Preventing respiratory infection-related cardiovascular disease events in primary care

Thesis submitted for the degree of Doctor of Philosophy by

Joseph Jonathan Lee

Nuffield Department of Primary Care Health Sciences

2025

Supervised by Professor James Sheppard, Professor FD Richard Hobbs, Dr Constantinos Koshiaris, Dr Susannah Fleming, and Dr Nicholas R Jones

Acknowledgements

To my supervisors, family, and colleagues: thank you for your forbearance, advice, and support.

Funding and disclaimer

I am also grateful to the NIHR who funded this project with both a doctoral research fellowship (NIHR300738) and further support via the ARC OTV.

The funders wish it to be known that the views expressed in this thesis are those of the author and not necessarily those of the NIHR or the Department of Health and Social Care.

Contents

Tables

Figures

Thesis abstract

Background: Cardiovascular disease events (CVD events, comprising coronary and cerebrovascular events) are major causes of morbidity and mortality. CVD can be prevented by medications that target the underlying pathological processes of thrombosis and atherosclerosis. When a patient is diagnosed with a respiratory infection their risk of CVD events is about four times higher than their background risk for the following four weeks. This infection-related CVD event risk is well characterised by epidemiological research, but clinical practice guidelines for primary care do not address it. Prior to this thesis there were no tools for predicting an individual's risk of an infection-related CVD event and, apart from vaccines, no established interventions for this scenario.

Overall aim: To investigate strategies for preventing infection-related CVD events in primary care.

Approach:

1. Developing statistical models to identify patients with respiratory infections who are at risk of CVD events
2. Validating the prediction models, using them to derive a clinical risk prediction score
3. Estimating the effect of aspirin on infection-related cardiovascular events
4. Estimating the effect of statin use on infection-related cardiovascular events

Methods: Four epidemiological studies using large cohorts from coded UK primary care records held by the Clinical Practice Research Datalink (CPRD). These data were linked to datasets of NHS hospital and Office of National Statistics (ONS) mortality and relative deprivation datasets. The first two studies used prediction modelling methods, and the next two used propensity modelling methods with logistic regression to estimate causal effects.

Results: I developed two statistical models and derived a clinical prediction points-based tool, the DASHI score. DASHI comprises five clinical variables: Diabetes, Age, Smoking status, Heart failure and Infection diagnosis. External validation showed DASHI can predict risk of infection-related CVD events with good calibration and discrimination (both C statistic and observed to expected ratios were 0.85 with IQR 0.85-0.85). This performance was very similar to the regression models. Aspirin and statins were estimated to increase infection-related CVD events; Relative Risk 2.52 (95% CI 2.26 to 2.81) for aspirin and 3.17 (95% CI 2.41 to 4.08) for statins. Aspirin increased bleeding with a relative risk of 1.31 (95% CI 1.06 to 1.16).

Conclusion: The DASHI score can predict risk of primary infection-related CVD events. The absolute risks are low for most people due to the short prediction period. It is unlikely that aspirin and statins increase CVD events given what we know about their effects in other settings. It is more likely the results are inaccurate because of confounding or coding problems in the datasets. In particular, prescriptions are recorded immediately in the clinical record, but there are delays before CVD events enter the datasets. This timing difference may have led to biases exacerbated by the short follow-up period. A definitive answer is likely to require a different approach, and may require different datasets, or a randomised controlled clinical trial.

1 Chapter One: Introduction and background

"All I am dreaming about now which seems to me so impossible and unearthly is really quite an ordinary thing," thought Ryabovitch, looking at the clouds of dust racing after the general's carriage. "It's all very ordinary, and everyone goes through it. . . . That general, for instance, has once been in love; now he is married and has children. Captain Vahter, too, is married and beloved, though the nape of his neck is very red and ugly and he has no waist. . . . Salrnanov is coarse and very Tatar, but he has had a love affair that has ended in marriage. . . . I am the same as everyone else, and I, too, shall have the same experience as everyone else, sooner or later. . . ."

And the thought that he was an ordinary person, and that his life was ordinary, delighted him and gave him courage.

Anton Chekhov, The Kiss, 15 December 1887

1.1 Motivating clinical case

The subject of this thesis is cardiovascular disease (CVD) events that occur shortly after respiratory tract infections (RTI). This is an important pattern for CVD events, but there are no specific practices in primary care guidelines dedicated to it. An illustrative vignette (with fictional patient details) describes the clinical scenario that motivates this work:

A dull December day in 2019. Mr Jones, a good-humoured man in late middle-age, wakes with a cough and fever. He aches. His wife wants to call the GP. He experiences a vague concern that something might be different about this illness, and anyway, he thinks, he wants to get better before Christmas. After all, he rarely bothers the doctor. He is cheered by this thought and, by protesting more feebly than usual, allows his wife to call the doctor.

Later, the GP, after listening to his history and his chest, offers sympathy and advice – this is a viral illness, probably, and antibiotics are not required. Take it easy, take paracetamol and all will be well, it might take ten days, maybe less, come back if things get worse.

Two days later things are indeed worse, and our patient is in an ambulance on his way to hospital. His chest pain is better having just received some treatment, and he apologises to the paramedic riding beside him - he has never had any problems before so is probably wasting their time.

Within half an hour, he will be told he is having a heart attack. If he survives, he will not survive unchanged. He may avoid heart failure and depression. But he will take half a dozen medications each day, and whilst he will still consider himself healthy, he will find thinking of his good health is somehow less reassuring.

This scenario plays out many times each year because acute respiratory infections are both common and associated with increased risk of cardiovascular events in the following weeks. [1,2]. In the UK respiratory tract infection is the commonest cause of illness, and CVD is the commonest cause of death and adult disability.[3,4] Although the risk associated with each individual infection is low, in aggregate infections may be an important driver of CVD.[5,6] Clinically, the two overarching questions I set out to explore in this thesis were - was it possible for the GP to tell if Mr Jones was at risk of CVD? and could the GP have prescribed something to change the outcome?

1.1.1. Introduction

In this chapter I define and describe infection-related CVD events for the purposes of this thesis, and give an overview of current practice in primary care for respiratory infections and CVD. The theme, and central hypothesis of this thesis, is that respiratory infections are a potential opportunity to reduce CVD events. It is not necessary for the association between infections and CVD to be causative, but to make use of this opportunity one would have to be able to identify people at risk of CVD events following their respiratory infections, and then to intervene with prescriptions or other prevention strategies to reduce their risk.

Chapter two is an overview of the methods I have used in my attempts to obtain evidence towards this end using routinely collected data. In chapters three to six I have described the research work. As there was no clinical prediction tool or statistical model for predicting respiratory infection-related CVD events before this thesis, I attempted to fill this gap by developing statistical prediction models in chapter three.[7] Chapter four details my external validation of these models, and how I developed and externally validated a clinical points score, the DASHI score (which stands for the clinical variables involved - Diabetes, Age, Smoking, Heart Failure, Infection type).

In chapters five and six I used the DASHI score to identify people who could be eligible for interventions that might reduce their risk of CVD events after infections and included them in the studies.

As there is little evidence of acute treatments for infection-related CVD events, I examined the effect of two medications that have established uses in CVD prevention.

In chapter five I examined the effects of aspirin, which is known to reduce primary and secondary CVD events by reducing thrombosis.[8,9] Because aspirin causes bleeding, its use is limited to situations where the risk of CVD events is higher than the risk of bleeding.[10] It seemed possible that the peak of risk around respiratory infections could be a period where the benefits could outweigh the risk. I examined the effect with logistic regression combined with propensity modelling methods, in a high-risk population identified with the DASHI score.

The last experimental chapter, chapter six, describes my measurement of the effects of statins on infection-related CVD. Statins are another well-established long-term medication for reducing CVD risk, which target atherosclerosis.[9–11] Statins are much safer than aspirin, so can be used in lower risk populations.[10] I used DASHI to identify a medium-risk population for this study, and explored different propensity modelling weighting methods to ameliorate confounding.

1.2 Definitions

1.2.1 Definition: Cardiovascular disease

In this thesis I used 'cardiovascular disease' (CVD) to refer to stroke, transient ischaemic attack (TIA), and symptomatic coronary heart disease (CHD).[12] For CHD I included angina pectoris (hereafter angina), acute coronary syndromes (ACS) and myocardial infarction (MI) as well as ischaemic cardiomyopathy.

1.2.2 Definition: Respiratory infections

I used respiratory tract infection (RTI) to refer to symptomatic acute respiratory tract infections thought to be of viral or bacterial aetiology. These are divided into upper respiratory tract infections (URTIs) which affect the respiratory system above the lungs, and lower respiratory tract infections (LRTI) of the lungs.

Chronic respiratory tract infections, for example tuberculosis, are not included in this definition. Pneumonia is defined by consolidation in one or more parts of the lungs, the commonest cause is LRTI (non-infective causes are outside the scope of this thesis).[13]

1.2.3 Definition: Infection-related CVD

For this thesis I defined infection-related CVD as cardiovascular events up to 28 days after diagnosis of respiratory infection. Four weeks is the period in which studies have consistently found patients have a higher risk of CVD events.[1,2] The duration of increased risk may persist for up to six months, but the relationship between infections and CVD events is most consistent and stronger at shorter time frames.

1.3 Infection – related cardiovascular disease

1.3.1 Epidemiology

Acute respiratory infections increase the short-term risk of MI and stroke in primary care patients.[14] Acute respiratory infection approximately quadruples the background risk of an acute cardiovascular event occurring in the following four weeks and the risk is highest in the first few days.[1,2,15–18] Most of the evidence comes from self-controlled case series which cannot estimate the absolute risk as they include only people with events. Never-the-less, this temporal relationship between cardiovascular disease events and infections is long established, and has been shown repeatedly in populations across the world, and with different infections including influenza and pneumonia.[1,2,15,18–25] Ecological studies using time series of laboratory confirmed viral infections and CVD events calculated respiratory infections were responsible for over ten thousand cardiovascular deaths per year in the UK.[26,27]

There is also evidence from a study using routinely collected UK data that compared to people with lower risk, people with ten-year risk of CVD over 10% have a higher incidence of respiratory infections, and pneumonia in particular, but are lower-risk for influenza-like illness (IRR 1.39, 95% CI 1.37 to 1.40 for any respiratory infection, 2.32, 95% CI 2.25 to 2.40 for pneumonia and 0.88, 95% CI 0.86 to 0.90 for influenza-like illness).[28] In this study they also had a higher risk of infection-related CVD events over the 30 days following respiratory infection (HR 3.65, 95% CI 3.42 to 3.89). The associated risk of CVD events was higher among people who were not prescribed antihypertensives, statins and antiplatelets (HR 4.04, 95% CI 3.76 to 4.34 for antihypertensives; HR 3.93, 95% CI 3.67 to 4.21 for statins and 3.68, 95% CI 3.45 to 3.93 for antiplatelets).[28]

Post-infection CVD events are not restricted to a particular infectious organism. Significant epidemiological evidence relates to undifferentiated acute RTIs in primary care, and there is

evidence of a variety of specific organisms having this relationship with CVD events.[2] COVID-19 also increases short-term CVD event risk (Incidence rate ratio 8.44, 95% CI 5.45 to 13.08 for days 0-7 in a Swedish study of February to September 2020), but most of the existing evidence predates this pandemic, as does the conception of this thesis and the data used throughout.[16] Influenza and pneumococcal infections are both associated with CVD events, but many other organisms are also implicated. For example incidence rate ratios for MI within seven days of laboratory confirmed infections were 10.11 (95% CI, 4.37 to 23.38) for influenza A, 5.17 (95% CI, 3.02 to 8.84) for influenza B, 3.51 (95% CI, 1.11 to 11.12) for RSV and 2.77 (95% CI, 1.23 to 6.24) for other respiratory viruses (a combination of adenovirus, coronavirus, enterovirus including rhinovirus, parainfluenza virus, human metapneumovirus, and bocavirus) in a Canadian study.[17]

There is also evidence that preventing infection mitigates CVD event risk, including trials that show influenza vaccine reduces CVD events, and observational evidence of protection from pneumococcal vaccines.[28–31]

1.3.2 Pathology

The temporal association between infection and CVD events raises the question as to why this should be – could infections trigger CVD events? The statistical relationship does not mean there is a causative relationship, but there are plausible pathological pathways from infection to CVD events.

The main pathological processes involved in MI and stroke are atherosclerosis and thrombosis. Atherosclerosis is the underlying cause of most myocardial infarctions and strokes, and thrombosis is a more immediate cause.[13] Infections may affect both atherosclerosis and thrombosis via inflammation or functional pathophysiological changes.

1.3.2.1 Pre-existing atherosclerosis

Atherosclerosis is common and increases with age.[32] About 40% of adults have asymptomatic coronary atheroma.[32] Patients with extant coronary disease could be compromised by respiratory infections. Atherosclerotic plaques are usually asymptomatic (and can be asymptomatic even when they rupture).[32] However, they can cause symptoms by narrowing the artery with their bulk, by dynamically narrowing the vessel due to inappropriate vasoconstriction, by leading to dissection of the artery, or by rupturing, which can lead to thrombosis which can occlude the vessel at that site or cause emboli.[33,34]

In the joint European Society for Cardiology/American Heart Association (ESC/AHA) definition of MI, type one MI is due to coronary thrombus, whereas type two is due to inadequate oxygen supply.[35]

Patients with stenosed coronary arteries are at risk of angina and type two MI if the cardiac demands outstrip the available supply.[36] Respiratory infections can increase cardiac demand, and cause hypoxia, vasospasm, and reductions in perfusion.[35]

Vessels with existing atheroma may exhibit maladaptive vasospasm in response to stress.[13] Vasospastic or variant angina, angina brought on by spasm of otherwise patent vessels, is responsible for 2% of hospital admissions with angina.[37] Mostly this is low risk, but it can cause arrythmias and MI.[38] There are case reports of coronary spasm in respiratory infection leading to cardiac arrest.[39]

The functional changes seen in respiratory infections could also cause plaque rupture, leading to thrombotic events. There is evidence that atheroma becomes unstable when cells in plaques become necrotic.[40] Necrosis can be triggered by inflammation or hypoxia, both of which can occur in respiratory infections.[34,41]

It is likely that the severity of the respiratory infection is important – the symptoms of infection are largely due to inflammation, which is an important driver of CVD. In a Texan study of hospitalised patients those with pneumonia had an estimated incidence rate ratio (IRR) of 47.6 (95% CI 24.5 to 92.5) for ACS in the 15 days after diagnosis, much higher than is seen in undifferentiated outpatient infections.[42] More severe infections with more inflammation also have worse derangements of physiology – for example, hypoxia is not likely with a cold, but more common in infective exacerbations of Chronic Obstructive Pulmonary Disease (COPD) and pneumonia.[43,44]

1.3.2.2 Accelerated atherosclerosis

Infection might also accelerate atherosclerosis. Atheroma development is a chronic inflammatory process, but lesions form quickly at times, and more slowly at other times.[33,34] Measuring atheroma in humans usually requires invasive methods and/or ionising radiation, which limits how often the vessels can be investigated. Nevertheless, there is evidence of atherosclerosis progressing over as little as a few months in native coronary vessels.[45] Accelerated atherosclerosis was first identified in patients with coronary artery bypass grafts.[46] There are two types of accelerated atherosclerosis in grafts, both can cause vessels to be completely occluded within a year. [46] Type one shows lesions dotted about the vessel like the pattern in 'normal' atheroma. [46] Type two has global narrowing of the entire graft, thought to result from endothelial damage leading to thrombosis, as the deeper layers of the vessel are normal.[46] It is plausible that infections could worsen or accelerate the progression of atheroma in native vessels. Worsening of classical risk factors for atheroma, such as increased lipids, glucose and blood pressure can occur in respiratory infections (and are seen in response to iatrogenic inflammation with interferon therapy).[47] In addition, inflammatory markers including C reactive protein (CRP) are associated with rapid progression of atheroma in people having multiple coronary angiography sessions, particularly if there is a rise

immediately after the first procedure.[45] Elevated erythrocyte sedimentation rate (ESR) is associated with rapidly progressing carotid atheroma in people with rheumatoid arthritis (RA).[45] Inflammation from infections could conceivably have a similar effect.

Infection can cause atheroma in animal models. A long list of organisms can do this to APO-E deficient mice (Apolipoprotein E is a protein involved in the transport of cholesterol; mice without this are a model for atheroma).[48] There is laboratory evidence from mice that specific organisms may have particular pathways that cause atheroma, for example *Streptococcus pneumoniae* (pneumococcus) shares molecular features with oxidised low-density lipoprotein (LDL) cholesterol, and vaccinating mice against pneumococcus prevents the progression of atherosclerosis.[49] Similarly, infecting model mice with influenza results in infiltration of arterial plaques by influenza-specific immune cells.[50] However, as other organisms can also cause atheroma there must be more general mechanisms at work, and there is also laboratory evidence of this. Macrophages dominate the chronic inflammatory processes in atheroma and can be activated by infections.[51] Immunisations targeting the innate immune system, and immunization with foam cells both reduced atheroma in model mice.[51,52]

1.3.2.3 Thrombosis

The most common cause of myocardial infarction related to atheroma is thrombosis.[13] Thrombus can stenose or occlude a vessel and can break apart and cause thromboembolic events downstream.[53] Antiplatelets are an important class of therapy for preventing CVD events.[9]

Thrombosis can be triggered by abnormalities in the elements of Virchow's triad – vessel walls, blood flow, and the constituents of the blood.[54] Of these the most relevant to infection-related CVD events may be loss of endothelium by plaque rupture, and several pro-coagulant effects including increasing viscosity, platelet count, inflammation, and dehydration.[55] When activated platelets tend to aggregate, leading to clotting or thrombosis.[13] Infections increase

platelet activation, which has been suggested as a cause of infection-related CVD.[56,57] Platelets are activated by systemic inflammation, and both bacterial and viral infections.[56] Platelets also have less appreciated immune functions and links with the respiratory system; half of platelets originate in the lungs, they also accumulate there in response to infection and are involved in clearing infections. [56]

1.3.2.4 Arrhythmias

Arrythmias are a cause of MI, stroke and cardiac death. Arrhythmias are more likely to occur when the myocardium is in a pro-arrhythmic state –more susceptible to degeneration of the electrical activity. Infections are pro-arrhythmic via hypoxia, ischaemia, and direct inflammatory effects on the muscle - myocarditis and pericarditis are pro-arrhythmic inflammatory responses to infections.[58]

Tachyarrhythmias are a risk for myocardial ischaemia – a pathologically rapid heart has lower output and higher demand which can lead to angina or MI.[59]

Pneumonia is a cause of Atrial Fibrillation (AF).[60] AF is a major cause of TIA and ischaemic stroke.[61] It is likely that some infection-related strokes come from this sequence of events.

1.3.2.5 Pathology overview

CVD is mostly caused by atherosclerosis and thrombosis, and there are reasons to think infections could influence these processes. Inflammation can worsen atherosclerosis, and pre-existing coronary disease is vulnerable to type one and two MI during the kinds of physiological disturbance that occur during infections. Infections are prothrombotic, and prothrombotic states increase the risk of MI and stroke. Infections area also pro-arrhythmic, which can lead to cardiac events and stroke. There are plausible biological mechanisms that could link respiratory infections with CVD, but it is possible some other unknown mechanism is responsible for the temporal association between infections and CVD. It is also possible that the association is due to some other, non-causal confounders.

1.4 Current practice in UK primary care – infection-related CVD

Current UK primary care guidance for respiratory infections does not acknowledge infection-related CVD, or make any suggestions for CVD risk stratification or mitigation during acute respiratory infections.[62,63] There is a preventative action in secondary prevention guidance - patients with a history of CVD events are recommended to receive influenza and pneumococcal vaccines.[64,65] However the UK vaccination guidance and information for patients describe this in terms of preventing severe influenza rather than CVD event prevention.[64] In contrast the ESC guidance for prevention of CVD recognises the importance of infection-related CVD events and recommends influenza vaccination on this basis.[19]

Infections and CVD are dealt with as separate entities in UK primary care, so I have briefly reviewed current practice in these two areas. The overall aim of this thesis – exploring how to intervene to prevent CVD events in acute respiratory infections – may be more successful if the CVD event outcomes have similar underlying pathology, likely to be both caused by the same processes and amenable to the same potential interventions. This also aligns with the previous epidemiological evidence which focussed on myocardial infarction and stroke, which share some causes and preventative strategies.[1,18,66]

1.5 Current practice in UK primary care – respiratory infections

1.5.1 Epidemiology of respiratory infections

Acute respiratory infections are the commonest illnesses throughout people's lives, with a mean of more than one per year even in the over sixties, who were found to have the lowest rates.[4] Young children have the highest incidence (with a mean over eight per year), but respiratory infections remain the commonest cause of ill health at all ages.[4]

In UK primary care, and consequently the scope of this thesis, only symptomatic respiratory infections are detected, coded and included in the datasets. Respiratory infections are driven by infection dynamics, so vary greatly over the year. Pre-Covid-19 the number of

presentations with respiratory infections for a typical 10,000 patient practice varied from 12 to 89 per week.[67] The largest contributor to the variation was influenza, but many other agents contribute, including respiratory syncytial virus (RSV), parainfluenza virus, rhinovirus, human metapneumovirus (HMPV), *Mycoplasma pneumoniae* and *S. pneumoniae, Haemophilus influenzae*, and adenoviruses.[6,67]

Most RTIs dealt with in primary care are self-limiting viral illnesses, but there are also presentations of more severe infections, up to and including fatal pneumonia.[68] Presentations to primary care with respiratory infections declined for many years before the Covid-19 pandemic, but there were still five million GP consultations each year for respiratory infections, and half a million hospital admissions, with peaks during winter epidemics.[4,6,69,70]

1.5.2 Prevention

Vaccination is the mainstay of prevention for respiratory infections in UK primary care. This is predominantly influenza, Covid-19, and pneumococcal vaccination.[64] These are targeted at higher risk adults, as identified by age or the presence of comorbidities.[64] Children are also vaccinated against influenza and *Haemophilus Influenzae*.[64] Respiratory Syncytial Virus (RSV) vaccination was introduced for the over 75s in 2024.[64]

1.5.3 Diagnosis

Diagnosis of respiratory infections in UK primary care is predominantly based upon the history and clinical examination, without the use of diagnostic tests.[62,63] However, because respiratory infections have shared symptoms clinical examination is extremely poor at identifying the causative infectious agents, or even if there is a pneumonia or not.[63,71] Identifying the organism is therefore not the focus of primary care assessment.[62,63] Instead the aim is to assess severity and arrange for transfer for further care for those who need it.[63] Since the Covid-19 pandemic there has been more testing for SARS-CoV-2, and various attempts to

introduce testing for other agents, however these remain the exception. There are surveillance practices, where testing is routine, but the results are not received acutely, so cannot be used for clinical management.[67,72,73] This is also true for diagnosing pneumonia - routine chest radiography can be requested from UK primary care, but if same day imaging is required this would usually be done via referral to secondary care and therefore does not typically impact upon the primary care assessment or treatment decision.

1.5.4 Acute treatment

Antibiotics are mostly ineffective for viral infections, and bacterial infections are usually self-limiting.[62,63,74,75] However, partly because of the diagnostic uncertainty, primary care presentations with respiratory infections are the number one cause for antibiotic prescribing over the whole NHS (about 10 patients in 100 receive one or more prescriptions per year).[76,77]

A 2023 paper examined Japanese medical records and found that patients who tested positive for influenza who were prescribed antibiotics in addition to antivirals had higher risk of hospitalisation than those only prescribed antivirals.[75] Pneumonia admissions in particular were elevated, so the prescriptions were probably partly related to treating secondary infections.

The other antimicrobial therapies available in primary care for respiratory tract infections are antivirals for influenza, which GPs can only prescribe for influenzas-like illness once the NHS has notified them that influenza cases have passed a threshold in the population.[9] A 2018 systematic review of trials of point of care tests for influenza in ambulatory care settings showed that antivirals are used more when point of care tests for influenza are used (Relative risk 2.65, 95% CI 1.95–3.60; $I^2 = 0\%$).[78] Unlike in Japan, these tests are not widely available in UK primary care. Guideline based care is therefore mostly assessment of severity, and advice regarding self-treatment of symptoms.[63]

1.6 Current practice in UK primary care – CVD

1.6.1 Epidemiology of cardiovascular disease

In the UK cerebrovascular deaths (29,000 per year) and ischaemic heart disease deaths (59,000 per year) are together the commonest cause of death, accounting for a quarter of all deaths.[3,79] There are 200,000 hospital attendances each year for MI and 42,000 premature, preventable CVD deaths in people under the age of 75 in the UK.[3] However, seven out of ten CVD events are not fatal and can lead to chronic morbidity, disability, depression, and heart failure.[3] Stroke, as well as being a major cause of death, is the biggest cause of disability in adults in the UK.[3] MI and stroke are predominantly treated in hospital, but they often present to primary care.[80]

1.6.2 Diagnosis

MI and stroke are medical emergencies, and even in cases where there the urgency is not realised and the patient presents after the fact, the diagnoses are mostly confirmed in secondary care, relying on imaging, ECG changes, and blood tests.[81,82] When people present to primary care with suspected stable angina (based on history, examination and ECG) they may be treated in primary care and referred to outpatient services, where the diagnosis is confirmed or refuted.[83] NICE do not suggest referral to cardiology is mandatory for stable angina, although the CT coronary angiograms required by the guidance are not available in most primary care settings. These diagnostic pathways mean that secondary care data is required to identify CVD events.[84]

1.6.3 Primary and secondary prevention

Primary prevention (i.e. before a first CVD event) centres on risk-factor control to prevent atherosclerosis progressing.[19] This consists of disease specific guidance for managing conditions that are themselves risks for CVD (e.g. diabetes, hypertension, and atrial fibrillation), and management of CVD risk independently, mostly by managing lipids.[10,85,86]

Secondary prevention is specific to the first event type, and as well as risk factor control usually includes the use of antiplatelets or anticoagulants, depending on the type of prior event and risk factors.[81,87] The secondary prevention strategy is to try to prevent progression of atherosclerosis, whilst also adding agents to reduce the risk of thrombosis.

1.6.4 Prediction of CVD risk guides prevention

In both primary and secondary prevention, prediction models are used to identify people at risk of CVD events, to target medical interventions to reduce the risk. In this thesis I have extended this approach to the problem of infection-related CVD events. Risk prediction models estimate CVD risk based on variables that weight each patient's characteristics, this typically includes age, blood pressure, some measure of cholesterol, and other risk factors.[7,88] A systematic review found hundreds of CVD risk models, but none addressed the short-term risk of infection-related CVD events or include infection specific variables.[7]

1.6.5 Primary CVD prevention

Current UK practice in primary prevention is for GPs to estimate each patient's risk of a first CVD event over the next ten years.[89,90] NICE recommend QRISK models for this, but there are many others, including SCORE and ASSIGN.[12] If ten-year risk of a CVD event is estimated to be above 10% then patients should be offered lifestyle advice and statin therapy, although the most recent guidance states that anyone with lower risk who wishes to have statins can do so, as they are cost effective at any level of risk.[10]

Controlling blood pressure with antihypertensive medications is an effective and widespread practice for preventing CVD.[91] It is possible that the use of antihypertensives could reduce infection-related CVD, however there is also evidence suggestive of harm during infections. Early in the Covid-19 pandemic I contributed to a paper that found the risk of Covid-19 mortality was lower in people whose hypertension had not been controlled (odds ratio, 0.76,

95% CI 0.62 to 0.92).[92] It is possible that this was because the treated population had worse underlying health due to the hypertension being present for longer leading to underlying atherosclerosis or organ damage.

1.6.6 Acute treatment of CVD

Acute treatment of CVD events in primary care is limited to conditions not sent immediately to emergency care - people with stable angina, or TIA. Initial medical treatment is for symptoms (antianginals) and secondary prevention – statins, blood pressure control and antiplatelets (aspirin for angina, and clopidogrel or short-term dual antiplatelets for TIA).[82,93,94] There are also people who will not be referred to secondary care for any reason, for example those nearing the end of life or with advanced care plans that preclude admission. Such patients are also less likely to receive medications aimed at long term prevention.[10]

The importance of preventing thrombosis is demonstrated by acute treatments for MI. Emergency care including aspirin can be given in primary care whilst awaiting patient transfer, and a variety of antiplatelets are used in patients undergoing emergency hospital care.[95] After cardiac stenting patients require dual antiplatelet therapy (often aspirin and clopidogrel, but others are used) usually for a year, and aspirin alone thereafter. [95]

1.7 Potential therapies for infection-related CVD events

There are many medications that can reduce the risk of CVD, and some of these might be relevant to preventing infection-related CVD. Broadly these could target thrombosis, atherosclerosis or the severity of symptoms.

1.7.1 Aspirin as a potential intervention to prevent infection-related CVD events

Aspirin (acetyl salicylic acid, or ASA) is an antiplatelet agent.[9] It reduces platelet aggregation by irreversibly inhibiting the enzyme cyclooxygenase one (COX 1).[96] Inhibiting platelet

aggregation is helpful for preventing CVD events occurring via thrombosis, such as MI.[13] The same mechanism is unhelpful if there is bleeding, when platelet aggregation is needed to achieve haemostasis.[13] This is a consequence of the function of the drug, so there is necessarily a trade-off between the benefits and risks.

Aspirin is effective at reducing both coronary events, and stroke. Clinical trials have consistently demonstrated that long-term aspirin reduces the risk of CVD events but increases the risk of bleeding.[8] A meta-analysis in 2009 by Collins *et al* reported that aspirin reduced the risk of myocardial infarction in both primary prevention and secondary prevention, in both men and women, and in people at different levels of estimated CVD risk.[8]

The underlying risk of CVD event determines the use case for aspirin. In long-term primary prevention trials, the risk in the placebo arms was 0.6% per year, compared with 8.2% in the secondary prevention trials.[8] As a result, the absolute reduction in major coronary events from aspirin use is higher in the secondary prevention trials (-1% per year for secondary prevention versus -0.06% in primary prevention).[8] Long-term aspirin is recommended in the population who have already had a myocardial infarction, but not for primary prevention. [19,97] The absolute benefit is related to the underlying risk of CVD, and in primary prevention the absolute benefit is finely balanced with the risk of major extracranial haemorrhage.[8]

The most common serious adverse outcome for aspirin is bleeding.[9] Collins and colleagues found that aspirin increases the annual risk of extracranial major bleeding by 0.03% in primary prevention trials (Rate Ratio (RR) 1.54, 95% CI 1.30 to 1.82).[8] In secondary prevention trials the increase in major bleeding was 0.19% in the 4/16 trials that reported this (RR 2.69, 95% CI 1.25 to 5.76).[8] The variables they identified that increase the risk of MI (age, hypertension, and low-density lipoprotein (LDL) cholesterol levels) were also associated with an increased risk of bleeding. They found the absolute risk of bleeding in

primary prevention is approximately balanced with the absolute benefit in vascular events, even if these risk factors were used to stratify people. They also observed that because statins reduce the absolute risk of CVD events, co-prescribing statins reduces the absolute benefit of aspirin.

1.7.1.1 Aspirin and stroke

Treating patients for short periods of increased risk is a strategy for stroke prevention.[98] The weeks after a Transient Ischaemic Attack (TIA) are high risk for stroke.[98] Randomized controlled trials have measured the effects of clopidogrel (an antiplatelet agent that irreversibly inhibits a platelet adenosine diphosphate (ADP) receptor) plus aspirin versus aspirin alone following high risk TIA and minor stroke.[9,81,94] Meta-analysis of the trials showed absolute risk reductions in stroke of about two percent (6.3% in the single antiplatelet arm versus 4.4% for dual antiplatelets; IRR 0.69, 95% CI 0.60 to 0.79) with increases in major or moderate bleeding of about 0.2% (0.5% vs 0.3%, IRR 1.71, 95% CI 0.92 to 3.20) and that the most favourable balance of risk and benefit was when the duration of the intervention was 10-21 days.[99] Following these trials, twenty-one days of combined aspirin and clopidogrel followed by long term single antiplatelet therapy has been incorporated into clinical guidance.[81,94,98]

There are no primary care trials of aspirin in respiratory infections, but there are some data from hospital-based trials. Various antiplatelets were trialled in hospitalised Covid-19 patients.[100] The largest was an arm of the RECOVERY platform trial that examined aspirin versus usual care.[101] They found a shorter duration of stay in hospital (eight days vs nine), with a higher proportion discharged at 28 days (rate ratio 1.06 (95% CI 1.02 to 1.10), but no change in the composite of requiring mechanical ventilation or death (the primary outcome). However, two thirds of the patients in both arms of this trial were prescribed thromboprophylaxis with low molecular weight heparin (LMWH). Almost all the benefit in

hospital discharge by day 28 came from the group with no anticoagulation (RR 1.19, 95% CI 1.01 to 1.41), and the point estimate for 28-day mortality in this group was 0.83 (95% CI 0.89 to 1.04, 129/466 vs 169/513).[101,102]

There has been one small trial of aspirin in people hospitalised with pneumonia.[103] It was a Turkish trial, published in 2013, and included patients with a minimum of two chronic risk factors for CVD. Inpatients with radiological pneumonia were randomized to 300mg aspirin daily for 30 days, or no aspirin. In the aspirin group 1/91 had acute coronary syndrome within 30 days, compared to 10/94 in the control group (RR 0.10, 95% CI 0.01 to 0.75).[103] There was no detectable change in mortality and no gastrointestinal bleeding was reported.[103] However there are some concerns about this result. The study was small, open label, and excluded 90% of the potentially eligible patients, which raised questions about selection and ascertainment biases.[103,104] The study only included doubling of cardiac troponins as ACS if there was also history of anginal chest pains or ECG changes, however increase in troponins without these findings was more common in the control group.

In summary, aspirin is a possible treatment for infection-related CVD. It is cheap and has been shown to be effective in reducing CVD events in primary prevention, but it carries a risk of bleeding. This bleeding risk limits the use of aspirin in populations at low risk for CVD. To use aspirin in primary prevention would require identifying a high-risk population to target, where the benefits may outweigh the risk. During the post-infection period CVD event risk is elevated, and so the risk-benefit balance may be in favour of using aspirin.

1.7.2 Statin therapy as a potential intervention to prevent infection-related CVD events

1.7.2.1 Cholesterol and statins

Cholesterol is a fat produced in the liver, it is a cause of atherosclerosis, and both cholesterol and CVD event risk can be reduced with statin therapy.[105,106] Because cholesterol is hydrophobic it is transported in blood in association with proteins. It forms particles with

water soluble surface phospholipids and the cholesterol in the centre. Plasma cholesterol is classified according to the density of the particles in which it is transported. The two subtypes relevant to this thesis are low density lipoprotein (LDL) and high-density lipoprotein (HDL) cholesterol.[107]

LDL cholesterol is a cause of atherosclerosis.[13] The liver releases VLDL (very low-density lipoprotein) cholesterol which is converted to LDL cholesterol. LDL circulates and is taken up by the liver and other tissues. [13] High density lipoprotein (HDL) cholesterol is associated with different transport proteins and is taken back to the liver and excreted in the biliary system. [13] HDL cholesterol is protective against atherosclerosis.[107]

LDL cholesterol is more difficult to measure than HDL cholesterol, so in the UK guidance focusses on 'non-HDL cholesterol' on the assumption that most of this is LDL.[10] Possibly because LDL cholesterol may not be reported in primary care databases the QRISK models use the ratio of total cholesterol to HDL cholesterol as a variable, and I have followed this practice in this thesis.[108]

Statins, properly '3-hydroxy-3-methylglutaryl Coenzyme A (HMG-CoA) reductase inhibitors', are a group of medications that interfere with cholesterol synthesis.[9] Most plasma cholesterol is synthesised in the liver, from saturated fats contained in the diet. The rate limiting step is performed by HMG-CoA reductase, so inhibiting this reduces cholesterol production, and plasma cholesterol levels.[9]

There is extensive evidence from randomised clinical trials that statins reduce CVD events in both primary and secondary prevention.[11] They are also very safe, with few significant side effects.[109] The combination of effectiveness and safety is why long-term statin use is the mainstay of primary prevention strategies.[10,19]

1.7.2.2 Statin use for other conditions

Statins have been suggested as impacting many different clinical outcomes, but there is minimal evidence that they affect anything other than vascular diseases.[66] An important result for this thesis is that a systematic review of randomised trials found no evidence that statins prevented either infections in general, or respiratory infections.[110,111] Statins have also been trialled for people who are critically ill with acute respiratory distress syndrome (ARDS), the severe syndrome of non-cardiogenic pulmonary oedema and diffuse lung inflammation that can accompany infections. A systematic review of statin trials in intensive care patients with ARDS found no change in the (extremely high) 28-day mortality, ventilator free days, or serious adverse events.[112] This population, with a 21% death rate, is far removed from primary care. There is some observational evidence that suggests reduced one year mortality in people taking statins following a respiratory infection, but this was driven by non-CVD mortality, suggesting an inaccurate measurement is likely.[113]

Statins are a therefore a possible intervention for infection-related CVD events. They target plaque stability, and atherosclerosis, and are very safe with few off-target effects. This means they could be safely offered to a large population of lower risk people. The variation in their use is also likely to be exploitable as a source of variation in observational studies.

1.8 Summary of infection-related CVD event background

Respiratory infection is associated with increased risk of CVD events.[1,2] This holds true for respiratory infections caused by different agents. There are hundreds of established models for predicting CVD events, but before this thesis there was no validated method for predicting infection-related CVD events, nor any unvalidated attempts to model this relationship.[7] It is biologically plausible that respiratory infection causes CVD events, given what we know about the pathology of infections and CVD. However, the association does not need to be causative to be predictive.[114]

1.8.1 Treatments for infection-related CVD

There are effective medications for preventing CVD, both over the long term, and short-term interventions for periods of higher risk (e.g. dual antiplatelet therapy to prevent stroke post TIA).[98] Vaccines can prevent infections, severe infections and CVD, but there are no established interventions for intervening to reduce infection-related CVD, once the infection is underway.[29] However, as respiratory infections are extremely common, long-term trials of medications must necessarily have included people with intercurrent, and multiple respiratory infections. As the two main processes leading to CVD events are thrombosis and atherosclerosis, this was my focus for identifying potential interventions.

1.8.2 Background conclusion

We did not know if infection-related CVD events are predictable, nor if there are interventions that could reduce the risk once the infection has been diagnosed. These evidence gaps are the subject of this thesis.

1.9 Hypotheses, aim, research questions, and objectives

1.9.1 Main hypothesis

My hypothesis was that acute respiratory infections are an opportunity to intervene to reduce associated CVD events in the following weeks.

This leads to two further ideas.

1.9.1.1 Secondary hypotheses

CVD events have been shown to be predictable in many different settings, for primary and secondary events, for specific clinical scenarios, and over the long term.[7] There are many interventions that mitigate CVD event risk.[19] I therefore hypothesized that

1. CVD event risk in patients with respiratory infection is predictable
2. Medications that protect against CVD events in other settings can also reduce infection-related CVD events, and in particular there could be a benefit associated with aspirin or statin use

1.9.2 Thesis aim and research questions

The overall aim of the work in this thesis was to investigate the potential for predicting and preventing infection-related CVD events in primary care.

The corresponding research questions are:

Firstly: are infection-related CVD events predictable?

Secondly: could aspirin use mitigate the risk of infection-related CVD events?

Thirdly: could statin use mitigate the risk of infection-related CVD events?

1.9.3 Thesis objectives

I addressed these questions through four experimental chapters, addressing four main objectives:

1. The first objective was to model the risk of CVD events in people with respiratory infection in UK primary care
2. The second objective was to externally validate the models derived in objective one
3. The third objective was to measure the effect of aspirin in people with higher infection-related CVD event risk
4. The fourth objective was to measure the effect of statin use in people with moderate to high infection-related CVD event risk

2 Chapter Two: Thesis methods overview

In this thesis I used two main methods, prediction modelling for the first two objectives (chapters three and four), and causal inference modelling for the second two (chapters five and six). In the final chapter, chapter seven, I discussed the findings and their implications.

2.3.1 Data source

All the data used in this thesis came from routinely collected sources, described in more detail in the experimental chapters. These are databases of codes from medical records from primary and secondary care, and UK government statistics on the cause of death, and relative socioeconomic deprivation.[115,116] The advantage to this data is primarily that it is representative of the UK population from which they were obtained, and where the results were intended to be used. These types of data have been used extensively for medical research, including prediction and causal inference studies.[117] There are problems with routinely collected data. Coded data has variations in quality, and misses uncoded 'free-text' comments and diagnoses. These missing data is not recoverable or detectable and generally is ignored. However, there are also missing data within variables that are identifiable, unknown values for measurements and characteristics (e.g. smoking status or blood pressure measurements that have not been recorded) and there are methods for mitigate these in analyses.

2.3.2 Modelling methods

There are two main modelling approaches I used in this thesis. Prediction modelling, which seeks to predict the probability of some future event happening, and causal inference modelling, which seeks to estimate a causal relationship.

2.3.3 Prediction modelling overview

The purpose of clinical prediction models is to provide diagnostic or prognostic predictions, which can be used clinically.[118] This depends on the predictions being reasonably accurate but does not require a causal link between the variables in the model and the outcome.[118] The typical process is to develop statistical models in a population, usually with a regression model.[118] Simpler scores can be derived from the regression models, based upon the risk attributed to levels of the variables in the regression models.[118,119] The model or score must then be validated – the performance assessed. This can be done internally (in the same population in which it was derived), but as internal validation is prone to overoptimistic performance estimates they should also be validated externally (in another population) before they can be used.[120]

2.3.3.1 Model derivation – model selection

Derivation of prediction models is the subject of chapter two. As there were no prior models I was faced with the problem of how complex to make my model. I considered this as two related problems - which variables to include, and how many variables to include.

2.3.3.1.1 Which variables?

There are statistical selection methods for choosing variables. The 'stepwise' approaches compare models with and without variables included.[121] They have been criticised as having potential to introduce or magnify the magnitude of bias whilst failing to select the best combination of variables.[122,123] When using statistical selection methods variables more strongly correlated with the outcome have a higher probability of being included in the final model. Unfortunately, the strength of the correlation in the dataset might be spurious – it might be influenced by some random or systematic factor. [122,123] The model choice might also be unstable – small differences in the variables could lead to different models being chosen.

[122,123] A further problem is that in very large datasets almost every association will have a high level of statistical significance, so the methods must be adapted.[121]

Another approach, and the one I chose, is to use external sources of information. I chose to use knowledge from prior studies to identify potential predictors and expert opinion to rank them, which has the advantage of avoiding being data driven. [122,123] This can lead to conservative choices but has the benefit of simplicity and face validity. To paraphrase George Box – 'all methods for developing models are wrong, but some of them are useful'.[124]

My approach was to generate a list of potential predictors from prior literature and ask clinical experts to rate each variable's importance. I then combined the expert ratings to obtain an ordered list of potential predictors, so that I could build models from those considered most important.

2.3.3.1.2 How many variables?

The upper limit to the number of terms in a model (including levels of categorical variables, higher order terms, and interaction terms) should be determined by a sample size calculation and depends on the data available.[125] I present this in the model derivation chapter. The 'big data' I used would enable very complex models.

However, a model should be simple by default and only complex if this is necessary to achieve acceptable performance.[124] Complex models, with many variables or higher order terms, are impractical for clinical use, and there are diminishing returns from increasing complexity.[114] Whilst many factors may be individually strongly correlated with disease or prognosis, when combined they may not add much to the predictions performance.[114] For example, variation in blood pressure is associated with cardiovascular disease, but the ability of it to improve CVD event prediction when in combination with other factors is minimal, as was seen with its inclusion in QRISK3.[126,127]

My own personal preference is towards simpler models, because in clinical practice simple scores seem to dominate. Simpler models are easier to use. There are some demonstrations of this. The first CVD risk factors were identified in the Framingham study which ran from 1948. [128–130] Framingham was first designed to identify the causes of cardiovascular disease but later was also used for prediction.[128,129] A logistic regression model was developed to predict the risk of CVD events, but it only became popular amongst clinicians once converted to paper based risk tables in 1998.[128–130] The WHO still issues CVD risk tables for the different epidemiological sub-regions of the world so that it can be used without access to technology, although a previous project I contributed to converted the old WHO score to an R package to aid research.[131] The SCORE tool which the European Society of Cardiology recommended for risk estimation in primary prevention also uses a table format.[132]

Complex models may be implemented by computer systems, but this also has had problems. The clinical software used by GPs tends to calculate a QRISK2 score automatically but there was an incident because the software was not calculating the risk accurately. In 2016 the MHRA issued an alert that the GP record software 'SystmOne' had been calculating QRISK2 wrongly since 2009.[133] Nobody noticed because it was being done by the computer, the workings were opaque to the users, and the predictions seemed reasonable.

Because there are no earlier published models of infection-related CVD event risk to build on, I chose to investigate the number of variables pragmatically.[7] I built two models, one with only the most highly rated variables from expert opinion, and another including the same variables plus more variables that were ranked less highly. I then compared these models in validation, to see if including more variables improved performance.

2.3.3.1.3 Prediction model validation

The performance of a model is more important than how it was derived, and so validation is an important step.[119]

Internal validation is assessing the model in the population used to derive the model, and external validation the process of doing this in another dataset.[120] Internal validation can be used to estimate the performance of a model, but the gold standard is external validation. I used internal validation as a check for problems with the modelling process and presented it in the model development chapter (chapter three). External validation provides the results needed to assess a tool for potential use in practice.[120] That is the subject of the external validation chapter (chapter four).

The clinical utility of prediction models is also important.[134] This is the overall effect of using the tool. It depends on the predictive ability of the tool, but also on the actions taken because of the predictions, and the consequences that follow these actions. As an example, scores for predicting stroke after TIA were developed and validated, and then were used to decide how quickly patients with a possible TIA should be assessed by stroke services. Because there was a high risk of harm from missing strokes the 2019 NICE guidance stated that they should no longer be used:

"*Evidence showed that risk prediction scores (ABCD2 and ABCD3) used in isolation are poor at discriminating low and high risk of stroke after TIA… …Arranging specialist assessment less urgently for some people based on a tool with poor discriminative ability for stroke risk has the potential for harm.* " [82]

The possible negative consequences of a treatment or investigation undertaken because of a clinical prediction tool depend on the action taken, and the risk of missing cases is different for different events. In the example above the action taken, expert assessment, carries few (if

any) inherent risks to the patient, whereas missing an opportunity to prevent a stroke is a serious problem. To help clinicians assess the equivalent trade-offs I used net benefit analysis to assess the clinical utility of the models in chapter four.[134] Net-benefit analyses compare the ratio of true and false positives at different probability thresholds, which allows the assessment of the trade-off between correctly treating people who will be cases, and unnecessarily treating people who would not.[134]

The prediction of risk, and the balance of risks and harms is also fundamental to identifying the populations in the causal inference chapters of this thesis. In chapter fiveI estimated the effects of aspirin, which carries a risk of bleeding. I restricted the study population to people for whom it might be worth accepting the risk of bleeding – people at higher risk of infection-related CVD events, as identified by the prediction models. For chapter six's analysis of the effects of statins (which have very low risk of side effects) I used a larger population, only excluding patients with very low risk of infection-related CVD events.

2.3.4 Causal inference – propensity modelling

Causal inference epidemiology aims to estimate causal relationships. Causal inference is challenging because it requires estimating counterfactuals – things that did not occur.

I used propensity modelling methods and regression models in chapters five and six, to assess the effects of aspirin and statins in different populations and with different implementation strategies. Propensity models aim to model each participant's propensity to become exposed to a treatment. If this propensity can be estimated then analysis strategies taking it into account can give unbiased measurements of the causal effect, providing the necessary assumptions are met.[135] Propensity models were originally developed from the idea of counterfactual treatments and outcomes, and the assumptions that they required follow from this framework.[135] In this thesis the counterfactual treatments are aspirin or statin exposures

that are opposite to what actually happened, and the counterfactual outcomes are CVD events or bleeding that would (or would not) have happened if the treatment had been counterfactual.[135,136]

2.3.4.1 Propensity modelling assumptions

The assumptions required for propensity modelling are no interference, consistency of effect, positivity, and conditional exchangeability.[135]

No interference is the assumption that treating one person does not alter another person's risk.[135] This could be violated by treatment for infectious diseases, where treating one person protects others from being infected too.

Consistency of effect is met if the method of allocation does not change the effect of the treatment.[135] For most medications this is a reasonable idea, behavioural interventions provide an example of a violation – people seeking out support for behaviour change are likely to be more motivated than those offered it out of the blue.

Positivity is the assumption that everyone in the population has some possibility of being treated (i.e. the propensity is positive). In a trial this is achieved by randomisation but in observational studies it is more difficult to be sure positivity has not been violated.[137]

The other major problem with observational studies measuring causal effects is conditional exchangeability.[135] This is the assumption that conditional on each patient's confounders, the treated and untreated could be swapped, and the outcomes would be the same. In other words, there is no confounding that has not been accounted for. In a trial this is achieved, on average, by randomisation. In observational data exchangeability is not a verifiable assumption, but it is possible to examine the balance of known confounders.[138,139]

2.3.4.2 Propensity model implementation

Propensity estimates for binary treatments are usually calculated with logistic regression modelling, and I took this approach in chapters five and six.[139] The propensity estimates can be used to obtain effect estimates in many ways; by regression, stratification, matching, and weighting.[140,141] I used regression and stratification in the development of and selection of propensity models. I used different methods for estimating the effects in chapters five and six. In chapter five I used regression in combination with inverse probability weighting - the 'doubly robust' approach - to estimate the effects of aspirin.[142] Inverse probability weighting allocates each participant a weight that is the inverse of the estimated probability of receiving the exposure that they experienced.[142]

In chapter six I explored the effects of weighting methods on covariate balance, and for the final analysis of the effect of statins I used regression with overlap weighting.[143] Overlap weights are the estimated probability of being allocated to the opposite treatment group.[143] As this is not a standard weighting in the software I used, I coded an implementation that would work with multiple imputation.[144] To obtain the final effect estimate I used weighted regression adjustment to model counterfactual outcomes, and then bootstrapped the procedure 500 times to estimate the confidence intervals.[145]

2.3.5 Missing data methods

The simplest way to deal with missing data is complete case analysis, but this reduces the available data, gives biased results unless the data are missing completely at random, and is recommended against.[146] The alternatives that I used are missing indicators, and multiple imputation.[147–151]

2.3.5.1 Multiple imputation

Multiple imputation is the process of creating copies of the dataset, each with imputed values for the missing data, drawn from a distribution defined by modelling the variable. It relies on

the missing data being 'missing completely at random', meaning that there was something external to the question that caused loss of information, or 'missing at random', meaning that the missing data are related to some data that are observed in the dataset.[118,147,151]. The data could also be 'missing not at random', meaning the missingness is not predictable from the data in the dataset, in which case multiple imputation can give biased results.

Where I thought the data likely to be missing at random I used Multiple Imputation by Chained Equations (MICE), an algorithm that iteratively estimates the missing data.[152] The first step is filling all the missing data with randomly chosen values. Variables with missing data are all then modelled one at a time. The first variable (using only its non-missing data) is modelled with regression on all the other variables including the outcome. The regression model is defined for each variable and so varies for different types of data in the dataset, logistic regression for binary variables, ordinal regression for ordinal variables and so on. Missing data in the first variable are then replaced with numbers drawn from the modeled distribution. The next variable with missing data is then treated in the same way, and so on until all the variables with missing data have gone through this process. These estimates are not accurate after one cycle, so the cycle is repeated several times (burn in) so that the regression results converge (I repeated ten times, the default that has been shown to be adequate unless the data are highly correlated).[152] The result is a single copy of the dataset, with no missing values known as an imputation. This is repeated m times to obtain m imputed datasets, each with filled values for the missing data.[152]

Some CVD risk modelling studies have used small number of imputations, five for the QRISK models for example.[108] Framingham used full cases only, and SCORE, which was derived from multiple cohorts did not appear to address missing values.[129,153] There have been concerns that there smaller numbers of imputations may not be adequate, and a heuristic of the number of imputations being equal to the number of incomplete cases could be used.[152]

Unfortunately this is not always practical because imputation is computationally expensive (i.e. slow) for the process of building the multiple imputed datasets, and dealing with multiple analyses in each of the datasets. I used a pragmatic approach. For the prediction modelling chapters I used five imputations for a dataset with only one missing variable, and ten for a dataset with five missing variables. For the causal inference modelling I used 20 imputations – this was feasible because the datasets were smaller, and a more powerful computer was available to me.

2.3.5.2 Missing indicator methods

Multiple imputation assumes it is possible to model the distribution of the missing data from the non-missing data in the dataset.[147,152,154] This is not always the case. I chose to use missing indicator variables for acute blood tests, CRP and full blood counts. These tests are requested for a clinical reason, rather than screening, and very high proportions are 'missing', meaning untested. I thought these patients were probably qualitatively different to those who were tested, and it would be impractical to either model the indications for testing or impute such high levels of 'missing' data. Another advantage to missing indicator methods is that they reflect clinical reality – when the model comes to be used the data may be missing for these tests.

In chapter six, which deals with prescriptions of statins, I combined a missing indicator with multiple imputation for cholesterol because I wanted to include information about being tested or not, as well as the level of cholesterol.[149]

2.4 Dissemination of findings

The chapters of this thesis detailing prediction modelling (chapters three and four) were based on a paper that has been accepted for publication with eClinicalMedicine in May 2025 (Appendix – Publication based on chapters three and four). Early results from the prediction modelling I presented at the GRIN (General practice Research into INfection) conference in

Lund in 2022, and further modelling was accepted as a poster presentation at the European Society of Hypertension in Milan in 2023. I also published an editorial on respiratory infection-related CVD events in BJGP.[155]

3 Chapter Three: Development of clinical risk models for cardiovascular events following acute respiratory infections: a retrospective cohort study

When trouble is sensed well in advance it can easily be remedied; if you wait for it to show itself any medicine will be too late because the disease will have become incurable. As the doctors say of consumptive illnesses, to start with it is easy to cure but difficult to diagnose; after a time, unless it has been diagnosed and treated at the outset, it becomes easy to diagnose but difficult to cure.

Niccoló Machiavelli, The Prince, 1532

3.3 Introduction

3.3.1 Overview within thesis – how this chapter fits in

In the introductory chapters I outlined the main problem – that there is an increase in the risk of CVD events following respiratory infections. CVD events are a serious complication of acute respiratory infection but current primary care guidelines for acute respiratory infections do not consider the risk of CVD events.[14,62,63,68] This means there may be an opportunity for intensifying CVD prevention during infections and for the following weeks. Prior to this thesis, there was no validated way to identify which patients with infections should be targeted with preventative measures.

