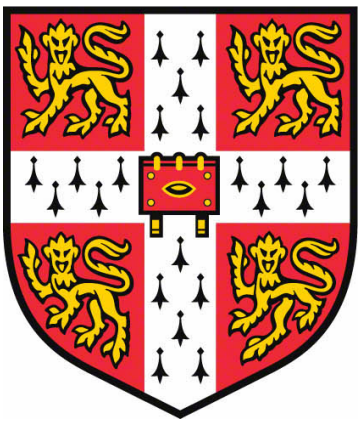
Personalising Predictive Prevention of Cardiovascular Disease using Electronic Health Records and Genomics



Ryan Kai-Yin Chung

Downing College

Department of Public Health and Primary Care

University of Cambridge

This thesis is submitted for the degree of

Doctor of Philosophy

Supervisors: Professor Angela M. Wood and Dr Juliet A. Usher-Smith

July 2023

Declarations

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text. It does not exceed the prescribed 60,000-word limit for the Faculties of Clinical Medicine and Veterinary Medicine Degree Committee.

Chapter 1

I drafted the chapter and produced all tables and figures, unless otherwise stated. Professor Angela Wood, Dr Juliet Usher-Smith provided helpful feedback.

Chapter 2

I performed the main statistical analysis and wrote the first draft of the chapter, producing all tables and figures. I performed the data management based on the UK Biobank dataset provided by Dr Matt Arnold. I produced all R code used for the modelling. Professor Angela Wood, Dr Juliet Usher-Smith provided helpful feedback.

Chapter 3

I performed the data management based on the UK Biobank dataset provided by Dr Matt Arnold and performed the population health modelling within the main analysis of the chapter. I produced the R programming code for analysis and wrote the eHEART R script shared on GitHub. Dr Ellie Paige, Dr Zhe Xu, Dr Lisa Pennells, Dr Matt Arnold and Professor Ruth Keogh derived the eHEART model. Dr Zhe Xu prepared all tables and figures using the CPRD dataset (Tables 3.1-3.5, Figures 3.5-3.16). I prepared all remaining tables and figures relating to the population health modelling, and wrote the first draft of the chapter. Prof Angela Wood, Dr Lois Kim and Dr Juliet Usher-Smith provided helpful comments during the writing of the chapter.

This chapter was prepared as a manuscript and is currently under review, of which I am the first author. Dr Hannah Harrison, Dr Juliet Usher-Smith, Dr Zhe Xu, Dr Matt Arnold, Dr Jessica Barrett and Dr Lois Kim reviewed the manuscript and provided detailed feedback.

Chapter 4

I developed the analysis plan, performed the data management based on the UK Biobank dataset provided by Dr Matt Arnold, performed all statistical analyses, prepared all tables and figures and wrote the first draft of the chapter. I produced the R programming code for analysis. Professor Mike Inouye and Dr Scott Ritchie provided expertise on the polygenic risk scores and Dr Zhe Xu assisted with summary level data from CPRD. Professor Angela Wood, Dr Lois Kim and Dr Juliet Usher-Smith provided helpful comments during the writing of the chapter.
This chapter was prepared as a manuscript and is currently under review, of which I am the first author. Dr Hannah Harrison, Dr Juliet Usher-Smith, Dr Zhe Xu, Dr Scott Ritchie, Professor Mike Inouye, Dr Jessica Barrett and Dr Lois Kim reviewed the manuscript and provided detailed feedback.

Chapter 5

Dr Lisa Pennells and I developed the analysis plan. I performed the data management, performed all statistical analysis, prepared all tables and figures and wrote the first draft of the chapter. Dr Zhe Xu assisted with summary level data from CPRD. Professor Angela Wood, Dr Lisa Pennells, Dr Lois Kim and Dr Juliet Usher-Smith provided helpful comments during the writing of the chapter.

Chapter 6

I drafted the chapter, and Professor Angela Wood, Dr Juliet Usher-Smith provided helpful feedback.

Appendix

Dr Zhe Xu compiled the code lists and drafted the landmark model methods and mixed-effects regression models, and I drafted the statistical methods used to recalibrate and rescale eHEART for the population health modelling.

Summary

Personalising Predictive Prevention of Cardiovascular Disease using Electronic Health Records and Genomics

Ryan Chung

Cardiovascular diseases (CVD) remain one of the leading causes of morbidity and mortality in the world. The development of CVD risk prediction models has been pivotal in helping identify high risk individuals who may benefit the most from appropriate treatment. In England, clinical guidelines provide recommendations of the appropriate care and treatments needed to manage CVD. In particular, the guidelines recommend using a CVD risk prediction model during a full formal risk assessment for all individuals between 40 and 75 years. To manage health resources, the guidelines also recommend systematically prioritising individuals for risk assessments using historical information already recorded in primary care records.

However, there are limitations of existing guidelines. **First**, a dedicated risk model designed for prioritisation to risk assessments does not exist, or is currently recommended for use in current primary care systems. In addition, it is unknown how implementing a fixed risk threshold for prioritisation would affect the effectiveness of formal CVD risk assessments. Therefore, the **first aim** of this thesis is to develop a novel prioritisation model and evaluate its public health impact, by comparing a fixed risk threshold against age- and sex- specific risk thresholds to determine whether individuals would be deemed at high risk. **Second,** the majority of research of genetic data, genomics, has focussed on improving risk model performance using genetic-based risk factors called polygenic risk scores (PRS). However, little is known as to whether PRS will benefit CVD prioritisation. Therefore, the **second aim** of this thesis is to investigate the potential benefits of PRS when used for both prioritisation and formal assessments. **Third,** a future healthcare system that incorporates widespread genetic profiling has the ability to personalise preventative medicine. Therefore, the **third aim** is to investigate and estimate the lifetime impact of novel PRS-based personalised invitation strategies.

Key finding 1: By utilising all available primary care records in a large, national database, a novel prioritisation model, eHEART, was developed. We showed that prioritisation, in addition to using optimised age-and sex specific risk thresholds, can be used to make formal CVD risk assessments more efficient. For example, a formal CVD risk assessment on all adults would identify 76% and 49% of future CVD events amongst men and women respectively. However, prioritisation with eHEART could identify 73% and 47% of future events amongst men and

women respectively, with a 19% and 42% reduction in the number needed to screen to prevent one CVD event respectively. The results suggest that optimising the risk thresholds used can lead to a more efficient CVD risk assessment programme, with the biggest improvements amongst younger individuals.

Key finding 2: To understand how PRS could improve CVD risk prioritisation, a comparison of how prioritisation differed when using either only primary care records, age and PRS, or primary care records enhanced with PRS. The results showed that prioritising using primary care records can reduce the number needed to screen to prevent one CVD event (NNS), and enhancing it with PRS can further improve this whilst saving the same number of events. Prioritisation with only age and PRS should not be used in isolation due to poor performance.

Key finding 3: The impact of using PRS to decide statin initiations across a lifetime is unknown. We devised four strategies for determining the first age of invitation, followed by a formal risk assessment at which treatment would be allocated if the individual is at high risk were created, each with increasing levels of PRS implementation. Compared to a population-wide invitation strategy followed by assessment using conventional CVD risk factors and PRS, a strategy using PRS to personalise the age of first invitation prior to an assessment led to a 43% and 39% reduction in the NNS in men and women respectively whilst saving a similar number of events over a lifetime.

Overall, this thesis has identified the potential benefits of prioritisation for CVD risk assessments, using existing primary care records within the framework of current guideline recommendations, as well as considering the potential that PRS may have in future healthcare systems.

Publications authored during PhD

Peer reviewed journal articles

- **Chung, R.** et al. (2023). Using Polygenic Risk Scores for Prioritizing Individuals at Greatest Need of a Cardiovascular Disease Risk Assessment. *Journal of the American Heart Association,* 12(15), e029296.

- ESC Cardiovasc Risk Collaboration, & SCORE2 working group. (2021). SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. *European Heart Journal*, *42*(25), 2439-2454.

- Sofianopoulou, E., Kaptoge, S. K.,... **Chung, R.**, ... & Burgess, S. (2021). Estimating dose-response relationships for vitamin D with coronary heart disease, stroke, and all-cause mortality: observational and Mendelian randomisation analyses. *The Lancet Diabetes & Endocrinology*, *9*(12), 837-846.

- Xu, Z., Arnold, M., Sun, L., Stevens, D., **Chung, R**., Ip, S., ... & Wood, A. M. (2022). Incremental value of risk factor variability for cardiovascular risk prediction in individuals with type 2 diabetes: results from UK primary care electronic health records. *International journal of epidemiology*, *51*(6), 1813-1823.

Conference abstracts and presentations

- British Heart Foundation Cambridge CRE Annual Research Symposium 2021 (Poster and lightning talk)

- British Heart Foundation Cambridge CRE Annual Research Symposium 2022 (Poster and lightning talk)

- International Society for Clinical Biostatistics 2022 (poster presentation)

Acknowledgements

This thesis has been conducted using the UK Biobank Resource under Application Number 26865. Data from the Clinical Practice Research Datalink (CPRD) were obtained under licence from the UK Medicines and Healthcare products Regulatory Agency (protocol 162RMn2). This work uses data provided by patients and collected by the NHS and Public Health England as part of their care and support.

I would first like to thank my supervisors Professor Angela Wood and Dr Juliet Usher-Smith, and my advisor Dr Lisa Pennells for their fantastic support throughout my PhD at Cambridge. They have all offered incredible academic guidance, and been patient and understanding, and have all had a part in making my PhD at Cambridge incredibly rewarding. I feel incredibly fortunate to have been supported by these fantastic individuals. In particular, I would also like to thank Professor Angela Wood for our regular catch-up meetings. Her enthusiasm and positivity spurred me on when times were tough, especially throughout the pandemic, and she has consistently encouraged and challenged me to become a better researcher.

I would also like to thank colleagues from Cardiovascular Epidemiology Unit, in particular Professor Emanuele Di Angelantonio, Dr Stephen Kaptoge, Dr Matt Arnold, Dr Lois Kim and Dr Scott Ritchie for their important guidance and insights in helping answer important research questions. I thank the British Heart Foundation for supporting my MPhil and PhD studies with a 4-year studentship (FS/18/56/34177).

I am also grateful for the friends I have made in the department, in particular Jilles Fermont, Carmen Petitjean, Anna Ramond, Owen Taylor and Zhe Xu, and friends in Downing College, in particular Julia Cabanas, Basia Hadrys, MayHsim Lai, Kathi Lauer, Scott Lee, Livia Lisi-Vega, Wing Man, Rhea Parande, Jess Pitchforth, Sarah Tan, Hannah Wynton and Aiwei Zeng.

Finally, I would like to thank my family for their never-ending support, encouragement and love.

Table of contents

List of abbreviations

ACA - American College of Cardiology

AF - Atrial fibrillation

AHA - American Heart Association

ASCVD - Atherosclerotic cardiovascular disease

ASSIGN score - ASSessing cardiovascular risk using SIGN guidelines

BLUPS - Best linear unbiased predictors

BMI - Body mass index

BP - Blood pressure

CAD - Coronary artery disease

CARDIA - Coronary Artery Risk Development in Young Adults

CHD - Coronary heart disease

CHS - Cardiovascular Health Study

CI - Confidence interval

CPRD - Clinical Practice Research Datalink

CRP - C-reactive protein

CTV3 - Clinical terms version 3

CVD - Cardiovascular disease

EMIS - Egton Medical Information Systems

ESC - European Society of Cardiology

GWAS - Genome wide association studies

HES - Hospital Episode Statistics

HDL cholesterol - High density lipoprotein cholesterol

IDI - Integrated discrimination improvement

IQR - Interquartile range

LASSA - Lipid and Atherosclerosis Society of Southern Africa

LDL cholesterol - Low density lipoprotein

LOCF - Last observation carried forward

MI - Myocardial infarction

NICE - National Institute for Health and Care Excellence

NHS - National Health Service

NNI - Number needed to invite to prevent one event

NNS - Number needed to screen to prevent one event

NNT - Number needed to treat to prevent one event

NRI - Net reclassification index

ONS - Office for National Statistics

PAD - Peripheral artery disease

PRS - Polygenic risk scores

QALY - Quality adjusted life year

QOF - Quality and Outcomes Framework

SAHA - South African Heart Association

SCORE2 - Systematic COronary Risk Evaluation 2

SCORE2-OP - Systematic COronary Risk Evaluation 2 for Old Persons

SD - Standard deviation

SIGN - Scottish Intercollegiate Guidelines Network

SBP - Systolic blood pressure

SLE - Systemic lupus erythematosus

SNP - Single nucleotide polymorphism

TIA - Transient ischaemic attack

TPP - The Phoenix Partnership

UKB - UK Biobank

WHO - World Health Organisation

Chapter 1

Introduction

Chapter summary

Cardiovascular diseases (CVD) remain one of the leading causes of morbidity and mortality in the world. The development of CVD risk prediction models has been pivotal in helping identify high risk individuals that may benefit the most from appropriate intervention. In England, health guidelines recommend prioritising individuals for risk assessments using historical electronic health record information. However, research into the best methods for systematically prioritising individuals is limited. In addition, whilst advances in genomics have led to improvements in CVD risk prediction, its impact in enabling better prioritisation is currently unknown. This thesis investigates the population health impact of prioritisation prior to a formal CVD risk assessment.

This introduction provides a background of CVD risk assessment and current health guidelines. It is followed by a summary of advancements in genomic research that may lead to more effective prioritisation, and key statistical methods used throughout this thesis. We then provide the rationale for three aims that the thesis will focus on: 1) the lack of a recommended prioritisation model in England; 2) the impact of including genomic information within prioritisation; 3) utilising genomic information to personalise invitation and treatment strategies.

1.1 Epidemiology of cardiovascular disease

CVD is a group of disorders of the heart and blood vessels and include coronary heart disease (CHD), cerebrovascular disease and heart failure,[1] and remains one of the leading causes of morbidity and mortality in the world. Prevalent CVD cases have risen from 271 million in 1990 to 523 million in 2019, and an estimated 18.6 million CVD-related deaths in 2019, accounting for approximately one-third of global deaths.[2] In England, whilst CVD mortality rates have reduced over recent decades due to improvements in treatments and lifestyle changes, progress in reducing premature CVD mortality has slowed.[3] As such, CVD remains a large causes of premature mortality, with over 37,000 deaths (21%) under the age of 75 caused by CVD in

2020 **(Figure 1.1)**.[4] In 2017, it was estimated that the direct and indirect costs of premature CVD to England's health service and the wider society was estimated to cost £15.8 billion and £7.4 billion per year in non-healthcare and healthcare costs respectively.[5]

The World Health Organisation (WHO) estimates that 80% of premature CVD is preventable[6]; therefore a key target in reducing the burden and costs of CVD is through the management of modifiable risk factors. These risk factors include high blood pressure, high levels of low-density lipoprotein (LDL) cholesterol, diabetes, obesity, smoking, and low physical activity. These can be modified using primordial preventions, such as smoking bans in public spaces, and primary prevention, such as health education campaigns, regular blood pressure checks and cholesterol checks. However, in contrast to the improvement in cardiovascular health for middle aged and elderly individuals over the past 20 years, younger individuals have developed an increased CVD risk profile, due to an increasing prevalence of obesity and diabetes.[7] It is therefore important to reduce the level of risk factors in younger individuals to reduce future morbidity and mortality.[8]

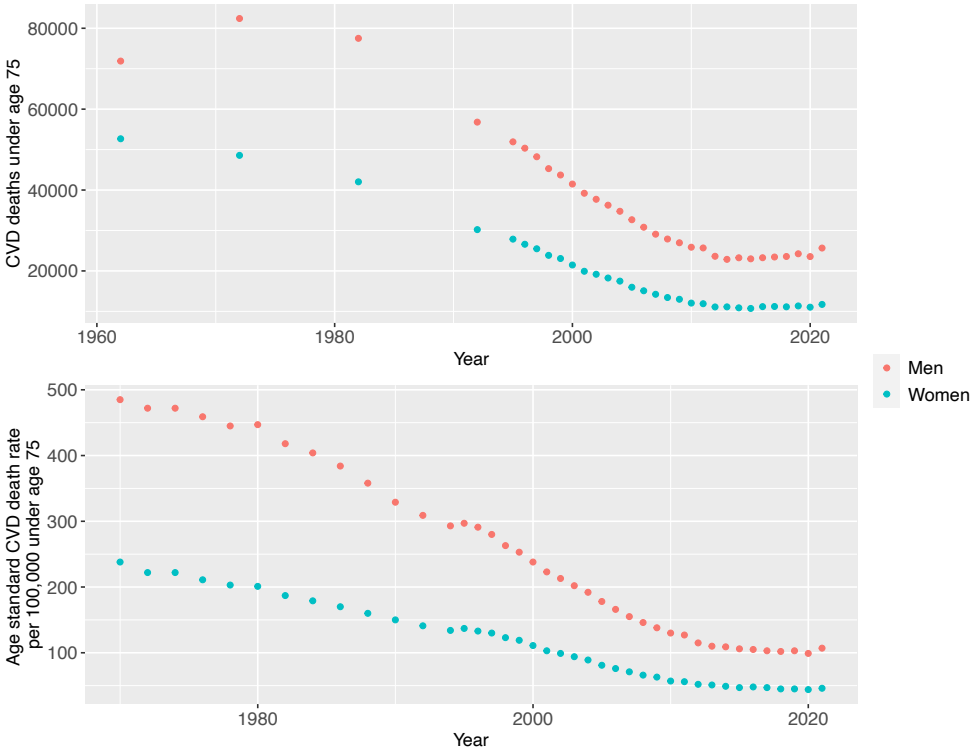


Figure 1.1: CVD deaths and age standardised CVD death rates in England by year and sex. Source: British Heart Foundation, 2022[4]

1.2 Cardiovascular disease risk prediction models

CVD risk prediction models are tools that use an individual's risk factor profile to quantify the risk of a CVD event occurring in a given time period in the future (usually 5 or 10 years).[9,10] They have become a widely implemented tool, as risk prediction models can stratify individuals into low and high-risk groups. In conjunction with clinical guidelines, it can assist clinicians in making appropriate decisions of whether an individual requires additional health interventions or pharmacotherapy. Interventions may include dietary advice, smoking cessation advice, anti-hypertensive medication for high blood pressure and cholesterol lowering medication (e.g., statins) for high cholesterol levels.[11]

As an example, if a model estimated a 10-year CVD risk of 13.8% for an individual, the absolute risk can be interpreted such that if there were 100 individuals with an identical risk factor profile, then approximately 14 individuals would experience an incident CVD event over the next 10 years **(Figure 1.2)**.

Whilst CVD risk prediction models provide a quantitative measure of risk, they have also become an aid in risk communication. By improving the way risk and risk factors are communicated with the general public, compliance to interventions can be improved, thus improving patient safety.[12–14] However, whilst a systematic review showed evidence that providing CVD risk model estimates to professionals and patients improved perceived CVD risk and medical prescribing, with little evidence of harm on psychological well-being,[15] the impact of risk communication may be limited as a trial showed that it did not substantially change risk perception of behaviour.

Since the introduction of the first CVD risk prediction model, the Framingham Risk Score in 1998[16], multiple risk prediction models have been developed catering for different populations and countries. Risk prediction models have also grown in complexity over time with additional novel risk factors being researched and implemented, in addition to larger datasets being used to derive the models. We reviewed a selected group of models, recommended in different international guidelines, in the following sections.

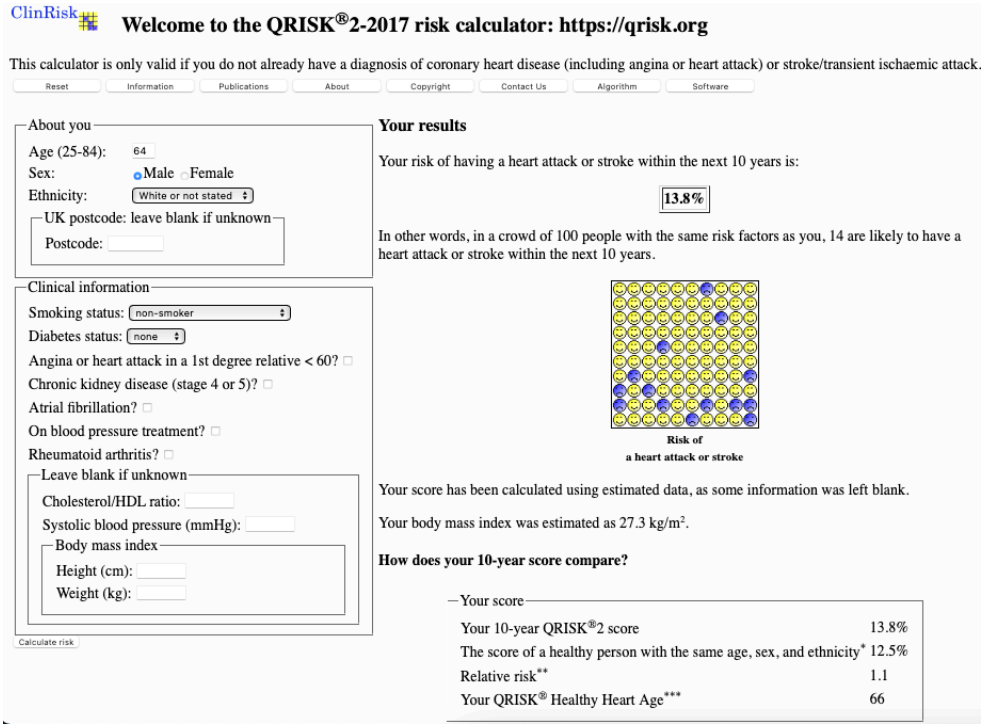


Figure 1.2: Screenshot of QRISK2 online risk calculator estimating the 10-year CVD risk of an example 64-year-old Caucasian man.

1.2.1 Data sources used for developing CVD risk prediction models

A selection of CVD risk prediction models that are recommended within current, and upcoming, healthcare guidelines in middle to high income countries were summarised **(Table 1.1).** In the selection of 11 guidelines, 8 CVD risk prediction models were recommended. Data sources for model derivation can be grouped into three categories: a traditional cohort study, a combination of cohorts, or a primary-care records-based cohort. The largest data sources used for developing risk models were primary care records based. These include QRISK2 and QRISK3 which used 1.28 million individuals and 7.89 million individuals respectively.

The age range of individuals used for derivation varied, however 7 risk models were developed in populations that included individuals aged between at least 40 and 70 years. QRISK3 included the greatest range of individuals with those aged between 25 years and 85 years. An exception to the other 7 models was the SCORE2-OP model which focussed only on individuals older than 70 years.

Generally, the risk models recommended by each guideline chose risk models designed for use in the population of interest. For example, the QRISK2 and QRISK3 risk models, as

recommended by the NICE guidelines in England, were derived in a database of English primary care records. The SCORE2 risk score, as recommended by the ESC guidelines, was derived in a combined dataset of 45 cohorts from 13 European countries, followed by recalibration to adjust for different population characteristics in 55 countries. However, guidelines in Australia, Canada, Singapore and South Africa all recommend the Framingham Risk Score, a risk model developed in a US-based population.

1.2.2 Risk factors used in CVD risk prediction models

Current CVD risk prediction models recommended for clinical practice use risk factors that are well understood to be major contributors to the risk of developing CVD and are easily measured. These include: age, SBP, total, LDL and high-density lipoprotein (HDL) cholesterol (usually a combination of total and HDL, or LDL and HDL), blood pressure medication, diabetes status and smoking status **(Table 1.1)**.

Some risk scores may include additional known risk factors, including BMI, ethnicity, family history of CVD and deprivation, as found in the QRISK2 and QRISK3 family of risk scores. Over the past two decades, a search to discover novel risk factors and other non-conventional risk factors has led to a set of candidate biomarkers in potentially causal pathways. These include triglycerides and lipoprotein(a) (markers of hyperlipidaemia), C-reactive protein (CRP) and IL6 (markers of low-grade inflammation), fibrinogen (a marker of haemostatic activity) and glucose levels and HbA1c (markers of metabolic dysfunction). Some recent risk models have incorporated a few novel biomarkers, including CRP and lipoprotein(a) in the Reynolds Risk Score published in 2007, and triglycerides in PROCAM published in 2002 and 2007. The inclusion of novel biomarkers however requires additional laboratory-based costs.

1.2.3 Outcomes used in CVD risk prediction models

The choice of outcomes used in CVD risk prediction models varied between each of the selected risk models **(Table 1.1).** Some of the models chosen estimated the future risk of a combination, or a composite, of multiple CVD outcomes.

A composite outcome typically includes at least one primary, or "hard", outcome and can be defined as events that have permanent consequences, including non-fatal CHD, myocardial infarction (MI), stroke and death. In addition, the defined outcome may be combined with secondary, or "soft", outcomes, such as hospitalisation for angina or a transient ischemic attack (TIA).[17,18] The majority of the chosen risk models estimate the risk of both fatal and non-fatal CHD and stroke over 10 years. The Framingham score, QRISK2, QRISK3, and PREDICT also included soft outcomes including: angina, TIA, and peripheral artery disease (PAD).

