]	2.4 [1.8,3.0]
5	613	88	3.9 [2.6,4.9]	5.0 [3.1,5.3]	5.7 [4.7,6.7]	5.1 [4.1,6.1]	4.7 [3.7,5.7]	4.8 [3.7,5.8]	3.0 [1.6,4.2]	4.4 [3.3,5.5]
6	363	26	7.1 [5.2,9.1]	8.3 [5.6,9.4]	8.9 [7.3,10.7]	8.4 [6.6,10.1]	8.0 [6.2,9.8]	8.0 [6.2,9.8]	6.2 [4.0,8.4]	7.6 [5.8,9.5]
7	156	15	8.6 [4.7,13.0]	10.1 [4.8,13.2]	11.1 [7.7,15.0]	10.4 [6.9,14.3]	9.8 [6.2,13.9]	9.8 [6.1,13.9]	6.9 [2.3,11.6]	9.2 [5.4,13.5]
8	21	3	15.9 [7.0,25.7]	16.0 [6.4,25.7]	16.3 [8.0,25.8]	16.3 [7.8,25.8]	16.1 [7.4,25.8]	16.0 [7.3,25.7]	15.1 [5.0,25.7]	15.9 [7.0,25.7]
	2963	**463**	**1.2 [1.0,1.4]**	**1.8 [1.3,1.7]**	**2.1 [1.9,2.3]**	**1.8 [1.6,1.9]**	**1.6 [1.4,1.8]**	**1.7 [1.5,1.9]**	**1.0 [0.8,1.3]**	**1.5 [1.3,1.7]**

Women										
1	16	6	0.3 [-0.4,0.8]	0.3 [-0.5,0.8]	0.4 [-0.2,0.8]	0.4 [-0.2,0.8]	0.3 [-0.3,0.8]	0.3 [-0.3,0.8]	0.1 [-0.8,0.8]	0.3 [-0.4,0.8]
2	72	16	-0.1 [-0.6,0.3]	0.3 [-0.2,0.5]	0.4 [0.0,0.7]	0.1 [-0.2,0.5]	0.1 [-0.3,0.5]	0.2 [-0.2,0.5]	0.0 [-0.5,0.4]	0.1 [-0.3,0.5]
3	357	62	1.5 [1.1,1.9]	1.9 [1.4,2.2]	2.0 [1.7,2.4]	1.8 [1.4,2.2]	1.8 [1.4,2.1]	1.8 [1.5,2.2]	1.6 [1.1,2.0]	1.7 [1.4,2.1]
4	1065	144	2.4 [2.0,2.8]	3.0 [2.4,3.1]	3.3 [2.9,3.6]	2.9 [2.6,3.3]	2.8 [2.4,3.2]	2.9 [2.5,3.2]	2.3 [1.9,2.8]	2.7 [2.4,3.1]
5	803	92	3.1 [2.3,3.8]	3.7 [2.5,3.9]	4.1 [3.6,4.7]	3.8 [3.2,4.4]	3.6 [2.9,4.2]	3.6 [2.9,4.2]	2.5 [1.6,3.3]	3.4 [2.7,4.0]
6	997	78	8.0 [6.6,9.3]	9.2 [7.0,9.6]	9.8 [8.7,11.0]	9.2 [8.0,10.4]	8.9 [7.6,10.1]	8.9 [7.6,10.1]	7.0 [5.5,8.5]	8.5 [7.2,9.8]
7	536	30	6.8 [4.4,9.1]	8.2 [4.5,9.1]	9.4 [7.5,11.4]	8.7 [6.7,10.8]	8.0 [6.0,10.2]	7.9 [5.9,10.1]	4.5 [1.8,7.3]	7.3 [5.1,9.6]
8	164	15	12.4 [7.9,17.3]	14.3 [8.8,18.0]	15.2 [11.4,19.5]	14.2 [10.3,18.8]	13.7 [9.7,18.4]	13.9 [9.8,18.4]	11.7 [6.6,16.8]	13.4 [9.1,18.0]
9	32	0	16.8 [3.1,30.5]	17.5 [1.5,30.3]	18.6 [6.6,31.3]	18.4 [6.1,31.2]	17.7 [4.8,30.9]	17.4 [4.2,30.8]	13.7 [-3.9,29.3]	16.8 [3.1,30.5]
	4042	**443**	**2.7 [2.4,3.0]**	**3.4 [2.7,3.3]**	**3.7 [3.5,3.9]**	**3.4 [3.1,3.6]**	**3.2 [2.9,3.5]**	**3.2 [3.0,3.5]**	**2.4 [2.1,2.7]**	**3.0 [2.8,3.3]**

Abbreviations: IS - ischaemic stroke, HS – haemorrhagic stroke, AT – as treated, ITT – intention to treat.

Chapter 8

Overall discussion of novel contributions, strengths, limitations and conclusion

No supplementary material.

Abbreviations

ACS	Acute Coronary Syndrome
AF	Atrial Fibrillation
AF+/-	Atrial Fibrillation with/without intercurrent cardiovascular disease
AGES	Age, Gene and Environment–Reykjavik study
ARIC	Atherosclerosis Risk in Communities
ARISTOTLE	Apixaban for Reduction in Stroke and Other Thromboembolic Events in Atrial Fibrillation
ATC [codes]	Anatomical Therapeutic Chemical Classification System
AV	Atrioventricular [node]
BHS	Busselton Health Study
BMI	Body Mass Index
BNF	British National Formulary
BP	Blood Pressure
CALIBER	ClinicAl research using LInked Bespoke studies and Electronic health Records
CATCH ME	Characterizing Atrial fibrillation by Translating its Causes into Health Modifiers in the Elderly
CCHS	Copenhagen City Heart Study
CHA_2DS_2-VASC	Congestive heart failure, Hypertension, Age ≥75 years, Diabetes mellitus, history of Stroke or thromboembolism, Vascular disease, Age 65–74 years, and Sex category
CHARGE-AF	Cohorts for Heart and Aging Research in Genomic Epidemiology – AF
CHD	Coronary Heart Disease
CHS	Cardiovascular Health Study
CI	Confidence Intervals
CIRCS	Circulatory Risk in Communities Study
COSM	Cohort of Swedish Men
CPRD / CPRD GOLD	Clinical Practice Research Datalink / GP OnLine Database
CRP	C-reactive protein
CVD	Cardiovascular Disease
DBP	Diastolic Blood Pressure
DCHS	Diet Cancer and Health study
D–EHR	Denmark Electronic Health Record cohort

DOAC	Direct Oral Anticoagulant
ECG	electrocardiography; electrocardiogram
ECHOES	Echocardiographic Heart Of England Screening
EHR	Electronic Health Records
ENGAGE AF-TIMI 48	Effective Anticoagulation with Factor Xa Next Generation in Atrial Fibrillation–Thrombolysis in Myocardial Infarction 48.
ESC	European Society of Cardiology
FHS	Framingham Heart Study
GP	General Practice; General Practitioners
GPPS	Göteborg Primary Prevention Study
HABC	Health, Aging, and Body Composition
HAS-BLED	Hypertension, Abnormal renal and liver function, Stroke, Bleeding, Labile INR, Elderly, Drugs or alcohol
HCUP	Healthcare Cost and Utilization Project
HDL	High–Density Lipoprotein [cholesterol]
HES	Hospital Episode Statistics
HR	Hazard Ratio
HS	Haemorrhagic Stroke
ICD / ICDCM	International Statistical Classification of Diseases and Health-Related Problems / Clinical Modification
IMI	Innovative Medicines Initiative
INR	International Normalised Ratio
IPHS	Ibaraki prefectural health study
IQI/R	Interquartile Interval/Range
IS	Ischaemic Stroke
ISAC	Independent Scientific Advisory Committee
L85PS	Leiden 85–Plus Study
LDL	Low Density Lipoprotein [cholesterol]
LOESS	LOcally wEighted Scatterplot Smoothing
LV	Left Ventricular
MCS	Malmö Cardiovascular Screening
MDCS	Malmö Diet and Cancer study
MESA	Multi–Ethnic Study of Atherosclerosis
MINAP	Myocardial Ischaemia National Audit Project