This chapter describes how I derived two models to identify people at increased risk of CVD following a respiratory infection. I developed two models to help assess the effect of increasing the complexity of the statistical modelling on predictive performance.

Only externally validated prediction models should be used in clinic.[156] I describe the external validation in the next chapter, chapter four. The later chapters, five and six, describe my exploration of the effects of aspirin and statins on higher risk patients identified by the prediction modelling.

3.3.2 Rationale

Prediction is important for informed discussion, research, and decision making about therapeutic options.[120] A GP and their patient may struggle to have an informed conversation about a risk which is not quantifiable.

The probability of CVD events limits the interventions that could be used to try to prevent them occurring. For example, one would not wish to expose a patient to the bleeding risk associated with taking antiplatelets if the risk of CVD events was lower than the risk of

bleeding. Any research into therapeutic options would benefit from being able to identify the likely event-rate in the population.

A risk prediction model is necessary because the use of other epidemiological evidence for prediction is not informative.[114] Risk factors strongly associated with diseases in the population can be extremely poor for prediction.[114] Odds ratios, or other measures of association, are not diagnostic accuracy measures, and even strong associations with very high odds ratios may be extremely poor when it comes to making predictions.[114]

There are many CVD risk prediction models, but these typically aim to predict long-term or lifetime risk based on demographic and lifestyle factors, and chronic health conditions.[7] None considers fluctuating short-term increases in CVD risk such as that associated with acute respiratory infections.[7] None of them include infection-specific factors.[88] I sought to address this evidence gap in these prediction modelling chapters.

3.4 Methods

3.4.1 Modelling aim

The overall aim was to produce a prediction tool for 28-day CVD event risk for people with acute respiratory infection.

3.4.2 Chapter objectives

As there were no pre-existing models, I examined the trade-off between model complexity and performance by building more than one model.

1. The first objective was to derive a simple clinical prediction model; Model One, comprising a small number of clinical variables.
2. Objective two was to derive another more comprehensive model; Model Two, including all the predictors from Model One, and some extra predictors.

3.4.3 Data source

The data source for all the experimental chapters was routinely collected data. For prediction modelling I used two retrospective observational cohorts, one to derive the statistical models, and another to externally validate them. External validation is the subject of the next chapter and will not be covered here, but as the cohorts were entirely analogous, this section is also relevant to the validation cohort. The cohorts were derived from routinely collected primary care data in the Clinical Practice Research Datalink (CPRD) Aurum and CPRD Gold databases. These are databases of routinely collected primary care medical records, containing coded anonymised data from participating practices. CPRD provides databases of UK primary care records from different electronic clinical records systems. Aurum comes from practices in England that used EMIS® software (Egton Medical Information Systems, Leeds, UK), it covers about 13% of the population of England.[157] GOLD data comes from different practices across the UK that used Vision® software (Cegedim Healthcare Solutions,

London, UK).[116] Gold covers about 7% of the population.[116] The model development dataset came from CPRD Aurum. I used CPRD GOLD data for validation (see next chapter, section 4.4). I excluded patients from the larger Aurum database if they appeared in both datasets. Both datasets are representative of the wider patient population in terms of deprivation, ethnicity and age.[116,157] They provide coded data, rather than free text. Clinicians and their teams code the data in the process of routine clinical care. CPRD provided the primary care data to the Nuffield Department of Primary Care Health Sciences under a departmental licence. The CPRD Independent Scientific Advisory Committee (ISAC) approved the protocol for all the analyses in this thesis (21_000380). CPRD provided person-level data linkages to the study population after extraction.

3.4.4 Data Linkages

I used data from 1st January 1999 to 31st December 2019 because CPRD datasets are linked to Hospital Episode Statistics (HES) and Office of National Statistics (ONS) data from the start of 1999. The end of the follow-up period was the latest available data at the start of the project and preceded the start of the UK's Covid-19 pandemic. The ONS datasets were Index of Multiple Deprivation (IMD) data, and ONS mortality data.

3.4.4.1 Index of Multiple deprivation

IMD is the official measure of relative deprivation and is reported based upon small areas in England.[158] The areas have about 1,500 people living in them. [158] ONS estimates deprivation using a measure that covers seven domains: income, employment, education, skills and training, health and disability, crime, barriers to housing and services, and living environment. [158] They then rank areas by score; IMD is a relative, rather than absolute measure. [158] I used patient level IMD data.

Primary care data were linked to IMD data by CPRD. Small area ONS data is applicable to England only, so patients living in Wales or Scotland are not linkable. CPRD Aurum data comes only from England, so has a higher proportion with IMD linkages than CPRD Gold.

3.4.4.2 Mortality

All patient deaths occurring during the follow-up period were ascertained through linkage to civil registrations of death data held by ONS, linked by CPRD. This data includes medically certified causes of death.

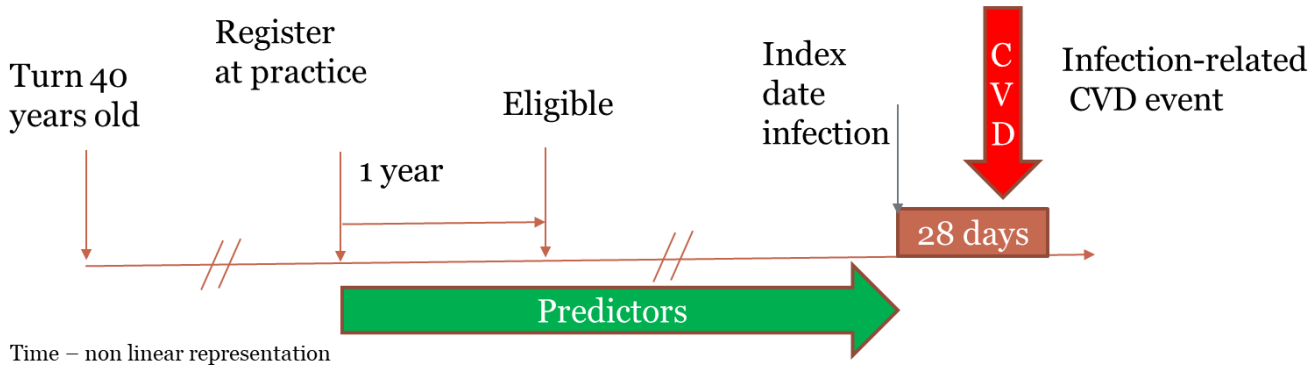
3.4.4.3 Hospital Episode Statistics (HES)

NHS hospital episodes of care are included in the HES dataset, which is linked to CPRD. HES was used to identify CVD outcomes. Linkage to HES revealed some participants had prior CVD which was not revealed in CPRD, and the HES data was used to refine the cohorts by excluding these patients.

3.4.5 Population

For all the studies in this thesis patients became eligible for inclusion at 40 years of age (Figure 1: Patient timeline in cohorts), I excluded younger people because of the low prevalence of CVD. This cut-off aligns with other CVD models.[108] Each patient's index date, when they entered the cohort, was the date of the first diagnosis of an acute respiratory infection (RTI) coded in their primary care record after the age of 40 years. The first respiratory infection was used to simplify the modelling, which would otherwise have had to account for within person correlations. I excluded patients with a history of CVD events in their CPRD or HES records at any point before the index date. Prior CVD events were defined in the same way as the outcome (section 3.4.7). I excluded people with less than one year of CPRD data before the index date to avoid recording old CVD events as new ones, which could happen when patients register at CPRD practices for the first time and their past medical history is reviewed.([117,159])

Figure 1: Patient timeline in cohorts



3.4.6 Respiratory infections

I identified acute respiratory infections by the presence of codes in the primary care record. I have made my codelists freely available online as this is currently a common practice (https://github.com/Protocols-For-Research/CPRD-codes-CVD-infection-risk). I excluded respiratory infections I considered probably chronic (such as mycobacterial infections) from the codelists. I included infections of any severity. I defined three mutually exclusive categories of respiratory tract infection (upper, lower, and lower with pneumonia codes), plus influenza-like illnesses (ILI) which was not exclusive to the other three categories (Table 1: Respiratory infection classification system). I defined upper respiratory tract infections (URTIs) as affecting anatomical sites at or above the larynx. Lower respiratory infections (LRTIs) included any codes for acute infections below the larynx. Infective exacerbations of chronic obstructive pulmonary disease (COPD) I included as LRTIs. I defined LRTI with a diagnosis of pneumonia (hereafter pneumonia) as a LRTI with one or more codes identifying pneumonia. If a patient had codes for more than one category then I selected the most severe with the following hierarchy: pneumonia, LRTI and URTI, making these mutually exclusive. Whilst I was able to identify codes for ILI, these were not exclusive from the other

categories. Some of the codes for influenza and influenza-like illness do not indicate an anatomical site; other codes specify influenza causing an URTI or a pneumonia.

Table 1: Respiratory infection classification system

Exclusive hierarchical categories of respiratory tract infection (RTI):	Possible additional category
1. Pneumonia	Influenza or influenza like illness (Lower RTI unless specified otherwise)
2. Lower RTI	
3. Upper RTI	

This table illustrates a hierarchy of respiratory infection codes. All patients have a hierarchical code indicating site. Respiratory infection codes indicating an anatomical site trump one another, so each patient is allocated to the most severe category of infection they have a code for. If they have a code indicating influenza-like illness they are classified as having influenza as well. If there is only an influenza/ILI code, without indicating a site, they are allocated to influenza and lower RTI.

3.4.7 Outcome

The studies in this thesis all followed patients for 28 days from the index date, a time duration consistently associated with increased risk of CVD.[1,2]

The primary outcome was a composite of a new MI, angina, stroke, TIA, or deaths associated with these. These were chosen as major types of atherosclerotic diseases.[13] The outcome was identified with a list of code which I have published online. The codelists are designed to identify new myocardial ischaemia (myocardial infarction, angina, acute coronary syndromes, or ischaemic cardiomyopathy), new cerebrovascular events (stroke, and TIA), and deaths from these causes. There are multiple definitions of major adverse cardiovascular events (MACE), these diagnoses align with the four-point MACE definition used in some clinical trials of diabetes therapies.[160]

Generic heart failure diagnoses are often included in outcomes used in CVD models, but I excluded them.[7] Heart failure is commonly due to chronic processes unrelated to atheroma. Chronic causes such as alcohol, infections such as HIV, and hypertension are not likely to be preventable by intervening at the point of acute respiratory infection.[161] Acute respiratory

virus infections can cause heart failure via infective or inflammatory myocarditis.[161] I excluded these rare events because they are a separate clinical entity unrelated to atherosclerosis.[58]

I used both primary and secondary care clinical codes, and ONS mortality records to identify outcomes (https://github.com/Protocols-For-Research/CPRD-codes-CVD-infection-risk and chapter appendix section 9.1.2). CPRD recommend the use of search strategies rather than specific codelists.[117] I searched CPRD code browser files using Stata 17 (Stata Corp. College Station, Texas). I aimed to make code lists specific, rather than over-sensitive. As an example, I did not use codes for referrals to chest pain or TIA clinics in the outcome, as many of these patients go on have CVD events ruled out.

3.4.8 Predictors and modelling

3.4.8.1 Predictor identification and selection

I identified predictors from a process of reviewing previous publications and expert clinical assessment.[1,7,44,65,162–169] I first identified clinical variables associated with CVD, severe acute respiratory infections, or both. Tables of these are included in the supplementary materials (Section 9.1.1). The potential variables included demographics, medications, laboratory tests and physical measurements. To help prioritise variables for inclusion, four general practitioners with a special interest in cardiology assessed their perceived importance. I asked them to consider both how well they expected the variable to predict, and the likely accuracy of coding in the GP record.

I standardised each GP's ratings into Z scores to express the relative importance they gave each variable on the same scale, with a mean of zero and a standard deviation of one. I then combined these scaled scores by arithmetic mean across GPs to give a mean Z score for each variable. I used this overall mean Z score to rank the variables from most to least relevant.

3.4.8.2 Predictor implementation

All the predictors were derived from the clinical records and linked datasets (I did not collect extra data). Predictors all came from the clinical record before the index date (Figure 1: Patient timeline in cohorts). I had to define a relevant time window before the index infection to search for the codes. I used different time windows for different variables (Table S28: Clinical variables considered for inclusion in models, and rankings by clinical experts). For cholesterol to HDL ratio, body mass index (BMI), and systolic blood pressure, which I felt might change gradually I used the most recent record in the five years before the index date. Cancers also had a five-year limit as this is the timeframe for follow-up for many cancer trials. For medications I used six months prior to the index date, so that they were relatively up to date. I took codes for other diagnoses that were unlikely to change, and family history, from the entire record prior to the index date. To better differentiate ex-smokers and never-smokers, and quantify the amount people smoked, I used the two most recent smoking records. I carried forward smoking quantities if they were missing in the most recent record, and classified non-smokers as ex-smokers if they had a prior record of smoking.

3.4.8.3 Modelling strategy

I developed two statistical models predicting infection-related CVD. This was to assess the trade-off between complexity and performance in the absence of previously published models.[7] I included infection type in both models. Model One also included the most important variables as identified by clinical experts. Model Two included all the variables in Model One, plus additional clinical variables from further down the expert's ranking. Model One included those variables with a mean Z score >1 and Model Two included variables with a mean Z score >0. Variables with a mean Z score of 0 or less were not included in models (Table S28: Clinical variables considered for inclusion in models, and rankings by clinical experts).

By ‘clinical variable’, I mean a diagnosis, demographic, or test, as presented to and ranked by the experts, rather than a term in a statistical model. For example, one of the clinical variables I presented to clinicians for ranking was smoking status. This is a single clinical variable but is represented in the statistical models as multiple mutually exclusive categories (i.e. never smoked, ex-smoker, smokes <10 per day, smokes 10-19 per day, smokes 20+ per day and smokes an unquantified amount per day).

3.4.9 Sample size calculations

I calculated sample sizes for the study protocol using methods published by Riley *et al* which I implemented in a spreadsheet.[125] I based the calculation on preliminary counts from CPRD Gold data; this gave a conservative outcome prevalence of 0.089%. I aimed for a global shrinkage factor of >0.995, though I planned to externally validate, rather than apply a shrinkage factor. I assumed a maximum of 50 candidate variables (including transformations and interactions). I also checked I would meet the criteria of an absolute difference of <0.05 in apparent and adjusted Nagelkerke's R^2, and a margin of error in outcome proportion estimates for null model <0.05. I calculated 61,198 patients would be required to meet these criteria.

3.4.10 Statistical analyses

I used Stata versions 17 and 18 for all the statistical analyses in this thesis [Stata Corp, College Station, TX].

3.4.10.1 Descriptive statistics

I calculated baseline descriptive statistics in the cohorts, according to outcome status. I estimated means and standard deviations for continuous variables, and numbers and percentages for categorical variables.

3.4.10.2 Missing data

If there were no codes for binary variables prior to the index date, I assumed the diagnosis or prescription was absent.

For continuous and categorical variables with missing data I used multiple imputation and missing indicator methods.[150,152] I imputed continuous variables (total serum cholesterol to HDL ratio, systolic blood pressure and body mass index (BMI)) after log transformation. Smoking status and IMD deciles I imputed as ordinal variables. I used five imputations for Model One, which had only one imputed clinical variable (smoking), and ten imputations for Model Two, which included five imputed variables. I used a larger number of imputations to increase precision with the greater number of missing variables, the choice of five and ten iterations was pragmatic, balancing precision with speed of computation. I used multiple imputation chained equations after ten burn-in iterations, with Stata command *mi impute*.[148] I used ordinal logistic regression models to impute ordinal variables and linear regression for log transformed continuous variables. I assessed imputations for consistency by examining summary statistics and density plots for imputed data and the original dataset.

I also used missing indicator methods for recent blood tests. These were platelet count and C reactive protein. For these I used categorical variables, with missing indicators for unknown values. I categorised CRP results using thresholds used in clinical trials of point-of-care CRP testing: unknown, <5mg/L, 5 to 19 mg/L, and 20 mg/L or more.[170] I categorised platelet results by reference range into: unknown, <150 $x10^9$/L (thrombocytopaenia), 150 to 450 $x10^9$/L (within reference range), and >450 $x10^9$/L (thrombocythaemia).

3.4.10.3 Refinement of variables

I refined, combined and dropped some of the clinical categories after the clinical experts had prioritized them. These were COPD, hypertension, dementia, diabetes subtypes, non-steroidal anti-inflammatories, erectile dysfunction, chronic kidney disease, peripheral vascular disease,

cancer subtypes, family history of CVD, and vaccination status. Their treatment is detailed below.

The top ranked variables for Model One, before refinement into the final model were: Age, heart failure, diabetes, smoking status, chronic kidney disease, peripheral vascular disease and COPD. For the initial Model Two, I included those in Model One plus variables further down the ranking. The ranking continued: systolic blood pressure, sex, cholesterol to HDL ratio, BMI, atrial arrhythmias, dementia, anticoagulants, NSAIDS, antiplatelets, antihypertensives, rheumatoid arthritis, statin use, platelets, CRP, erectile dysfunction, other chronic heart diseases (including valve disease, congenital disease), IMD decile, haematological cancers, solid cancers, family history of CVD in first degree relative less than 60 years of age, and pneumococcal vaccine.

Because exacerbations of COPD were classed as lower respiratory tract infection, I did not include COPD as a separate variable.

A history of hypertension was strongly correlated with both systolic blood pressure and diastolic blood pressure, and so I included only systolic blood pressure.

I initially defined family history of CVD as an event in a first-degree relative aged less than sixty, reflecting the definition used by QRISK3.[108] Unfortunately, the coding systems do not include codes for this. Instead, I defined a high-risk family history using the available codes as being CVD aged less than 65 years if the first-degree relative was female, and 55 years if they were male.

I simplified variables by combining rare and overlapping categories. I combined diabetes mellitus type one (which was rare) with type two diabetes, and diabetes of other and unspecified types. I did not include the strongly correlated ‘glucose lowering medications’. Haematological cancers were uncommon, so I combined them with solid cancers. I also

combined peripheral vascular disease with chronic kidney disease and erectile dysfunction. These three predictors were sparse and are diagnoses that can have common cause in underlying atherosclerosis. Over the course of the study, the UK was introducing various pneumococcal vaccines against different serotypes, and for different populations (for over 65's from 2003 for example), so were tightly correlated with age. I did not include pneumococcal vaccination as a predictor.[64] *Haemophilus* vaccination is only given to children and so was excluded.[64] Non-steroidal anti-inflammatory drugs are widely used without prescriptions during respiratory infections and therefore prone to misclassification bias, so were excluded. Dementia was the only variable excluded based on an odds ratio, which was 1.00.

3.4.10.4 Model development

For both regression models I used logistic regression, with fractional polynomials to model continuous variables, and did not use stepwise variable selection algorithms.

Because external validation was planned, I used apparent performance for the internal validation.[118] To do this I estimated the measures using the whole of the derivation dataset, without splitting, bootstrapping or other adjustments for overperformance or optimism.[171] For internal calibration, I estimated apparent calibration and discrimination in the derivation cohort and used these primarily to check the models.

I estimated the C statistic, observed to expected ratios, and drew calibration curves for internal validation. The C statistic (concordance statistic) came from the development of radar.[172] Plotting the false positive proportion on the x axis against the true positives on the y axis gave a receiver operating characteristic (ROC) curve, and the area under it was the C statistic.[114,173] A C statistic of one is perfect and 0.5 is the proportion achieved by random guesses. I tried to calculate this directly, but it was impossible due to the size of the dataset. It is calculated by establishing all possible pairs of patients where one had the event and one did

not, then finding the proportion of these where the lower risk patient was the one without the event. The number of possible pairs increased nonlinearly with of the number of individuals, so became too computationally expensive. Instead, I split the datasets into random groups, calculated the C statistic in each group, and undertook meta-analyses of the results.

I used observed to expected ratios to examine the performance of the model overall and calculated this as the number of outcomes observed divided by the number predicted by the model. Ideal performance would have been one, ratios greater than one describe underprediction of events by the model, and numbers less than one overprediction.[174]

I used calibration curves to display the level of over or under prediction at different levels in the range of predicted probabilities.[174] I made the plots by dividing the predicted probabilities into quantiles (fiftieths, or where there were fewer levels, deciles). These groups I plotted on a graph with the expected proportion (the mean prediction in the group) on the x axis and the observed proportion of events in each group on the y axis. A perfectly calibrated model would have aligned these groups on the line y=x, which I also plotted. I also included a cubic spline fitted to the points to aid interpretation.[175]

3.5 Results

3.5.1 Study population characteristics

The derivation dataset comprised 3,789,293 patients with first acute respiratory infections (Table 2: Characteristics of patients by CVD event outcome status in derivation dataset). Of these 63% (2,393,312) were URTIs, 35% (1,312,569) were LRTIs and 2.1% (78,412) were pneumonia. In addition, 189,567 (5%) had influenza-like illness codes. There were 11,996 cardiovascular events in the following 28 days (0.3%) and the outcome included 1,441 CVD deaths. The variables with the most missing data were Cholesterol : HDL ratio of which 40.6% was missing and BMI, which was 45.3% missing. Systolic BP was 15.8% missing. The most deprived decile had 12% of the patients who had CVD events, and 9% of those without CVD events. IMD data was missing for 2.6% (97,941). The mean age was 56.5 years (standard deviation, SD, 13.7) but patients who went on to have CVD had a mean age of 75 years (SD 13.7). Mean cholesterol: HDL ratio was the same in both groups (3.9, SD 1.3). Never smokers were 36% of the total population and 29% of the population who had CVD. Pneumonia was found in 2.0% of those who did not have events and 22.5% of those who had CVD.

Table 2: Characteristics of patients by CVD event outcome status in derivation dataset

Variable type	Total		No CVD outcome		CVD outcome	
Continuous	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Age in years	56.5	13.7	56.4	13.6	75.0	13.7
Cholesterol : HDL ratio	3.9	1.3	3.9	1.3	3.9	1.3
Systolic blood pressure mmHg	131.2	17.1	131.2	17.1	138.0	20.2
BMI KgM^{-2}	27.8	5.8	27.8	5.8	27.0	6.2
Categorical	**n**	**%**	**n**	**%**	**n**	**%**
Total	3,789,293	100%	3,777,297	99.7%	11,996	0.3%
Female	2,185,255	57.7%	2,179,316	57.5%	5,939	49.5%
URTI	2,398,312	63.3%	2,395,640	63.2%	2,672	22.3%
LRTI	1,312,569	34.6%	1,305,943	34.5%	6,626	55.2%
Influenza	189,567	5.0%	188,851	5.0%	716	6.0%
Pneumonia	78,412	2.1%	75,714	2.0%	2,698	22.5%
Smoking status:						
Never smoked	1,359,179	35.9%	1,355,722	35.8%	3,457	28.8%
Ex-smoker	860,961	22.7%	857,629	22.6%	3,332	27.8%
Light smoker (<10/day)	131,370	3.5%	130,899	3.5%	471	3.9%
Moderate smoker (11-19/day)	170,701	4.5%	170,211	4.5%	490	4.1%
Heavy smoker (20+/day)	122,481	3.2%	122,102	3.2%	379	3.2%
Smoker, amount unknown	285,596	7.5%	284,684	7.5%	912	7.6%
Smoking data missing	859,005	22.7%	856,050	22.6%	2,955	24.6%
Diabetes Mellitus	300,963	7.9%	298,905	7.9%	2,058	17.2%
Heart failure	36,325	1.0%	35,324	0.9%	1,001	8.3%
Chronic heart disease	49,348	1.3%	48,793	1.3%	555	4.6%
Atrial arrhythmia	78,807	2.1%	77,480	2.0%	1,327	11.1%
Markers of atherosclerosis	289,681	7.6%	287,102	7.6%	2,579	21.5%
Cancer	123,686	3.3%	122,875	3.2%	811	6.8%
Family history of CVD	26,370	0.7%	26,298	0.7%	72	0.6%
Anticoagulants	65,776	1.7%	64,900	1.7%	876	7.3%
Antiplatelets	15,572	0.4%	15,198	0.4%	374	3.1%

Antihypertensives	747,893	19.7%	742,457	19.6%	5,436	45.3%
Rheumatoid arthritis	48,098	1.3%	47,779	1.3%	319	2.7%
Statins	424,842	11.2%	422,107	11.1%	2,735	22.8%
Platelets x 10^9/L						
Unknown platelets	2,654,098	70%	2,647,469	69.9%	6,629	55.3%
Thrombocytopenia (<150)	34,487	0.9%	34,176	0.9%	311	2.6%
Platelets normal (150-450)	1,078,154	28.5%	1,073,325	28.3%	4,829	40.3%
Thrombocytosis (>450)	22,554	0.6%	22,327	0.6%	227	1.9%
CRP mg/L						
CRP unknown	3,510,766	92.6%	3,500,121	92.4%	10,645	88.7%
CRP <5	139,785	3.7%	139,380	3.7%	405	3.4%
CRP 5 to <20	109,479	2.9%	108,943	2.9%	536	4.5%
CRP >=20	29,263	0.8%	28,853	0.8%	410	3.4%
Index of Multiple deprivation:						
Most deprived decile	336,655	8.9%	335,215	8.8%	1,440	12%
IMD missing	97,941	2.6%	97,810	2.6%	131	1.1%

CVD = composite outcome (myocardial infarction, coronary syndromes, transient ischaemic event, stroke, ischaemic cardiomyopathy) in the 28 days following respiratory infection. All patients have a Respiratory Tract Infection (RTI) categorised into upper (URTI), lower (LRTI) and pneumonia. Influenza is a separate, non exclusive category and can be included separately in addition to LRTI (default for influenza, unless coded to another site). Heart failure includes all non-ischaemic diagnoses. Chronic heart disease incluses valvular disease, hypertensive disease and congenital disease. Atrial arrhythmias include atrial tachicardias, atrial fibrillation, and flutter. Diabetes mellitus includes type i, type ii, and other/unrecorded type. Markers of atherosclerosis includes: chronic kidney disease, peripheral arterial disease, and erectile dysfunction. Cholesterol : HDL ratio is serum total cholesterol/serum HDL cholesterol, 40.6% was missing. BMI was 45.3% missing. Systolic BP was 15.8% missing.

3.5.2 Model development

The regression models comprised the clinical variables detailed in Table 3: Variables included in models, and contributions to infection-related CVD event risk prediction (for model equations including categories and fractional polynomial transformation of variables see appendix section 9.2.7). The variable with the strongest association with increased CVD risk in models one and two was pneumonia (Odds ratio 10.59, 95% CI 9.98 to 11.24, in Model One and 9.82, 95% CI 9.21 to 10.46 in Model Two respectively). Other variables strongly associated with increased CVD risk were lower respiratory tract infection (OR 2.59, 95% CI 2.48 to 2.71, and 2.50 95% CI 2.38 to 2.63, respectively) and age (OR for each year increase 1.07, 95% CI 1.07 to 1.07, and 1.06 95% CI 1.06 to 1.07 respectively). Most variables were associated with an increased predicted risk of CVD events to a lesser extent. Some variables reduced the predicted probabilities: influenza diagnosis lowered predicted risk in both models (OR 0.84, 95% CI 0.78 to 0.91, and 0.85, 95% CI 0.78 to 0.92, respectively). In Model Two, CRP less than 5 mg/L was associated with reduced risk, compared to the baseline of unknown CRP (OR 0.86, 95% CI 0.77 to 0.96). In Model Two there were associations with increased risk from the use of antiplatelets (OR 2.12, 95% CI 1.03 to 2.38), and anticoagulants (OR 1.13, 95% CI 1.03 to 1.24).

Table 3: Variables included in models, and contributions to infection-related CVD event risk prediction

Variables	**Model**	
Continuous	Model One OR (95% CI)	Model Two OR (95% CI)
Age*	1.07 (1.07 to 1.07)	1.06 (1.06 to 1.07)
Cholesterol : HDL ratio*		1.10 (1.08 to 1.13)
Systolic blood pressure*		
Fractional polynomial term one		3.67×10^{-6} (3.99×10^{-8} to 3.33×10^{-4})
Fractional polynomial term two		406.59 (58.56 to 2822.97)
BMI*		
Fractional polynomial term one		0.12 (0.07 to 0.12)
Fractional polynomial term two		1.08 (1.06 to 1.11)
Categorical or binary		
RTI diagnosis		
URTI	Reference	Reference
LRTI	2.59 (2.48 to 2.71)	2.50 (2.38 to 2.63)
Pneumonia	10.59 (9.98 to 11.24)	9.82 (9.21 to 10.46)
Separate influenza diagnosis	0.84 (0.78 to 0.91)	0.85 (0.78 to 0.92)
Smoking status:		
Never smoked	Reference	Reference
Ex-smoker	1.24 (1.19 to 1.30)	1.14 (1.08 to 1.20)
Light smoker (<10/day)	1.69 (1.54 to 1.87)	1.48 (1.32 to 1.66)
Moderate smoker (11-19/day)	1.70 (1.52 to 1.89)	1.50 (1.36 to 1.65)
Heavy smoker (20+/day)	1.90 (1.70 to 2.12)	1.60 (1.44 to 1.79)
Smoker, amount unknown	1.61 (1.50 to 1.74)	1.44 (1.33 to 1.56)
Diabetes	1.49 (1.42 to 1.57)	1.32 (1.25 to1.40)
Heart failure	1.92 (1.79 to 2.05)	1.67 (1.55 to 1.81)
Atherosclerosis marker	1.21 (1.15 to 1.26)	1.04 (0.99 to 1.10)
Male sex		1.43 (1.37 to 1.50)
Chronic heart disease		1.30 (1.18 to 1.43)
Atrial arrhythmia		1.19 (1.10 to 1.29)
Cancer		0.93 (0.86 to 1.00)
Family history of CVD		1.28 (1.00 to 1.65)
Anticoagulants		1.13 (1.03 to 1.24)
Antiplatelets		2.12 (1.03 to 2.38)
Antihypertensives		1.30 (1.24 to 1.36)
Rheumatoid arthritis		1.20 (1.06 to 1.36)
Statins		1.15 (1.09 to 1.22)
Platelets x 10^9/L		
Unknown platelets		Baseline
Thrombocytopenia (<150)		1.23 (1.08 to 1.39)
Platelets normal (150-450)		1.10 (1.06 to 1.16)
Thrombocytosis (>450)		1.32 (1.14 to 1.53)
CRP mg/L		
CRP unknown		Baseline
CRP <5		0.86 (0.77 to 0.96)
CRP 5 to <20		1.10 (1.06 to 1.16)

CRP >=20	1.32	(1.14 to 1.53)
Index of Multiple Deprivation**		
Least deprived decile	Baseline	
Most deprived decile	1.52	(1.40 to 1.66)

*Variable tranformed with fractional polynomials, so not directly interpretable: Age = age in years -56.5. **model for each decile presented in supplementary materials. Models predict CVD = composite outcome (myocardial infarction, coronary syndromes, transient ischaemic event, stroke, ischaemic cardiomyopathy) in the 28 days following respiratory infection diagnosis. All patients have a Respiratory Tract Infection (RTI) categorised into upper (URTI), lower (LRTI) and pneumonia. Influenza is a separate, non exclusive category and can be included separately in addition to LRTI (default for influenza, unless coded to another site). Heart failure includes all non-ischaemic diagnoses. Chronic heart disease incluses valvular disease, hypertensive disease and congenital disease. Atrial arrhythmias include atrial tachicardias, atrial fibrillation, and flutter. Diabetes mellitus includes type i, type ii, and other/unrecorded type. Markers of atherosclerosis includes: chronic kidney disease, peripheral arterial disease, and erectile dysfunction.

3.5.3 Internal validation

3.5.4 Apparent internal calibration

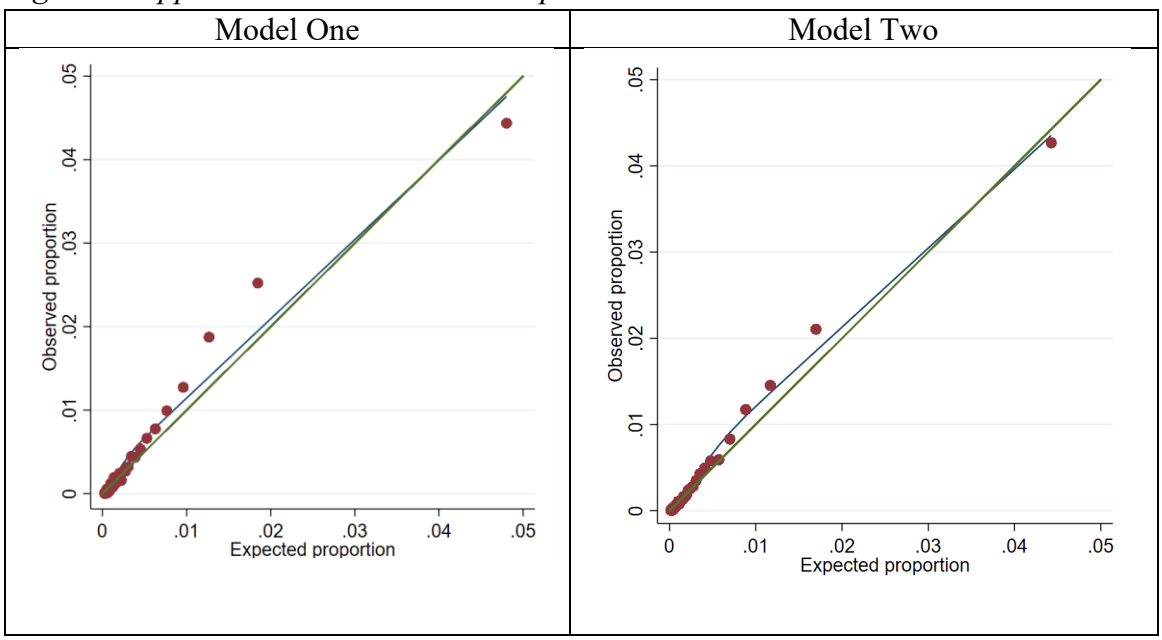
The observed to expected ratios were close to one, with slight overprediction from Model One (0.98, IQR 0.98 to 0.98), but underprediction from Model Two (1.09, IQR 1.09 to 1.09) (Table 4: Apparent observed to expected ratios). The apparent calibration plots demonstrate good performance, with predicted risks being very similar to the observed risks (Figure 2: Apparent internal calibration plots).

Table 4: Apparent observed to expected ratios

Model	Observed to expected ratio, median (IQR)*
Model One	0.98 (0.98 to 0.98)
Model Two	1.09 (1.09 to 1.09)

Median and IQR are across the multiple imputation datasets, five datasets for Model One and ten for Model Two Model One had five clinical variables, Model Two had 20 clinical variables as described in table three.

Figure 2: Apparent internal calibration plots



Internal apparent calibration plots for two models of infection-related CVD events. Proportions expected and observed in groups of predicted risk (red markers). Green line – expected equals observed. Blue line – Cubic spline based on red markers. Groups are 50ths of expected risk. Model One has five clinical variables, Model Two has 20 clinical variables as described in table three.

3.5.5 Apparent internal discrimination

Both models showed the same discrimination performance to two decimal places, with median C statistics of 0.88 (IQR 0.88 to 0.88 with a difference only at further decimal places) (Table 5: Apparent internal discrimination).

Table 5: Apparent internal discrimination

Model	Apparent C statistic, median (IQR)
Model One	0.88 (0.88 to 0.88)
Model Two	0.88 (0.88 to 0.88)

C statistic = concordance statistic. Median and IQR are across the datasets made by multiple imputation, five datasets for Model One and ten for Model Two. Model One has five clinical variables, Model Two has 20 clinical variables as described in table 3.

3.6 Discussion

3.6.1 Principal findings

This chapter describes the derivation of novel prediction models for post-acute respiratory infection CVD event risk in a population of patients aged 40 years or older. Age and pneumonia were the variables that conferred the greatest risk of CVD events. Internal (apparent) discrimination and calibration were satisfactory for both models, indicating the models performed well in the datasets used to create them.

3.6.2 Comparison with prior studies

Prior research has shown individual variables to be associated with infection-related CVD, but these studies were not seeking to predict risk, and clinicians should not be tempted to use these associations to guide treatment decisions.[114] Studies that identify risk factors rather than assessing predictive performance cannot be used for prediction - even strong associations make poor predictors and can behave in unexpected ways.[114] Another problem is the comparator is different in different types of study. In this study we are stratifying risk in people who present with a respiratory infection, whereas most studies are attempting to measure the risk associated with infections (or another risk factor) compared to not having the infection. As an example, influenza infection is a risk factor for CVD, when compared to being well.[1] There is very strong evidence for this association, it has been shown in systematic reviews of case-control studies, and self-controlled case series, and a protective effect has been shown in systematic reviews of influenza vaccine trials.[1,29] This evidence is one of the bases for this thesis. Nevertheless, in this study, among people who are all presenting with respiratory infections, a diagnosis of influenza or influenza-like illness decreased the predicted risk of CVD. This individual variable cannot be interpreted – it could be for any number of reasons – for example because their risk is being described by the other respiratory tract infection variable, unmeasured associations, or because clinicians use

influenza-like illness to mean a less severe illness – someone unwell but without chest signs for example.[71]

This study gives further examples of how causal relationships do not necessarily operate as predictors in the way one might expect. Very strong trial evidence shows antiplatelets and anticoagulants reduce CVD events.[8,176] A clinician who is reassured that their patient has low risk because they are taking this medication would be making an error. Whilst the patient's risk is lower than if they were not taking the medication, Model Two showed that in our population these patients are at increased risk compared to people who are not taking these medications. Antiplatelets increase the predicted risk over and above their protective effect, probably because the people taking them have higher underlying risk of CVD.

Pneumonia is an important variable in the models and is known to be associated with CVD events.[42,104,164] An observational study examining the effect of aspirin on CVD events in pneumonia suggests a protective effect, as did a small trial.[103,104] Pneumonia could be used in this way as a diagnostic tool because the modelling shows it identifies high risk individuals, however a large proportion of post-infection CVD events occur in people with other diagnoses (78% of events in this cohort). This means that it would miss most of the people who go on to have CVD events.

Prior CVD risk prediction studies have not targeted post-infection CVD, nor included infection-related variables.[7] NICE guidance has not advised acute infection-related risk prediction. Instead, they advise GPs in the UK to estimate overall ten-year CVD risk using the QRISK prediction models.[10,108] These models were developed and validated in UK primary care datasets, with comparable methods to this study. QCovid™ is another prediction tool developed using routinely collected data, it predicts risk of hospitalisation or death following Covid-19, but it is not specific to cardiovascular disease.[177]

3.6.3 Strengths and limitations of this study

There are limitations related to the data. Clinicians collected these data for a different purpose, over 20 years, during which there were changes in coding, patient behaviour, and clinical practice. I had to combine some categories of exposure and impute some data, more than 40% for two variables. I have missed acute respiratory infections that did not present to primary care or were not coded. These are the majority of infections.[4] This reduced the size of the dataset, which has remained sufficiently large. Some of the data included in the models may have changed by the time of the index date, for example blood pressure and BMI measurements. These limitations are shared with other CVD risk prediction models that are developed using routinely collected data. The results are only generalizable to presentations recorded in primary care, which is the population for which it is intended.

The event rate is relatively low, and so the C statistic is less informative than it would be in a less rare outcome (because simply predicting near zero probability for each person would also have a high C statistic) and this contributes to the C statistics for Model One and Model Two being similar (identical at two decimal places). In the next chapter, for external validation, to address this I report other metrics for assessing model performance, including the ratio of true and false negatives and positives.[173,178]

I used diagnoses recorded in clinical records. I cannot tell how pneumonia was diagnosed, for example, but it is likely that some cases have been diagnosed following chest radiography whilst others will not. Pneumonia might be more likely to be diagnosed in older patients, and diagnoses may be made to justify prescribing decisions. Clinical trials or prospective studies might instead have strict diagnostic criteria, probably including tests.[103] The reality in UK primary care is that diagnoses of respiratory infections are largely clinical. Patients who have more severe infection are also more likely to be assessed in hospital or ambulatory care settings, which may not be coded in the primary care record until discharge.[179] They would

then enter the cohort later than those with URTIs. It is also possible that patients with severe infections present sooner, compared to people with minor illnesses. Either way, the presentation to primary care is representative of the population in which the model is designed should be used.

I have used the first respiratory infection presenting to primary care after the age of forty years, but it is possible that more presentations with infections change the risk. The short follow-up period is a strength in that it means the recorded CVD events are more likely to be causally linked to the index infection, as was found in prior epidemiological studies, and minimised competing events.[2] However, I will also have missed some events as some studies with a longer follow-up after infections found a longer tail of increased risk.[1,104] If the follow-up became too long the risk would revert the background primary prevention risk, which would require different methods and is already extensively modelled.[7]

3.6.3.1 Clinical implications

I selected variables with physicians' expert opinion. This is necessarily a biased process, as clinicians bring their own subjective views and experiences. A corresponding strength is that it resulted in models that have face validity with the target users, clinicians. However, these models cannot be used in clinic or research without external validation, which is the subject of the next chapter.

3.6.4 Conclusions

I have derived two models for predicting risk of primary CVD events in acute respiratory infections. They apply to primary care populations over the age of 40 and have acceptable internal calibration.

These new tools could potentially identify individuals who are most likely to have an infection-related cardiovascular event. However, all the potential uses depend on adequate external validation, which is the subject of the next chapter.

4 Chapter Four: Prediction model external validation: a retrospective cohort study

"*Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena.*"

George Box, 1976 [124]

4.3 Introduction

4.3.1 Overview within thesis – how this chapter fits in

The previous chapter describes the process of deriving two statistical models in data from CPRD Aurum. These were designed to predict the 28-day risk of primary cardiovascular events following a patient's presentation to primary care with respiratory infection. They apply to individuals over the age of forty years.

The usefulness of predictive models depends on how well they perform in their intended population, and so this chapter sets out how I assessed the models in external data. First, I examined their external calibration, and external discrimination. I then used Model One to derive a clinical prediction score, the DASHI score (Diabetes Age Smoking Heart failure Infection type), which I also externally validated. I then performed diagnostic accuracy measurements using thresholds which could be used as clinical cut-offs. I also performed clinical utility analyses using the external validation population.

The following chapters, five and six, use the prediction modelling work to identify higher risk populations in which to examine potential interventions to prevent infection-related cardiovascular events.

4.3.2 Background: serious diseases have low incidence in primary care

Infection-related cardiovascular events are rare in primary care. In the previous chapter I reported that one in 133 people with respiratory infections went on to have an outcome event in the derivation cohort. This low prevalence of serious disease is a practical problem shared with most presentations to primary care. Buntinx and colleagues discussed this problem.[180] As an example, for males over the age of 80 presenting with rectal bleeding only 1 in 22 end up with colorectal cancer diagnoses.[181] CVD and cancers are some of the commonest causes of death but are rare diagnoses in the primary care population.[79] This project uses the strategy

of developing diagnostic algorithms, which could be used to initiate treatment. Potential interventions could include things used for other CVD prevention scenarios, short courses of antiplatelets, blood pressure control, statin therapy, safety-netting, and vaccinations. [10,19,91,98]

An inherent property of diagnostic algorithms is the trade-off between the ideal high sensitivity (not missing serious disease) and low specificity, which leads to over-referral or over-treatment. Infection-related CVD is uncommon, but potential preventative actions could be performed in primary care without overwhelming secondary care with referrals because primary care already implements most of the possible interventions for primary CVD event prevention.

All treatment strategies have costs as well as benefits. The risks of a treatment contribute to deciding the amount of overtreatment that is acceptable. Extreme examples can illustrate this principle. Cytotoxic chemotherapy is a very harmful treatment. It is not given to all the patients with localised breast cancer. Instead, surgical specimens are sent to the USA for tumour profiling tests to identify people at higher risk of progression, in whom the benefit outweighs the risk.[182] In contrast, because it is very safe, fluoride is added to the water supply to prevent dental caries.[183]

Understanding the risk of an event is also important for research.[118] It allows identification and stratification of populations by event risk. This can be useful as inclusion criteria (I used the modelling to identify populations in the following chapters), or for comparisons between higher and lower risk populations.[28,118]

4.3.3 Rationale

All treatments have both benefits and harms, and targeting treatments can improve this balance. Before a tool is used to identify people at higher risk of serious disease there should

be an assessment of its diagnostic performance to assess these trade-offs. This is the purpose of the external validation in this chapter.

GPs need very rapid tools to enable assessment in pressured clinical situations. It is possible to implement regression models in the clinical software, if they have been cleared for use by the UK Medicines and Healthcare products Regulatory Agency (MRHA).[133] An alternative is to derive a points-based clinical prediction tool.[119,184] In this chapter I used the validation results to select a model to be converted into a clinical points score, the DASHI score. I then externally validated this score.

Clinical decisions are based on some risk threshold, over which an action can be justified.[134] To aid these decisions I did further analyses in the external validation data, using thresholds for action in both modelled probabilities and DASHI scores. This gave estimates of diagnostic accuracy including sensitivity, specificity, likelihood ratios, and predictive values, and the number of people per 100,000 people expected to have true and false negatives and positives.

To allow visual assessment of the trade-off between true and false positives I compared the net benefit of the models and DASHI score with decision curve analyses in the external validation population.[134]

4.4 Methods:

4.4.1 Aim

The aim of the work presented in this chapter was to develop and validate a tool that could be used to predict infection-related CVD clinically and for research.

4.4.2 Chapter Four objectives

1. To externally validate Models One and Two
2. To derive a clinical points score from the statistical modelling
3. To externally validate the points score
4. To compare the score and statistical models' clinical utility using decision curve analysis

4.4.3 Brief overview of cohort design and validation methods

To achieve the validation objectives one and three I used methods described in the previous chapter. The data source, definitions, and outcomes were also the same as described previously, only using CPRD Gold data for external calibration rather than CPRD Aurum. I have therefore only included a brief overview of these methods in this chapter.

I used separate retrospective observational cohorts to derive and externally validate the statistical models and prediction score. The cohort described in the previous chapter was extracted from CPRD Aurum and is used for the internal apparent validation of the DASHI score. The cohort used for external validation is extracted from CPRD Gold. The data were different, but to ensure comparable datasets the processes for extracting each cohort were analogous. The methods were described in the previous chapter.

4.4.4 Population

The different databases use different coding systems, and the codes I used are online (https://github.com/Protocols-For-Research/CPRD-codes-CVD-infection-risk). I designed

these code lists to identify populations as close to each other as possible in terms of search strategies, conditions coded, and prevalence of the variables in each dataset. I developed the searches simultaneously. I first identified codes with a search in one coding system, before refining them. I then tried the refined search in the other coding system, and updated the terms again according to the results, before repeating it in the first coding system. I followed this iterative process until the code list results were stable. On applying the code lists to the datasets, if the prevalence was markedly different between the datasets, I revisited the searches and refined the code lists further.

The population definition is described in the previous chapter (section 3.4.5). As a reminder, patients became eligible for inclusion when they reached 40 years of age, with no upper age limit (Previous chapter, Figure 1: Patient timeline in cohorts). Their index date, when they entered the cohort, was the date of the first time after they turned 40 that they had a coded diagnosis of acute respiratory infection. People with less than one year of data before the index date or prior CVD events were excluded. I followed patients for 28 days from the index date.

4.4.5 Outcome

The outcome was the same composite of cardiovascular outcomes as for the model derivation. It included new myocardial ischaemia (myocardial infarction, angina, acute coronary syndromes, or ischaemic cardiomyopathy), new cerebrovascular events (stroke, and TIA), and deaths from these causes. I used primary and secondary care clinical codes, and ONS mortality records to identify outcomes.

4.4.6 Statistical analyses

I used Stata versions 17 and 18 for statistical analyses [Stata Corp, College Station, TX]. I calculated baseline descriptive statistics in the validation cohort according to outcome status

in the same way as the derivation cohort. I used estimated means and standard deviations for continuous variables, and numbers and percentages for categorical variables.

4.4.6.1 Missing data

I used the same methods in each cohort. These are described in the previous chapter, so I will summarise here. If there were no codes for binary variables prior to the index date, the disease or prescription was considered absent.

For continuous and categorical variables with missing data, I used multiple imputation and missing indicator methods. I imputed continuous variables after log transformation. Smoking status and IMD deciles were imputed as ordinal variables. I used five imputations for Model One, and ten for Model Two, with chained equations (MICE) with Stata command *mi impute*.[148]

Missing indicator methods were used for recent blood tests (platelets, and C reactive protein), using categorical variables based on the reference ranges.

4.4.6.2 Validation methods

The same methods were used for internal and external validation, only applied to different datasets. For external performance, the external dataset was used. For objective two I also performed internal apparent calibration on the score.

I applied the models and score to the individuals in each dataset, after imputing missing values, to estimate predicted probabilities. I used these predictions to estimate C statistics, expected/observed ratios, and draw calibration plots (see chapter three). I calculated measures in each imputed dataset and combined them with Rubin's rules where appropriate and report median predicted probabilities and expected to observed ratios over the imputed datasets.

As described in the previous chapter, deriving C statistics had high time complexity (the computational time required increases non-linearly with the number of patients). This meant

it was impossible for me to calculate C statistics directly in these large populations. Instead, I divided each imputed dataset into 20 random subsets and derived the C statistic in each. I then used random effects meta-analysis to combine these results to get the overall C statistic for each imputed dataset (the default was random effects - there was no heterogeneity so the results would be identical with fixed effects models).

4.4.6.3 Objective two: clinical risk score development

To choose a statistical model to convert to a clinical points score I considered calibration and classification performance, and parsimony. This meant choosing the best model, or if they had the same performance, choosing the simpler model.

To derive a points score I specified a scaling factor such that two points in the score were equivalent to the risk from being 20 years older (the same approach as used in CHA_2DS_2-VASc).[119] I categorised age into 20-year periods and used the midpoints to calculate risk due to age. I then calculated the risk attributed to other variables in points on this scale, and rounded points to the nearest point.[119] I calculated predicted probabilities for the full range of possible scores.[119]

4.4.6.4 Objective four: assessing clinical utility

I examined the clinical utility of the statistical models and score using net benefit analysis.[134] Net benefit is calculated by subtracting scaled false positives from true positives, at each predicted probability. The scaling of false positives is necessary to put them on the same scale as the true positives. Scaling is determined by the level of risk a clinician or patient would be willing to take. This is the amount of unnecessary treatment that would be acceptable (i.e. the number of false positives treated) for each necessary treatment (in true positives), so varies according to the treatment being considered. This threshold is the x axis, and different diagnostic strategies can be compared by plotting their net benefit on the y axis. Net benefit analysis includes plotting the default strategies of treating everyone, or no one

(reflecting current practice in infection-related CVD). One can then use the graph by first deciding what level of overuse one would accept for a given treatment, find this point on the x axis, and find the diagnostic strategy with the highest net benefit at that threshold.

4.4.6.5 Diagnostic test performance measures – post-hoc analyses

To aid comparisons between each of the models and the score I applied probability thresholds as if they were to be used for a clinical decision, and calculated numbers of patients with true and false positives and negatives, above the threshold per 100,000 people. As the ratio of true to false results is the basis of clinical utility analyses, I also give these ratios. These are also helpful in situations with rare outcomes, where the C statistic can be flattering.[173] As previous studies have concentrated on people with pneumonia as a high-risk group, I evaluated the performance of this single covariate as a predictor of acute CVD events.[104] I also calculated diagnostic performance characteristics for the DASHI score in the external data. To give statistics that may be familiar to clinicians, and to demonstrate the effect of using different thresholds, I calculated sensitivity, specificity, positive and negative likelihood ratios, and predictive values for the DASHI score.[119]

4.5 Results

4.5.1 Study population characteristics

The validation dataset comprised 2,630,113 patients. The mean age was 56.7 years (SD 13.6) and 6,868 (0.3%) had CVD events in the 28-day follow-up period (Table 6: Characteristics of patients by CVD event outcome status in the validation dataset). The patients with CVD events were older than those who did not have events (74.7 years SD 13.4 vs 56.6 years SD 13.5) with higher systolic blood pressure (140.7mmHg, SD 20.3) but had lower total cholesterol: HDL cholesterol ratios (3.2, SD 0.8 Vs 3.6, SD 1.7). They had a higher use of statin therapy (16.8% Vs 9.2%), antihypertensives (42.2% Vs 18.8%), antiplatelets (2.8% Vs 0.4%) and anticoagulants (5.5% Vs 1.6%). Fewer people with CVD were never smokers (25.7% Vs 33.1%) or female (50.1% Vs 57.9%) or had a family history of CVD (0.1% Vs 0.2%). CVD cases had high prevalence of conditions such as diabetes (14.3% Vs 6.1%), atrial arrythmias (9.6% Vs 1.9%), markers of atherosclerosis (17.2% Vs 6.6) and cancers (5.4% Vs 2.7%). CVD cases were also more likely to have had their platelets and CRP measured, and these were more likely elevated. A difference was seen in the types of respiratory infection – CVD cases had similar levels of influenza-like illness to non-cases (5.8% Vs 5.2%) but higher diagnoses of LRTI (52.0% Vs 34.7%) and pneumonia (23.9% Vs 1.8%).