Table 1.1: Characteristics of current guidelines on CVD risk assessment and primary prevention and recommended risk prediction models implemented in a selection of middle to high income countries

Guideline (country/region)	Risk model recommended	Country/region model originally developed for	Data source	Participants and age range	Risk factors	Outcome	Recommended threshold(s) for initiation of lipid lowering medication*	Prioritisation of individuals for assessment
2017 SIGN Risk estimation and the prevention of cardiovascular disease (Scotland)[19]	ASSIGN score 2007[20]	United Kingdom	Scottish Heart Health Extended Cohort (Cohort recruitment between 1984-1995)	N = 13,297 (30-74 years)	Age, sex, total and HDL cholesterol, SBP, diabetes, smoking, family history of CVD, deprivation.	10-year CVD (CHD, cerebrovascular, coronary artery interventions)	10-year CVD risk ≥20%	N/A
2012 National Stroke Foundation Guidelines for the management of absolute CVD risk (Australia)[21]	Framingham Risk Score 2008[22]	United States	Framingham Heart Study and Framingham Offspring Study (Cohort recruitment between 1968-1971, 1971-1975, 1984-1987)	N = 8,491 (30-74 years)	Age, sex, total and HDL cholesterol, SBP, blood pressure medication, diabetes, smoking.	10-year CVD (CHD, stroke, PAD, heart failure)	5-year CVD risk ≥15%. Consider if 5-year risk 10%-15%, and 3-6 months of lifestyle intervention does not reduce risk, or BP ≥160/100mmHg, familial history of premature CVD, or certain ethnic groups.	N/A
2017 Ministry of Health Clinical Practice Guidelines on lipids (Singapore)[23]	Framingham Risk Score 2008[22]	United States	Framingham Heart Study and Framingham Offspring Study (Cohort recruitment	N = 8,491 (30-74 years)	Age, sex, total and HDL cholesterol, SBP, blood pressure medication, diabetes, smoking.	10-year CVD (CHD, stroke, PAD, heart failure)	10-year CAD risk > 20%	N/A

			between 1968-1971, 1971-1975, 1984-1987)					
2018 SAHA/LASSA South African Dyslipidaemia Guideline Consensus Statement (South Africa)[24]	Framingham Risk Score 2008[22]	United States	Framingham Heart Study and Framingham Offspring Study (Cohort recruitment between 1968-1971, 1971-1975, 1984-1987)	N = 8,491 (30-74 years)	Age, sex, total and HDL cholesterol, SBP, blood pressure medication, diabetes, smoking.	10-year CVD (CHD, stroke, PAD, heart failure)	10-year CVD risk ≥15% and LDL cholesterol level ≥ 2.5 mmol/L or 10-year CVD risk ≥30% and LDL cholesterol level ≥ 1.8 mmol/L	N/A
2021 Canadian Cardiovascular Society Guidelines for the Prevention of CVD (Canada)[25]	Modified Framingham Risk Score 2008[22]	United States	Framingham Heart Study and Framingham Offspring Study (Cohort recruitment between 1968-1971, 1971-1975, 1984-1987)	N = 8,491 (30-74 years)	Age, sex, total and HDL cholesterol, SBP, blood pressure medication, diabetes, smoking.	10-year CVD (CHD, stroke, PAD, heart failure)	LDL cholesterol ≥ 5.0 mmol/L or 10-year CVD risk ≥ 20% or 10-year CVD risk 10-19% and LDL cholesterol ≥ 3.5 mmol/L	N/A
2019 ACC/AHA Guideline on the Primary Prevention of CVD (United States)[26]	Pooled Cohort Equations 2013[27]	United States	Selection of cohorts: ARIC, CHS, CARDIA, Framingham Original and Offspring Study. (Cohort recruitment between 1968-1990)	N = 24,626 (40-79 years)	Age, sex, total and HDL cholesterol, SBP, blood pressure medication, diabetes, smoking.	10-year ASCVD (Non-fatal MI, CHD death, fatal/non-fatal stroke)	10-year CVD risk ≥7.5% and LDL cholesterol level of 1.8–4.8 mmol/L	N/A

2018 New Zealand CVD Risk Assessment and Management for Primary Care (New Zealand)[28]	PREDICT 2018[29]	New Zealand	Primary care providers (Cohort recruitment between 2002-2015)	N = 401,752 (30-74 years)	Age, sex, total and LDL cholesterol, SBP, blood pressure medication, diabetes, smoking, family history of CVD, deprivation.	5-year CVD (Ischaemic heart disease (including angina); ischaemic or haemorrhagic cerebrovascular events (including TIA); or peripheral vascular disease, congestive heart failure, or other ischaemic CVD deaths)	5-year CVD risk ≥15%	N/A
2014 NICE guidelines on CVD risk assessment and lipid modification (England)[30]	QRISK2 2008[31]	United Kingdom (England and Wales)	Primary care providers contributing to QResearch (Cohort recruitment between 1993-2008)	N = 1.28 million in derivation cohort (35-74 years)	Age, sex, ethnicity, total and HDL cholesterol, SBP, blood pressure medication, diabetes, smoking, family history of CVD, BMI, deprivation, renal disease, AF, rheumatoid arthritis	10-year CVD (CHD (MI and angina), stroke, TIA)	10-year CVD risk ≥ 10%	Yes – using existing primary care records to systematically prioritise for a full formal risk assessment.
(Upcoming) 2023 NICE guidelines on CVD risk assessment and lipid modification (England)	QRISK3 2017[32]	United Kingdom	Primary care providers contributing to QResearch. (Cohort recruitment between 1998-2015)	N = 7.89 million in derivation cohort. (25-84 years)	Age, sex, ethnicity, total and HDL cholesterol, SBP, blood pressure medication, diabetes, smoking, family history of CVD, BMI, deprivation, renal disease, AF, rheumatoid arthritis, SBP variability, migraine, corticosteroids, SLE, atypical antipsychotics, severe mental illness.	10-year CVD (CHD (MI and angina), stroke, TIA)	10-year CVD risk ≥ 5%	Yes – using existing primary care records to systematically prioritise for a full formal risk assessment.

2021 ESC Guidelines on CVD prevention in clinical practice (Europe)[33]	SCORE2 2021[34]	Europe	45 European cohorts recalibrated to 55 countries (Cohort recruitment between 1990-2018)	N = 677,684 in derivation cohort. (40-69 years)	Age, sex, total and HDL cholesterol, SBP, diabetes, smoking	10-year CVD (CVD mortality, non-fatal MI, non-fatal stroke)	10-year CVD risk: • ≥ 7.5% for under 50-year-olds, • ≥10% for 50–69-year-olds,	N/A
2021 ESC Guidelines on CVD prevention in clinical practice (Europe)[33]	SCORE2-OP 2021[35]	Europe	Cohort of Norway study (Cohort recruitment between 1994-2013)	N = 28,503 in derivation cohort. (70+ years)	Age, sex, total and HDL cholesterol, SBP, diabetes, smoking	10-year CVD (CVD mortality, non-fatal MI, non-fatal stroke)	10-year CVD risk ≥15% for over 70-year-olds	N/A

Abbreviations: AF, atrial fibrillation; ACC, American College of Cardiology; AHA, American Heart Association; ARIC, Atherosclerosis Risk in Communities; ASCVD, atherosclerotic cardiovascular disease; ASSIGN score, ASSessing cardiovascular risk using SIGN guidelines; BMI, body mass index; CAD, coronary artery disease; CARDIA, Coronary Artery Risk Development in Young Adults; CHD, coronary heart disease; CHS, Cardiovascular Health Study; CVD, Cardiovascular disease; ESC, European Society of Cardiology; LASSA, Lipid and Atherosclerosis Society of Southern Africa; LDL, low-density lipoprotein cholesterol; HDL, high-density lipoprotein cholesterol; MI, myocardial infarction; NICE, National Institute for Health and Care Excellence; PAD, peripheral artery disease; SAHA, South African Heart Association; SBP, systolic blood pressure; SCORE2, Systematic Coronary Risk Evaluation; SCORE2-OP, Systematic Coronary Risk Evaluation 2 for Old Persons; SIGN, Scottish Intercollegiate Guidelines Network; SLE, systemic lupus erythematosus; TIA, transient ischaemic attack.

*Individuals without pre-existing disease

1.3 Primary prevention of cardiovascular disease guidelines in England

Clinical guidelines on the primary prevention of CVD in England were first introduced in 2008 by the National Institute for Health and Care Excellence (NICE), with a subsequent major update in 2014.[30] These evidence-based guidelines were created to facilitate the improvement in healthcare and provide recommendations of the appropriate care and treatments needed to manage CVD. A major component of the guidelines is the identification of individuals for a CVD risk assessment and its subsequent management, including lifestyle modifications and lipid modification therapy **(Box 1.1)**.

Box 1.1: NICE guidelines (2014) - Lipid modification: Cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease[30]

Identifying and assessing cardiovascular disease (CVD) risk

Identifying people for full formal risk assessment

1. For the primary prevention of CVD in primary care, use a systematic strategy to identify people who are likely to be at high risk.
2. Prioritise people on the basis of an estimate of their CVD risk before a full formal risk assessment. Estimate their CVD risk using CVD risk factors already recorded in primary care electronic medical records.
3. People older than 40 should have their estimate of CVD risk reviewed on an ongoing basis.
4. Prioritise people for a full formal risk assessment if their estimated 10-year risk of CVD is 10% or more.
5. Discuss the process of risk assessment with the person identified as being at risk, including the option of declining any formal risk assessment.
6. Do not use opportunistic assessment as the main strategy in primary care to identify CVD risk in unselected people.

The guidelines have two steps for assessing CVD risk; the first is to identify and prioritise high-risk individuals using existing information, followed by a second step of estimating an individual's current CVD risk by using current risk factor levels obtained at an in-person formal risk assessment.

Clinicians are recommended to identify individuals who may benefit the most from a full formal risk assessment. This is to be done using *a systematic strategy to identify people who are likely to be at high risk.* It is recommended *to prioritise people on the basis of an estimate of their CVD risk before a full formal risk assessment by estimating their CVD risk using CVD risk factors already recorded in primary care electronic medical records.* Individuals should then be *prioritised for a full formal risk assessment if the estimated 10-year risk of CVD is greater than 10%,* and *those older than 40 are expected to have their estimated CVD risk reviewed on an ongoing basis*.

Upon acceptance of an invitation to a formal risk assessment, *the QRISK2 risk assessment tool* is used *to assess CVD risk for the primary prevention of CVD in people up to and including age 84 years*, where current risk factor measurements would be taken. Statins would be offered to those who have a 10% or greater 10-year risk of developing CVD. The 10% formal risk assessment threshold was revised from 20% in 2014 due to a reduction in the cost of statins, allowing the health service to offer statins to a greater number of individuals.[30]

Throughout this thesis, we will use the term "prioritisation" to indicate the prioritisation of individuals using an estimate of their CVD risk calculated with existing data, and the term "formal CVD risk assessment" to indicate a risk assessment performed in person using current risk factor levels.

In 2009, the National Health Service (NHS) Health Check, a national preventative healthcare programme was introduced. The programme was introduced to systematically measure and manage health outcomes by assessing the risk, create awareness and manage CVD risk factors[36], including diabetes, heart disease, kidney disease, stroke and dementia, and became a mandated service in 2013, with the health check being offered to all individuals without a prior history of CVD aged between 40 and 75 years every five years.[37]

The programme contains best practice guidance to support local public health commissioners and providers of the NHS Health Check with information needed to commission and deliver

the programme. Importantly, it is designed to be used in conjunction with the recommendations from the NICE guidelines, and many of the recommendations made in the NICE guidelines are also recommended within the best practice guidance. Whilst the latest guidance from 2019 does not mention the use of prioritisation, it was previously mentioned up until 2017 and was further mentioned in a review by the Office for Health Improvement and Disparities, stating that: "As people look for new ways to participate in their health and wellbeing, technologies and digital innovations provide the means for improvements…It also offers the potential to redirect resulting efficiencies to prioritise those at greatest risk, who would benefit most, in order to 'level up' outcomes."[38]

Whilst the recommendations of prioritising using existing primary care records were guided by research that showed the potential to reduce running costs and health inequality, there are current limitations with the guidelines.[30] First, no dedicated prioritisation tool or existing risk tool is currently recommended. Second, a lack of a recommended tool, in particular one not built into a primary care-based system, may limit a practice's ability to prioritise individuals. Third, no recommendations are made in how the existing primary care records should be handled. For example, whether only the most recent record should be used or whether all historical records should be used. Fourth, the guidelines make no recommendations in how individuals without primary care records should be prioritised. Finally, a fixed 10% threshold is recommended for both during prioritisation and during formal assessment.

With recent economic analysis suggesting that optimising the NHS Health Check programme could be modestly effective and cost-effective, the use of a systematic and automated tool to prioritise individuals could be of importance.[39,40]

1.3.1 Implementation of QRISK2 CVD risk prediction model

The QRISK2 CVD risk prediction model is currently recommended for formal risk assessments by the NICE guidelines (see **Chapter 1, Section 1.2.1).** It was developed using a database of primary care records in England and Wales, and was published in 2008, and the model coefficients are available for researchers to use.[31] The risk model was externally validated in an independent database of primary care records in the UK, and showed good discriminatory performance and calibration.[41]

The QRISK2 risk model is currently implemented into major primary care systems and is also accessible as a web-calculator. QRISK2 also includes a method for imputing missing information. Missing values for SBP, total and HDL cholesterol, and BMI are imputed using age-, sex- and ethnicity-specific means.[42] By incorporating a method for imputing missing data, the scores can be used with primary care data where missing risk factor information is more likely.

1.3.2 Comparison with international cardiovascular disease guidelines

CVD guidelines across the world compare similarly with those in England. However, key differences can be observed in a number of areas: the CVD risk model used, the risk threshold used to determine statin initiation, and overall recommendations relating to the identification of individuals at high risk of CVD **(Table 1.1).**

The choice of a single fixed absolute risk threshold is commonly recommended across all guidelines. However, these vary by both risk score and country. For example, in Australia, a 5-year CVD risk greater than 15% is currently recommended whereas a 10-year risk threshold of 10% is recommended in England and a 10-year risk threshold of 7.5% in the United States of America. Similar to the reduction of the recommended risk threshold in England from 20%, published in 2008, to 10% in 2014, US guidelines also reduced the threshold from 20%, published in 2007, and lowered to 7.5% in 2013. In comparison with the single fixed thresholds recommended, the 2021 European Society of Cardiology (ESC) guidelines implemented varying risk thresholds determined by age-group and was chosen due to the influence of age in a risk model. Individuals were defined as at very-high risk of CVD, with treatment being recommended, if their 10-year CVD risk was greater than 7.5% if the individual is younger than 50 years, greater than 10% if the individual is between 50-69 years, and greater than 15% if the individual is older than 70 years.

Whilst the NICE guidelines in England makes explicit recommendations for the prioritisation of individuals using existing primary care records, no other guidelines implement similar recommendations.

1.3.3 Upcoming primary prevention of cardiovascular disease draft guidelines in England

During the preparation of the thesis, a draft copy of upcoming NICE guidelines in England was published in early 2023 **(Table 1.1)**. The major changes between the current and the upcoming guidelines is the lowering of the recommended 10-year CVD risk threshold for statin initiation, from 10% to 5%, and a change in the recommend CVD risk model from QRISK2 to QRISK3. As such, additional analyses will be conducted exploring how changing the risk thresholds will impact the population health.

1.4 Emerging risk factors

1.4.1 Utilising repeated measures from primary care records

Utilising repeated measures from primary care records has the potential to improve CVD risk prediction in a cost-effective manner. Whilst large databases of primary care records exist, many existing CVD risk models do not implement methods to handle repeated measures. As such, this presents an opportunity to systematically assess individuals at high risk of CVD, by harnessing all risk factor measurements already recorded in their primary care records.

Generally, the majority of current CVD risk models are derived using traditional prospective cohorts, where single observed measurements for each risk factor, taken at study entry, are used to derive the model. This was done for risk models including the Framingham Risk Score, SCORE2, SCORE2-OP and Pooled Cohort Equations score. An exception to this is the PREDICT, QRISK2 and QRISK3 risk models, which were derived using a primary care records-based cohort. However, whilst primary care records were utilised for model derivation, the PREDICT, QRISK2 or QRISK3 models do not fully utilise the repeated measurements during model derivation or within clinical use. In particular, all three models generally use a single, most recent, measurement for each risk factor observed before study entry for the model derivation. One exception is for QRISK3, which includes a measure for historical SBP variability.

Harnessing repeated measures can offer additional information by capturing individual level risk trajectories that may not be captured by using a single measurement.[43,44] However, challenges in the handling of such data should be considered.[43,45,46] First, unlike single risk factor measurements recorded in a prospective cohort, primary care records are dynamic over

time. There may exist a greater number of measurements in certain risk factors than others, and the time between measurements may be non-constant. Second, most current risk models often require complete information for all risk factor measurements. Although QRISK2 and QRISK3 substitute missing measurements with age-and sex-specific average risk factor values, it does not use other known information to impute. Since data collected in a primary care system is likely to be used for patient management, missingness of risk factor data is likely and needs to be considered.[43] This is more likely in younger individuals or men who may not attend primary care services as often, and may altogether skew the representativeness of the data used to derive the model.[47–49] Third, due to the longitudinal nature of the data, primary care records can be large and computationally intensive during the model derivation.

1.4.2 Genetic information

Due to ever-reducing costs and technological developments, genome wide association studies (GWAS) have made measuring variations across millions of genetic markers, also known as single nucleotide polymorphisms (SNPs), across the human genome feasible. These advancements have led to large-scale studies of genes and has improved the understanding of genetic factors and the biological mechanisms of risk factors and diseases. It has also led to a better understanding that a large number of genetic, or polygenic, variants can contribute to diseases or risk factor.[50]

Polygenic risk scores (PRS) are one way to summarise this genetic susceptibility to the disease or risk factor of interest. PRS summarise the estimated effect of many genetic markers associated with a genetic trait of interest. After identifying the genetic markers of interest, the polygenic risk score is created as a weighted sum of the number of trait-associated alleles in an individual. The weights chosen is typically proportional to the odds ratio of the SNP to the disease or risk factor trait. The combined weighted sum of the SNPs, multiplied by whether an individual has zero, one or two copies of the SNP of interest creates a unique, personalised polygenic risk score of that trait. As genetic risk is accumulated continuously over the entire lifespan, genetic risk scores may be able to capture lifetime risk at the individual level.[51]

A recent development of improving PRS is by meta-analysing multiple polygenic risk scores of the same trait.[52,53] By combining multiple studies, each of which may: involve a limited number of genetic variants used, have small statistical power to provide precise effect sizes, or

limited diversity in ethnicities, a meta-analysed score allows for the creation of a more statistically powerful risk score.

The inclusion of CVD-based PRS in a risk model has shown promising improvements in the area of risk prediction.[51–57] PRS have been created for many diseases, typically affected by many small-effect variants (as opposed to a handful of variants with large effects), including coronary artery disease (CAD) and stroke.[53,55] PRS have also been explored for conventional CVD risk factors including blood pressure, cholesterol levels and BMI.[57–59]

Disease-based PRS have shown the potential to improve existing risk scores based on conventional risk factors alone due to the risk factors orthogonal nature to existing risk factors. It has been consistently shown that the addition of disease-based genetic risk score can not only improve a model's ability to discriminate between high and low risk individuals, but also its ability to stratify risk (**Figure 1.3).**

In one study by Inouye et al (2018), researchers created a polygenic risk score (PRS) for coronary artery disease (CAD) and demonstrated that the CAD PRS alone was more associated with 10-year risk and outperformed individual conventional risk factors[52]. After combining the conventional CVD risk factors together, the model using only the CAD PRS performed similarly. It also showed that the combination of traditional risk factors with genetic data improved on a risk score with only traditional risk with an increase in the C-index, the probability a randomly selected individual who had an event was correctly ranked than an individual who did not have an event, by 3.7%. Other studies have shown similar improvements in performance.[60]

PRS for CAD have also demonstrated the ability to stratify individuals based on their genetic risk alone, identifying different trajectories of risk across all ages when stratified by quintiles of genetic risk.[51,52,55,60] As the genome remains constant from birth to death, an individual's genetic risk score can be calculated and used at any time in the life course. Consequently, PRS has the potential to estimate an individual's lifetime risk of CVD. As such PRS may be better for younger individuals who may not have developed conventional risk factors yet.[61]

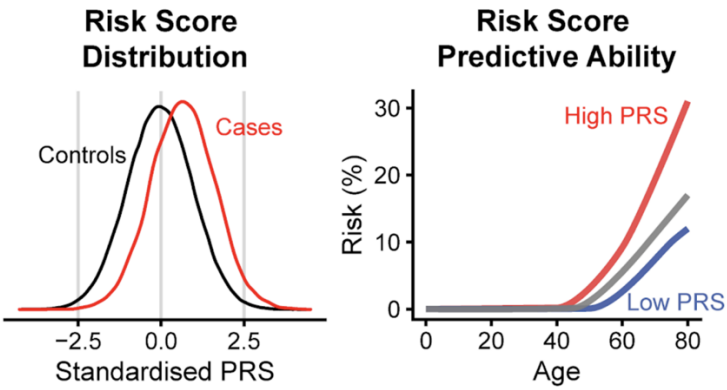


Figure 1.3: Illustration of polygenic risk score distribution and predictive ability.
Source: Wand et al., 2021[62]

1.5 Statistical methods in risk model development

The use of core statistical methods has played a pivotal role in the development and progression of CVD risk models. Since the inception of the Framingham Score, the majority of risk scores continue to rely on key statistical principles to develop, validate and assess the models. However, as research on risk prediction continues, new and more novel statistical methods have emerged, better utilising larger datasets and harnessing the increased computational resources available.

In this section, we discuss key statistical methods used in risk prediction, along with emerging statistical methods.

1.5.1 Established methods

1.5.1.1 Cox proportional hazards

Survival analysis is a key branch of statistics that allows researchers to model time to event data and make inferences of the rate of an event over time. Whilst statistical techniques such as Kaplan-Meier plots are effective if comparing the survival of two groups (e.g., smokers versus non-smokers or placebo versus treatment), the Cox proportional hazards model is a key method in the estimation of risk in individuals with one of more risk factors.[63]

The development of the Cox proportional hazards model has in turn led to the creation and ongoing research of risk calculators. By using the estimated survival function and risk factor model coefficients, the risk of an event within a specified period of time for every individual can be estimated. It can be estimated using the following formula for individual i:

$$h_i(t) = h_0(t)\, e^{(\beta x_i)}$$

Where $h_0(t)$ represents the baseline hazard function at time t, and βx_i represents a linear combination of the model coefficients and the individual's risk factor levels. As such, the Cox model forms the fundamental building blocks for most risk calculators.

Whilst the Cox proportional hazards model is commonly used to estimate risk, the model was designed to estimate the model coefficients without needing to estimate the baseline hazard, i.e., a semi-parametric model. As such, a major disadvantage of using a Cox proportional hazards model is the need to separately estimate the baseline hazard function after, which can be done using methods including the Breslow estimate. An alternative approach to the semi-parametric model is to fit a full parametric model, such as the Weibull model. However, in this case the baseline hazard must satisfy a Weibull shape for the model to be valid.

Another disadvantage of using the Cox proportional hazards model is that uncertainties in both the estimated baseline hazard and the coefficients are typically not accounted for, resulting in uncertainty in the risks estimated by a risk model.

1.5.1.2 Competing risks adjusted survival models

Censoring is a core concept of the Cox proportional hazards model. In time to event data, an individual is censored if they do not experience the event of interest during the study's follow up period.[64] This can happen if they were lost to follow up or if the event did not occur between study entry and at the end of follow up. Conventional statistical analysis makes a key assumption in that censoring is non informative, where if an individual is censored, the risk of an event during analysis is not affected as if the censored individuals who dropped out are random. However, the presence of competing risks, defined as events not of the primary interest which occur before an event of the primary interest, can affect analyses and requires additional adjustments to the Cox model.[65] An example of a competing risk is if a study was investigating first CVD events, then non-CVD related deaths is the competing event.

As conventional statistical analysis assumes non informative censoring, the existence of competing risks may violate the assumptions made, leading to overestimated risks of the primary event. The goal therefore is to calculate the probability of an event accounting for competing risks. For example, assuming no competing risks, estimating the crude incidence of an event can be performed using a Kaplan-Meier estimate is frequently used.[65] Using such a method in the presence of competing risks may lead to overestimated risks. A common approach to deal with competing risks is to calculate the cumulative incidence function, which measures the marginal probabilities of an event occurring.[65,66]

When fitting regression models to estimate associations, two alternative modelling approaches are typically implemented: cause-specific hazards, and Fine and Gray.[67,68] The former estimates the associations within each cause-specific hazard, whilst the latter estimates the associations on the cumulative incidence function.

1.5.1.3 Model performance metrics

The assessment of performance plays an important role during the development of a new risk model. Researchers are often interested in aims including: evaluating predictive performance in a new target population, quantifying performance gains upon the introduction of a novel risk factor to an existing risk score, and comparing multiple risk models against each other.[69] Subsequently, metrics have been designed to quantify certain attributes of a model to answer these questions.

A model's calibration is defined as the agreement between the predicted risks and the frequency of observed outcomes.[70] Calibration can be assessed by plotting the observed frequencies against the mean predicted risks and is often performed within risk groups, such as quintiles or deciles of predicted risks. A well calibrated model would have each of the points on the line of equality. Calibration can also be assessed using a Brier score, which is calculated as the mean square error within the prediction estimates.[71]

To quantify a model's discriminative ability i.e., the ability to distinguish between individuals who go on to have an event versus those who do not, the C-index (or C-statistic) is a widely used metric.[72,73] The C-index is assessed by calculating the probability that the predicted risks are correctly ranked in a randomly selected pair of individuals and is defined using the formula:

$$Cindex = \frac{\sum_{i \neq j} 1\{n_i < n_j\} 1\{T_i > T_j\} d_j}{\sum_{i \neq j} 1\{T_i > T_j\} d_j}$$

Where i and j represent different pairs of n individuals, T represent the time-to-event response, and d represent whether an event occurred. The C-index ranges from 0 to 1, where a value of 0.5 is equivalent to random chance and 1 being equal to a perfect test to distinguish events and non-events.

Risk reclassification is used to summarise the improvement in a risk model after the inclusion of a new risk factor. It aims to be more clinically relevant by quantifying the movement of individuals between risk groups. Correct reclassifications are classed as events moving up towards higher risk groups and non-events moving down towards lower risk groups. One summary metric is the net reclassification index (NRI); this is interpreted as the net proportion of individuals correctly reclassified, and can be summarised within events and non-events separately. The NRI can be used by specifying categorical groups (e.g., whether the risk is below or above a 10% threshold) or continuously (whether the risk is greater or lower after the inclusion of a risk factor). Another metric is the integrated discrimination improvement (IDI), which also quantifies reclassification without the use of a threshold, and is interpreted as the number of individuals whose risk predictions correctly increase or decrease in value.

1.5.1.4 Population health modelling

Population health modelling is a method used to estimate the benefits and harms of implementing interventions at a population level.[54,74] In the field of disease screening and intervention, the aim is to measure the impact adding a new risk factor variable to an existing risk score. One way of measuring the impact is to summarise the number of events saved due to the intervention. Further metrics based off the number events of saved include:

1) NNT: the number of individuals needed to treat to prevent one event,
2) NNS: the number of individuals needed to screen to prevent one event,
3) NNI: the number of individuals needed to invite for screening to prevent one event.

1.5.2 Emerging statistical methods

1.5.2.1 Handling repeated risk factor measurements

Multiple statistical methods have been developed to better optimise repeated risk factor measurements for use in CVD risk prediction. The methods, listed in terms of statistical complexity are summarised as follows.