MPP	Malmö Preventive Project
NCB	Net Clinical Benefit
NHS	National Health Service
NICOR	National Institute for Cardiovascular Outcomes Research
NIH	National Institutes of Health
NorPD	Norwegian Prescription Database
NOS	Not Otherwise Specified
NPMS	Niigata preventive medicine study
NR	Not Reported
OCS	Oslo Cardiovascular Survey
ONS	Office for National Statistics
OPCS	Office of Population Censuses and Surveys' Classification of Interventions and Procedures
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analyses
PY	Person Years
QOF	Quality and Outcomes Framework
RE-LY	Randomized Evaluation of Long-Term Anticoagulation Therapy
ROCKET AF	Rivaroxaban Once Daily Oral Direct Factor Xa Inhibition Compared with Vitamin K Antagonism for Prevention of Stroke and Embolism Trial in Atrial Fibrillation
RR	Relative Risk
RS	Rotterdam Study
SA	Sinoatrial [node]
SBP	Systolic Blood Pressure
SD	Standard Deviation
SE	Systemic Embolism
S–EHR	Sweden Electronic Health Record cohort
SES	Socio-Economic Status
SHIP	Study of Health in Pomerania
S–HS	Stockholm Health Screening cohort
SLMS	[UCL] School of Life and Medical Sciences
SMC	Swedish Mammography Cohort
SNOMED CT	Systematised Nomenclature Of Medicine Clinical Terms
TIA	Transient Ischaemic Attack

T–NHIRD	Taiwan National Health Insurance Research Database
TS	Tromsø Study
TSH	Thyroid Stimulating Hormone
TSS	The Suita Study
UK	United Kingdom
US	United States
VHD	Valvular Heart Disease
WHI–OS	Women's Health Initiative Observational Study
WHS	Women's Health Study

Bibliography

1. Cottrell C. Atrial fibrillation part 1: pathophysiology. Practice Nursing 2012;23(1):16-21.
2. Allan V, Honarbakhsh S, Casas JP, et al. Are cardiovascular risk factors also associated with the incidence of atrial fibrillation? A systematic review and field synopsis of 23 factors in 32 population-based cohorts of 20 million participants. Thrombosis and haemostasis 2017;117(5):837-50.
3. Kirchhof P, Benussi S, Kotecha D, et al. 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS. European heart journal 2016;37(38):2893-962.
4. Joundi RA, Cipriano LE, Sposato LA, et al. Ischemic Stroke Risk in Patients With Atrial Fibrillation and CHA2DS2-VASc Score of 1: Systematic Review and Meta-Analysis. Stroke 2016;47(5):1364-7.
5. Denaxas SC, George J, Herrett E, et al. Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). International journal of epidemiology 2012;41(6):1625-38.
6. De Caterina R, Camm AJ. What is 'valvular' atrial fibrillation? A reappraisal. European heart journal 2014;35(47):3328-35.
7. Allan V, Banerjee A, Shah AD, et al. Net clinical benefit of warfarin in individuals with atrial fibrillation across stroke risk and across primary and secondary care. Heart (British Cardiac Society) 2017;103(3):210-18.
8. Chugh SS, Havmoeller R, Narayanan K, et al. Worldwide epidemiology of atrial fibrillation: a Global Burden of Disease 2010 Study. Circulation 2014;129(8):837-47.
9. Miyasaka Y, Barnes ME, Gersh BJ, et al. Secular trends in incidence of atrial fibrillation in Olmsted County, Minnesota, 1980 to 2000, and implications on the projections for future prevalence. Circulation 2006;114(2):119-25.
10. Krijthe BP, Kunst A, Benjamin EJ, et al. Projections on the number of individuals with atrial fibrillation in the European Union, from 2000 to 2060. European heart journal 2013;34(35):2746-51.
11. Lane DA, Skjoth F, Lip GYH, et al. Temporal Trends in Incidence, Prevalence, and Mortality of Atrial Fibrillation in Primary Care. Journal of the American Heart Association 2017;6(5).
12. MacLeod KT. *An Essential Introduction to Cardiac Electrophysiology*, 2013.
13. Wolf PA, Abbott RD, Kannel WB. Atrial fibrillation as an independent risk factor for stroke: the Framingham Study. Stroke 1991;22(8):983-8.
14. Marini C, De Santis F, Sacco S, et al. Contribution of atrial fibrillation to incidence and outcome of ischemic stroke: results from a population-based study. Stroke 2005;36(6):1115-9.
15. Benjamin EJ, Wolf PA, D'Agostino RB, et al. Impact of atrial fibrillation on the risk of death: the Framingham Heart Study. Circulation 1998;98(10):946-52.
16. Kassianos G, Arden C, Hogan S, et al. Current management of atrial fibrillation: an observational study in NHS primary care. BMJ open 2013;3(11):e003004.
17. NHS Improvement. Atrial fibrillation in primary care: making an impact on stroke prevention. October 2009. Available at: http://webarchive.nationalarchives.gov.uk/20130221101407/http://www.improvement.nhs.uk/LinkClick.aspx?fileticket=%2bLIKN1gSgOA%3d&tabid=62 [accessed 28 November 2017].
18. January CT, Wann LS, Alpert JS, et al. 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society. Journal of the American College of Cardiology 2014;64(21):e1-76.
19. Jones C, Pollit V, Fitzmaurice D, et al. The management of atrial fibrillation: summary of updated NICE guidance. BMJ (Clinical research ed) 2014;348:g3655.
20. Arrhythmia Alliance. Know Your Pulse. Available at: http://www.heartrhythmalliance.org/aa/uk/know-your-pulse [accessed 28 November 2017].
21. Kotecha D, Chua WWL, Fabritz L, et al. European Society of Cardiology smartphone and tablet applications for patients with atrial fibrillation and their health care providers. Europace : European pacing, arrhythmias, and cardiac electrophysiology : journal of the

working groups on cardiac pacing, arrhythmias, and cardiac cellular electrophysiology of the European Society of Cardiology 2017.
22. January CT, Wann LS, Alpert JS, et al. 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on practice guidelines and the Heart Rhythm Society. Circulation 2014;130(23):2071-104.
23. Aguilar MI, Hart R. Oral anticoagulants for preventing stroke in patients with non-valvular atrial fibrillation and no previous history of stroke or transient ischemic attacks. The Cochrane database of systematic reviews 2005(3):Cd001927.
24. Saxena R, Koudstaal PJ. Anticoagulants for preventing stroke in patients with nonrheumatic atrial fibrillation and a history of stroke or transient ischaemic attack. The Cochrane database of systematic reviews 2004(2):Cd000185.
25. Zimetbaum P. Antiarrhythmic drug therapy for atrial fibrillation. Circulation 2012;125(2):381-9.
26. Kotecha D, Calvert M, Deeks JJ, et al. A review of rate control in atrial fibrillation, and the rationale and protocol for the RATE-AF trial. BMJ open 2017;7(7):e015099.
27. Honarbakhsh S, Finlay M, Earley MJ, et al. Management of atrial fibrillation: when are invasive approaches useful? British journal of hospital medicine (London, England : 2005) 2016;77(8):460-6.
28. Nyong J, Amit G, Adler AJ, et al. Efficacy and safety of ablation for people with non-paroxysmal atrial fibrillation. The Cochrane database of systematic reviews 2016;11:Cd012088.
29. Lip GY, Nieuwlaat R, Pisters R, et al. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest 2010;137(2):263-72.
30. Pisters R, Lane DA, Nieuwlaat R, et al. A novel user-friendly score (HAS-BLED) to assess 1-year risk of major bleeding in patients with atrial fibrillation: the Euro Heart Survey. Chest 2010;138(5):1093-100.
31. Connolly SJ, Ezekowitz MD, Yusuf S, et al. Dabigatran versus warfarin in patients with atrial fibrillation. The New England journal of medicine 2009;361(12):1139-51.
32. Patel MR, Mahaffey KW, Garg J, et al. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. The New England journal of medicine 2011;365(10):883-91.
33. Granger CB, Alexander JH, McMurray JJ, et al. Apixaban versus warfarin in patients with atrial fibrillation. The New England journal of medicine 2011;365(11):981-92.
34. Giugliano RP, Ruff CT, Braunwald E, et al. Edoxaban versus warfarin in patients with atrial fibrillation. The New England journal of medicine 2013;369(22):2093-104.
35. Savelieva I, Kakouros N, Kourliouros A, et al. Upstream therapies for management of atrial fibrillation: review of clinical evidence and implications for European Society of Cardiology guidelines. Part I: primary prevention. Europace : European pacing, arrhythmias, and cardiac electrophysiology : journal of the working groups on cardiac pacing, arrhythmias, and cardiac cellular electrophysiology of the European Society of Cardiology 2011;13(3):308-28.
36. Moran PS, Flattery MJ, Teljeur C, et al. Effectiveness of systematic screening for the detection of atrial fibrillation. The Cochrane database of systematic reviews 2013(4):Cd009586.
37. Levy S, Camm AJ, Saksena S, et al. International consensus on nomenclature and classification of atrial fibrillation; a collaborative project of the Working Group on Arrhythmias and the Working Group on Cardiac Pacing of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Europace : European pacing, arrhythmias, and cardiac electrophysiology : journal of the working groups on cardiac pacing, arrhythmias, and cardiac cellular electrophysiology of the European Society of Cardiology 2003;5(2):119-22.
38. Benjamin EJ, Chen PS, Bild DE, et al. Prevention of atrial fibrillation: report from a national heart, lung, and blood institute workshop. Circulation 2009;119(4):606-18.
39. Hemingway H, Asselbergs FW, Danesh J, et al. Big data from electronic health records for early and late translational cardiovascular research: challenges and potential. European heart journal 2017:ehx487-ehx87.
40. Alonso A, Krijthe BP, Aspelund T, et al. Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: the CHARGE-AF consortium. Journal of the American Heart Association 2013;2(2):e000102.
41. NHS Digital. Hospital Episodes Statistics (HES). Available at: http://content.digital.nhs.uk/hes [accessed 28 November 2017]. .