Table 6: Characteristics of patients by CVD event outcome status in the validation dataset

Variable type	Total		No CVD outcome		CVD outcome	
Continuous	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Age in years	56.7	13.6	56.6	13.5	74.7	13.4
Cholesterol to HDL ratio	3.6	1.7	3.6	1.7	3.2	0.8
Systolic BP mmHg	132.0	17.5	132.0	17.5	140.7	20.3
BMI KgM^{-2}	27.7	5.7	27.7	5.7	27.0	5.9
Categorical	**n**	**%**	**n**	**%**	**n**	**%**
Total	2,636,981	100%	2,630,113	99.7%	6,868	0.3%
Female	1,530,454	58.0%	1,527,014	57.9%	3,440	50.1%
URTI	1,669,855	63.3%	1,668,200	63.3%	1,655	24.1%
LRTI	918,981	34.8%	915,408	34.7%	3,573	52.0%
Influenza	138,216	5.2%	137,818	5.2%	398	5.8%
Pneumonia	48,145	1.8%	46,505	1.8%	1,640	23.9%
Smoking status:						
Never smoked	875,671	33.2%	873,908	33.1%	1,763	25.7%
Ex-smoker	426,837	16.2%	425,609	16.1%	1,228	17.9%
Light smoker (<10/day)	118,481	4.5%	118,175	4.5%	306	4.5%
Moderate smoker (11-19/day)	184,054	7.0%	183,645	7.0%	409	6.0%
Heavy smoker (20+/day)	155,509	5.9%	155,108	5.9%	401	5.8%
Smoker, amount unknown	160,415	6.1%	159,899	6.1%	516	7.5%
Smoking data missing	716,014	27.2%	713,769	27.1%	2,245	32.7%
Diabetes	161,397	6.1%	160,414	6.1%	983	14.3%
Heart failure	27,973	1.1%	27,408	1.0%	565	8.2%
Chronic heart disease	30,684	1.2%	30,438	1.2%	246	3.6%
Atrial arrhythmia	50,838	1.9%	50,181	1.9%	657	9.6%
Markers of atherosclerosis	174,204	6.6%	173,022	6.6%	1,182	17.2%
Cancer	70,839	2.7%	70,470	2.7%	369	5.4%
Family history of CVD	4,339	0.2%	4,330	0.2%	9	0.1%
Anticoagulants	42,357	1.6%	41,979	1.6%	378	5.5%
Antiplatelets	10,815	0.4%	10,623	0.4%	192	2.8%
Antihypertensives	498,441	18.9%	495,540	18.8%	2,901	42.2%
Rheumatoid arthritis	34,223	1.3%	34,040	1.3%	183	2.7%
Statins	244,312	9.3%	243,156	9.2%	1,156	16.8%

Platelets x 10^9/L						
Unknown platelets	2,023,120	76.7%	2,018,449	76.5%	4,671	68%
Thrombocytopenia (<150)	18,250	0.7%	18,108	0.7%	142	2.1%
Platelets normal (150-450)	583,543	22.1%	581,591	22.1%	1,952	28.4%
Thrombocytosis (>450)	12,068	0.5%	11,965	0.5%	103	1.5%
CRP mg/L						
CRP unknown	2,481,348	94.1%	2,475,011	93.9%	6,337	92.3%
CRP <5	76,949	2.9%	76,818	2.9%	131	1.9%
CRP 5 to <20	62,726	2.4%	62,485	2.4%	241	3.5%
CRP >=20	15,958	0.6%	15,799	0.6%	159	2.3%
Index of Multiple deprivation:						
Most deprived decile	92,357	3.5%	91,971	3.5%	386	5.6%
Deprivation missing	1,310,635	49.7%	1,307,710	49.6%	2,925	42.6%

CVD = composite outcome (myocardial infarction, coronary syndromes, transient ischaemic event, stroke, ischaemic cardiomyopathy) in the 28 days following respiratory infection. All patients have a Respiratory Tract Infection (RTI) categorised into upper (URTI), lower (LRTI) and pneumonia. Influenza is a separate, non exclusive category and can be included separately in addition to LRTI (default for influenza, unless coded to another site). Heart failure includes all non-ischaemic diagnoses. Chronic heart disease incluses valvular disease, hypertensive disease and congenital disease. Atrial arrhythmias include atrial tachicardias, atrial fibrillation, and flutter. Diabetes mellitus includes type i, type ii, and other/unrecorded type. Markers of atherosclerosis includes: chronic kidney disease, peripheral arterial disease, and erectile dysfunction. Cholesterol : HDL ratio is serum total cholesterol/serum HDL cholesterol.

4.5.2 Objective one: External validation of models one and two - calibration

Model One is the simpler model, Model Two includes all the variables in Model One, plus further predictors (see chapter three Table 3: Variables included in models, and contributions to infection-related CVD event risk prediction).

The observed to expected ratios showed the models were overpredicting risk, they were 0.83 (IQR 0.83 to 0.83) for Model One and 0.78 (IQR 0.77 to 0.78) for Model Two (Table 7: Observed to expected ratios for Model One and Model Two).

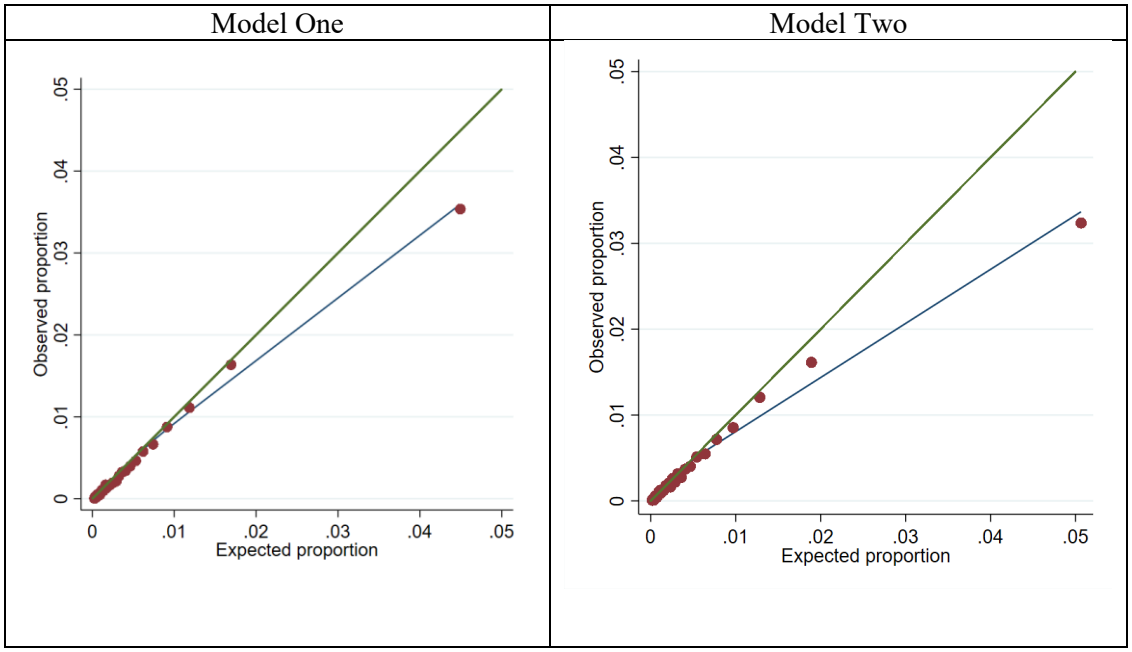
Table 7: Observed to expected ratios for Model One and Model Two

Model	Observed to expected ratio, median (IQR)*
Model One	0.83 (0.83 to 0.83)
Model Two	0.78 (0.77 to 0.78)

Median and IQR are across the multiple imputation datasets, five datasets for Model One and ten for Model Two Model One had five clinical variables, Model Two had 20 clinical variables as described in table 3.

Calibration curves showed good external calibration for both models (Figure 3: External calibration plots). However, Model Two had a slightly worse performance in both O/E ratio and the external calibration plot, which showed worse overprediction at higher predicted probabilities.

Figure 3: External calibration plots – Model One and Model Two



External calibration plots: Proportions expected and observed in groups of predicted risk (red markers). Green line – expected equals observed. Blue line – Cubic spline based on red markers. Groups are 50ths of expected risk. Model One has five clinical variables, Model Two has 20 clinical variables as described in table 3.

4.5.3 Objective one: External validation of models one and two - discrimination

Each model demonstrated excellent external discrimination with median C statistics of 0.86 (IQR 0.860 to 0.860) for Model One, 0.85 (IQR 0.849 to 0.852) for Model Two (Table 8: External validation - discrimination of models one and two). This was slightly lower than the values for internal apparent discrimination (See chapter three).

Table 8: External validation - discrimination of models one and two

Model	External C statistic, median (IQR)
Model One	0.86 (0.86 to 0.86)
Model Two	0.85 (0.85 to 0.85)

Median and IQR are across the datasets made by multiple imputation, five datasets for Model One and ten for Model Two. Model One has five clinical variables, Model Two has 20 clinical variables as described in table 3.

4.5.4 Objective two: Clinical prediction score derivation - the DASHI score

Given these performance statistics I derived a clinical score from the simpler model, Model One. The result was the DASHI score (Table 9). DASHI is an acronym of the five variables that confer points. They are Diabetes (1 point for diabetes mellitus of any type), Age (2 points for age 60-79, 4 points for 80+ years), Smoking (1 point for current smokers), Heart failure (1 point for a diagnosis), and Infection type (1 point for LRTI, 4 points for pneumonia). The influenza variable dropped out in the derivation process because it conferred less than half of one point to the score and so rounded to zero. The DASHI score predicts 28-day risks from 0.04% for zero points, to 35.6% for the maximum of 11 points (Table 10). To aid interpretation of these risks I included the ten-year risk required to achieve the same level of risk for 28 days.

Table 9: DASHI - scoring system

Variable		Points
Diabetes	No	0
	Yes	1
Age (years)	40-59	0
	60-79	2
	80+	4
Smoking	Never, or ex-smoker	0
	Current smoker	1
Heart failure	No	0
	Yes	1
Infection	Upper tract	0
	Lower tract	1
	Pneumonia	4

Table shows points awarded for clinical characteristics in a patient presenting to primary care with a respiratory tract infection over the age of 40 years

Table 10: DASHI points - predicted risk

Points scored	Predicted CVD within 28 days (%)	Equivalent if risk maintained over 10 years (%)
0	0.04	5.42
1	0.08	10.14
2	0.16	18.53
3	0.30	32.48
4	0.58	52.86
5	1.10	76.28
6	2.09	93.58
7	3.93	99.46
8	7.28	99.99
9	13.09	99.99
10	22.40	99.99
11	35.64	99.99

Table shows predicted risk of a primary CVD event in the 28 days following diagnosis with a respiratory infection in primary care by DASHI points (see previous table). Equivalent risk calculated as 130 independent 28-day periods using formula Risk = 1-(1-28 day predicted probability)^{130}

Objective three: validation of the DASHI score – internal and external calibration and discrimination

In internal apparent calibration the DASHI score had a median observed to expected ratio of 1.07 (IQR 1.07 to 1.07) and the C statistic was 0.84 (IQR 0.84 to 0.84). In external calibration the DASHI score had a median observed to expected ratio of 0.85 (IQR 0.85 to 0.85), this was the same as the C statistic (0.85, IQR 0.85 to 0.85) (Table 11: DASHI score external validation statistics).

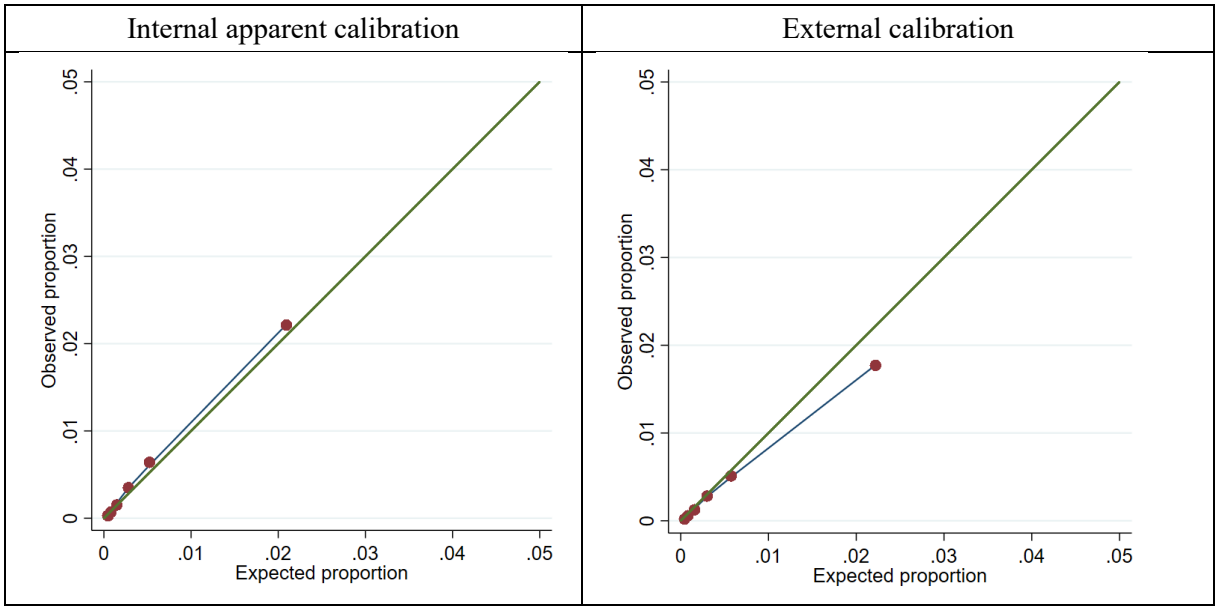
The calibration curves showed good calibration, with some over prediction in the top 10% of predicted risk in the external dataset (Figure 4: DASHI score internal and external calibration plots).

Table 11: DASHI score external validation statistics

DASHI score estimate	Internal (apparent)	External
Observed to expected ratio: median (IQR)	1.07 (1.07 to 1.07)	0.85 (0.85 to 0.85)
C statistic: median (IQR)	0.84 (0.84 to 0.84)	0.85 (0.85 to 0.85)

C statistic = concordance statistic. Median and IQR are across five datasets made by multiple imputation.

Figure 4: DASHI score internal and external calibration plots

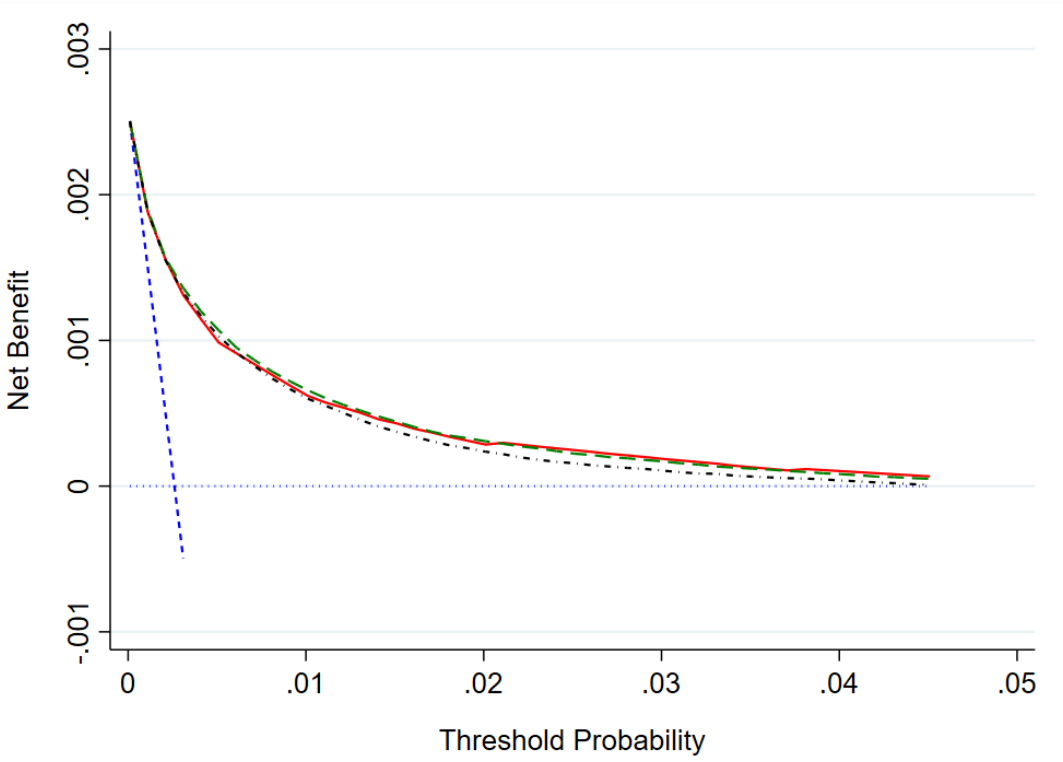


Calibration plots for the DASHI score: proportions expected and observed in groups of predicted risk (red markers). Green line – expected equals observed. Blue line joining groups – Cubic spline based on red markers. Groups are deciles of expected risk.

4.5.5 Objective four, clinical utility - decision curve analyses

Decision curve analysis showed little difference in net benefit between the two models and DASHI score (Figure 5). The two statistical prediction models and the DASHI score all outperformed the default options of assuming all or none of the patients would have CVD events.

Figure 5 Decision curves for Model One, Model Two and DASHI score



Lines are net benefit for different strategies: blue dash: 'treat none' assumes no-one will have an event (net benefit=0), bold blue dash: 'treat all' assuming everyone will have events, solid red: DASHI score, green dash: Model One, black alternating dots and dashes: Model Two. Net benefit = true positive proportion minus false positive proportion multiplied by threshold probability/1-threshold probability.

4.5.6 Discrimination at thresholds of predicted probability

For each model there is a trade-off between the threshold applied and the performance statistics, but the models are broadly comparable (Table 12: External performance: model discrimination at thresholds of predicted probability). In the validation population, which had 260 events per 100,000 patients, if one used a threshold of 0.1%, or one DASHI point (0.08%), Model One would identify 246 people with events, Model Two would identify 241, and DASHI 254. One would include 15, 19, and 6 false negatives respectively. A pneumonia diagnosis occurred in 1.8% of the population, and one could expect 3.4% of these to have CVD events. Pneumonia diagnosis would identify 62 CVD events and miss 198 per 100,000 patients. This is a similar performance to seven DASHI points (3.93% predicted risk, 51 true positives and 209 false negatives).

Table 12: External performance: model discrimination at thresholds of predicted probability

	Per 100,000 patients				Ratio	
Predicted percentage with outcome at threshold	**True positives**	**False Positives**	**False Negatives**	**True negatives**	**Ratio of false positives to true positives**	**Ratio of true negatives to false negatives**
Model One						
0.1%	246	51,796	15	47,944	211	3,258
0.2%	222	31,222	39	68,518	141	1,768
1%	127	6,023	133	93,717	47	703
2%	75	2,143	186	97,596	29	526
3%	53	1,143	207	98,597	22	474
Model Two						
0.1%	241	48,703	19	51,037	202	2,681
0.2%	218	30,109	42	69,631	138	1,656
1%	127	6,595	133	93,144	52	700
2%	77	2,608	184	97,132	34	529
3%	53	1,390	207	98,349	26	476
Pneumonia diagnosis						
3.4%	62	1,764	198	97,976	28	494
DASHI points						
1 (0.08%)	254	70,598	6	29,142	278	4,686
2 (0.16%)	240	47,280	20	52,460	197	2,605
3 (0.30%)	213	27,190	48	72,549	128	1,528
4 (0.58%)	172	14,025	88	85,714	81	972
5 (1.10%)	132	6,761	129	92,979	51	721
6 (2.09%)	84	2,651	177	97,089	32	549
7 (3.93%)	51	1,053	209	98,686	20	473
8 (7.28%)	37	647	223	99,093	17	444
9 (13.09%)	14	204	247	99,356	15	403
10 (22.40%)	2	27	259	99,712	14	386
11 (35.64%)	<1	1	260	99,738	17	383

Results including all patients with a probability or score >= threshold values at presentation to primary care with a respiratory infection. Pneumonia diagnosis used as a stand-alone one variable rule. All results calculated in the external validation data, which has an overall prevalence of primary CVD event outcome 260/100K (0.26%) in the 28 days following respiratory infection. CVD outcomes are a composite of cerebrovascular and coronary events.

The sensitivity, specificity, likelihood ratios, and predictive values of the DASHI score are presented in Table 13: Diagnostic performance measures for DASHI score over threshold of points scored.

Sensitivity was very high for one DASHI point (97.74%, 95% CI 96.08 to 99.39), but with very low specificity (27.74%, 95% CI 27.64 to 27.85). At ten DASHI points sensitivity reduced to less than one percent (0.84%, 95% CI 0.63 to 1.04) and specificity was very high at 99.97 (95% CI 99.96 to 99.97). Negative predictive values were very high (>99%) for all the possible scores. Positive predictive values ranged from 0.35% (95% 0.34% to 0.36%) for one point to 6.55% (95% CI 4.58% to 8.52%) for ten points.

Table 13: Diagnostic performance measures for DASHI score over thresholds of points scored

DASHI Points Threshold (Predicted %)	Sensitivity % (95% CI)	Specificity % (95% CI)	Positive likelihood ratio (95% CI)	Negative likelihood ratio (95% CI)	PPV % (95% CI)	NPV % (95% CI)
1 (0.08%)	97.74 (96.08 to 99.39)	27.74 (27.64 to 27.85)	1.35 (1.35 to 1.36)	NC	0.35 (0.34 to 0.36)	99.98 (99.98 to 99.98)
2 (0.16%)	92.33 (90.73 to 94.95)	52.16 (52.07 to 52.24)	1.93 (1.92 to 1.94)	0.15 (0.06 to 0.23)	0.50 (0.49 to 0.51)	99.96 (99.96 to 99.96)
3 (0.30%)	82.21 (80.68 to 83.75)	72.12 (72.06 to 72.19)	2.95 (2.93 to 2.96)	0.25 (0.20 to 0.30)	0.76 (0.74 to 0.78)	99.94 (99.93 to 99.94)
4 (0.58%)	66.54 (65.16 to 67.91)	85.54 (85.49 to 85.59)	4.60 (4.58 to 4.63)	0.39 (0.36 to 0.42)	1.19 (1.15 to 1.22)	99.90 (99.89 to 99.90)
5 (1.10%)	50.80 (49.57 to 52.03)	93.10 (93.07 to 93.13)	7.36 (7.30 to 7.42)	0.53 (0.50 to 0.55)	1.89 (1.82 to 1.95)	99.86 (99.86 to 99.87)
6 (2.09%)	33.01 (32.00 to 34.01)	97.18 (97.16 to 97.20)	11.71 (11.55 to 11.86)	0.68 (0.67 to 0.71)	2.97 (2.84 to 3.09)	99.82 (99.82 to 99.83)
7 (3.93%)	20.08 (19.30 to 20.86)	98.90 (98.89 to 98.92)	18.33 (18.04 to 18.63)	0.80 (0.80 to 0.82)	4.57 (4.33 to 4.81)	99.79 (99.78 to 99.80)
8 (7.28%)	14.32 (13.68 to 15.00)	99.35 (99.34 to 99.36)	22.00 (21.77 to 22.23)	0.86 (0.85 to 0.87)	5.43 (5.10 to 5.77)	99.78 (99.77 to 99.78)
9 (13.09%)	5.81 (5.40 to 6.23)	99.77 (99.77 to 99.78)	25.66 (25.17 to 26.14)	0.94 (0.94 to 0.95)	6.28 (5.67 to 6.88)	99.75 (99.74 to 99.75)
10 (22.40%)	0.84 (0.63 to 1.04)	99.97 (99.96 to 99.97)	26.83 (20.56 to 33.10)	0.99 (0.98 to 0.99)	6.55 (4.58 to 8.52)	99.74 (99.74 to 99.75)
11 (35.64%)	0.03 (0.006 to 0.06)	99.99 (99.99 to 1.00)	19.19 (16.08 to 22.31)	0.99 (0.99 to 1.00)	NC	99.74 (99.73 to 99.75)

Diagnostic performance of DASHI score for primary CVD events in the 28 days following presentation with respiratory infection to primary care. CVD events are a composite of cerebrovascular and coronary event diagnoses. Estimates derived in the external calibration dataset. DASHI points threshold = patients with this number of points or more on the DASHI score. NC: estimates not calculable due to low numbers.

4.6 Discussion

4.6.1 Principal findings

This chapter presents the validation of two prediction models and the DASHI risk score for estimating post-acute respiratory infection CVD event risk in a population of patients aged 40 years or older. The DASHI score showed excellent discrimination and good calibration in external validation. The performance of DASHI was very similar to the more complex models. Given this, I considered the simplicity of DASHI made it preferable for use in clinical practice.

The DASHI score could be used in primary care to estimate cardiovascular risk in patients with acute respiratory infection, for identifying cohorts for research, and potentially to target primary prevention measures.

4.6.2 Comparison with prior studies

DASHI is a novel tool for predicting post-infection CVD events.[7] Prior research has been focussed on either CVD risk prediction in other contexts, or epidemiological studies demonstrating the CVD risks associated with respiratory infection.[1,7,88,185,186] CVD risk prediction has focussed on longer term risks and has not included variables relating to infection. Prior studies have also sought to look at infection-related CVD in people with increased CVD risk: people with pneumonia or Covid-19, people with higher QRISK2 scores or risk factors for CVD [28,103,104,187].

Prediction-modelling studies estimating CVD risk are many and varied, but none appear to apply specifically to the post-infection period.[7] NICE guidance has not advised acute, infection-related risk prediction. Instead, they recommend GPs estimate overall ten-year CVD risk using the QRISK prediction models.[10,108] These models were developed and

validated in UK primary care datasets, with comparable methods to this study. There is evidence from a CPRD cohort study that people with QRISK2 CVD risk of greater than 10% are higher risk for both respiratory infections (Incidence rate ratio (IRR) 1.39 95% CI 1.37 to 1.40) and subsequent CVD events (IRR 3.65, 95% CI 3.42 to 3.89) than people with ten year risk below 10%.[28] There is a self-controlled case series and matched cohort study from population wide medical records in Sweden that provides evidence that Covid-19 increases the risk of myocardial infarction and stroke, with an incidence rate ratio of 2.89 in the first week (95% CI 1.51 to 5.55).[16] None of these studies are validated as a prediction tool for infection-related CVD, and so we do not know how they would perform for this outcome.

Pneumonia was a strong predictor of outcome in the models and the DASHI score. Pneumonia was previously known to be associated with CVD events.[42,104,164] An observational study examining the effect of aspirin on CVD events in pneumonia suggested a protective effect of aspirin in this population, as did a small open-label randomised trial in people with pneumonia and two risk factors for CVD.[103,104] Pneumonia can be used to identify some people at high risk of CVD, but the results in this chapter show this approach will miss many patients who go on to have infection-related CVD events.

Pneumonia was diagnosed in 1.8% of the validation population and based on the derivation population, 3.4% of these were predicted to have CVD events. This 'pneumonia diagnosis rule' performs well, it is comparable to the models and score at similar levels of predicted probability. However, people who get pneumonia would generally score highly in these models as well – they would get 4 DASHI points for the diagnosis, and many of them are likely to be older or have diabetes, which would earn them further points. This is why DASHI predicts only 0.58% risk for pneumonia alone – in DASHI this means the patient doesn't smoke, or have heart failure, or diabetes, and is under 60 years of age. These are not the characteristics of most people with pneumonia in this population, or in general.[164] There are

downsides to using pneumonia alone as a rule to identify high risk people. It is not flexible, and the 3.4% predicted risk is higher than might be ideal for instigating some interventions. Consequently, false negatives are high. Using the diagnosis of pneumonia alone there were 62 true positives and 198 false negatives per 100,000 population. For an intervention to have much impact on infection-related CVD, a lower threshold would be needed to include more of the cases.

In comparison to CVD risk prediction tools and strategies for other settings DASHI is a simpler to use, predicts shorter-term risk, and considers respiratory infection related variables.

4.6.3 Strengths and limitations of this study

The strengths and limitations of the data and approaches to imputation and model derivation are described in the previous chapter.

One potential strength of this external validation is the choice of using a different CPRD database to the derivation cohort. This is a separate set of people, and the baseline statistics are close to those in the derivation cohort, and representative of the population for which it was intended. A corresponding weakness is that one might argue this population was too similar to the derivation cohort. One could argue that being a subset of the UK population it is part of the same population as the derivation cohort. This would make the validation closer to being internal. One approach to this criticism would be to validate DASHI in another population, for example using UK primary care data from a later time or using data from abroad. However, the results show there is a difference between internal and external performance; there was evidence of slight overfitting in Model Two, which lead to slightly impaired calibration in the external dataset. If one were to validate the score in a completely different population, this would provide validation data for that population, rather than

demonstrating performance for the target population of UK primary care. The limitation is that it has not been validated for other populations, such as different countries.

The calibration of the models and score was slightly suboptimal at the highest predicted probabilities. For example, in the top 2% of the population by risk, Model One over-predicts by less than one percent. This magnitude of miscalibration is small, it applies to relatively few patients, and, as they would be high-risk anyway, is unlikely to make much difference in clinical use. The DASHI score, which has only 12 possible predicted probabilities (0 to 11 points), is very well calibrated for 90% of patients, with slight overprediction in the top ten percent of predicted risk.

A weakness of creating a points score is that deriving points scores necessarily loses information; they are in effect new models, and need validation, which I have done.[184] In this case the performance appears to be very similar to the more complex models. Scores are also easier to implement in clinical practice than statistical models. They can be remembered, and implemented mentally, whereas statistical models usually cannot. They also do not require the MHRA approves them as medical devices.

4.6.4 Implications for research and practice

Primary care clinicians could use DASHI to estimate and discuss primary CVD event risk. DASHI is valid for patients over the age of 40 presenting to primary care with an acute respiratory tract infection. We don't have evidence regarding acute treatments yet, but many high-risk patients will already be eligible for long-term primary prevention. A potential use is that clinicians could take the opportunity to offer primary prevention with statins to those who are already eligible, and to check adherence and optimise the dose in the 9% of people already prescribed them. The sixth chapter explores the effect of statin use on infection-related CVD. As the risks from statins are very low, a low threshold would be appropriate to

trigger this action; an equivalent to the 10% ten-year CVD risk used for statin prescriptions is one DASHI point.[10] The latest guidance from NICE suggests the use of statins at any level of ten year CVD risk, if patients wish to take them, as they are likely to be cost effective.[10]

Antiplatelets are not recommended for primary prevention, and exploring the possible effects of this option is the subject of the next chapter. Because the risks of bleeding with antiplatelets are higher than any risks with statins I used DASHI to define a higher risk population for research. The predicted risk at three DASHI points exceeds the risk of major bleeding from short duration single antiplatelet therapy (about 0.3%).[98] The CVD event risk at four DASHI points exceeds the risk of major bleeding with dual antiplatelet therapy (0.5%).[98]

Clinicians could also use DASHI as a tool to encourage influenza vaccination to prevent cardiovascular events, though the vaccination itself would not be administered during acute illness.[29,30]

Meta-analysis of previous trials of antibiotics in stable coronary disease shows at best no benefit, and at worst increased mortality.[188] Acute respiratory infection can be an indication for antibiotic prescribing, but clinicians should not prescribe antibiotics for CVD risk.

4.6.5 Conclusions

In this work I have developed and validated the DASHI score. DASHI is a simple clinical risk prediction score for primary CVD events during and following acute respiratory infections. It applies to primary care populations over the age of 40, presenting with respiratory infections, and predicts 28-day CVD risk.

This new tool is a way to identify individuals who are most likely to have an infection-related cardiovascular event. It enables research aiming to reduce CVD events in people with acute respiratory infections.

The following two chapters will use the DASHI score to stratify the risk of post infection CVD risk in studies seeking to measure the effect of two potential interventions, aspirin and statins.

5 Chapter Five: Aspirin and acute infection-related cardiovascular events: a retrospective cohort study

"In order to be in control, you have to have a definite plan for at least a reasonable period of time. So how, may I ask, can man be in control if he can't even draw up a plan for a ridiculously short period of time, say a thousand years, and is moreover unable to ensure his own safety for even the next day? . . . . Yes, man is mortal, but that isn't so bad. What's bad is that sometimes he's unexpectedly mortal, that's the rub!"

Michael Bulgakov, The Master and Margharita, 1938

5.3 Introduction

5.3.1 Overview within thesis – how this chapter fits in

The previous two chapters described the derivation and validation of the DASHI score. The DASHI score predicts 28-day risk of primary cardiovascular events in people over the age of 40 presenting to primary care with acute respiratory infections. The next clinical problem is what to do with this information. The subject of the chapters five and six is assessing interventions which might mitigate that risk. I chose to focus on possible repurposing of interventions used to mitigate the risk of CVD events in other settings. In this chapter I tried to measure what I hoped would be a protective effect of aspirin on infection-related CVD events. The next chapter, chapter six, explores mitigating the risk with statins. Both these chapters use DASHI to identify higher-risk patients, and causal inference epidemiology methods to attempt to measure the effect of these interventions on respiratory infection-related CVD event risk.

5.3.2 Background

Guidelines recommend long term aspirin should not be used for primary prevention of CVD events but should be used for secondary prevention because the underlying risk is higher.[19] The risk of infection-related CVD events is an intermediate case – for a short time the CVD event risk exceeds the background primary prevention risk. We can now estimate the risk with DASHI. This means that there may be an opportunity to use short-term aspirin to reduce CVD events during the period of temporarily elevated risk.

The high-risk period around respiratory infections might be an exceptional time when the risk of CVD events is high enough to justify the risks of aspirin use. Previous chapters show we can estimate short term CVD event risk following respiratory infection. This chapter describes how I explored the use of aspirin as an intervention for patients at higher risk of infection-related CVD events.

5.4 Methods

5.4.1 Chapter aim

The overall aim of this chapter was to measure the effect of aspirin on respiratory infection-related CVD events and bleeding.

5.4.2 Chapter objectives

1. The first objective was to develop and assess four candidate propensity models to predict the probability of exposure to aspirin.
2. The second objective was to use one of these models to estimate the effect of aspirin on infection-related CVD events
3. The third objective was to use the same methods to assess the risk of bleeding due to aspirin in the 28 days following respiratory infection.

5.4.3 Data source

As described in the previous chapters, data for this thesis was provided by the Clinical Practice Research Datalink (CPRD) GOLD and Aurum databases (Section 3.4.3). Aurum is larger than Gold, and so I used Aurum for the causal inference analyses to maximise precision. As before I used data from 1999 to 2019 which was linked to ONS and HES data as described in the earlier chapters. [158]

5.4.4 Population

I selected the study population to include people who could theoretically be recruited to a future trial, to help clarify the question and reduce immortal time bias.[189] Immortality time bias in observational studies occurs when participants receive treatment prior to the study period, and therefore have to have been event-free for longer than the controls – and as a result are lower risk, biasing the effect of the treatment towards lower events.[189] In addition, patients being recruited to a future trial would have to be eligible to receive the potential

intervention without contraindications, and have a risk of outcomes likely to exceed the risk of bleeding.

As the risk of bleeding seen in trials of short term antiplatelets is of the order of 0.3%, I included patients with a higher risk of infection-related CVD events, defined as a DASHI score of three or more (three DASHI points predicts 0.3% 28-day risk).[99] The other inclusion criteria were the same as the previous chapters – they had to have no prior CVD events and one year of data at a CPRD practice (Chapter Three: Figure 1: Patient timeline in cohorts).

Further exclusion criteria were: people established on aspirin prescriptions at six months before the index date (who were therefore not eligible to become newly exposed to aspirin), those taking anticoagulants or heparins (in whom aspirin is contraindicated), and people with low platelet counts (another contraindication).[9,176] .

Patients entered the study at the diagnosis of their first respiratory infection after the age of forty years, which defined the index date, and were followed for 28 days (Chapter three Figure 1: Patient timeline in cohorts).

5.4.5 Exposure

If I had designed a trial, the intervention would have been similar to treatment of acute stroke: 75mg of aspirin daily with a 300mg loading dose to ensure quick action, prescribed at the index date, and continued for the 28-day study period.[9] Because this is observational data, and it is not current practice to prescribe aspirin for respiratory infections, I used a compromise exposure definition of new aspirin use. I defined this as patients exposed to a new prescription of aspirin at the index date, which allowed it to be newly prescribed beforehand. To ensure they could still be exposed at the index date I required the number of doses (tablets, capsules etc.) had been prescribed the same (or fewer) number of days before the index date. For example, if they were prescribed seven tablets, a prescription eight days

before the index date would not count as exposed to aspirin, whereas a 28-capsule prescription 27 days before the index date would count as an exposure. I do not know the reason for the aspirin prescriptions, but the BNF lists several indications: secondary prevention cardiovascular indications, secondary prevention of deep vein thrombosis, pulmonary embolus, pain and migraine.[9] These 'as required' indications might lead to an under-estimation of the effects of aspirin. Aspirin is also widely used for primary prevention of cardiovascular disease, for which it is taken regularly.

I undertook two post-hoc sensitivity analyses examining the effect of aspirin that had been started as repeat prescriptions either i) six to three months before the index date, or ii) likely established at six months before the index date (as I did not have data from this period, I used a repeat prescription fewer days from six months before the index date than the number of doses in the prescription).

5.4.6 Control

The control group comprised people who did not start aspirin. This meant anyone starting aspirin after the index date (I did not have data identifying these patients) would have been allocated to the control group, because I expected people with an event in the study period to start aspirin for secondary prevention. I also undertook a post-hoc sensitivity analysis moving those prescribed aspirin on day zero into the control group.

5.4.7 Outcomes

5.4.7.1 Primary outcome

The primary outcome was the same as the previous chapters: infection-related CVD events (Section 3.4.7). This was a composite of diagnoses likely related to coronary or cerebral atherosclerosis, occurring within 28 days of a first diagnosis of respiratory infection. The events were: myocardial infarction, new angina, unstable angina or acute coronary syndrome, ischaemic cardiomyopathy, stroke or transient ischaemic attack (TIA). These were defined by

primary and secondary care codelists, which included interventions for these (e.g. percutaneous intervention) and deaths from these causes.

5.4.7.2 Secondary outcomes

Secondary outcomes were bleeding diagnoses. These were pre-specified as any bleeding, and major bleeding. Any bleeding was a pathological haemorrhage of any sort recorded in the medical record. Major bleeding I defined based on the International Society on Thrombosis and Haemostasis definition of major bleeding for non-surgical patients, which I adapted for the coding systems and data available.[190] I included bleeding that was associated with blood product transfusion, iron infusion, or a procedure to arrest haemorrhage. I also included bleeding in anatomical sites I thought likely to require further care e.g. intracranial, intrathoracic, intraocular, gastrointestinal, peritoneal, arterial dissections, haemarthroses, and those affecting solid organs. The full codelist is online.

5.4.7.3 Negative control outcomes

Negative controls are controls that are not plausibly linked to the causal relationship of interest. They can help detect inaccuracies in the analysis if it measures a causal association for these biologically implausible outcomes where none is expected.[191] I selected two negative outcomes I expected to have no causal association with aspirin (or statins) but be correlated with CVD events via confounders. I chose skin lesions and constipation.[104] These are minor, relatively common conditions often seen in primary care, which are associated with age, attendance at the GP, and various medications, but not classically aspirin.[9] I assessed their occurrence in the 28-day follow-up period by the presence of codes in the primary care record, and their causal association with aspirin via the same analysis as the main outcomes. A non-zero causal association would imply a problem with the analysis.

5.5 Statistical methods

5.5.1 Propensity score overview

As discussed in the introductory chapters, the propensity score approach is to model the process by which an exposure was allocated to patients.[135] If this is successful, then one can account for the propensity to be exposed, and obtain an estimate of the effect of exposure on the outcome, contingent on satisfying the necessary assumptions.[135,192]

Best practice for propensity modelling is to define candidate propensity models first, and assess their performance, before using a single analysis to assess the exposure-outcome relationship of interest.[121,139] This prevents the models being built with knowledge of the final causal estimate. I have followed this approach; Objective One, the largest part of the analysis, involved building multiple candidate propensity models for aspirin exposure, and assessing their performance. I have given an overview of these methods and results in this chapter and described the details in the supplementary materials (Appendix – Supplementary materials for chapter five). The Objective Two methods section describes how, after selecting a single propensity model, I used it to estimate the effects of aspirin on CVD events and bleeding within 28 days of a respiratory infection.

5.5.2 Measures of causal effects

I have followed the counterfactual approach to causal inference, which was developed from the observation that one cannot observe what would have happened to people if they had received a different treatment.[135] Potential outcomes are the outcomes people would have had if they had been assigned a different treatment or exposure status than they really received.[144,193] This is theoretical and in practice modelled based on an individual's covariates.[144,193] I report Potential Outcome Means (POMs) – the mean outcomes over the study population if they had all been allocated to a particular level of exposure.[144,193] There are two POMs, according to the treatment exposure. 'POM treated' (also POM_1) is the

estimated outcome over the study population modelled as if they had all received a treatment. [144,193] 'POM untreated' (POM_0) is the estimate of outcomes had no-one in the population received the treatment. [144,193] POM_1-POM_0 gives the Average Treatment Effect (ATE) (also called Average Causal Effect (ACE) in some publications) - the estimated absolute effect due to treatment, equivalent to the risk difference and directly comparable to trial results.[144,193] The risk ratio estimate is POM_1/POM_0. [144,193]

5.5.3 Missing data

5.5.3.1 Multiple imputation

I used multiple imputation methods to mitigate missing data, as described in the earlier chapters (Section 2.3.5.1). I again used chained equations with augmentation to generate 20 imputations including the outcomes in the models, and included the outcomes in the models.[147,152] I log transformed skewed continuous variables (Cholesterol to HDL ratio, Systolic BP and BMI) before modelling with linear regression, and then back transforming them. I modelled ethnicity with multinomial logistic regression, and ordinal variables with ordered logistic regression. I examined the distributions of these variables in the imputed datasets and the original data before proceeding to use the data for the analyses.

5.5.3.2 Missing indicators

I used missing indicators for CRP and platelet counts for two reasons. Firstly, these are blood tests carried out for symptoms, rather than screening (as is the case for lipids).[10,194] This makes imputation more difficult and less appropriate. Whilst it may be possible to identify a complete set of reasons for these tests being requested, and to use these to impute the values, this would be extremely impractical.[147–149] Secondly, variables with high missingness require more imputations which is computationally expensive to impute.[152] As reported in the previous chapters, CRP and platelet count both had a large amount of missing data (92% missing for CRP and 70% missing for platelets see Table 2: Characteristics of patients by CVD event outcome status in derivation dataset). The ideal number of imputations might be

more than 90, which could in theory be calculated given extra computational resources but was not practical for this project.

5.5.4 Objective one methods: Propensity modelling and weights

To assess the suitability of different modelling approaches I used four nested sets of variables to build four candidate propensity models (supplementary methods section 10.3.1).[195] I started with the simplest model first and then added groups of covariates sequentially. This allowed me to compare the candidate propensity models and select one for the final analysis.

5.5.4.1 Overview of methods for objective one

To group variables for the four candidate propensity models, I assessed the strength of association between the variables, the outcome and the exposure (section 10.3.1).[121] Group one variables were demographics, group two had greater magnitude of association with the outcome than the exposure, group three had stronger associations with the exposure than the outcome, and group four was recent blood test results.

For each of the four candidate models, I then:

1. Predicted the probabilities (propensity) for each individual
2. Examined the overlap of probabilities between exposed and unexposed groups
3. Looked for evidence of violation of the positivity assumption and interaction
4. Estimated the apparent (internal) C statistic
5. Produced apparent calibration plots
6. Generated truncated stabilised inverse probability weights
7. Examined the mean and range of inverse probability weights
8. Examined the covariate balance before and after weighting
9. Assessed the crude effect of aspirin on the negative controls constipation and skin lesions

Further details of the methods for these steps are in the supplementary materials (Appendix – Supplementary materials for chapter five).

5.5.5 Objective two and three methods: Final modelling of effect of aspirin on CVD events and bleeding

To model the effect of aspirin on the outcomes, I selected one candidate propensity model for a single analysis. To choose a model I first ensured it performed adequately without major concerns about validity, using the process above. I made the final decision based on the covariate balance achieved by the model.[192]

I used Stata 18 (College Station, Texas) and the command *teffects* to estimate the effects of aspirin on the outcomes.[144] This package uses linear regression to model potential outcomes. I weighted the regression models by the truncated stabilised inverse probability weightings and adjusted for covariates that were less well balanced by these weightings.[192] I combined these approaches because doing so has the attractive property of double robustness – that if either the regression model or the propensity modelling is correct then the results are unbiased.[142] *teffects* does not provide risk ratios, I obtained these from the marginal probabilities option *coefflegend*, and combined them with the *nlcom* command.[144]

5.5.5.1 Post-hoc sensitivity analyses

I performed post-hoc analyses to assess the robustness of the results.

1. To assess the possibility that the propensity modelling was insufficiently complex, I performed further analyses after introducing some interaction terms with the propensity and more imbalanced variables.
2. I then added the type of infection to the models
3. I examined different definitions for the exposure variable
    a. Aspirin initiated three to six months prior to the index date
    b. Aspirin use at six months prior to the index date

c. Moving those prescribed aspirin on day zero into the control group

5.6 Results

5.6.1 Population

In the study population there were 807,389 patients, of whom 9,809 (1.2%) started aspirin prior to the index date (Table 14: Characteristics of participants by aspirin status). CVD events occurred in 6,533 (0.8%) patients, and 276 occurred in people taking aspirin (2.8%). Aspirin users were older (mean age 74 years Vs 72 years), had higher systolic blood pressure (mean 141 mmHg vs 138 mmHg) and BMI (28 KgM^{-2} Vs 27 KgM^{-2}). Aspirin users also had higher use of antihypertensives and statins.

Bleeding events occurred 2,749 times, with 344 of these being major bleeding events. In the new aspirin users there 43 bleeds, but only five major bleeds. I was therefore unable to do further analyses using major bleeds as an outcome.

Table 14: Characteristics of participants by aspirin status

Variables*	Total		No aspirin		Aspirin	
Continuous	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Age in years	71.6	10.9	71.6	10.9	73.9	10.1
Cholesterol to HDL ratio	3.8	1.2	3.8	1.2	3.9	1.3
Systolic BP mmHg	137.6	17.7	137.5	17.7	140.8	18.7
BMI KgM^{-2}	27.3	5.6	27.3	5.6	28.0	5.7
DASHI score	3.9	1.2	3.9	1.2	4.0	1.3
Categorical	**n**	**%**	**n**	**%**	**n**	**%**
CVD event	6,533	0.8%	6,257	0.8%	276	2.8%
Negative controls:						
Skin lesions	6,484	0.8%	6,418	0.8%	66	0.7%
Constipation	2,772	0.3%	2,727	0.3%	45	0.5%
Any bleed outcome	2,749	0.3%	2,706	0.3%	43	0.4%
Major bleed outcome	344	<0.1%	339	<0.1%	5	0.1%
Other antiplatelets	7,963	1.0%	7,797	1.0%	166	1.7%
Female	465,625	57.7%	460,487	57.7%	5,138	52.4%
Index infection:						
URTI	193,850	24.0%	190,548	23.6%	3,302	33.7%
LRTI	549,637	68.1%	543,780	67.3%	5,857	59.7%
Influenza	55,062	6.8%	54,472	6.7%	590	6.0%
Pneumonia	63,911	7.9%	63,261	7.8%	650	6.6%
Smoking status:						
Nonsmoker	246,502	30.5%	243,158	30.1%	3,344	34.1%
Ex-smoker	202,228	25.0%	199,686	24.7%	2,542	25.9%
Light smoker	38,408	4.8%	37,932	4.7%	476	4.9%
Moderate smoker	43,623	5.4%	43,114	5.3%	509	5.2%
Heavy smoker	29,667	3.7%	29,338	3.6%	329	3.4%
Current smoker	78,714	9.7%	77,835	9.6%	879	9.0%
Smoking data missing	168,256	20.8%	166,526	20.6%	1,730	17.6%
Diabetes mellitus	125,385	15.5%	122,304	15.1%	3,081	31.4%
Heart failure	16,789	2.1%	16,281	2.0%	508	5.2%
Chronic heart disease	13,343	1.7%	13,020	1.6%	323	3.3%
Atrial arrhythmia	14,691	1.8%	13,823	1.7%	868	8.8%

Markers of atherosclerosis	108,728	13.5%	106,426	13.2%	2,302	23.5%
Cancer	48,529	6.0%	47,902	5.9%	627	6.4%
Family history of CVD	4,868	0.6%	4,837	0.6%	31	0.3%
Antihypertensive use	275,615	34.1%	269,969	33.4%	5,646	57.6%
Rheumatoid Arthritis	18,376	2.3%	18,218	2.3%	158	1.6%
Statin use	149,491	18.5%	145,635	18.0%	3,856	39.3%
Platelets x 10^9/L:						
Unknown platelets	520,154	64.4%	514,708	63.7%	5,446	55.5%
Platelets normal (150-450)	278,634	34.5%	274,402	34.0%	4,232	43.1%
Thrombocytosis (>450)	8,610	1.1%	8,479	1.1%	131	1.3%
C Reactive Protein mg/L						
CRP unknown	737,131	91.3%	728,061	90.2%	9,070	92.5%
CRP <5	29,514	3.7%	29,231	3.6%	283	2.9%
CRP 5 to <20	28,489	3.5%	28,176	3.5%	313	3.2%
CRP >=20	12,264	1.5%	12,121	1.5%	143	1.5%
Index of Multiple deprivation:						
Least deprived decile	83,045	10.3%	82,120	10.2%	925	9.4%
Most deprived decile	77,753	9.6%	76,746	9.5%	1,007	10.3%
Total N	807,398	100.0%	797,589	98.8%	9,809	1.2%

* All patients have a Respiratory Tract Infection (RTI) categorised into upper (URTI), lower (LRTI) and pneumonia. Influenza is a separate, non exclusive category and can be included separately in addition to LRTI (default for influenza, unless coded to another site). BP = blood pressure. BMI body mass index. CVD event = composite outcome (myocardial infarction, coronary syndromes, transient ischaemic event, stroke, ischaemic cardiomyopathy) in the 28 days following respiratory infection. Negative controls and bleeding outcomes have the same follow-up time. Skin lesions are any non-melanoma skin lesions. Major bleeding is bleeding in a body site thought to require admission/interventions. Heart failure includes all non-ischaemic diagnoses. Chronic heart disease incluses valvular disease, hypertensive disease and congenital disease. Atrial arrhythmias include atrial tachicardias, atrial fibrillation, and flutter. Diabetes mellitus includes type i, type ii, and other/unrecorded type. Markers of atherosclerosis includes: chronic kidney disease, peripheral arterial disease, and erectile dysfunction. Family history of CVD is in first degree relatives. Cholesterol : HDL ratio is serum total cholesterol/serum HDL cholesterol, 40.6% was missing. BMI was 45.3% missing. Systolic BP was 15.8% missing .

5.6.2 Objective one results: candidate propensity model variables

The variables associated more strongly with CVD events than aspirin were smoking, total cholesterol to HDL ratio, systolic blood pressure, heart failure and cancer diagnoses. These were included in candidate models two three and four, whereas those associated more strongly with aspirin than CVD events were included only in candidate models three and four (Table 15: Groups of variables included in propensity models).

Table 15: Groups of variables included in propensity models for aspirin

Propensity models using group of confounders	Group Description	Clinical variables in group
Models 1, 2, 3 and 4	Demographics	Age, sex, index of multiple deprivation decile, and ethnicity
Models 2, 3 and 4	Confounders more strongly associated with cardiovascular outcomes than aspirin	Smoking, total cholesterol to HDL cholesterol ratio, systolic blood pressure, heart failure, and cancer diagnoses
Models 3 and 4	Confounders more strongly associated with aspirin exposure than cardiovascular outcomes	Body mass index, valvular and congenital heart disease, markers of atherosclerosis, atrial arrhythmias, statin prescriptions and diabetes mellitus diagnosis
Model 4 only	Recent blood tests	Platelet count, C reactive protein

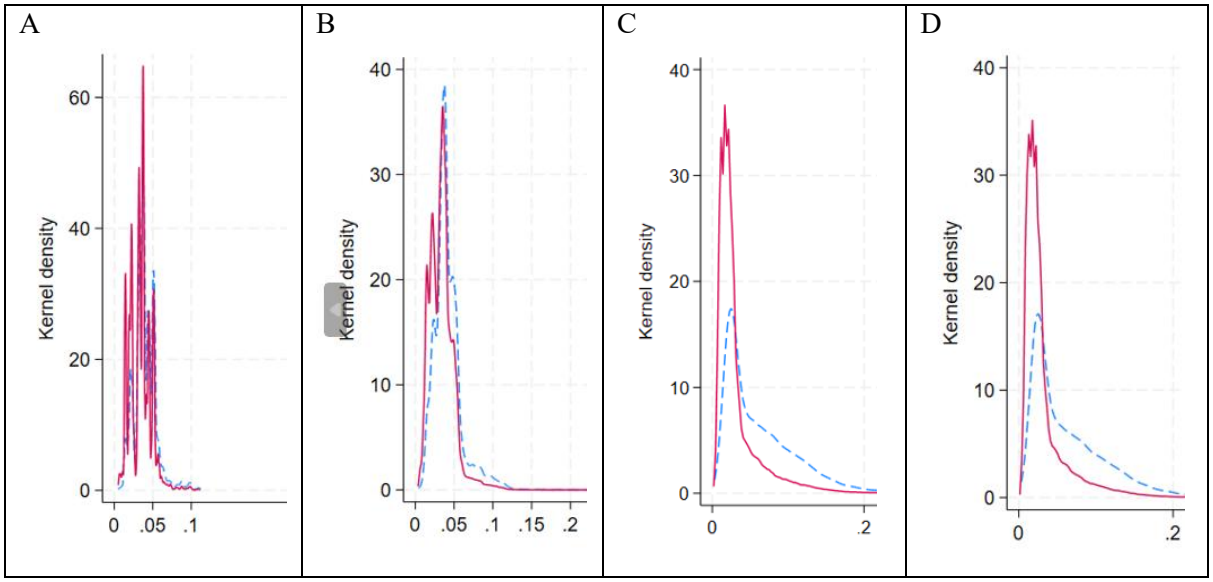
Models one through four are propensity models for aspirin exposure at diagnosis with respiratory infection in primary care. Aspirin exposure defined as prescriptions started within six months before the index date and ongoing at the index date. Cardiovascular outcomes are a composite outcome (myocardial infarction, coronary syndromes, transient ischaemic event, stroke, ischaemic cardiomyopathy) in the 28 days following respiratory infection

5.6.3 Objective one results: assessment of candidate propensity models

5.6.3.1 Propensity overlaps

I assessed the distributions of propensity graphically (Figure 6: Histograms of propensity model predicted probabilities by aspirin exposure). The range of predicted probabilities was lower with propensity Model One than the other models, but all models had overlapping regions.

Figure 6: Histograms of propensity model predicted probabilities by aspirin exposure



Overlap of propensity (predicted probability, x axis) by exposure status for each of four sequential propensity models. Blue dashed line is scores in the exposed population, red solid line is scores in those unexposed. A; model with demographics only (age, sex, IMD, ethnicity), B; model with demographics and further confounders (Smoking, total cholesterol to HDL cholesterol ratio, systolic blood pressure, heart failure, and cancer diagnoses), C; model with further variables added (Body mass index, valvular and congenital heart disease, markers of atherosclerosis, atrial arrhythmias, statin prescriptions and diabetes mellitus diagnosis), D; model with recent blood tests added (CRP, platelets).). Kernel density is smoothed estimate of probability density (in units of 1/propensity where area under curve = 1). Note different axis scales.