1) Last observation carried forward

Last observation carried forward (LOCF) is one of the simplest methods in handling repeated measurements and can be used for both continuous and categorical risk factors. It assumes the last observed measurement is used, regardless of other historical information and often the amount of time since the last measurement. In risk models that use a single measurement for each risk factor, the LOCF method can be used in scores such as: QRISK2, QRISK3, SCORE2, Pooled Cohort Equations, and Framingham.[22,27,31,32,34]

2) Average of multiple measurements

An extension to the LOCF method is to allow a risk factor measurement to be a summary of multiple risk factor measurements over time. Compared to using a single measurement, using multiple measurements can reduce measurement error and long term within person variation.[75] One example is to use a cumulative mean of historical measurements. This is specified where the oldest observation is the starting point, and a cumulative, or running, mean is calculated for each new, more recent risk factor measurement. Another example is to simply the cumulative mean approach and to specify and restrict the calculation to a handful of the most recent risk factor measurements.

3) Mixed-effects linear models

The use of linear regression modelling allows the modelling and prediction of risk factor values. However, a key assumption when fitting a linear regression model is that the risk factor measurements used to derive the model are independent and identically distributed. Since multiple responses from the same individual cannot be assumed to be independent and identically distributed, using simple linear regression with repeated measurements can result in misleading inferences. By adapting the methodology of linear regression, mixed-effects linear regression model can be used to fit on all available risk factor measurements and estimate future risk factor values allowing for variation between and within individuals.

The mixed-effects linear model combines both fixed effects and random effects.[75–78] The most common model is the random-intercept mixed model, where in addition to a fixed intercept and fixed slopes for the risk factor, a random intercept is also estimated. It is normally distributed with mean 0 and an estimated variance, and allows individuals to have a varying intercept, where each prediction is shifted up or down relative to using only a fixed intercept. Similarly, a random slope may also be included in the model to vary the person-level slope.

The model may be fitted either as a univariate model, fitting onto one risk factor of interest, or as a multivariate model, fitting onto multiple risk factors simultaneously. An advantage of the multivariate mixed effects model is its ability to handle missing values by using the intra-correlation structure of risk factor information. This means that to have a prediction of all of the risk factors from the multivariate model, individuals must have at least one measurement from at least one risk factor.

4) Landmark modelling

Dynamic risk prediction takes into account time-dependent risk factors. One approach is to use a landmark framework.[43,79] By defining a set of reference, landmark ages (e.g. 40, 45, …, 70 years), landmark-age specific Cox proportional hazards models are derived on eligible individuals who before the landmark age, at which predictions will be made from. Individuals may contribute to multiple landmark-age specific models. By doing so, the landmark model can be updated dynamically using the most relevant data.

The landmark model can be further extended and optimised using a two-stage process. As risk factor information is typically summarised up until the specified landmark-age, landmark models typically use LOCF risk factor values. This however can be improved upon by using a mixed-effects model to handle the repeated measurements found in primary care records, prior to fitting the Cox model. This was demonstrated by Paige et al.[45]

5) Joint modelling

Joint modelling allows for both the longitudinal component and the survival component of a risk model to be modelled simultaneously, and allows the estimated risk to change dynamically.[75,77] Joint modelling can lead to improvements in the efficiency of statistical inferences and reduce biases. They can also be used to improve prediction, because they are tailored to account for individual variability, and determine whether a longitudinal process is a surrogate for a time-to-event process.[80,81] However, due to its computationally intensive nature,

its use with primary care records may not be suitable for practical consideration until computational resources improve in the near future.[82]

1.5.2.2 Model recalibration

As mentioned in **Section 1.2.3,** applying a previously developed CVD risk model to a new target population may lead to inaccurate estimated risks, due to differences in the characteristics between the risk model's derivation population and the new target population. One solution that has been widely used is to use a simple linear transformation to recalibrate the risk model (**Figure 1.4).**

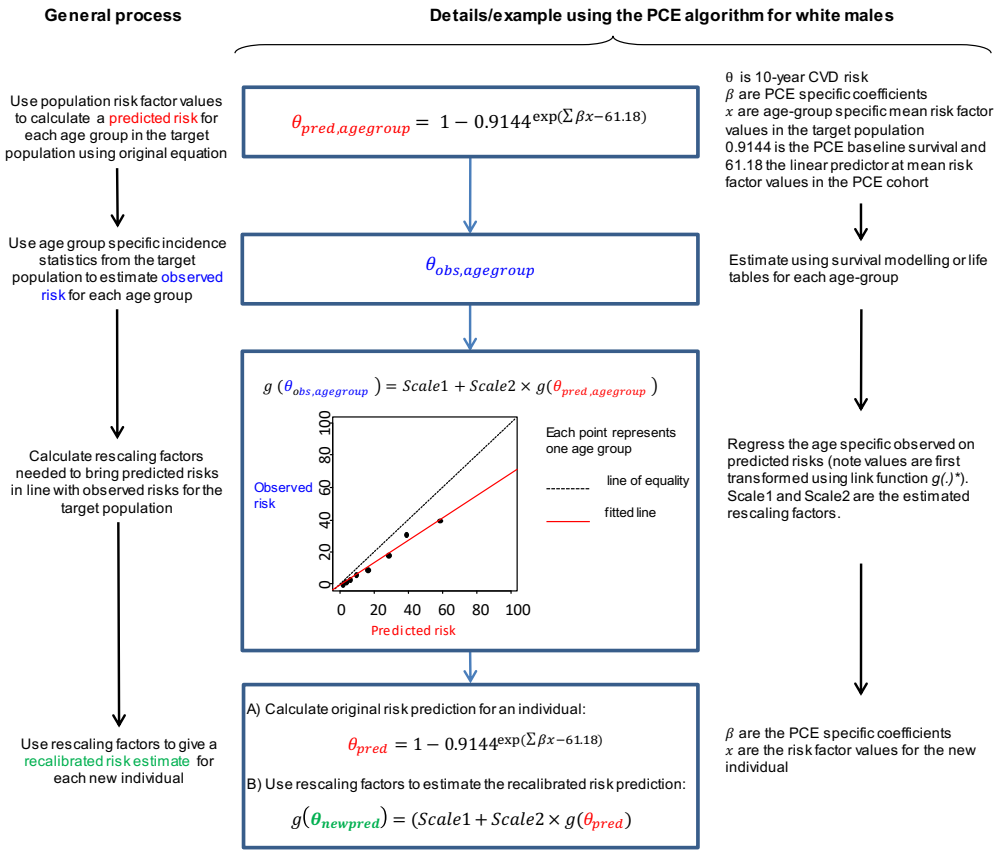


Figure 1.4: Methods used for recalibration of risk models.

Abbreviations: PCE, pooled cohort equations.

$g(.) = \ln(-\ln(1-.))$

Source: Pennells et al., 2019[83]

To recalibrate the estimated risks, population-level risk factor values are used with the risk model to calculate a predicted risk for each age group in the target population. The predicted risks are compared to age-group specific observed risks, estimated using observed CVD incidence rates in the target population. Rescaling factors are calculated by regressing the

observed risks against the predicted risks and the recalibrated risk estimates are obtained by applying the rescaling factors to the original risk estimates.

As summarising a country's average risk factor levels and incidence rates can prove challenging, and using cohorts to summarise data can be misleading due to cohorts often being non-representative of the general population, it is common to use summary data collated from large collaborations. These resources, including the Non-Communicable Disease Risk Factor Collaboration, and the Global Burden of Disease have modelled the average risk factor profile and incidence rates globally using a large collection of official statistics and cohort data.

Examples of recalibration have been demonstrated in research[83], and has been used in the development of the SCORE2, SCORE2-OP and the WHO CVD risk charts to recalibrate the model to multiple geographic regions.[34,35,84]

1.5.2.3 Dynamic population health modelling

Recently, efforts have been made to generalise population health modelling to estimate the health impacts of implementing primary prevention programmes, such as the NHS health check, dynamically and over a longer time period. As mentioned in **Section 1.5.1.4,** population health modelling often summarises the health impact at a single time point due to constraints in data. As such, comparing different strategies and changing variables including the frequency of assessments, and the age at which they are carried out, is limited.

By implementing a dynamic population health modelling approach, the impact over a wider timescale can be estimated. For example, a mixed-effects model could be used to estimate individual risk factor trajectories over time. This can then be used to simulate multiple risk assessments over time for each individual. As such, different interventions and strategies can be directly compared. Another is to create a microsimulation model which simulates the impact of interventions on individual trajectories structured around multiple health states and transition probabilities.[85–87]

1.6 Current research gaps

The aim of my thesis is to provide quantitative evidence of optimising invitation to formal CVD risk assessments in the context of current guidelines in England. The thesis focusses on three challenges:

1) Lack of recommended prioritisation model in England.

As discussed in **Section 1.3,** current guidelines in England recommend using existing primary care records for systematic prioritisation of individuals before a full formal CVD risk assessment. However, no such prioritisation tool exists or is recommended for use in current primary care systems. In addition, it is unknown how implementing a fixed risk threshold for both the prioritisation and formal risk assessment would impact the number of individuals prioritised for formal CVD risk assessments.

2) The unknown impact of including PRS within prioritisation

As discussed in **Section 1.4,** the majority of past research in CVD-based PRS has found that the inclusion of disease based PRS with conventional CVD risk factors can identify a greater number of future CVD events, improves the ability to discriminate between individuals with future events and non-events, and improves stratification of estimated risks. However, little is understood about the impact including PRS will have if used during both the prioritisation stage and at the formal assessment stage.

3) Lack of evidence on utilising PRS to personalise invitation and treatment strategies

As discussed in **Section 1.3,** guidelines in England recommend an approach of formally assessing all individuals every 5 years from the age of 40 onwards, using existing records to prioritise individuals. For the majority of younger individuals, and especially younger women, statins are unlikely to be prescribed to those between the ages of 40 and 50.

With the rise of PRS, proposals from large foundations and the UK government envision a future healthcare system that incorporates widespread genetic profiling.[88–90] PRS therefore has the ability to be implemented in ways to personalise predictive medicine, where invitation strategies can be tailored using PRS. Opportunities also exist where individuals are invited and treated using only genetically predicted information, potentially saving additional costs associated with measuring biomarkers at a formal assessment.

1.7 Rationale and aims of thesis

To summarise, the specific aims of the thesis are:

1) Develop a prioritisation tool and evaluate its public health impact comparing a fixed 10% threshold and age- and sex- specific thresholds;

2) Investigate the potential benefits of including CVD-based PRS when used for both prioritisation and formal CVD risk assessments;

3) Investigate novel, PRS based methods for invitation to a formal CVD risk assessment using a dynamic population health modelling approach.

To be able to achieve these aims, data from the UK Biobank cohort are used due to the unique structure of detailed risk factor information at study baseline, linked primary care records before and after baseline, genetic data required for PRS, and linkage with outcomes recorded in hospital and death registries for a large number of individuals. Data from the Clinical Practice Research Datalink (CPRD) will also be harnessed to supplement analysis, by using the data to estimate risk factor levels and CVD incidence rates representative of the general population.

As the linked primary care records in UK Biobank will play a pivotal role in achieving the aims of the thesis, data will first be analysed to understand the characteristics between the measurements collected at baseline and the measurements within the primary care records. Since current literature regarding this comparison is limited, the analysis will inform the appropriate approaches when modelling with the different data sources within UK Biobank.

The thesis will use a range of statistical techniques, including commonly used approaches to assess the predictive performance and using population health modelling to validate a risk model. The thesis will also use emerging methods to harness the repeated measurements exhibited in primary care data and to generalise the findings to the general population.

1.8 Outline of thesis

Chapter 2 describes the data structure of UK Biobank and linked data from Hospital Episode Statistics (HES) and Office for National Statistics (ONS). **Chapter 2** will assess whether differences exist between the risk factor measurements collected at baseline, and the measurements recorded in the primary care records. The aims, data, methods, results and discussions of the three aims of the thesis are outlined in **Chapters 3-5**. **Chapter 6** summarises the findings in a wider context, explaining potential clinical implications, overall strengths and limitations and potential future work.

References

1. Stewart J, Manmathan G, Wilkinson P. Primary prevention of cardiovascular disease: A review of contemporary guidance and literature. doi:10.1177/2048004016687211
2. Roth GA, Mensah GA, Johnson CO, et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study. *J Am Coll Cardiol*. 2020;76(25):2982-3021. doi:10.1016/J.JACC.2020.11.010
3. British Heart Foundation. Cardiovascular Disease Statistics 2017. Accessed November 29, 2022. https://www.bhf.org.uk/what-we-do/our-research/heart-statistics/heart-statistics-publications/cardiovascular-disease-statistics-2017
4. British Heart Foundation. Heart & Circulatory Disease Statistics 2022 - BHF. Accessed April 12, 2023. https://www.bhf.org.uk/what-we-do/our-research/heart-statistics/heart-statistics-publications/cardiovascular-disease-statistics-2022
5. European Heart Network. European Cardiovascular Disease Statistics 2017. Accessed November 29, 2022. https://ehnheart.org/cvd-statistics/cvd-statistics-2017.html
6. World Health Organization. Cardiovascular diseases: Avoiding heart attacks and strokes. Accessed November 30, 2022. https://www.who.int/news-room/questions-and-answers/item/cardiovascular-diseases-avoiding-heart-attacks-and-strokes
7. Andersson C, Vasan RS. Epidemiology of cardiovascular disease in young individuals. *Nat Rev Cardiol*. 2018;15(4):230-240. doi:10.1038/nrcardio.2017.154
8. World Health Organization. *Prevention of Cardiovascular Disease Guidelines for Assessment and Management of Cardiovascular Risk*.; 2007. Accessed February 20, 2020. www.inis.ie
9. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ*. 2016;353. doi:10.1136/bmj.i2416
10. Moons KGM, Kengne AP, Woodward M, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart*. 2012;98(9):683-690. doi:10.1136/HEARTJNL-2011-301246
11. Lloyd-Jones DM, Braun LT, Ndumele CE, et al. Use of Risk Assessment Tools to Guide Decision-Making in the Primary Prevention of Atherosclerotic Cardiovascular Disease: A Special Report From the American Heart Association and American College of Cardiology. *Circulation*. 2019;139(25):E1162-E1177. doi:10.1161/CIR.0000000000000638
12. Naik G, Ahmed H, Edwards AGK. Communicating risk to patients and the public. *Br J Gen Pract*. 2012;62(597):213. doi:10.3399/BJGP12X636236

13. Vision D. Risk communication: a pillar of shared decision making. *Prescriber*. 2022;33(6):24-28. doi:10.1002/PSB.1993
14. Zipkin DA, Umscheid CA, Keating NL, et al. Evidence-based risk communication: a systematic review. *Ann Intern Med*. 2014;161(4):270-280. doi:10.7326/M14-0295
15. Usher-Smith JA, Silarova B, Schuit E, Moons KGM, Griffin SJ. Impact of provision of cardiovascular disease risk estimates to healthcare professionals and patients: a systematic review. *BMJ Open*. 2015;5(10). doi:10.1136/BMJOPEN-2015-008717
16. Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation*. 1998;97(18):1837-1847. doi:10.1161/01.CIR.97.18.1837
17. Gomez G, Gomez-Mateu M, Dafni U. Informed Choice of Composite End Points in Cardiovascular Trials. *Circ Cardiovasc Qual Outcomes*. 2014;7(1):170-178. doi:10.1161/CIRCOUTCOMES.113.000149
18. Lin JS, Evans C V., Johnson E, Redmond N, Coppola EL, Smith N. Nontraditional risk factors in cardiovascular disease risk assessment: Updated evidence report and systematic review for the US preventive services task force. *JAMA - J Am Med Assoc*. 2018;320(3):281-297. doi:10.1001/jama.2018.4242
19. Scottish Intercollegiate Guidelines Network. SIGN 149 • Risk estimation and the prevention of cardiovascular disease. Published online 2017. Accessed January 21, 2023. www.nice.org.uk/accreditation
20. Woodward M, Brindle P, Tunsfall-Pedoe H. Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). *Heart*. 2007;93(2):172-176. doi:10.1136/HRT.2006.108167
21. Foundation NS, Lalor E, Boyden A, et al. Guidelines for the management of absolute cardiovascular disease risk. Published online January 1, 2012:124. Accessed January 19, 2023. http://strokefoundation.com.au/site/media/AbsoluteCVD_GL_webready.pdf
22. D'Agostino RB, Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: The Framingham heart study. *Circulation*. 2008;117(6):743-753. doi:10.1161/CIRCULATIONAHA.107.699579
23. Tai ES, Chia BL, Bastian AC, et al. Ministry of Health Clinical Practice Guidelines: Lipids. *Singapore Med J*. 2017;58(3):155-166. doi:10.11622/SMEDJ.2017018
24. Klug E, Raal F, Marais A, et al. South African dyslipidaemia guideline consensus statement: 2018 update A joint statement from the South African Heart Association (SA Heart) and the Lipid and Atherosclerosis Society of Southern Africa (LASSA). *S Afr*

Med J. Published online 2018.

25. Pearson GJ, Thanassoulis G, Anderson TJ, et al. 2021 Canadian Cardiovascular Society Guidelines for the Management of Dyslipidemia for the Prevention of Cardiovascular Disease in Adults. *Can J Cardiol*. 2021;37(8):1129-1150. doi:10.1016/J.CJCA.2021.03.016
26. Arnett DK, Blumenthal RS, Albert MA, et al. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation*. 2019;140(11):e596-e646. doi:10.1161/CIR.0000000000000678
27. Goff DC, Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: A report of the American college of cardiology/American heart association task force on practice guidelines. *Circulation*. 2014;129(25 SUPPL. 1):49-73. doi:10.1161/01.CIR.0000437741.48606.98/-/DC1
28. New Zealand Ministry of Health. Cardiovascular Disease Risk Assessment and Management for Primary Care 2018.
29. Pylypchuk R, Wells S, Kerr A, et al. Cardiovascular disease risk prediction equations in 400 000 primary care patients in New Zealand: a derivation and validation study. *Lancet*. 2018;391(10133):1897-1907. doi:10.1016/S0140-6736(18)30664-0
30. National Institute for Health and Care Excellence (NICE). Lipid modification: Cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease (clinical guideline CG181). Published online 2014.
31. Hippisley-Cox J, Coupland C, Vinogradova Y, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ*. 2008;336(7659):1475-1482. doi:10.1136/BMJ.39609.449676.25
32. 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*. 2017;357. doi:10.1136/bmj.j2099
33. Visseren FLJ, MacH F, Smulders YM, et al. 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice. *Eur Heart J*. 2021;42(34):3227-3337. doi:10.1093/EURHEARTJ/EHAB484
34. collaboration S working group and EC risk, Hageman S, Pennells L, et al. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. *Eur Heart J*. 2021;42(25):2439-2454. doi:10.1093/EURHEARTJ/EHAB309
35. collaboration S-O working group and EC risk, de Vries TI, Cooney MT, et al. SCORE2-

OP risk prediction algorithms: estimating incident cardiovascular event risk in older persons in four geographical risk regions. *Eur Heart J*. 2021;42(25):2455-2467. doi:10.1093/EURHEARTJ/EHAB312

36. Usher-Smith J, Mb MA, Mphil B. NHS Health Check Programme rapid evidence synthesis. Published online 2017.
37. Public Health England. Guidance overview: NHS Health Checks: applying All Our Health - GOV.UK. Accessed November 23, 2021. https://www.gov.uk/government/publications/nhs-health-checks-applying-all-our-health
38. NHS. NHS Health Check - National guidance. Accessed January 21, 2023. https://www.healthcheck.nhs.uk/commissioners-and-providers/national-guidance/
39. NHS Digital. Appointments in General Practice August 2021. Accessed November 24, 2021. https://digital.nhs.uk/data-and-information/publications/statistical/appointments-in-general-practice/august-2021
40. O'Flaherty M, Lloyd-Williams F, Capewell S, et al. Using the workHORSE model to explore and compare the effectiveness, cost-effectiveness and equity impact of different implementations of the NHS Health Check programme. Published online 2021. Accessed January 24, 2023. https://www.ncbi.nlm.nih.gov/books/NBK570866/
41. Collins GS, Altman DG. Predicting the 10 year risk of cardiovascular disease in the United Kingdom: independent and external validation of an updated version of QRISK2. *BMJ*. 2012;344:e4181. doi:10.1136/bmj.e4181
42. Pate A, Emsley R, Ashcroft DM, Brown B, Van Staa T. The uncertainty with using risk prediction models for individual decision making: An exemplar cohort study examining the prediction of cardiovascular disease in English primary care. *BMC Med*. 2019;17(1):134. doi:10.1186/s12916-019-1368-8
43. Paige E, Barrett J, Stevens D, et al. Landmark Models for Optimizing the Use of Repeated Measurements of Risk Factors in Electronic Health Records to Predict Future Disease Risk. *Am J Epidemiol*. 2018;187(7):1530-1538. doi:10.1093/AJE/KWY018
44. Paige E, Barrett J, Pennells L, et al. Use of Repeated Blood Pressure and Cholesterol Measurements to Improve Cardiovascular Disease Risk Prediction: An Individual-Participant-Data Meta-Analysis. *Am J Epidemiol*. 2017;186(8):899. doi:10.1093/aje/kwx149
45. Cifuentes M, Davis M, Fernald D, Gunn R, Dickinson P, Cohen DJ. Electronic Health Record Challenges, Workarounds, and Solutions Observed in Practices Integrating Behavioral Health and Primary Care. *J Am Board Fam Med*. 2015;28(Suppl 1):S63.

doi:10.3122/JABFM.2015.S1.150133

46. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic. *J Am Med Inform Assoc*. 2017;24(1):198-208. doi:10.1093/jamia/ocw042
47. Wells BJ, Nowacki AS, Chagin K, Kattan MW. Strategies for handling missing data in electronic health record derived data. *EGEMS (Washington, DC)*. 2013;1(3):7. doi:10.13063/2327-9214.1035
48. Wang Y, Hunt K, Nazareth I, Freemantle N, Petersen I. Do men consult less than women? An analysis of routinely collected UK general practice data. *BMJ Open*. 2013;3(8):e003320. doi:10.1136/BMJOPEN-2013-003320
49. Thompson AE, Anisimowicz Y, Miedema B, Hogg W, Wodchis WP, Aubrey-Bassler K. The influence of gender and other patient characteristics on health care-seeking behaviour: A QUALICOPC study. *BMC Fam Pract*. 2016;17(1):1-7. doi:10.1186/S12875-016-0440-0/TABLES/4
50. Price AL, Spencer CCA, Donnelly P. Progress and promise in understanding the genetic basis of common diseases. *Proceedings Biol Sci*. 2015;282(1821). doi:10.1098/RSPB.2015.1684
51. Levin MG, Rader DJ. Polygenic Risk Scores and Coronary Artery Disease. *Circulation*. 2020;141(8):637-640. doi:10.1161/CIRCULATIONAHA.119.044770
52. Inouye M, Abraham G, Nelson CP, et al. Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. *J Am Coll Cardiol*. 2018;72(16):1883-1893. doi:10.1016/j.jacc.2018.07.079
53. Abraham G, Malik R, Yonova-Doing E, et al. Genomic risk score offers predictive performance comparable to clinical risk factors for ischaemic stroke. *Nat Commun 2019 101*. 2019;10(1):1-10. doi:10.1038/s41467-019-13848-1
54. Sun L, Pennells L, Kaptoge S, et al. Polygenic risk scores in cardiovascular risk prediction: A cohort study and modelling analyses. Hindy G, ed. *PLOS Med*. 2021;18(1):e1003498. doi:10.1371/journal.pmed.1003498
55. Abraham G, Havulinna AS, Bhalala OG, et al. Genomic prediction of coronary heart disease. *Eur Heart J*. 2016;37(43):3267-3278. doi:10.1093/eurheartj/ehw450
56. Dudbridge F, Pashayan N, Yang J. Predictive accuracy of combined genetic and environmental risk scores. *Genet Epidemiol*. 2018;42(1):4-19. doi:10.1002/gepi.22092
57. Vaura F, Kauko A, Suvila K, et al. Polygenic Risk Scores Predict Hypertension Onset and Cardiovascular Risk. *Hypertension*. 2021;77(4):1119-1127. doi:10.1161/HYPERTENSIONAHA.120.16471

58. Wu H, Forgetta V, Zhou S, Bhatnagar SR, Paré G, Richards JB. Polygenic Risk Score for Low-Density Lipoprotein Cholesterol Is Associated With Risk of Ischemic Heart Disease and Enriches for Individuals With Familial Hypercholesterolemia. *Circ Genomic Precis Med*. 2021;14(1):E003106. doi:10.1161/CIRCGEN.120.003106

59. Khera A V., Chaffin M, Wade KH, et al. Polygenic prediction of weight and obesity trajectories from birth to adulthood. *Cell*. 2019;177(3):587. doi:10.1016/J.CELL.2019.03.028

60. Khera A V., Chaffin M, Aragam KG, et al. Genome-wide polygenic score to identify a monogenic risk-equivalent for coronary disease. *bioRxiv*. Published online November 15, 2017:218388. doi:10.1101/218388

61. Isgut M, Sun J, Quyyumi AA, Gibson G. Highly elevated polygenic risk scores are better predictors of myocardial infarction risk early in life than later. *Genome Med*. 2021;13(1):1-16. doi:10.1186/S13073-021-00828-8/FIGURES/6

62. Wand H, Lambert SA, Tamburro C, et al. Improving reporting standards for polygenic scores in risk prediction studies. *Nat 2021 5917849*. 2021;591(7849):211-219. doi:10.1038/s41586-021-03243-6

63. Cox DR. Regression Models and Life-Tables. *J R Stat Soc Ser B*. 1972;34(2):187-202. doi:10.1111/J.2517-6161.1972.TB00899.X

64. Kalbfleisch J, Prentice R. *The Statistical Analysis of Failure Time Data*.; 2011. Accessed January 27, 2023. https://books.google.com/books?hl=en&lr=&id=BR4Kq-a1MIMC&oi=fnd&pg=PR7&ots=xEqjcGQR6W&sig=nXh1z43rZdvcdCuZa_jnfXtS1Lw

65. Austin PC, Lee DS, Fine JP. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. *Circulation*. 2016;133(6):601-609. doi:10.1161/CIRCULATIONAHA.115.017719/-/DC1

66. Zhang Z. Survival analysis in the presence of competing risks. *Ann Transl Med*. 2017;5(3). doi:10.21037/ATM.2016.08.62

67. Austin PC, Fine JP. Practical recommendations for reporting Fine-Gray model analyses for competing risk data. *Stat Med*. 2017;36(27):4391. doi:10.1002/SIM.7501

68. Wolbers M, Koller MT, Stel VS, et al. Competing risks analyses: objectives and approaches. *Eur Heart J*. 2014;35(42):2936-2941. doi:10.1093/EURHEARTJ/EHU131

69. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology*. 2010;21(1):128. doi:10.1097/EDE.0B013E3181C30FB2

70. Hilden J, Habbema JDF, Bjerregaard B. The measurement of performance in

probabilistic diagnosis. II. Trustworthiness of the exact values of the diagnostic probabilities. *Methods Inf Med*. 1978;17(4):227-237. doi:10.1055/S-0038-1636442/ID/JR6442-20

71. Brier GW, Brier, W. G. Verification of Forecasts Expressed in Terms of Probability. *MWRv*. 1950;78(1):1. doi:10.1175/1520-0493(1950)078
72. Longato E, Vettoretti M, Di Camillo B. A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models. *J Biomed Inform*. 2020;108:103496. doi:10.1016/J.JBI.2020.103496
73. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15(4):361-387. doi:10.1002/(sici)1097-0258(19960229)15:4<361::aid-sim168>3.0.co;2-4
74. Xu Z, Arnold M, Stevens D, et al. Prediction of Cardiovascular Disease Risk Accounting for Future Initiation of Statin Treatment. *Am J Epidemiol*. 2021;190(10):2000-2014. doi:10.1093/AJE/KWAB031
75. Sweeting MJ, Barrett JK, Thompson SG, Wood AM. The use of repeated blood pressure measures for cardiovascular risk prediction: a comparison of statistical models in the ARIC study. *Stat Med*. 2017;36(28):4514-4528. doi:10.1002/sim.7144
76. Laird NM, Ware JH. Random-Effects Models for Longitudinal Data. *Biometrics*. 1982;38(4):963. doi:10.2307/2529876
77. Sweeting MJ, Thompson SG. Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biom J*. 2011;53(5):750-763. doi:10.1002/BIMJ.201100052
78. Barrett JK, Huille R, Parker R, Yano Y, Griswold M. Estimating the association between blood pressure variability and cardiovascular disease: An application using the ARIC Study. *Stat Med*. 2019;38(10):1855-1868. doi:10.1002/sim.8074
79. Teramukai S, Okuda Y, Miyazaki S, Kawamori R, Shirayama M, Teramoto T. Dynamic prediction model and risk assessment chart for cardiovascular disease based on on-treatment blood pressure and baseline risk factors. *Hypertens Res 2016 392*. 2015;39(2):113-118. doi:10.1038/hr.2015.120
80. Long JD, Mills JA. Joint modeling of multivariate longitudinal data and survival data in several observational studies of Huntington's disease. *BMC Med Res Methodol*. 2018;18(1):1-15. doi:10.1186/S12874-018-0592-9/FIGURES/5
81. Ibrahim JG, Chu H, Chen LM. Basic concepts and methods for joint models of longitudinal and survival dataIbrahim, J. G., Chu, H., & Chen, L. M. (2010, June 1).