42. Schmidt M, Schmidt SA, Sandegaard JL, et al. The Danish National Patient Registry: a review of content, data quality, and research potential. Clinical epidemiology 2015;7:449-90.
43. Ludvigsson JF, Andersson E, Ekbom A, et al. External review and validation of the Swedish national inpatient register. BMC public health 2011;11:450.
44. Denaxas SC, Morley KI. Big biomedical data and cardiovascular disease research: opportunities and challenges. European Heart Journal - Quality of Care and Clinical Outcomes 2015;1(1):9-16.
45. World Health Organisation. International Classification of Diseases (ICD). Available at: http://www.who.int/classifications/icd/en/ [accessed 28 November 2017].
46. Wang Z, Shah AD, Tate AR, et al. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. PloS one 2012;7(1):e30412.
47. Vezyridis P, Timmons S. Evolution of primary care databases in UK: a scientometric analysis of research output. BMJ open 2016;6(10):e012785.
48. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. Journal of the American Medical Informatics Association : JAMIA 2013;20(1):117-21.
49. Morley KI, Wallace J, Denaxas SC, et al. Defining disease phenotypes using national linked electronic health records: a case study of atrial fibrillation. PloS one 2014;9(11):e110900.
50. Lyons RA, Jones KH, John G, et al. The SAIL databank: linking multiple health and social care datasets. BMC medical informatics and decision making 2009;9:3.
51. Hemingway H, Feder GS, Fitzpatrick NK, et al. Programme Grants for Applied Research. Using nationwide 'big data' from linked electronic health records to help improve outcomes in cardiovascular diseases: 33 studies using methods from epidemiology, informatics, economics and social science in the ClinicAl disease research using LInked Bespoke studies and Electronic health Records (CALIBER) programme. Southampton (UK): NIHR Journals Library

52. The National Health Service. The NHS in England. Available at: https://www.nhs.uk/NHSEngland/thenhs/about/Pages/overview.aspx [accessed 28 November 2017].
53. Herrett E, Gallagher AM, Bhaskaran K, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). International journal of epidemiology 2015;44(3):827-36.
54. Denaxas SC, Asselbergs FW, Moore JH. The tip of the iceberg: challenges of accessing hospital electronic health record data for biological data mining. BioData mining 2016;9:29.
55. Davis RC, Hobbs FD, Kenkre JE, et al. Prevalence of atrial fibrillation in the general population and in high-risk groups: the ECHOES study. Europace : European pacing, arrhythmias, and cardiac electrophysiology : journal of the working groups on cardiac pacing, arrhythmias, and cardiac cellular electrophysiology of the European Society of Cardiology 2012;14(11):1553-9.
56. Ioannidis JP, Boffetta P, Little J, et al. Assessment of cumulative evidence on genetic associations: interim guidelines. International journal of epidemiology 2008;37(1):120-32.
57. Blomstrom Lundqvist C, Lip GY, Kirchhof P. What are the costs of atrial fibrillation? Europace : European pacing, arrhythmias, and cardiac electrophysiology : journal of the working groups on cardiac pacing, arrhythmias, and cardiac cellular electrophysiology of the European Society of Cardiology 2011;13 Suppl 2:ii9-12.
58. Moran PS, Teljeur C, Ryan M, et al. Systematic screening for the detection of atrial fibrillation. The Cochrane database of systematic reviews 2016(6):Cd009586.
59. Perk J, De Backer G, Gohlke H, et al. European Guidelines on cardiovascular disease prevention in clinical practice (version 2012). The Fifth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease

Prevention in Clinical Practice (constituted by representatives of nine societies and by invited experts). European heart journal 2012;33(13):1635-701.
60. Goldstein LB, Bushnell CD, Adams RJ, et al. Guidelines for the primary prevention of stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. Stroke 2011;42(2):517-84.
61. Samokhvalov AV, Irving HM, Rehm J. Alcohol consumption as a risk factor for atrial fibrillation: a systematic review and meta-analysis. European journal of cardiovascular prevention and rehabilitation : official journal of the European Society of Cardiology, Working Groups on Epidemiology & Prevention and Cardiac Rehabilitation and Exercise Physiology 2010;17(6):706-12.
62. Kodama S, Saito K, Tanaka S, et al. Alcohol consumption and risk of atrial fibrillation: a meta-analysis. Journal of the American College of Cardiology 2011;57(4):427-36.
63. Larsson SC, Drca N, Wolk A. Alcohol consumption and risk of atrial fibrillation: a prospective study and dose-response meta-analysis. Journal of the American College of Cardiology 2014;64(3):281-9.
64. Wu N, Xu B, Xiang Y, et al. Association of inflammatory factors with occurrence and recurrence of atrial fibrillation: a meta-analysis. International journal of cardiology 2013;169(1):62-72.
65. Huxley RR, Filion KB, Konety S, et al. Meta-analysis of cohort and case-control studies of type 2 diabetes mellitus and risk of atrial fibrillation. The American journal of cardiology 2011;108(1):56-62.
66. Wanahita N, Messerli FH, Bangalore S, et al. Atrial fibrillation and obesity--results of a meta-analysis. American heart journal 2008;155(2):310-5.
67. Ofman P, Khawaja O, Rahilly-Tierney CR, et al. Regular physical activity and risk of atrial fibrillation: a systematic review and meta-analysis. Circulation Arrhythmia and electrophysiology 2013;6(2):252-6.
68. Kwok CS, Anderson SG, Myint PK, et al. Physical activity and incidence of atrial fibrillation: a systematic review and meta-analysis. International journal of cardiology 2014;177(2):467-76.
69. Zimmerman D, Sood MM, Rigatto C, et al. Systematic review and meta-analysis of incidence, prevalence and outcomes of atrial fibrillation in patients on dialysis. Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association 2012;27(10):3816-22.
70. Kuper H, Nicholson A, Kivimaki M, et al. Evaluating the causal relevance of diverse risk markers: horizontal systematic review. BMJ (Clinical research ed) 2009;339:b4265.
71. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ (Clinical research ed) 2009;339:b2535.
72. Hemingway H, Philipson P, Chen R, et al. Evaluating the quality of research into a single prognostic biomarker: a systematic review and meta-analysis of 83 studies of C-reactive protein in stable coronary artery disease. PLoS medicine 2010;7(6):e1000286.
73. Azarbal F, Stefanick ML, Salmoirago-Blotcher E, et al. Obesity, physical activity, and their interaction in incident atrial fibrillation in postmenopausal women. Journal of the American Heart Association 2014;3(4).
74. Perez MV, Wang PJ, Larson JC, et al. Risk factors for atrial fibrillation and their population burden in postmenopausal women: the Women's Health Initiative Observational Study. Heart (British Cardiac Society) 2013;99(16):1173-8.
75. Drca N, Wolk A, Jensen-Urstad M, et al. Atrial fibrillation is associated with different levels of physical activity levels at different ages in men. Heart (British Cardiac Society) 2014;100(13):1037-42.
76. Watanabe H, Watanabe T, Sasaki S, et al. Close bidirectional relationship between chronic kidney disease and atrial fibrillation: the Niigata preventive medicine study. American heart journal 2009;158(4):629-36.
77. Watanabe H, Tanabe N, Watanabe T, et al. Metabolic syndrome and risk of development of atrial fibrillation: the Niigata preventive medicine study. Circulation 2008;117(10):1255-60.
78. Watanabe H, Tanabe N, Yagihara N, et al. Association between lipid profile and risk of atrial fibrillation. Circulation journal : official journal of the Japanese Circulation Society 2011;75(12):2767-74.
79. Drca N, Wolk A, Jensen-Urstad M, et al. Physical activity is associated with a reduced risk of atrial fibrillation in middle-aged and elderly women. Heart (British Cardiac Society) 2015;101(20):1627-30.