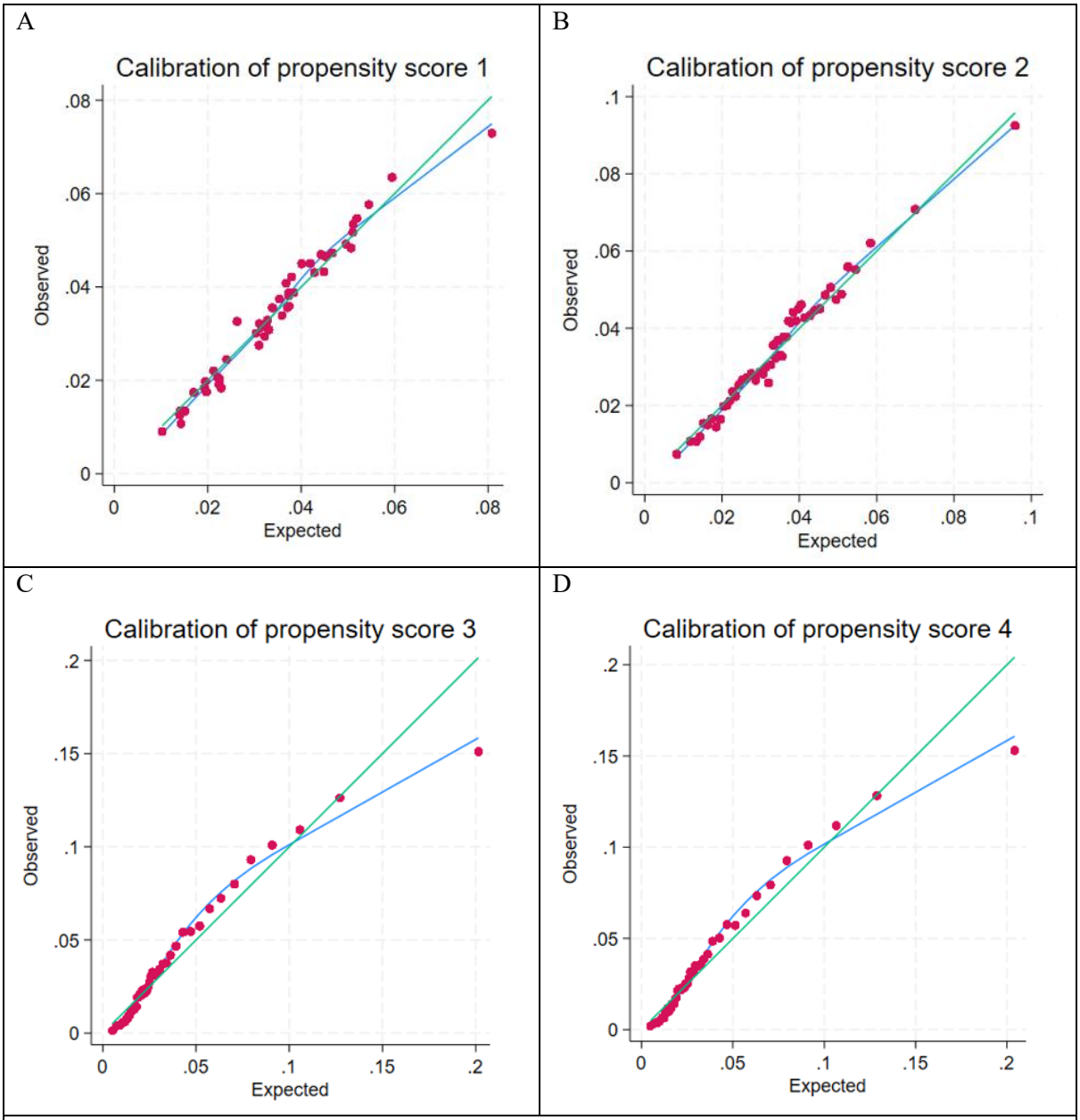
5.6.3.2 Assessment of predictive properties of the propensity models

The estimated C statistics for discriminating people exposed to aspirin from those not exposed showed similar performance for Model One and two and (0.62 and 0.63 with

overlapping confidence intervals), and three and four (0.75 and 0.76 with overlapping confidence intervals, Table S31: Discrimination of propensity models, internal apparent estimates).

Internal calibration of the propensity models showed good calibration in the range of up to 10% predicted probability, but minor overprediction in scores three and four, at higher predicted probabilities, (predicted ~20% probability for the 2% of the population with the highest predicted probabilities but observed ~15%). Models one and two predicted ~8% and ~10% probability of aspirin prescription respectively for the 2% with the highest predicted probabilities, and the observed outcomes were about 7.5% and 9% (Figure 7: Calibration curves for propensity).

Figure 7: Calibration curves for propensity models for aspirin exposure



Internal calibration curves for propensity models for aspirin exposure. Green (straight) line is ideal performance; blue (curved) line is a cubic spline of actual performance, fit to the red markers (dots), which represent 50ths of predicted exposure risk plotted against actual exposure in these groups. A; model with demographics only (age, sex, IMD, ethnicity), B; model with demographics and further confounders (Smoking, total cholesterol to HDL cholesterol ratio, systolic blood pressure, heart failure, and cancer diagnoses), C; model with further variables added (Body mass index, valvular and congenital heart disease, markers of atherosclerosis, atrial arrhythmias, statin prescriptions and diabetes mellitus diagnosis), D; model with recent blood tests added (CRP, platelets).

5.6.3.3 Covariate balance under different propensity models

The raw data had many covariates with large imbalances between those prescribed aspirin and those not using aspirin. After weighting the population by the weights from each model, none achieved standardized mean difference balance (SMD of less than 0.1) in diabetes, statins, or markers of atherosclerosis (Table 16: Covariate balance under different propensity models). Models three and four reduced these imbalances the most but worsened the imbalance in ethnicities and age compared to models one and two.

Table 16: Covariate balance under different propensity models

Model		Raw data			Propensity model 1 weighted			Propensity model 2 weighted			Propensity model 3 weighted			Propensity model 4 weighted		
Variable and categories		Percent no aspirin	Percent in aspirin	Standardised difference	Percent no aspirin	Percent in aspirin	Standardised difference	Percent no aspirin	Percent in aspirin	Standardised difference	Percent no aspirin	Percent in aspirin	Standardised difference	Percent no aspirin	Percent in aspirin	Standardised difference
Age																
	Decile 1	0.49	0.09	-0.06	0.48	0.11	-0.07	0.48	0.11	-0.07	0.48	0.13	-0.06	0.48	0.13	-0.06
	Decile 2	0.49	0.21	-0.04	0.49	0.25	-0.04	0.49	0.25	-0.04	0.49	0.24	-0.04	0.49	0.24	-0.04
	Decile 3	0.63	0.22	-0.05	0.63	0.26	-0.06	0.63	0.27	-0.05	0.63	0.29	-0.05	0.63	0.29	-0.05
	Decile 4	0.60	0.46	-0.02	0.60	0.52	-0.01	0.60	0.50	-0.01	0.60	0.45	-0.02	0.60	0.46	-0.02
	Decile 5	0.80	0.51	-0.03	0.80	0.58	-0.03	0.80	0.59	-0.03	0.80	0.58	-0.03	0.80	0.59	-0.03
	Decile 6	0.82	0.60	-0.02	0.82	0.66	-0.02	0.82	0.67	-0.02	0.82	0.61	-0.02	0.82	0.62	-0.02
	Decile 7	18.47	11.92	-0.17	18.39	13.41	-0.14	18.39	13.60	-0.13	18.39	12.83	-0.15	18.39	12.83	-0.15
	Decile 8	22.61	19.41	-0.08	22.57	21.16	-0.03	22.57	21.18	-0.03	22.57	19.65	-0.07	22.57	19.65	-0.07
	Decile 9	22.29	25.07	0.07	22.33	24.61	0.05	22.33	24.53	0.05	22.33	23.78	0.03	22.33	23.75	0.03
	Decile 10	32.81	41.50	0.19	32.91	38.44	0.12	32.91	38.32	0.11	32.90	41.43	0.18	32.90	41.43	0.18
Sex Female		57.73	52.38	-0.11	57.67	55.63	-0.04	57.67	55.10	-0.05	57.68	55.26	-0.05	57.68	55.27	-0.05
Index of multiple deprivation																
	Decile 1	10.50	9.61	-0.03	10.29	10.04	-0.01	10.29	9.93	-0.01	10.29	9.82	-0.02	10.29	9.82	-0.02

Decile 2	9.92	8.96	-0.03	9.72	9.30	-0.01	9.72	9.34	-0.01	9.72	9.40	-0.01	9.72	9.40	-0.01
Decile 3	10.22	9.52	-0.02	10.02	9.74	-0.01	10.02	9.77	-0.01	10.02	9.79	-0.01	10.02	9.80	-0.01
Decile 4	10.57	10.12	-0.01	10.37	10.23	0.00	10.37	10.18	-0.01	10.37	10.12	-0.01	10.37	10.13	-0.01
Decile 5	9.44	8.92	-0.02	9.26	8.92	-0.01	9.26	8.88	-0.01	9.26	9.13	0.00	9.26	9.11	-0.01
Decile 6	10.12	10.27	0.00	9.93	10.01	0.00	9.93	10.04	0.00	9.93	10.16	0.01	9.93	10.15	0.01
Decile 7	10.38	10.32	0.00	10.19	10.05	0.00	10.19	10.01	-0.01	10.19	9.92	-0.01	10.19	9.89	-0.01
Decile 8	9.38	9.90	0.02	9.21	9.39	0.01	9.21	9.40	0.01	9.21	9.39	0.01	9.21	9.38	0.01
Decile 9	9.65	11.91	0.08	9.50	10.38	0.03	9.50	10.51	0.03	9.49	10.50	0.03	9.49	10.53	0.03
Decile 10	9.81	10.47	0.02	9.63	9.92	0.01	9.63	9.90	0.01	9.63	9.87	0.01	9.63	9.88	0.01
Ethnicity															
Bangladeshi	0.41	0.60	0.03	0.29	0.34	0.01	0.29	0.34	0.01	0.29	0.34	0.01	0.29	0.34	0.01
Black African	0.88	1.83	0.10	0.61	0.78	0.02	0.61	0.82	0.02	0.62	0.80	0.02	0.62	0.80	0.02
Black Caribbean	1.34	3.72	0.21	0.94	1.35	0.04	0.94	1.37	0.04	0.95	1.34	0.04	0.95	1.35	0.04
Chinese	0.24	0.33	0.02	0.17	0.20	0.01	0.17	0.20	0.01	0.17	0.17	0.00	0.17	0.17	0.00
Indian	1.98	4.16	0.16	1.39	1.79	0.03	1.39	1.84	0.04	1.39	1.90	0.04	1.39	1.89	0.04
Other Asian	0.82	1.36	0.06	0.58	0.72	0.02	0.57	0.73	0.02	0.58	0.72	0.02	0.58	0.72	0.02
Other enthicity	1.20	1.77	0.05	0.83	0.98	0.02	0.83	1.01	0.02	0.84	0.93	0.01	0.84	0.93	0.01
Pakistani	0.89	2.19	0.14	0.62	0.86	0.03	0.62	0.87	0.03	0.62	0.85	0.03	0.62	0.85	0.03
White	92.24	84.04	-0.31	64.05	58.53	-0.11	64.04	59.17	-0.10	64.08	55.79	-0.17	64.08	55.80	-0.17
Smoking status															
Never smoked	38.53	41.39	0.06	30.50	33.09	0.06	30.51	32.90	0.05	30.53	31.51	0.02	30.53	31.53	0.02
Ex-smoker	31.64	31.46	0.00	25.05	25.85	0.02	25.03	26.59	0.04	25.05	25.12	0.00	25.05	25.08	0.00
Light smoker	6.01	5.89	-0.01	4.75	4.93	0.01	4.75	4.91	0.01	4.76	4.85	0.00	4.76	4.85	0.00
Moderate smoker	6.83	6.30	-0.02	5.40	5.54	0.01	5.40	5.45	0.00	5.40	5.39	0.00	5.40	5.39	0.00
Heavy smoker	4.65	4.07	-0.03	3.67	3.61	0.00	3.67	3.55	-0.01	3.67	3.53	-0.01	3.67	3.52	-0.01
Smoking unknown quantity	12.33	10.88	-0.04	9.75	9.30	-0.02	9.74	9.42	-0.01	9.75	9.02	-0.03	9.75	9.05	-0.02

Cholesterol to HDL ratio														
Decile 1	12.70	12.31	-0.01	5.69	6.64	0.04	5.68	6.74	0.04	5.70	5.77	0.00	5.70	5.76	0.00
Decile 2	9.83	9.68	-0.01	4.40	5.21	0.04	4.40	5.38	0.05	4.41	4.63	0.01	4.41	4.59	0.01
Decile 3	10.77	10.50	-0.01	4.82	5.74	0.04	4.82	5.90	0.05	4.83	5.12	0.01	4.83	5.12	0.01
Decile 4	10.65	10.19	-0.01	4.77	5.59	0.04	4.77	5.67	0.04	4.78	4.98	0.01	4.78	4.99	0.01
Decile 5	10.47	9.73	-0.02	4.69	5.38	0.03	4.68	5.59	0.04	4.69	4.92	0.01	4.69	4.91	0.01
Decile 6	10.02	9.75	-0.01	4.49	5.25	0.04	4.48	5.42	0.04	4.50	4.80	0.01	4.50	4.80	0.01
Decile 7	9.70	10.14	0.01	4.34	5.64	0.06	4.34	5.58	0.06	4.35	4.96	0.03	4.35	4.99	0.03
Decile 8	9.56	9.34	-0.01	4.28	5.19	0.04	4.28	5.32	0.05	4.29	4.63	0.02	4.29	4.60	0.02
Decile 9	8.52	9.53	0.04	3.81	5.37	0.07	3.81	5.32	0.07	3.82	4.71	0.04	3.82	4.70	0.04
Decile 10	7.77	8.83	0.04	3.48	5.00	0.08	3.48	4.91	0.07	3.49	4.12	0.03	3.49	4.12	0.03
Systolic blood pressure														
Decile 1	6.59	4.56	-0.08	5.87	4.26	-0.07	5.85	4.80	-0.05	5.85	4.78	-0.05	5.85	4.78	-0.05
Decile 2	4.43	3.67	-0.04	3.94	3.50	-0.02	3.93	3.79	-0.01	3.94	3.50	-0.02	3.94	3.48	-0.02
Decile 3	6.80	6.04	-0.03	6.06	5.74	-0.01	6.05	5.88	-0.01	6.05	5.68	-0.02	6.05	5.69	-0.02
Decile 4	6.89	5.86	-0.04	6.14	5.44	-0.03	6.13	5.98	-0.01	6.13	5.68	-0.02	6.13	5.68	-0.02
Decile 5	12.53	10.86	-0.05	11.16	10.22	-0.03	11.14	10.88	-0.01	11.15	10.25	-0.03	11.15	10.32	-0.03
Decile 6	6.20	5.95	-0.01	5.52	5.60	0.00	5.52	5.85	0.01	5.52	5.52	0.00	5.52	5.52	0.00
Decile 7	20.99	20.85	0.00	18.70	19.71	0.03	18.70	20.05	0.03	18.70	19.43	0.02	18.70	19.38	0.02
Decile 8	4.01	4.36	0.02	3.58	4.07	0.03	3.58	4.04	0.02	3.58	3.78	0.01	3.58	3.79	0.01
Decile 9	14.70	15.71	0.03	13.10	14.88	0.05	13.11	14.77	0.05	13.11	14.62	0.04	13.11	14.63	0.04
Decile 10	16.85	22.14	0.14	15.02	20.99	0.16	15.06	18.31	0.09	15.07	19.75	0.12	15.07	19.70	0.12
Heart Failure	2.04	5.18	0.22	2.05	4.95	0.16	2.07	2.51	0.03	2.07	3.35	0.08	2.07	3.35	0.08
Cancer	6.01	6.39	0.02	6.01	6.31	0.01	6.01	6.39	0.02	6.01	6.55	0.02	6.01	6.58	0.02
Body mass index deciles														
Decile 1	12.34	10.05	-0.07	6.99	6.27	-0.03	6.98	6.35	-0.03	6.97	6.78	-0.01	6.97	6.78	-0.01

Decile 2	9.98	8.34	-0.05	5.65	5.10	-0.02	5.65	5.16	-0.02	5.65	5.41	-0.01	5.65	5.42	-0.01
Decile 3	9.90	8.32	-0.05	5.61	5.08	-0.02	5.60	5.11	-0.02	5.60	5.26	-0.01	5.60	5.26	-0.02
Decile 4	10.09	9.19	-0.03	5.71	5.61	0.00	5.71	5.73	0.00	5.71	5.68	0.00	5.71	5.64	0.00
Decile 5	9.93	10.08	0.00	5.62	6.23	0.03	5.62	6.27	0.03	5.63	6.14	0.02	5.63	6.13	0.02
Decile 6	10.14	11.01	0.03	5.74	6.79	0.04	5.74	6.86	0.05	5.75	6.14	0.02	5.75	6.12	0.02
Decile 7	9.96	11.17	0.04	5.64	6.95	0.05	5.64	7.01	0.06	5.65	6.18	0.02	5.65	6.17	0.02
Decile 8	9.82	10.57	0.03	5.56	6.61	0.04	5.56	6.58	0.04	5.57	5.89	0.01	5.57	5.88	0.01
Decile 9	9.47	11.74	0.08	5.35	7.56	0.09	5.36	7.56	0.09	5.38	6.19	0.03	5.38	6.20	0.04
Decile 10	8.37	9.54	0.04	4.73	6.24	0.07	4.74	6.22	0.07	4.76	5.14	0.02	4.76	5.15	0.02
Chronic heart disease	1.63	3.29	0.13	1.63	3.27	0.11	1.64	3.15	0.10	1.65	2.62	0.07	1.65	2.63	0.07
Atherosclerosis marker	13.34	23.47	0.30	13.36	22.86	0.25	13.36	23.12	0.25	13.45	18.75	0.14	13.45	18.79	0.15
Atrial arrhythmia	1.73	8.85	0.55	1.74	8.74	0.32	1.74	8.36	0.31	1.79	3.09	0.08	1.79	3.11	0.09
Statins	18.26	39.31	0.54	18.27	39.30	0.48	18.27	39.95	0.49	18.50	26.39	0.19	18.50	26.40	0.19
Diabetes mellitus	15.33	31.41	0.45	15.35	30.60	0.37	15.35	31.05	0.38	15.52	22.16	0.17	15.52	22.17	0.17
Platelets															
Platelets unknown	64.53	55.52	-0.19	64.52	55.83	-0.18	64.52	55.61	-0.18	64.44	58.66	-0.12	64.43	59.40	-0.10
Platelets normal	34.40	43.14	0.18	34.42	42.83	0.17	34.42	43.03	0.18	34.49	39.96	0.11	34.50	39.35	0.10
Platelets high	1.06	1.34	0.03	1.06	1.34	0.03	1.06	1.36	0.03	1.06	1.38	0.03	1.07	1.25	0.02
C reactive protein															
CRP unknown	91.28	92.47	0.04	91.28	92.47	0.04	91.28	92.37	0.04	91.27	92.92	0.06	91.30	91.96	0.02
CRP <5	3.66	2.89	-0.04	3.66	2.88	-0.04	3.66	2.91	-0.04	3.67	2.60	-0.06	3.66	3.11	-0.03
CRP 5 to <20	3.53	3.19	-0.02	3.53	3.17	-0.02	3.53	3.21	-0.02	3.54	3.05	-0.03	3.53	3.39	-0.01
CRP 20+	1.52	1.46	-0.01	1.52	1.48	0.00	1.52	1.51	0.00	1.52	1.43	-0.01	1.52	1.53	0.00

Weighted standardised mean differences in population weighted with stabilized truncated inverse probability weights from each propensity model. Percentages were calculated over the whole population, including those with missing values in the denominator. Chronic heart disease incluses valvular disease, hypertensive disease and congenital disease. Atrial arrhythmias include atrial tachicardias, atrial fibrillation, and flutter. Diabetes mellitus includes type i, type ii, and other/unrecorded type. Markers of atherosclerosis includes: chronic kidney disease, peripheral arterial disease, and erectile dysfunction. Family history of CVD is in first degree relatives. Cholesterol : HDL ratio is serum total cholesterol/serum HDL cholesterol

5.6.3.4 Overall assessment of the propensity models

I selected propensity Model Two to take forward for the final analysis. This was because there was some evidence that the more complex models (three and four) may have been introducing bias by overfitting. They both increased covariate imbalance in some variables, and calibration was slightly less good than models one and two. I therefore opted to use propensity score two, which offered a higher range of probabilities than Model One.

5.6.4 Objective two results: Estimated effect of aspirin on infection-related CVD events

5.6.4.1 Negative control results

The negative controls, skin lesions and constipation, did not have associations with aspirin in the raw data (Table S34: Association between aspirin and negative controls). This remained the case in the final analysis, with the relative risk of skin lesions being 0.91 (95% CI 0.78 to 1.07) and constipation 0.93 (95% CI 0.74 to 1.16) (Table 17: Estimated effects of aspirin on CVD events, bleeding and negative outcomes).

5.6.4.2 Effects of aspirin on infection-related CVD events

The potential outcome mean (POM) for infection-related CVD events in the unexposed (i.e. the estimated baseline risk of CVD events if no-one had been prescribed aspirin) was 0.80% (95% CI 0.78 to 0.82%), the average treatment effect due to aspirin (ATE - absolute risk difference) was an increase of 1.22% (95% CI 1.00 to 1.43%) and the risk ratio (RR) was 2.52 (95% CI 2.26 to 2.81) (Table 17: Estimated effects of aspirin on CVD events, bleeding and negative outcomes).

5.6.5 Objective three results: Estimated effect of aspirin on bleeding

The estimated 28-day risk of bleeding without aspirin was 0.34% (95% CI 0.33 to 0.36) with an absolute risk difference (ATE) of 0.01% (95% CI 0.01 to 0.02) and a risk ratio of 1.31 (95% CI 1.06 to 1.63).

Table 17: Estimated effects of aspirin on CVD events, bleeding and negative outcomes

	Estimate with no aspirin		Estimate with aspirin		Estimated effect due to aspirin		Estimated effect due to aspirin	
Outcome	POM	95% CI	POM	95% CI	ATE	95% CI	RR	95% CI
CVD events	0.80%	0.78 to 0.82	2.02%	1.81 to 2.24	1.22%	1.00 to 1.43	2.52	2.26 to 2.81
Bleeding	0.34%	0.33 to 0.36	0.45%	0.36 to 0.54	0.10%	0.01 to 0.20	1.31	1.06 to 1.63
Skin lesions	0.81%	0.79 to 0.83	0.74%	0.62 to 0.85	-0.07%	-0.19 to 0.05	0.91	0.78 to 1.07
Constipation	0.35%	0.33 to 0.36	0.32%	0.25 to 0.39	-0.03%	-0.01 to 0.05	0.93	0.74 to 1.16

Results from inverse probability weighting with propensity from Model Two, with further adjustment for age, statin use, diabetes, valvular and congenital heart disease, atrial fibrillation and platelet count. POM is potential outcome mean, ATE is average treatment effect , CI is confidence interval, RR is risk ratio.

5.6.6 Sensitivity analyses

5.6.6.1 Interaction terms

There was evidence of interaction between propensity to be exposed to aspirin and the effect estimate (Appendix, section 10.4). Accordingly, I examined the effect of using more complex models, including terms for interactions between allocation and the square of propensity, and for models including interaction terms between these and diabetes and atherosclerosis. Sensitivity analyses using interaction terms in the final model reduced the point estimate of risk with aspirin, but did not alter the direction of the results. The lowest was for the most complex model (ATE 0.72%, RR 2.33; Table 18: post-hoc sensitivity analyses).

I considered if infection types should have been included in the model, as they would have been if there was a trial. I did not do this in the main analysis because I expected the infection to be post-exposure in most cases and so could not confound the exposure to outcome relationship. A post-hoc sensitivity analysis including the type of infection as well as interactions did not change the direction of the results (ATE 0.71%, RR 2.48; Table 18: post-hoc sensitivity analyses).

5.6.6.2 Duration of aspirin exposure

The two analyses examining sensitivity to duration of aspirin were performed using Model Four. This was because on re-running the objective one analyses with different exposure definitions the best covariate balance was obtained with this more complex model. For aspirin established between three to six months prior to the respiratory infection the ATE was 0.63 (95% CI 0.007 to 1.21) and the RR 1.81 (95% CI 1.21 to 2.71; Table 18: post-hoc sensitivity analyses). For aspirin use thought to be ongoing six months before the index date (as identified by repeat prescriptions) the ATE was 0.51% (95% CI 0.35 to 0.84%) and the RR 1.62 (95% CI 1.43 to 1.82).

5.6.6.3 Aspirin on the index date

There were 2,171 prescriptions on the index date and once these were allocated to the controls there were 7,466 people who started aspirin of whom 192 had CVD events (2.6%). In this analysis aspirin was estimated to give an ATE of 1.76% (95% CI 1.33 to 2.19%), with a risk ratio of 3.15 (95% CI 2.67 to 3.73, Table 18: post-hoc sensitivity analyses).

Table 18: post-hoc sensitivity analyses

Sensitivity analysis	ATE (absolute risk difference)	95% range*	Risk ratio	95% range*
Interaction between allocation and square of propensity **	0.91%	0.52 to 1.10%	2.51	1.29 to 4.56
Interaction between allocation and square of propensity adjusted ***	0.75%	0.36 to 2.22%	2.23	1.29 to 4.06
Interactions between allocation, propensity, diabetes and atherosclerosis†	0.72%	0.21 to 2.38%	2.33	1.22 to 2.93
Interactions and infections included‡	0.71%	0.17 to 3.40%	2.48	1.31 to 3.15
Prevalent aspirin use (at 6m before index)	0.51%	0.35 to 0.84%	1.62	1.43 to 1.82
Aspirin started 6-3m before infection	0.63%	0.07 to 1.21%	1.81	1.21 to 2.71
Index date aspirin allocated to control	1.76%	1.33 to 2.19%	3.15	2.67 to 3.73

*Sensitivity analyses for propensity Model Two examining more complex modelling were not bootstrapped, so 95% range of results in the population presented. ** without further adjustment. ***further adjusted for age, diabetes, markers of atherosclerosis, chronic heart disease, atrial fibrillation, statins and platelet categories. †adjusted for age, chronic heart disease, atrial fibrillation, statins and platelet categories ‡ Adjusted for the same variables, with additional adjustment for infection diagnosis type. Aspirin exposure sensitivity models used propensity Model Four, which gave better balance in covariates for these modelling tasks.

5.7 Discussion

5.7.1 Summary of findings

I this chapter I set out to explore the potential of aspirin as a potential intervention for patients with a DASHI score of three or more. I developed four propensity models, and assessed their performance, before selecting one for the final analysis.

The main result was that aspirin increased infection-related CVD events by 1.22% (95% CI 1.00 to 1.43%) in absolute terms and increased the absolute risk of bleeding by 0.1% (95% CI 0.01 to 0.20%). I used the same propensity modelling analysis to look for a causal effect of aspirin use on constipation and skin lesions and found no evidence of an effect on either.

I performed post-hoc sensitivity analyses to explore the CVD event results. I increased the complexity of the propensity modelling to attempt to increase the adjustment for the confounders, used different definitions of aspirin exposure to examine the effects of longer exposure periods, and moved the index-date prescriptions to the control arm.

5.7.1.1 Unexpected findings

The increase in the risk of bleeding is similar to what has been seen in trials of short-term aspirin (0.2%).[99] The increase in the risk of CVD events with aspirin is unexpected because it contrasts with the very well-established decrease in CVD event risk by about 20% with aspirin use in many different settings (Introduction section 1.7.1).[8,99,102–104,196]

One explanation for the result is that it is an accurate measurement of a true effect. It is possible that the use of aspirin during acute infections increases the risk of CVD events. One can speculate as to mechanisms. It is possible that aspirin worsened the respiratory symptoms and increased physiological stresses, leading to CVD events. Up to six percent of the population are intolerant of aspirin, particularly those with asthma.[197] An extreme phenotype is aspirin-exacerbated respiratory disease (AERD), which can cause a range of vasomotor and bronchoconstriction symptoms.[198] AERD is uncommon, but conceivably even mild

bronchoconstriction could increase the risk of CVD events. Similarly, aspirin reduces renal function, which conceivably could reduce the physiological reserve required to cope with infections.[199] Overall, I think these possible explanations are unlikely. These effects of aspirin would have to be specific to the period around the infections and less than the benefit at other times, as they have not been seen in trials of long-term aspirin. The strengths and limitations of this study suggest alternative explanations.

5.7.2 Strengths and limitations

One strength of the study is that the large sample size helps provide precise estimates of these relatively uncommon events. Other strengths include setting out the assumptions in advance, being clear that this is an attempt to model the causal effect and developing the propensity models independently before the final estimate.[195,200]

I think that the modelling approach, incorporating diagnostics for the propensity modelling was a strength, as it provided some evidence that the modelling itself was not grossly mis-specified or otherwise problematic. If the propensity modelling was mis-specified, then this could have introduced bias.[201] However, the Objective One results did not point to major problems in the models. There was evidence of interaction, and I have explored the sensitivity of the results to different propensity modelling structures, including using squared terms for the propensity, and various interaction terms. These sensitivity analyses reduced the point estimate of effect compared to the main analysis, but none of these sensitivity analyses reduced the estimate of harm to zero or less (Table 18: post-hoc sensitivity analyses).

5.7.2.1 Exposure variable

A strength of the exposure variable is that prescription codes are automated by clinical records systems, so they are relatively accurate. As with any prescribed medication in clinical practice, I cannot tell if the drugs have been dispensed or consumed, but the clinical decision I hoped to inform with this study is well recorded: to prescribe or not. Another limitation is

that aspirin is available without prescription, and although non-prescribed aspirin is coded in the medical record, coding may not be complete, or timely, leading to underestimation of the exposure. The extent of the misclassification in a similar UK medical records database (THIN) was found to be minimal, although in a small sample size (1 of 69 people who had never used aspirin in THIN reported OTC use).[202] Overall, although there some inevitable misclassification, coded aspirin prescriptions are likely to be closely correlated with real world aspirin use.

I decided to use new aspirin prescribing as the exposure because this was a close approximation to the clinical use scenario of offering aspirin to high-risk people with respiratory infections.[189] A strength of this approach is minimising immortal time bias.[203–205] This bias occurs because to be included in the study people must survive up to the index date without a CVD event, or else they are excluded from the analysis because they are no longer part of the primary prevention population. The people exposed to aspirin must 'survive' whilst exposed to get into the study and the longer the definition period of exposure the greater the bias. The highest risk people will have had events before the index date and will have been excluded from the population. Those with the lowest risk of CVD events will have the longest CVD-event-free survival, even if aspirin had no effect. The longer the window used to define aspirin exposure, the longer the survival needed to enter the study, and the more biased towards a reduction in CVD events the result will be. Defining the exposure as newly prescribed aspirin is therefore conservative. This is consistent with the results of the sensitivity analyses – the increased risk of CVD events with aspirin was lowest for prevalent aspirin users at six months before the index date (absolute risk difference 0.51%, 95% CI 0.35 to 0.84%), and intermediate for those who started aspirin six to three months prior to the index date (ARD 0.63%, 95% CI 0.07 to 1.21) (Table 18: post-hoc sensitivity analyses).

The sensitivity analysis that reclassified people who received aspirin on the index date to the control group increased the estimate of harm. This group had proportionally more events than the other aspirin users, which was not what I expected. In an analysis without propensity modelling increasing the proportion of events in the unexposed group would reduce the risk ratio. The explanation must therefore be due to the modelling process. Two things are modelled – the propensity, and the potential outcomes are modelled. Patients who were prescribed aspirin on day zero have a higher propensity to receive aspirin (they did in fact receive it) than the rest of the control group who did not receive aspirin. Moving them into the control arm will reduce the difference in propensity between the study arms. The modelling attributed more of the difference in the outcomes to the effect of aspirin, because the propensity model was responsible for less of the difference. Examining risk factors and propensity in this group did show this pattern, with levels of risk factors in the day zero group closer to the exposed groups than the unexposed group. The mean propensity was 0.14 in the day zero group, 0.17 in the exposed group, and 0.009 in rest of the unexposed group.

5.7.2.2 Data limitations - Coding, delays in recording events, and reverse causality

Routinely collected medical record databases are imperfect, and this may have biased the results. The study design, with a short duration of follow-up, may have exacerbated some of these biases.

5.7.2.3 Missed events

Routinely collected datasets consist of clinical codes.[116] They don't include the free-text notes GPs (and others) write (or don't), and this loss of information could lead to misclassification.

GPs may record their suspicions or working diagnoses in free text but only commit to a clinical code once there is a more certain diagnosis.[179] In contrast prescription codes are automatically coded and can appear before diagnosis codes.[179] For example, if a patient

presents with chest pains of uncertain cause, and angina is considered likely, aspirin may be prescribed.[83] The aspirin prescription is automatically coded, but suspicion of angina can remain in the free text comments pending further investigations. The GP has recorded their working diagnosis in the clinical record, but as they have not coded a CVD event the research database would show only the aspirin prescription.

5.7.2.4 Delayed event recording

Even when events are recorded with codes there can be a delay between clinical events happening and their coding. The delay is different for prescriptions and diagnoses, and between primary care and secondary care events. Prescriptions and diagnoses made in secondary care are only coded in the primary care record when the discharge letter reaches the practice and someone enters the codes into the primary care record. This might be weeks or months after the event and the delay varies by condition.[206] The 28-day follow-up may have been too short for some hospital events to be recorded.

Some diagnoses are made and coded immediately in primary care. Minor bleeding outcomes mostly fall into this category but CVD events, which are likely to require secondary care and to present directly to hospitals, do not.[84] A study examining the delay in coding of four outcomes of interest for vaccine surveillance in CPRD found delays were longer for secondary care diagnoses.[206] Bell's Palsy, a primary care 'spot diagnosis' was coded on the day of diagnosis in 72% of the cases, and 93% within a month. Febrile seizures are acute events that carers find alarming and usually result in emergency department attendance. Only 39% of febrile seizures were recorded on the day, and 61% within a month. Insidious conditions that require secondary care diagnosis had even longer delays – 33% of Guillan-Barré syndrome diagnoses were recorded on the day, and 45% within a month. Optic neuritis was coded on the day for only 28% of cases, and 76% within a month. All four diagnoses were more than 95% recorded within six months of their occurrence.

I used secondary care linkage to HES, which increases the completeness of the record, and probably reduces the delays. But secondary care records are also usually coded after the discharge of the patient, by a hospital coding team working from the hospital notes and discharge letter, we cannot be sure HES data is free from this effect.[207] With hindsight it would be ideal to have a longer period of follow-up as a sensitivity analysis, but unfortunately this data is not available to me, and such an analysis may be measuring long-term effects of aspirin on CVD events (which is already known) rather than infection-related events.

The choice of new aspirin as the exposure could have exacerbated this effect and the sensitivity analyses were consistent with this - using prevalent aspirin use six months before the infection as exposure gave a lower estimate of harm (Absolute increase in risk of CVD events 0.51%, 95% CI 0.35 to 0.84%) as did using new aspirin starters in the period six to three months prior to the infection (0.63% 95% CI 0.07 to 1.21%).

5.7.2.5 Reverse causality

Delayed recording of events also opens the door to reverse causality. Post CVD event pneumonia is a recognised phenomenon, with 10% of stroke patients being diagnosed with pneumonia during their acute admission.[208] It is possible that some of the prescriptions of aspirin were associated with a CVD event that was not coded immediately. If some of these patients then then later presented with a CVD-event-related respiratory infection, the results would be biased towards overestimating risk with aspirin use. This could be contributing to the reduction of the effect estimates in the sensitivity analyses using earlier aspirin exposures.

5.7.2.6 Other missing data

Aspirin is not recommended for stroke prophylaxis in atrial fibrillation, because it doesn't work and increases the risk of bleeding.[85,209] This practice persists (see introduction section 1.7.1.1) and it is commoner in patients with peripheral arterial disease, dyslipidaemia and

coronary disease.[210] One might also expect a greater amount of missing data and worse coding from those practices with less-than-optimal treatments. If this is true, it would bias the effect estimate towards harm because people taking aspirin inappropriately would be higher risk for CVD events than the modelling would be able to estimate, because of these missing diagnoses.

5.7.2.7 Assumptions for propensity modelling: no interference, consistency, positivity and no unmeasured confounding

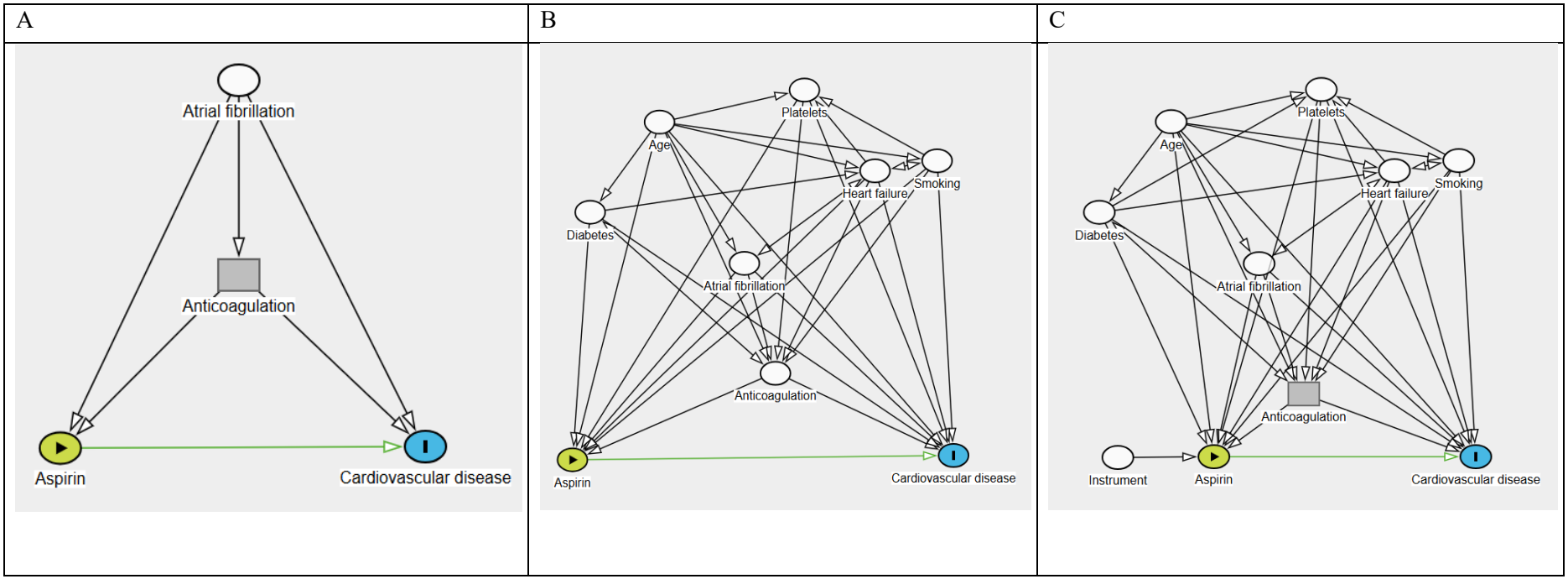
The objective one results assessing the candidate models showed adequate performance in predicting the exposure, and reduced covariate imbalances. Unfortunately, this doesn't necessarily mean they met the underlying assumptions for propensity models.

The assumptions of no interference (treating one person doesn't affect another person's risk) and consistency of effect (the exposure is the same regardless of the method of allocation) seem reasonable for the aspirin intervention in this study.[135]

There is no evidence that positivity (the requirement that the propensity probability is positive for all the participants) has been violated, in that the modelling did not predict zero probability of allocation. Unfortunately, this is not sufficient to eliminate the possibility of violations in some combinations of levels of confounders, and there were many patients with near zero probabilities of being prescribed aspirin.[137,201] The more confounders included in a propensity model the higher the chances of violating positivity, so having chosen a less complex model is a strength.[201] Using truncated and stabilized weights also avoided large weights and reduced the likelihood of violating positivity.[201] It is also possible, or likely, that the population contains people who would never accept aspirin, or who have unmeasured contraindications that mean they should have been excluded. These patients are undetectable and would violate positivity irrespective of any modelling approach.[211]

5.7.2.8 Structure of study - selection bias

There is a trade-off between positivity and selection bias and there is no analytical solution to this within this study design. (Figure 8: Directed Acyclic Graphs).[209,212]



Directed Acyclic Graphs drawn with dagitty.net. Arrows show direction of causal paths. Panel A shows causal associations between atrial arrhythmias (AF), aspirin, anticoagulation and CVD events. Selection on anticoagulation prevents unbiased assessment of the aspirin to CVD event effect. Only adjusting for both anticoagulation and AF can give an unbiased estimate of the effect of aspirin on CVD events. As anticoagulation is a contraindication for aspirin, this would violate the positivity requirement that all the patients are eligible for the exposure. Panel B includes other important confounders – including and adjusting for all the covariates can give an unbiased effect estimate, but violates positivity, as it would include people not eligible for aspirin use. Panel C demonstrates how a valid instrument such as randomisation can estimate an effect irrespective of adjustment or selection by confounders.

Figure 8: Directed Acyclic Graphs

This study has some selection bias because atrial fibrillation (AF) is associated with both aspirin use and anticoagulation use.[210,213] Anticoagulation is a contraindication for aspirin, and reduces CVD events.[85,176] To meet the positivity assumption I excluded patients who were taking anticoagulants on the basis that they would not be eligible to start aspirin. I think this is a reasonable choice, but selection bias may be partly responsible for the unexpected results.[214] The problem arises because patients with AF and who have high risk of stroke should be offered anticoagulation, since 2014 at least, so can only enter the study population if they are treated inadequately (with no treatment or with aspirin).[85] This could induce an association between aspirin and infection-related CVD events in the study population, even if no such association exists in the wider population.[214,215]

Including all the patients and adjusting for both atrial fibrillation and anticoagulation, could remove this selection bias, but including patients not eligible to receive the exposure would violate the positivity requirement. An alternative, excluding patients with atrial arrhythmias, might mitigate this, but this selection criterion also risks opening other confounding causal paths.[214,215] Another possibility would be restricting the analysis to the period before the 2014 NICE guidance that recommended against use of aspirin.[85]

This trade-off between selection and positivity can be avoided with alternative study designs using instrumental variables, for example randomised trials (where randomisation is the instrument) or instrumental variables in observational analyses.[136,216] These study designs replace the assumptions needed for propensity analyses with requirements about the validity of the instrument.[217]

5.7.2.9 Confounding

Another limitation is possible violation of the assumption of no unmeasured confounding, or 'conditional exchangeability'.[212,218,219] It assumes no difference between exposed and unexposed, conditional on the set of confounders being accounted for in the study.[135] If there

are important confounders that are not included, or that are included but with residual confounding, then the exchangeability assumption has not been met, and the results may be biased.[218,219] There is evidence this is a risk in this study.

5.7.2.10 Measured confounding

The imperfect covariate balance in the weighted pseudopopulation means the known confounders have not been completely accounted for by the propensity models. For example, 33% of patients with no aspirin were in the oldest decile of the age range, but this was 39% of the aspirin users. Age is one of the most important risks for CVD events, so residual confounding would bias towards overestimating risk in the aspirin group. Other major risk factors for CVD events - diabetes, atrial arrythmias, and markers of atherosclerosis, were all imbalanced after weighting. Although I also adjusted the final regression model for these factors, it is possible the imbalance was still not adequately addressed.

5.7.2.11 Unmeasured confounding

There may be many unmeasured confounders. An example is suboptimal care.

As discussed earlier, it is also likely that suboptimal care in one field is associated with other suboptimal care practices. A study of GP practices in the UK found higher quality outcome framework (QoF) scores associated with training practices, group practices, and those in less deprived areas.[220] I did not include a term for practice identity in the models, and training and group status was not available. Similarly, in Swedish data, patients with multiple comorbidities were less likely to have guideline-consistent prescriptions than people with single diseases, not least because guidelines are based on single diseases and usually do not account for comorbidity.[213] They found the more comorbidities patients with atrial fibrillation had, the less likely they were to receive anticoagulation.[213] If this is in effect it would bias the results towards harm in the aspirin group. There may be many other quality associations like

this that are unknown, cannot be accurately modelled, and could lead risk being overestimated.

5.7.2.12 Bleeding and confounding

Most clinical characteristics that increase the risk of bleeding also increase the risk of CVD events.[8] If there is confounding of the aspirin to CVD event relationship, I would expect it to increase the risk of bleeding in line with CVD event risk. Long-term aspirin has been shown to increase the risk of bleeding, but a meta-analysis of trials of short-term aspirin (<=10 days treatment) found no increase in serious bleeding, ulcers or perforation of the GI tract.[8,221] The estimated increase in the risk of bleeding is what I had expected given the long-term trial evidence, but it could also be attributed to uncontrolled confounding.

5.7.3 Comparison with prior studies

The estimated effect of aspirin on CVD events is not compatible with meta-analyses of clinical trials of long-term aspirin use, which show a 20% relative reduction in CVD events.[8] Nor is it comparable with the results from the single trial of aspirin in people hospitalized with pneumonia, which suggested one month of 300mg aspirin might prevent acute coronary syndrome.[103] Observational evidence from a self-controlled case series also showed a reduction in coronary events over six months in people with pneumonia who were long term aspirin users.[104]

Many aspirin trials excluded older adults, who would have higher CVD event risk, but there is evidence that the overall effect of aspirin, and the trade-off between risks and benefits is similar in older people. The ASPREE trial randomised over 19,000 older people (> 65 years or >70-year-old white people) to aspirin 100mg or placebo.[196] Although it was not powered for CVD event outcomes, they found aspirin had a non-significant point estimate of a 20% reduction in cardiovascular deaths (Hazard ratio 0.82 (95% CI 0.62 to 1.08)) and an increase in haemorrhage (HR 1.13 (95% CI 0.66-1.94)).[196]

The RECOVERY platform trial tested daily aspirin 100mg until discharge as an intervention in people hospitalized by Covid-19.[102] More than 90% of the population were treated with low molecular weight heparin (LMWH) during their admissions (6,885/7,351, 91% of those allocated aspirin and 7,028/7,541 93% of those allocated to usual care).[101] This is very different to primary care, where only a very tiny proportion of people are treated with LMWH. In the subgroup without antithrombotic therapy there was a signal towards reduced 28-day mortality (RR 0.83, 95% CI 0.66-1.04) and an increase in successful hospital discharge (RR 1.19, 95% CI 1.01 to 1.41). Similarly, in the whole trial population aspirin caused a significant increase in live discharge within 28 days (RR 1.06, 95% CI 1.02 to 1.10), with no benefit in those with higher dose LMWH (RR 1.01, 95% CI 0.94 to 1.08), and a lesser estimate for those on standard dose LMWH.[101]

A recently updated Cochrane review found the difference in results between randomised trials and observational studies is small.[222] However, this study may fall into the group of observational studies with different results to clinical trials.[223] There are various reasons for the discrepancies.

A famous example of the effects of hormone replacement therapy (HRT) on CVD events was due to duration of exposure.[224] Randomised trials showed an increase in the risk of CVD events in the first years of HRT use, but observational studies seemed to show a protective effect instead.[224] Once it was appreciated that the effect of HRT might vary with the amount of time people were exposed, and this was accounted for in observational studies, they agreed with the trials - there was an increase in CVD events, particularly in older women.[224] Duration of exposure is unlikely to be a major problem in this study as there is no reason to think the effects of aspirin on CVD events are cumulative. The absolute benefit from aspirin on CVD events probably increases slowly over time, as the underlying risk accumulates with

age. It does not seem likely that this increases enough to cause measurement errors over the short duration of exposure in this study.

The HRT controversy went the other way for the association with breast cancer outcomes - trials estimated there was no risk from oestrogens, but the observational evidence showed an increased risk of breast cancer.[224] In trials of HRT the women were mostly post-menopausal for some time before being exposed to oestrogens, whereas the observational evidence was based upon initiation at or near menopause. Once the trials were analysed considering both cumulative oestrogen exposure before the trials started, and time since menopause, they agreed with the observational studies – there is an increased risk of breast cancer on starting HRT, which ameliorates with time. A difference between exposure to oestrogen, a hormone with systemic effects, and aspirin, is that the total dose of aspirin exposure would not be expected to be protective against CVD events – the mechanism of action is to prevent acute thrombosis, not to prevent the chronic build-up of atheroma. Another major difference is that the risk of CVD events or breast cancer with oestrogen therapy is an unintended side effect – so there is no confounding by indication, unlike for the intended protective effects of aspirin in CVD events.[223] The bleeding result would not be affected by these biases and so may be a more reliable measurement.

5.7.4 Implications for practice

As aspirin is an intervention associated with hazardous bleeding, there needs to be good evidence of benefit outweighing this risk before using it. This study has not provided this evidence but instead estimated aspirin to be harmful in terms of both CVD events and bleeding events. On current evidence clinicians should not offer short term aspirin during acute respiratory infection for primary prevention of CVD events, unless part of a trial.

5.7.5 Implications for research

Estimating the effects of aspirin on infection-related CVD events needs another approach, that relies on different assumptions. The most convincing method would be instrumentation by randomisation in a clinical trial, and it seems likely that a trial would be needed to change practice. The evidence from this study might give pause for thought before initiating a trial, but I do not think that it is strong enough evidence to prevent a trial, particularly when viewed in the context of prior evidence with contrary findings.

5.7.6 Conclusions

The results suggest the question of antiplatelets for respiratory infection is not settled. The results can be viewed in the context of the previous chapters. The DASHI score can identify people at higher risk of infection-related CVD events, but causal questions are more challenging than prediction. The choice of a short-term outcome is appropriate for prediction, because it gives the meaningful association with the respiratory infection, but this may cause problems in these causal analyses because the ascertainment of aspirin use has less delay in recording in the CPRD data than outcome events.

There are many other explanations for why the results of this study could be incorrect. One of the most important explanations is unmeasured or incompletely mitigated confounding by indication. If a GP perceives a patient has high CVD event risk, they are both more likely to have aspirin prescribed, and to have infection-related CVD events. Aspirin is an attractive candidate intervention for infection-related CVD events because it is not indicated for the primary prevention population, so could be a meaningful change to practice. This also makes observational analysis more difficult because a significant proportion of those taking it should either have been offered alternatives, or to have been excluded from the primary prevention population.

6 Chapter Six: Statins and acute infection-related cardiovascular events: a retrospective cohort study

"*One of the things I've said before in interviews is: 'Without deviation (from the norm), 'progress' is not possible." In order for one to deviate successfully, one has to have at least a passing acquaintance with whatever norm one expects to deviate from."*

Frank Zappa, The real Frank Zappa book, 1989

6.3 Introduction

6.3.1 Overview within thesis – how this chapter fits in

The overarching theme of this thesis is investigating infection-related cardiovascular events, and their prevention. In chapters three and four I described prediction modelling and validation, resulting in the DASHI score. Next, in chapter five, I used the DASHI score to identify higher risk patients (with a DASHI score >=3) and explored the effect of aspirin on infection-related CVD events. The results were unexpected – aspirin seemingly increased the risk of cardiovascular events. This chapter describes how I examined the effect of statins on infection-related CVD using propensity modelling. Statins are cardioprotective, but unlike aspirin they are indicated in primary prevention because they have a better balance of benefits and risks. The study population is different to the previous chapters. It includes people with a DASHI score of one or more to reflect the low risks of statin use. I also explored alternative weighting methods, and a used a different approach to implementing propensity modelling.

The initial results showed these approaches had variable effectiveness in balancing covariates between the weighted study arms. For the final analysis I chose to use the most comprehensive propensity model and overlap weighting, a combination which best balanced the covariates. In the final analysis I estimated an absolute increase in the risk of infection-related CVD events in people taking statins, a result that I explored further with sensitivity analyses.

6.3.2 Rationale

Statins are effective in the long-term but also have acute effects on atherosclerotic plaques.[225,226] Plaques physically regress with statin therapy, so that the regression is detectable on MRI imaging after six months of treatment with simvastatin.[225] Statins also help transform plaques from fatty lesions into the stabler, calcified form, which can be demonstrated on CT scans.[226] Statins should be started early in acute coronary syndromes, as

trials have shown early intensive statin therapy reduces mortality compared to later initiation.[227] However, guidelines don't recommended statins for use in acute respiratory infections.[10,68] I could find no observational or trial evidence regarding the short-term effects of statins on CVD following acute respiratory infections in primary care.

The previous chapter, chapter five, demonstrated some of the difficulties estimating causal effects in observational data. I have used these difficulties to inform the approach in this chapter. Compared to aspirin, statins have several attractive properties for analyses of this kind. Clinically statins are attractive because they are effective at preventing CVD, and they have a very favourable side effect profile.[109] Unlike aspirin, statins are indicated in primary prevention of CVD.[10,19] This means there is no suspicion that people taking statins should have been offered another treatment. There are few contraindications, so there is no need to exclude people taking anticoagulants or other medications.[9] Statins are more widely prescribed than aspirin, so the numbers should help with precision.[228] I also used insights from the objective one results in the previous chapter to inform the methods to try to mitigate the imbalances of covariates. Another attraction to statins is that some of the variability is likely to be due to causes other than the indication. Confounding by indication is always a problem in observational studies of medications, but this may be less for statins because they are relatively unpopular with patients, so are frequently declined, introducing variability less subject to confounding by indication.[229] This could mean lower levels of misclassification, and fewer differences in indication status between people taking the medication and not. As there is no need to exclude people with anticoagulation, the concerns about balancing positivity and selection bias are also reduced.

This chapter therefore describes how I sought to measure the effect of statins on infection-related CVD in CPRD data.

6.4 Methods

6.4.1 Chapter aim

The overall aim of this chapter is to measure the effect of starting statins on infection-related CVD events.

6.4.2 Chapter objectives

1. To develop and assess candidate propensity models and weighting methods to predict the propensity for being prescribed statins
2. To use as single propensity model and weighting method to estimate the effect of statins on infection-related CVD events

6.4.3 Data source

I used data from CRPD Aurum.[157] As described in the previous chapters, the cohorts were from data in the Clinical Practice Research Datalink (CPRD) databases from 1999 to 2019 and linked to ONS and HES datasets (section 2.3.1).

The CPRD Independent Scientific Advisory Committee (ISAC) approved the protocol for all the CPRD work in this thesis (21_000380), but the statin analyses were approved as an amendment, also prior to the analyses being undertaken.

6.4.4 Population

I designed the study as if to inform a potential future clinical trial and so the included population consisted of people who could be recruited into a theoretical randomised trial of statins for primary prevention of infection-related CVD.[189] The population had to be eligible to receive the potential intervention, have a reasonable risk of outcomes, and not have contraindications.[189] As with the prior chapters patients enter the study at the diagnosis of a respiratory infection, the index date, and are followed for 28 days (Chapter Three: Figure 1: Patient timeline in cohorts).