Basic concepts and methods for joint models of longitudinal and survival data. Journal of Clinical Oncology, Vol. 28, pp. 2. *J Clin Oncol*. 2010;28(16):2796-2801. doi:10.1200/JCO.2009.25.0654

82. Suresh K, Taylor JMG, Spratt DE, Daignault S, Tsodikov A. Comparison of joint modeling and landmarking for dynamic prediction under an illness-death model. *Biometrical J*. 2017;59(6):1277-1300. doi:10.1002/bimj.201600235
83. Pennells L, Kaptoge S, Wood A, et al. Equalization of four cardiovascular risk algorithms after systematic recalibration: Individual-participant meta-analysis of 86 prospective studies. *Eur Heart J*. 2019;40(7):621-631. doi:10.1093/eurheartj/ehy653
84. Kaptoge S, Pennells L, De Bacquer D, et al. World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions. *Lancet Glob Heal*. 2019;7(10):e1332-e1345. doi:10.1016/S2214-109X(19)30318-3
85. ten Haaf K, Tammemägi MC, Bondy SJ, et al. Performance and Cost-Effectiveness of Computed Tomography Lung Cancer Screening Scenarios in a Population-Based Setting: A Microsimulation Modeling Analysis in Ontario, Canada. *PLoS Med*. 2017;14(2). doi:10.1371/JOURNAL.PMED.1002225
86. Kypridemos C, Collins B, McHale P, et al. Future cost-effectiveness and equity of the NHS Health Check cardiovascular disease prevention programme: Microsimulation modelling using data from Liverpool, UK. Sheikh A, ed. *PLOS Med*. 2018;15(5):e1002573. doi:10.1371/journal.pmed.1002573
87. Krijkamp EM, Alarid-Escudero F, Enns EA, Jalal HJ, Hunink MGM, Pechlivanoglou P. Microsimulation modeling for health decision sciences using R: a tutorial. *Med Decis Making*. 2018;38(3):400. doi:10.1177/0272989X18754513
88. PHG. Implementing polygenic scores for cardiovascular disease into NHS Health Checks. Published 2021. Accessed April 25, 2023. https://www.phgfoundation.org/report/prs-implementation-and-delivery
89. Secretary PU, Health P, Care P, Majesty H, July P. Advancing our health: prevention in the 2020s. 2019;(July). Accessed January 29, 2023. https://www.gov.uk/government/consultations/advancing-our-health-prevention-in-the-2020s
90. GOV.UK. Genome UK: 2022 to 2025 implementation plan for England - GOV.UK. Accessed May 3, 2023. https://www.gov.uk/government/publications/genome-uk-2022-to-2025-implementation-plan-for-england/genome-uk-2022-to-2025-implementation-plan-for-england

Chapter 2

Concordance of primary care records in cohorts – an exploration in UK Biobank

Chapter summary

Objective: UK Biobank, with its unique multi-modal dataset of detailed risk factor information at study baseline, and linked primary care records, will play a pivotal role in deriving models using both sources of data. This chapter will explore the characteristics of UK Biobank, by analysing the within-person level concordance of commonly recorded chronic disease risk predictors in primary care records with self-reported and recorded baseline cohort data.

Methods: 176,220 individuals aged between 40 and 70 years in UK Biobank, a UK-based population-based cohort with recruitment between 2006 and 2010, with at least one primary care record and complete baseline measurements, were used. For continuous risk factors, we quantified the differences between the cohort-baseline measurements and (i) last observed primary care records and (ii) predicted values from univariate mixed-effects models utilising repeated measures of primary care records. Agreement in categorical risk factors was assessed using simple agreement and kappa statistics between the cohort-baseline measure and the last observed primary care record.

Results: In primary care records, women had more repeat measurements than men. BMI, creatinine, HDL cholesterol, systolic blood pressure and total cholesterol were all commonly recorded, with the percentage of at least one risk factor measurement in the primary care records exceeding 90% in individuals aged between 60 and 70 years. Systolic blood pressure was the most commonly recorded risk factor, with over 70% of both men and women having at least 2 SBP measurements across all ages. Primary care continuous measurements for BMI, glucose and total cholesterol were on average lower than the UK Biobank cohort-baseline measurements, and higher for HDL cholesterol, creatinine, HbA1c and glucose. Systolic blood pressure measurements were consistently lower in primary care records than measured at the UK Biobank cohort-baseline, with differences increasing with the underlying blood pressure measurements. A mean difference of -0.05 and -0.20 standard deviations was observed in the

youngest men and women respectively which increased to -0.50 and -0.66 standard deviations as age increased respectively. Agreement in categorical risk factors, including smoking, diabetes, atrial fibrillation and rheumatoid arthritis was high (>90%) regardless of time since the last observed primary care record.

Conclusion: There were high levels of concordance for most chronic disease markers between primary care records and cohort measurements. Systolic blood pressure measures were generally lower in primary care records compared to cohort baseline measurements, likely due to a combination of guidelines and differences in measurement methodologies. Post-hoc adjustments should be considered if using linked primary care records with cohort data.

2.1 Introduction

Generally, electronic health records contain a wealth of longitudinal, real-world patient data that includes demographics, diagnoses, biomarker measurements and medication history. Along with facilitating the risk assessment and management of disease[1,2], electronic health records have facilitated researchers to conduct observational research for over 30 years, such as developing risk scores in large samples with similar characteristics to the intended target population[3,4], to enhance disease identification[5,6] and to accelerate genomic discovery[7–9]. More recently, the linkage of primary care records with traditional cohorts have allowed cohorts to improve the breadth of the data available for research and offers a longitudinal view of an individual's health status before and after study entry.

Whilst the benefits of linking primary care records with cohorts exists, challenges remain in its implementation, including selection biases, noise, missing and sporadic data, and unmeasured confounding[10]. Recent research has aimed to better utilise the repeated measurements using risk factor specific algorithms designed to integrate primary care records into cohorts, such as for diabetes and alcohol.[11,12] However, quantitative evidence of the similarity for a range of common risk factors in primary care measurements and cohort-baseline measurements in the same population is limited.

As the linked primary care records in UK Biobank (UKB) will be used extensively throughout the thesis to achieve its aims, the UKB data will first be analysed to understand the characteristics between the measurements collected at baseline and the measurements within the primary care records (see **Chapter 1, Section 1.7)**

In this chapter, we will investigate the concordance of traditional chronic disease risk factor measurements recorded in the linked primary care records from multiple providers and at cohort-baseline measurements of UKB by evaluating the differences and agreement within the same individuals.

2.2 Methods

2.2.1 Data source

The UKB study is a large prospective cohort study of over 500,000 individuals aged between 40 and 69 years in the United Kingdom, and is now a "*large scale biomedical database and research resource, containing in-depth genetic and health information"*. Study participants were recruited across 22 centres in England, Wales and Scotland between 2006 and 2010. Over 9 million individuals were invited, resulting in an overall participation rate of 5.45%. At study entry, a baseline survey was conducted, consisting of a detailed questionnaire, physical measurements and biological measurements. Further follow-up assessments have been conducted since the study started, with repeated biological measurements and genotyping.

Primary care record data (from The Phoenix Partnership (TPP), Egton Medical Information Systems (EMIS) and Vision GP system suppliers) have been linked to approximately 230,000 (45%) of the UKB cohort. The data available in the primary care records includes primary care events recorded, including consultations, diagnoses, symptoms and laboratory tests. The largest provider linked with UKB is the TPP system, with 165,000 individuals from England. Different clinical coding classification systems were used for each system, with TPP using Clinical Terms Version 3 (CTV3) and Read v3, and EMIS and Vision using the Read v2 coding system. Prescription data is also available in the linked primary care records, and is coded using a combination of the British National Formulary, Read v2 and Dictionary of Medicines and Devices systems.

The study is linked to secondary care admissions from Hospital Episode Statistics (HES) and mortality records from the Office for National Statistics (ONS). HES is a database containing details about admissions, accident and emergency attendances, and outpatient appointments in NHS hospitals in England.[13] ONS mortality records contains details related to an individual's death, taken from the death certificate, for all deaths recorded in England and Wales.[14] Linking with both HES and ONS allows deaths to be captured within and outside of hospitals.

Generally, UKB has been shown to be not representative of the general population, with evidence of a health-volunteer selection bias.[15] UKB participants were more likely to be older, female and less socioeconomically deprived. They were also generally healthier, with participants less likely to be obese, smokers, drink alcohol on a daily basis, have fewer self-reported health conditions and have lower mortality rates. However, additional research has shown that despite these differences, risk factor associations for mortality due to CVD were generalisable to the general population.[16]

At the time of writing, 176,888 individuals were extracted for use in the analysis. In addition, we restricted to using records from between the 1st April 2004, the introduction of the Quality and Outcomes Framework (QOF), and 5th July 2019, the end of data linkage.

2.2.2 Risk factors

2.2.2.1 Primary care records

Conventional and commonly recorded chronic disease risk factors in primary care were pre-selected for comparison against the cohort-baseline measurements, and included: systolic blood pressure (SBP) (mmHg), total cholesterol (mmol/litre) and high-density lipoprotein (HDL) cholesterol (mmol/litre), body mass index (BMI) (kg/m2), glucose (mmol/litre), creatinine (mmol/litre), HbA1c (mmol/mol), diabetes mellitus (ever or never diagnosed), smoking status (current or non-current smoker), hypertension treatment (ever or never), statin treatment (ever or never), rheumatoid arthritis (ever or never diagnosed), chronic kidney disease stages 4 or 5 (ever or never diagnosed) and atrial fibrillation (ever or never diagnosed).

2.2.2.2. Cohort-baseline measurements

Risk factors in UKB were collected by a standardised questionnaire and with standardised equipment. SBP was measured using two automatic readings and the mean value was recorded[17]. Total cholesterol, HDL cholesterol, glucose and creatinine were measured using blood samples and a Beckman Coulter AU5800 analyser, and HbA1c was Bio-Rad VARIANT II Turbo haemoglobin testing system[18–22]. Diabetes status, smoking status, hypertension treatment, statin treatment, rheumatoid arthritis and chronic kidney disease were recorded using a self-reported information.

2.2.3 Statistical analysis

To compare measurements of systolic blood pressure, total cholesterol, HDL cholesterol, BMI, creatinine, glucose and HbA1c in primary care records with those recorded at UKB cohort-baseline survey, we performed two analyses. First, we took the last observed primary care measurement before the cohort-baseline. Second, we used sex-specific univariate mixed-effects models with the longitudinal primary care records, described below, to estimate the expected risk factor level at the time of the cohort-baseline survey (**Figure 2.1**). Individual level differences between the cohort-baseline measurements and the last observed primary care measurement, and the cohort-baseline measurements and univariate mixed-effects model estimates were summarised with mean differences and Bland Altman plots.

For each risk factor, separate sex-specific univariate mixed-effects models were fitted on the longitudinal primary care measurements. For each risk factor, each individual had a measurement at the cohort-baseline survey, and a minimum of two primary care measurements, with at least one measurement observed before the cohort-baseline survey. Each sex-specific model included fixed and random-intercepts, and linear and quadratic fixed-effects for age at cohort-baseline survey and the number of years the primary care measurement was made before or after the cohort-baseline survey, with 0 years being equivalent to the cohort-baseline survey. For example, BMI was modelled such that:

$$BMI_{ij} = a_1 + \left(b_1 * time_{ij}\right) + \left(c_1 * time_{ij}^2\right) + \left(d_1 * age_i\right) + \left(e_1 * age_i^2\right) + u_i + \varepsilon_{ij}$$

for $i = 1 \ldots N, j = 1 \ldots M_i, u_i \sim N(0, \sigma_u^2)$ and $\varepsilon_{ij} \sim N(0, \sigma_e^2)$, where N is defined as the number of individuals included in the model, and M_i is defined as the total number of longitudinal primary care measurements observed, for individual i. BMI_{ij} denotes the repeated BMI measurements for individual i and measurement j, $time_{ij}$ denotes the number of years the primary care measurement was made before or after the cohort-baseline survey, for individual i and measurement j, and age_i denotes the age at cohort-baseline survey in years for individual i. The parameters a_1, b_1 , c_1, d_1 and e_1 represents fixed coefficients, u_i represents the random intercept and is normally distributed with variance σ_u^2, and represents the difference in the individual average risk factor level above the population average risk factor level, and e_{ij} represents the uncorrelated residual errors.

Best linear unbiased predictors (BLUPS) can be estimated for each risk factor from the random intercept u_{I} and at $time_{ij}$=0.[23,24] When modelling systolic blood pressure and total cholesterol, an additional interaction between systolic blood pressure and a time-varying status of antihypertensive medication, and total cholesterol and a time-varying status of statin medication was included such that:

$$SBP_{ij} = a_2 + \left(b_2 * time_{ij}\right) + \left(c_2 * time_{ij}^2\right) + \left(d_2 * age_i\right) + \left(e_2 * age_i^2\right) + \left(f_2 * AHM_{ij}\right) + u_i + \varepsilon_{ij}$$

$$Total\ cholesterol_{ij} = a_3 + \left(b_3 * time_{ij}\right) + \left(c_3 * time_{ij}^2\right) + \left(d_3 * age_i\right) + \left(e_3 * age_i^2\right) + \left(f_3 * statin_{ij}\right) + u_i + \varepsilon_{ij}$$

For $i = 1 \ldots N, j = 1 \ldots M_i$, $u_i \sim N(0, \sigma_u^2)$ and $\varepsilon_{ij} \sim N(0, \sigma_e^2)$

Where, in addition to the previously mentioned risk factors and coefficients, we let AHM_{ij} and $statin_{ij}$ denote whether anti-hypertensive medication and statins had been prescribed prior to time ij, and let f_2 and f_3 represent fixed coefficients.

For categorical variables (ever reported diabetes mellitus, smoking status, rheumatoid arthritis, chronic kidney disease, atrial fibrillation and family history of CVD disease), the simple agreement and Cohen's Kappa statistic[25] were used to summarise the agreement between primary care records and self-reported baseline values in UKB. We summarised the agreement by grouping individuals based on the time between the last observed primary care measurement and the self-reported values. We grouped individuals into those without a positive record by cohort-baseline, and then included those with a record in the past 6 months, 1 year, 3 years, 5 years and 10 years. All analyses were stratified by sex.

Data were cleaned and analysed with R x64 3.6.1.

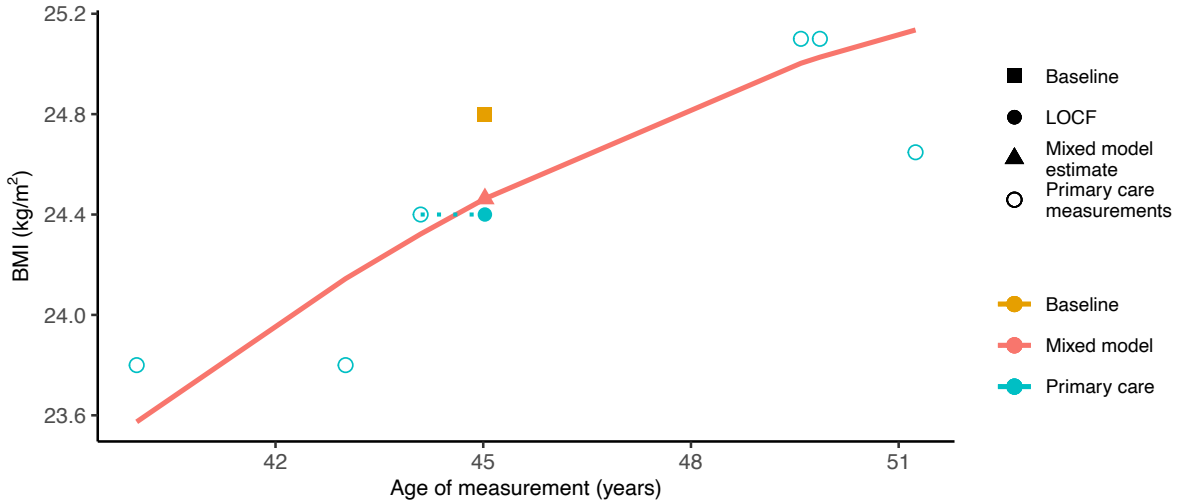


Figure 2.1: Diagram illustrating comparison between primary care and UK Biobank baseline measurements, with the last observed measurement and expected levels from the univariate mixed effects model, shown for BMI in an example individual.

Abbreviations: BMI, body mass index; LOCF, last observed carried forward

2.3 Results

2.3.1 Baseline characteristics

Of the 502,511 individuals recruited in UKB, 176,888 individuals had at least one linked primary care record. We excluded 664 individuals who did not have any primary care records after excluding records before the 1st April 2004 and two individuals who entered UKB outside the intended study age range of between 40 and 70 years. Overall, 176,220 individuals were included in the main analysis dataset **(Figure 2.2).** Of the 176,220 individuals, 45% were men and 55% were women. Generally, compared to individuals without a linked primary care record, those with linked primary care records were marginally older, had marginally elevated risk factor levels for SBP, were more likely to be non-smokers, and a greater proportion were a current user of antihypertensive medication, had rheumatoid arthritis, and atrial fibrillation **(Table 2.1a-2.1c)**. Risk factor values at cohort-baseline survey were generally higher at older ages for the majority of risk factors **(Figure 2.3).** Cohort-baseline risk factor values had a wider spread (i.e., larger standard deviations) at older ages for SBP and HDL cholesterol, but narrower spread at older ages for BMI **(Figure 2.4).**

BMI, creatinine, HDL cholesterol, SBP and total cholesterol were commonly recorded in the linked primary care records, with the percentage of at least one risk factor measurement in the

primary care records exceeding 90% in men and women aged between 60 and 70 years at the cohort-baseline **(Figure 2.5 and Figure 2.6).** SBP was the most measured risk factor, with over 70% of both men and women having at least 2 SBP measurements across all ages. Smoking status was the most commonly recorded categorical risk factor, which was consistently recorded in over 75% of men and women between 40 years and 70 years.

The maximum percentage of at least one risk factor measurement, in men and women, for all cohort-baseline ages were low for diabetes (14.8% and 7.2%), rheumatoid arthritis (1.4% and 1.7%), chronic kidney disease (1.6% and 1.0%) and atrial fibrillation (10.3% and 8.2%).

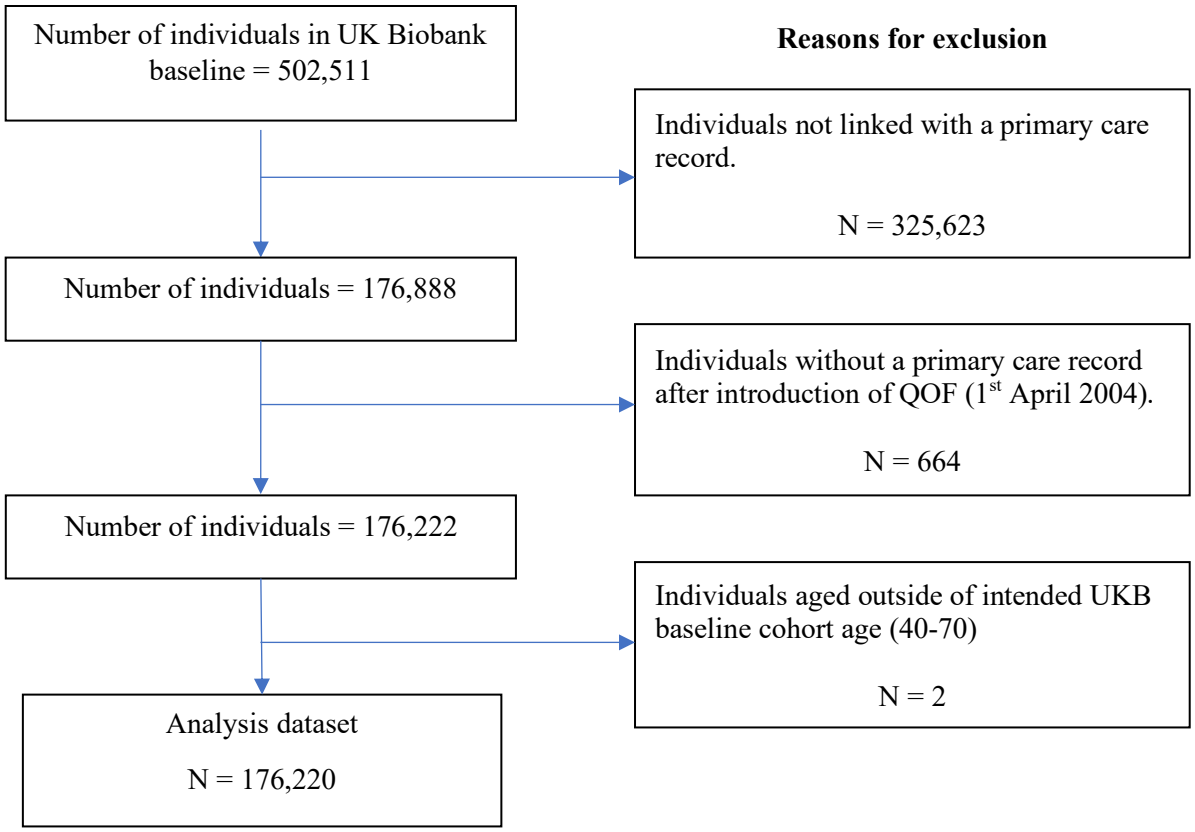


Figure 2.2: Flowchart showing selection of patient records in UK Biobank

Abbreviations: QOF, Quality and Outcomes Framework; UKB, UK Biobank

Table 2.1a: Key baseline characteristics of men in UK Biobank with and without linked primary care records

Characteristic	Linked men, N = 80,075 (45%)	Unlinked men, N = 149,044 (46%)
Age, mean (SD)	57.5 (8.1)	57.1 (8.2)
Ethnicity — White, N (%)	76,080 (95.5%)	139,178 (94.1%)
Townsend, mean (SD)	-1.4 (3.0)	-1.2 (3.2)
Systolic blood pressure – mmHg, mean (SD)	141.4 (17.6)	140.3 (17.4)
Total cholesterol – mmol/litre, mean (SD)	5.48 (1.14)	5.48 (1.12)
HDL cholesterol – mmol/litre, mean (SD)	1.28 (0.31)	1.28 (0.31)
BMI – kg/m^2, mean (SD)	27.9 (4.3)	27.8 (4.2)
Smoking status – current/ever smoker, N (%)	9,713 (12.2%)	18,914 (12.7%)
History of diabetes, N (%)	5,609 (7.0%)	10,407 (7.0%)
Antihypertensive medication – current user, N(%)	20,165 (25.5%)	35,914 (24.7%)
Rheumatoid arthritis, N (%)	795 (0.99%)	1,303 (0.87%)
Atrial fibrillation, N (%)	2,277 (2.84%)	3,533 (2.37%)

Abbreviations: BMI, body mass index; HDL, high density lipoprotein; SD, standard deviation

Table 2.1b: Key baseline characteristics of women in UK Biobank with and without linked primary care records