80. Frost L, Benjamin EJ, Fenger-Gron M, et al. Body fat, body fat distribution, lean body mass and atrial fibrillation and flutter. A Danish cohort study. Obesity (Silver Spring, Md) 2014;22(6):1546-52.
81. Frost L, Hune LJ, Vestergaard P. Overweight and obesity as risk factors for atrial fibrillation or flutter: the Danish Diet, Cancer, and Health Study. The American journal of medicine 2005;118(5):489-95.
82. Frost L, Frost P, Vestergaard P. Work related physical activity and risk of a hospital discharge diagnosis of atrial fibrillation or flutter: the Danish Diet, Cancer, and Health Study. Occupational and environmental medicine 2005;62(1):49-53.
83. Fedorowski A, Hedblad B, Engstrom G, et al. Orthostatic hypotension and long-term incidence of atrial fibrillation: the Malmo Preventive Project. Journal of internal medicine 2010;268(4):383-9.
84. Misialek JR, Rose KM, Everson-Rose SA, et al. Socioeconomic status and the incidence of atrial fibrillation in whites and blacks: the Atherosclerosis Risk in Communities (ARIC) study. Journal of the American Heart Association 2014;3(4).
85. Huxley RR, Misialek JR, Agarwal SK, et al. Physical activity, obesity, weight change, and risk of atrial fibrillation: the Atherosclerosis Risk in Communities study. Circulation Arrhythmia and electrophysiology 2014;7(4):620-5.
86. Huxley RR, Lopez FL, Folsom AR, et al. Absolute and attributable risks of atrial fibrillation in relation to optimal and borderline risk factors: the Atherosclerosis Risk in Communities (ARIC) study. Circulation 2011;123(14):1501-8.
87. Lopez FL, Agarwal SK, Maclehose RF, et al. Blood lipid levels, lipid-lowering medications, and the incidence of atrial fibrillation: the atherosclerosis risk in communities study. Circulation Arrhythmia and electrophysiology 2012;5(1):155-62.
88. Alonso A, Tang W, Agarwal SK, et al. Hemostatic markers are associated with the risk and prognosis of atrial fibrillation: the ARIC study. International journal of cardiology 2012;155(2):217-22.
89. Alonso A, Agarwal SK, Soliman EZ, et al. Incidence of atrial fibrillation in whites and African-Americans: the Atherosclerosis Risk in Communities (ARIC) study. American heart journal 2009;158(1):111-7.
90. Marcus GM, Alonso A, Peralta CA, et al. European ancestry as a risk factor for atrial fibrillation in African Americans. Circulation 2010;122(20):2009-15.
91. Alonso A, Lopez FL, Matsushita K, et al. Chronic kidney disease is associated with the incidence of atrial fibrillation: the Atherosclerosis Risk in Communities (ARIC) study. Circulation 2011;123(25):2946-53.
92. Chamberlain AM, Agarwal SK, Folsom AR, et al. A clinical risk score for atrial fibrillation in a biracial prospective cohort (from the Atherosclerosis Risk in Communities [ARIC] study). The American journal of cardiology 2011;107(1):85-91.
93. Jensen PN, Thacker EL, Dublin S, et al. Racial differences in the incidence of and risk factors for atrial fibrillation in older adults: the cardiovascular health study. Journal of the American Geriatrics Society 2013;61(2):276-80.
94. Mozaffarian D, Furberg CD, Psaty BM, et al. Physical activity and incidence of atrial fibrillation in older adults: the cardiovascular health study. Circulation 2008;118(8):800-7.
95. Aviles RJ, Martin DO, Apperson-Hansen C, et al. Inflammation as a risk factor for atrial fibrillation. Circulation 2003;108(24):3006-10.
96. Cappola AR, Arnold AM, Wulczyn K, et al. Thyroid function in the euthyroid range and adverse outcomes in older adults. The Journal of clinical endocrinology and metabolism 2015;100(3):1088-96.
97. Deo R, Katz R, Kestenbaum B, et al. Impaired kidney function and atrial fibrillation in elderly subjects. Journal of cardiac failure 2010;16(1):55-60.
98. Psaty BM, Manolio TA, Kuller LH, et al. Incidence of and risk factors for atrial fibrillation in older adults. Circulation 1997;96(7):2455-61.
99. Smith JG, Platonov PG, Hedblad B, et al. Atrial fibrillation in the Malmo Diet and Cancer study: a study of occurrence, risk factors and diagnostic validity. European journal of epidemiology 2010;25(2):95-102.
100. Smith JG, Newton-Cheh C, Almgren P, et al. Assessment of conventional cardiovascular risk factors and multiple biomarkers for the prediction of incident heart failure and atrial fibrillation. Journal of the American College of Cardiology 2010;56(21):1712-9.
101. Rosengren A, Hauptman PJ, Lappas G, et al. Big men and atrial fibrillation: effects of body size and weight gain on risk of atrial fibrillation in men. European heart journal 2009;30(9):1113-20.

102. Xu D, Murakoshi N, Sairenchi T, et al. Anemia and reduced kidney function as risk factors for new onset of atrial fibrillation (from the Ibaraki prefectural health study). The American journal of cardiology 2015;115(3):328-33.
103. Schoen T, Pradhan AD, Albert CM, et al. Type 2 diabetes mellitus and risk of incident atrial fibrillation in women. Journal of the American College of Cardiology 2012;60(15):1421-8.
104. Everett BM, Conen D, Buring JE, et al. Physical activity and the risk of incident atrial fibrillation in women. Circulation Cardiovascular quality and outcomes 2011;4(3):321-7.
105. Tedrow UB, Conen D, Ridker PM, et al. The long- and short-term impact of elevated body mass index on the risk of new atrial fibrillation the WHS (women's health study). Journal of the American College of Cardiology 2010;55(21):2319-27.
106. Mora S, Akinkuolie AO, Sandhu RK, et al. Paradoxical association of lipoprotein measures with incident atrial fibrillation. Circulation Arrhythmia and electrophysiology 2014;7(4):612-9.
107. Sandhu RK, Kurth T, Conen D, et al. Relation of renal function to risk for incident atrial fibrillation in women. The American journal of cardiology 2012;109(4):538-42.
108. Conen D, Ridker PM, Everett BM, et al. A multimarker approach to assess the influence of inflammation on the incidence of atrial fibrillation in women. European heart journal 2010;31(14):1730-6.
109. Conen D, Tedrow UB, Cook NR, et al. Alcohol consumption and risk of incident atrial fibrillation in women. Jama 2008;300(21):2489-96.
110. Conen D, Tedrow UB, Koplan BA, et al. Influence of systolic and diastolic blood pressure on the risk of incident atrial fibrillation in women. Circulation 2009;119(16):2146-52.
111. Thelle DS, Selmer R, Gjesdal K, et al. Resting heart rate and physical activity as risk factors for lone atrial fibrillation: a prospective study of 309,540 men and women. Heart (British Cardiac Society) 2013;99(23):1755-60.
112. Nyrnes A, Mathiesen EB, Njolstad I, et al. Palpitations are predictive of future atrial fibrillation. An 11-year follow-up of 22,815 men and women: the Tromso Study. European journal of preventive cardiology 2013;20(5):729-36.
113. Nyrnes A, Njolstad I, Mathiesen EB, et al. Inflammatory biomarkers as risk factors for future atrial fibrillation. An eleven-year follow-up of 6315 men and women: the Tromso study. Gender medicine 2012;9(6):536-47.e2.
114. Mitchell GF, Vasan RS, Keyes MJ, et al. Pulse pressure and risk of new-onset atrial fibrillation. Jama 2007;297(7):709-15.
115. Schnabel RB, Sullivan LM, Levy D, et al. Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study. Lancet (London, England) 2009;373(9665):739-45.
116. Alonso A, Yin X, Roetker NS, et al. Blood lipids and the incidence of atrial fibrillation: the Multi-Ethnic Study of Atherosclerosis and the Framingham Heart Study. Journal of the American Heart Association 2014;3(5):e001211.
117. Sawin CT, Geller A, Wolf PA, et al. Low serum thyrotropin concentrations as a risk factor for atrial fibrillation in older persons. The New England journal of medicine 1994;331(19):1249-52.
118. Schnabel RB, Larson MG, Yamamoto JF, et al. Relation of multiple inflammatory biomarkers to incident atrial fibrillation. The American journal of cardiology 2009;104(1):92-6.
119. Adamsson Eryd S, Smith JG, Melander O, et al. Inflammation-sensitive proteins and risk of atrial fibrillation: a population-based cohort study. European journal of epidemiology 2011;26(6):449-55.
120. Sinner MF, Stepas KA, Moser CB, et al. B-type natriuretic peptide and C-reactive protein in the prediction of atrial fibrillation risk: the CHARGE-AF Consortium of community-based cohort studies. Europace : European pacing, arrhythmias, and cardiac electrophysiology : journal of the working groups on cardiac pacing, arrhythmias, and cardiac cellular electrophysiology of the European Society of Cardiology 2014;16(10):1426-33.
121. Schnabel RB, Aspelund T, Li G, et al. Validation of an atrial fibrillation risk algorithm in whites and African Americans. Archives of internal medicine 2010;170(21):1909-17.
122. Chaker L, Heeringa J, Dehghan A, et al. Normal Thyroid Function and the Risk of Atrial Fibrillation: the Rotterdam Study. The Journal of clinical endocrinology and metabolism 2015;100(10):3718-24.
123. Heeringa J, Kors JA, Hofman A, et al. Cigarette smoking and risk of atrial fibrillation: the Rotterdam Study. American heart journal 2008;156(6):1163-9.