Inclusion criteria were people over the age of forty years, with no prior cardiovascular disease, presenting with a respiratory infection. They had to have a DASHI score of one or more. I used a lower threshold of one point (equivalent to 0.08% 28-day risk) for this chapter because the risks associated with statins are very low, so a trial could justify exposing patients with a relatively low risk of infection-related CVD.[109]

Exclusion criteria were: people already established on repeat prescriptions of statins six months before the index date, those with prior CVD diagnoses, those with a DASHI score of zero, and those under the age of 40 years.

6.4.5 Exposure

To decide on the exposure, I considered what I would suggest for a future trial: starting a statin at the diagnosis of respiratory infection, for example 40mg of atorvastatin daily, for 28 days, (patients would have the option to continue long term in accordance with the guidelines for primary prevention).[10] In the observational data I used newly starting statin therapy, meaning a first prescription on the index date or up to six months prior. I included any statin for several reasons, firstly atorvastatin only became widely used in the NHS after becoming available as a generic in 2011, secondly statins have a class effect on CVD outcomes, reducing the long term risk of CVD by about 20% for each 1mMol reduction in LDL cholesterol.[105] Patients needed to be exposed to statins at the index date, so if X was the number of days before the index date that it was prescribed, the prescription needed to be for more than X doses of statin to ensure they were still able to take it on the index date (i.e. I expected a 28 day prescription to have run out had it been prescribed more than 28 days before the index date). To identify prevalent users I measured the time from six months before the index date to the first prescription date. As repeat prescriptions were mostly for the same number of doses, patients were considered current users if the days prior to the prescription could have been covered by a prescription for the same number of doses (i.e. if a

repeat prescription for 28 tables was made less than 28 days after the start of a six month period ending on the index date, I assumed it was a continuation). The statin prescriptions were identified in the primary care record using CPRD codes, and these were derived from the list of statins in the British National Formulary.[9] Codelists are online (https://github.com/Protocols-For-Research/CPRD-codes-CVD-infection-risk).

6.4.6 Control

The control group were people who did not start a statin prior to the index date. People starting statins following the index date were allocated to the control arm because I would expect people with new diagnoses of CVD would be started on secondary prevention medications, including statins.

6.4.7 Outcomes

6.4.7.1 Primary outcome

The primary outcome was infection-related CVD events. This is the same composite of events as in the other chapters (Section 3.4.7).

6.4.7.2 Negative control outcomes

I used the same two negative control outcomes as in chapter five – constipation, and skin lesions (section 5.4.7). These outcomes are not causally related to statin use, but are likely related to frequency of attendance, age and other medications that may be associated with CVD.[9]

6.4.8 Statistical methods

6.4.9 Missing data

Multiple imputation

As described in chapter five, I used multiple imputation methods to deal with non-binary missing data (section 5.5.3) using Stata's 'MI' suite of commands to generate 20 imputations. [147]

6.4.10 Missing indicators

As described previously, multiple imputation was not suitable or practical for all the missing variables, for these I used missing indicators. I did this for variables where I thought missingness was informative, and likely to be poorly explained by the data in the dataset, or impractical to impute (section 5.5.3). These were CRP and platelet count as previously, but also total cholesterol to HDL ratio.

I included both the imputed cholesterol to HDL ratio and a missing indicator because of the likely implications of missingness on exposure status. Doctors often prescribe statins because of a cholesterol result.[230] The guideline specified primary prevention pathway involves testing cholesterol, then using the cholesterol to HDL ratio in a QRISK model and if the 10-year risk is above 10% offering a statin.[10] Having no cholesterol result is likely to be associated with different statin prescribing behaviour. I therefore added this missing indicator to the dataset. I also used the imputed levels of cholesterol:HDL ratio, because these are known to be strongly associated with the outcome.[108]

6.4.11 Propensity model overview

As described in the previous chapter, the propensity score approach is to model the exposure allocated to patients. If this is successful, then one can account for the propensity to be exposed, and obtain an estimate of the effect of exposure on the outcome.[135,192] I have followed the approach suggested by Brookhart *et al*; the first part of the methods describes building multiple candidate propensity models for aspirin exposure, and assessing their performance, the details of which are in the supplementary materials (Appendix section 10.3).[121,139] In this chapter I also examined different weighting methods to implement the modelling, and their effect on covariate balance.[139,143,146,231,232] The second part describes how, after selecting one propensity model and one weighting method, I estimated the effects of statins on infection-related CVD.

6.4.12 Objective one methods: propensity modelling and weights

6.4.12.1 Overview of methods for objective one

As in the previous chapter I used four sets of variables to build four candidate propensity models before choosing one for the final analysis (Appendix section 10.3).[195] To allow for nonlinearity I first converted continuous variables to categorical variables, with ten categories (deciles). To group variables for the four candidate propensity models, I assessed the strength of association between the variables and both the CVD outcome and the statin exposure with logistic regression.[121] Group one variables were demographics, groups two had greater magnitude of association with CVD events than statin use, group three variables had stronger associations with statins than CVD events. Group four was CRP and platelets.

6.4.12.2 Weights

I explored using different weighting systems. I used the propensity models to generate stabilized inverse probability weights and then truncated inverse probability weights in the same way as the previous chapter (section 5.5.4).[232] I also calculated overlap weights (OW) using the formula w = 1-p for those exposed, and w = p for the unexposed, where w is the weight for each individual and p is the estimated propensity (probability) for that individual to receive a statin.[143] Overlap weights are the probability of being allocated to the opposite group.[143] They have several helpful characteristics. They downweigh the extremes and increases the contribution of those most likely to be allocated to either group – the region of overlapping propensity. This is likely to be the most clinically interesting group who might be treated either way, and those who are closest to meeting the required assumptions for propensity analyses.[135] Also, as logistic regression was used to estimate the propensity, overlap weighting is guaranteed to balance the included covariates in the pseudopopulation.[143]

For each of the four candidate models I:

1. Predicted the probabilities (propensity) for each individual
2. Examined the overlap of probabilities between exposed and unexposed groups
3. Looked for evidence of violation of the positivity assumption and interaction in quartiles of propensity
4. Estimated the apparent (internal) C statistic
5. Produced apparent calibration plots
6. Generated
    a. Stabilised inverse probability weights
    b. Truncated stabilised inverse probability weights
    c. Overlap weights
7. Examined the mean and range of the weights
8. Examined the covariate balance before and after weighting with IPW and overlap weights
9. Assessed the crude effect of statins on the negative controls constipation and skin lesions

Further details of the methods for these steps are in the supplementary materials (Appendix section 11).

6.4.13 Objective two methods: Final modelling of exposure-outcomes association

I selected one of the four candidate propensity models and weighting methods based on the model performance being adequate, and satisfactory covariate balance being achieved. I then used the propensity model to perform weighted logistic regression.[192] Because the procedure I chose was not implemented by Stata's built in commands, I used bootstrapping to estimate confidence intervals, with 500 resampling cycles.[145] The analytical procedure I used is outlined in supplementary materials (Appendix, section 11).

6.4.13.1 Post-hoc sensitivity analyses

I undertook post-hoc sensitivity analyses, to investigate the robustness of the results to changing the inclusion criteria. I undertook a complete case analysis as a less computationally expensive approach. I then undertook complete case analysis with an interaction term. Finally, I used different cut-off values of DASHI score for the population to which I applied the final model, from a DASHI of one to eleven, again in complete cases.

6.5 Results

6.5.1 Population

The population comprised 2,457,350 patients, of whom 13,530 (0.6%) started statins before their first infection (Table 19: Characteristics of participants by statin status). CVD events occurred in 7,921 (0.3%) overall, and in 110 of those prescribed statins (0.8%). Statin use was associated with older age (mean 62 years vs 58 years), higher systolic blood pressure (137mmHg Vs 132mmHg) and aspirin use (13% vs 1%). Mean DASHI score was 2.57 (SD 1.33) in statin starters, compared to 2.25 (SD 1.41) in non-statin users.

Table 19: Characteristics of participants by statin status

Variables	Total		No statin exposure		Statin exposure	
Continuous	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Age in years	57.82	13.79	57.80	13.80	61.96	11.55
Cholesterol to HDL ratio	3.95	1.25	3.94	1.24	4.54	1.52
Systolic BP mmHg	132.39	17.58	132.35	17.58	136.92	17.41
BMI KgM^{-2}	27.38	5.71	27.36	5.71	29.34	5.91
DASHI score	2.25	1.41	2.25	1.41	2.57	1.33
Categorical	**n**	**%**	**n**	**%**	**n**	**%**
CVD event	7,912	0.3%	7,802	0.3%	110	0.8%
Negative controls:						
Skin lesions	17,094	0.7%	16,989	0.7%	105	0.8%
Constipation	4,206	0.2%	4,176	0.2%	30	0.2%
Aspirin use	27,430	1.1%	25,635	1.0%	1,795	13.3%
Other antiplatelet use	5,359	0.2%	5,163	0.2%	196	1.4%
Anticoagulant use	41,944	1.7%	41,561	1.7%	383	2.8%
Any bleed outcome	7,261	0.3%	7,218	0.3%	43	0.3%
Major bleed outcome	573	<0.01%	566	<0.01%	7	0.1%
Female	1,395,708	56.8%	1,389,227	56.8%	6,481	47.9%
Index infection:						
URTI	1,290,028	52.5%	1,282,406	52.2%	7,622	56.3%
LRTI	1,105,109	45.0%	1,099,496	44.7%	5,613	41.5%
Influenza	169,497	6.9%	168,786	6.9%	711	5.3%
Pneumonia	62,213	2.5%	61,918	2.5%	295	2.2%
Smoking status:						
Nonsmoker	607,420	24.7%	602,341	24.5%	5,079	37.5%
Ex-smoker	407,315	16.6%	403,615	16.4%	3,700	27.3%
Light smoker	114,407	4.7%	113,691	4.6%	716	5.3%
Moderate smoker	151,754	6.2%	150,941	6.1%	813	6.0%
Heavy smoker	108,365	4.4%	107,735	4.4%	630	4.7%
Current smoker	247,967	10.1%	246,493	10.0%	1,474	10.9%
Smoking data missing	820,122	33.4%	819,004	33.3%	1,118	8.3%
Diabetes	138,150	5.6%	133,693	5.4%	4,457	32.9%
Heart failure	22,874	0.9%	22,687	0.9%	187	1.4%

Chronic heart disease	27,616	1.1%	27,325	1.1%	291	2.2%
Atrial arrhythmia	43,658	1.8%	43,243	1.8%	415	3.1%
Markers of atherosclerosis	142,170	5.8%	139,532	5.7%	2,638	19.5%
Cancer	85,502	3.5%	85,010	3.5%	492	3.6%
Family history of CVD	13,258	0.5%	13,096	0.5%	162	1.2%
Antihypertensive use	388,187	15.8%	381,685	15.5%	6,502	48.1%
Rheumatoid Arthritis	33,083	1.3%	32,852	1.3%	231	1.7%
Platelets x 10^9/L:						
Platelets unknown	1,822,284	74.2%	1,816,414	73.9%	5,870	43.4%
Thrombocytopenia (<150)	20,844	0.8%	20,663	0.8%	181	1.3%
Platelets normal (150-450)	599,774	24.4%	592,423	24.1%	7,351	54.3%
Thrombocytosis (>450)	14,448	0.6%	14,320	0.6%	128	0.9%
C Reactive Protein mg/L						
CRP unknown	2,298,007	93.5%	2,285,955	93.0%	12,052	89.1%
CRP <5	76,581	3.1%	75,853	3.1%	728	5.4%
CRP 5 to <20	63,508	2.6%	62,910	2.6%	598	4.4%
CRP >=20	19,254	0.8%	19,102	0.8%	152	1.1%
Index of Multiple deprivation:						
Least deprived decile	274,395	11.2%	273,315	11.2%	1,080	8.0%
Most deprived decile	223,592	9.1%	222,104	9.0%	1,488	11.0%
IMD missing	61,038	2.5%	60,675	2.5%	363	2.7%
Cholesterol missing	1,699,005	69.1%	1,696,985	69.1%	2,020	14.9%
Total N	2,457,350	100%	2,443,820	99.4%	13,530	0.6%

Statin exposure is a new prescription of statins continued through the index date and started in the six months before the index date. CVD = composite outcome (myocardial infarction, coronary syndromes, transient ischaemic event, stroke, ischaemic cardiomyopathy) in the 28 days following respiratory infection. All patients have a Respiratory Tract Infection (RTI) categorised into upper (URTI), lower (LRTI) and pneumonia. Influenza is a separate, non exclusive category and can be included separately in addition to LRTI (default for influenza, unless coded to another site). Heart failure includes all non-ischaemic diagnoses. Chronic heart disease incluses valvular disease, hypertensive disease and congenital disease. Atrial arrhythmias include atrial tachicardias, atrial fibrillation, and flutter. Diabetes mellitus includes type i, type ii, and other/unrecorded type. Markers of atherosclerosis includes: chronic kidney disease, peripheral arterial disease, and erectile dysfunction. Cholesterol : HDL ratio is serum total cholesterol/serum HDL cholesterol. Family history of CVD is in first degree relatives. Percentages are of complete variable (excluding missing).

6.5.2 Objective one results

6.5.2.1 Candidate propensity model variables

Variables more strongly associated with CVD events than statin use were smoking, atrial arrhythmias, heart failure, rheumatoid arthritis, other heart diseases (congenital and valvular) and cancers. These were included in models two through four. Variables more strongly associated with statin prescribing were included in models three and four only (BMI, total cholesterol:HDL cholesterol, systolic BP, markers of atherosclerosis, family history of early CVD and diabetes). Model four also included recent CRP and platelet counts (Table 20: Groups of variables added sequentially in propensity models for statin exposure).

Table 20: Groups of variables added sequentially in propensity models for statin exposure

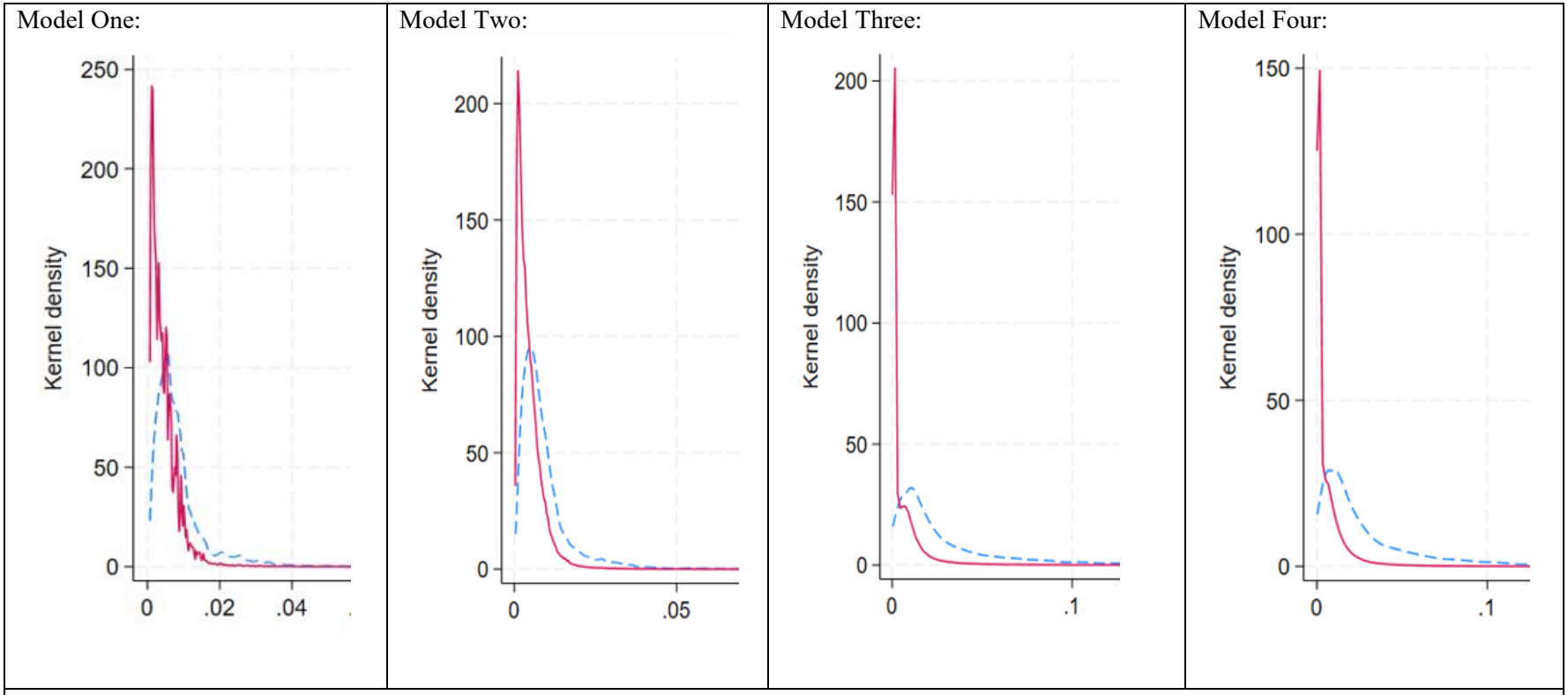
Propensity models using group	Group Description	Clinical variables in group
Models 1, 2, 3 and 4	Demographics	Age, sex, index of multiple deprivation decile, and ethnicity
Models 2, 3 and 4	Confounders more strongly associated with cardiovascular event outcomes than statin exposure	Smoking, atrial arrythmias, heart failure, rheumatoid arthritis, valvular and congenital heart disease, and cancer diagnoses
Models 3 and 4	Confounders more strongly associated with statin exposure than cardiovascular event outcomes	Body mass index, total cholesterol to HDL cholesterol ratio, missing indicator for cholesterol measurement, systolic blood pressure, markers of atherosclerosis, family history of early CVD, and diabetes diagnosis
Model 4 only	Recent blood tests	Platelet count, C reactive protein

Models one through four are propensity models for statin exposure at diagnosis with respiratory infection diagnosis in primary care. Statin exposure defined as prescriptions started within six months before the index date and ongoing at the index date. Cardiovascular outcomes are a composite outcome (myocardial infarction, coronary syndromes, transient ischaemic event, stroke, ischaemic cardiomyopathy) in the 28 days following respiratory infection

6.5.2.2 Propensity probability overlap

The four models all had regions of overlapping propensity for the exposed and unexposed, with long tails for the distributions and larger ranges of predicted probability for the more complex Model Three and Model Four (Figure 9: Histograms of propensity overlap by statin exposure).

Figure 9: Histograms of propensity overlap by statin exposure for propensity models one to four



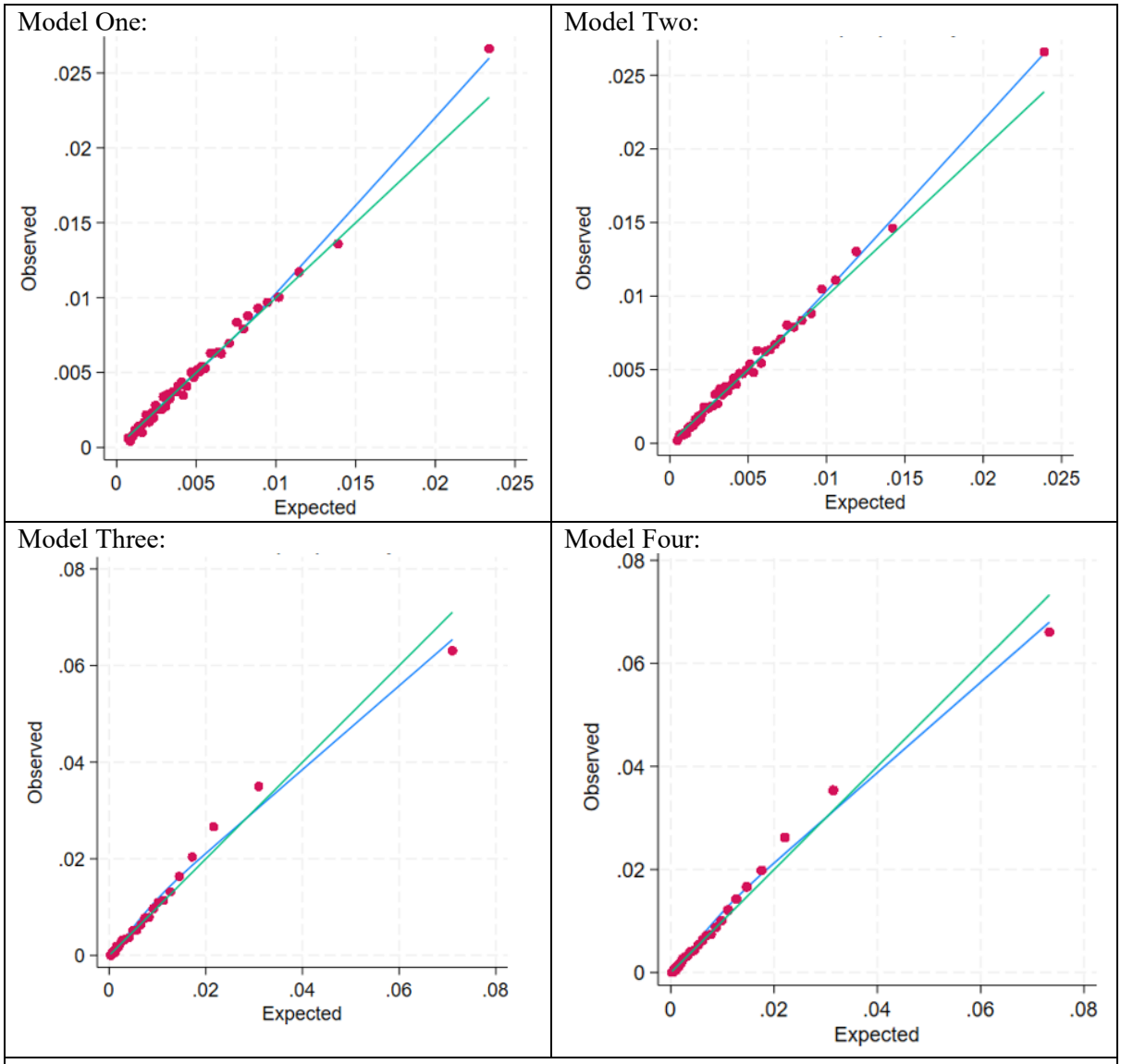
Histograms display overlap of propensity models by exposure status for each of four sequential models. Blue dashed line is modelled probability in the exposed population, red solid line is probability in those unexposed. Model One is demographics only (age, sex, IMD, ethnicity), Model Two with demographics and further confounders (Smoking, atrial arrythmias, heart failure, rheumatoid arthritis, valvular and congenital heart disease, and cancer diagnoses), Model Three includes further confounders added (Body mass index, total cholesterol to HDL cholesterol ratio, missing indicator for cholesterol measurement, systolic blood pressure, markers of atherosclerosis, family history of early CVD, and diabetes diagnosis), Model Four includes additional recent blood tests (CRP, platelets). Kernel density is smoothed estimate of probability density (in units of 1/propensity where area under curve = 1). Note different axis scales.

6.5.2.3 Assessment of predictive properties of the propensity models

The estimated C statistics showed similar performance in discriminating between those exposed to statins and those unexposed for Models One and Two (0.73 and 0.73 with overlapping confidence intervals), and Models Three and Four (0.87 and 0.88 with overlapping confidence intervals). (Table S38: Discrimination of propensity models, internal estimates).

Internal calibration of the propensity models showed similar good calibration in the range of up to 10%, with only very minor overprediction in Models Three and Four at higher predicted probabilities (Figure 10: Calibration curves for propensity models).

Figure 10: Calibration curves for propensity models for statin exposure



Axes: probability of exposure. Calibration curves for propensity models, internal calibration. Green line is ideal performance; blue line is a cubic spline of actual performance, fit to the red markers, which represent 50ths of predicted exposure risk plotted against actual exposure in these groups. One; model with demographics only (age, sex, IMD, ethnicity), Two; model with demographics and further confounders (Smoking, Cholesterol:HDL ratio, Systolic BP), Three; model with further variables added (BMI, Chronic heart disease, peripheral atherosclerosis), Four; model with recent blood tests added (CRP, platelets).

6.5.2.4 Weights

The distribution of stabilised inverse probability weights was satisfactory for all the models, with means close to one. The highest weight was 26.5 in Model Four (Table S39: Weight distributions by weighting type and propensity model). Truncated stabilised IP weights were closer to one, with a maximum of 1.056 in propensity Model Four. Overlap weights were bounded by zero and one as expected.

6.5.2.5 Covariate balance under candidate propensity models

The raw data had covariate imbalance of greater magnitude than 0.1 for some of the age categories, sex, index of multiple deprivation, smoking status, systolic blood pressure and CRP (Table S41: Covariate balance by models with stabilised truncated inverse probability weights). There were imbalances of greater magnitude than 0.3 for higher BMI categories. The highest imbalances were in diabetes (1.21), markers of atherosclerosis (0.59) and cholesterol to HDL ratio (-1.18 for the missing indicator). I initially compared the models using truncated inverse probability weights (Table S41: Covariate balance by models with stabilised truncated inverse probability weights). Model Four reduced the imbalances for the most variables under this weighting system, but large imbalances remained in some of the variables, the standardised mean difference was 0.75 for the missing indicator for cholesterol to HDL ratio, and 0.37 for missing platelets, with 0.35 for diabetes and normal platelets.

6.5.2.6 Covariate balance with different weighting systems

All the weighting systems reduced the covariate imbalances between the groups of patients defined by exposure. Stabilized inverse probability weights without truncation reduced the covariate imbalances more than the truncated version, for example it gave a standarised mean difference of magnitude 0.1 or less for the cholesterol measurements and missing indicator. The most complete balance was from using overlap weighting, which achieved standardised mean differences with magnitude of <0.1 for each confounder included in each of the models (Table 21: Covariate balance by propensity model and weighting system). Details of the

weighted pseudo-population estimated with overlap weighting are in the supplementary materials (Table S42: Covariate balance by propensity model with overlap weighting).

Table 21: Covariate balance by propensity model and weighting system

Propensity model	Raw data			Model One weighting			Model Two weighting			Model Three weighting			Model Four weighting		
Variable and categories	Percent no statin	Percent in statin	Standardised difference	stabilised IPW	truncated stabilised IPW	Overlap	stabilised IPW	truncated stabilised IPW	Overlap	stabilised IPW	truncated stabilised IPW	Overlap	stabilised IPW	truncated stabilised IPW	Overlap
Age															
Decile 1	9.54	3.44	-0.21	0.01	-0.20	0.00	0.00	-0.20	0.00	-0.13	-0.21	0.00	-0.13	-0.21	0.00
Decile 2	8.51	3.53	-0.18	0.00	-0.16	0.00	-0.01	-0.16	0.00	-0.09	-0.17	0.00	-0.09	-0.17	0.00
Decile 3	9.63	5.06	-0.15	0.00	-0.12	0.00	-0.01	-0.12	0.00	-0.11	-0.15	0.00	-0.11	-0.16	0.00
Decile 4	8.00	5.42	-0.10	-0.02	-0.07	0.00	-0.02	-0.06	0.00	-0.06	-0.10	0.00	-0.06	-0.10	0.00
Decile 5	9.38	7.94	-0.05	-0.01	-0.01	0.00	-0.01	-0.01	0.00	-0.02	-0.06	0.00	-0.02	-0.06	0.00
Decile 6	8.32	8.95	0.02	-0.01	0.04	0.00	-0.01	0.04	0.00	0.03	0.00	0.00	0.04	0.00	0.00
Decile 7	12.29	16.27	0.12	0.00	0.10	0.00	0.00	0.09	0.00	0.02	0.08	0.00	0.01	0.08	0.00
Decile 8	12.02	19.29	0.22	0.01	0.12	0.00	0.01	0.12	0.00	0.07	0.15	0.00	0.07	0.15	0.00
Decile 9	10.75	19.42	0.28	0.01	0.12	0.00	0.02	0.13	0.00	0.10	0.19	0.00	0.10	0.20	0.00
Decile 10	11.55	10.68	-0.03	0.01	0.04	0.00	0.02	0.04	0.00	0.11	0.08	0.00	0.12	0.08	0.00
Female Sex	56.85	47.90	-0.18	-0.03	-0.09	0.00	-0.02	-0.09	0.00	-0.03	-0.07	0.00	-0.03	-0.07	0.00
Index of multiple deprivation															
Decile 1	11.47	8.20	-0.10	-0.03	-0.05	0.00	-0.03	-0.05	0.00	-0.01	-0.05	0.00	-0.01	-0.05	0.00
Decile 2	10.37	8.10	-0.07	-0.03	-0.03	0.00	-0.02	-0.03	0.00	0.00	-0.03	0.00	0.01	-0.03	0.00
Decile 3	10.40	8.07	-0.08	0.01	-0.03	0.00	0.00	-0.03	0.00	-0.02	-0.04	0.00	-0.02	-0.04	0.00
Decile 4	10.64	8.52	-0.07	0.00	-0.03	0.00	-0.01	-0.03	0.00	-0.04	-0.04	0.00	-0.04	-0.04	0.00
Decile 5	9.33	8.92	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	-0.02	-0.01	0.00	-0.03	-0.01	0.00
Decile 6	9.90	10.34	0.01	-0.01	0.00	0.00	-0.01	0.00	0.00	-0.01	0.01	0.00	-0.01	0.00	0.00
Decile 7	9.93	10.64	0.02	-0.01	0.01	0.00	-0.01	0.01	0.00	0.03	0.01	0.00	0.02	0.01	0.00
Decile 8	9.24	12.42	0.11	0.01	0.04	0.00	0.02	0.04	0.00	0.03	0.05	0.00	0.03	0.05	0.00
Decile 9	9.41	13.49	0.14	0.02	0.05	0.00	0.02	0.05	0.00	0.01	0.05	0.00	0.01	0.05	0.00
Decile 10	9.32	11.30	0.07	0.02	0.03	0.00	0.02	0.03	0.00	0.00	0.03	0.00	0.00	0.03	0.00
Ethnicity															
Bangladeshi	0.45	1.69	0.18	0.01	0.04	0.01	0.02	0.04	0.01	0.01	0.04	0.00	0.01	0.04	0.00
Black African	1.61	3.81	0.18	0.01	0.04	0.03	0.02	0.05	0.03	0.01	0.06	0.01	0.00	0.06	0.01

Black Caribbean	1.39	3.19	0.15	0.03	0.05	0.03	0.03	0.06	0.03	0.03	0.07	0.01	0.02	0.07	0.01
Chinese	0.30	0.52	0.04	0.01	0.02	0.01	0.01	0.02	0.01	-0.01	0.01	0.00	-0.01	0.01	0.00
Indian	2.13	5.68	0.25	0.03	0.07	0.04	0.03	0.07	0.04	0.05	0.09	0.01	0.06	0.09	0.01
Other Asian	1.05	2.86	0.18	0.02	0.05	0.03	0.02	0.05	0.03	0.01	0.06	0.01	0.01	0.06	0.01
Other enthicity	1.81	3.07	0.09	0.01	0.04	0.03	0.02	0.04	0.03	0.02	0.04	0.01	0.02	0.04	0.01
Pakistani	1.14	3.63	0.23	0.02	0.05	0.03	0.02	0.05	0.03	0.03	0.07	0.01	0.03	0.07	0.01
White	90.13	75.56	-0.49	0.24	0.17	0.17	0.24	0.17	0.17	0.06	0.10	0.04	0.04	0.09	0.03
Smoking status															
Never smoked	37.07	40.92	0.08	0.20	0.23	0.18	0.24	0.26	0.21	0.12	0.21	0.04	0.11	0.21	0.04
Ex-smoker	24.84	29.81	0.12	0.21	0.26	0.19	0.18	0.24	0.16	0.14	0.23	0.02	0.13	0.23	0.02
Light smoker	7.00	5.77	-0.05	0.05	0.03	0.04	0.06	0.04	0.05	-0.01	0.01	0.01	-0.01	0.01	0.01
Moderate smoker	9.29	6.55	-0.09	0.05	0.01	0.04	0.06	0.02	0.05	-0.01	-0.01	0.02	-0.02	-0.02	0.02
Heavy smoker	6.63	5.08	-0.06	0.07	0.04	0.05	0.06	0.03	0.04	0.01	0.00	0.02	0.02	0.01	0.01
Smoking unknown quantity	15.17	11.88	-0.09	0.08	0.05	0.07	0.06	0.04	0.07	0.03	0.03	0.02	0.03	0.03	0.01
Cholesterol to HDL ratio															
Decile 1	3.07	5.45	0.14	0.11	0.11	0.10	0.12	0.12	0.10	-0.06	0.06	-0.10	-0.05	0.06	-0.09
Decile 2	2.65	4.86	0.14	0.10	0.11	0.10	0.11	0.11	0.10	0.01	0.13	0.00	0.01	0.12	0.00
Decile 3	3.03	5.40	0.14	0.09	0.11	0.09	0.10	0.11	0.09	0.01	0.13	-0.01	0.01	0.14	-0.01
Decile 4	3.11	5.88	0.16	0.12	0.13	0.10	0.12	0.13	0.10	0.02	0.16	0.00	0.02	0.15	0.00
Decile 5	3.17	6.73	0.20	0.13	0.15	0.13	0.14	0.15	0.13	0.02	0.16	0.00	0.03	0.16	0.00
Decile 6	3.17	7.42	0.24	0.16	0.18	0.15	0.16	0.18	0.15	0.03	0.17	0.00	0.03	0.17	0.00
Decile 7	3.19	8.51	0.30	0.21	0.22	0.19	0.21	0.22	0.19	0.03	0.18	0.01	0.04	0.18	0.00
Decile 8	3.31	10.57	0.41	0.28	0.29	0.25	0.28	0.29	0.25	0.04	0.20	0.01	0.04	0.20	0.01
Decile 9	3.03	12.22	0.54	0.35	0.35	0.32	0.35	0.35	0.31	0.05	0.20	0.01	0.05	0.20	0.01
Decile 10	2.83	18.03	0.92	0.58	0.54	0.49	0.56	0.53	0.49	0.07	0.24	0.03	0.07	0.24	0.03
Missing indicator	69.44	14.93	-1.18	-1.29	-1.29	-1.16	-1.28	-1.30	-1.14	-0.09	-0.75	0.02	-0.10	-0.75	0.02
Systolic Blood Pressure															
Decile 1	9.51	5.70	-0.13	-0.12	-0.14	-0.08	0.06	-0.08	0.07	-0.03	-0.08	0.01	-0.04	-0.09	0.01
Decile 2	5.50	5.03	-0.02	0.00	-0.02	0.00	0.07	0.01	0.06	0.02	0.01	0.01	0.02	0.01	0.01
Decile 3	7.64	6.31	-0.05	-0.02	-0.04	-0.03	0.03	-0.02	0.02	0.00	-0.03	0.00	0.00	-0.03	0.00
Decile 4	6.74	8.18	0.06	0.07	0.05	0.06	0.07	0.06	0.06	0.01	0.05	0.01	0.01	0.05	0.01
Decile 5	10.67	12.25	0.05	0.06	0.05	0.04	0.05	0.06	0.04	0.03	0.05	0.01	0.03	0.05	0.01
Decile 6	4.67	7.58	0.14	0.11	0.12	0.10	0.06	0.10	0.06	0.04	0.08	0.01	0.03	0.08	0.01
Decile 7	13.95	20.06	0.18	0.14	0.16	0.11	0.07	0.13	0.05	0.07	0.13	0.01	0.08	0.13	0.01
Decile 8	2.53	4.34	0.12	0.08	0.09	0.07	0.05	0.08	0.04	0.02	0.08	0.01	0.02	0.08	0.01
Decile 9	8.87	12.92	0.14	0.11	0.13	0.08	0.07	0.12	0.04	0.08	0.11	0.01	0.08	0.11	0.01
Decile 10	9.04	15.22	0.22	0.16	0.19	0.12	0.09	0.15	0.06	0.09	0.15	0.01	0.09	0.15	0.01
Heart Failure	0.93	1.38	0.05	0.03	0.04	0.03	0.01	0.03	0.00	0.07	0.06	0.00	0.07	0.06	0.00

Cancer	3.48	3.64	0.01	-0.01	0.01	-0.03	0.03	0.03	0.00	0.04	0.05	0.00	0.05	0.05	0.00
BMI															
Decile 1	5.44	3.70	-0.08	-0.10	-0.09	-0.09	-0.08	-0.09	-0.07	-0.08	-0.06	-0.07	-0.08	-0.07	-0.07
Decile 2	5.01	4.67	-0.02	-0.03	-0.03	-0.02	-0.02	-0.02	-0.02	0.01	0.00	-0.03	0.01	0.00	-0.03
Decile 3	4.89	5.63	0.03	0.01	0.02	0.01	0.02	0.02	0.02	0.01	0.03	-0.01	0.00	0.03	-0.01
Decile 4	4.85	6.28	0.07	0.03	0.04	0.04	0.04	0.05	0.04	0.00	0.05	0.00	-0.01	0.05	0.00
Decile 5	4.71	7.58	0.14	0.08	0.10	0.09	0.09	0.10	0.09	0.06	0.10	0.03	0.05	0.10	0.03
Decile 6	4.67	7.39	0.13	0.09	0.11	0.08	0.09	0.11	0.08	0.07	0.09	0.01	0.06	0.09	0.00
Decile 7	4.52	8.82	0.21	0.15	0.16	0.14	0.16	0.17	0.14	0.08	0.13	0.04	0.07	0.13	0.04
Decile 8	4.44	9.75	0.26	0.20	0.20	0.18	0.18	0.20	0.17	0.11	0.16	0.06	0.10	0.16	0.06
Decile 9	4.29	9.79	0.27	0.23	0.22	0.20	0.22	0.21	0.19	0.10	0.16	0.04	0.10	0.16	0.04
Decile 10	4.14	10.61	0.33	0.34	0.28	0.26	0.32	0.27	0.24	0.12	0.17	0.05	0.12	0.17	0.05
Chronic heart disease	1.12	2.15	0.10	0.07	0.08	0.07	0.01	0.03	0.00	0.07	0.07	0.00	0.07	0.08	0.00
Atherosclerosis	5.71	19.50	0.59	0.38	0.41	0.36	0.37	0.40	0.35	0.16	0.25	0.00	0.16	0.26	0.00
Atrial arrhythmia	1.77	3.07	0.10	0.06	0.09	0.06	0.01	0.04	0.00	0.11	0.11	0.00	0.11	0.11	0.00
Diabetes mellitus	5.47	32.94	1.21	0.79	0.72	0.68	0.78	0.72	0.68	0.15	0.34	0.00	0.16	0.35	0.00
Family history	0.54	1.20	0.09	0.08	0.08	0.07	0.08	0.08	0.07	0.02	0.06	0.00	0.02	0.06	0.00
Platelets unknown	74.33	43.39	-0.71	-0.63	-0.64	-0.60	-0.64	-0.64	-0.59	-0.34	-0.54	-0.23	-0.12	-0.37	0.01
Platelets low	0.85	1.34	0.05	0.03	0.03	0.02	0.03	0.03	0.02	0.02	0.03	-0.05	0.05	0.05	0.00
Platelets normal	24.24	54.33	0.70	0.62	0.63	0.59	0.62	0.63	0.58	0.33	0.53	0.24	0.10	0.35	0.00
Platelets high	0.59	0.95	0.05	0.05	0.05	0.04	0.05	0.05	0.04	0.05	0.04	0.01	0.03	0.04	0.00
CRP unknown	93.54	89.08	-0.18	-0.16	-0.16	-0.15	-0.17	-0.16	-0.15	-0.06	-0.14	0.00	-0.05	-0.14	0.00
CRP <5	3.10	5.38	0.13	0.11	0.11	0.11	0.12	0.11	0.11	0.03	0.10	0.02	0.01	0.09	0.00
CRP 5 to <20	2.57	4.42	0.12	0.10	0.10	0.09	0.11	0.10	0.09	0.05	0.09	-0.01	0.05	0.09	0.00
CRP 20+	0.78	1.12	0.04	0.03	0.04	0.02	0.03	0.04	0.03	0.01	0.03	-0.02	0.02	0.05	0.00

Covariate balance in pseudopopulations created with overlap weighting. Percentages were calculated as the mean of the binary variable over the population, including those with missing values in the denominator. Chronic heart disease incluses valvular disease, hypertensive disease and congenital disease. Atrial arrhythmias include atrial tachicardias, atrial fibrillation, and flutter. Diabetes mellitus includes type i, type ii, and other/unrecorded type. Markers of atherosclerosis includes: chronic kidney disease, peripheral arterial disease, and erectile dysfunction. Family history of CVD is in first degree relatives. Cholesterol : HDL ratio is serum total cholesterol/serum HDL cholesterol

6.5.2.7 Overall assessment of the propensity models

I selected propensity Model Four to take forward for the final analysis and performed this final analysis using overlap weighting. This was because Model Four included all the major confounders, had very good calibration and discrimination, and overlap weighting achieved the best performance in terms of covariate balance.

6.5.3 Objective two results

6.5.3.1 Estimate of effect of statin therapy on infection-related CVD events and negative controls

The estimated baseline risk of CVD events without statin therapy (the potential outcome mean in the untreated, POM_0) was 0.32% (95% CI 0.31 to 0.32%), the potential outcome mean in the treated (POM_1) was 1.01% (95% CI 0.77 to 1.30%), and the average treatment effect (ATE or Risk Difference) estimate was 0.69% (95% CI 0.45 to 0.98%) (Table 22: Estimated effects of statin therapy on CVD and negative outcomes). The risk ratio for CVD events was 3.17 (95% CI 2.41 to 4.08) for those taking statins.

Skin lesions and constipation were not associated with statins in the raw data, nor was an effect estimated by the final propensity model (RR 1.21; 95% CI 0.86 to 1.38 and 1.35 95% CI 0.8 to 2.01 respectively (Table 22: Estimated effects of statin therapy on CVD and negative outcomes).

Table 22: Estimated effects of statin therapy on CVD and negative outcomes

	Estimate with no statin		Estimate with statin		Estimated effect due to statin		Estimated effect due to statin	
Outcome	POM	95% CI	POM	95% CI	ATE	95% CI	RR	95% CI
CVD events	0.32%	0.31 to 0.33	1.01%	0.77 to 1.30	0.69%	0.45 to 0.98	3.17	2.41 to 4.08
Skin lesions	0.70%	0.69 to 0.71	0.78%	0.60 to 0.96	0.08%	-0.09 to 0.26	1.12	0.86 to 1.38
Constipation	0.17%	0.17 to 0.18	0.23%	0.15 to 0.34	0.06%	-0.02 to 0.17	1.35	0.88 to 2.01

Results after overlap weighting with predictions of statin exposure from propensity Model Four, with further adjustment for interaction. Estimate with no statin models every patient if they were not taking statin, estimate with statin is the estimate from modelling all patients as taking a statin. POM potential outcome mean, ATE average treatment effect. RR risk ratio. 95% CI is 95% Confidence Intervals obtained by bootstrapping the imputed datasets 500 times.

6.5.3.2 Sensitivity analyses

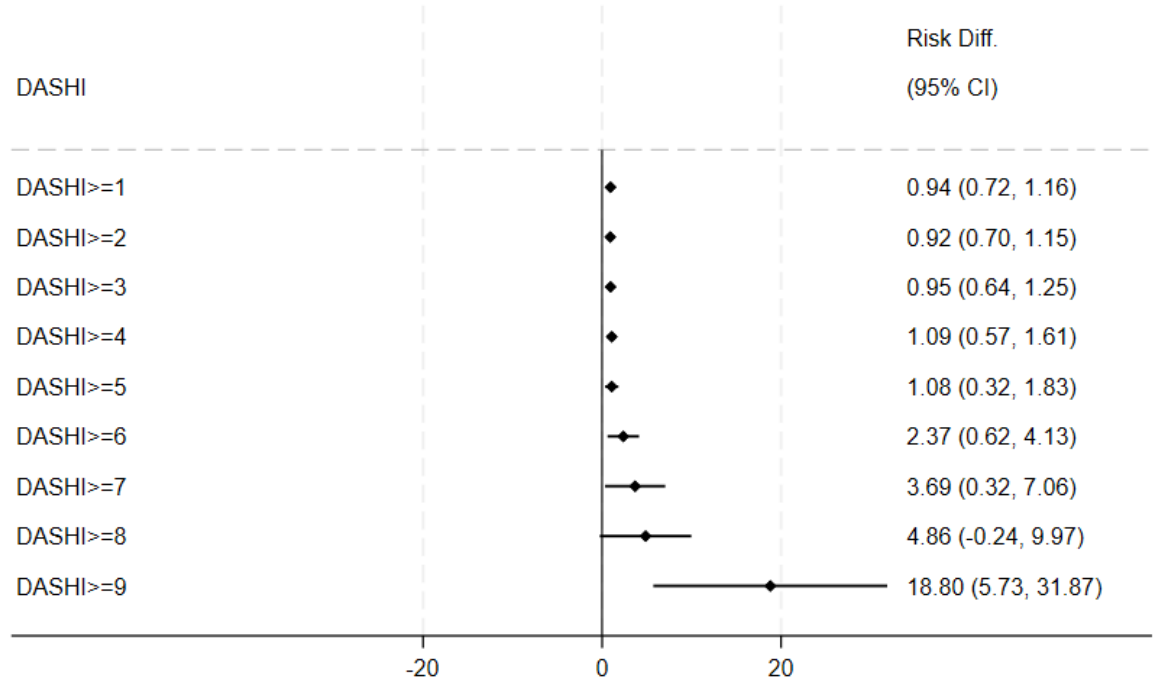
Post hoc sensitivity analyses to explore the effects of interaction terms and DASHI threshold made some difference to the magnitude of results, but the estimates continued to be in the same direction as the main result, indicating harm. A complete case analysis gave a similar estimate to the full modelling using multiple imputation (0.57%, 95% CI 0.44 to 0.70%), despite including only 1.2 million patients (Table 23: Sensitivity analysis results). Including an interaction term between the outcome and the propensity prediction increased the estimated risk difference to 0.94% (95% CI 0.73 to 1.15%). The point estimates for harm from statins were greater as the DASHI increased from one to nine (18.8%, 95% CI 5.73 to 31.87). Restricting the analysis to patients with higher DASHI scores failed with DASHI >=10 due to small numbers. (Table 23: Sensitivity analysis results, and Figure 11: Sensitivity analyses restricting by DASHI score).

Table 23: Sensitivity analysis results

Sensitivity analysis	N in analysis	ATE (absolute risk difference, %)	95% CI
Complete case analysis	1,246,787	0.57	0.44 to 0.70
Complete case with interaction between allocation and propensity	1,246,787	0.94	0.73 to 1.15
Complete case with interaction, minimum DASHI scores as inclusion criteria			
DASHI >= 2	754,839	0.92	0.70 to 1.15
DASHI >= 3	386,143	0.95	0.64 to 1.25
DASHI >= 4	117,502	1.09	0.57 to 1.61
DASHI >= 5	83,268	1.08	0.32 to 1.83
DASHI >= 6	30,206	2.37	0.62 to 4.13
DASHI >= 7	11,391	3.69	0.32 to 7.06
DASHI >= 8	6,929	4.86	-0.24 to 9.97
DASHI >= 9	1,656	18.80	5.73 to 31.87

Sensitivity analyses using complete case analyses, logistic regression with propensity adjustment, with bootstrapping for confidence intervals, 500 selection cycles intended, min achieved 470 for DASHI >=5. Analysis not possible for DASHI of 10 or more due to low numbers.

Figure 11: Sensitivity analyses restricting by DASHI score



Axis: Absolute risk difference (%) in CVD outcomes within 28 days of respiratory infection

Visual representation of estimated effect of new statins on 28-day infection-related CVD risk, by different populations defined by minimum DASHI score. Results are from complete case analyses, logistic regression with propensity adjustment, with bootstrapping for confidence intervals, 500 selection cycles intended, min achieved 470 for DASHI >=5. Analysis not possible for DASHI 10 or more due to low numbers. Risk diff = risk difference calculated as average treatment effect.

6.6 Discussion

6.6.1 Summary of findings

In this chapter I used propensity modeling methods to estimate the effects of statins on infection-related CVD. I derived four candidate propensity models, which were acceptable in terms of discrimination and calibration. I investigated different weighting systems for achieving covariate balance. I selected the most comprehensive model and overlap weighting to estimate the final effect. The overall effect estimate was a small absolute increase in the risk of CVD in people taking statins in the 28 days following respiratory infection, compared to those not taking statins (Absolute risk difference/Average Treatment Effect (ATE) 0.69%, 95% CI 0.45 to 0.98%). This was approximately tripling the risk ratio (3.17, 95% CI 2.41 to 4.08).

This finding was not in the direction expected. It is possible that it is an accurate measurement, and statins are harmful in the context of acute respiratory infections, but this seems unlikely. Existing evidence is that statins reduce cardiovascular disease in primary and secondary prevention.[11,233] The previous data comes from multiple randomized controlled trials and is very robust. The trials are of long-term statin therapy, which necessarily includes periods around acute respiratory infection, so if the findings of this study represent an accurate measurement of effect, it must be offset by a greater protective effect at other times. An alternative, more likely explanation, is that my result is an inaccurate measurement.

6.6.2 Strengths and limitations

A strength of this study is the use of a large dataset, including more of the population and allowing estimation of rare events with some statistical precision. Propensity modelling is a robust approach, that acknowledges the question is causal, the challenges that this brings, and gives results interpretable as a risk-difference on the same scale as trial results.[195,200]

The drawbacks of my approach may have outweighed the benefits. As discussed in the previous chapter, propensity modelling relies on adequate modelling and several assumptions being met. I examined these possible causes for error.

The propensity modelling has performed well in terms of classification and calibration, so there were no grossly unacceptable models. Unfortunately, the purpose of propensity modelling is not only prediction, and so unlike in chapters three and four, good predictive performance is not sufficient to allow accurate estimates.[201] Propensity prediction can even be too good - perfect prediction would be equivalent to just using the raw exposure data without any modelling.

6.6.2.1 Data limitations

This study suffers from some of the same data limitations as the previous chapter (section 5.7.2). Delays in coding diagnoses may have affected the results, compounded by the immediate recording of prescription codes.[116,207] Whilst statins are indicated in primary prevention, there may be suspected vascular events that prompt both assessment of CVD risk and statin prescribing, but which are recorded in free-text and may be missing from the codes, or coded later.[179] Delays and missing data would both lead to overestimation of the risk in the group exposed to statins.

6.6.2.2 Assumptions for propensity modelling: no interference, consistency, positivity and exchangeability

The assumption of no interference between patient's risk seems safe for statins.[135] There are interventions that contravene no interference – for example vaccination of one patient also protects others, but I would not expect treatment of one patient with statins would change the risk of someone else in the cohort. Consistency of effect – that the effect should be the same despite how the exposure is allocated, also seems a safe assumption (unlike for example the effect of being allocated to a smoking cessation support group versus the effect in motivated people seeking out that same support).[135]

Unfortunately, the choice of methods involved a trade-off between the exchangeability assumption and the positivity assumption.[135] Balancing the measured confounders was only achieved with the overlap weighting derived from the most complex propensity model. The more complex the model, the greater the chance of patients with zero (or one) probability of allocation to exposure. This was seen in the results of the weightings, although covariate balance was achieved with the overlap weighting, there were patients with near zero weighting. I chose this approach because of the excellent covariate balance, and because overlap weights minimize the influence of those in the least overlapping regions of propensity, but this risks violation of the positivity assumption. Making a different choice, prioritizing the positivity assumption by using a less complex model or a different weighting system could have led to less risk of violating positivity. Unfortunately, it would also have led to imbalances in confounders, violating the conditional exchangeability requirement. The population also includes people who would never accept statins (every clinician has met these folks), which is also an unavoidable violation of positivity under this study design.

6.6.2.3 Structure of study - selection bias

Compared to the aspirin study of the last chapter, there is less risk of selection bias. There are few contraindications for statins, and they are indicated in primary prevention. However, there were people in the study with very low probability of receiving statins. One approach would be to have tighter selection criteria, to identify people at greater likelihood of being allocated to statins. This could be achieved by including only patients with an elevated cholesterol to HDL ratio for example. The difficulty with this is that it would then introduce the same problem found in the aspirin study – the only people eligible to start statins would be those who should already be taking them and are therefore a higher risk group than usual, being under-treated or refusing treatment. The sensitivity analysis looking at people with higher DASHI scores, so higher risk of CVD, appeared to show increased absolute risk of

events attributed to statin use at higher baseline risk. This is the opposite to what would be seen if statins give a reduction in risk proportional to the underlying risk, as they have in the trials.[11] This is strongly suggestive of a biased result (although it could be interpreted as statins increasing the underlying risk).

6.6.2.4 Exposure variable

New statin use is likely to be very well recorded and is less likely to be taken as an over-the-counter medication than aspirin.[202] As with the aspirin exposure, using new prescriptions is a tradeoff between minimizing immortal time bias, but worsening the effect of delayed recording of events.[205] An additional problem with statins is that their mechanism of action is to address atheroma formation. Whilst there is evidence of relatively short-term effects of statins, it is possible that the time-period in this study is too short to have an effect, or at least a measurable effect. However, in an unbiased study this would only result in an estimate of no effect and not give the increase in risk estimated here.

I also considered the possibility of designing a different study, comparing people who had declined or discontinued statins with those taking them.[234] This should reduce confounding by indication but introduces difficulties in modelling the ways that these patients are different to people who meekly accept their statins. This might include refusal of other medications, perceived levels of CVD risk, and factors not found in the medical record. As discussed in the previous chapter suboptimal care is associated with factors not available in the dataset, such as training practice status and group practice status, and reasons for discontinuation are also lacking.[220,234–236]

6.6.2.5 Comparison with prior studies

There are no trials of statins for infection-related CVD specifically, but there is evidence that statins do not reduce the risk of infections. [110,111]

High-quality clinical trials have shown statins are effective in both primary and secondary prevention.[66,105,233] A meta-analysis of high intensity LDL reduction trials found that after one year there was a 15% reduction in CVD deaths at one year with a 10% reduction in deaths per 1mMol/L reduction in LDL.[105] These were highly significant results, and the effectiveness of statins is beyond doubt. However, the absolute risk at one year was only 4.0% in the control arms, with a risk difference of 0.8. This is equivalent to a baseline risk of 0.3% over 28 days (similar to this study), with a risk difference of six events in 10,000 people over 28 days, if the risk were evenly distributed throughout the year (which it is not). Large numbers of participants are needed to identify this small amelioration of risk. However, this is not an explanation for the results in this study, which is large, and showed harm from statins rather than no effect.