Characteristic	Linked women, N = 96,145 (55%)	Unlinked women, N = 177,247 (54%)
Age, mean (SD)	57.0 (7.9)	56.8 (8.0)
Ethnicity — White, N (%)	91,797 (95.8%)	165,648 (93.9%)
Townsend, mean (SD)	-1.4 (2.9)	-1.3 (3.1)
Systolic blood pressure – mmHg, mean (SD)	135.8 (19.3)	134.6 (19.2)
Total cholesterol – mmol/litre, mean (SD)	5.90 (1.14)	5.86 (1.12)
HDL cholesterol – mmol/litre, mean (SD)	1.59 (0.38)	1.60 (0.38)
BMI – kg/m^2, mean (SD)	27.2 (5.2)	27.0 (5.2)
Smoking status – current/ever smoker, N (%)	8,432 (8.8%)	15,942 (9.0%)
History of diabetes, N (%)	3,307 (3.5%)	6,343 (3.6%)
Antihypertensive medication – current user, N(%)	17,322 (18.1%)	30,596 (17.6%)
Rheumatoid arthritis, N (%)	1,687 (1.75%)	2,972 (1.68%)
Atrial fibrillation, N (%)	984 (1.02%)	1,542 (0.87%)

Abbreviations: BMI, body mass index; HDL, high density lipoprotein; SD, standard deviation

Table 2.1c: Key baseline characteristics of all individuals in UK Biobank

Characteristic	Men, N = 229,119 (46%)	Women, N = 273,392 (54%)
Age, mean (SD)	57.2 (8.2)	56.8 (8.0)
Ethnicity — White, N (%)	215,258 (94.6%)	257,445 (94.6%)
Townsend, mean (SD)	-1.2 (3.2)	-1.3 (3.0)
Systolic blood pressure – mmHg, mean (SD)	140.7 (17.5)	135.1 (19.2)
Total cholesterol – mmol/litre, mean (SD)	5.48 (1.13)	5.87 (1.13)
HDL cholesterol – mmol/litre, mean (SD)	1.28 (0.31)	1.59 (0.38)
BMI – kg/m^2, mean (SD)	27.8 (4.2)	27.1 (5.2)
Smoking status – current/ever smoker, N (%)	28,627 (12.5%)	24,374 (8.9%)
History of diabetes, N (%)	16,016 (7.0%)	9,650 (3.5%)
Antihypertensive medication – current user, N(%)	56,079 (25.0%)	47,918 (17.8%)
Rheumatoid arthritis, N (%)	2,098 (0.92%)	4,659 (1.70%)
Atrial fibrillation, N (%)	5,810 (2.54%)	2,526 (0.92%)

Abbreviations: BMI, body mass index; HDL, high density lipoprotein; SD, standard deviation

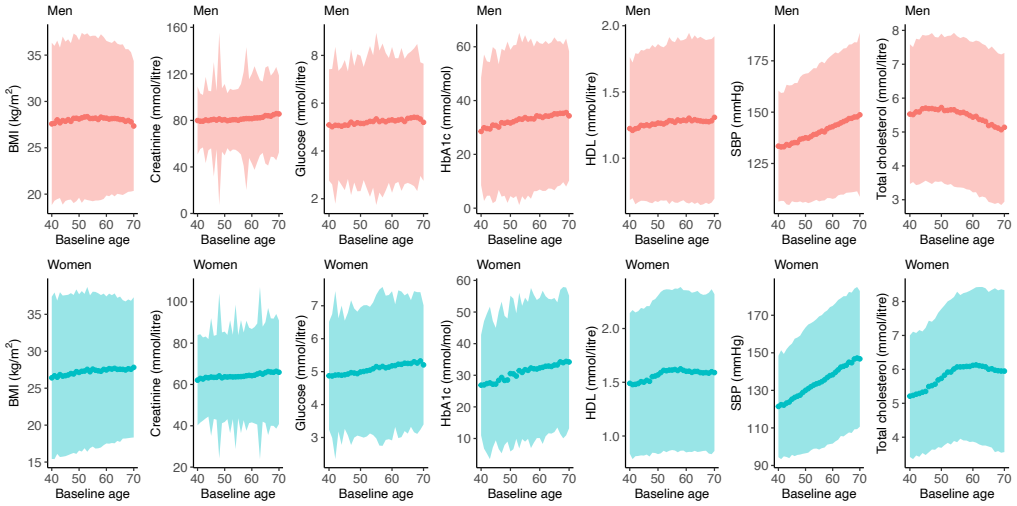


Figure 2.3: Mean risk factor measurements at UK Biobank baseline survey by age and sex

Abbreviations: BMI, body mass index; HDL, high density lipoprotein; SBP, systolic blood pressure

Individuals were restricted to those with both a historical primary care record measurement and complete data at baseline for each risk factor. Shaded region represents +/- 2SD from the mean.

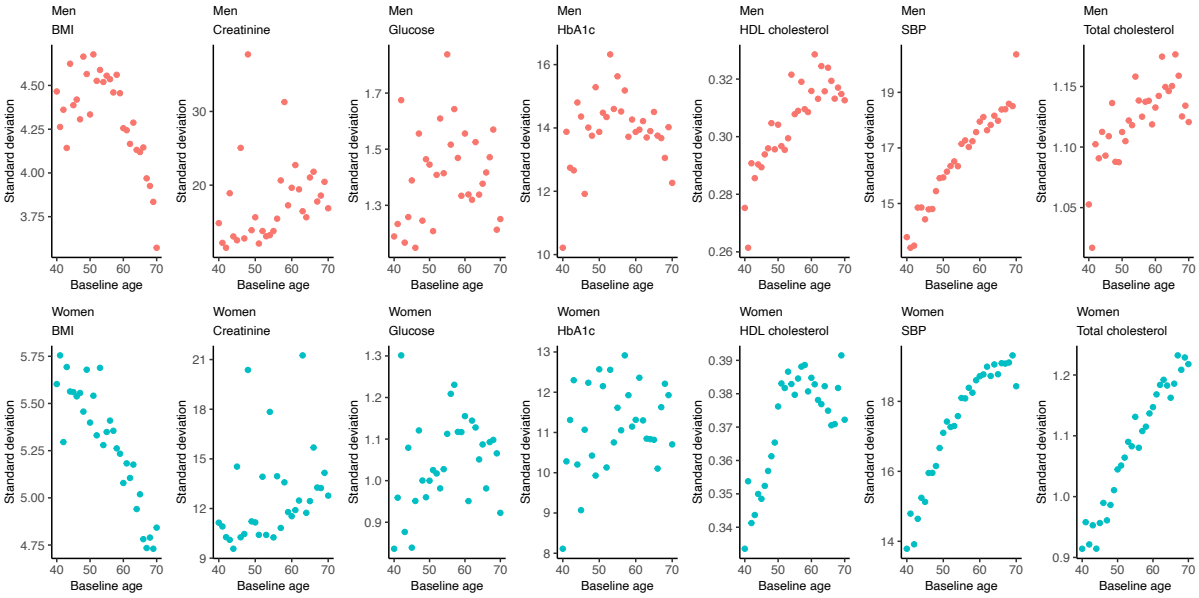


Figure 2.4: Age and sex-specific standard deviations of risk factors at UK Biobank baseline survey.

Abbreviations: BMI, body mass index; HDL, high density lipoprotein; SBP, systolic blood pressure

Individuals were restricted to those with both a historical primary care record measurement and complete data at baseline for each risk factor.

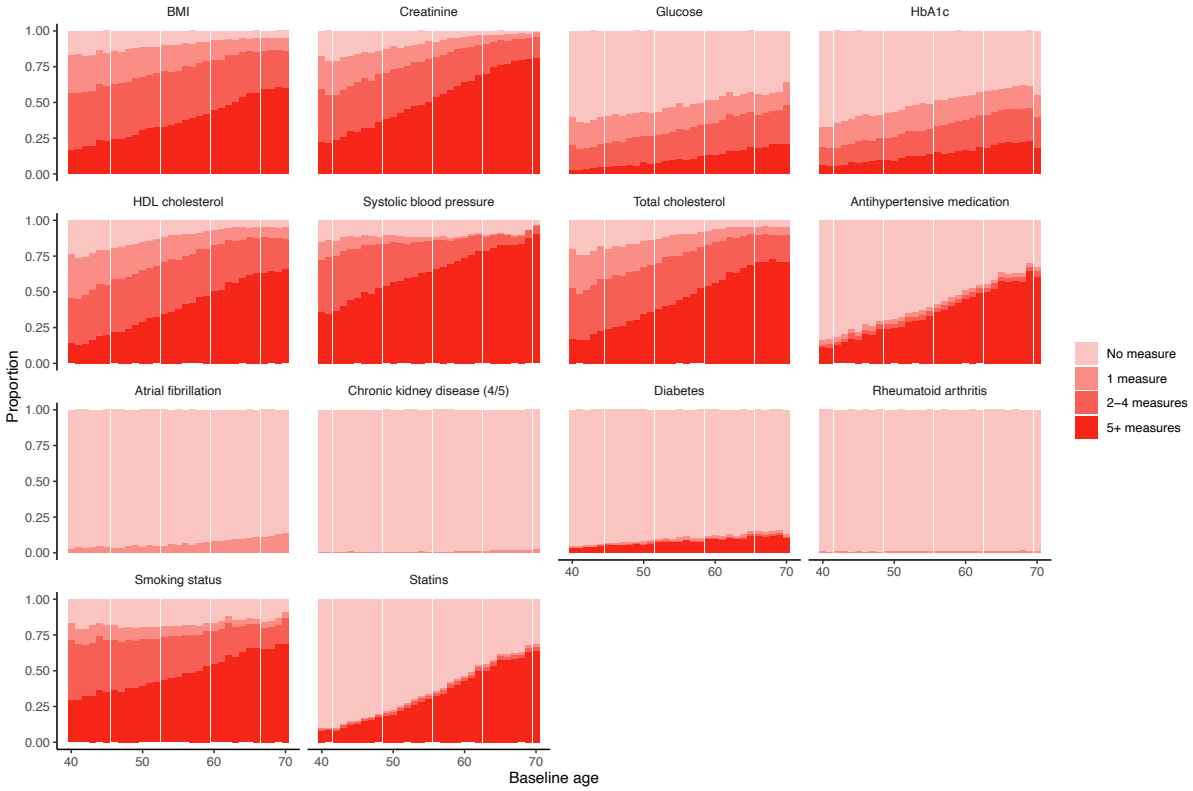


Figure 2.5: Distribution of observed risk factor measures in primary care records in UK Biobank individuals among men

Abbreviations: BMI, body mass index; HDL, high density lipoprotein.

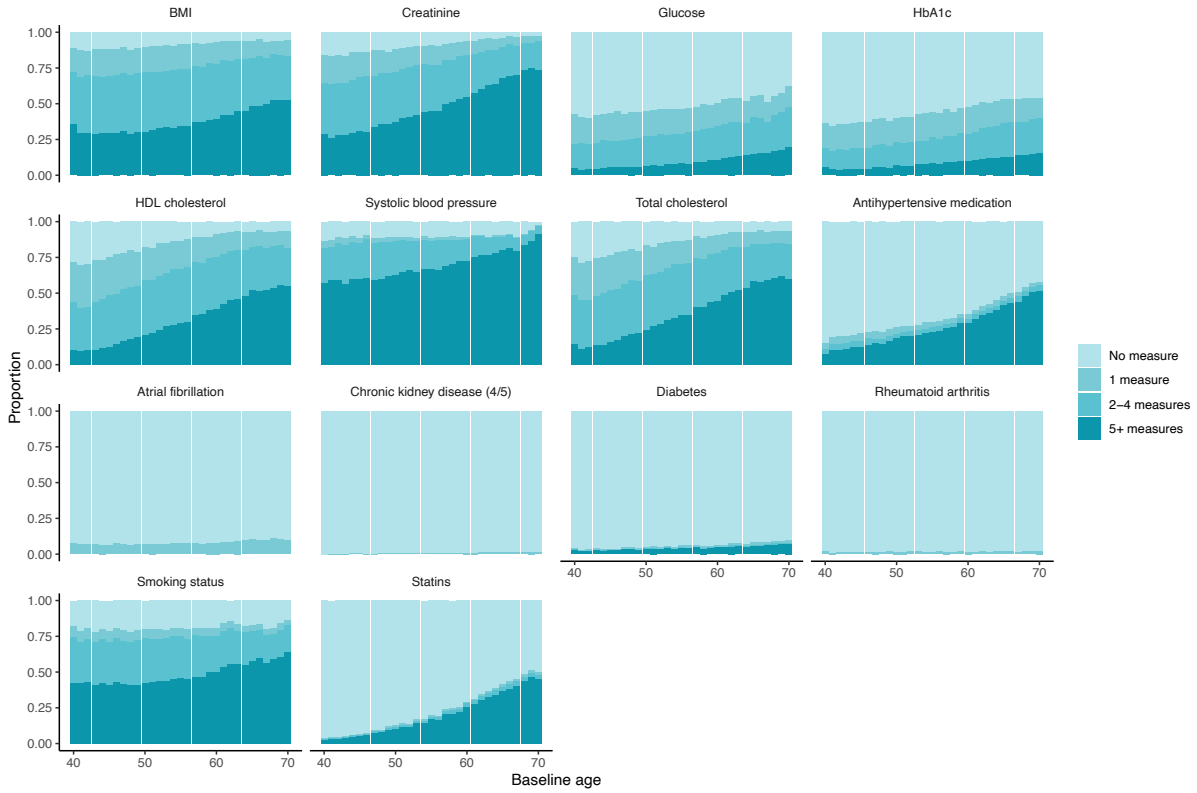


Figure 2.6: Distribution of observed risk factor measures in primary care records in UK Biobank individuals among women

Abbreviations: BMI, body mass index; HDL, high density lipoprotein.

Figure drops one individual aged 71 at cohort-baseline for presentation purposes.

2.3.2 Agreement in continuous risk factors

In individuals with a minimum of two primary care risk factor measurements, compared to the measurements observed at the cohort-baseline, age- and sex-specific mean differences using either the last observed measurement or the mixed effects estimation showed generally lower values for BMI, SBP and total cholesterol and generally higher values for HDL cholesterol, creatinine, HbA1c and glucose levels **(Figure 2.7)**.

As age increased, the differences between the cohort-baseline measurements and the primary care measurements increased (i.e., attenuated away from zero-difference) for BMI, creatinine, SBP and total cholesterol. In contrast, the differences for HbA1c became smaller as age increased (i.e., attenuated towards zero-difference), and the differences remained consistent across all ages for glucose and HDL cholesterol. Generally, the magnitude of the mean differences was smaller when using expected levels from the mixed effects model compared to using the last observed measurement. Comparing standardised mean differences for each risk factor shows that the expected levels from the mixed effects model for SBP had the greatest amount of variability across the full age range. We observed a mean difference of -0.05 standard deviations in the youngest women, with the mean difference increasing to -0.50 standard deviations as age increased. Similar results were observed in men, with a mean difference of -0.20 standard deviations in the youngest men, increasing to -0.66 standard deviations in the oldest men (**Figure 2.8)**.

Bland Altman plots, where the absolute differences between the cohort-baseline measurements and the expected levels from the mixed effects model is plotted against the arithmetic mean of the two for each individual, showed that the differences varied with the underlying risk factor level. **(Figure 2.9)**. For SBP, the expected levels from the mixed effects were greater than the cohort-baseline measurement when the average between the two were low. However, when the average increased, i.e., when the underlying SBP increased, the expected levels from the mixed effects were lower than the cohort-baseline measurement, and this difference would increase as the average increased. This pattern was also observed for total cholesterol.

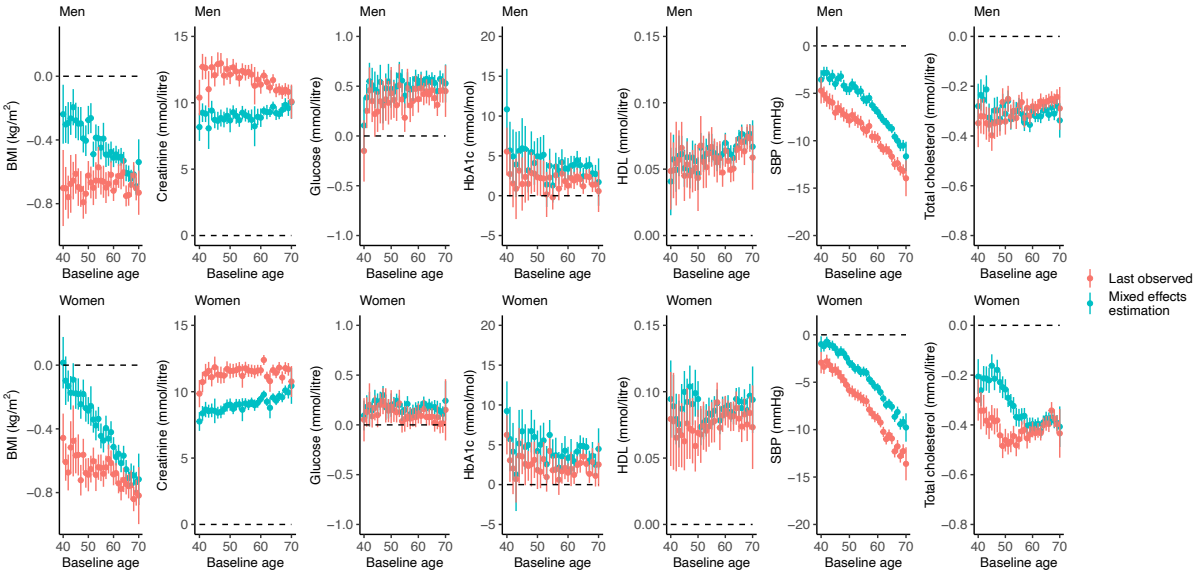


Figure 2.7: Mean differences between continuous chronic disease risk factors in primary care records and at UK Biobank baseline by age and sex

Abbreviations: BMI, body mass index; HDL, high density lipoprotein.

95% confidence interval are represented by vertical lines. First row uses the last observed measurement and the second row uses a sex and risk factor specific univariate mixed model. Negative/positive values indicate primary care record estimation is lower/greater than baseline measurement. Individuals were included for the risk-factor specific comparison if an individual had a measurement at baseline, and a minimum of two primary care measurements, with at least one measurement before the cohort-baseline survey.

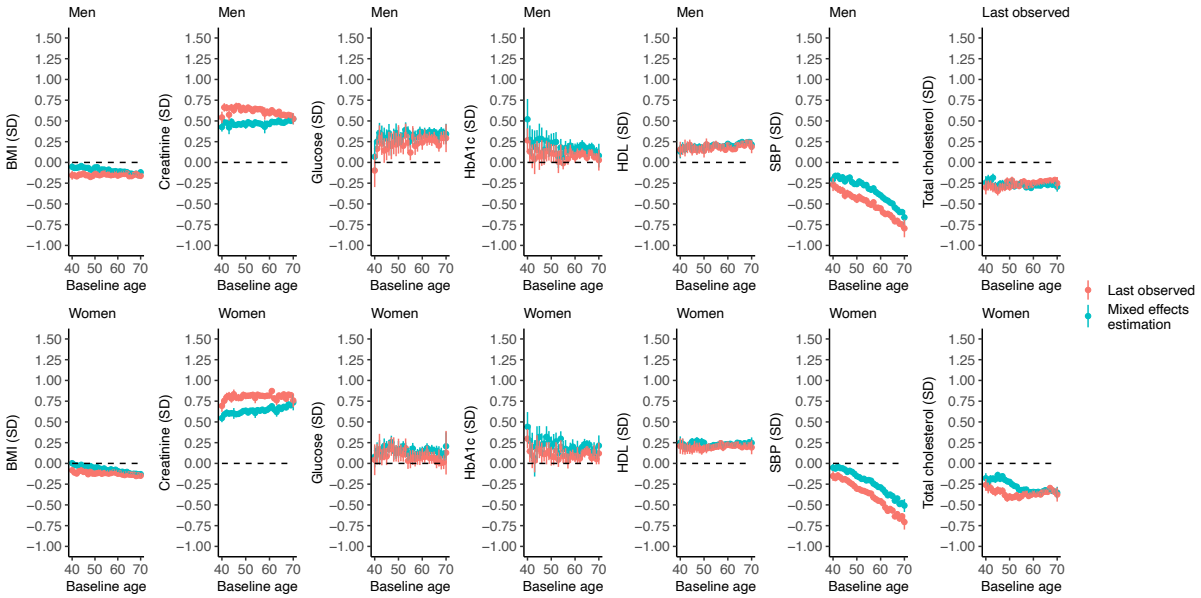


Figure 2.8: Mean standardised differences between continuous risk factors in primary care records and at UK Biobank baseline by age and sex

Abbreviations: BMI, body mass index; HDL, high density lipoprotein; SD, standard deviation

95% confidence interval are represented by vertical lines. First row uses the last observed measurement and the second row uses a sex and risk factor specific univariate mixed model. Negative/positive values indicate primary care record estimation is lower/greater than baseline measurement. Measurements were standardised using age, sex and exposure specific mean and standard deviations. Individuals were included for the risk-factor specific comparison if an individual had a measurement at baseline, and a minimum of two primary care measurements, with at least one measurement before the cohort-baseline survey.

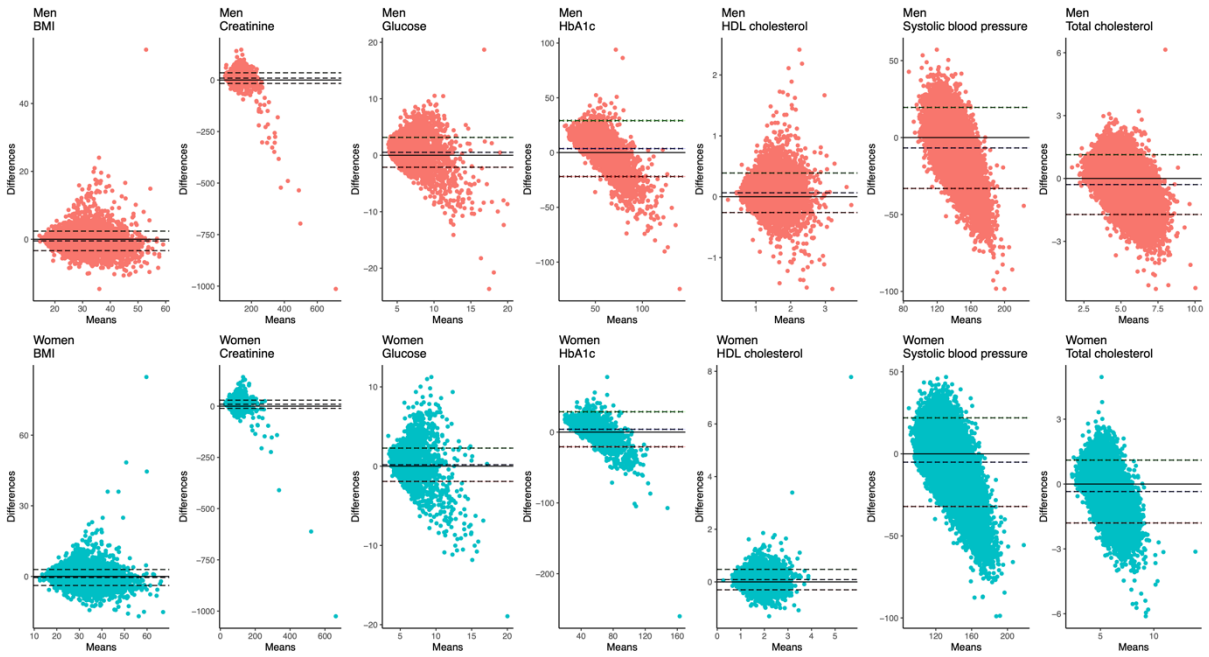


Figure 2.9: Bland Altman plot of the expected levels from a univariate mixed model using primary care records and baseline measurement by sex and risk factor

Abbreviations: BMI, body mass index; HDL, high density lipoprotein.

95% confidence intervals are represented by dotted lines. The means (x-axis) represents the arithmetic mean of the expected levels from the mixed model, and the baseline measurement. A negative value for the difference indicates the expected levels from the mixed model using primary care measurements is lower relative to the cohort baseline measurement. Individuals were included for the risk-factor specific comparison if an individual had a measurement at baseline, and a minimum of two primary care measurements, with at least one measurement before the cohort-baseline survey.

2.3.3 Agreement in categorical risk factors

The overall agreement between the last observed primary care measurement and at UKB baseline was over 90% for smoking diabetes, atrial fibrillation and rheumatoid arthritis **(Table 2.2).** Agreement marginally decreased as the time between the last observed measurement and baseline increased from 6 months to all records within the past 10 years. Low Kappa levels were observed for atrial fibrillation and rheumatoid arthritis whilst higher Kappa levels were observed for smoking and diabetes status. The high level of agreement suggests the Kappa levels were affected due to low prevalence of the underlying risk factor.

Table 2.2: Agreement of last observed risk factor status in primary care records with self-reported diagnosis at UK Biobank baseline survey by sex.

Risk factor	Time frame between last observed primary care measurement and self-reported baseline value	Men			Women		
		Individuals, N	Agreement, N (%)	Unweighted Kappa	Individuals, N	Agreement, N (%)	Unweighted Kappa
Smoking	Negative/missing at baseline	21156	19720 (93%)	0.00	26055	25044 (96%)	0
	6 months	39487	37190 (94%)	0.64	46354	44592 (96%)	070
	1 year	52391	49454 (94%)	0.69	60736	58378 (96%)	0.73
	3 years	72837	68752 (94%)	0.72	86431	83001 (96%)	0.74
	5 years	79143	74744 (94%)	0.71	94898	91186 (96%)	0.74
	10 years	79933	75496 (94%)	0.71	95987	92256 (96%)	0.74
Diabetes	Negative/missing at baseline	75084	73909 (98%)	0.00	93022	92279 (99%)	0.00
	6 months	78650	77415 (98%)	0.84	95103	94291 (99%)	0.83
	1 year	78935	77653 (98%)	0.84	95298	94441 (99%)	0.83
	3 years	79487	78147 (98%)	0.85	95688	94738 (99%)	0.83
	5 years	79665	78316 (98%)	0.86	95801	94835 (99%)	0.84
	10 years	79694	78343 (98%)	0.86	95813	94843 (99%)	0.84
Atrial fibrillation	Negative/missing at baseline	78045	76710 (98%)	0.00	93073	92471 (99%)	0
	6 months	78265	76809 (98%)	0.12	93381	92506 (99%)	0.07
	1 year	78470	76905 (98%)	0.19	93646	92532 (99%)	0.09
	3 years	79336	77317 (97%)	0.36	94974	92702 (98%)	0.16
	5 years	79959	77599 (97%)	0.41	95969	92833 (97%)	0.18
	10 years	80073	77649 (97%)	0.42	96144	92853 (97%)	0.18
Rheumatoid arthritis	Negative/missing at baseline	79757	79090 (99%)	0.00	95587	94181 (99%)	0.00
	6 months	79787	79103 (99%)	0.04	95636	94205 (99%)	0.03
	1 year	79826	79121 (99%)	0.08	95684	94226 (98%)	0.06
	3 years	79957	79177 (99%)	0.18	95889	94340 (98%)	0.17
	5 years	80056	79212 (99%)	0.22	96098	94439 (98%)	0.23
	10 years	80074	79217 (99%)	0.22	96144	94461 (98%)	0.24

A risk factor status not observed in primary care records was assigned a negative risk factor status.