124. Friberg J, Buch P, Scharling H, et al. Rising rates of hospital admissions for atrial fibrillation. Epidemiology (Cambridge, Mass) 2003;14(6):666-72.
125. Mukamal KJ, Tolstrup JS, Friberg J, et al. Fibrinogen and albumin levels and risk of atrial fibrillation in men and women (the Copenhagen City Heart Study). The American journal of cardiology 2006;98(1):75-81.
126. Aronis KN, Wang N, Phillips CL, et al. Associations of obesity and body fat distribution with incident atrial fibrillation in the biracial health aging and body composition cohort of older adults. American heart journal 2015;170(3):498-505.e2.
127. Collet TH, Gussekloo J, Bauer DC, et al. Subclinical hyperthyroidism and the risk of coronary heart disease and mortality. Archives of internal medicine 2012;172(10):799-809.
128. Knuiman M, Briffa T, Divitini M, et al. A cohort study examination of established and emerging risk factors for atrial fibrillation: the Busselton Health Study. European journal of epidemiology 2014;29(3):181-90.
129. Roetker NS, Chen LY, Heckbert SR, et al. Relation of systolic, diastolic, and pulse pressures and aortic distensibility with atrial fibrillation (from the Multi-Ethnic Study of Atherosclerosis). The American journal of cardiology 2014;114(4):587-92.
130. Rodriguez CJ, Soliman EZ, Alonso A, et al. Atrial fibrillation incidence and risk factors in relation to race-ethnicity and the population attributable fraction of atrial fibrillation risk factors: the Multi-Ethnic Study of Atherosclerosis. Annals of epidemiology 2015;25(2):71-6, 76.e1.
131. Bapat A, Zhang Y, Post WS, et al. Relation of Physical Activity and Incident Atrial Fibrillation (from the Multi-Ethnic Study of Atherosclerosis). The American journal of cardiology 2015;116(6):883-8.
132. O'Neal WT, Soliman EZ, Qureshi W, et al. Sustained pre-hypertensive blood pressure and incident atrial fibrillation: the Multi-Ethnic Study of Atherosclerosis. Journal of the American Society of Hypertension : JASH 2015;9(3):191-6.
133. Sano F, Ohira T, Kitamura A, et al. Heavy alcohol consumption and risk of atrial fibrillation. The Circulatory Risk in Communities Study (CIRCS). Circulation journal : official journal of the Japanese Circulation Society 2014;78(4):955-61.
134. Nystrom PK, Carlsson AC, Leander K, et al. Obesity, metabolic syndrome and risk of atrial fibrillation: a Swedish, prospective cohort study. PloS one 2015;10(5):e0127111.
135. Grundvold I, Skretteberg PT, Liestol K, et al. Importance of physical fitness on predictive effect of body mass index and weight gain on incident atrial fibrillation in healthy middle-age men. The American journal of cardiology 2012;110(3):425-32.
136. Kokubo Y, Watanabe M, Higashiyama A, et al. Interaction of Blood Pressure and Body Mass Index With Risk of Incident Atrial Fibrillation in a Japanese Urban Cohort: The Suita Study. American journal of hypertension 2015;28(11):1355-61.
137. Dewland TA, Olgin JE, Vittinghoff E, et al. Incident atrial fibrillation among Asians, Hispanics, blacks, and whites. Circulation 2013;128(23):2470-7.
138. Lindhardsen J, Ahlehoff O, Gislason GH, et al. Risk of atrial fibrillation and stroke in rheumatoid arthritis: Danish nationwide cohort study. BMJ (Clinical research ed) 2012;344:e1257.
139. Ahlehoff O, Gislason GH, Jorgensen CH, et al. Psoriasis and risk of atrial fibrillation and ischaemic stroke: a Danish Nationwide Cohort Study. European heart journal 2012;33(16):2054-64.
140. Pallisgaard JL, Schjerning AM, Lindhardt TB, et al. Risk of atrial fibrillation in diabetes mellitus: A nationwide cohort study. European journal of preventive cardiology 2016;23(6):621-7.
141. Selmer C, Olesen JB, Hansen ML, et al. The spectrum of thyroid disease and risk of new onset atrial fibrillation: a large population cohort study. BMJ (Clinical research ed) 2012;345:e7895.
142. Emilsson L, Smith JG, West J, et al. Increased risk of atrial fibrillation in patients with coeliac disease: a nationwide cohort study. European heart journal 2011;32(19):2430-7.
143. Chiang CH, Huang CC, Chan WL, et al. Herpes simplex virus infection and risk of atrial fibrillation: a nationwide study. International journal of cardiology 2013;164(2):201-4.
144. Gorenek B, Pelliccia A, Benjamin EJ, et al. European Heart Rhythm Association (EHRA)/European Association of Cardiovascular Prevention and Rehabilitation (EACPR) position paper on how to prevent atrial fibrillation endorsed by the Heart Rhythm Society (HRS) and Asia Pacific Heart Rhythm Society (APHRS). Europace : European pacing, arrhythmias, and cardiac electrophysiology : journal of the working groups on cardiac pacing, arrhythmias, and cardiac cellular electrophysiology of the European Society of Cardiology 2017;19(2):190-225.

145. Schneider MP, Hua TA, Bohm M, et al. Prevention of atrial fibrillation by Renin-Angiotensin system inhibition a meta-analysis. Journal of the American College of Cardiology 2010;55(21):2299-307.
146. Emdin CA, Callender T, Cao J, et al. Effect of antihypertensive agents on risk of atrial fibrillation: a meta-analysis of large-scale randomized trials. Europace : European pacing, arrhythmias, and cardiac electrophysiology : journal of the working groups on cardiac pacing, arrhythmias, and cardiac cellular electrophysiology of the European Society of Cardiology 2015;17(5):701-10.
147. Rahimi K, Emberson J, McGale P, et al. Effect of statins on atrial fibrillation: collaborative meta-analysis of published and unpublished evidence from randomised controlled trials. BMJ (Clinical research ed) 2011;342:d1250.
148. Rapsomaniki E, Timmis A, George J, 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 (London, England) 2014;383(9932):1899-911.
149. Lokaj P, Parenica J, Goldbergova M, et al. Pulse Pressure in Clinical Practice. European Journal of Cardiovascular Medicine 2011;2(1):66-68.
150. National Institute for Health and Care Excellence. Hypertension in adults: diagnosis and management. Available at: http://www.nice.org.uk/guidance/cg127 [accessed 28 November 2017].
151. Danesh J, Erqou S, Walker M, et al. The Emerging Risk Factors Collaboration: analysis of individual data on lipid, inflammatory and other markers in over 1.1 million participants in 104 prospective studies of cardiovascular diseases. European journal of epidemiology 2007;22(12):839-69.
152. Heneghan C, Blacklock C, Perera R, et al. Evidence for non-communicable diseases: analysis of Cochrane reviews and randomised trials by World Bank classification. BMJ open 2013;3(7).
153. Chugh SS, Roth GA, Gillum RF, et al. Global burden of atrial fibrillation in developed and developing nations. Global heart 2014;9(1):113-9.
154. Montori VM, Wilczynski NL, Morgan D, et al. Optimal search strategies for retrieving systematic reviews from Medline: analytical survey. BMJ (Clinical research ed) 2005;330(7482):68.
155. Kwon Y, Powelson SE, Wong H, et al. An assessment of the efficacy of searching in biomedical databases beyond MEDLINE in identifying studies for a systematic review on ward closures as an infection control intervention to control outbreaks. Systematic reviews 2014;3:135.
156. Stevinson C, Lawlor DA. Searching multiple databases for systematic reviews: added value or diminishing returns? Complementary therapies in medicine 2004;12(4):228-32.
157. Lemeshow AR, Blum RE, Berlin JA, et al. Searching one or two databases was insufficient for meta-analysis of observational studies. Journal of clinical epidemiology 2005;58(9):867-73.
158. Horton R. The less acceptable face of bias. Lancet (London, England) 2000;356(9234):959-60.
159. Nelson CP, Hamby SE, Saleheen D, et al. Genetically determined height and coronary artery disease. The New England journal of medicine 2015;372(17):1608-18.
160. Office for National Statistics. Mortality Statistics in England and Wales. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/qmis/mortalitystatisticsinenglandandwalesqmi [accessed 28 November 2017].
161. Herrett E, Smeeth L, Walker L, et al. The Myocardial Ischaemia National Audit Project (MINAP). Heart (British Cardiac Society) 2010;96(16):1264-7.
162. 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).
163. Chisholm J. The Read clinical classification. BMJ (Clinical research ed) 1990;300(6732):1092.
164. Kendall M, Enright D. Provision of medicines information: the example of the British National Formulary. British journal of clinical pharmacology 2012;73(6):934-8.
165. NHS Digital. NHS Classifications OPCS-4. Available at: https://isd.digital.nhs.uk/trud3/user/guest/group/0/pack/10 [accessed 28 November 2017].
166. Flynn MR, Barrett C, Cosio FG, et al. The Cardiology Audit and Registration Data Standards (CARDS), European data standards for clinical cardiology practice. European heart journal 2005;26(3):308-13.