6.7 Conclusions

6.7.1 Implications for practice

Prior research clearly shows statins are life-saving medications with an excellent safety profile, and that they should be used widely.[10,19,237] There was no evidence of a particular benefit to starting statins at respiratory infections, and this study has not provided evidence of such a benefit. The clinical implications are that current practice should not change based on this study. Instead, statins should be introduced at the earliest opportunity for primary prevention, used in sufficient doses as to reduce the lifetime risk of CVD as much as possible, and maximised for secondary prevention.[19]

The purpose of risk prediction models is to identify groups of patients, so they are only useful when there is some uncertainty about the groups. This only makes sense in the context where some people are not offered the intervention, and those who are offered the intervention are not immediately identifiable. As examples, in secondary CVD prevention there is no need for a tool, nor is a tool required for people with diabetes.[10,86] This may become the case for

primary prevention as well. The threshold for offering statins for primary prevention has reduced over the years in UK guidance – it used to be 20% 10 year CVD risk, then 10%, and now, because statins are cost effective at any level of risk, to anyone who wishes to take them.[10,194] As the threshold decreases it approaches what may be the most efficient scenario: to offer statins to everyone without bothering to assess CVD risk. This in the context of the leading cause of death globally being CVD.[10] This approach is advocated by poly-pill enthusiasts who advocate offering statins combined with multiple antihypertensives to everyone over the age of 50.[238] This more public health based approach would also cover the large majority of people with elevated DASHI score, and therefore the question of effectiveness in this groups becomes of minor interest, probably relating only to a small group of people who are unwilling to take statins long-term, but might acquiesce for a shorter time period, or people under the age of 50 with lower respiratory tract infections.

6.7.2 Implications for research

The effect of statins in acute respiratory infections is not clear, and alternative methods could be used to investigate further. A clinical trial would be the gold-standard way to identify the effect of statins in infections, but it would need to be large and consequently expensive. The current evidence base is probably insufficient to provide the equipoise needed for a trial – obtaining further observational data might provide evidence either in favour or against further investigation.

Using routinely collected data has challenges. Study designs relying on the positivity and exchangeability assumptions could provide more evidence, but there is likely to be a trade-off between these. The results from the different weighting systems show the difficulty in choosing specific methods for implementing propensity modelling methods. Truncating weights may be more appropriate where the range of weights includes extreme values, but in this study truncating produced more imbalanced pseudopoulations than using the full range of

weights. Researchers should use truncation cautiously, particularly in studies where the exposure is uncommon.

Confounding by indication is a particularly difficult problem, because risk of the outcome is directly related to the prescribing decision (unlike for pharmacovigilance studies looking for side effects, which are unintended and may be unknown and unpredictable).[223,224] However, there may be natural experiments that could be used, for example examining different thresholds for statin prescribing, or variation in the introduction of atorvastatin to the NHS in 2012 after coming off patent.[228] These, or prior prescribing preference, could be used as instruments, which would be subject to different assumptions.[239–241]

Another design for investigating the effect of medications is combining self-controlled case series in prior event-rate ratio analyses (PERR and an alternative PERR-ALT).[104,242] To minimise unmeasured confounding these designs compare the event rate in periods of time before and after exposure, within the same people. This is repeated with people who were not exposed. The idea is that the difference in event rate between the groups reflects confounding. In PERR-ALT the rate ratio for the exposed group is divided by the rate ratio from unexposed group to obtain the effect estimate.[242] The need to identify exposed and unexposed groups of people requires identifying and matching patients at similar risk of being exposed at a particular time. These methods also rely on the presence of prior events, which means they can only be used in high risk scenarios, reduce the available sample size, and cannot be used for deaths.[242] These different approaches have attractive properties and could be applied to the use of statins, but are not without their problems - secondary prevention would have to be included, which is a population for which statins are already indicated.

6.7.3 Overall conclusion

These results, together with a lack of previous studies, means there is no evidence that initiating statins at or shortly before respiratory infections is beneficial for preventing CVD in the following weeks.

7 Chapter Seven: Discussion

"All people wear out and die, this is natural; but nowhere did people wear out so quickly as here in our Russia"

Maxim Gorky, My Childhood, 1914

7.1 Introduction

The overall aim of the work in this thesis was to investigate the potential for preventing infection-related CVD events in primary care. To do this I first attempted to identify people at higher risk of infection-related CVD by prediction modelling, followed by validation of the models. Then I used the prediction models to identify cohorts at increased risk of respiratory infection-related CVD events. I investigated the effects of first aspirin, then statins, on infection-related CVD events, using regression combined with propensity modelling.

In this chapter I give a summary of the previous chapters, the overall findings and the implications as I see them. I discuss the results in the context of prior research and assesses the strengths and weaknesses of the evidence this work has produced. I discuss the clinical implications of the results, and the research implications that I believe this work points towards.

7.2 Summary of main findings

7.2.1 Prediction modelling

The first research chapters, three and four, described deriving and validating the DASHI score for predicting infection-related CVD event risk in primary care patients. For most primary care patients with respiratory tract infections the absolute risk of CVD events in the next 28 days is low (0.3%) and DASHI estimates risks between 0.04% for zero points and 36% for eleven points. DASHI is the only validated model for this clinical scenario, and as it is published (Appendix – Publication based on chapters three and four) it is available for use in clinical practice and research.[243] DASHI is simple and has good performance in terms of calibration and discrimination, although the positive predictive values were limited by the low background prevalence.[243]

Having developed and validated a way to identify people at increased risk, the natural next question was to ask if anything could be done to prevent people going on to have CVD events.

7.2.2 Causal inference epidemiology

The approach I took was to try to measure the effect of aspirin and statins in retrospective cohorts of people at higher risk of CVD events. I used the DASHI score to identify patients, and regression with propensity modelling to estimate causal effects. The effect of aspirin on infection-related CVD events appeared to be harmful (ATE 1.22%, 95% CI 1.00 to 1.43, RR 2.52, 95% CI 2.26 to 2.81). Aspirin also seemed to cause an increase in bleeding (ATE 0.10%, 95% CI 0.01 to 0.20, RR 1.31, 95% CI 1.06 to 1.63). I also estimated statins increase infection-related CVD events (ATE 0.69%, 95% CI 0.45 to 0.98%, RR 3.17, 95% CI 2.41 to 4.08).

Taken at face value these results put the primary care clinician in an uncomfortable position– able to tell someone their risk of a CVD event, but also to inform them the potential remedies appear to be worse than doing nothing.

7.3 Changes in clinical practice and context

The medical world has moved on since this project was originally conceived and funded in 2019. There have been many changes to the treatment and prevention of CVD, and for the risk factors that make up DASHI. Prevention of respiratory infections has also changed, and there has been a pandemic.

7.3.1 Covid-19

I have been working on this project part-time since April 2020, shortly after the start of the first Covid-19 lockdown in the UK. The emergence of a new respiratory infection was a demonstration that respiratory infections are important, and sometimes of critical importance. However, despite this novelty, the old patterns reasserted themselves; diabetes mellitus, heart failure, and age were found to be risk factors for severe Covid-19 and death.[16,177,244] CVD deaths during the pandemic increased, people with CVD were at high risk of severe disease, and Covid-19 caused infection-related CVD events which were barely recognised clinically.[16,245,246]

DASHI was derived and validated using data from a long period of time (1999-2019), which included the 2009-10 influenza A H1N1 pandemic, and ongoing variation in the circulating strains and severity of respiratory infections. However, DASHI has not been validated in this post-Covid-19 world, and it is possible that it would perform differently now. The first waves of the pandemic were more deadly than later infections because immunity, both natural and from vaccination, reduced the risk of severe disease.[247] More severe Covid-19 was higher risk for poor outcomes, and as this presented as pneumonia or LRTI they would also have a higher DASHI score. In this context the rudimentary definitions of upper and lower respiratory tract infections as practiced in UK primary care are helpful for DASHI – they are proxies for severity and are used post Covid-19 much as they were before the pandemic. The waning deadliness of Covid-19 could lead to miscalibration of prediction models that do not

have markers of severity in them – derivation in data from the first waves would lead to over-prediction of severe outcomes in the later waves. A testable hypothesis is that by focussing on the clinical picture DASHI would predict higher average risk in the earlier waves, and lower average risk in the later ones. DASHI is not affected by the changing 'variants of concern', or the vaccination status or infection history of the individual, only the clinical presentation. It may therefore remain reasonably well calibrated, but only more validation studies can tell if this is the case. Revalidation in a routinely collected dataset would first mean making decisions to classify clinical codes for various Covid-19 presentations into the appropriate classes of infection.

7.3.2 Changes to NHS vaccination recommendations

Changes include the introduction of Covid-19 vaccines (including the advent of mRNA vaccine technology), respiratory syncytial virus (RSV) vaccines, and a reduction in the age for routine influenza vaccines in the UK to 50 years from 65.[64] These changes may have reduced the number of severe respiratory infections that occur (relative to the potential number in the untreated at least), and subsequently infection-related CVD.[29,248] Public health is usually the most efficient way to reduce disease, reducing the age at which vaccination is offered gradually increases the number of at-risk people included.[248] However, people still get ill with respiratory infections and go on to have CVD events, and it is these clinical presentations that this thesis was attempting to address.

7.3.3 Changes in CVD event prevention with statins

Since the start of this thesis NICE changed their guidance to its most inclusive yet:

"Do not rule out treatment with atorvastatin 20 mg for the primary prevention of CVD just because the person's 10-year QRISK3 score is less than 10% if they have an informed preference for taking a statin…" [10]

Essentially allowing anyone who wishes to take statins to do so, irrespective of their estimated risk. Even before this change statin prescribing was increasing. From January 2020 to November 2024 UK atorvastatin prescriptions have increased by about a quarter, to nearly six million per month.[228] This will reduce CVD events and also reduce the opportunity for statins to be used for infection-related CVD event risk, even in the context of a trial.

7.3.4 Changes to medical treatments for diabetes and heart failure

Recent trials have shown the benefits of SGLT2 (Sodium-Glucose Cotransporter-2) inhibitors and GLP-1 (Glucagon-like peptide-1) agonists for CVD event prevention and in type two diabetes and CKD, and they are used much more commonly.[161,228] SGLT2 inhibitors reduce CVD deaths in trial populations at higher risk due to heart failure, diabetes or chronic kidney disease by about 20%.[249] SGLT2 inhibitors are also now established treatments for heart failure with preserved ejection fraction, a condition with a particularly stubborn lack of options until recently.[161,250] Since January 2020 dapagliflozin prescriptions have increased to about six times the baseline in the UK.[228]

There has been a rise in the use of GLP-1 receptor agonists, these medications can be used for diabetes and for weight loss.[9,228] Meta-analysis of trials of GLP-1 receptor agonists in diabetes suggest a reduction in myocardial infarction and stroke of about 10%.[251] A trial in people with overweight or obesity but not diabetes found a 20% reduction in a composite of CVD events and deaths.[252] NHS Semaglutide prescriptions have increased by about five times since January 2020, but it is difficult to tell how widespread the use of GLP-1 receptor agonists for weight loss is, as much of it seems to be private prescribing.[228] In the USA GLP-1 agonist use increased by about seven times between 2019 and 2023, with a decreasing proportion of patients having diabetes (71% Vs 88%).[253]

Overall, the effect of these changes may have been to reduce some of the risk factors for CVD events, which likely includes infection-related CVD events. Patients taking agents that reduce CVD event risk have lower risks to mitigate. This would reduce the absolute benefit of aspirin in primary prevention, analogous to the reduction in absolute benefit seen in trials of aspirin once statin therapy was introduced.[8]

7.4 Results in context

7.4.1 DASHI in context

The first objectives, to derive and validate prediction models, have been met by DASHI. It is possible there are other risk factors or other combinations that would give a more effective risk prediction model, although finding them was outside the scope of this thesis.

The DASHI score is the only score of its type, and the only validated way to assess the 28-day risk of CVD events in primary care patients with RTI. However, DASHI was built on the back of the experience of hundreds of pre-existing long-term CVD event prediction models.[7] This is why the variables that make up DASHI are unsurprising - they were selected by clinical experts, from a list I derived from prior research. This process was bound to result in conformity. The benefit of conformity is that the variables appear rational and reasonable to clinicians and have a long history of being predictive of CVD.

I think the simple approach is justified, as many of the new and updated CVD event prediction models often are attempts to identify or incorporate new variables, which don't usually make much difference.[7,108,114] Most new variables add little to age, which is a risk factor in itself and is related to developing other risk factors, and the duration of exposure to them.[254] For example in CHA_2DS_2VASC, used for stroke prediction in atrial fibrillation, being over 75 gives the same number of points as having had a previous stroke. Similarly with DASHI age adds as many points as any other variable (four for the over 80s, the same as

a pneumonia diagnosis). The modifiable 'risk factors' for CVD events are well characterised, and it is likely that we are close to the limits of what can be achieved with routinely collected data, which doesn't include novel biomarkers or imaging. One of the things I learned from developing the DASHI score is that the more complex models did not perform better, in my view, not sufficiently better to be worth the effort of a clinician collecting the variables, or operating a computer program to calculate the risk, nor applying to the MRHA for permission to use a calculator as a 'medical device'.

The causal inference chapters, five and six, demonstrated problems partly due to the delayed recording of cardiovascular events. There is a distribution of delays in recording CVD where the recorded dates in the databases represent the latest possible date for the CVD events. This delay must also apply to the derivation and validation of DASHI. Some of the patients included in the datasets have already had CVD events before the index date, and some of the patients who had infection-related CVD events in the 28-day follow-up period will not have had these recorded until after the study period, if ever. This is likely to be similar between the two datasets given the validation results, and means the risk predicted by the model is not perfectly accurate but is likely to be proportional to the actual risk. The direction of bias is not clear because these inclusions have opposite effects. Including people with pre-existing CVD events in the derivation dataset (who are at higher risk of secondary events and likely to have diabetes, be older, smoke, and have heart failure) may have increased the estimated risk attributable to these characteristics. Including people in the dataset who had infection-related CVD events that were not recorded will have reduced the estimated risk attributable to these factors. I note that these problems also affect other routinely collected datasets use for deriving and validating cardiovascular risk scores.[108] Using longer follow-up times can ensure events affected by delayed recording are not completely missed. The problem of identifying participants who have not had events (i.e. for primary prevention studies)

remains. In time to event analyses, it means ineligible patients who have already had CVD events contributing their person time to the analysis up until those events are recorded.[255]. It may be that cohort studies such as Framingham, designed specifically for the purpose of collecting this information give more accurate risk estimates, but it is also possible that these are also subject to recording delays or inaccurate dates for CVD events.[129]

I think we can be confident that DASHI can give us some information about the risk of infection-related CVD events.

7.4.2 Aspirin results in context

The protective effects of aspirin on CVD events are well established.[8,19] Aspirin reduces the risk of MI by about 20% in long term use, whether for primary or secondary infection.[8] It also increases the risk of bleeding.[8] Bleeding limits aspirin's use in low prevalence settings.[19,85] These results are extremely robust, backed with meta-analyses of multiple high quality randomised controlled trials. The attempts to assess aspirin in respiratory infection are few, but all pointed to a reduction in events.[102–104] In this context, it is unlikely that aspirin increases the risk of CVD events during respiratory infections.

7.4.3 Statin results in context

Statins are also well established as preventative medications for CVD events when used long term.[10,11,19] They work for long-term primary and secondary prevention.[10,19] This evidence is as robust as that for the use of aspirin. The use of statins over very short timescales is less clear, and more difficult to measure due to fewer events. Statins work by preventing and stabilising atherosclerosis, and it is possible that this takes more than a few weeks of treatment to prevent MI or stroke.[9] A lack of effect is not sufficient explanation for my results, which showed harm. Inaccurate measurement of the effect seems the most likely explanation for the estimate of increased CVD events with statin use.

Overall, I do not believe the measurements of the effects of aspirin and statins from this thesis can be relied upon.

7.5 Strengths and limitations

7.5.1 Routinely collected medical record databases

The whole of this thesis used CPRD data linked to other databases. A strength of these data is the size – millions of patients were included. Large datasets help with precision of estimates, but large numbers don't make any difference to the structural problems that are the most important limitations.

In my view the most important problem with the data is the difficulty establishing the timing of the CVD events, and the difference between the delays in recording CVD events and prescriptions. This is not a problem that can easily be addressed using databases of medical records. Using a longer duration of follow-up reduces the likelihood that the CVD event is related to the respiratory infection, and if the duration is too long ends up measuring 'standard' primary prevention. As primary prevention is well established with clinical trial data, such an analysis would not add to the evidence base.

A longer follow-up, or another method for accounting for the pre-infection exposure to medications would not ameliorate the possibility of reverse causality – that the medication could be started because of a CVD event prior to the infection. The use of prescriptions to identify CVD events is a possible remedy. This approach was used in Dutch primary care data to identify CVD events found in a hospital database and improved the recording of events from 43% in primary care data to 94%.[256] They did not assess the timing of the CVD events. In this thesis I have included the data from hospital databases, so the issue is different, but conceivably the approach could be adapted to help identify timings of CVD events for studies with longer follow-up periods. When someone is prescribed a combination of antiplatelets, statins, ACE

inhibitors and beta blockers they have probably had a recent myocardial infarction, but this prescription still doesn't give us an exact date for the event. Estimating the timing based on these prescriptions would require further methodological work, and may be unverifiable without further datasets.

It is possible that using dates of hospital admissions could be used to estimates of the dates of CVD events. This would not be foolproof, as many patients (I am thinking of my own patients and relatives) sit at home having cardiovascular events and must be strongly persuaded to attend hospital. Another approach would be to change the outcome. Some outcomes, notably death, are less subject to incomplete recording, and have specific dates.[206]

7.5.2 Causal inference

Causal inference is challenging because it means measuring an estimate of something that didn't happen.[135,136] It always relies on unverifiable assumptions. The causal inference modelling seems not to have worked well. Using instrumental variable (IV) analyses could be a way to take this forward but may have similar problems arising from the data. IV analyses can avoid the problems of confounding, provided there is a valid instrument.[136,217] The CONSORT guidelines for the reporting of clinical trials emphasise those elements of trial conduct that attempt to ensure the assumptions of instrumental variable analysis are met, although they don't describe them in such terms.[257] Finding a valid instrument is a more of a challenge for observational data of prescriptions where there is no clear decision date or moment of choice between prescriptions (as in the case of aspirin where it is aspirin or no aspirin rather than a choice between medications), which can lead to biases.[240,241,258] It is difficult to find an instrument that is not associated with the outcomes except via the exposure.[136] One possibility for statins would be to use the intensity of the treatment as an instrument (dose equivalence) but this also introduces the difficulties of dealing with changes in treatment in response to cholesterol measurements, and clinical events.[259,260]

If there is a serious problem with identifying the timing of the events compared to the exposures, and I believe there is, then this could also bias an observational IV analysis of infection-related CVD events.

7.5.3 DASHI score

A limitation of the score is highlighted by the causal inference chapters – if we are not able to clearly identify people with events within 28 days the predicted risks may be mis-calibrated. The absolute risks might be underestimated because some of the events might have been delayed in their recording. Alternatively, delays in recording events prior to the infections may have caused secondary prevention patients to be wrongly included. These would be the highest risk patients, potentially increasing the predicted risks. An approach to this would be to revalidate the DASHI score in another dataset, if one could be found which is subject to less recording delay. However, in the clinical situation for which it is intended the available data is the same clinical record as was used to derive the models.

7.6 Implications

7.6.1 Clinical implications

When faced with a patient with respiratory infection there is insufficient evidence to change clinical practice, which is driven by interventions rather than prediction. Whilst DASHI can predict risk, which may influence a consultation in terms of discussion and safety netting, clinicians should therefore not offer statins or aspirin to patients because of an elevated DASHI score. However, CVD event prevention is a major priority, and clinicians should implement established both primary and secondary prevention measures. The offer of influenza vaccines to people aged 50 based on age alone, and of statins to almost any informed person who wants them, reduce the risk of CVD events, and I welcome these changes to the guidance.[10,64]

7.6.2 Implications for policy

DASHI tells us something about the features of people who are at higher risk of infection-related CVD events, and these conditions are already targets for CVD prevention efforts. It is not controversial to suggest that people who have diabetes, smoke, and have heart failure are at increased risk of CVD events, and that treating these risk factors is likely to prevent CVD events. Similarly reducing severe respiratory infection could be expected to reduce infection-related CVD events. The traditional way to do this is with vaccination.[64] These conditions are all amenable to public health measures and there have been clinical steps in this direction during the last five years, as detailed above. A recent opinion article in the BMJ argued that all prevention should be removed from primary care, and GPs should focus only on people with symptoms.[261] In the unlikely event of this palliative approach being widely adopted, more than just CVD event prevention will have to be re-organised, but there could be benefits to a more public health style approach, for example with the 'poly-pill' of statins and antihypertensives being offered to everyone over the age of fifty years, thus normalising rather than medicalising prevention.[262]

7.6.3 Research implications

7.6.3.1 DASHI

The implications of the DASHI score are positive for research – it allows the identification of higher risk cohorts for involvement in clinical trials and observational studies. Further studies validating DASHI in the post-Covid-19 world, and using only recent data would be welcome, as well as in different settings.

DASHI is a score that is context specific – the UK primary care setting is one where infections are not routinely subject to extensive testing, and CVD events are relatively uncommon. DASHI is likely to have different predictive properties in other countries or in hospitals. Inpatients are more likely to have chest radiography and blood tests if there is

suspicion of lower respiratory tract infections, they are more likely to have CVD events, and in a setting where ECGs and troponin tests are easily available smaller events are more likely to be detected. DASHI would therefore need to be validated in other settings before being applied there.[120] However, being simple means it is less likely to be overfitted than higher dimensional models.[263] Because the variables in DASHI are known to be causative for CVD events there is no obvious reason why it could not be refitted and revalidated for other settings. A calibration and validation for secondary CVD events could also be attempted, as the variables are likely to apply to secondary CVD events as well. It is possible that it would be underfitted, but a validation study could identify this. A bigger difficulty is the following clinical question – what to do for people who have already had CVD events and should already be on maximal preventative therapies. It is possible that different treatments for the respiratory infection could have an effect, for example antibiotics or antivirals, or intensification of antiplatelet or anticoagulation therapy could be evaluated.

Re-fitting and revalidating in different settings could address another problem – low prevalence. There are few people with higher DASHI risk estimates in primary care, and it might be that a higher prevalence secondary care setting would have a greater clinical utility. Ultimately though, the only use for prediction is to be able to make an informed change in clinical management. Finding something that can help reduce risk should be a priority. I have not managed to come up with new answers in this thesis.

7.6.3.2 Causal inference study designs

One solution for observational causal inference studies of vaccines, which I originally conceived as part of this project, was to use vaccine effectiveness as an instrument. The attractiveness of this idea was that one can compare between years, when the intervention was the same as far as the clinicians and patients were concerned, and indications for use are similar. Vaccine effectiveness is an instrument because it is only associated with the outcome

via the exposure – the risk of infections.[217] It is only calculated at the end of the influenza season so knowledge of effectiveness can't have affected what went on in the prior six months (although here have been years where it was obvious the vaccine was not very effective, although as influenza vaccine effectiveness is typically about 40%, and CVD event risk from influenza is more or less unacknowledged clinically, this wouldn't be much of a change).[264] The duration of the effectiveness of influenza vaccines (and Covid-19 vaccines) starts about 14 days after vaccination and lasts until six months later.[265] This longer 'exposure' period would allow time for the CVD events to be recorded, and the negative control period, the first 14 days, would allow assessment of the baseline risk of events. Delays in recording CVD events, particularly those occurring later in the follow-up, could be assessed by looking for apparent effects after the six months of exposure, as they could not be attributed to vaccine effects.

One approach that could be effective for assessing the effects of antiplatelets on infection-related CVD events is Mendelian randomisation.[266] There are established genetic proxies for antiplatelet (and statin) activity, which could be applied to infection-related CVD events.[267,268] This avoids the problems of confounding by indication and the timing of the exposure to the proxy for medication, which is permanent. This long duration of exposure could still be a proxy for short term antiplatelet exposure (as the effects are not cumulative), but would not help with statins which have long-term effects.

A trial would be one way to settle the causal questions about aspirin and statins in respiratory infections. However, such a trial would have to be innovative and/or enormous because of the small absolute risks involved. I imagined trials to help identify the challenges. The results from this thesis help to define the population – assuming a primary care population with a cut off at three DASHI points - the minimum baseline risk would be 0.3% over 28 days, and the population average would be about four DASHI points (0.6% risk). If we optimistically

assume the effects of aspirin 75mg are no worse than in other primary prevention scenarios, we can expect a 20% reduction in CVD events, so 0.48% having a CVD event in the intervention arm. These are small numbers, and to achieve 80% power with an alpha of 0.05 would require 58,548 participants in each arm. This is unlikely to be feasible. If aspirin were only to reduce events by 15%, more than 110,000 would be needed in each arm. Another approach would be to randomise high-risk individuals to a strategy of taking aspirin whenever they got ill with respiratory infections. The average number of COPD exacerbations per year is three, and patients with COPD are likely to have a DASHI score is likely to be at least three.[43] Using within-person randomisation would be most efficient, but organising this so the randomisation was achieved and adhered to at each exacerbation would be challenging. Treating each individual with one strategy over many exacerbations, in effect as their own one-person cluster, also reduces the number of participants. Following them up for an average of three years, an average of nine exacerbations each, would require 23,527 participants in each arm. This is still a very large trial, which highlights why most clinical trials tend to take place in higher risk populations, typically in hospitals, or over many years of follow-up. Restricting the population to a DASHI of eight or more (e.g. over 80s with pneumonia) gives a minimum risk of 7.3%, which would need 4,520 in each arm (with a 20% effectiveness of the intervention, versus 8,243 for 15% reduction in risk and 19,011 for a 10% reduction in risk). Ultimately these would all have to very big trials, for very small risks. It is possible that prevention in this context might not be worth pursuing, which is itself a question that might be amenable to health economics research.

7.6.3.3 Caution regarding timing of events

Possibly the most important finding in this thesis for future research is that the results from studies using routinely collected databases cannot be trusted if they are sensitive to the timing of events. This bias was exaggerated in this thesis because of the nature of the events, which

are subject to delayed recording, the contrast with more immediate recording of prescription data, and by the short follow-up period. This has implications for both observational studies and trials that use primary care records to ascertain events.

As an example, the self-controlled case series methods that first identified the increased risk after respiratory infections may also be affected by these delays. Smeeth *et al* excluded events that they thought might have been recorded retrospectively, 'including on the date of new or well-person visits, or discharge from hospital', but did not employ hospital data or prescribing to identify timing of events.[2] They used a 91 day follow-up period and showed an increase in the risk of MI and stroke following respiratory and urinary infections, but not vaccinations. This risk was estimated as being highest in the first few days (IRR 4.95, 95% CI 4.43 to 5.53 for MI in the first 3 days after a respiratory infection) and ameliorated over the follow-up (IRR 1.40, 95% CI 1.33 to 1.48 for MI for 29-91 days post respiratory infection). It is possible that delays in recording events could be responsible for this long tail of increased risk.

Similarly, a CPRD study looking at the effects of aspirin on infection-related CVD events reported an increased risk for six months post pneumonia in both intervention and control groups.[104] This could be a measurement of a long-term effect from the inflammatory insult of the infection, or it could be another case of delayed recording extending the risk period. The study used propensity modelling to match patients who were using aspirin with those who were not. The aspirin users had higher levels of important confounders such as prior CVD events, and these were well balanced by the propensity model matching.[104] It is also possible that aspirin users were receiving a higher standard of care, which probably includes more timely recording of events. Rapid recording of events in the aspirin groups could have led to a higher incidence of CVD events in the control period (one year prior to the pneumonia diagnosis to the six months before the pneumonia). This pattern was present (Aspirin users

2% stroke and 0.8% myocardial infarction risk in the control period compared with 1.8% and 0.4% for the non-aspirin users).[104] It is possible that a pneumonia diagnosis is a trigger for recording of events that had happened earlier, which, if the aspirin users were more likely to have already had events recorded, would bias the estimates towards benefit.

7.7 Conclusion

Clinically, long-term primary prevention of CVD events is an immediate priority, with a strong evidence base. Respiratory infection-related CVD events are a low-prevalence, high-harm clinical problem. The DASHI score can be used to inform future research and itself be the subject of further validation studies. The short-term effect of acute interventions is something that is challenging to investigate fully in primary care records, not least because of the difficulty of establishing the timing of outcome events. In future researchers should carefully consider these potential limitations.

Clinicians can now predict respiratory infection-related CVD event risk, but there is insufficient evidence to recommend acute interventions to mitigate this risk.

8 References

1. Warren-Gash, C., Smeeth, L. & Hayward, A. C. Influenza as a trigger for acute myocardial infarction or death from cardiovascular disease: a systematic review. *Lancet Infect Dis* **9**, 601–610 (2009).

2. Smeeth, L. *et al.* Risk of Myocardial Infarction and Stroke after Acute Infection or Vaccination. *New England Journal of Medicine* **351**, 2611–2618 (2004).

3. Health Intelligence Team, B. *BHF UK CVD Factsheet*. (2024).

4. Monto, A. S. Epidemiology of viral respiratory infections. *Am J Med* **112 Suppl**, 4S-12S (2002).

5. Nguyen, J. L. *et al.* Seasonal influenza infections and cardiovascular disease mortality. *JAMA Cardiol* **1**, 274–281 (2016).

6. Pitman, R. J. *et al.* Assessing the burden of influenza and other respiratory infections in England and Wales. *Journal of Infection* **54**, 530–538 (2007).

7. Damen, J. A. A. G. *et al.* Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ (Online)* **353**, (2016).

8. Collins, R. *et al.* Aspirin in the primary and secondary prevention of vascular disease: collaborative meta-analysis of individual participant data from randomised trials. *The Lancet* **373**, 1849–1860 (2009).

9. Joint Formulary Committee. *British National Formulary*. (BMJ Group and Pharmaceutical Press, London, 2019).

10. National Institute for Clinical Excellence. Cardiovascular disease: risk assessment and reduction, including lipid modification. *NICE Guidance CG238* (2023).

11. Collins, R. *et al.* Interpretation of the evidence for the efficacy and safety of statin therapy. *The Lancet* **388**, 2532–2561 (2016).

12. Hippisley-Cox, J., Coupland, C. & Brindle, P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: Prospective cohort study. *BMJ (Online)* **357**, 1–21 (2017).

13. Robbins, S., Cotran, R., Kumar, V. & Collins, T. *Pathologic Basis of Disease*. (W.B. Saunders Company, 1999).

14. Visseren, F. L. J. *et al.* 2021 ESC Guidelines on cardiovascular disease prevention in clinical practiceDeveloped by the Task Force for cardiovascular disease prevention in clinical practice with representatives of the European Society of Cardiology and 12 medical societies With the special contribution of the European Association of Preventive Cardiology (EAPC). *Eur Heart J* **42**, 3227–3337 (2021).

15. Barnes, M. *et al.* Acute myocardial infarction and influenza: A meta-analysis of case-control studies. *Heart* **101**, 1738–1747 (2015).

16. Katsoularis, I., Fonseca-Rodríguez, O., Farrington, P., Lindmark, K. & Fors Connolly, A. M. Risk of acute myocardial infarction and ischaemic stroke following COVID-19 in Sweden: a self-controlled case series and matched cohort study. *The Lancet* **398**, 599–607 (2021).

17. Kwong, J. C. *et al.* Acute myocardial infarction after laboratory-confirmed influenza infection. *New England Journal of Medicine* **378**, 345–353 (2018).

18. Warren-Gash, C. *et al.* Influenza infection and risk of acute myocardial infarction in england and wales: A CALIBER self-controlled case series study. *Journal of Infectious Diseases* **206**, 1652–1659 (2012).

19. Visseren, F. *et al.* 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice. *Eur Heart J* **42**, 3227–3337 (2021).

20. Rodríguez, A. *et al.* Recommendations of the Infectious Diseases Work Group (GTEI) of the Spanish Society of Intensive and Critical Care Medicine and Coronary Units (SEMICYUC) and the Infections in Critically Ill Patients Study Group (GEIPC) of the Spanish Society of Infectiou. *Medicina Intensiva (English Edition)* **36**, 103–137 (2012).

21. Razonable, R. R. *et al.* A Collaborative Multidisciplinary Approach to the Management of Coronavirus Disease 2019 in the Hospital Setting. *Mayo Clin Proc* **95**, 1467–1481 (2020).

22. Muscente, F. & De Caterina, R. Causal relationship between influenza infection and risk of acute myocardial infarction: pathophysiological hypothesis and clinical implications. *European Heart Journal Supplements* **22**, E68–E72 (2020).

23. Cantan, B., Luyt, C. E. & Martin-Loeches, I. Influenza Infections and Emergent Viral Infections in Intensive Care Unit. *Semin Respir Crit Care Med* **40**, 488–497 (2019).

24. Caldeira, D. *et al.* The association of influenza infection and vaccine with myocardial infarction: systematic review and meta-analysis of self-controlled case series. *Expert Rev Vaccines* **18**, 1211–1217 (2019).

25. Blackburn, R., Zhao, H., Pebody, R., Hayward, A. & Warren-Gash, C. Laboratory-confirmed respiratory infections as predictors of hospital admission for myocardial infarction and stroke: Time-series analysis of English data for 2004-2015. *Clinical Infectious Diseases* **67**, 8–17 (2018).

26. Pitman, R. J. *et al.* Assessing the burden of influenza and other respiratory infections in England and Wales. *Journal of Infection* **54**, 530–538 (2007).

27. Blackburn, R., Zhao, H., Pebody, R., Hayward, A. & Warren-Gash, C. Laboratory-confirmed respiratory infections as predictors of hospital admission for myocardial infarction and stroke: Time-series analysis of English data for 2004-2015. *Clinical Infectious Diseases* **67**, 8–17 (2018).

28. Davidson, J. A. *et al.* Risk of acute respiratory infection and acute cardiovascular events following acute respiratory infection among adults with increased cardiovascular risk in England between 2008 and 2018 : a retrospective , population-based cohort study. *Lancet Digit Health* **3**, e773–e783 (2021).

29. Clar, C., Oseni, Z., Flowers, N., Keshtkar-Jahromi, M. & Rees, K. Influenza vaccines for preventing cardiovascular disease. *Cochrane Database of Systematic Reviews* **133**, 384 (2015).

30. Jaiswal, V. *et al.* Effect of Pneumococcal Vaccine on Mortality and Cardiovascular Outcomes: A Systematic Review and Meta-Analysis. *J Clin Med* **11**, (2022).

31. Udell, J. A. *et al.* Association between influenza vaccination and cardiovascular outcomes in high-risk patients: a meta-analysis. *JAMA* **310**, 1711–1720 (2013).

32. Toth, P. P. Subclinical atherosclerosis: What it is, what it means and what we can do about it. *International Journal of Clinical Practice* vol. 62 1246–1254 Preprint at https://doi.org/10.1111/j.1742-1241.2008.01804.x (2008).

33. Robbins, S., Cotran, R., Kumar, V. & Collins, T. *Pathologic Basis of Disease*. (W.B. Saunders Company, 1999).

34. Woolf, N. *Cell, Tissue and Disease*. (W. B. Saunders, London, 2000).

35. Thygesen, K. *et al.* Fourth Universal Definition of Myocardial Infarction (2018). *J Am Coll Cardiol* **72**, 2231–2264 (2018).

36. Thygesen, K. *et al.* Fourth Universal Definition of Myocardial Infarction (2018). *J Am Coll Cardiol* **72**, 2231–2264 (2018).

37. Lanza, G. & Crea, F. Vasospastic Angina. *E-journal of Cardiology Practice* **2**, (2003).

38. Jenkins, K. *et al.* Vasospastic angina: a review on diagnostic approach and management. *Therapeutic Advances in Cardiovascular Disease* vol. 18 Preprint at https://doi.org/10.1177/17539447241230400 (2024).

39. Lee, A. E., Mizrahi, I., Nagamine, T., Kawamoto, K. & Devendra, G. Complex Clinical Cases SEVERE CORONARY VASOSPASM AND CARDIAC ARREST IN COVID-19 INFECTION. in vol. 81 2686 (Elsevier, 2023).

40. Geng, Y. J. & Libby, P. Progression of atheroma: A struggle between death and procreation. *Arterioscler Thromb Vasc Biol* **22**, 1370–1380 (2002).

41. Geng, Y. J. & Libby, P. Progression of atheroma: A struggle between death and procreation. *Arterioscler Thromb Vasc Biol* **22**, 1370–1380 (2002).

42. Corrales-Medina, V. F. *et al.* Acute bacterial pneumonia is associated with the occurrence of acute coronary syndromes. *Medicine* **88**, 154–159 (2009).

43. Donaldson, G. C. & Wedzicha, J. A. COPD exacerbations · 1: Epidemiology. *Thorax* vol. 61 164–168 Preprint at https://doi.org/10.1136/thx.2005.041806 (2006).

44. Lin, S. H. *et al.* Increased risk of community-acquired pneumonia in COPD patients with comorbid cardiovascular disease. *International Journal of COPD* **11**, 3051–3058 (2016).

45. Shah, P., Bajaj, S., Virk, H., Bikkina, M. & Shamoon, F. Rapid Progression of Coronary Atherosclerosis: A Review. *Thrombosis* 1–6 (2015) doi:http://dx.doi.org/10.1155/2015/634983 Review.

46. Shah, P., Bajaj, S., Virk, H., Bikkina, M. & Shamoon, F. Rapid Progression of Coronary Atherosclerosis: A Review. *Thrombosis* 1–6 (2015) doi:http://dx.doi.org/10.1155/2015/634983 Review.

47. Joint Formulary Comittee. *British National Formulary*. (BMJ Group and Pharmaceutical Press, London, 2019).

48. Madjid, M., Aboshady, I., Awan, I., Litovsky, S. & Casscells, S. W. Influenza and Cardiovascular Disease: Is There a Causal Relationship? *Tex Heart Inst J* **31**, 4–13 (2004).

49. Binder, C. J. *et al.* Pneumococcal vaccination decreases atherosclerotic lesion formation: Molecular mimicry between Streptococcus pneumoniae and oxidized LDL. *Nat Med* **9**, 736–743 (2003).

50. Naghavi, M. *et al.* Influenza infection exerts prominent inflammatory and thrombotic effects on the atherosclerotic plaques of apolipoprotein E-deficient mice. *Circulation* **107**, 762–768 (2003).

51. Wang, F. *et al.* Macrophage foam cell-targeting immunization attenuates atherosclerosis. *Front Immunol* **10**, (2019).

52. Xia, M. *et al.* Modulation of recombinant Antigenic constructs containing multi-epitopes towards effective reduction of atherosclerotic lesion in B6;129S-Ldlrtm1HerApobtm2Sgy/J mice. *PLoS One* **10**, (2015).

53. *Robbins Basic Pathology*. (Elsevier, 2012).

54. Bagot, C. N. & Arya, R. Virchow and his triad: A question of attribution. *Br J Haematol* **143**, 180–190 (2008).

55. Bagot, C. N. & Arya, R. Virchow and his triad: A question of attribution. *Br J Haematol* **143**, 180–190 (2008).

56. Gomez-Casado, C. *et al.* Understanding platelets in infectious and allergic lung diseases. *International Journal of Molecular Sciences* vol. 20 Preprint at https://doi.org/10.3390/ijms20071730 (2019).

57. Kreutz, R. P., Bliden, K. P., Tantry, U. S. & Gurbel, P. A. Viral respiratory tract infections increase platelet reactivity and activation: an explanation for the higher rates of myocardial infarction and stroke during viral illness. *Journal of Thrombosis and Haemostasis* **3**, 2108–2109 (2005).

58. Yokomichi, H. *et al.* Incidence of hospitalisation for severe complications of influenza virus infection in Japanese patients between 2012 and 2016 : a cross-sectional study using routinely collected administrative data. 1–11 (2019) doi:10.1136/bmjopen-2018-024687.

59. Patel, J. Tachycardia-Induced Heart Failure. *Perm J* **11**, 50–52 (2007).

60. Mortensen, E., Metersky, M., Atuegwu, N. & Anzueto, A. New onset atrial fibrillation in patients hospitalised with pneumonia. *European Respiratory Journal* **54**, (2019).

61. Lip, G. Y. H. *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* **137**, 263–272 (2010).

62. NICE. Influenza - seasonal - NICE CKS. *NICE Clinical Knowledge Summaries* http://cks.nice.org.uk/influenza-seasonal#!diagnosissub (2014).

63. NICE. Respiratory tract infections (self-limiting): prescribing antibiotics | Guidance and guidelines | NICE. *NICE Guidelines* 20 (2008).

64. Ramsay, M. (UK H. S. A. *The Green Book*. (2023). doi:10.1007/BF01975422.

65. Public Health England. *Influenza: The Green Book, Chapter 19*. (2019).

66. Smeeth, L., Douglas, I., Hall, A. J., Hubbard, R. & Evans, S. Effect of statins on a wide range of health outcomes: A cohort study validated by comparison with randomized trials. *Br J Clin Pharmacol* **67**, 99–109 (2009).

67. Morbey, R. A. *et al.* Estimating the burden on general practitioner services in England from increases in respiratory disease associated with seasonal respiratory pathogen activity. *Epidemiol Infect* **146**, 1389–1396 (2018).

68. *Suspected Acute Respiratory Infection in over 16s: Assessment at First Presentation and Initial Management NICE Guideline*. www.nice.org.uk/guidance/ng237 (2023).

69. Bou-Antoun, S. *et al.* Age-related decline in antibiotic prescribing for uncomplicated respiratory tract infections in primary care in England following the introduction of a national financial incentive (the Quality Premium) for health commissioners to reduce use of antibiotics in the community: An interrupted time series analysis. *Journal of Antimicrobial Chemotherapy* **73**, 2883–2892 (2018).

70. Phe. Surveillance of influenza and other respiratory viruses in the United Kingdom: winter 2014 to 2015. 29 (2015).

71. Ebell MH, Afonso A. A Systematic Review of Clinical Decision Rules for the Diagnosis of Influenza. *Ann Fam Med* **9**, 69–77 (2011).

72. Surveillance of influenza and other respiratory pathogens in the UK. (2012).

73. Nicholson, B. D. *et al.* Rapid community point-of-care testing for COVID-19 (RAPTOR-C19): protocol for a platform diagnostic study. *Diagn Progn Res* **5**, (2021).

74. Del Mar, C. Antibiotics for acute respiratory tract infections in primary care. *The BMJ* **354**, i3482 (2016).

75. Yokomichi, H. *et al.* Antibiotic prescription for outpatients with influenza and subsequent hospitalisation: A cohort study using insurance data. *Influenza Other Respir Viruses* **17**, (2023).

76. Brauer, R. *et al.* Prevalence of antibiotic use : a comparison across various European health care data sources. **25**, 11–20 (2016).

77. Tonkin-Crine, S., Yardley, L. & Little, P. Antibiotic prescribing for acute respiratory tract infections in primary care: A systematic review and meta-ethnography. *Journal of Antimicrobial Chemotherapy* **66**, 2215–2223 (2011).

78. Lee, J. J. *et al.* The Clinical Utility of Point-of-Care Tests for Influenza in Ambulatory Care : A Systematic Review and Meta-analysis. *Clinical Infectious Diseases* **69**, 24–33 (2018).

79. *Deaths Registered in England and Wales: 2022*. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/deathregistrationsummarystatisticsenglandandwales/2022 (2023).

80. George, J. *et al.* How does cardiovascular disease first present in women and men? *Circulation* **132**, 1320–1328 (2015).

81. Powers, W. J. *et al. Guidelines for the Early Management of Patients with Acute Ischemic Stroke: 2019 Update to the 2018 Guidelines for the Early Management of*

Acute Ischemic Stroke a Guideline for Healthcare Professionals from the American Heart Association/American Stroke A. *Stroke* vol. 50 (2019).

82. NICE. *Stroke and Transient Ischaemic Attack in over 16s: Diagnosis and Initial Management Management*. *Nice Guideline* http://www.nice.org.uk/guidance/CG68 (2019).

83. NICE. *Recent-Onset Chest Pain of Suspected Cardiac Origin: Assessment and Diagnosis Clinical Guideline*. www.nice.org.uk/guidance/cg95 (2016).

84. Davidson, J. A., Banerjee, A., Muzambi, R., Smeeth, L. & Warren-Gash, C. Validity of acute cardiovascular outcome diagnoses in European electronic health records: A systematic review protocol. *BMJ Open* **9**, 1095–1111 (2019).

85. NICE. NICE guideline - Atrial fibrillation: management. *NICE GUidelines* 38 (2014).

86. NICE. Diabetes (type 1 and type 2) in children and young people: diagnosis and management. *National Institute for Health and Care Excellence* 1–92 (2015).

87. NICE. Myocardial infarction: cardiac rehabilitation and prevention of further cardiovascular disease. *Clinical guideline [CG172]* Key priorities for implementation (2013).

88. Damen, J. A. A. G. *et al.* Supplemental Material: Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ : British Medical Journal* **353**, s1–s32 (2016).

89. Piepoli, M. F. *et al.* 2016 European Guidelines on cardiovascular disease prevention in clinical practice. *Eur Heart J* **37**, 2315–2381 (2016).

90. NICE guideline development group. *Cardiovascular Disease: Risk Assessment and Reduction Including Lipid Modification*. (2014).

91. Bakris, G., Ali, W. & Parati, G. ACC/AHA Versus ESC/ESH on Hypertension Guidelines: JACC Guideline Comparison. *J Am Coll Cardiol* **73**, 3018–3026 (2019).

92. Sheppard, J. P. *et al.* Association Between Blood Pressure Control and Coronavirus Disease 2019 Outcomes in 45 418 Symptomatic Patients With Hypertension: An Observational Cohort Study. *Hypertension* **77**, 846–855 (2021).

93. Hypertension in adults: Diagnosis and management. *NICE Clinical Guideline* Preprint at https://www.nice.org.uk/guidance/ng136 (2019).

94. Johnston, S. C. *et al.* Clopidogrel and Aspirin in Acute Ischemic Stroke and High-Risk TIA. *New England Journal of Medicine* **379**, 215–225 (2018).

95. NICE. Myocardial infarction with ST-segment elevation: acute management. *NICE Clinical Guideline* 1–27 (2013).

96. Grahame-Smith, D. G. & Aronson, J. K. *The Oxford Textbook of Clinical Pharmacology and Drug Therapy*. (Oxford University Press, Oxford, 2002).

97. NICE guideline development group. *Myocardial Infarction : Cardiac Rehabilitation and Prevention of Further Cardiovascular Disease*. (2013).

98. Prasad, K. *et al.* Dual antiplatelet therapy with aspirin and clopidogrel for acute high risk transient ischaemic attack and minor ischaemic stroke: A clinical practice guideline. *BMJ (Online)* **363**, (2018).

99. Hao, Q. *et al.* Clopidogrel plus aspirin versus aspirin alone for acute minor ischaemic stroke or high risk transient ischaemic attack: Systematic review and meta-analysis. *BMJ (Online)* **363**, (2018).

100. Duo, H. *et al.* Effect of antiplatelet therapy after COVID-19 diagnosis: A systematic review with metaanalysis and trial sequential analysis. *PLoS One* **19**, 1–18 (2024).

101. Horby, P. W. (RECOVERY C. group). Supplementary appendix: Aspirin in patients admitted to hospital with COVID-19 (RECOVERY): a randomised , controlled , open-label , platform trial. *Lancet* (2022).

102. Horby, P. W. (RECOVERY C. group). Aspirin in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial. *The Lancet* **399**, 143–151 (2022).

103. Oz, F. *et al.* Does aspirin use prevent acute coronary syndrome in patients with pneumonia: Multicenter prospective randomized trial. *Coron Artery Dis* **24**, 231–237 (2013).

104. Hamilton, F., Arnold, D., Henley, W. & Payne, R. A. Aspirin reduces cardiovascular events in patients with pneumonia: a prior event rate ratio analysis in a large primary care database. *European Respiratory Journal* 2002795 (2020) doi:10.1183/13993003.02795-2020.

105. Baigent, C. *et al.* Efficacy and safety of more intensive lowering of LDL cholesterol: A meta-analysis of data from 170 000 participants in 26 randomised trials. *The Lancet* **376**, 1670–1681 (2010).

106. Bulbulia, R. *et al.* Effects on 11-year mortality and morbidity of lowering LDL cholesterol with simvastatin for about 5 years in 20,536 high-risk individuals: a randomised controlled trial. *Lancet* **378**, 2013–20 (2011).

107. Barter, P. *et al.* HDL Cholesterol, Very Low Levels of LDL Cholesterol, and Cardiovascular Events. *N Engl J Med* 1301–1310 (2007).

108. Hippisley-Cox, J., Coupland, C. & Brindle, P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: Prospective cohort study. *BMJ (Online)* **357**, 1–21 (2017).

109. Cai, T. *et al.* Associations between statins and adverse events in primary prevention of cardiovascular disease: Systematic review with pairwise, network, and dose-response meta-analyses. *The BMJ* **374**, (2021).

110. Alsanosi, S. M. & Alshanberi, A. M. Do statins protect against Respiratory Tract Infection: A systematic review and meta-analysis. *Med Sci* **27**, 1–10 (2023).

111. Van Den Hoek, H. L., Bos, W. J. W., De Boer, A. & Van De Garde, E. M. W. Statins and prevention of infections: Systematic review and meta-analysis of data from large randomised placebo controlled trials. *BMJ (Online)* vol. 343 1242 Preprint at https://doi.org/10.1136/bmj.d7281 (2011).

112. Nagendran, M. *et al.* Statin therapy for acute respiratory distress syndrome: an individual patient data meta-analysis of randomised clinical trials. *Intensive Care Medicine* vol. 43 663–671 Preprint at https://doi.org/10.1007/s00134-016-4649-0 (2017).

113. Franco-Peláez, J. A. *et al.* Statin use is associated with reduced mortality after respiratory viral infection. *ERJ Open Res* **7**, 1–8 (2021).

114. Pepe, M. S., Janes, H., Longton, G., Leisenring, W. & Newcomb, P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol* **159**, 882–890 (2004).

115. Office of National Statistics.

116. Herrett, E. *et al.* Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* **44**, 827–836 (2015).

117. CPRD. *CPRD Aurum Frequently Asked Questions (FAQs)*. (2021).

118. Steyerberg, E. W. *Clinical Prediction Models*. (Cham, 2019). doi:10.1007/978-3-030-16399-0_20.

119. Bonnett, L. J., Snell, K. I. E., Collins, G. S. & Riley, R. D. Guide to presenting clinical prediction models for use in clinical settings. *BMJ (Online)* **365**, 1–8 (2019).

120. Steyerberg, E. W. & Harrell, F. E. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* **69**, 245–247 (2016).

121. Brookhart, M. A. *et al.* Variable Selection for Propensity Score Models. *Am J Epidemiol* **163**, 1149–1156 (2006).

122. Steyerberg, E. W. *et al.* Poor performance of clinical prediction models: the harm of commonly applied methods. *J Clin Epidemiol* **98**, 133–143 (2018).

123. Sauerbrei, W. & Schumacher, M. A bootstrap resampling procedure for model building: Application to the cox regression model. *Stat Med* **11**, 2093–2109 (1992).

124. Box, G. E. P. Science and Statistics. *J Am Stat Assoc* **71**, 791–799 (1976).

125. Riley, R. D. *et al.* Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med* **38**, 1276–1296 (2019).

126. Stevens, S. L. *et al.* Blood pressure variability and cardiovascular disease: Systematic review and meta-analysis. *BMJ (Online)* **354**, 14–16 (2016).

127. Stevens, S. L., McManus, R. J. & Stevens, R. J. The utility of long-term blood pressure variability for cardiovascular risk prediction in primary care. *J Hypertens* **37**, 522–529 (2019).

128. Mahmood, S. S., Levy, D., Vasan, R. S. & Wang, T. J. The Framingham Heart Study and the epidemiology of cardiovascular disease: A historical perspective. *Lancet* **383**, 999–1008 (2013).

129. D'Agostino, R. B. *et al.* General cardiovascular risk profile for use in primary care: The Framingham heart study. *Circulation* **117**, 743–753 (2008).

130. Anderson, K. M., Odell, P. M., Wilson P W F & Kannel, W. B. Cardiovascular disease risk profiles. *Am Heart J* 293–298 (1990).

131. Koshiaris, C. *et al.* Simple and adaptable R implementation of WHO/ISH cardiovascular risk charts for all epidemiological subregions of the world. *F1000Res* **5**, 2522 (2016).

132. Piepoli, M. F. *et al.* 2016 European Guidelines on cardiovascular disease prevention in clinical practice. *Eur Heart J* **37**, 2315–2381 (2016).

133. Wilkinson, J. (MHRA). MHRA alert - QRISK2 Calculator in SystmOne. 1 Preprint at https://www.gov.uk/government/news/mhra-information-on-tpp-and-qrisk2 (2016).

134. Vickers, A. J., Van Calster, B. & Steyerberg, E. W. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ (Online)* **352**, (2016).

135. Rosenbaum, P. R. & Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55 (1983).

136. Angrist, J. D., Imbens, G. W. & Rubin, D. B. Identification of Causal Effects Using Instrumental Variables. **91**, 444–455 (2010).

137. Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y. & Van Der Laan, M. J. Diagnosing and responding to violations in the positivity assumption. *Stat Methods Med Res* **21**, 31–54 (2012).

138. Imai, K. & Ratkovic, M. Covariate balancing propensity score. *J R Stat Soc Series B Stat Methodol* **76**, 243–263 (2014).

139. Austin, P. C. & Stuart, E. A. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med* **34**, 3661–3679 (2015).

140. Seeger, J. D., Kurth, T. & Walker, A. M. Use of Propensity Score Technique to Account for Exposure-Related Covariates. *Med Care* **45**, S143–S148 (2007).

141. Rubin, D. B. & Waterman, R. P. Estimating the Causal Effects of Marketing Interventions Using Propensity Score Methodology. *Statistical Science* **21**, 206–222 (2006).

142. Funk, M. J. *et al.* Doubly robust estimation of causal effects. *Am J Epidemiol* **173**, 761–767 (2011).

143. Li, F. & Thomas, L. E. Addressing Extreme Propensity Scores via the Overlap Weights. *Am J Epidemiol* **188**, 250–257 (2018).

144. STATA TREATMENT-EFFECTS REFERENCE MANUAL: potential outcomes/counterfactual outcomes release 13. Preprint at (2013).