2.4 Discussion

This study has evaluated the concordance of a wide range of chronic disease markers recorded in primary care records and at UKB cohort-baseline survey. Noteworthy, compared to the baseline measurements, our study observed consistently lower values of systolic blood pressure in the primary care records in comparison to those recorded at the UKB baseline, with the differences increasing with age. We found that the remaining chronic disease markers were generally concordant between the data sources.

Higher values in primary care were generally observed for HbA1c and creatinine as expected due to the targeted nature of these measurements as disease indicators for diabetes and impaired kidney function respectively. The lower values of systolic blood pressure observed in primary care records may be explained by a combination of current UK guidelines,[26] differences in the way SBP is measured at cohort-baseline and high inter-person variability in SBP at older ages. Current guidelines recommend taking a second measurement if the blood pressure measured in the clinic is greater than 140/90mmHG, and a third measurement if the second is substantially different from the first. The lowest of the last two measurements is then recorded in the electronic health system. Compared to the average of two blood pressure readings used in UKB baseline, repeated testing may cause the reported reading to be lower than the underlying blood pressure.[17,27] Other possible explanations include high inter-person variability, which could cause more deviation due to potential regression to the mean,[28] and that some SBP measurements were likely taken at home and would be expected to be lower than those measured in a clinic.[29]

We also observed high levels of concordance between the last observed record before cohort-baseline and the self-reported information at cohort-baseline for categorical risk factors, suggesting that self-reported information is a suitable disease indicator in cohort studies.

My findings should inform researchers of the similarities and also potential differences between primary care records and cohort measurements when enriching studies with electronic health records data that were not intended for research, for example as done in previous studies,[30–32] and in later chapters of this thesis. My findings suggest that whilst using the last observed primary care record for categorical risk factors may serve as an adequate substitute due to high

levels of agreement, utilising all repeated measurements using a univariate mixed effects model can improve agreement.

The findings may however have implications on the development of risk models and its applications in clinical practice. In particular, models derived using primary care records and applied to cohorts (for example derivation of a model using the lower primary care SBP values relative to the higher values observed at the cohort-baseline survey), or vice versa, may result in a different distribution of risks. These differences may be largely corrected by a simple recalibration process after model derivation assuming the relative risks for the risk factors are similar (see **Chapter 1, Section 1.5.2.2**), however further care will be needed for risk factors including HbA1c and creatinine due to its targeted nature when measured within primary care. Another potential implication is the underdiagnosis of hypertension in primary care. With the lower primary care SBP values, compared to those measured in cohorts, individuals may end up being below the set threshold for hypertension. However, as it is unclear whether the primary care or cohort measurements are closer to the true underlying blood pressure, future research should continue investigating this.

Our study has some key strengths. First, we quantified individual level differences between linked primary care records with a cohort for a range of chronic disease risk factors in a large population. Second, we evaluated the use of a mixed-effects model to utilise all repeated measures for continuous risk factors, highlighting the potential benefits of utilising a more complex model and harnessing all longitudinal data. Third, we used different metrics to quantify the agreement of categorical risk factors, where we saw similar levels of agreement even after changing the amount of time between the last observed primary care measurement and the cohort-baseline measurement.

Our study has some limitations. As UKB is a healthy cohort and more homogenous than the general population, the simple agreement statistic used for the categorical risk factors may not be directly generalisable to the whole population and greater differences may exist between primary care and measured risk factors. In particular, due to the low prevalence of risk factors such as atrial fibrillation or rheumatoid arthritis in UKB, the high agreement observed is partly due to the majority of individuals remaining negative.

Another key limitation is that the analysis focussed predominantly within individuals with linked primary care records and with non-missing data. It has been shown that the recording,

and any associated missingness, of health indicators is affected by demographics.[33] For example, women are more likely to have weight, and therefore BMI, recorded at a younger age. Since adjustments were not made to account for this, further work is required to understand whether the concordance of primary care records in cohorts varies by demographics.

2.5 Conclusion

We have explored the concordance between longitudinal measurements in a cohort of individuals with linked primary care records. We observed high levels of agreement for binary disease markers and good agreement of the majority of chronic disease markers evaluated. Our results suggest care should be taken when linking systolic blood pressure and HbA1c from primary care records to cohorts due to significant differences between the two data sources.

These findings will inform the modelling decisions made in **Chapters 3-5.** First, any missing binary disease markers can be substituted with last observed values**.** Second, as differences exist between the primary care and cohort-baseline data, deriving a model in the different datasets will lead to differences between the estimated 10-year CVD risks. As such, recalibration will be performed to ensure that the models are comparable. Third, the differences observed will be necessary to implement a dynamic population health modelling approach in **Chapter 5**, where the primary care records will be used to estimate individual risk factor trajectories over time, necessary for modelling repeated formal risk assessments.

References

1. Abul-Husn NS, Kenny EE. Leading Edge Perspective Personalized Medicine and the Power of Electronic Health Records. *Cell*. 2019;177:58-69. doi:10.1016/j.cell.2019.02.039
2. National Institute for Health and Care Excellence (NICE). Lipid modification: Cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease (clinical guideline CG181). Published online 2014.
3. 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*. 2017;357. doi:10.1136/bmj.j2099
4. Hippisley-Cox J, Coupland C, Vinogradova Y, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ*. 2008;336(7659):1475-1482. doi:10.1136/BMJ.39609.449676.25
5. Aliabadi A, Sheikhtaheri A, Ansari H. Electronic health record–based disease surveillance systems: A systematic literature review on challenges and solutions. *J Am Med Inform Assoc*. 2020;27(12):1977. doi:10.1093/JAMIA/OCAA186
6. Wood A, Denholm R, Hollings S, et al. Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. *BMJ*. 2021;373. doi:10.1136/BMJ.N826
7. Bowton E, Field JR, Wang S, et al. Biobanks and Electronic Medical Records: Enabling Cost-Effective Research. *Sci Transl Med*. 2014;6(234):234cm3. doi:10.1126/SCITRANSLMED.3008604
8. Hall JL, Ryan JJ, Bray BE, et al. Merging Electronic Health Record Data and Genomics for Cardiovascular Research: A Science Advisory From the American Heart Association. *Circ Cardiovasc Genet*. 2016;9(2):193-202. doi:10.1161/HCG.0000000000000029
9. Kullo IJ, Jarvik GP, Manolio TA, Williams MS, Roden DM. Leveraging the electronic health record to implement genomic medicine. *Genet Med 2013 154*. 2012;15(4):270-271. doi:10.1038/gim.2012.131
10. Manemann SM, St Sauver JL, Liu H, et al. Longitudinal cohorts for harnessing the electronic health record for disease prediction in a US population. *BMJ Open*. 2021;11(6):e044353. doi:10.1136/BMJOPEN-2020-044353
11. Fraile-Navarro D, Azcoaga-Lorenzo A, Agrawal U, et al. Development of an algorithm to classify primary care electronic health records of alcohol consumption: experience

using data linkage from UK Biobank and primary care electronic health data sources. *BMJ Open*. 2022;12(2):e054376. doi:10.1136/BMJOPEN-2021-054376

12. Darke P, Cassidy S, Catt M, Taylor R, Missier P, Bacardit J. Curating a longitudinal research resource using linked primary care EHR data—a UK Biobank case study. *J Am Med Informatics Assoc*. 2022;29(3):546-552. doi:10.1093/JAMIA/OCAB260
13. NHS Digital. Hospital Episode Statistics (HES). Accessed May 15, 2020. https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics
14. NHS Digital. Linked HES-ONS mortality data. Accessed May 15, 2020. https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/linked-hes-ons-mortality-data
15. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants with Those of the General Population. *Am J Epidemiol*. 2017;186(9):1026-1034. doi:10.1093/aje/kwx246
16. Batty GD, Gale CR, Kivimäki M, Deary IJ, Bell S. Comparison of risk factor associations in UK Biobank against representative, general population based studies with conventional response rates: prospective cohort study and individual participant meta-analysis. *BMJ*. 2020;368. doi:10.1136/bmj.m131
17. UK Biobank. Data-Field 4080. Accessed June 14, 2022. https://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=4080
18. Biobank U. Data-Field 30690. Accessed September 23, 2022. https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=30690
19. Biobank U. Data-Field 30760. Accessed September 23, 2022. https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=30760
20. Biobank U. Data-Field 30740. Accessed September 23, 2022. https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=30740
21. Biobank U. Data-Field 30700. Accessed September 23, 2022. https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=30700
22. Biobank U. Data-Field 30750. Accessed September 23, 2022. https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=30750
23. Goldberger AS. Best Linear Unbiased Prediction in the Generalized Linear Regression Model. *J Am Stat Assoc*. 1962;57(298):369. doi:10.2307/2281645
24. Paige E, Barrett J, Stevens D, et al. Landmark Models for Optimizing the Use of Repeated Measurements of Risk Factors in Electronic Health Records to Predict Future Disease Risk. *Am J Epidemiol*. 2018;187(7):1530-1538. doi:10.1093/AJE/KWY018

25. Cohen J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychol Bull*. 1968;70(4):213-220. doi:10.1037/H0026256
26. National Institute for Health and Care Excellence (NICE). Hypertension in adults: diagnosis and management NICE guideline. Published online 2019. Accessed June 14, 2022. www.nice.org.uk/guidance/ng136
27. Chung RK, Wood AM, Sweeting MJ. Biases incurred from nonrandom repeat testing of haemoglobin levels in blood donors: Selective testing and its implications. *Biometrical J*. 2019;61(2):454-466. doi:10.1002/bimj.201700268
28. Barrett JK, Huille R, Parker R, Yano Y, Griswold M. Estimating the association between blood pressure variability and cardiovascular disease: An application using the ARIC Study. *Stat Med*. 2019;38(10):1855-1868. doi:10.1002/sim.8074
29. Miao H, Yang S, Zhang Y. Differences of blood pressure measured at clinic versus at home in the morning and in the evening in Europe and Asia: A systematic review and meta-analysis. *J Clin Hypertens*. 2022;24(6):677. doi:10.1111/JCH.14487
30. Bhaskaran K, Rentsch CT, Hickman G, et al. Overall and cause-specific hospitalisation and death after COVID-19 hospitalisation in England: A cohort study using linked primary care, secondary care, and death registration data in the OpenSAFELY platform. *PLoS Med*. 2022;19(1). doi:10.1371/JOURNAL.PMED.1003871
31. Wells S, Riddell T, Kerr A, et al. Cohort Profile: The PREDICT Cardiovascular Disease Cohort in New Zealand Primary Care (PREDICT-CVD 19). *Int J Epidemiol*. 2017;46(1):22-22. doi:10.1093/IJE/DYV312
32. Gaziano JM, Concato J, Brophy M, et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol*. 2016;70:214-223. doi:10.1016/J.JCLINEPI.2015.09.016
33. Petersen I, Welch CA, Nazareth I, et al. Health indicator recording in UK primary care electronic health records: key implications for handling missing data. *Clin Epidemiol*. 2019;11:157-167. doi:10.2147/CLEP.S191437

Chapter 3

Prioritising cardiovascular disease risk assessment to high risk individuals based on primary care records

Chapter summary

Background: This chapter aims to quantify the efficiency gains from systematically prioritising individuals for full formal cardiovascular disease (CVD) risk assessment. I develop a novel tool (eHEART) which utilises primary care records and assess the impact of using such a tool using population health modelling, along with age- and sex- specific risk thresholds.

Methods: eHEART was derived using landmark Cox models for incident CVD with repeated measures of conventional CVD risk predictors from primary care records in 1,642,498 individuals using the Clinical Practice Research Datalink. We then used 119,137 individuals from UK Biobank to model the implications of initiating guideline-recommended statin therapy using eHEART with age- and sex-specific prioritisation thresholds corresponding to 5% false negative rates to prioritise adults aged 40-69 years in a population in England for invitation to a formal CVD risk assessment with QRISK2.

Results: Formal CVD risk assessment with QRISK2 on all adults would identify 74% and 46% of future CVD events amongst men and women respectively, and 188 (95% CI: 185, 192) men and 602 (95% CI: 579, 623) women would need to be screened (NNS) to prevent one CVD event. In contrast, if eHEART was first used to prioritise individuals for formal CVD risk assessment with QRISK2, we would identify 72% and 44% of future events amongst men and women respectively, and a NNS of 152 (95% CI: 150, 155) men and 346 (95% CI: 332, 358) women. Replacing the age- and sex-specific prioritisation thresholds with a 10% threshold identify around 10% less events.

Conclusions: The use of prioritisation tools with age- and sex-specific thresholds could lead to more efficient CVD assessment programmes with only small reductions in effectiveness at preventing new CVD events.

3.1 Introduction

The WHO estimate that the majority of premature cardiovascular disease (CVD) events are preventable through lifestyle choices and pharmacotherapy that target modifiable risk factors.[1] Many countries implement population-wide CVD risk assessment programmes in primary care that recommend use of established risk prediction tools to identify individuals at high risk of CVD and guide patient and clinical decision-making.[2–8] More recently, to reduce programme running costs and address rising health inequalities in access to preventative care, guidelines and policy makers have advocated using primary care records for systematic prioritisation of individuals *before* formal CVD risk assessment (see **Chapter 1, Section 1,3 Box 1.1**).[7]

Previous studies have shown that compared to using a universal approach, pre-stratifying individuals prior to a more expensive formal risk assessment can be similarly effective at preventing new cases, be more cost-effective and be more equitable. Alternative approaches that have been investigated, include using a simple risk score incorporating routine data[9], using age- and sex-specific invitation strategies[10], and strategies that target key cardiovascular risk factors.[11] Indeed, Kypridemos et al. modelled the current implementation of the NHS Health Check and believes the programme is neither equitable or cost saving in a city with high levels of deprivation and CVD.[11]

In addition to the recommendations of formally assessing all individuals in England, individuals are recommended to be prioritised for a formal CVD risk assessment if their estimated risk, using existing data from primary care health records, is greater than 10% risk over ten years.[7] However, it is currently unknown how using a fixed 10% CVD risk threshold impacts prioritisation and as a consequence formal CVD risk assessments. In addition, no specific risk tool is recommended for systematic prioritisation of risk.

Therefore, to help in the formulation of evidence-based CVD risk assessment programmes, we aim to evaluate two key aspects. First, we compare the *population-wide* formal CVD risk assessment approach against approaches that *systematically prioritise* individuals before full formal CVD risk assessment. Here, we evaluate prioritisation based on primary care records using a novel tool (eHEART), developed in this chapter to leverage the sparse and sporadically observed longitudinal data on conventional CVD risk factors recorded primary care records. Second, we compare the use of a fixed 10% prioritisation threshold versus age- and sex-specific prioritisation thresholds.

We will use primary care data from the Clinical Practice Research Datalink GOLD (CPRD) to derive the eHEART prioritisation tool. We will then use the UK Biobank (UKB) baseline data and the linked primary care records to evaluate eHEART in a representative population using population health modelling (see **Chapter 2, Section 2.2.1**).

3.2 Methods

3.2.1 Data sources

3.2.1.1 CPRD data source

Initially established in 1987 and launched in 2012, CPRD is a large, ongoing database of routinely collected data from UK general practices, with longitudinal primary care records covering over 60 million total patients from over 2000 general practices in the UK.[12,13] As of 2023, 18 million patients (over 25% of the total UK population) are currently registered and alive, with 25% of the patients having at least 20 years of follow-up.[13] It is therefore one of the largest research databases of medical records in the UK and the world.[14]

CPRD currently collects coded data from practices that use the Vision GP patient management software, as part of the CPRD GOLD dataset available to researchers since 2012, and data from the EMIS GP software, as part of the CPRD Aurum dataset, available to researchers since 2018. As of April 2023, the total number of research acceptable patients is over 21 million in CPRD GOLD and over 41 million in CPRD Aurum.[13,15,16] During the preparation of this thesis only data from CPRD GOLD was available.

The spatial distributions of the Vision and EMIS systems used in general practices across England was assessed in a previous study **(Figures 3.1-3.2)**.[17] In general, Vision had a strong presence in London, the South, Greater Manchester and Birmingham, whereas EMIS had a stronger presence in the West of England, London and the South. Although CPRD is unable to provide comprehensive coverage in the North and East of England, in 2015 the active patients in CPRD were shown to be broadly representative of the general population in the UK in terms of age, sex and ethnicity.[15]

To derive eHEART we used CPRD GOLD with linked information on hospital episodes from Hospital Episode Statistics (HES), and death registrations from the Office of National Statistics

(ONS) (see **Chapter 2, Section 2.2.1**).[18] Primary care records were extracted for individuals from the latest of: date of their registration at general practice plus 6 months, their 30th birthday, date when the general practice provided "up-to-standard" data,[19] or 1st April 2004 (the date of introduction of the Quality and Outcomes Framework),[20] until the earliest of: date of their first (i.e. "incident") newly recorded CVD event, date of de-registration at the general practice, their 85th birthday, date of death, last contact date for the practice with CPRD, or 31 May 2019 (the end of data availability). All analyses focus on individuals without known pre-existing CVD and without prescribed statins.

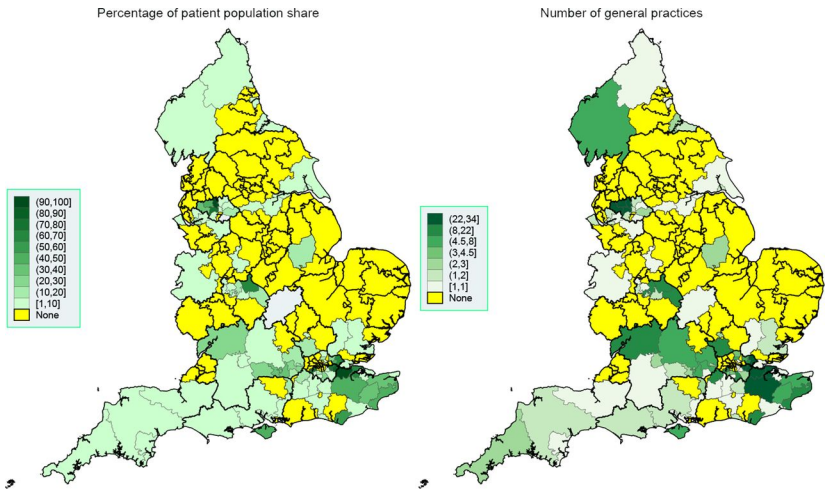


Figure 3.1: Spatial map at the Clinical Commissioning Group level, September 2016: Vision.

Left graph: percentage of patient population share. Right graph: Number of general practices. Thicker border lines correspond to the 14 National Health Service regions, left graph uses equidistant class breaks; right graph uses class breaks based on distribution of variable of interest, with each class having approximately the same number of spatial polygons.

Source: Kontopantelis et al., 2018[17]

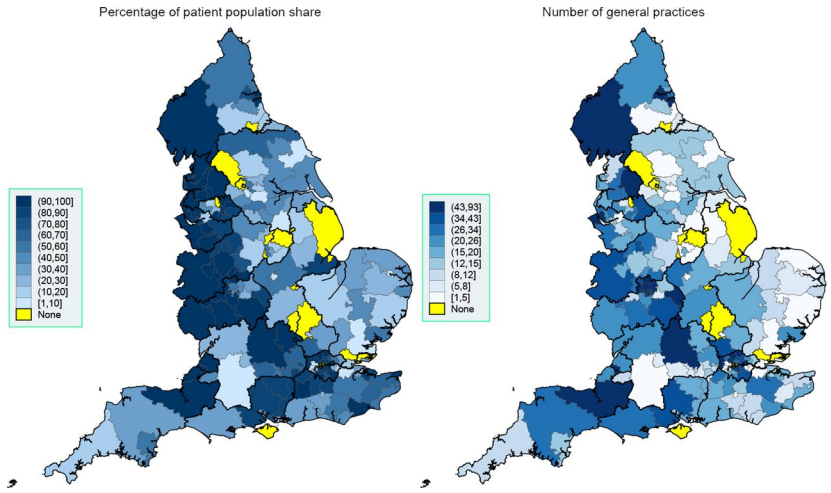


Figure 3.2: Spatial map at the Clinical Commissioning Group level, September 2016: EMIS

Left graph: percentage of patient population share. Right graph: Number of general practices. Thicker border lines correspond to the 14 National Health Service regions, left graph uses equidistant class breaks; right graph uses class breaks based on distribution of variable of interest, with each class having approximately the same number of spatial polygons.

Source: Kontopantelis et al., 2018[17]

3.2.1.2 UK Biobank data source

177,359 individuals with linked primary care records from UKB were used to model the implications of prioritising individuals for formal CVD assessment and subsequent initiation of guideline-recommended statin therapy in a primary care setting (see **Chapter 2, Section 2.2.1**) UKB was chosen due being able to use to the linked primary care records before baseline to evaluate eHEART as a prioritisation tool, in addition to the complete data available at baseline to replicate being able to perform a full risk assessment. Population health modelling focussed on individuals without known pre-existing CVD and without prescribed statins at baseline.

3.2.2 Outcomes and risk factors

We ascertained individuals in both CPRD and UKB cohorts with a first ever incident CVD, defined as nonfatal or fatal events of coronary heart disease (CHD) (including myocardial infarction and angina), stroke, and transient ischemic attack, as those with a relevant Read-code or ICD-10 code (listed in **Appendix 1**) appearing in the primary care data, hospital episodes (main or secondary diagnostic code position in the admitted patient care component of the

Hospital Episode Statistics data), or death registry (underlying or contributing cause of death) during follow-up.

Conventional and commonly recorded CVD risk predictors[5,21] in primary care records were pre-selected for inclusion into the estimated eHEART prioritisation tool and included: age (in years), sex (men or women), diabetes status (yes or no), hypertension medication (yes or no), systolic blood pressure (SBP) (mmHg), total cholesterol (mmol/litre) and high-density lipoprotein (HDL) cholesterol (mmol/litre) and smoking status (current smoker or not) (details of measurements have been previously described).[15]

As in **Chapter 2**, the following measurements were considered biologically implausible and were changed to missing (~0.4% of measurements): SBP <60 or >250 mmHg; total cholesterol <1.75 or >20 mmol/litre; and HDL cholesterol <0.3 or >3.1 mmol/litre.[22]

3.2.3 Statistical modelling

3.2.3.1 The eHEART prioritisation tool

To optimise the nature of longitudinal primary care records, we used a landmark-age approach to derive 90 age- (i.e., 40, 41, 42, …to 84 years) and sex-specific Cox models as previously described (see **Chapter 1, Section 1.5.2.1**)[23] Each model utilised primary care records from persons alive without pre-existing CVD and without prescribed statins at the landmark-age by using available risk predictors recorded before the landmark-age to predict 10-year risk of incident CVD outcomes **(Figure 3.3)**.

Each model was developed in two stages. In the first stage, a multivariate mixed-effects model[24] was fitted to repeat measures of smoking status, SBP, total cholesterol and HDL cholesterol recorded before the landmark-age and used to estimate risk predictor values at the landmark-age amongst individuals who had at least one past record of any one predictor. Unlike the univariate mixed-effects models previously used (see **Chapter 2**, **Section 2.2.3**) a single multivariate model is fit on the chosen risk factors instead of using risk-factor specific models. The model included fixed- and correlated random-intercepts and linear age fixed-effects (fixed quadratic terms and random slope terms were not significant) for each predictor, as well as an interaction between SBP and an age-varying binary covariate for hypertension treatment, and an interaction between total cholesterol and an age-varying binary covariate for statin treatment (but excluded those with pre-statins for Cox model). The model intrinsically accounts for

missing data in the predictors with the assumption that predictor values from individuals with incomplete data are from the same multivariate normal distribution for predictor values as individuals with observed data. Further details are given in **Appendix 2**.

In the second stage, we derived Cox models with time since landmark age as the underlying time scale and the following predictors: diabetes status, treatment for hypertension, and the estimated values of SBP, total cholesterol, HDL cholesterol, and smoking status from the multivariate mixed-effects model. The four estimated predictors were entered as linear terms. The outcome was incident CVD and censoring occurred for individuals at end of their follow-up, either date of CVD event (cases), the study end date or date of death (from non-CVD causes). No violation of the proportional hazards assumption was identified. The Cox model was used to estimate 10-year CVD risks, constituting the "eHEART risk".