167. Lea NC, Nicholls J, Dobbs C, et al. Data Safe Havens and Trust: Toward a Common Understanding of Trusted Research Platforms for Governing Secure and Ethical Health Research. JMIR medical informatics 2016;4(2):e22.
168. Jones KH, Laurie G, Stevens L, et al. The other side of the coin: Harm due to the non-use of health-related data. International journal of medical informatics 2017;97:43-51.
169. Carter P, Laurie GT, Dixon-Woods M. The social licence for research: why care.data ran into trouble. Journal of medical ethics 2015;41(5):404-9.
170. Jinks C, Carter P, Rhodes C, et al. Patient and public involvement in primary care research - an example of ensuring its sustainability. Research involvement and engagement 2016;2:1.
171. University College London. SLMS Information Governance Framework. Available at: https://www.ucl.ac.uk/isd/itforslms/services/handling-sens-data/info-gov-framework/ [accessed 28 November 2017].
172. HM Government. Health and Social Care Act 2012. Available at: http://www.legislation.gov.uk/ukpga/2012/7/contents [accessed 28 November 2017].
173. The Medical Research Council. 2016 Max Perutz Science Writing Award shortlist. Available at: https://www.mrc.ac.uk/news/browse/2016-max-perutz-science-writing-award-shortlist-announced/ [accessed 28 November 2017].
174. Wood L, Martinez C. The general practice research database: role in pharmacovigilance. Drug safety 2004;27(12):871-81.
175. Gillam S, Steel N. The Quality and Outcomes Framework--where next? BMJ (Clinical research ed) 2013;346:f659.
176. Duncan ME, Pitcher A, Goldacre MJ. Atrial fibrillation as a cause of death increased steeply in England between 1995 and 2010. Europace : European pacing, arrhythmias, and cardiac electrophysiology : journal of the working groups on cardiac pacing, arrhythmias, and cardiac cellular electrophysiology of the European Society of Cardiology 2014;16(6):797-802.
177. NHS Digital. The processing cycle and HES data quality. Availabe at: http://content.digital.nhs.uk/article/1825/The-processing-cycle-and-HES-data-quality [accessed 28 November 2017].
178. Welch, CA. Implementation, evaluation and application of multiple imputation for missing data in longitudinal electronic health record research. Doctoral thesis , UCL (University College London). Available at: http://discovery.ucl.ac.uk/1464072/ [accessed 28 November 2017].
179. Welch CA, Petersen I, Bartlett JW, et al. Evaluation of two-fold fully conditional specification multiple imputation for longitudinal electronic health record data. Statistics in medicine 2014;33(21):3725-37.
180. Ruigomez A, Johansson S, Wallander MA, et al. Incidence of chronic atrial fibrillation in general practice and its treatment pattern. Journal of clinical epidemiology 2002;55(4):358-63.
181. George J, Rapsomaniki E, Pujades-Rodriguez M, et al. How Does Cardiovascular Disease First Present in Women and Men? Incidence of 12 Cardiovascular Diseases in a Contemporary Cohort of 1,937,360 People. Circulation 2015;132(14):1320-8.
182. George J, Mathur R, Shah AD, et al. Ethnicity and the first diagnosis of a wide range of cardiovascular diseases: Associations in a linked electronic health record cohort of 1 million patients. PloS one 2017;12(6):e0178945.
183. Pujades-Rodriguez M, Timmis A, Stogiannis D, 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;9(8):e104671.
184. Pujades-Rodriguez M, George J, Shah AD, et al. Heterogeneous associations between smoking and a wide range of initial presentations of cardiovascular disease in 1937360 people in England: lifetime risks and implications for risk prediction. International journal of epidemiology 2015;44(1):129-41.
185. Bell S, Daskalopoulou M, Rapsomaniki E, et al. Association between clinically recorded alcohol consumption and initial presentation of 12 cardiovascular diseases: population based cohort study using linked health records. BMJ (Clinical research ed) 2017;356:j909.
186. Moayyeri A, Riyaz P, Pujades-Rodriguez M, et al. LDL-cholesterol and the primary prevention of heart failure and atrial fibrillation: evidence from a new observational cohort of 550,000 people and meta-analysis of statin trials. Manuscript submitted for publication 2017.

187. Shah AD, Langenberg C, Rapsomaniki E, et al. Type 2 diabetes and incidence of cardiovascular diseases: a cohort study in 1.9 million people. The lancet Diabetes & endocrinology 2015;3(2):105-13.
188. Bhaskaran K, Forbes HJ, Douglas I, et al. Representativeness and optimal use of body mass index (BMI) in the UK Clinical Practice Research Datalink (CPRD). BMJ open 2013;3(9):e003389.
189. Pujades-Rodriguez M, Duyx B, Thomas SL, et al. Rheumatoid Arthritis and Incidence of Twelve Initial Presentations of Cardiovascular Disease: A Population Record-Linkage Cohort Study in England. PloS one 2016;11(3):e0151245.
190. Haukoos JS, Newgard CD. Advanced statistics: missing data in clinical research--part 1: an introduction and conceptual framework. Academic emergency medicine : official journal of the Society for Academic Emergency Medicine 2007;14(7):662-8.
191. Schnabel RB, Yin X, Gona P, et al. 50 year trends in atrial fibrillation prevalence, incidence, risk factors, and mortality in the Framingham Heart Study: a cohort study. Lancet (London, England) 2015;386(9989):154-62.
192. Lee S, Monz BU, Clemens A, et al. Representativeness of the dabigatran, apixaban and rivaroxaban clinical trial populations to real-world atrial fibrillation patients in the United Kingdom: a cross-sectional analysis using the General Practice Research Database. BMJ open 2012;2(6).
193. Lacoin L, Lumley M, Ridha E, et al. Evolving landscape of stroke prevention in atrial fibrillation within the UK between 2012 and 2016: a cross-sectional analysis study using CPRD. BMJ open 2017;7(9):e015363.
194. Chung SC, Gedeborg R, Nicholas O, et al. Acute myocardial infarction: a comparison of short-term survival in national outcome registries in Sweden and the UK. Lancet (London, England) 2014;383(9925):1305-12.
195. Shah AD, Thornley S, Chung SC, et al. White cell count in the normal range and short-term and long-term mortality: international comparisons of electronic health record cohorts in England and New Zealand. BMJ open 2017;7(2):e013100.
196. Wright FL, Green J, Canoy D, et al. Vascular disease in women: comparison of diagnoses in hospital episode statistics and general practice records in England. BMC medical research methodology 2012;12:161.
197. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS medicine 2015;12(3):e1001779.
198. Mathur R, Bhaskaran K, Chaturvedi N, et al. Completeness and usability of ethnicity data in UK-based primary care and hospital databases. Journal of public health (Oxford, England) 2014;36(4):684-92.
199. Nieuwlaat R, Capucci A, Camm AJ, et al. Atrial fibrillation management: a prospective survey in ESC member countries: the Euro Heart Survey on Atrial Fibrillation. European heart journal 2005;26(22):2422-34.
200. Kakkar AK, Mueller I, Bassand JP, et al. International longitudinal registry of patients with atrial fibrillation at risk of stroke: Global Anticoagulant Registry in the FIELD (GARFIELD). American heart journal 2012;163(1):13-19.e1.
201. Banerjee A, Taillandier S, Olesen JB, et al. Pattern of atrial fibrillation and risk of outcomes: the Loire Valley Atrial Fibrillation Project. International journal of cardiology 2013;167(6):2682-7.
202. Wei WQ, Teixeira PL, Mo H, et al. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. Journal of the American Medical Informatics Association : JAMIA 2016;23(e1):e20-7.
203. Christophersen IE, Rienstra M, Roselli C, et al. Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation. Nature genetics 2017;49(6):946-52.
204. Casey JA, Schwartz BS, Stewart WF, et al. Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. Annual review of public health 2016;37:61-81.
205. Munoz-Price LS, Frencken JF, Tarima S, et al. Handling Time-dependent Variables: Antibiotics and Antibiotic Resistance. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 2016;62(12):1558-63.
206. Teuwen CP, Ramdjan TT, Gotte M, et al. Time Course of Atrial Fibrillation in Patients With Congenital Heart Defects. Circulation Arrhythmia and electrophysiology 2015;8(5):1065-72.
207. Townsend N, Wilson L, Bhatnagar P, et al. Cardiovascular disease in Europe: epidemiological update 2016. European heart journal 2016;37(42):3232-45.