145. Bittmann, F. Applied Bootstrap Analysis with Imputed Data in Stata. *www.preprints.org* (2024) doi:10.20944/preprints202401.0813.v1.

146. Seaman, S. & White, I. Inverse probability weighting with missing predictors of treatment assignment or missingness. *Commun Stat Theory Methods* **43**, 3499–3515 (2014).

147. *STATA MULTIPLE-IMPUTATION REFERENCE MANUAL RELEASE 18*. www.stata.com.

148. van Buuren, S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res* **16**, 219–242 (2007).

149. Sperrin, M. & Martin, G. P. Multiple imputation with missing indicators as proxies for unmeasured variables: Simulation study. *BMC Med Res Methodol* **20**, 1–11 (2020).

150. Morris, T. P., White, I. R., Carpenter, J. R., Stanworth, S. J. & Royston, P. Combining fractional polynomial model building with multiple imputation. *Stat Med* **34**, 3298–3317 (2015).

151. Marshall, A., Altman, D. G., Holder, R. L. & Royston, P. Combining estimates of interest in prognostic modelling studies after multiple imputation: Current practice and guidelines. *BMC Med Res Methodol* **9**, 1–8 (2009).

152. Royston, P. & White, I. R. Journal of Statistical Software Multiple Imputation by Chained Equations (MICE): Implementation in Stata. *J Stat Softw* **45**, 1–20 (2011).

153. Conroy, R. M. *et al.* Estimation of ten-year risk of fatal cardiovascular disease in Europe: The SCORE project. *Eur Heart J* **24**, 987–1003 (2003).

154. Sterne, J. A. C. *et al.* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* **338**, b2393 (2009).

155. Lee, J. J., Koshiaris, C., Hobbs, F. D. R. & Sheppard, J. P. Beyond COVID-19: Respiratory infection and cardiovascular events. *British Journal of General Practice* vol. 71 342–343 Preprint at https://doi.org/10.3399/bjgp21X716477 (2021).

156. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *Eur Urol* **67**, 1142–1151 (2015).

157. Wolf, A. *et al.* Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int J Epidemiol* **48**, 1740-1740G (2019).

158. GOV.UK. The English Indices of Deprivation 2019- Statistical Release. *Ministry of Housing, Communities and Local Government* **2019**, 1–12 (2019).

159. Lee, J. J. *et al.* Risk factors for influenza-related complications in children during the 2009 / 10 pandemic : a UK primary care cohort study using linked routinely collected data. *Epidemiol Infect* **146**, 817–823 (2018).

160. Sharma, A. *et al.* Impact of Regulatory Guidance on Evaluating Cardiovascular Risk of New Glucose-Lowering Therapies to Treat Type 2 Diabetes Mellitus: Lessons Learned and Future Directions. *Circulation* **141**, 843–862 (2020).

161. McDonagh, T. A. *et al.* 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. *Eur Heart J* **42**, 3599–3726 (2021).

162. Millett, E. R. C., De Stavola, B. L., Quint, J. K., Smeeth, L. & Thomas, S. L. Risk factors for hospital admission in the 28 days following a community-acquired pneumonia diagnosis in older adults, and their contribution to increasing hospitalisation rates over time: A cohort study. *BMJ Open* **5**, (2015).

163. Mertz, D. *et al.* Populations at risk for severe or complicated influenza illness: systematic review and meta-analysis. *Bmj* **347**, f5061–f5061 (2013).

164. Lange, P., Vestbo, J. & Nyboe, J. Risk factors for death and hospitalization from pneumonia. A prospective study of a general population. *European Respiratory Journal* **8**, 1694–1698 (1995).

165. Nguyen-Van-Tam, J. S. *et al.* Risk factors for hospitalisation and poor outcome with pandemic A/H1N1 influenza: United Kingdom first wave (May-September 2009). *Thorax* **65**, 645–651 (2010).

166. Ope, M. O. *et al.* Risk factors for hospitalized seasonal influenza in rural western Kenya. *PLoS One* **6**, (2011).

167. Juhn, Y. J. Risks for infection in patients with asthma (or other atopic conditions): Is asthma more than a chronic airway disease? *Journal of Allergy and Clinical Immunology* **134**, (2014).

168. Hayward, A. C. *et al.* Comparative community burden and severity of seasonal and pandemic influenza: Results of the Flu Watch cohort study. *Lancet Respir Med* **2**, 445–454 (2014).

169. Johnston, S. C. *et al.* Validation and refinement of scores to predict very early stroke risk after transient ischaemic attack. *Lancet* **369**, 283–292 (2007).

170. Verbakel, J. Y. *et al.* Impact of point-of-care C reactive protein in ambulatory care : a systematic review and meta-analysis. *BMJ Open* **9**, (2019).

171. Steyerberg, E. W. *et al.* Internal validation of predictive models. *J Clin Epidemiol* **54**, 774–781 (2001).

172. Peterson, W. W., Birdsall, T. G. & Fox, W. C. The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory* **4**, 171–212 (1954).

173. Caetano, S. J., Sonpavde, G. & Pond, G. R. C-statistic: A brief explanation of its construction, interpretation and limitations. *European Journal of Cancer* vol. 90 130–132 Preprint at https://doi.org/10.1016/j.ejca.2017.10.027 (2018).

174. Riley, R. D. *et al.* Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *BMJ* (2024) doi:10.1136/bmj-2023-074820.

175. Huang, Y., Li, W., Macheret, F., Gabriel, R. A. & Ohno-Machado, L. A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association* **27**, 621–633 (2021).

176. Steffel, J. *et al.* The COMPASS Trial: Net Clinical Benefit of Low-Dose Rivaroxaban plus Aspirin as Compared with Aspirin in Patients with Chronic Vascular Disease. *Circulation* 40–48 (2020) doi:10.1161/CIRCULATIONAHA.120.046048.

177. Clift, A. K. *et al.* Living risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus 19 in adults: national derivation and validation cohort study. *The BMJ* **371**, 1–20 (2020).

178. Romero-Brufau, S., Huddleston, J. M., Escobar, G. J. & Liebow, M. Why the C-statistic is not informative to evaluate early warning scores and what metrics to use. *Crit Care* **19**, (2015).

179. Ford, E. *et al.* What evidence is there for a delay in diagnostic coding of RA in UK general practice records? An observational study of free text. *BMJ Open* **6**, (2016).

180. Buntinx, F., Mant, D., Van Den Bruel, A., Donner-Banzhof, N. & Dinant, G. J. Dealing with low-incidence serious diseases in general practice. *British Journal of General Practice* **61**, 43–46 (2011).

181. Hamilton, W. *et al.* The risk of colorectal cancer with symptoms at different ages and between the sexes: A case-control study. *BMC Med* **7**, 1–9 (2009).

182. Harnan, S. *et al.* Tumour profiling tests to guide adjuvant chemotherapy decisions in early breast cancer: A systematic review and economic analysis. *Health Technol Assess (Rockv)* **23**, 1–327 (2019).

183. Department of Health and Social Care. Health and Care Bill: water fluoridation - GOV.UK. 1–12 (2022).

184. Sullivan, L. M., Massaro, J. M. & D'Agostino, R. B. Presentation of multivariate data for clinical use: The Framingham Study risk score functions. *Stat Med* **23**, 1631–1660 (2004).

185. Violi, F. *et al.* Cardiovascular complications and short-term mortality risk in community-acquired pneumonia. *Clinical Infectious Diseases* **64**, 1486–1493 (2017).

186. Feldman, C., Normark, S., Henriques-Normark, B. & Anderson, R. Pathogenesis and prevention of risk of cardiovascular events in patients with pneumococcal community-acquired pneumonia. *J Intern Med* **285**, 635–652 (2019).

187. Horby, P. W. *et al.* Aspirin in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial. *medRxiv* 2021.06.08.21258132 (2021) doi:10.1101/2021.06.08.21258132.

188. Gluud, C. *et al.* Clarithromycin for 2 weeks for stable coronary heart disease: 6-Year follow-up of the CLARICOR randomized trial and updated meta-analysis of antibiotics for coronary heart disease. *Cardiology* **111**, 280–287 (2008).

189. Hernán, M. A., Sauer, B. C., Hernández-Díaz, S., Platt, R. & Shrier, I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J Clin Epidemiol* **79**, 70–75 (2016).

190. Schulman, S. & Kearon, C. Definition of major bleeding in clinical investigations of antihemostatic medicinal products in non-surgical patients. *Journal of Thrombosis and Haemostasis* vol. 3 692–694 Preprint at https://doi.org/10.1111/j.1538-7836.2005.01204.x (2005).

191. Lipsitch, M., Tchetgen Tchetgen, E. & Cohen, T. Negative Controls: A tool for detecting confounding and bias in observational studies. *Epidemiology* **21**, 383–388 (2010).

192. Vansteelandt, S. & Daniel, R. M. On regression adjustment for the propensity score. *Stat Med* **33**, 4053–4072 (2014).

193. Hernán, M. A. A definition of causal effect for epidemiological research. *Journal of Epidemiology and Community Health* vol. 58 265–271 Preprint at https://doi.org/10.1136/jech.2002.006361 (2004).

194. NICE guideline development group. *Cardiovascular Disease: Risk Assessment and Reduction Including Lipid Modification*. http://www.nice.org.uk/guidance/cg181/chapter/Introduction (2014).

195. Rubin, D. B. For objective causal inference, design trumps analysis. *Annals of Applied Statistics* **2**, 808–840 (2008).

196. McNeil, J. J. *et al.* Effect of Aspirin on All-Cause Mortality in the Healthy Elderly. *New England Journal of Medicine* **379**, 1519–1528 (2018).

197. Suresh Babu, K. & Salvi, S. S. *Aspirin and Asthma**. *CHEST* vol. 118 (2000).

198. Lee, R. U. & Stevenson, D. D. Aspirin-exacerbated respiratory disease: Evaluation and management. *Allergy, Asthma and Immunology Research* vol. 3 3–10 Preprint at https://doi.org/10.4168/aair.2011.3.1.3 (2010).

199. Segal, R. *et al.* Early and late effects of low-dose aspirin on renal function in elderly patients. *American Journal of Medicine* **115**, 462–466 (2003).

200. Hernán, M. A. The C-word: Scientific euphemisms do not improve causal inference from observational data. *Am J Public Health* **108**, 616–619 (2018).

201. Cole, S. R. & Hernán, M. A. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol* **168**, 656–664 (2008).

202. Cea Soriano, L., Soriano-Gabarró, M. & García Rodríguez, L. A. Validation of low-dose aspirin prescription data in The Health Improvement Network: How much misclassification due to over-the-counter use? *Pharmacoepidemiol Drug Saf* **25**, 392–398 (2016).

203. Maringe, C. *et al.* Reflection on modern methods: Trial emulation in the presence of immortal-time bias. Assessing the benefit of major surgery for elderly lung cancer patients using observational data. *Int J Epidemiol* **49**, 1719–1729 (2020).

204. Jones, M. & Fowler, R. Immortal time bias in observational studies of time-to-event outcomes. *J Crit Care* **36**, 195–199 (2016).

205. Lévesque, L. E., Hanley, J. A., Kezouh, A. & Suissa, S. Problem of immortal time bias in cohort studies: Example using statins for preventing progression of diabetes. *BMJ (Online)* **340**, 907–911 (2010).

206. Leite, A., Andrews, N. J. & Thomas, S. L. Assessing recording delays in general practice records to inform near real-time vaccine safety surveillance using the Clinical Practice Research Datalink (CPRD). *Pharmacoepidemiol Drug Saf* **26**, 437–445 (2017).

207. Herrett, E. *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 (Online)* **346**, 1–12 (2013).

208. Badve, M. S., Zhou, Z., van de Beek, D., Anderson, C. S. & Hackett, M. L. Frequency of post-stroke pneumonia: Systematic review and meta-analysis of observational studies. *International Journal of Stroke* vol. 14 125–136 Preprint at https://doi.org/10.1177/1747493018806196 (2019).

209. Mant, J. *et al.* Warfarin versus aspirin for stroke prevention in an elderly community population with atrial fibrillation (the Birmingham Atrial Fibrillation Treatment of the Aged Study, BAFTA): a randomised controlled trial. *Lancet* **370**, 493–503 (2007).

210. Hsu, J. C. *et al. Aspirin Instead of Oral Anticoagulant Prescription in Atrial Fibrillation Patients at Risk for Stroke*. www.ncdr.com. (2016).

211. Petersen, I., Douglas, I. & Whitaker, H. Self controlled case series methods: an alternative to standard epidemiological study designs. *BMJ* **354**, i4515 (2016).

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

213. Maun, A., Björkelund, C. & Arvidsson, E. Primary care utilisation, adherence to guideline-based pharmacotherapy and continuity of care in primary care patients with chronic diseases and multimorbidity – a cross-sectional study. *BMC Primary Care* **24**, (2023).

214. Griffith, G. J. *et al.* Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nat Commun* **11**, 1–12 (2020).

215. Nguyen, T. Q., Dafoe, A. & Ogburn, E. L. The magnitude and direction of collider bias for binary variables. *Epidemiol Methods* **8**, (2019).

216. Swanson, S. A. & Hernán, M. A. Commentary: How to report instrumental variable analyses (suggestions welcome). *Epidemiology* **24**, 370–374 (2013).

217. Hernán, M. A. & Robins, J. M. Instruments for causal inference: An epidemiologist's dream? *Epidemiology* **17**, 360–372 (2006).

218. Robins, J. M., Epidemiology, S., Mar, N., Robins, M. & Greenland, S. Identifiability and Exchangeability for Direct and Indirect Effects. *Epidemiology* **3**, 143–155 (1992).

219. Greenland, S. & Robins, J. M. Identifiability, exchangeability, and epidemiological confounding. *Int J Epidemiol* **15**, 413–418 (1986).

220. Ashworth, M. & Armstrong, D. The relationship between general practice characteristics and quality of care: A national survey of quality indicators used in the UK Quality and Outcomes Framework, 2004-5. *BMC Fam Pract* **7**, (2006).

221. Baron, J. A. *et al.* Gastrointestinal adverse effects of short-term aspirin use: A meta-analysis of published randomized controlled trials. *Drugs in R and D* vol. 13 9–16 Preprint at https://doi.org/10.1007/s40268-013-0011-y (2013).

222. Toews, I. *et al.* Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials: a meta-epidemiological study. *Cochrane Database of Systematic Reviews* **2024**, (2024).

223. Vandenbroucke, J. P. When are observational studies as credible as randomised trials? *Lancet* **363**, 1728–1731 (2004).

224. Vandenbroucke, J. P. The HRT controversy: observational studies and RCTs fall in line. *The Lancet* vol. 373 1233–1235 Preprint at https://doi.org/10.1016/S0140-6736(09)60708-X (2009).

225. Lima, J. A. C. *et al.* Statin-induced cholesterol lowering and plaque regression after 6 months of magnetic resonance imaging-monitored therapy. *Circulation* **110**, 2336–2341 (2004).

226. Van Rosendael, A. R. *et al.* Association of Statin Treatment with Progression of Coronary Atherosclerotic Plaque Composition. *JAMA Cardiol* **6**, 1257–1266 (2021).

227. Hulten, E., Jackson, J. L., Douglas, K., George, S. & Villines, T. C. The effect of early, intensive statin therapy on acute coronary syndrome: A meta-analysis of randomized controlled trials. *Arch Intern Med* **166**, 1814–1821 (2006).

228. DataLab, E. OpenPrescribing.net. *University of Oxford* (2020).

229. Matthews, A. *et al.* Impact of statin related media coverage on use of statins: Interrupted time series analysis with UK primary care data. *BMJ (Online)* **353**, (2016).

230. Finnikin, S., Willis, B. H., Ryan, R., Evans, T. & Marshall, T. Factors predicting statin prescribing for primary prevention: A historical cohort study. *British Journal of General Practice* **71**, E219–E225 (2021).

231. Seaman, S. R. & White, I. R. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res* **22**, 278–295 (2013).

232. Chesnaye, N. C. *et al.* An introduction to inverse probability of treatment weighting in observational research. *Clin Kidney J* **15**, 14–20 (2022).

233. Taylor, F. *et al.* Statins for the primary prevention of cardiovascular disease. *Cochrane Database of Systematic Reviews* **2017**, CD004816 (2013).

234. Zhang, H. *et al.* Discontinuation of statins in routine care settings, A cohort study. *Ann Intern Med* **158**, 526–534 (2013).

235. Round, T., Ashworth, M., Esperance, V. L. & Moller, H. Cancer detection via primary care urgent referral and association with practice characteristics: A retrospective cross-sectional study in England from 2009/2010 to 2018/2019. *British Journal of General Practice* **71**, E826–E835 (2021).

236. Gulliford, M. C. *et al.* Continued high rates of antibiotic prescribing to adults with respiratory tract infection: survey of 568 UK general practices. *BMJ Open* **4**, e006245 (2014).

237. Fulcher, J. *et al.* Efficacy and safety of LDL-lowering therapy among men and women: Meta-analysis of individual data from 174 000 participants in 27 randomised trials. *The Lancet* **385**, 1397–1405 (2015).

238. Agarwal, A. *et al.* Fixed-dose combination therapy for the prevention of atherosclerotic cardiovascular disease. *Nat Med* **30**, 1199–1209 (2024).

239. Brookhart, M. A., Wang, P., Solomon, D. H. & Schneeweiss, S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology* **17**, 268–275 (2006).

240. Davies, N. M. *et al.* Physicians' prescribing preferences were a potential instrument for patients' actual prescriptions of antidepressants. *J Clin Epidemiol* **66**, 1386–1396 (2013).

241. Rassen, J. A., Brookhart, M. A., Glynn, R. J., Mittleman, M. A. & Schneeweiss, S. Instrumental variables II: instrumental variable application—in 25 variations, the physician prescribing preference generally was strong and reduced covariate imbalance. *J Clin Epidemiol* **62**, 1233–1241 (2009).

242. Yu, M., Xie, D., Wang, X., Weiner, M. G. & Tannen, R. L. Prior event rate ratio adjustment: Numerical studies of a statistical method to address unrecognized confounding in observational studies. *Pharmacoepidemiol Drug Saf* **21**, 60–68 (2012).

243. Lee, J., Wright-drakesmith, C. & Warren-gash, C. The Lancet Development and external validation of a risk prediction score (DASHI) for cardiovascular events following acute respiratory infections : derivation and validation retrospective cohort study. **85**,.

244. Williamson, E. J., Walker, A. J. & Bhaskaran, K. et al. Factors associated with COVID-19-related death using OpenSAFELY. *Nature* (2020) doi:https://doi.org/10.1038/s41586-020-2521-4.

245. Banerjee, A. *et al.* Excess deaths in people with cardiovascular diseases during the COVID-19 pandemic. *Eur J Prev Cardiol* (2021) doi:10.1093/eurjpc/zwaa155.

246. Zarifkar, P. *et al.* Clinical Characteristics and Outcomes in Patients with COVID-19 and Cancer: a Systematic Review and Meta-analysis. *Clin Oncol* **33**, e180–e191 (2021).

247. Ip, S. *et al.* Cohort study of cardiovascular safety of different COVID-19 vaccination doses among 46 million adults in England. *Nat Commun* **15**, (2024).

248. Aballéa, S. *et al.* The cost-effectiveness of influenza vaccination for people aged 50 to 64 years: An international model. *Value in Health* **10**, 98–116 (2007).

249. Usman, M. S. *et al.* Effect of SGLT2 Inhibitors on Cardiovascular Outcomes Across Various Patient Populations. *J Am Coll Cardiol* **81**, 2377–2387 (2023).

250. Anker, S. D. *et al.* Empagliflozin in Heart Failure with a Preserved Ejection Fraction. *New England Journal of Medicine* **385**, 1451–1461 (2021).

251. Kristensen, S. L. *et al.* Cardiovascular, mortality, and kidney outcomes with GLP-1 receptor agonists in patients with type 2 diabetes: a systematic review and meta-analysis of cardiovascular outcome trials. *Lancet Diabetes Endocrinol* **7**, 776–785 (2019).

252. Lincoff, A. M. *et al.* Semaglutide and Cardiovascular Outcomes in Obesity without Diabetes. *New England Journal of Medicine* **389**, 2221–2232 (2023).

253. Yeo, Y. H. *et al.* Shifting Trends in the Indication of Glucagon-like Peptide-1 Receptor Agonist Prescriptions: A Nationwide Analysis. *Ann Intern Med* **177**, 1289–1291 (2024).

254. Dhingra, R. & Vasan, R. S. Age As a Risk Factor. *Medical Clinics of North America* vol. 96 87–91 Preprint at https://doi.org/10.1016/j.mcna.2011.11.003 (2012).

255. Schober, P. & Vetter, T. R. Survival analysis and interpretation of time-to-event data: The tortoise and the hare. *Anesth Analg* **127**, 792–798 (2018).

256. Pouwels, K. B., Voorham, J., Hak, E. & Denig, P. Identification of major cardiovascular events in patients with diabetes using primary care data. *BMC Health Serv Res* **16**, (2016).

257. Campbell, M., Piaggio, G., Elbourne, D., Altman, D. & Group., for the C. Consort 2010 statement: extension to cluster randomised trials. *BMJ : British Medical Journal* **345**, (2010).

258. Swanson, S. A., Robins, J. M., Miller, M. & Hernï¿½n, M. A. Selecting on treatment: A pervasive form of bias in instrumental variable analyses. *Am J Epidemiol* **181**, 191–197 (2015).

259. Robins, J. M., Greenland, S. & Hu, F. C. Estimation of the Causal Effect of a Time-Varying Exposure on the Marginal Mean of a Repeated Binary Outcome. *J Am Stat Assoc* **94**, 687–700 (1999).

260. Shinozaki, T., Matsuyama, Y. & Ohashi, Y. Estimation of controlled direct effects in time-varying treatments using structural nested mean models: Application to a primary prevention trial for coronary events with pravastatin. *Stat Med* **33**, 3214–3228 (2014).

261. Martin, S. A., Johansson, M., Heath, I., Lehman, R. & Korownyk, C. Sacrificing patient care for prevention: Distortion of the role of general practice. *BMJ* (2025) doi:10.1136/bmj-2024-080811.

262. Wald, N. J., Hingorani, A. D., Vale, S. H., Bestwick, J. P. & Morris, J. The polypill in the primary prevention of heart attacks and strokes: Overcoming barriers to implementation. *Journal of Medical Screening* vol. 31 66–69 Preprint at https://doi.org/10.1177/09691413241235486 (2024).

263. Briscoe, E. & Feldman, J. Conceptual complexity and the bias/variance tradeoff. *Cognition* **118**, 2–16 (2011).

264. England, P. H. *Influenza Vaccine Effectiveness in Adults and Children in Primary Care in the UK: Provisional End-of-Season Results 2015-16*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/779476/Influenza_vaccine_effectiveness_in_primary_care_2015_2016.pdf (2016).

265. Young, B., Sadarangani, S., Jiang, L., Wilder-Smith, A. & Chen, M. I. C. Duration of influenza vaccine effectiveness: A systematic review, meta-analysis, and meta-regression of test- negative design case-control studies. *Journal of Infectious Diseases* **217**, 731–741 (2018).

266. Palmer, T. M. *et al.* Using multiple genetic variants as instrumental variables for modifiable risk factors. *Stat Methods Med Res* **21**, 223–242 (2012).

267. Xiao, W. *et al.* Effects of genetically proxied lipid-lowering drugs on acute myocardial infarction: a drug-target mendelian randomization study. *Lipids Health Dis* **23**, (2024).

268. Fan, Y. xiang *et al.* Genetically Proxied Antiplatelet Drug Target Perturbation and Risk of Aneurysmal Subarachnoid Hemorrhage: A Mendelian Randomization Analysis. *World Neurosurg* **196**, (2025).

269. Smeeth, L. *et al.* Risk of Myocardial Infarction and Stroke after Acute Infection or Vaccination. *New England Journal of Medicine* **351**, 2611–2618 (2004).

270. Sebastian, S., Stein, L. K. & Dhamoon, M. S. Infection as a Cardiovascular Trigger: Associations Between Different Organ System Infections and Cardiovascular Events. *Am J Med* (2020) doi:10.1016/j.amjmed.2020.04.033.

271. Sebastian, S., Stein, L. K. & Dhamoon, M. S. Infection as a Stroke Trigger: Associations between Different Organ System Infection Admissions and Stroke Subtypes. *Stroke* **50**, 2216–2218 (2019).

272. Corrales-Medina, V. F. *et al.* Acute bacterial pneumonia is associated with the occurrence of acute coronary syndromes. *Medicine* **88**, 154–159 (2009).

273. Moberley, S., Holden, J., Tatham, D. P. & Andrews, R. M. Vaccines for preventing pneumococcal infection in adults. *Cochrane Database of Systematic Reviews* **2013**, (2013).

274. Ihara, H. *et al.* 23-Valent Pneumococcal Polysaccharide Vaccine Improves Survival in Dialysis Patients By Preventing Cardiac Events. *Vaccine* **37**, 6447–6453 (2019).

275. Caldeira, D. *et al.* The association of influenza infection and vaccine with myocardial infarction: systematic review and meta-analysis of self-controlled case series. *Expert Rev Vaccines* **18**, 1211–1217 (2019).

276. Clar, C., Oseni, Z., Flowers, N., Keshtkar-Jahromi, M. & Rees, K. Influenza vaccines for preventing cardiovascular disease. *Cochrane Database of Systematic Reviews* **133**, 384 (2015).

277. Chen, Y., Williams, E. & Kirk, M. Risk factors for acute respiratory infection in the Australian community. *PLoS One* **9**, 1–7 (2014).

278. Robinson, C. H. *et al.* The relationship between duration and quality of sleep and upper respiratory tract infections: A systematic review. *Family Practice* vol. 38 802–810 Preprint at https://doi.org/10.1093/fampra/cmab033 (2021).

279. Millett, E. R. C., De Stavola, B. L., Quint, J. K., Smeeth, L. & Thomas, S. L. Risk factors for hospital admission in the 28 days following a community-acquired

pneumonia diagnosis in older adults, and their contribution to increasing hospitalisation rates over time: A cohort study. *BMJ Open* **5**, e008737 (2015).

9 Appendix – Supplementary materials for chapter three

9.1 Supplementary methods for chapter three

9.1.1 Identification of potential predictors

I performed a rapid review to look for evidence of factors that influence risks infection-related CVD, but because there was no pre-existing modelling, I also looked to previous general CVD prediction, and for evidence of evidence for elevated risks associated with respiratory infections. I looked for evidence of factors that alter risk of being infected, the risk of symptomatic infection, and the risk of severe or complicated infections. These necessarily follow on from each other, and all might increase the risk of CVD events. I looked for systematic reviews and clinical guidelines, as well as epidemiological evidence.

9.1.2 Codes employed:

CPRD codes can be found online here: https://github.com/Protocols-For-Research/CPRD-codes-CVD-infection-risk. Here follows the ICD and OPCS codes used for the outcome:

9.1.2.1 ICD codes used to identify CVD deaths in ONS data:

Cerebrovascular: I63 I636 I634 I631 I633 I630 I635 I632 I639 I61 I613 I614 I611 I610 I612 I615 I616 I619 I629 I676 I621 I661 I651 I652 I663 I66 I660 I664 I653 I668 I658 I662 I65 I669 I659 I650 I638 I618 I62 I608 I693 I69 I691 I692 I694 I690 I64 I60 I602 I604 I600 I607 I601 I606 I603 I605 I609

Cardiac: I46 I460 I469 I516 I249 I21 I219 I214 I210 I211 I212 I213 I201 I209 I250 I251 I231 I23 I25 I259 I254 I240 I241 I230 I255 I24 I238 I248 I208 I258 I233 I234 I235 I256 I22 I220 I221 I228 I229 I461 I200 I232

9.1.2.2 NHS ‘opcs’ procedure codes used in HES data to identify CVD events

K424 K412 K454 K431 K442 K458 K499 K402 K758 K429 K432 K434 K498 K483 K451 K754 K45 K40 K503 K43 K459 K455 K456 K433 K493 K414 K46 K41 K404 K759 K501

K453 K438 K42 K409 K491 K403 K423 K44 K468 K408 K509 K494 K751 K421 K452

K449 K75 K419 K418 K422 K469 K428 K401 K49 K413 K753 K411 K492 K508 K50

K441 K439 K482 K448 K504 K752 L372 K471 L318 L303 L314 L31 K502 L311 L319

9.2 Supplementary results chapter three:

9.2.1 Variable identification – CVD risk

I used Damen's 2016 systematic review as a source for potential predictors of CVD.[7] They found 363 prediction model developments, and 473 external validations.[7] Overall more than 100 different predictors were used in different models, the mean number of predictors per model was seven, but some models had as many as 80. Most models rely on some demographic features, particularly age, sex, family history of CVD, and ethnicity. Measures of deprivation or location also feature. Smoking and BMI were the most frequently included 'lifestyle factors'. Diabetes was the most featured comorbidity, and a measure of blood pressure (BP) was usually included, most commonly systolic BP. Some scores included measurements from echocardiography or other imaging, or genetics, but these are not available for the UK primary care setting, and I did not consider them further. A measure of blood lipids was common, most commonly scores used total cholesterol as a predictor, despite our understanding that LDL cholesterol is the dangerous fraction. Some scores include blood biomarkers such as CRP, and markers of renal function such as creatinine. I tabulated the predictors and methods used in five primary care CVD prediction models (Table S24: Properties of five cardiovascular risk calculators).

Table S24: Properties of five cardiovascular risk calculators

Characteristic	Primary Care CVD prediction model				
	WHO/ISH	**SCORE**	**Framingham 2008**	**QRISK2**	**QRISK3**
Outcome	Fatal or nonfatal major cardiovascular event (myocardial infarction or stroke)	Fatal cardiac event	Composite CVD* and individual outcomes of: CHD, stroke, intermittent claudication and congestive heart failure	Composite of coronary heart disease, ischaemic stroke or TIA	Composite of coronary heart disease, ischaemic stroke or TIA
Timeframe (main)	10 years	10 years	10 years	10 years	10 years (1-15)
Data source	Unpublished	12 prospective studies in European countries	Prospective study	Routinely collected GP records	Routinely collected records
Missing values	Unpublished	Not addressed	Full cases at baseline	Five imputations	Five imputations
Derivation method	Unpublished	Weibull proportional hazards model checked with Cox regression	Cox regression	Cox regression	Cox regression
Ethnicity	-	-	-	9 categories	9 categories
Age	Banded, over 18	40-65 years banded in charts, continuous time at risk in model	Continuous, 30-74 years	Continuous 25-84 years	Continuous 25-84 years
Sex	Separate models	Separate models	Separate models	Separate models	Separate models
Body habitus	-	-	-	BMI	BMI
Smoking	Binary	Binary	Binary	Non-smoker, former smoker, 1-9 per day, 10-19 per day, 20+	Non-smoker, former smoker, 1-9 per day, 10-19 per day, 20+
Systolic Blood pressure	Banded, at assessment	Banded, various	Systolic measured biannually	Most recent	Most recent
Antihypertensive use	-	-	Binary	Binary	Binary
Blood pressure variation	-	-	-	-	SD of last 5 years measurements
Diabetes Mellitus	Binary present/absent	-	Binary present/absent	None, type 1, or 2	None, type 1, or 2
Cholesterol	Total, banded	Total or total:HDL ratio, banded	Total and HDL	Total: HDL ratio	Total: HDL ratio

Geography or socioeconomic indicator	WHO sub-region	Separate versions for high and low risk countries	-	UK Postcode level Townsend deprivation score	UK Postcode level Townsend deprivation score
Family history of CVD	-	-	-	Coronary disease in a first degree relative < 60 years	Coronary disease in a first degree relative < 60 years
Rheumatoid Arthritis	-	-	-	Binary	Binary
Atrial fibrillation	-	-	--	Binary, including flutter and paroxysmal	Binary, including flutter and paroxysmal
Renal disease	-	-	-	CKD 4 or 5 or major disease **	CKD 3, 4 or 5 or major disease**
Migraine	-	-	-	-	Binary: any diagnosis***
Corticosteroid use	-	-	-	-	Binary: BNF chapter 6.3.2
SLE	-	-	-	-	Binary (inc Libman Sacks endocarditis)
Atypical antipsychotics	-	-	-	-	Binary†
Severe mental illness diagnosis	-	-	-	-	Binary - psychosis, schizophrenia, bipolar
HIV/AIDS	-	-	-	-	Binary
Erectile dysfunction	-	-	-	-	Diagnosis or treatment (BNF chapter 7.4.5)

WHO/ISH is World Health Organisation/ International Society of Hypertension.
*Composite includes coronary disease (coronary death, myocardial infarction, coronary insufficiency, and angina), cerebrovascular events (including ischemic stroke, hemorrhagic stroke, and transient ischemic attack), peripheral artery disease (claudication), and heart failure.
**Nephrotic syndrome, glomerulonephritis, pyelonephritis, transplant, dialysis
***Classic migraine, atypical migraine, abdominal migraine, cluster headaches, basilar migraine, hemiplegic migraine, and migraine with or without aura
† Amisulpride, aripiprazole, clozapine, lurasidone, olanzapine, paliperidone, quetiapine, risperidone, sertindole, and zotepine

9.2.2 Variable identification – infections with evidence of association with CVD events

There is evidence that infections are associated with CVD events from different settings and infections.

9.2.2.1 Undifferentiated infections

Smeeth examined the risk of prior undifferentiated respiratory infection in primary care patients with myocardial infarction and stroke, using a case series design that split each patient's clinical record into risk and control periods.[269] The incidence ratio for a first MI or stroke was highest in the three days following a diagnosis of a systemic respiratory infection. The risk was five times higher for MI and three times higher for stroke.

Sebastian used routinely collected data from New York State community hospitals to examine the risk of MI following different infections.[270] They used a similar design to Smeeth and included infections of the skin, urinary tract (UTI), respiratory tract, abdomen and sepsis. The risk of MI was elevated following all infection types. The risk was highest following respiratory infections, and lowest following abdominal infections. Again the risk was highest in the first seven days after infection was diagnosed, with an odds ratio of 4.00 (95% CI 3.31-4.83) for respiratory infection, and 1.75 (95% CI 0.73- 4.17) following abdominal infection.

Sebastian and colleagues also investigated the risk of stroke following infections in another paper.[271] Ischaemic strokes were more likely immediately following all of the infection types. The infection with the greatest magnitude of association with stroke in the first week was UTI (OR 5.32 95% CI 3.69-7.68). Respiratory infection was associated with an OR of 3.20 in the first week (95% CI 2.47–4.15). Smaller numbers of other types of stroke were observed. Intracranial haemorrhage was associated with respiratory infection in the first seven days (OR 2.1 95% CI 1.21-3.70). Sepsis had the association of greatest magnitude with intracranial haemorrhage (OR 3.24 95% CI 1.06–9.97). Subarachnoid haemorrhage was the rarest stroke

subtype and was only associated with respiratory infections (OR 3.67 95% CI 1.49-9.04 at 14 days).

9.2.2.2 Bacterial pneumonia

I found epidemiological evidence that bacterial pneumonia is associated with CVD events. A Texan study employed both self-controlled case-series and case-control designs to explore the association between acute coronary syndrome (ACS) and pneumonia due to *Streptococcus pneumoniae* or *Haemophilus pneumoniae,* in the 15 days after hospital admission.[272] They found that pneumonia was a strong risk factor for developing ACS (OR 8.52, 95% CI 3.35-22.23) in an analysis of 206 pneumonia cases. This was a stronger association than prior CVD, or peripheral vascular disease, or than having more than two traditional risk factors. The case-series showed a high incidence rate ratio for ACS in the three and 15 day periods following admission (three day IRR 132.0, 95% CI 69.2 - 255.6; 15 day IRR 47.6, 95% CI 24.5 - 92.5). This analysis included only 21 ACS events, hence the wide confidence intervals.

A Cochrane review pooled 18 RCTs of 23-valent pneumococcal vaccine and found strong evidence of a reduction in invasive pneumococcal disease (OR 0.26, 95% CI 0.14 - 0.45, I^2 = 0%).[273] Unfortunately, they didn't include cardiovascular events as an outcome, and there were no adverse events included. There is some evidence that vaccination against pneumococcus appears to reduce cardiovascular deaths. In a retrospective study of dialysis patients in Japan, use of 23-valent vaccine reduced all cause death, but not pneumonia.[274] Lower deaths were driven by fewer cardiovascular deaths in the vaccinated group; (Hazard Ratio (HR) 0.36, 95% CI 0.18–0.71).

9.2.2.3 Influenza

A 2019 systematic review of self-controlled studies of influenza, influenza vaccine and MI combined three studies.(Caldeira et al., 2019b) The risk of MI was highest 1-3 days post infection (IRR 5.79 95%CI 3.59–9.38, $I^2 = 0\%$) and similar between 4-7 days (IRR 4.52; 95%CI: 2.80–7.32, $I^2 = 0\%$), but not detectable beyond that. In the first 4 weeks following influenza vaccination the risk of MI was reduced compared to baseline (IRR 0.84 95%CI 0.78–0.91). The authors felt this may represent a bias analogous to the healthy worker effect – healthy user effect bias – where people who are suitable for an influenza vaccine, or choose to have it, tend to be in better health.

A Cochrane review of randomised controlled trials (RCTs) of vaccines found that in secondary prevention there was likely a reduction in both cardiovascular death, and events.[276] The risk ratio (RR) for cardiovascular deaths was 0.45 (95% CI 0.26 - 0.76, $I^2 = 58\%$). There were too few events in primary prevention trials to determine an effect.

9.2.2.4 Other respiratory viruses

A 2018 paper modelled population level CVD events with the level of laboratory confirmed viruses in the population, in a time series analysis.[25] Associations with MI were highest in the oldest age strata – those over 75 years – and they identified associations between MI and influenza, RSV, Human Metapneumovirus, adenovirus and rhinovirus. These viruses were also associated with ischaemic stroke admission. They estimated up to 16 percent of MIs in the over 75s were attributable to these viruses taken together, and 19.7% of MIs in the 65-74 year olds. This is hundreds of excess cases in weeks with peaks of viral infections. An earlier case series which included Office of National Statistics mortality data found influenza, RSV and bacterial respiratory infections appeared to account for 10,000 circulatory deaths per year.[6]

A study of Canadian inpatient records with laboratory confirmed viruses estimated risks of acute MI using a self-controlled design.[17] They compared a 'risk period' of seven days post laboratory-based diagnosis with a period spanning from one year earlier to one year later. They found influenza B was the agent with the greatest risk: IRR 10.11 (95% CI 4.37 - 23.38). Influenza A was also a risk (IRR 5.17, 95% C 3.02 - 8.84), as was RSV (IRR 3.51 95% CI 1.11 - 11.12), and a combined category of other viruses (2.77, 95% CI 1.23 - 6.24). They didn't find an association in time periods longer than two weeks.

9.2.3 Variable identification - risk factors for symptomatic respiratory infection

Most symptomatic respiratory infections are mild, self-limiting viral illnesses, but there was little evidence regarding risk factors for mild infections. A survey of randomly selected people in Australia found acute respiratory infection was commonest in children and people with contact with children and decreased with age in a linear fashion. [277] The only chronic disease identified as a risk factor for symptomatic colds or 'flu was asthma.[277]

We know that infections can trigger asthma, and so it is possible that reactive airways make an infection more obvious rather than more likely to occur in the first place. I have not been able to find evidence that addresses this, thought there is some circumstantial evidence. A review found there are some studies that show asthma is associated with greater colonisation with potentially pathogenic bacteria in the nasopharynx.[167] There might be some immunological reason that asthma is associated with infection in other systems as well; enteric E. coli infections are associated with asthma, but overall there is a paucity of evidence that asthma causes infection.[167] Never the less, I included asthma in the potential predictors.

Other evidence related to factors not available in routinely collected records. A systematic review that I supervised (published in 2021) found that sleep duration and quality was

associated with upper respiratory infections, and prior systematic reviews showed a similar finding for pneumonia.[278]

9.2.4 Risk factors for hospitalisation or death with pneumonia

There were several epidemiological cohort studies examining risk factors for pneumonia hospitalisations and deaths. A study used CPRD to identify risk factors for hospitalisation within a month of a diagnosis of community acquired pneumonia (CAP).[279] Age was a risk factor, with the median age of 81 in this population, and risk of admission highest in those aged 80 to 84. However, the point estimates for the likelihood of admission seemed to decline for patients older then 84, and especially over 90 years. This may represent a reluctance of these patients and their clinicians to admit to hospital, rather than a lower risk of severe infection. The same reduction in admissions exists in care home residents, people with dementia, bedsores, incontinence and terminal illness (Table S25: Risk factors associated with hospital admission following a diagnosis of Community Acquired Pneumonia). This seems likely to be describing a cohort of patients who are not going to get admitted to hospital, irrespective of their illnesses.

Table S25: Risk factors associated with hospital admission following a diagnosis of Community Acquired Pneumonia

Decreased risk of admission after CAP diagnosis	No change in risk of admission after CAP diagnosis	Increased Risk of admission after CAP diagnosis
Demographics		
Female sex Vs male		
		Age over 70 years Vs younger
Residential care	Lives alone or in sheltered accommodation	Recent carer
Diagnoses coded		
Dementia	Cerebrovascular disease	Heart failure
Terminal illness	Neurological diseases	Chronic lung disease
Bedsore/ulcers	Excessive alcohol consumption	Peripheral Vascular disease
Low weight/poor nutrition	Fatigue	Immune disorders
Catheter/incontinence	Anxiety/depression	Ischaemic heart disease (pre or post MI)
	Renal disease - mild	Renal disease - severe
	Hemiplegia	Connective tissue disease
	Falls	Peptic ulcer
		Liver disease (mild or severe)
		Diabetes (with or without complications)
		Solid cancer (metastases or not)
		Visual impairment
Prescriptions		
	Immunosuppressant drugs	
		Statins in previous 6 months
		PO steroids in last 90 days
Antibiotics in previous 8-28 days	Antibiotics in previous week	
Inhaled steroids 61-180 or > 365 days ago	Inhaled steroids within 60 or 181-365 days ago	
Vaccines		
Last influenza vaccine 14 days to 5 years pre CAP (vs none)	Last influenza vaccine >5 years ago	
	Last pneumococcal vaccine 14 days to 5 years before CAP (vs none)	Pneumococcal vaccine > 5 years before CAP (vs none)

Data from Millett ERC, De Stavola BL, Quint JK, Smeeth L, Thomas SL. Risk factors for hospital admission in the 28 days following a community-acquired pneumonia diagnosis in older adults, and their contribution to increasing hospitalisation rates over time: A cohort study. BMJ Open 2015; 5:e008737. Available at: http://bmjopen.bmj.com/content/5/12/e008737.abstract.

A prospective population based cohort study from Denmark, originally designed for cardiovascular epidemiology, was used to examine risk for pneumonia causing death or hospitalisation.[164] Apart from increasing age, which was strongly correlated with hospitalisation, they found low FEV1 (Forced Expiratory Volume in one second, measure of lung function showing likely obstructive lung disease, such as asthma or COPD) was associated with pneumonia hospitalisation, as was self-reported mucus production, smoking status (women only), asthma (women only) and stroke (men only). Deaths from pneumonia were much less common, with only 27 in the cohort. Age and FEV1 were was strongly associated with death, as was a BMI less than 20.

A cohort study of COPD patients in Taiwan found further evidence that COPD patients with lower FEV1 are more likely to be diagnosed with community acquired pneumonia, as were patients taking inhaled steroids, those with more frequent exacerbations, older patients, and those with comorbid CVD.[44]

9.2.5 Risk factors for severe outcomes of influenza

A systematic review of risk factors for severe influenza which formed the basis for some of the UK influenza vaccination guidance recommendations found the overall evidence base was poor.[163] Nevertheless they were able to identify some risk factors for seasonal influenza (Table S26: Summary of systematic review of risk factors for severe influenza outcomes). Increasing age was a risk for death and admission to hospital. Obesity was only identified as a risk for death by a single study. The compound category of 'chronic lung disease' was associated with admission to hospital, ICU admission and respiratory support. For specific respiratory conditions, asthma was associated with pneumonia only, but COPD was also associated with respiratory support. CVD increased the risk of death, respiratory support, ICU and hospital admissions. Patients with immunocompromised status seemed to be at

higher risk of death but not pneumonia. Diabetes seemed to be a risk for admission, but not other outcomes. Neuromuscular diseases was associated with death only.

The same study dealt with pandemic influenza separately. Perhaps because pandemics tend to focus research there were more studies of pandemic ‘flu than seasonal ‘flu, mostly from H1N1 in 2009-10. There was a complex picture in the evidence for age, with young people being at particularly low risk of death in community studies, presumably because to get into hospital one must be closer to death than the population at large. Elderly adults had a lower chance of translating their hospital admission into an ICU visit, a finding that could be because getting ICU care can be based on the patient’s functional baseline.

There was slim evidence addressing ethnicity, but some pointers to a higher admission rate for Black and Hispanic people, but lower for Australian Aboriginal people and Torres Strait Islanders. Black people seemed less likely to get admitted to ICU. These findings come from societies where these ethnicities are disadvantaged, which might explain these findings.

Table S26: Summary of systematic review of risk factors for severe influenza outcomes

Risk factor	Pneumonia		Hospitalisation		ICU admission or respiratory support		Death	
	Seasonal	Pandemic	Seasonal	Pandemic	Seasonal	Pandemic	Seasonal	Pandemic
Age (elderly Vs adult)	=	-	+	+	NA	-	+	+
Sex (male)	=	=	+	=	=	=	=	=
Ethnicity vs White								
Asian	=	NA	=	=	NA	=	NA	=
Black	=	NA	=	+	NA	+	=	=
Hispanic	=	NA	-	+	NA	=	NA	=
American	NA	=	=	=	NA	=	NA	=
Australian	NA	NA	NA	-	NA	=	NA	=
Pregnancy	NA	=	NA	+	NA	-	=	=
Obesity	NA	=	NA	+	NA	+	+	+
Chronic Lung disease	=	=	+	+	+	+	+	+
Asthma	+	=	NA	=	=	=	=	=
COPD	NA	=	NA	+	+	+	=	+
OSA	NA	NA	NA	NA	NA	=	NA	+
CVD - any	+	=	+	+	+	+	+	+
Hypertension	NA	=	NA	=	NA	=	=	=
CVA	NA	NA	NA	+	NA	NA	=	+
Malignancy	=	=	+	+	NA	=	=	+
Neuromuscular disease	+	=	NA	+	NA	+	+	+
Neurocognitive	+	NA	NA	+	=	+	=	=
Immunosuppression	=	=	=	+	=	=	+	+
Steroid use	NA	NA	+	=	NA	=	=	=
Endocrine disorders	NA	NA	NA	+	NA	=	+	NA
Diabetes	=	=	+	+	NA	+	=	+
Anaemia	=	=	NA	+	=	=	NA	+

Liver disease	NA	NA	NA	=	NA	+	=	+
Metabolic disease	=	=	NA	=	NA	+	=	+
Renal disease	=	=	NA	+	=	=	=	+

NA – not applicable (no data), + risk factor associated with increased risk of outcome, = no association found, - associated with reduced risk of outcome
Data summarized from: Mertz D, Kim TH, Johnstone J, et al. Populations at risk for severe or complicated influenza illness: systematic review and meta-analysis. Bmj 2013; 347:f5061–f5061. Available at: http://www.bmj.com/cgi/doi/10.1136/bmj.f5061.

9.2.5.1 Vaccination recommendations

A challenge for policymakers, especially given the paucity of evidence from community settings, is to identify people at high risk of severe outcomes, to target them for influenza vaccination. The UK maintains a list of risk factors in a document known as the 'green book.[65] This defines the groups of people eligible for free NHS influenza vaccinations, and for oseltamivir treatment when surveillance indicates influenza has passed a threshold in the community. The UK also used this list to identify people who might be at risk of Covid-19, in the absence of evidence when it was a new disease. The document is based on reviews of published research, and if you follow the references carefully back, you find the recommendation for vaccination in CVD is based on Smeeths NEJM study, despite it not addressing this question.[2]

Public Health England, prior to their dissolution, also used national statistics to identify conditions associated with laboratory confirmed influenza cases being admitted to intensive care units or dying (Table S27 UK influenza vaccination categories and estimated relative risk of death with a laboratory confirmed influenza infection published by Public Health England).[65] This approach doesn't address the thousands of deaths each year caused by but not attributed to influenza, and many CVD deaths would come into this category. To get a laboratory confirmed influenza test is quite difficult, particularly in 2010 when the study was done. One would need to be tested in hospital, which was unusual, or at a surveillance GP practice. Apart from the fact one would have to be in the 25% of influenza patients with a symptomatic infection, there is likely a bias towards testing in people who are sicker and have underlying conditions that would qualify them for oseltamivir or other antiviral treatments.

Table S27 UK influenza vaccination categories and estimated relative risk of death with a laboratory confirmed influenza infection published by Public Health England

Risk factors	Definition	Age adjusted relative risk of death (95% CI)*
No risk factors	No diagnoses	Baseline
Renal	CKD3 or higher, nephrotic syndrome, transplant	18.5 (NA)
CVD and heart disease	Congenital, failure, IHD with medications, hypertension with CVD complications	10.7 (7.3-15.7)
Respiratory	Asthma with regular meds, COPD, CF, bronchiectasis, bronchopulmonary dysplasia, interstitial, pneumoconiosis, children with prior LRT admissions	7.4 (5.5-10.0)
Chronic liver disease	Cirrhosis, biliary atresia, hepatitis	48.2 (32.8-70.6)
Diabetes	Type 1 or 2 including diet controlled	5.8 (3.8-8.9)
Immunosuppression	Chemo, BMT, HIV, steroids > 1 month at >20mg/day, immune system disorders,	47.3 (35.5-63.1)
Neurological disease	Stroke/TIA, respiratory compromise, CP, LD, MS, hereditary or degenerative neuromuscular disease/disability	40.4 (28.7-56.8)
Splenic disease	Any including coeliac and sickle cell	NA
Pregnancy	All durations	NA
Obesity	BMI>40 kg/m^2	NA
One or more risk factors	Any	11.3 (9.1-14.0)

*Patients 6 months to 65 years old with laboratory confirmed influenza in the 2010-11 influenza season, when the predominant strain was A H1N1. Derived from 352 deaths. Data collated from Public Health England. Influenza: the green book, chapter 19. London: 2019.

9.2.6 Variable ranking

After combining and ranking of the lists of potential variables by GPs with a special interest in cardiology, the most highly ranked variable was age, followed by heart failure, diabetes and smoking status (Table S28: Clinical variables considered for inclusion in models, and rankings by clinical experts).

Table S28: Clinical variables considered for inclusion in models, and rankings by clinical experts

Clinical variable*	Time frame**	Standardised z score mean***
Age	At day 0	1.9
Heart Failure	ever	1.6
Diabetes	ever	1.6
Smoking status	5 year/latest	1.6
Renal disease (CKD 3 and above)	ever	1.3
Hypertension history	ever	1.3
Peripheral Vascular disease	ever	1.2
COPD	ever	1.1
Blood pressure - Systolic	5 year/latest	1.1
Sex	ever	1.0
Blood pressure - Diastolic	5 year/latest	0.9
Cholesterol (Total:HDL ratio)	ever/latest	0.9
BMI	5 year/latest	0.8
Atrial Arrhythmia (Fibrillation/flutter)	ever	0.8
Dementia	ever	0.7
Anticoagulants	3 months	0.6
NSAIDS	6 months	0.6
Obstructive sleep apnoea	5 year/latest	0.6
Antiplatelets	6 months	0.6
Antihypertensive treatment	6 months	0.5
Rheumatoid arthritis	ever	0.5
Statin treatment	6 months	0.5
FBC - Platelets	6 months	0.5
CRP	6 months	0.4
Erectile dysfunction	5 year/latest	0.4
Learning disability	ever	0.4
Other Chronic or congenital heart disease	ever	0.3
Index of multiple deprivation score	5 year/latest	0.3

Cancer - haematological	5 year latest	0.2
Cancer - non haematological	5 year latest	0.2
Family history of CVD age< 60	ever	0.2
Pneumococcus vaccine	10 years	0.1
Glucose lowering medications	6 months	0.0
Chemotherapy last 6 months	6 months	0.0
Haemophilus vaccine	ever	-0.1
Influenza vaccine	6 months	-0.1
FBC - WCC	6 months	-0.1
Full blood count - Haemoglobin	6 months	-0.2
SLE (Lupus)	ever	-0.2
Asthma - active	5 years	-0.2
Antibiotics on day of consultation	day 0	-0.2
Splenic dysfunction	ever	-0.2
Immunosuppression	6 months	-0.2
Conective tissue disorders	ever	-0.2
Ethnicity	ever	-0.2
Prior pneumonia	excluded	-0.3
Alcohol moderate or heavy (>3 units/day)	5 year/latest	-0.4
Other chronic lung disease (ex asthma/ COPD)	ever	-0.5
Chronic liver disease	ever	-0.5
Migraine	ever	-0.5
Coeliac disease	ever	-0.6
Other bleeding drugs	6 months	-0.7
Recent cellulitis	5 year/latest	-0.7
Chronic neurological conditions (except seizure)	ever	-1.2
Hepatitis C infection	ever	-1.2
Seizure disorder	ever	-1.2
UTI (number in last 5 years)	5 year/latest	-1.4

*Clinical variables presented to four GPs with a special interest in cardiology to prioritize for inclusion in a prediction model. ** Time frame is the clinical record prior to the index date searched for codes used to define the variables. *** mean of the four clinician's Z scores for this variable. Z scores calculated as Z = (x-mean)/sd where x is the score for the variable from a clinician, mean is the mean score that clinician gave the variables, and sd is the standard deviation of the scores that clinician gave.

9.2.7 Model specification

9.2.7.1 Model One

Variable transformations:

IAge__1 = Age-56.70140054.