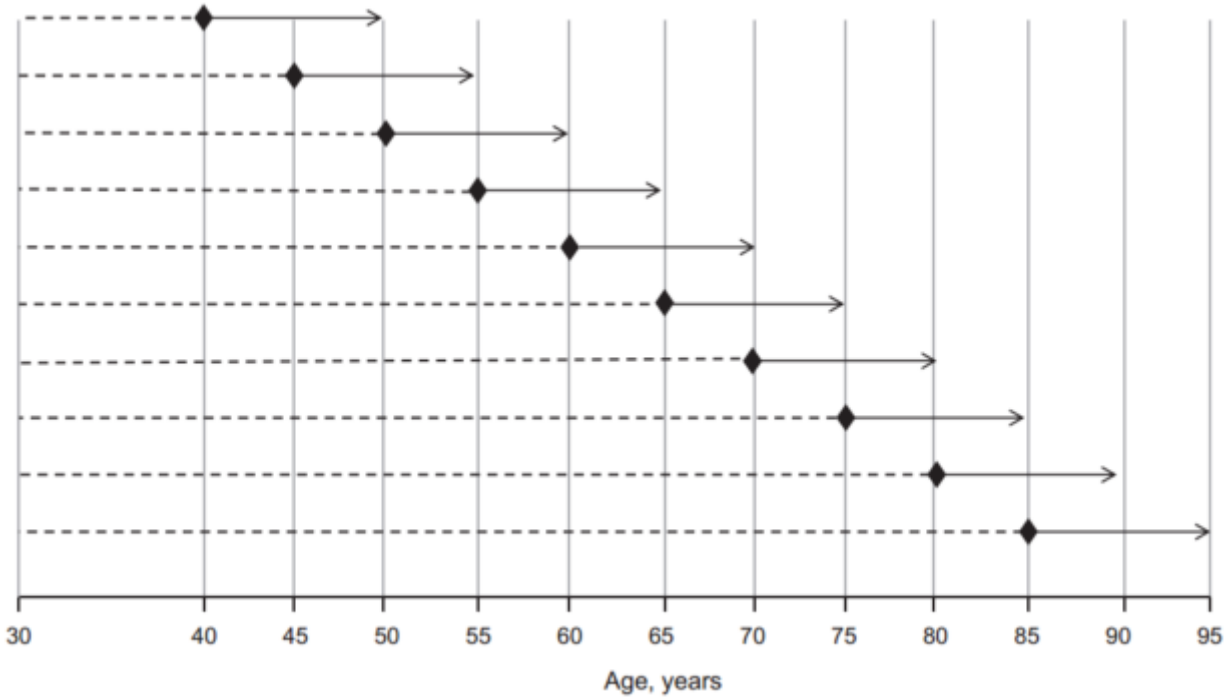


Figure 3.3: Schematic showing the landmark age approach of model derivation.

The dashed lines indicate historical repeat measures of smoking status, systolic blood pressure, total cholesterol, and HDL cholesterol, modelled by means of landmark-age–specific multivariate linear mixed-effects models. The diamonds show the landmark age (time of risk prediction). The arrows indicate the 10-year follow-up to the point of a cardiovascular disease event or censoring, modelled via a landmark Cox model. This figure was published in our previous work[23].

3.2.3.2 Model internal validation and assessment

The eHEART prioritisation model was derived using primary care records from 2/3 of general practices and internally validated on the remaining 1/3 of general practices. In the validation dataset, we calculated measures of predictive accuracy using the Brier score[25] and risk discrimination using the C-index.[26] Calibration was assessed visually by plotting mean predicted risk against mean observed risk by deciles of predicted risk[27,28] and by assessing the calibration slope, estimated from the linear regression of log transformed observed risk and predicted risk in predicted risk decile groups.

3.2.4 Population health modelling

We compared the potential population health impact of using eHEART to prioritise individuals for invitation to a formal CVD risk assessment with QRISK2 **(Figure 3.4)** versus the whole-population strategy to invite all adults for formal CVD risk assessment with QRISK2. We used 174,715 participants in UKB who had historical primary care records to calculate eHEART, as well as large complete data at the cohort-baseline assessment to calculate QRISK2 at the participant's age at cohort-baseline.

By implementing a two-stage sequential testing process, with prioritisation followed by a formal assessment, versus a single-stage test of only a formal risk assessment, prioritisation will improve the formal risk assessment's specificity i.e., deferring individuals that would have been deemed at low risk at the prioritisation stage. Consequently, this will result in a reduction in the formal risk assessment's sensitivity.

Formal 10-year CVD risks using QRISK2 were estimated using published coefficients (see **Chapter 1, Section 1.3.1**) and individual-level predictor values measured at the cohort-baseline. Chronic kidney disease, family history and historical prescriptions of statins from the linked primary care records were supplemented in addition to the baseline records. Missing values for SBP, cholesterol levels and BMI were imputed using age-, sex- and ethnicity-specific means estimated in UKB[29] (since the proposed approach for imputing QRISK2 missing values is unavailable - see **Chapter 1, Section 1.3.1)**.

Due to UKB being a cohort of healthier individuals than the UK primary care population, we recalibrated the eHEART and QRISK2 models using age-group- and sex-specific risk factor levels and CVD incidence rates from UKB (see **Chapter 1, Section 1.5.2.2**), and then *rescaled*

the predicted eHEART and QRISK2 risks to achieve 10-year CVD risk distributions that would be expected in a UK primary care setting, using methods previously described[28]. Details are provided in **Appendix 3**.

We modelled a population of 50,000 men and 50,000 women aged 40–69 years with age profiles matching that of the contemporary UK population (2017 mid-year population),[30] and CVD incidence rates as observed in CPRD individuals. We assumed a policy of statin initiation for individuals at ≥10% predicted QRISK2 10-year risk as recommended by National Institute for Health and Care Excellence (NICE) guidelines[7] and assumed statin allocation would reduce CVD risk by 20%.[31]

Using expert primary care physician advice and existing literature, we assumed 50% statin compliance and a 50% invitation uptake when formally inviting all individuals for a formal assessment.[32–34] We also assumed that the addition of prioritisation to the risk assessment process would increase the invitation uptake to 55%. Whilst we chose an arbitrary increase of 5%, the improvement was guided by existing literature; in particular, evidence has shown that, in the context of kidney cancer screenings, risk stratification prior to screening could improve uptake.[35] In addition, trials have been conducted highlighting improved uptake when using alternate forms of invitation.[36,37]

We then modelled the impact of prioritising individuals for QRISK2 formal assessment, based on a fixed prioritisation threshold of ≥10% estimated eHEART 10-year risk as currently recommended by NICE guidelines[7] (see **Chapter 1, Section 1.3**) and also on age- and sex-specific prioritisation thresholds selected to correspond to 5% false negative rates in UKB; these were chosen by ranking the estimated eHEART 10-year risks in UKB individuals who go on to have a CVD event in the next 10 years. The prioritisation thresholds were chosen as the minimum 10-year risk such that 5% of events would be missed. Population health impact was assessed using the number needed to screen (i.e., undergo formal assessment) to prevent one CVD event, the number of events identified and the number need to invite for formal assessment to prevent one CVD event assuming an increased invitation uptake of 55% after prioritisation (see **Chapter 1, Section 1.5.1.4**).[33]

3.2.4.1 Sensitivity Analyses

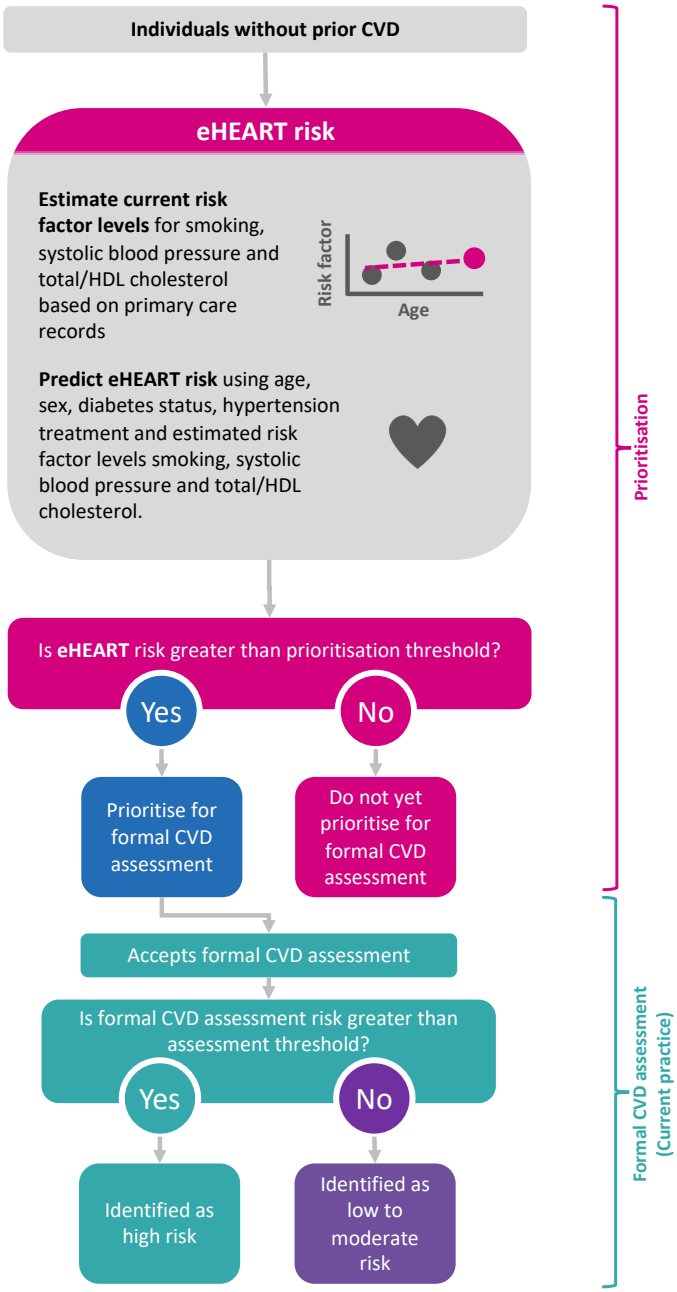
In sensitivity analyses, first we repeated population health analyses by replacing eHEART with an estimated QRISK2 risk based on historical primary care records. Estimated QRISK2 risks were calculated using last observed values within primary care records at the baseline age of UKB and published coefficients (see **Chapter 1, Section 1.2.1**). Ethnicity and deprivation information was obtained using the self-reported baseline data. Missing values for SBP, cholesterol levels and BMI were imputed using the approach previously described (see **Chapter 1, Section 3.2.3)**. The range of valid measurements for QRISK2 mirrors the range used in the published calculator and were restricted to: SBP 70 to 210 mmHg; cholesterol ratio 1 to 12; BMI 20 to 40kg /m^2.

Second, we repeated population-health analyses including all individuals, including those without a primary care record for any one of SBP, HDL, total cholesterol and smoking status. For the eHEART tool, we assumed all these individuals would be directly invited for a formal risk assessment, and for the estimated QRISK2 prioritisation tool, additional missing values were imputed as described above.

Third, due to the upcoming changes to the formal risk assessment thresholds recommended by NICE (see **Chapter 1, Section 1.3.3)**, we repeated the analyses assuming a 5% formal risk assessment threshold in combination with a fixed 5% prioritisation threshold and age- and sex-specific prioritisation thresholds selected to correspond to 2.5% false negative rates in UKB.

Analyses were performed with R x64 3.6.1[38] and Stata version 15. This study follows the RECORD and TRIPOD reporting guidelines.[39,40] R code to predict eHEART for new individuals is provided online.

Figure 3.4: Flowchart of using eHEART as a prioritisation tool prior to formal cardiovascular disease risk assessments



Abbreviations: BMI, body mass index; CVD, cardiovascular disease; HDL, high density lipoprotein.

3.3 Results

3.3.1 Study population and baseline characteristics using CPRD

We identified 2,154,089 individuals from 398 practices contributing to the CPRD database with linked hospital episode statistics and ONS death registrations in England and who met our inclusion criteria. For our primary analysis we excluded 511,591 (24%) without at least 1 record of SBP, total cholesterol, HDL cholesterol or smoking status, leaving 1,642,498 individuals. Compared to those included in our analysis, individuals without any measurement on SBP, cholesterol, or smoking were slightly older, more likely to be men, and had higher CVD incidence rate (**Table 3.1**). A total of 263 practices were randomly assigned to the derivation dataset and the remainder (n=135) to a validation cohort. Overall, 1,120,053 and 522,445 individuals were included in the derivation and validation cohorts respectively (**Figure 3.5**).

The baseline characteristics of individuals are summarised in **Table 3.2** (split by the derivation and validation cohorts in **Table 3.3**). Overall, 45% were men and 55% were women; 3% of men and 3% of women had a reported diabetes diagnosis and 16% of men and 23% of women were prescribed anti-hypertension medication before the study entry (**Table 3.2**). The number of individuals with CVD risk factors recorded in primary care varied by age and sex (for example, from 21,674 men at age 85 to 188,959 women at age 46) (**Figures 3.6-3.7**). There were more repeated SBP than cholesterol measurements recorded (**Table 3.1, Figures 3.6-3.7**). Characteristics were similar for individuals in the derivation and validation cohorts (**Table 3.3**).

In the derivation cohort, 104,830 (6%) individuals had an incident CVD event recorded in either primary care, HES or death registry (**Figure 3.8**) over a median 9 years of follow-up (IQR: 5-11). The crude CVD incidence rates per 1000 person years increased with age (from 2.13 in 40-year-olds to 54.88 in those aged 85) and were higher among men than women (**Figure 3.9**). In the validation cohort, 32,862 (6%) individuals had an incident CVD event over a median 9 years of follow-up (IQR: 5-11).

Figure 3.5: Flow chart showing selection of patient records for analysis in CPRD

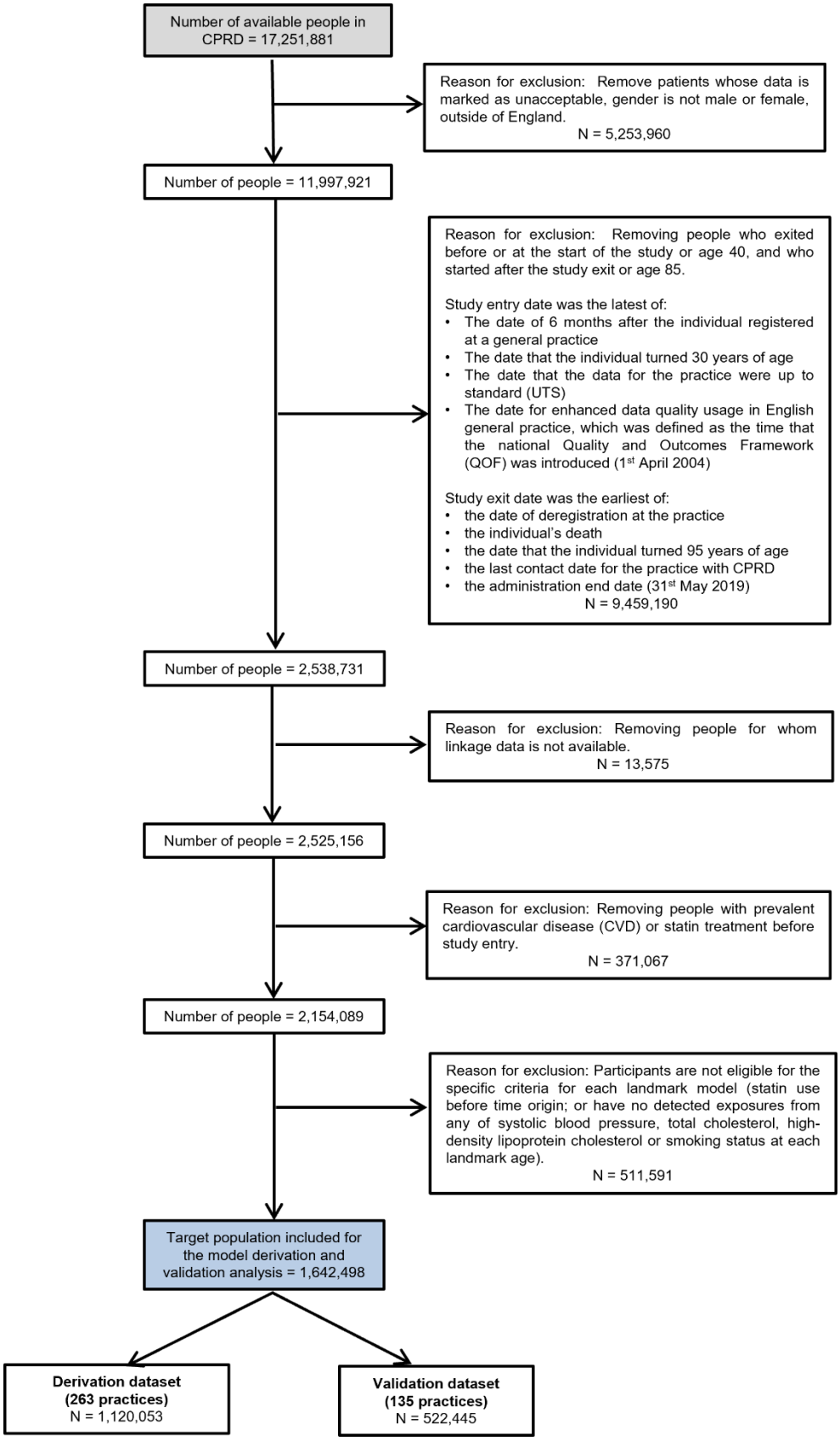


Table 3.1: Comparison of characteristics of 1,642,498 individuals included in the derivation and validation datasets* and 511,591 individuals with missing predictor measurements§

Characteristics	Individuals with at least 1 measurement on any of the risk predictors (n = 1,642,498)	Individuals without any measurement on the risk predictors (n = 511,591)
Age at study entry, mean (SD), years	50.9 (13.2)	51.04 (12.5)
Men, n (%)	746,386 (45.4)	309,542 (60.5)
Incidence of CVD, rate (95% CI), 1,000 person-years	7.50 (7.5, 7.5)	14.9 (14.8, 15.1)

Abbreviations: CI, confidence interval; CVD, cardiovascular disease.

* Included 1,642,498 individuals from Clinical Practice Research Datalink, Hospital Episode Statistics, and the Office for National Statistics, England, United Kingdom, 2004-2019, aged 40-85 years, without prevalent CVD and statin treatment before study entry, and had least 1 measurement on any of systolic blood pressure, total cholesterol, HDL cholesterol, or smoking status between their study entry and study exit dates.

§ The 511,591 participants not included in the study were those without prevalent CVD and statin treatment before study entry, but have no detected measurements for complete risk predictors measurement on any of systolic blood pressure, total cholesterol, HDL cholesterol, or smoking status between their study entry and study exit dates

Table 3.2: Key characteristics of individuals in CPRD cohort.

Characteristics	Men, N=746,386			Women, N=896,112		
	Mean (SD) or n (%)	No. (%) of persons with value	Median (IQR) of measures per person	Mean (SD) or n (%)	No. (%) of persons with value	Median (IQR) of measures per person
Earliest age during follow-up, years	50.4 (12.7)	746,836 (100)	-	51.3 (13.6)	896,112 (100)	-
History of diabetes[§]	21,936 (3.0)	746,836 (100)	-	20,662 (3.0)	896,112 (100)	-
Blood pressure-lowering medication prescriptions^	119,230 (16.0)	746,836 (100)	-	203,394 (22.7)	896,112 (100)	-
Current smoker, n (%)[#]	192,509 (25.8)	438,638 (59)	3 (2-6)	176,719 (19.7)	424,268 (47)	4 (2-7)
Systolic blood pressure mm Hg, mean (SD)[#]	136.7 (18.5)	699,229 (93)	4 (2-11)	131.6 (20.3)	870,075 (97)	6 (3-14)
Total cholesterol mmol/litre, mean (SD)[#]	5.4 (1.0)	532,620 (71)	2 (1-5)	5.6 (1.0)	624,189 (70)	2 (1-5)
HDL cholesterol mmol/litre, mean (SD)[#]	1.3 (0.4)	490,003 (66)	2 (1-4)	1.6 (0.4)	573,842 (64)	2 (1-4)

* Included 1,642,498 individuals from Clinical Practice Research Datalink, Hospital Episode Statistics, and the Office for National Statistics, England, United Kingdom, 2004-2019, aged 40-85 years, without prevalent CVD and statin treatment before study entry, and had least 1 measurement on any of systolic blood pressure, total cholesterol, HDL cholesterol, or smoking status between their study entry and study exit dates.

[§]Defined as ever having recorded diagnosis of diabetes before the study entry

^Defined as ever having recorded use of blood-pressure lowering medications before the study entry

[#]Proportion or mean (standard deviation) of the first ever measurement

Table 3.3: Key characteristics of individuals in the derivation and validation cohorts

Characteristics	Derivation cohort						Validation cohort					
	Men, N=509,127			Women, N=610,926			Men, N=237,259			Women, N=285,186		
	Mean (SD) or n (%)	No. (%) of persons with value	Median (IQR) of measures per person	Mean (SD) or n (%)	No. (%) of persons with value	Median (IQR) of measures per person	Mean (SD) or n (%)	No. (%) of persons with value	Median (IQR) of measures per person	Mean (SD) or n (%)	No. (%) of persons with value	Median (IQR) of measures per person
Earliest age during follow-up, years	50.3 (12.7)	509,127 (100)	-	51.2 (13.6)	610,926 (100)	-	50.5 (12.6)	237,259 (100)	-	51.5 (13.6)	285,186 (100)	-
History of diabetes[§]	15,035 (3)	509,127 (100)	-	14,081 (2)	610,926 (100)	-	6,901 (3)	237,259 (100)	-	6,581 (2)	285,186 (100)	-
Blood pressure-lowering medication prescriptions^	80,896 (16)	509,127 (100)	-	138,254 (23)	610,926 (100)	-	38,334 (16)	237,259 (100)	-	65,140 (23)	285,186 (100)	-
Current smoker, n (%)[#]	132,314 (44)	299,787 (59)	3 (2-6)	121,987 (42)	289,273 (47)	4 (2-7)	60,195 (43)	138,851 (59)	3 (2-6)	54,732 (41)	134,995 (47)	4 (2-7)
Systolic blood pressure mm Hg, mean (SD)[#]	136.9 (18.5)	475,990 (93)	4 (2-11)	131.8 (20.3)	592,353 (97)	6 (3-14)	136.3 (18.4)	223,239 (94)	4 (2-11)	131.2 (20.3)	277,722 (97)	6 (3-14)
Total cholesterol mmol/litre, mean (SD)[#]	5.4 (1.0)	364,435 (72)	2 (1-5)	5.6 (1.0)	427,394 (70)	2 (1-5)	5.4 (1.0)	168,185 (71)	2 (1-5)	5.6 (1.1)	196,795 (69)	2 (1-5)
HDL cholesterol mmol/litre, mean (SD)[#]	1.3 (0.4)	337,436 (66)	2 (1-4)	1.6 (0.4)	395,568 (65)	2 (1-4)	1.3 (0.4)	152,567 (64)	2 (1-4)	1.6 (0.4)	178,274 (63)	2 (1-4)

* Included 1,642,498 individuals from Clinical Practice Research Datalink, Hospital Episode Statistics, and the Office for National Statistics, England, United Kingdom, 2004-2019, aged 40-85 years, without prevalent CVD and statin treatment before study entry, and had least 1 measurement on any of systolic blood pressure, total cholesterol, HDL cholesterol, or smoking status between their study entry and study exit dates.

[§]Defined as ever having recorded diagnosis of diabetes before the study entry

^Defined as ever having recorded use of blood-pressure lowering medications before the study entry

[#]Proportion or mean (standard deviation) of the first ever measurement

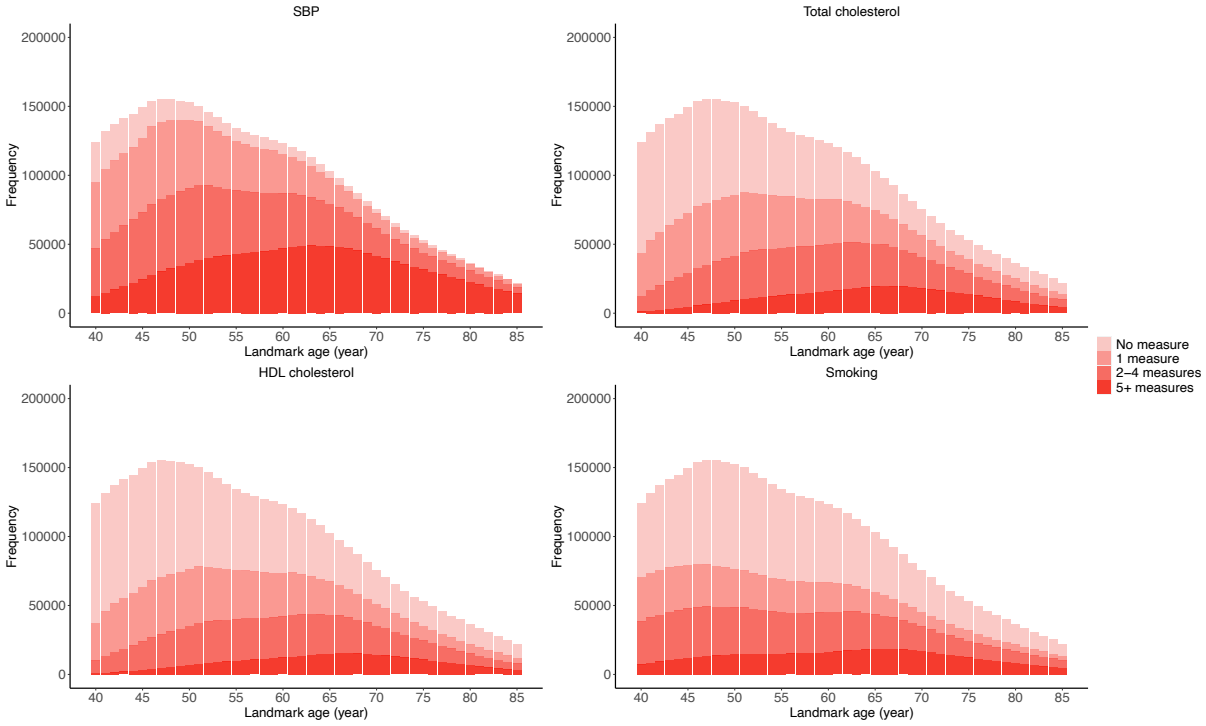


Figure 3.6: Number of participants and number of exposures at the start of each age among men in CPRD.