208. Cordoba G, Schwartz L, Woloshin S, et al. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. BMJ (Clinical research ed) 2010;341:c3920.
209. Koudstaal S, Pujades-Rodriguez M, Denaxas S, et al. Prognostic burden of heart failure recorded in primary care, acute hospital admissions, or both: a population-based linked electronic health record cohort study in 2.1 million people. European journal of heart failure 2017;19(9):1119-27.
210. Shah AD, Denaxas S, Nicholas O, et al. Neutrophil Counts and Initial Presentation of 12 Cardiovascular Diseases: A CALIBER Cohort Study. Journal of the American College of Cardiology 2017;69(9):1160-69.
211. Shah AD, Denaxas S, Nicholas O, et al. Low eosinophil and low lymphocyte counts and the incidence of 12 cardiovascular diseases: a CALIBER cohort study. Open heart 2016;3(2):e000477.
212. Daskalopoulou M, George J, Walters K, et al. Depression as a Risk Factor for the Initial Presentation of Twelve Cardiac, Cerebrovascular, and Peripheral Arterial Diseases: Data Linkage Study of 1.9 Million Women and Men. PloS one 2016;11(4):e0153838.
213. Herrett E, Shah AD, Boggon R, 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 (Clinical research ed) 2013;346:f2350.
214. Rapsomaniki E, Shah A, Perel P, et al. Prognostic models for stable coronary artery disease based on electronic health record cohort of 102 023 patients. European heart journal 2014;35(13):844-52.
215. Lewis JD, Bilker WB, Weinstein RB, et al. The relationship between time since registration and measured incidence rates in the General Practice Research Database. Pharmacoepidemiology and drug safety 2005;14(7):443-51.
216. Norby FL, Soliman EZ, Chen LY, et al. Trajectories of Cardiovascular Risk Factors and Incidence of Atrial Fibrillation Over a 25-Year Follow-Up: The ARIC Study (Atherosclerosis Risk in Communities). Circulation 2016;134(8):599-610.
217. Rahman F, Yin X, Larson MG, et al. Trajectories of Risk Factors and Risk of New-Onset Atrial Fibrillation in the Framingham Heart Study. Hypertension (Dallas, Tex : 1979) 2016;68(3):597-605.
218. Hoekstra T, Twisk JWR. The Analysis of Individual Health Trajectories Across the Life Course: Latent Class Growth Models Versus Mixed Models. In: Burton-Jeangros C, Cullati S, Sacker A, et al., eds. A Life Course Perspective on Health Trajectories and Transitions. Cham (CH): Springer
Copyright 2015, The Author(s). 2015:179-95.
219. Fabritz L, Guasch E, Antoniades C, et al. Expert consensus document: Defining the major health modifiers causing atrial fibrillation: a roadmap to underpin personalized prevention and treatment. Nature reviews Cardiology 2016;13(4):230-7.
220. Murphy A, Banerjee A, Breithardt G, et al. The Word Heart Federation Roadmap for Nonvalvular Atrial Fibrillation. Global heart 2017.
221. Curtis MJ, Hancox JC, Farkas A, et al. The Lambeth Conventions (II): guidelines for the study of animal and human ventricular and supraventricular arrhythmias. Pharmacology & therapeutics 2013;139(2):213-48.
222. Schotten U, Verheule S, Kirchhof P, et al. Pathophysiological mechanisms of atrial fibrillation: a translational appraisal. Physiological reviews 2011;91(1):265-325.
223. Nicholson A, Ford E, Davies KA, et al. Optimising use of electronic health records to describe the presentation of rheumatoid arthritis in primary care: a strategy for developing code lists. PloS one 2013;8(2):e54878.
224. Darby AE, Dimarco JP. Management of atrial fibrillation in patients with structural heart disease. Circulation 2012;125(7):945-57.
225. Bouchardy J, Therrien J, Pilote L, et al. Atrial arrhythmias in adults with congenital heart disease. Circulation 2009;120(17):1679-86.
226. Manuguerra R, Callegari S, Corradi D. Inherited Structural Heart Diseases With Potential Atrial Fibrillation Occurrence. Journal of cardiovascular electrophysiology 2016;27(2):242-52.
227. Arsenault KA, Yusuf AM, Crystal E, et al. Interventions for preventing post-operative atrial fibrillation in patients undergoing heart surgery. The Cochrane database of systematic reviews 2013(1):Cd003611.
228. Molteni M, Polo Friz H, Primitz L, et al. The definition of valvular and non-valvular atrial fibrillation: results of a physicians' survey. Europace : European pacing, arrhythmias, and cardiac electrophysiology : journal of the working groups on cardiac pacing,