Model:

logit (p) = intercept + (IAge__1 * .065120631558634) + (Ex-smoker * .216435085481023) + (Smoking <10 per day * .521069474752047) + (Smoking 11 to <20 per day * .6212762224915964) + (Smoking 20 or more per day * .6402038495993185) + (Smoker unknown amount * .479095173673624) + (Heart failure * .6504444223129626) + (Diabetes * .4020472246812231) + (Peripheral vascular disease/erectile dysfunction/CKD * .1866634297003109) + (Lower respiratory tract infection * .9504990322745389) + (Influenza * -.1766913795845679) + (Pneumonia * 2.360004208451382)

9.2.7.2 Model Two

Variable transformations:

IAge__1 = Age-56.4915727

IBMI__1 = X^.5-1.660968746 (where: X = BMI/10)

IBMI__2 = X^2-7.611072204 (where: X = BMI/10)

ISyst__1 = X^.5-1.143992165 (where: X = Systolic blood pressure/100)

ISyst__2 = X^.5*ln(X)-.3077889054 (where: X = Systolic blood pressure /100)

Ichol__1 = choloverHDL-3.928200556

Model:

logit (p) = intercept + (IAge__1 * .0624268656055371) + (Male * .3602814828580722) + (IBMI__1 * -2.11821951601376) + (IBMI__2 * .079928616816893) + (ISyst__1 * -

12.51473706712241) + (ISyst__2 * 6.00780703111644) + (Ichol__1 * .0994292795556542) + (Ex-smoker * .1285313176342534) + (Smoking <10 per day * .3903047381201049) + (Smoking 11 to <20 per day * .4053342599263999) + (Smoking 20 or more per day * .4722622955322949) + (Smoker unknown amount * .3665724928985887) + (Heart failure * .5156211258528431) + (Diabetes * .277791542462455) + (Peripheral vascular disease/erectile dysfunction/CKD * .0402272711850079) + (Lower respiratory tract infection * .9163506575774839) + (Influenza * -.1650086994783002) + (Pneumonia * 2.284118276574413) + (Chronic heart disease * .2620935008602651) + (Atrial arrhythmias * .1777328854785878) + (Anticoagulated * .1194084995233393) + (Antihypertensives * .2593188135199826) + (Antiplatelets * .7495876582572326) + (Rheumatoid Arthritis * .1847157658361301) + (Statins * .1440367605302275) + (Family history of CVD * .2495579573416379) + (Cancer * -.0752826901643886) + (IMD_decile2 * .0854324013423883) + (IMD_decile3 * .0499679427695114) + (IMD_decile4 * .0824179242752817) + (IMD_decile5 * .1562734961824674) + (IMD_decile6 * .178860458895174) + (IMD_decile7 * .1919099331725266) + (IMD_decile8 * .3073143511672598) + (IMD_decile9 * .3131420516485625) + (IMD_decile10 * .4202440161178939) + (CRP<5 * -.1564572668463038) + (CRP>=5 & CRP<20 * .0243215508572336) + (CRP >=20 * .2480620256490901) + (Platelets <150 * .2041245574197942) + (Platelets >=150 & <450 * .0998278924632811) + (Platelets >=450 * .2784685782020367)

I presented the equations for the models without the intercepts. This is to retain the intellectual property, to allow the possibility of implementation in clinical software in future.

10 Appendix – Supplementary materials for chapter five

10.3 Objective one supplementary methods

10.3.1 Identifying groups of variables for candidate propensity models

To allocate covariates to different groups I used logistic regression to regress each potential confounder against the outcome, adjusted for age and sex, and then regressed the potential confounder against the exposure, again with adjustment for age and sex. To compare the continuous variables with the categorical variables I obtained an estimate of the association standardised over two standard deviations of the continuous variable. I then took the ratio of the point estimates for association with outcome and exposure. I used these results to group the variables into those more strongly associated with the exposure or the primary outcome.[121]

10.3.2 Predicting the probabilities (propensity) for each individual for each candidate model

To allow for nonlinearity in the continuous variables I first converted them to categorical variables, each with ten values. I generated the candidate propensity models using all the imputed datasets to obtain a model, using the Stata command *mi estimate*. I then estimated the predicted probabilities across the imputed datasets and applied the predicted probability in the original dataset.

10.3.3 Displaying the distribution of propensity probabilities and overlap between groups

To ensure there were not extreme values I tabulated the ranges of predicted probabilities by exposure status. To ensure there were adequate regions of overlap between exposure groups I plotted histograms of the predicted probabilities by exposure status.

10.3.4 Examining evidence of violation of the positivity assumption, and interaction

I split the predicted probabilities into quartiles and examined the numbers of events in each quartile to ensure there were sufficient events at lower probabilities. To look for evidence of interaction I preformed unweighted regression in each quartile, adjusted for propensity but not for imbalanced covariates.[192]

10.3.5 Apparent (internal) C statistic

I used the same methods as described in chapter three (3.4.10.4) to estimate internal concordance statistics for each candidate model. Because this is not achievable as a single process in large datasets I split the dataset randomly, performing the operation which evaluated pairwise comparisons on each subset of the dataset, and meta-analysed the results.

10.3.6 Apparent calibration plots

For each candidate propensity model, I used predicted and observed probabilities of exposure to aspirin to generate apparent calibration plots, using the same methods as for internal calibration (section 4.4.6).

10.3.7 Calculating truncated inverse probability weights and examining distribution of weights

I used the predicted probability of exposure to generate stabilized truncated weights for each candidate model.[139,201,232] I did this to avoid extreme weights, because this is of particular concern for rare exposures.[232] I calculated these for each individual as: the proportion exposed/p for the exposed group and proportion unexposed/(1-p) for the unexposed group (where p = propensity model predicted probability).[232] I then truncated these weights at the 1st and 99th centile of probability, allocating the values of these centiles to individuals with predictions above or below these values.[232] I tabulated the range of weights to ensure weights

were not extreme, and calculated the mean weight to check it was close to one. I also examined the ranges of the weights by exposure status.

10.3.8 Examining covariate balance before and after weighting

For each candidate model I examined the covariate balance in the raw confounders, and the balance achieved in the pseudopopulation by inverse probability weighting.[139] I did this by coding an implementation of the formulae published in Austin 2015.[139] The percentages used were calculated as 100n/N where N includes those with missing values, and so the percentages may differ from those in table one. I considered variables with a standardized mean difference of 0.1 or less to be balanced and identified less well-balanced variables to be added as additional covariates in the final analyses.[192]

10.3.9 Assessment of the crude effect of aspirin on the negative controls

I checked that the propensity models were not introducing biases by estimating the effect of aspirin on the negative controls (skin lesions and constipation) occurring within the 28-day follow-up period. I did this with crude regression adjustment, without adjustment for imbalanced covariates.[104]

10.4 Objective one supplementary results

10.4.1 Evidence of positivity

All the models had patients prescribed aspirin in the lower quartiles of predicted probabilities. There were over three thousand in the lowest quartile of predictions from propensity models one and two, of whom about 50 had CVD events, but only around 1,200 in models three and four, and 33 events in those prescribed aspirin (Table S29: Numbers of events in strata of propensity probabilities by exposure status).

Table S29: Numbers of events in strata of propensity probabilities by exposure status

Propensity model	CVD events and Aspirin status			
Quartile of propensity	No CVD event		CVD event	
Model One	No aspirin	Aspirin	No aspirin	Aspirin
Quartile 1	212,290	3,495	711	55
Quartile 2	195,213	5,918	922	71
Quartile 3	198,533	8,074	2,400	178
Quartile 4	195,924	10,743	2,395	222
Model Two				
Quartile 1	205,405	3,154	679	48
Quartile 2	202,157	5,722	1,315	92
Quartile 3	199,142	8,084	1,922	138
Quartile 4	195,256	11,270	2,512	248
Model Three				
Quartile 1	207,246	1,238	769	33
Quartile 2	203,991	3,628	1,588	79
Quartile 3	200,660	6,341	2,149	136
Quartile 4	190,063	17,023	1,922	278
Model Four				
Quartile 1	207,180	1,274	799	33
Quartile 2	204,070	3,540	1,602	74
Quartile 3	200,730	6,318	2,100	138
Quartile 4	189,980	17,098	1,927	281

10.4.2 Interaction

All four propensity models showed evidence of interaction, with higher effect estimates in the lower quartiles (Table S30 Effect of aspirin in each quartile of four propensity models, linear and logistic regression adjusted for propensity to be prescribed aspirin). In the people least likely to be prescribed aspirin (according to the models) the odds ratio for CVD events was

between 4.7 (95% CI 3.6 to 6.2) in Model Two, and 7.2 (95% CI 5.0 to 10.2) in Model Three, but the crude odds ratio in the highest propensity quartile was lower (1.7 (95% CI 1.5 to 2.0) with Model Two and 1.6 (95% CI 1.4 to 1.8) for Model Three).

Table S30 Effect of aspirin in each quartile of four propensity models, linear and logistic regression adjusted for propensity to be prescribed aspirin

Effect of aspirin in each quartile of four propensity scores, linear and logistic regression

	Linear risk diff	95% Confidence interval	Odds Ratio	95% Confidence interval
Model 1				
Quartile one	0.012	0.010 to 0.014	4.699	3.564 to 6.194
Quartile two	0.007	0.005 to 0.009	2.540	1.993 to 3.238
Quartile three	0.010	0.007 to 0.012	1.824	1.564 to 2.127
Quartile four	0.012	0.012 to 0.013	1.690	1.471 to 1.942
Model 2				
Quartile one	0.012	0.010 to 0.014	4.604	3.428 to 6.182
Quartile two	0.009	0.007 to 0.011	2.472	1.998 to 3.058
Quartile three	0.007	0.005 to 0.009	1.769	1.486 to 2.105
Quartile four	0.009	0.007 to 0.011	1.710	1.499 to 1.951
Model 3				
Quartile one	0.022	0.019 to 0.026	7.184	5.048 to 10.224
Quartile two	0.014	0.011 to 0.016	2.797	2.226 to 3.515
Quartile three	0.010	0.008 to 0.013	2.003	1.681 to 2.386
Quartile four	0.006	0.004 to 0.008	1.615	1.423 to 1.833
Model 4				
Quartile one	0.021	0.018 to 0.025	6.717	4.721 to 9.555
Quartile two	0.013	0.010 to 0.016	2.663	2.104 to 3.370
Quartile three	0.011	0.008 to 0.014	2.088	1.754 to 2.485
Quartile four	0.006	0.005 to 0.008	1.620	1.428 to 1.838

10.4.3 Apparent (internal) concordance statistic

The candidate models C statistics were between 0.62 and 0.76 (Table S31: Discrimination of propensity models, internal apparent estimates)

Table S31: Discrimination of propensity models, internal apparent estimates

Propensity model	Concordance statistic (95% confidence interval)
Model 1	0.621 (0.618 to 0.625)
Model 2	0.636 (0.632 to 0.639)
Model 3	0.753 (0.750 to 0.755)
Model 4	0.755 (0.752 to 0.757)

10.4.4 Candidate propensity model weight distributions

None of the propensity models generated weights with very extreme values, Model Four had the largest range (0.104 to 1.161 see Table S32: Propensity model comparison: weight ranges by exposure status).

Table S32: Propensity model comparison: weight ranges by exposure status

Truncated stabilised weight ranges by exposure status

	Model One		Model Two		Model Three		Model Four	
Aspirin	Min	Max	Min	Max	Min	Max	Min	Max
Unexposed	.970	1.043	.969	1.062	.967	1.157	.967	1.161
Exposed	.365	1.043	.312	1.062	.107	1.157	.104	1.161

10.4.5 Truncated inverse probability weights distributions

The weights for each model had mean near one, without extreme weights (Table S33: Truncated stabilised weights distribution).

Table S33: Truncated stabilised weights distribution

Truncated stabilised weights distributions

Weights	N	Mean	Std. Dev.	Min	Max
Model One weights	837,144	.995	.044	.365	1.043
Model Two weights	837,144	.995	.049	.312	1.062
Model Three weights	837,144	.990	.091	.107	1.157
Model Four weights	837,144	.990	.091	.104	1.161

10.4.6 Assessment of the crude effect of aspirin on the negative controls

None of the candidate propensity models induced associations between aspirin and negative controls in the model building process (Table S34: Association between aspirin and negative controls).

Table S34: Association between aspirin and negative controls

	Skin lesions		Constipation	
Propensity model	Linear risk difference	95% Confidence interval	Linear risk difference	95% Confidence interval
Raw data	0.00	-0.001 to 0.001	0.00	-0.001 to 0.001
Model One	0.00	-0.001 to 0.001	0.00	-0.001 to 0.001
Model Two	0.00	-0.001 to 0.001	0.00	-0.001 to 0.001
Model Three	0.00	-0.002 to 0.001	0.00	-0.001 to 0.001
Model Four	0.00	-0.002 to 0.001	0.00	0.000 to 0.004

Risk difference estimates between aspirin and negative controls, linear regressions in unweighted data, adjusted for propensity estimates derived from each model

11 Appendix – Supplementary materials for chapter six

11.3.1 Propensity models – identifying groups of variables

To allocate covariates to different groups I used logistic regression to regress each potential confounder against the outcome, adjusted for age and sex, and then regressed the potential confounder against the exposure, again with adjustment for age and sex. To make the continuous variables comparable with the categorical variables I examined the association over two standard deviations of the continuous variable. I then took the ratio of these point estimates. I used these results to group the variables into those more strongly associated with the exposure or the primary outcome in the raw data.

11.3.2 Assessment of evidence for interaction

As part of the model building process, for each of the four candidate models I subdivided the predicted probabilities of being prescribed aspirin into quartiles and performed regression of CVD against statin exposure status adjusted for propensity score in each quartile.

11.3.3 Propensity model assessment

I assessed the four potential propensity models by examining the distribution and overlap of their estimates and the apparent calibration of the predictions. I examined the distribution of weights to ensure extreme weights were not present and the mean was close to one for stabilised inverse probability weights (this property is not required of non-stabilised weights).[143] I examined the ranges of the weights by exposure status.

I regressed the outcome against the exposure in quartiles of propensity probability to examine for interaction between the propensity models and the effect. I examined the numbers of outcomes and exposures in these quartiles to assess the positivity assumption (that there should be nonzero probability of being allocated to each exposure at every level of the confounders).[135,201]

11.3.4 Weighting assessment

I examined the covariate balance of confounders achieved by weighting under the different weight types. To do this I calculated weighted standardised mean difference for each variable and considered a difference of less than 0.1 as acceptable covariate balance.[139] I did this in the raw data, without imputation. The percentages used were calculated as 100n/N where N includes those with missing values, and so the percentages may differ from those in table one.

11.3.5 Implementation of overlap weighting with multiple imputation

I developed code to estimate the use of overlap weighting in the context of multiple imputation based on an algorithm published by Felix Bittmann (Table S35: Algorithm for bootstrapping confidence intervals with multiple imputation).[145] The challenge was that bootstrapping in multiple imputation requires the selected individuals to be selected in all the imputed datatsets and remain identifiable.

Table S35: Algorithm for bootstrapping confidence intervals with multiple imputation

This procedure starts with an imputed dataset, mine had 20 imputations (m=20 in Stata terms).

1. A new identifying variable is generated
2. A single bootstrap sample is generated, using both identifying variables to ensure patients are not moved from one imputation to another, and if selected are represented in each of the 20 imputed sets

In each of the 20 imputations

3. Fit the logistic regression propensity model with the variables already decided on
4. Use the propensity model to predict the log probability of each individual being prescribed a statin

Then, for each individual

5. Obtain the mean log probability of allocation over the 20 imputations
6. Exponentiate this to obtain an estimated probability of exposure
7. Use this probability to calculate the desired weights (overlap, inverse probability etc.)

In the base dataset (i.e. m=0, without imputations):

8. Perform logistic regression to predict an outcome, given exposure status, weighted by the weights calculated in step 7
9. Replace the actual exposure status with being unexposed for everyone
10. Use the model parameters from step 8 to predict the outcome for everyone
11. Obtain the mean – this is the potential outcome mean in the unexposed (POM_0)
12. Replace the actual exposure with exposed for everyone
13. Use the model parameters from step 8 to predict the outcome for everyone
14. Obtain the mean – this is the potential outcome mean in the exposed (POM_1)
15. Calculate the treatment effect for each individual (POM_1-POM_0), and obtain the population mean – the average treatment effect in the population (ATE or ACE Average Causal Effect)

Repeat steps 2-15 the desired number of times (I used 500)

16. The mean, 2.5, and 97.5 centiles are the point estimate and confidence intervals for the estimands

Based on methods published by Felix Bittman[145]

11.3.6 Sensitivity analyses

On order to use computationally efficient methods I ran sensitivity analyses using propensity adjustment to logistic regression in the raw dataset using complete case analyses, without weighting or multiple imputation.[192] I used the regression models to estimate the risk difference as the difference between predicted POM_0 and POM_1. I re-ran the analyses with an interaction term between exposure and propensity, and used bootstrapping, replicating fitting

the most complex propensity model and bootstrapping the results 500 times to estimate confidence intervals. I repeated this analysis with different DASHI thresholds for inclusion.

11.4 Objective one supplementary results for chapter six

11.4.1 Evidence regarding positivity assumption

All the models had some patients in the lowest propensity quartiles with statins and CVD events. However, the numbers were low in this quartile, with only single digit numbers of patients starting statins before having infection-related CVD events in all the models (Table S36: Numbers of events in four strata of propensity, by propensity model, and exposure status).

Table S36: Numbers of events in four strata of propensity, by propensity model, and exposure status

Propensity model	Number of people with CVD events and statin status			
Quartiles of propensity	No CVD event		CVD event	
Model One	No statin	Statin	No statin	Statin
Quartile 1	620,107	1,037	187	3
Quartile 2	603,290	2,088	2,163	20
Quartile 3	608,105	3,359	2,652	34
Quartile 4	604,516	6,936	2,800	53
Model Two				
Quartile 1	613,085	882	399	5
Quartile 2	610,371	2,026	1,927	16
Quartile 3	608,381	3,281	2,607	35
Quartile 4	604,181	7,231	2,869	54
Model Three				
Quartile 1	613,553	156	626	3
Quartile 2	611,512	457	2,356	12
Quartile 3	610,415	1,507	2,392	25
Quartile 4	600,538	11,300	2,428	70
Model Four				
Quantile 1	613,585	155	656	4
Quantile 2	611,718	431	2,206	9
Quantile 3	610,267	1,441	2,518	23
Quantile 4	600,448	11,393	2,422	74

11.4.2 Interaction between propensity models and effects of exposure

All four propensity models showed evidence of interaction, with higher crude effect estimates in the lower quartiles (Table S37: Crude estimate of effect of statin therapy in each quartile of four propensity model, linear and logistic regression adjusted for predicted probability).

Table S37: Crude estimate of effect of statin therapy in each quartile of four propensity model, linear and logistic regression adjusted for predicted probability

Effect of statin in each quartile of four propensity models, linear and logistic regression

Propensity model/ Quartile	Linear risk diff	95% Confidence interval	Odds Ratio	95% Confidence interval
Model 1				
Quartile 1	0.003	0.002 to 0.004	9.594	3.06 to 30.06
Quartile 2	0.006	0.003 to 0.008	2.672	1.72 to 4.16
Quartile 3	0.006	0.003 to 0.008	2.321	1.65 to 3.26
Quartile 4	0.003	0.001 to 0.005	1.650	1.26 to 2.17
Model 2				
Quartile 1	0.005	0.003 to 0.007	8.711	3.597 to 21.094
Quartile 2	0.005	0.002 to 0.007	2.501	1.526 to 4.099
Quartile 3	0.006	0.004 to 0.009	2.489	1.780 to 3.481
Quartile 4	0.003	0.001 to 0.004	1.573	1.200 to 2.061
Model 3				
Quartile 1	0.018	0.013 to 0.023	18.848	5.997 to 59.236
Quartile 2	0.022	0.016 to 0.027	6.815	3.837 to 12.107
Quartile 3	0.012	0.009 to 0.016	4.234	2.846 to 6.298
Quartile 4	0.002	0.001 to 0.003	1.532	1.207 to 1.945
Model 4				
Quartile 1	0.024	0.019 to 0.029	24.138	8.920 to 65.319
Quartile 2	0.017	0.011 to 0.022	5.790	2.989 to 11.219
Quartile 3	0.012	0.008 to 0.015	3.868	2.558 to 5.851
Quartile 4	0.002	0.001 to 0.004	1.610	1.277 to 2.031

11.4.3 Apparent (internal) concordance statistics

The candidate propensity models had C statistics between 0.72 and 0.88 (Table S38: Discrimination of propensity models, internal estimates).

Table S38: Discrimination of propensity models, internal estimates

Propensity model	Concordance statistic (95% confidence interval)
Model One	0.719 (0.715 to 0.723)
Model Two	0.734 (0.730 to 0.737)
Model Three	0.872 (0.870 to 0.875)
Model Four	0.876 (0.874 to 0.879)

11.4.4 Weight distributions

The IPW weights all had means close to one (Table S39: Weight distributions by weighting type and propensity model). Truncating the stabilised weights reduced the range of the weights. The range of overlap weights were very close to one and zero (within three decimal places) at the most extreme weights.

Table S39: Weight distributions by weighting type and propensity model

Weight distributions by weighting type, and propensity model

Weighting**	Propensity model*	Mean	Std. Dev.	Min	Max
Stabilised inverse probability	Model One	1.000	0.057	0.075	6.029
	Model Two	1.000	0.061	0.060	9.675
	Model Three	0.999	0.123	0.013	24.181
	Model Four	0.999	0.124	0.011	26.534
Truncated stabilized inverse probability	Model One	0.998	0.028	0.138	1.018
	Model Two	0.998	0.030	0.127	1.019
	Model Three	0.996	0.050	0.032	1.055
	Model Four	0.996	0.050	0.029	1.056
Overlap	Model One	0.011	0.073	0.001	0.999
	Model Two	0.011	0.073	<0.001	0.999
	Model Three	0.011	0.072	<0.001	1.000
	Model Four	0.011	0.072	<0.001	1.000

*Models are sequentially more complex logistic regression models for the propensity to be exposed (p). Model One included demographics only, Model Two also included confounders more strongly associated with CVD outcomes than statin exposure. Model Three included confounders more strongly associated with the exposure as well. Model Four also included recent blood tests for CRP and platelet count. **Weights (w_i) were calculated for each (i^{th}) individual using each models prediction for that individual (p_i). Stabilised inverse probability weights were calculated as w_i = proportion exposed/p_i in the exposed population, and w_i= proportion unexposed/(1-p_i) in the unexposed population. Truncated weights were calculated by identifying the 1st and 99th centile, assigning the value of this centile to those with more extreme w_i. Overlap weights were calculated as w_i = 1-p_i for the exposed, and w_i =p_i for the unexposed. N=2,457,350

The weight ranges were higher for the exposed patients than unexposed patients, under all the models and weighting schemes. The more complex models had broader ranges of weights than the simpler models (Table S40: Weight ranges by weighting type, propensity model, and exposure status).

Table S40: Weight ranges by weighting type, propensity model, and exposure status

Weight ranges by weighting type, propensity model, and exposure status

		Model One*		Model Two*		Model Three*		Model Four*	
Weighting	Statin status	Min	Max	Min	Max	Min	Max	Min	Max
Stabilised IP**	Unexposed	0.995	1.073	0.995	1.129	0.995	1.757	0.995	1.955
	Exposed	0.075	6.029	0.060	9.675	0.013	24.181	0.011	26.534
Truncated Stabilised IP**	Unexposed	0.995	1.018	0.995	1.019	0.995	1.055	0.995	1.056
	Exposed	0.138	1.018	0.127	1.019	0.032	1.055	0.029	1.056
Overlap	Unexposed	0.001	0.073	0.000	0.119	0.000	0.434	0.000	0.491
	Exposed	0.927	0.999	0.909	0.999	0.579	1.000	0.516	1.000

* Models are sequentially more complex logistic regression models for the propensity to be exposed (p). Model One included demographics only, Model Two also included confounders more strongly associated with CVD outcomes than statin exposure. Model Three included confounders more strongly associated with the exposure as well. Model Four also included recent blood tests for CRP and platelet count. ** IP = inverse probability. Weights (w_i) were calculated for each (i^{th}) individual using each models prediction for that individual (p_i) given their covariates. Stabilised inverse probability weights were calculated as w_i = proportion exposed/p_i in the exposed population, and w_i= proportion unexposed/(1-p_i) in the unexposed population. Truncated weights were calculated by identifying the 1st and 99th centile, assigning the value of this centile to those with more extreme w_i. Overlap weights were calculated as w_i = 1-p_i for the exposed, and w_i =p_i for the unexposed.

11.4.5 Covariate balance over candidate propensity models

When the stabilised truncated weights were used to weight the population the more complex models improved the balance of the largest number of covariates between those exposed to statins and those not exposed. However, even with the most complex candidate model, imbalances remained in important variables such as age and diabetes (Table S41: Covariate balance by models with stabilised truncated inverse probability weights).

Table S41: Covariate balance by models with stabilised truncated inverse probability weights

Propensity Model	Raw data			Model One weighted			Model Two weighted			Model Three weighted			Model Four weighted		
Variable and categories	Percent no statin	Percent in statin	Standardised difference	Percent no statin	Percent in statin	Standardised difference	Percent no statin	Percent in statin	Standardised difference	Percent no statin	Percent in statin	Standardised difference	Percent no statin	Percent in statin	Standardised difference
Age															
Decile 1	9.54	3.44	-0.21	9.51	4.47	-0.20	9.51	4.50	-0.20	9.51	4.23	-0.21	9.51	4.28	-0.21
Decile 2	8.51	3.53	-0.18	8.48	4.56	-0.16	8.48	4.60	-0.16	8.49	4.35	-0.17	8.49	4.36	-0.17
Decile 3	9.63	5.06	-0.15	9.61	6.27	-0.12	9.61	6.35	-0.12	9.61	5.59	-0.15	9.61	5.51	-0.16
Decile 4	8.00	5.42	-0.10	7.99	6.30	-0.07	7.99	6.42	-0.06	7.99	5.52	-0.10	7.99	5.49	-0.10
Decile 5	9.38	7.94	-0.05	9.37	9.11	-0.01	9.37	9.13	-0.01	9.37	7.78	-0.06	9.37	7.77	-0.06
Decile 6	8.32	8.95	0.02	8.33	9.57	0.04	8.33	9.42	0.04	8.33	8.37	0.00	8.33	8.35	0.00
Decile 7	12.29	16.27	0.12	12.31	15.60	0.10	12.31	15.31	0.09	12.31	15.17	0.08	12.31	15.09	0.08
Decile 8	12.02	19.29	0.22	12.06	16.39	0.12	12.06	16.26	0.12	12.05	17.27	0.15	12.05	17.35	0.15
Decile 9	10.75	19.42	0.28	10.80	14.94	0.12	10.80	15.18	0.13	10.79	17.45	0.19	10.79	17.68	0.20
Decile 10	11.55	10.68	-0.03	11.54	12.79	0.04	11.54	12.84	0.04	11.55	14.26	0.08	11.55	14.12	0.08
Sex Female	56.85	47.90	-0.18	56.80	52.43	-0.09	56.80	52.43	-0.09	56.81	53.34	-0.07	56.81	53.22	-0.07
Index of multiple deprivation															
Decile 1	11.47	8.20	-0.10	11.17	9.71	-0.05	11.17	9.66	-0.05	11.17	9.57	-0.05	11.17	9.52	-0.05
Decile 2	10.37	8.10	-0.07	10.10	9.21	-0.03	10.10	9.22	-0.03	10.10	9.25	-0.03	10.10	9.31	-0.03
Decile 3	10.40	8.07	-0.08	10.13	9.19	-0.03	10.13	9.19	-0.03	10.13	9.07	-0.04	10.13	9.03	-0.04
Decile 4	10.64	8.52	-0.07	10.37	9.48	-0.03	10.37	9.41	-0.03	10.37	9.28	-0.04	10.37	9.30	-0.04
Decile 5	9.33	8.92	-0.01	9.09	9.05	0.00	9.09	9.10	0.00	9.09	8.89	-0.01	9.09	8.83	-0.01
Decile 6	9.90	10.34	0.01	9.65	9.74	0.00	9.65	9.73	0.00	9.65	9.83	0.01	9.65	9.77	0.00
Decile 7	9.93	10.64	0.02	9.69	9.91	0.01	9.69	9.92	0.01	9.69	10.05	0.01	9.69	10.08	0.01
Decile 8	9.24	12.42	0.11	9.03	10.17	0.04	9.03	10.21	0.04	9.02	10.65	0.05	9.02	10.61	0.05
Decile 9	9.41	13.49	0.14	9.20	10.68	0.05	9.20	10.72	0.05	9.20	10.74	0.05	9.20	10.69	0.05

Decile 10	9.32	11.30	0.07	9.10	9.98	0.03	9.10	9.92	0.03	9.10	9.89	0.03	9.10	9.94	0.03
Ethnicity															
Bangladeshi	0.45	1.69	0.18	0.33	0.57	0.04	0.33	0.60	0.04	0.33	0.62	0.04	0.33	0.63	0.04
Black African	1.61	3.81	0.18	1.17	1.67	0.04	1.17	1.72	0.05	1.17	1.85	0.06	1.17	1.84	0.06
Black Caribbean	1.39	3.19	0.15	1.01	1.61	0.05	1.01	1.64	0.06	1.01	1.85	0.07	1.01	1.86	0.07
Chinese	0.30	0.52	0.04	0.22	0.31	0.02	0.22	0.30	0.02	0.22	0.26	0.01	0.22	0.26	0.01
Indian	2.13	5.68	0.25	1.55	2.55	0.07	1.55	2.57	0.07	1.55	2.96	0.09	1.55	2.94	0.09
Other Asian	1.05	2.86	0.18	0.76	1.21	0.05	0.76	1.25	0.05	0.76	1.35	0.06	0.76	1.33	0.06
Other enthicity	1.81	3.07	0.09	1.31	1.79	0.04	1.31	1.82	0.04	1.31	1.80	0.04	1.31	1.77	0.04
Pakistani	1.14	3.63	0.23	0.83	1.40	0.05	0.83	1.38	0.05	0.83	1.56	0.07	0.83	1.58	0.07
White	90.13	75.56	-0.49	64.98	72.76	0.17	64.98	72.65	0.17	65.02	69.66	0.10	65.02	69.38	0.09
Smoking status															
Never smoked	37.07	40.92	0.08	24.67	35.14	0.23	24.66	36.33	0.26	24.70	34.37	0.21	24.70	34.19	0.21
Ex-smoker	24.84	29.81	0.12	16.53	27.30	0.26	16.54	26.25	0.24	16.56	26.08	0.23	16.56	26.06	0.23
Light smoker	7.00	5.77	-0.05	4.65	5.35	0.03	4.65	5.45	0.04	4.65	4.97	0.01	4.65	4.93	0.01
Moderate smoker	9.29	6.55	-0.09	6.17	6.52	0.01	6.17	6.61	0.02	6.17	5.86	-0.01	6.17	5.80	-0.02
Heavy smoker	6.63	5.08	-0.06	4.40	5.26	0.04	4.41	5.12	0.03	4.41	4.46	0.00	4.41	4.52	0.01
Smoking unknown quantity	15.17	11.88	-0.09	10.08	11.54	0.05	10.08	11.35	0.04	10.09	10.91	0.03	10.09	10.90	0.03
Cholesterol to HDL ratio															
Decile 1	3.07	5.45	0.14	3.07	5.34	0.11	3.07	5.43	0.12	3.09	4.16	0.06	3.09	4.27	0.06
Decile 2	2.65	4.86	0.14	2.65	4.73	0.11	2.65	4.73	0.11	2.66	5.13	0.13	2.66	5.06	0.12
Decile 3	3.03	5.40	0.14	3.03	5.19	0.11	3.03	5.25	0.11	3.04	5.78	0.13	3.04	5.92	0.14
Decile 4	3.11	5.88	0.16	3.11	5.68	0.13	3.11	5.72	0.13	3.12	6.42	0.16	3.12	6.34	0.15
Decile 5	3.17	6.73	0.20	3.18	6.38	0.15	3.18	6.37	0.15	3.19	6.64	0.16	3.19	6.60	0.16
Decile 6	3.17	7.42	0.24	3.17	7.10	0.18	3.18	7.08	0.18	3.19	6.82	0.17	3.19	6.94	0.17
Decile 7	3.19	8.51	0.30	3.20	8.18	0.22	3.20	8.16	0.22	3.22	7.24	0.18	3.22	7.18	0.18
Decile 8	3.31	10.57	0.41	3.31	10.48	0.29	3.31	10.48	0.29	3.34	7.85	0.20	3.34	7.80	0.20
Decile 9	3.03	12.22	0.54	3.03	12.28	0.35	3.03	12.30	0.35	3.07	7.54	0.20	3.07	7.46	0.20
Decile 10	2.83	18.03	0.92	2.83	18.96	0.54	2.84	18.89	0.53	2.89	8.32	0.24	2.89	8.26	0.24
Systolic blood pressure															
Decile 1	9.51	5.70	-0.13	9.50	5.81	-0.14	9.49	7.21	-0.08	9.49	7.15	-0.08	9.49	7.04	-0.09
Decile 2	5.50	5.03	-0.02	5.50	5.03	-0.02	5.50	5.78	0.01	5.50	5.73	0.01	5.50	5.68	0.01
Decile 3	7.64	6.31	-0.05	7.63	6.53	-0.04	7.63	7.23	-0.02	7.63	6.89	-0.03	7.63	6.86	-0.03
Decile 4	6.74	8.18	0.06	6.73	7.91	0.05	6.73	8.29	0.06	6.74	8.01	0.05	6.74	7.99	0.05
Decile 5	10.67	12.25	0.05	10.67	12.36	0.05	10.67	12.67	0.06	10.68	12.39	0.05	10.68	12.32	0.05
Decile 6	4.67	7.58	0.14	4.68	7.47	0.12	4.68	6.94	0.10	4.69	6.55	0.08	4.69	6.56	0.08
Decile 7	13.95	20.06	0.18	13.96	19.95	0.16	13.97	18.78	0.13	13.98	18.72	0.13	13.98	18.87	0.13
Decile 8	2.53	4.34	0.12	2.53	4.18	0.09	2.53	3.86	0.08	2.54	3.94	0.08	2.54	3.97	0.08
Decile 9	8.87	12.92	0.14	8.88	12.88	0.13	8.88	12.48	0.12	8.89	12.14	0.11	8.89	12.16	0.11

Decile 10	9.04	15.22	0.22	9.05	15.27	0.19	9.06	13.94	0.15	9.06	13.68	0.15	9.06	13.71	0.15
Heart Failure	0.93	1.38	0.05	0.93	1.40	0.04	0.93	1.19	0.03	0.93	1.60	0.06	0.93	1.59	0.06
Cancer	3.48	3.64	0.01	3.48	3.71	0.01	3.48	4.06	0.03	3.48	4.45	0.05	3.48	4.51	0.05
Body mass index deciles															
Decile 1	5.44	3.70	-0.08	5.44	3.52	-0.09	5.44	3.67	-0.09	5.44	4.12	-0.06	5.44	4.05	-0.07
Decile 2	5.01	4.67	-0.02	5.01	4.45	-0.03	5.01	4.53	-0.02	5.01	5.12	0.00	5.01	5.08	0.00
Decile 3	4.89	5.63	0.03	4.90	5.35	0.02	4.90	5.41	0.02	4.90	5.55	0.03	4.90	5.52	0.03
Decile 4	4.85	6.28	0.07	4.86	5.83	0.04	4.86	5.93	0.05	4.86	6.07	0.05	4.86	5.97	0.05
Decile 5	4.71	7.58	0.14	4.71	7.09	0.10	4.71	7.16	0.10	4.72	7.04	0.10	4.72	7.08	0.10
Decile 6	4.67	7.39	0.13	4.67	7.18	0.11	4.67	7.17	0.11	4.68	6.70	0.09	4.68	6.67	0.09
Decile 7	4.52	8.82	0.21	4.53	8.57	0.16	4.53	8.60	0.17	4.54	7.65	0.13	4.54	7.59	0.13
Decile 8	4.44	9.75	0.26	4.44	9.62	0.20	4.44	9.46	0.20	4.45	8.25	0.16	4.45	8.23	0.16
Decile 9	4.29	9.79	0.27	4.29	9.84	0.22	4.29	9.71	0.21	4.31	8.17	0.16	4.31	8.17	0.16
Decile 10	4.14	10.61	0.33	4.13	11.61	0.28	4.14	11.41	0.27	4.16	8.35	0.17	4.16	8.35	0.17
Congenital or valve disease	1.12	2.15	0.10	1.12	2.15	0.08	1.12	1.52	0.03	1.12	2.05	0.07	1.12	2.11	0.08
Atherosclerosis	5.71	19.50	0.59	5.72	18.74	0.41	5.72	18.68	0.40	5.77	13.03	0.25	5.77	13.20	0.26
Atrial arrhythmia	1.77	3.07	0.10	1.77	3.13	0.09	1.78	2.40	0.04	1.78	3.52	0.11	1.78	3.48	0.11
Diabetes	5.47	32.94	1.21	5.48	31.83	0.72	5.48	31.97	0.72	5.59	16.15	0.34	5.58	16.37	0.35
Family history of CVD	0.54	1.20	0.09	0.54	1.30	0.08	0.54	1.31	0.08	0.54	1.11	0.06	0.54	1.09	0.06
Platelets															
Platelets unknown	74.33	43.39	-0.71	74.31	44.46	-0.64	74.31	44.42	-0.64	74.23	49.01	-0.54	74.18	56.91	-0.37
Platelets low	0.85	1.34	0.05	0.85	1.18	0.03	0.85	1.20	0.03	0.85	1.14	0.03	0.85	1.41	0.05
Platelets normal	24.24	54.33	0.70	24.25	53.36	0.63	24.25	53.39	0.63	24.33	48.87	0.53	24.38	40.74	0.35
Platelets high	0.59	0.95	0.05	0.59	1.00	0.05	0.59	1.00	0.05	0.59	0.97	0.04	0.59	0.94	0.04
C reactive protein															
CRP unknown	93.54	89.08	-0.18	93.54	89.17	-0.16	93.54	89.11	-0.16	93.52	89.57	-0.14	93.52	89.64	-0.14
CRP <5	3.10	5.38	0.13	3.10	5.29	0.11	3.10	5.34	0.11	3.11	5.13	0.10	3.12	4.78	0.09
CRP 5 to <20	2.57	4.42	0.12	2.57	4.40	0.10	2.58	4.41	0.10	2.58	4.20	0.09	2.58	4.30	0.09
CRP 20+	0.78	1.12	0.04	0.78	1.14	0.04	0.78	1.15	0.04	0.78	1.11	0.03	0.78	1.29	0.05
Rheumatoid arthritis	1.34	1.71	0.03	1.35	1.72	0.03	1.35	1.62	0.02	1.35	1.83	0.04	1.35	1.83	0.04

Weighted standardised mean differences in population weighted with stabilized truncated inverse probability weights from each propensity model. Percentages were calculated over the population, including those with missing values in the denominator, and so may vary from table 1. Covariate balance in pseudopopulations created with overlap weighting. Percentages are calculated as the mean of the binary variable over the population, including those with missing values. Model One is demographics only (age, sex, IMD, ethnicity), Model Two with demographics and further confounders (Smoking, atrial arrythmias, heart failure, rheumatoid arthritis, valvular and congenital heart disease, and cancer diagnoses), Model Three includes further confounders added (Body mass index, total cholesterol to HDL cholesterol ratio, missing indicator for cholesterol measurement, systolic blood pressure, markers of atherosclerosis, family history of early CVD, and diabetes diagnosis), Model Four includes additional recent blood tests (CRP, platelets).

Table S42: Covariate balance by propensity model with overlap weighting

Model	Raw data			Propensity Model One overlap weighted			Propensity Model Two overlap weighted			Propensity Model Three overlap weighted			Propensity Model Four overlap weighted		
Variable and categories	Percent no statin	Percent in statin	Standardised difference	Percent no statin	Percent in statin	Standardised difference	Percent no statin	Percent in statin	Standardised difference	Percent no statin	Percent in statin	Standardised difference	Percent no statin	Percent in statin	Standardised difference
Age															
Decile 1	9.54	3.44	-0.21	3.42	3.46	0.00	3.41	3.46	0.00	3.45	3.47	0.00	3.45	3.47	0.00
Decile 2	8.51	3.53	-0.18	3.52	3.55	0.00	3.51	3.55	0.00	3.55	3.57	0.00	3.55	3.57	0.00
Decile 3	9.63	5.06	-0.15	5.05	5.09	0.00	5.03	5.09	0.00	5.07	5.09	0.00	5.07	5.09	0.00
Decile 4	8.00	5.42	-0.10	5.41	5.43	0.00	5.39	5.43	0.00	5.40	5.42	0.00	5.40	5.42	0.00
Decile 5	9.38	7.94	-0.05	7.93	7.96	0.00	7.92	7.95	0.00	7.91	7.93	0.00	7.91	7.93	0.00
Decile 6	8.32	8.95	0.02	8.95	8.96	0.00	8.94	8.95	0.00	8.88	8.90	0.00	8.88	8.89	0.00
Decile 7	12.29	16.27	0.12	16.28	16.25	0.00	16.29	16.25	0.00	16.24	16.22	0.00	16.23	16.22	0.00
Decile 8	12.02	19.29	0.22	19.30	19.24	0.00	19.33	19.24	0.00	19.29	19.25	0.00	19.28	19.25	0.00
Decile 9	10.75	19.42	0.28	19.40	19.35	0.00	19.44	19.35	0.00	19.39	19.36	0.00	19.39	19.36	0.00
Decile 10	11.55	10.68	-0.03	10.72	10.71	0.00	10.75	10.71	0.00	10.81	10.80	0.00	10.82	10.81	0.00
Sex Female	56.85	47.90	-0.18	48.03	47.97	0.00	48.06	47.97	0.00	48.22	48.15	0.00	48.24	48.17	0.00
Index of multiple deprivation															
Decile 1	11.47	8.20	-0.10	8.13	8.01	0.00	8.13	8.01	0.00	8.10	8.06	0.00	8.10	8.06	0.00
Decile 2	10.37	8.10	-0.07	7.99	7.91	0.00	7.99	7.91	0.00	7.97	7.95	0.00	7.97	7.95	0.00
Decile 3	10.40	8.07	-0.08	8.00	7.87	0.00	8.00	7.87	0.00	7.98	7.91	0.00	7.98	7.91	0.00
Decile 4	10.64	8.52	-0.07	8.43	8.31	0.00	8.43	8.31	0.00	8.42	8.35	0.00	8.42	8.35	0.00
Decile 5	9.33	8.92	-0.01	8.76	8.69	0.00	8.76	8.69	0.00	8.76	8.72	0.00	8.76	8.72	0.00
Decile 6	9.90	10.34	0.01	10.11	10.06	0.00	10.11	10.05	0.00	10.08	10.05	0.00	10.08	10.05	0.00
Decile 7	9.93	10.64	0.02	10.37	10.35	0.00	10.37	10.35	0.00	10.37	10.34	0.00	10.37	10.34	0.00
Decile 8	9.24	12.42	0.11	11.99	12.05	0.00	11.99	12.05	0.00	11.98	11.99	0.00	11.97	11.99	0.00
Decile 9	9.41	13.49	0.14	13.02	13.08	0.00	13.02	13.08	0.00	12.97	12.99	0.00	12.97	12.99	0.00
Decile 10	9.32	11.30	0.07	10.98	10.98		10.98	10.98	0.00	10.96	10.94	0.00	10.96	10.94	0.00
Ethnicity															
Bangladeshi	0.45	1.69	0.18	1.24	1.41	0.01	1.24	1.41	0.01	1.35	1.37	0.00	1.35	1.37	0.00
Black African	1.61	3.81	0.18	2.65	3.20	0.03	2.65	3.20	0.03	2.95	3.15	0.01	2.96	3.15	0.01

Black Caribbean	1.39	3.19	0.15	2.20	2.69	0.03	2.19	2.69	0.03	2.44	2.67	0.01	2.45	2.67	0.01
Chinese	0.30	0.52	0.04	0.36	0.44	0.01	0.36	0.44	0.01	0.40	0.44	0.00	0.40	0.44	0.00
Indian	2.13	5.68	0.25	4.04	4.78	0.04	4.04	4.78	0.04	4.49	4.71	0.01	4.49	4.71	0.01
Other Asian	1.05	2.86	0.18	2.03	2.40	0.03	2.03	2.40	0.03	2.27	2.37	0.01	2.27	2.36	0.01
Other enthicity	1.81	3.07	0.09	2.17	2.60	0.03	2.16	2.60	0.03	2.43	2.57	0.01	2.43	2.57	0.01
Pakistani	1.14	3.63	0.23	2.57	3.04	0.03	2.56	3.03	0.03	2.82	2.98	0.01	2.81	2.97	0.01
White	90.13	75.56	-0.49	56.31	64.43	0.17	56.00	64.43	0.17	62.85	64.66	0.04	63.03	64.66	0.03
Smoking status															
Never smoked	37.07	40.92	0.08	29.21	37.49	0.18	27.65	37.51	0.21	35.52	37.42	0.04	35.66	37.42	0.04
Ex-smoker	24.84	29.81	0.12	19.34	27.35	0.19	20.66	27.33	0.16	26.25	27.34	0.02	26.36	27.35	0.02
Light smoker	7.00	5.77	-0.05	4.39	5.29	0.04	4.33	5.30	0.05	5.00	5.28	0.01	5.01	5.27	0.01
Moderate smoker	9.29	6.55	-0.09	5.11	6.02	0.04	4.88	6.02	0.05	5.63	6.01	0.02	5.64	6.01	0.02
Heavy smoker	6.63	5.08	-0.06	3.61	4.67	0.05	3.78	4.67	0.04	4.33	4.65	0.02	4.35	4.65	0.01
Smoking unknown quantity	15.17	11.88	-0.09	8.75	10.91	0.07	8.84	10.90	0.07	10.44	10.91	0.02	10.46	10.90	0.01
Cholesterol to HDL ratio															
Decile 1	3.07	5.45	0.14	3.32	5.45	0.10	3.33	5.45	0.10	7.86	5.44	-0.10	7.79	5.45	-0.09
Decile 2	2.65	4.86	0.14	2.98	4.85	0.10	2.99	4.85	0.10	4.90	4.88	0.00	4.90	4.88	0.00
Decile 3	3.03	5.40	0.14	3.49	5.39	0.10	3.51	5.39	0.09	5.60	5.44	-0.01	5.61	5.45	-0.01
Decile 4	3.11	5.88	0.16	3.68	5.87	0.13	3.72	5.87	0.10	5.99	5.92	0.00	6.00	5.92	0.00
Decile 5	3.17	6.73	0.20	3.83	6.73	0.15	3.89	6.73	0.13	6.74	6.77	0.00	6.75	6.77	0.00
Decile 6	3.17	7.42	0.24	3.89	7.41	0.19	3.96	7.41	0.15	7.46	7.44	0.00	7.47	7.45	0.00
Decile 7	3.19	8.51	0.30	3.92	8.51	0.25	4.01	8.51	0.19	8.39	8.53	0.01	8.40	8.54	0.00
Decile 8	3.31	10.57	0.41	4.07	10.57	0.32	4.19	10.57	0.25	10.27	10.54	0.01	10.28	10.54	0.01
Decile 9	3.03	12.22	0.54	3.70	12.23	0.49	3.81	12.23	0.31	11.64	12.12	0.01	11.63	12.11	0.01
Decile 10	2.83	18.03	0.92	3.29	18.04	0.06	3.41	18.04	0.49	16.37	17.58	0.03	16.35	17.55	0.03
Missing indicator	69.44	14.93	-1.18	63.83	14.94	-1.16	63.17	14.94	-1.14	14.78	15.34	0.02	14.80	15.36	0.02
Systolic BP															
Decile 1	9.51	5.70	-0.13	7.76	5.70	-0.08	4.30	5.72	0.07	5.50	5.75	0.01	5.54	5.75	0.01
Decile 2	5.50	5.03	-0.02	4.98	5.03	0.00	3.89	5.04	0.06	4.92	5.07	0.01	4.94	5.07	0.01
Decile 3	7.64	6.31	-0.05	6.98	6.31	-0.03	5.90	6.33	0.02	6.27	6.34	0.00	6.27	6.33	0.00
Decile 4	6.74	8.18	0.06	6.66	8.18	0.06	6.50	8.18	0.06	7.94	8.18	0.01	7.96	8.18	0.01
Decile 5	10.67	12.25	0.05	10.96	12.26	0.04	11.04	12.26	0.04	12.07	12.27	0.01	12.09	12.27	0.01
Decile 6	4.67	7.58	0.14	5.09	7.57	0.10	6.10	7.56	0.06	7.32	7.54	0.01	7.35	7.55	0.01
Decile 7	13.95	20.06	0.18	15.75	20.06	0.11	18.17	20.04	0.05	19.72	20.02	0.01	19.75	20.02	0.01
Decile 8	2.53	4.34	0.12	2.96	4.34	0.07	3.51	4.33	0.04	4.19	4.33	0.01	4.20	4.33	0.01
Decile 9	8.87	12.92	0.14	10.49	12.92	0.08	11.54	12.91	0.04	12.60	12.90	0.01	12.63	12.90	0.01
Decile 10	9.04	15.22	0.22	11.04	15.22	0.12	13.15	15.20	0.06	14.61	15.14	0.01	14.63	15.13	0.01
Heart Failure	0.93	1.38	0.05	1.07	1.38	0.03	1.38	1.38	0.00	1.38	1.38	0.00	1.38	1.39	0.00

Cancer	3.48	3.64	0.01	4.22	3.64	-0.03	3.66	3.64	0.00	3.68	3.67	0.00	3.68	3.67	0.00
BMI															
Decile 1	5.44	3.70	-0.08	5.56	3.70	-0.09	5.21	3.70	-0.07	5.12	3.72	-0.07	5.17	3.72	-0.07
Decile 2	5.01	4.67	-0.02	5.20	4.67	-0.02	5.02	4.67	-0.02	5.24	4.68	-0.03	5.24	4.68	-0.03
Decile 3	4.89	5.63	0.03	5.33	5.62	0.01	5.23	5.63	0.02	5.80	5.64	-0.01	5.82	5.64	-0.01
Decile 4	4.85	6.28	0.07	5.44	6.27	0.04	5.40	6.27	0.04	6.32	6.27	0.00	6.30	6.27	0.00
Decile 5	4.71	7.58	0.14	5.38	7.56	0.09	5.40	7.57	0.09	6.76	7.56	0.03	6.74	7.56	0.03
Decile 6	4.67	7.39	0.13	5.39	7.39	0.08	5.49	7.39	0.08	7.19	7.37	0.01	7.24	7.36	0.00
Decile 7	4.52	8.82	0.21	5.17	8.82	0.14	5.32	8.82	0.14	7.57	8.77	0.04	7.61	8.77	0.04
Decile 8	4.44	9.75	0.26	4.99	9.75	0.18	5.20	9.75	0.17	8.06	9.70	0.06	8.12	9.70	0.06
Decile 9	4.29	9.79	0.27	4.66	9.80	0.20	4.93	9.79	0.19	8.52	9.73	0.04	8.60	9.73	0.04
Decile 10	4.14	10.61	0.33	4.00	10.63	0.26	4.33	10.63	0.24	8.91	10.54	0.05	9.09	10.53	0.05
Congenital or valve disease	1.12	2.15	0.10	1.24	2.15	0.07	2.15	2.14	0.00	2.15	2.15	0.00	2.16	2.15	0.00
Atherosclerosis	5.71	19.50	0.59	7.47	19.48	0.36	7.65	19.48	0.35	19.22	19.12	0.00	19.22	19.12	0.00
Atrial arrhythmia	1.77	3.07	0.10	2.05	3.07	0.06	3.09	3.06	0.00	3.10	3.08	0.00	3.10	3.08	0.00
Diabetes	5.47	32.94	1.21	7.07	32.91	0.68	7.28	32.91	0.68	32.03	31.92	0.00	32.02	31.90	0.00
Family history	0.54	1.20	0.09	0.55	1.20	0.07	0.55	1.20	0.07	1.21	1.20	0.00	1.21	1.20	0.00
Platelets unknown	74.33	43.39	-0.71	71.80	43.41	-0.60	71.47	43.41	-0.59	54.77	43.54	-0.23	43.62	43.89	0.01
Platelets low	0.85	1.34	0.05	1.11	1.33	0.02	1.11	1.33	0.02	1.91	1.33	-0.05	1.38	1.34	0.00
Platelets normal	24.24	54.33	0.70	26.46	54.31	0.59	26.80	54.31	0.58	42.46	54.19	0.24	54.05	53.82	0.00
Platelets high	0.59	0.95	0.05	0.63	0.95	0.04	0.62	0.95	0.04	0.86	0.95	0.01	0.95	0.95	0.00
CRP unknown	93.54	89.08	-0.18	93.30	89.08	-0.15	93.29	89.08	-0.15	89.17	89.07	0.00	89.00	89.06	0.00
CRP <5	3.10	5.38	0.13	3.11	5.38	0.11	3.09	5.38	0.11	4.84	5.39	0.02	5.41	5.39	0.00
CRP 5 to <20	2.57	4.42	0.12	2.72	4.42	0.09	2.75	4.42	0.09	4.61	4.41	-0.01	4.45	4.43	0.00
CRP 20+	0.78	1.12	0.04	0.88	1.12	0.02	0.87	1.12	0.03	1.37	1.12	-0.02	1.14	1.13	0.00
Rheumatoid Arthritis	1.34	1.71	0.03	1.55	1.71	0.01	1.71	1.71	0.00	1.72	1.71	0.00	1.72	1.72	0.00

Covariate balance in pseudopopulations created with overlap weighting. Percentages are calculated as the mean of the binary variable over the population, including those with missing values. Model One is demographics only (age, sex, IMD, ethnicity), Model Two with demographics and further confounders (Smoking, atrial arrythmias, heart failure, rheumatoid arthritis, valvular and congenital heart disease, and cancer diagnoses), Model Three includes further confounders added (Body mass index, total cholesterol to HDL cholesterol ratio, missing indicator for cholesterol measurement, systolic blood pressure, markers of atherosclerosis, family history of early CVD, and diabetes diagnosis), Model Four includes additional recent blood tests (CRP, platelets).

11.4.6 Negative control assessment

The negative controls, skin lesions and constipation, did not have associations with statins in the raw data. Regression with adjustment for the propensity models did not induce an association between statin prescriptions and the negative controls for each of the propensity models (Table S43: Association between statin use and negative controls).

Table S43: Association between statin use and negative controls

	Skin lesions		Constipation	
Propensity model adjustment	Linear risk diff	95% Confidence interval	Linear risk diff	95% Confidence interval
Raw data				
Model One	0.001	-0.001 to 0.002	0.000	-0.001 to 0.001
Model Two	0.001	-0.001 to 0.002	0.000	-0.001 to 0.001
Model Three	0.000	-0.001 to 0.002	0.000	-0.001 to 0.001
Model Four	0.000	-0.002 to 0.001	0.000	-0.001 to 0.001

Risk difference estimates between statin and negative controls, linear regressions in raw data and adjusted for propensity models

12 Appendix – Publication based on chapters three and four