arrhythmias, and cardiac cellular electrophysiology of the European Society of Cardiology 2014;16(12):1720-5.
229. Hammad TA, Margulis AV, Ding Y, et al. Determining the predictive value of Read codes to identify congenital cardiac malformations in the UK Clinical Practice Research Datalink. Pharmacoepidemiology and drug safety 2013;22(11):1233-8.
230. Pujades-Rodriguez M, Guttmann O, Gonzalez-Izquierdo A, et al. Identifying unmet clinical need in hypertrophic cardiomyopathy using national electronic health records. Manuscript submitted for publication 2017.
231. Jordan KP, Timmis A, Croft P, et al. Prognosis of undiagnosed chest pain: linked electronic health record cohort study. BMJ (Clinical research ed) 2017;357:j1194.
232. Wallander MA, Johansson S, Ruigomez A, et al. Morbidity associated with sleep disorders in primary care: a longitudinal cohort study. Primary care companion to the Journal of clinical psychiatry 2007;9(5):338-45.
233. Gundlund A, Olesen JB, Staerk L, et al. Outcomes Associated With Familial Versus Nonfamilial Atrial Fibrillation: A Matched Nationwide Cohort Study. Journal of the American Heart Association 2016;5(11).
234. Rothnie KJ, Smeeth L, Herrett E, et al. Closing the mortality gap after a myocardial infarction in people with and without chronic obstructive pulmonary disease. Heart (British Cardiac Society) 2015;101(14):1103-10.
235. Hoffmann TJ, Ehret GB, Nandakumar P, et al. Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. Nature genetics 2017;49(1):54-64.
236. Melby SJ, George JF, Picone DJ, et al. A time-related parametric risk factor analysis for postoperative atrial fibrillation after heart surgery. The Journal of thoracic and cardiovascular surgery 2015;149(3):886-92.
237. Martins RP, Galand V, Colette E, et al. Defining nonvalvular atrial fibrillation: A quest for clarification. American heart journal 2016;178:161-7.
238. Camm AJ, Lip GY, De Caterina R, et al. 2012 focused update of the ESC Guidelines for the management of atrial fibrillation: an update of the 2010 ESC Guidelines for the management of atrial fibrillation. Developed with the special contribution of the European Heart Rhythm Association. European heart journal 2012;33(21):2719-47.
239. Fuster V, Ryden LE, Cannom DS, et al. ACC/AHA/ESC 2006 Guidelines for the Management of Patients with Atrial Fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the European Society of Cardiology Committee for Practice Guidelines (Writing Committee to Revise the 2001 Guidelines for the Management of Patients With Atrial Fibrillation): developed in collaboration with the European Heart Rhythm Association and the Heart Rhythm Society. Circulation 2006;114(7):e257-354.
240. Bloomfield P. Choice of heart valve prosthesis. Heart (British Cardiac Society) 2002;87(6):583-9.
241. Cleveland WS, Devlin SJ, Grosse E. Regression by local fitting: Methods, properties, and computational algorithms. Journal of Econometrics 1988;37(1):87-114.
242. Breithardt G, Baumgartner H, Berkowitz SD, et al. Native valve disease in patients with non-valvular atrial fibrillation on warfarin or rivaroxaban. Heart (British Cardiac Society) 2016;102(13):1036-43.
243. Iung B, Vahanian A. Epidemiology of acquired valvular heart disease. The Canadian journal of cardiology 2014;30(9):962-70.
244. Iung B, Baron G, Butchart EG, et al. A prospective survey of patients with valvular heart disease in Europe: The Euro Heart Survey on Valvular Heart Disease. European heart journal 2003;24(13):1231-43.
245. Nkomo VT, Gardin JM, Skelton TN, et al. Burden of valvular heart diseases: a population-based study. Lancet (London, England) 2006;368(9540):1005-11.
246. Potpara TS, Lip GY, Larsen TB, et al. Stroke prevention strategies in patients with atrial fibrillation and heart valve abnormalities: perceptions of 'valvular' atrial fibrillation: results of the European Heart Rhythm Association Survey. Europace : European pacing, arrhythmias, and cardiac electrophysiology : journal of the working groups on cardiac pacing, arrhythmias, and cardiac cellular electrophysiology of the European Society of Cardiology 2016;18(10):1593-98.
247. Lip GYH, Collet JP, Caterina R, et al. Antithrombotic therapy in atrial fibrillation associated with valvular heart disease: a joint consensus document from the European Heart Rhythm Association (EHRA) and European Society of Cardiology Working Group on Thrombosis, endorsed by the ESC Working Group on Valvular Heart Disease, Cardiac Arrhythmia Society of Southern Africa (CASSA), Heart Rhythm Society (HRS), Asia

Pacific Heart Rhythm Society (APHRS), South African Heart (SA Heart) Association and Sociedad Latinoamericana de Estimulacion Cardiaca y Electrofisiologia (SOLEACE). Europace : European pacing, arrhythmias, and cardiac electrophysiology : journal of the working groups on cardiac pacing, arrhythmias, and cardiac cellular electrophysiology of the European Society of Cardiology 2017;19(11):1757-58.
248. Kumar RK, Tandon R. Rheumatic fever & rheumatic heart disease: the last 50 years. The Indian journal of medical research 2013;137(4):643-58.
249. Baumgartner H, Falk V, Bax JJ, et al. 2017 ESC/EACTS Guidelines for the management of valvular heart disease. European heart journal 2017;38(36):2739-91.
250. Forslund T, Wettermark B, Wandell P, et al. Risk scoring and thromboprophylactic treatment of patients with atrial fibrillation with and without access to primary healthcare data: experience from the Stockholm health care system. International journal of cardiology 2013;170(2):208-14.
251. Gallagher AM, van Staa TP, Murray-Thomas T, et al. Population-based cohort study of warfarin-treated patients with atrial fibrillation: incidence of cardiovascular and bleeding outcomes. BMJ open 2014;4(1):e003839.
252. Go AS, Mozaffarian D, Roger VL, et al. Heart disease and stroke statistics--2013 update: a report from the American Heart Association. Circulation 2013;127(1):e6-e245.
253. Friberg L, Skeppholm M, Terent A. Benefit of anticoagulation unlikely in patients with atrial fibrillation and a CHA2DS2-VASc score of 1. Journal of the American College of Cardiology 2015;65(3):225-32.
254. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate behavioral research 2011;46(3):399-424.
255. Singer DE, Chang Y, Fang MC, et al. The net clinical benefit of warfarin anticoagulation in atrial fibrillation. Annals of internal medicine 2009;151(5):297-305.
256. Huang D, Anguo L, Yue WS, et al. Refinement of ischemic stroke risk in patients with atrial fibrillation and CHA2 DS2 -VASc score of 1. Pacing and clinical electrophysiology : PACE 2014;37(11):1442-7.
257. Chao TF, Liu CJ, Wang KL, et al. Should atrial fibrillation patients with 1 additional risk factor of the CHA2DS2-VASc score (beyond sex) receive oral anticoagulation? Journal of the American College of Cardiology 2015;65(7):635-42.
258. Olesen JB, Lip GY, Hansen ML, et al. Validation of risk stratification schemes for predicting stroke and thromboembolism in patients with atrial fibrillation: nationwide cohort study. BMJ (Clinical research ed) 2011;342:d124.
259. Olesen JB, Torp-Pedersen C, Hansen ML, et al. The value of the CHA2DS2-VASc score for refining stroke risk stratification in patients with atrial fibrillation with a CHADS2 score 0-1: a nationwide cohort study. Thrombosis and haemostasis 2012;107(6):1172-9.
260. Olesen JB, Lip GY, Lindhardsen J, et al. Risks of thromboembolism and bleeding with thromboprophylaxis in patients with atrial fibrillation: A net clinical benefit analysis using a 'real world' nationwide cohort study. Thrombosis and haemostasis 2011;106(4):739-49.
261. Lip GY, Skjoth F, Rasmussen LH, et al. Oral anticoagulation, aspirin, or no therapy in patients with nonvalvular AF with 0 or 1 stroke risk factor based on the CHA2DS2-VASc score. Journal of the American College of Cardiology 2015;65(14):1385-94.
262. Friberg L, Rosenqvist M, Lip GY. Evaluation of risk stratification schemes for ischaemic stroke and bleeding in 182 678 patients with atrial fibrillation: the Swedish Atrial Fibrillation cohort study. European heart journal 2012;33(12):1500-10.
263. Guo Y, Apostolakis S, Blann AD, et al. Validation of contemporary stroke and bleeding risk stratification scores in non-anticoagulated Chinese patients with atrial fibrillation. International journal of cardiology 2013;168(2):904-9.
264. Forslund T, Wettermark B, Wandell P, et al. Risks for stroke and bleeding with warfarin or aspirin treatment in patients with atrial fibrillation at different CHA(2)DS(2)VASc scores: experience from the Stockholm region. European journal of clinical pharmacology 2014;70(12):1477-85.
265. Friberg L, Rosenqvist M, Lip GY. Net clinical benefit of warfarin in patients with atrial fibrillation: a report from the Swedish atrial fibrillation cohort study. Circulation 2012;125(19):2298-307.
266. Chackery DG, Keshavjee K, Mirza K, et al. Integrating Clinical Decision Support into EMR and PHR: a Case Study Using Anticoagulation. Studies in health technology and informatics 2015;208:98-103.
267. Watts G. Will UK Biobank pay off? BMJ (Clinical research ed) 2006;332(7549):1052.

268. Hong KN, Fuster V, Rosenson RS, et al. How Low to Go With Glucose, Cholesterol, and Blood Pressure in Primary Prevention of CVD. Journal of the American College of Cardiology 2017;70(17):2171-85.
269. Deo RC. Machine Learning in Medicine. Circulation 2015;132(20):1920-30.
270. Burgess S, Timpson NJ, Ebrahim S, et al. Mendelian randomization: where are we now and where are we going? International journal of epidemiology 2015;44(2):379-88.
271. Li X, Liu H, Du X, et al. Integrated Machine Learning Approaches for Predicting Ischemic Stroke and Thromboembolism in Atrial Fibrillation. AMIA Annual Symposium proceedings AMIA Symposium 2016;2016:799-807.
272. Booth HP, Prevost AT, Gulliford MC. Validity of smoking prevalence estimates from primary care electronic health records compared with national population survey data for England, 2007 to 2011. Pharmacoepidemiology and drug safety 2013;22(12):1357-61.
273. Emdin CA, Anderson SG, Salimi-Khorshidi G, et al. Usual blood pressure, atrial fibrillation and vascular risk: evidence from 4.3 million adults. International journal of epidemiology 2017;46(1):162-72.