Abbreviations: HDL, high-density lipoprotein; SBP, systolic blood pressure

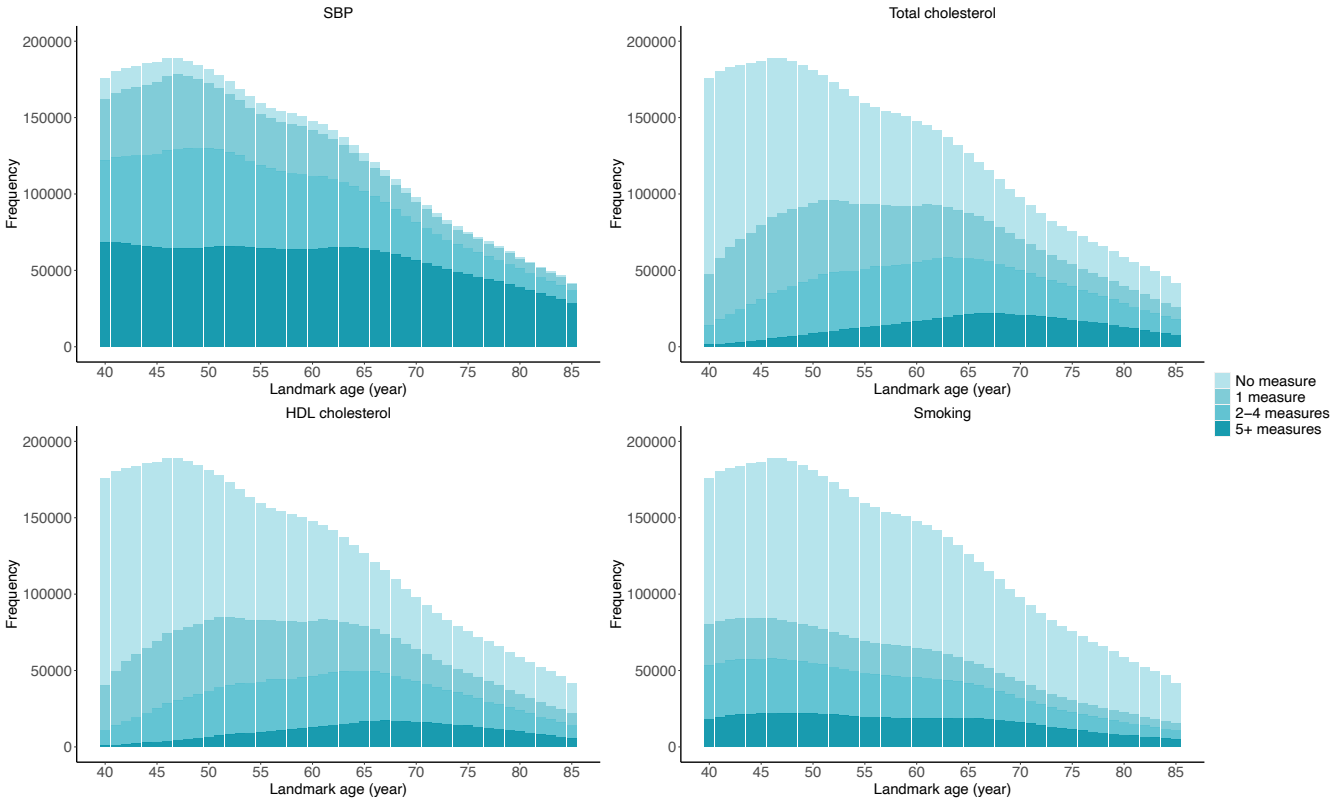


Figure 3.7: Number of participants and number of exposures at the start of each age among women in CPRD.

Abbreviations: HDL, high-density lipoprotein; SBP, systolic blood pressure

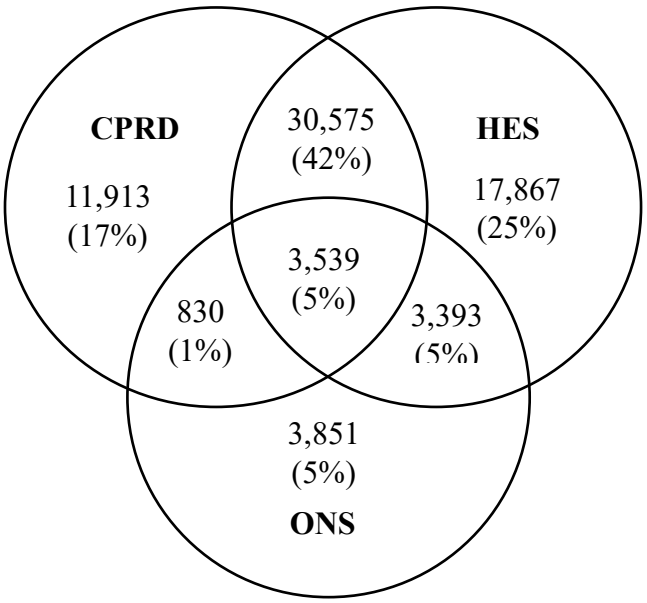


Figure 3.8: Venn diagram of incident cardiovascular events in the derivation dataset in CPRD

Events during follow-up in the derivation dataset recorded from primary care data in Clinical Practice Research Datalink (CPRD) (n=46,857 first events identified), secondary care data in Hospital Episode Statistics (HES) (n=55,374 first events identified), and mortality records in Office for National Statistics (ONS) (n=11,613 first events identified), England, United Kingdom, 2004-2019

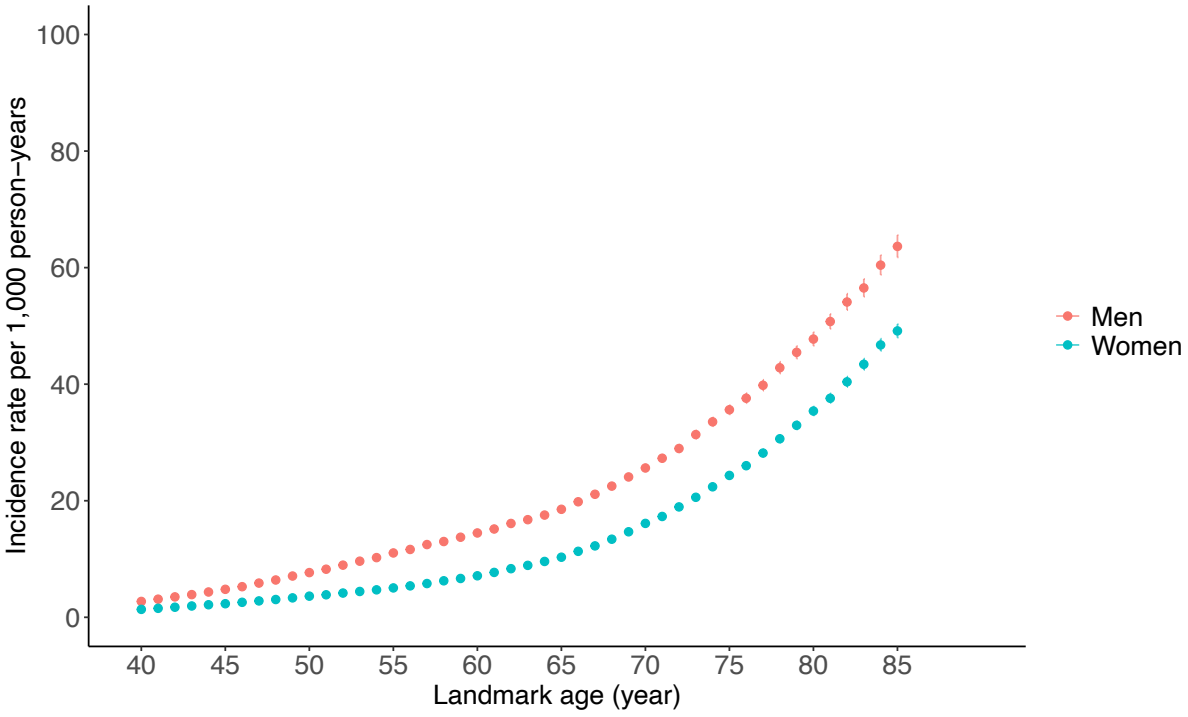


Figure 3.9: Crude cardiovascular disease incidence rate by age and sex in the derivation dataset in CPRD.

3.3.2 Derivation of the eHEART model

We estimated person-level risk predictor values for smoking status, SBP, HDL and total cholesterol at each age of follow-up; all estimated model coefficients are provided (**Table 3.4**). Mean levels of estimated SBP and HDL cholesterol were higher at older ages, as were the proportions of individuals on hypertensive treatment and with diagnosed diabetes. Estimated levels of total cholesterol remained stable across age, whilst the proportion of smokers declined with age (**Figure 3.10**).

The eHEART age- and sex-specific hazard ratios for SBP, HDL, total cholesterol, smoking and history of diabetes steadily attenuated towards 1 at older ages, whilst the hazard ratios for hypertensive treatment declined to a lesser extent, especially amongst women (**Figure 3.11**). Predicted eHEART risks increased at older ages and changed in distribution from a positive skew to bimodal (due to the contribution of the risk factor for hypertensive treatment) (**Figure 3.12**).

Table 3.4: Fixed intercepts and slopes from the age- and sex-specific multivariate mixed-effects models in derivation dataset

Men	Fixed intercepts						Fixed slopes			
Age	SBP	Total cholesterol	HDL cholesterol	Current smoker	Hypertension treatment	statin use	SBP	Total cholesterol	HDL cholesterol	Current smoker
40	-0.557	0.337	-0.178	0.549	0.067	-0.764	-0.001	0.011	0.000	-0.015
41	-0.543	0.351	-0.162	0.539	0.063	-0.808	-0.002	0.009	0.001	-0.015
42	-0.524	0.370	-0.147	0.527	0.061	-0.810	-0.002	0.007	0.002	-0.016
43	-0.506	0.390	-0.125	0.520	0.059	-0.828	-0.002	0.006	0.005	-0.016
44	-0.490	0.408	-0.115	0.512	0.046	-0.853	-0.002	0.004	0.005	-0.015
45	-0.471	0.430	-0.108	0.506	0.032	-0.883	-0.003	0.003	0.005	-0.015
46	-0.454	0.443	-0.097	0.499	0.020	-0.900	-0.004	0.001	0.004	-0.014
47	-0.437	0.461	-0.087	0.491	0.019	-0.910	-0.005	0.000	0.005	-0.014
48	-0.415	0.476	-0.074	0.482	0.012	-0.924	-0.005	-0.002	0.004	-0.014
49	-0.388	0.487	-0.060	0.473	0.001	-0.937	-0.004	-0.004	0.005	-0.014
50	-0.362	0.493	-0.045	0.465	-0.011	-0.949	-0.004	-0.005	0.004	-0.013
51	-0.336	0.501	-0.037	0.452	-0.027	-0.952	-0.004	-0.006	0.003	-0.014
52	-0.311	0.503	-0.025	0.441	-0.036	-0.964	-0.005	-0.008	0.003	-0.014
53	-0.283	0.506	-0.012	0.427	-0.044	-0.978	-0.006	-0.010	0.003	-0.014
54	-0.257	0.509	-0.003	0.414	-0.049	-0.993	-0.006	-0.011	0.002	-0.015
55	-0.224	0.501	0.011	0.401	-0.060	-0.995	-0.007	-0.014	0.002	-0.015
56	-0.196	0.494	0.021	0.391	-0.072	-0.999	-0.009	-0.017	0.002	-0.015
57	-0.169	0.495	0.035	0.378	-0.084	-1.004	-0.011	-0.017	0.002	-0.015
58	-0.148	0.493	0.046	0.362	-0.086	-1.020	-0.013	-0.019	0.002	-0.015
59	-0.122	0.490	0.059	0.351	-0.100	-1.031	-0.014	-0.020	0.002	-0.015
60	-0.095	0.482	0.074	0.336	-0.106	-1.035	-0.015	-0.022	0.003	-0.015
61	-0.069	0.476	0.091	0.322	-0.117	-1.045	-0.016	-0.022	0.003	-0.015
62	-0.052	0.472	0.097	0.307	-0.127	-1.057	-0.019	-0.022	0.002	-0.015
63	-0.034	0.467	0.104	0.293	-0.134	-1.061	-0.021	-0.022	0.001	-0.016
64	-0.023	0.456	0.109	0.278	-0.145	-1.069	-0.024	-0.023	0.001	-0.016
65	-0.010	0.447	0.118	0.265	-0.155	-1.073	-0.026	-0.023	0.000	-0.015
66	0.004	0.438	0.121	0.249	-0.159	-1.070	-0.027	-0.023	0.000	-0.016
67	0.012	0.428	0.128	0.238	-0.163	-1.075	-0.029	-0.023	0.000	-0.015
68	0.027	0.419	0.140	0.223	-0.169	-1.080	-0.029	-0.023	0.000	-0.015
69	0.035	0.406	0.143	0.208	-0.170	-1.082	-0.031	-0.023	0.000	-0.015
70	0.045	0.393	0.153	0.199	-0.174	-1.081	-0.033	-0.023	0.001	-0.015
71	0.053	0.372	0.160	0.187	-0.167	-1.074	-0.033	-0.024	0.001	-0.014

72	0.065	0.355	0.163	0.178	-0.171	-1.077	-0.034	-0.025	0.000	-0.014
73	0.070	0.339	0.165	0.168	-0.171	-1.079	-0.036	-0.026	0.000	-0.013
74	0.089	0.323	0.169	0.159	-0.178	-1.076	-0.035	-0.025	0.001	-0.012
75	0.097	0.307	0.171	0.151	-0.180	-1.076	-0.037	-0.025	0.001	-0.012
76	0.108	0.298	0.183	0.144	-0.178	-1.081	-0.037	-0.023	0.002	-0.011
77	0.114	0.288	0.186	0.140	-0.178	-1.082	-0.038	-0.022	0.002	-0.010
78	0.121	0.268	0.195	0.136	-0.175	-1.076	-0.038	-0.022	0.003	-0.010
79	0.125	0.249	0.212	0.128	-0.174	-1.068	-0.039	-0.022	0.004	-0.009
80	0.133	0.238	0.225	0.123	-0.183	-1.069	-0.042	-0.021	0.004	-0.009
81	0.130	0.216	0.234	0.117	-0.176	-1.062	-0.044	-0.022	0.004	-0.008
82	0.132	0.201	0.254	0.110	-0.171	-1.063	-0.045	-0.022	0.005	-0.008
83	0.126	0.186	0.267	0.105	-0.165	-1.062	-0.046	-0.022	0.004	-0.007
84	0.126	0.157	0.275	0.100	-0.175	-1.051	-0.048	-0.023	0.004	-0.007
85	0.112	0.125	0.288	0.097	-0.166	-1.043	-0.050	-0.026	0.004	-0.006
Women				**Fixed intercepts**					**Fixed slopes**	
Age	**SBP**	**Total cholesterol**	**HDL cholesterol**	**Current smoker**	**Hypertension treatment**	**statin use**	**SBP**	**Total cholesterol**	**HDL cholesterol**	**Current smoker**
40	-0.845	-0.324	-0.220	0.470	0.137	-0.560	0.018	0.013	0.005	-0.013
41	-0.810	-0.319	-0.192	0.470	0.131	-0.614	0.019	0.011	0.007	-0.012
42	-0.767	-0.281	-0.177	0.472	0.117	-0.641	0.021	0.015	0.007	-0.012
43	-0.723	-0.243	-0.160	0.474	0.097	-0.667	0.023	0.017	0.009	-0.011
44	-0.680	-0.209	-0.149	0.473	0.085	-0.702	0.023	0.017	0.009	-0.011
45	-0.638	-0.168	-0.134	0.472	0.078	-0.736	0.024	0.019	0.010	-0.011
46	-0.595	-0.132	-0.110	0.472	0.063	-0.758	0.024	0.019	0.011	-0.011
47	-0.553	-0.082	-0.091	0.470	0.057	-0.779	0.024	0.022	0.012	-0.011
48	-0.511	-0.028	-0.069	0.467	0.048	-0.803	0.023	0.025	0.013	-0.011
49	-0.472	0.026	-0.045	0.465	0.039	-0.819	0.022	0.028	0.014	-0.011
50	-0.435	0.087	-0.008	0.457	0.030	-0.848	0.020	0.031	0.017	-0.012
51	-0.400	0.147	0.024	0.451	0.025	-0.880	0.018	0.033	0.018	-0.012
52	-0.365	0.215	0.054	0.441	0.019	-0.905	0.015	0.036	0.019	-0.013
53	-0.333	0.281	0.082	0.429	0.008	-0.943	0.013	0.037	0.019	-0.014
54	-0.304	0.346	0.099	0.420	0.007	-0.973	0.010	0.038	0.018	-0.014
55	-0.273	0.404	0.110	0.407	-0.001	-0.998	0.007	0.039	0.017	-0.014
56	-0.243	0.447	0.125	0.397	0.001	-1.026	0.005	0.036	0.015	-0.014
57	-0.212	0.488	0.126	0.384	-0.006	-1.053	0.002	0.033	0.012	-0.015
58	-0.179	0.522	0.132	0.370	-0.005	-1.073	0.000	0.029	0.011	-0.016
59	-0.150	0.546	0.125	0.360	-0.007	-1.096	-0.001	0.024	0.007	-0.016

60	-0.123	0.560	0.124	0.346	-0.007	-1.112	-0.003	0.019	0.005	-0.016
61	-0.092	0.565	0.123	0.335	-0.012	-1.129	-0.005	0.014	0.003	-0.016
62	-0.061	0.575	0.117	0.321	-0.022	-1.136	-0.007	0.009	0.002	-0.016
63	-0.033	0.586	0.109	0.307	-0.025	-1.151	-0.008	0.006	0.000	-0.017
64	-0.007	0.593	0.110	0.295	-0.032	-1.155	-0.009	0.003	-0.001	-0.017
65	0.022	0.592	0.113	0.281	-0.042	-1.160	-0.010	0.000	-0.001	-0.017
66	0.049	0.590	0.120	0.268	-0.048	-1.171	-0.011	-0.003	0.000	-0.017
67	0.080	0.594	0.116	0.259	-0.057	-1.181	-0.012	-0.004	0.000	-0.017
68	0.113	0.595	0.121	0.247	-0.066	-1.190	-0.013	-0.006	0.001	-0.017
69	0.142	0.594	0.128	0.239	-0.071	-1.197	-0.014	-0.008	0.003	-0.016
70	0.171	0.593	0.134	0.229	-0.074	-1.204	-0.014	-0.009	0.004	-0.016
71	0.200	0.593	0.143	0.217	-0.084	-1.205	-0.016	-0.009	0.006	-0.016
72	0.225	0.590	0.145	0.212	-0.088	-1.207	-0.018	-0.011	0.005	-0.015
73	0.252	0.582	0.152	0.208	-0.091	-1.197	-0.019	-0.013	0.007	-0.014
74	0.276	0.579	0.155	0.199	-0.097	-1.197	-0.021	-0.013	0.006	-0.013
75	0.294	0.567	0.164	0.193	-0.099	-1.196	-0.023	-0.015	0.007	-0.013
76	0.317	0.560	0.171	0.185	-0.106	-1.188	-0.024	-0.015	0.008	-0.012
77	0.338	0.559	0.183	0.179	-0.114	-1.183	-0.025	-0.015	0.008	-0.011
78	0.355	0.552	0.191	0.173	-0.118	-1.180	-0.026	-0.015	0.008	-0.011
79	0.374	0.547	0.207	0.169	-0.129	-1.181	-0.028	-0.015	0.008	-0.010
80	0.386	0.536	0.214	0.166	-0.127	-1.174	-0.029	-0.016	0.008	-0.009
81	0.399	0.521	0.221	0.156	-0.131	-1.167	-0.031	-0.017	0.007	-0.009
82	0.416	0.503	0.233	0.147	-0.133	-1.158	-0.032	-0.018	0.007	-0.008
83	0.427	0.494	0.254	0.142	-0.133	-1.158	-0.032	-0.018	0.007	-0.008
84	0.439	0.481	0.260	0.137	-0.135	-1.161	-0.033	-0.019	0.007	-0.007
85	0.434	0.468	0.273	0.122	-0.129	-1.158	-0.037	-0.020	0.008	-0.008

Abbreviations: HDL, high-density lipoprotein; SBP, systolic blood pressure. Derivation dataset: Clinical Practice Research Datalink, Hospital Episode Statistics, and the Office for National Statistics, England, United Kingdom, 2004-2019

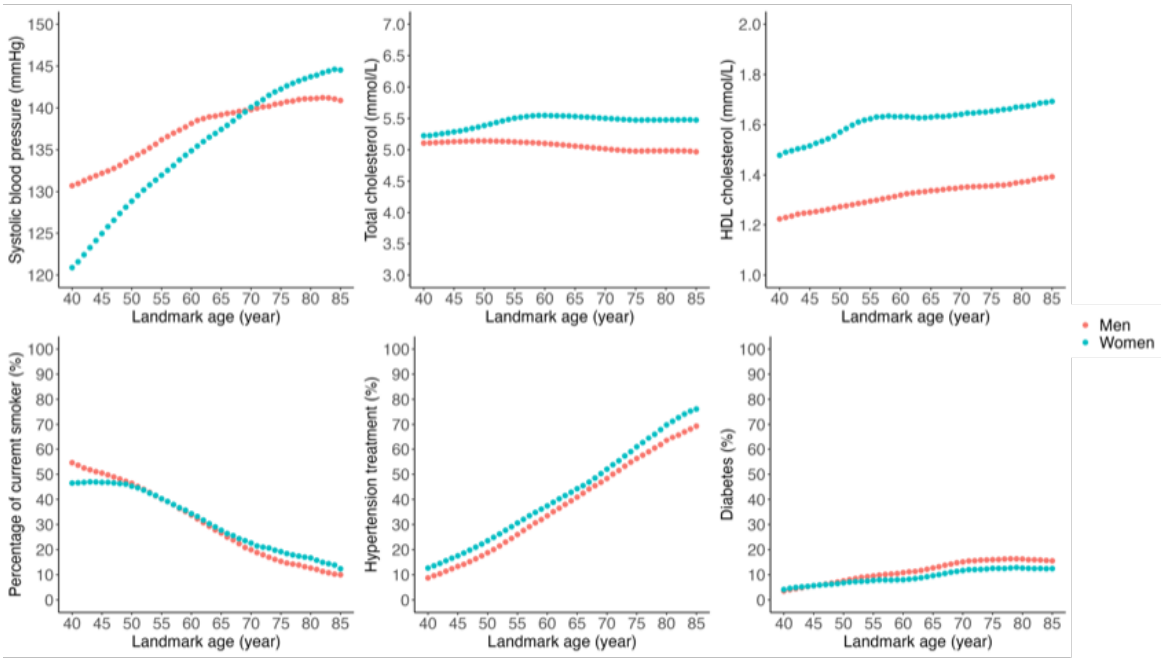


Figure 3.10: Mean level of estimated risk factors for SBP, total cholesterol, HDL cholesterol, percentage of current smokers, percentage of people on blood pressure lowering medication, and percentage of people with diabetes by age in the derivation dataset in CPRD.

Abbreviations: HDL, high-density lipoprotein

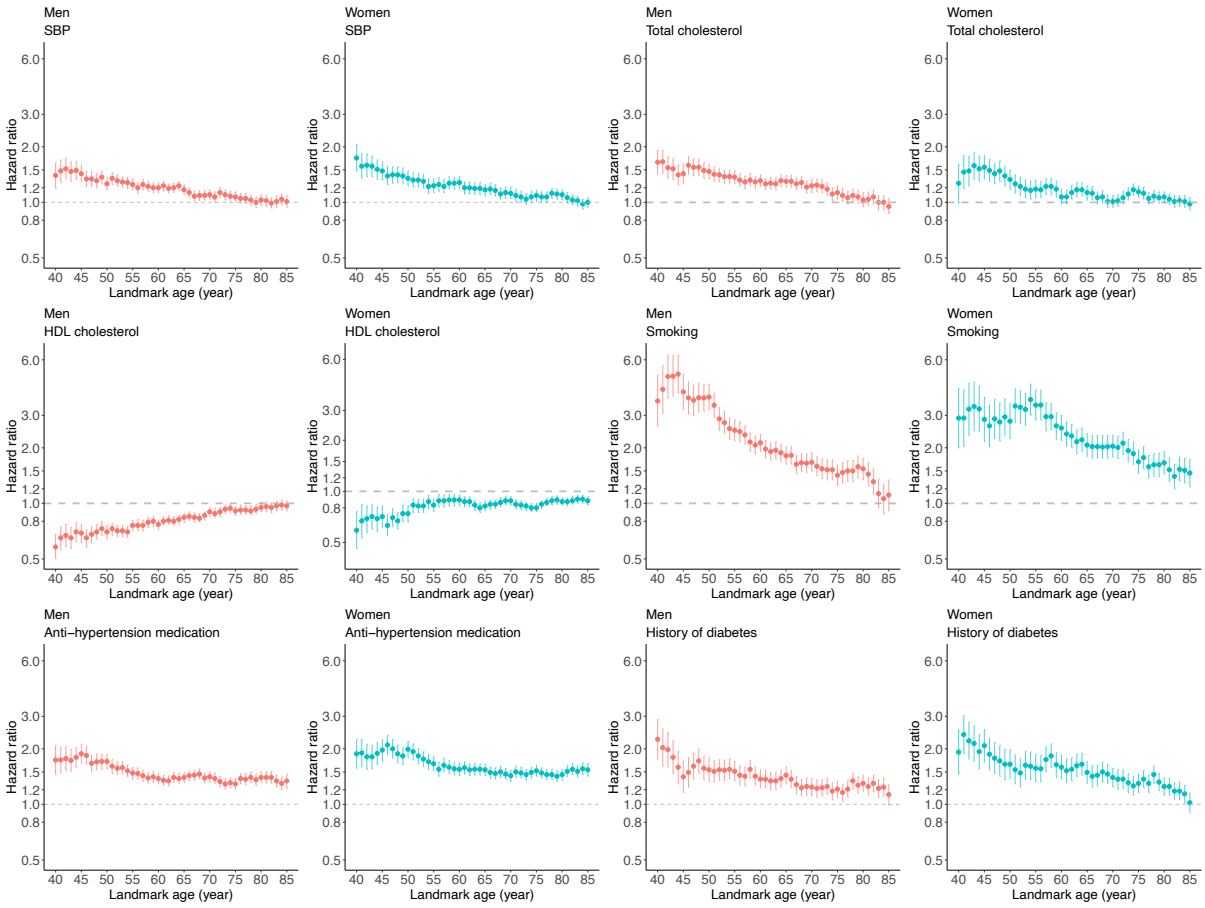


Figure 3.11: Age and sex-specific hazard ratios for association of eHEART risk predictors with cardiovascular disease in the derivation cohort

Abbreviations: HDL, high-density lipoprotein; SBP, systolic blood pressure

Hazard ratios and 95% confidence intervals (shown as vertical lines) for association of cardiovascular disease with: systolic blood pressure, total cholesterol, HDL cholesterol, smoking status, anti-hypertension medication and history of diabetes by age, in men and women in the derivation dataset, Clinical Practice Research Datalink, Hospital Episode Statistics, and the Office for National Statistics, England, United Kingdom, 2004–2019. Hazard ratios are given per standard-deviation increase for SBP, total cholesterol, and HDL cholesterol. Hazard ratios and 95% confidence intervals are shown on the natural log scale.

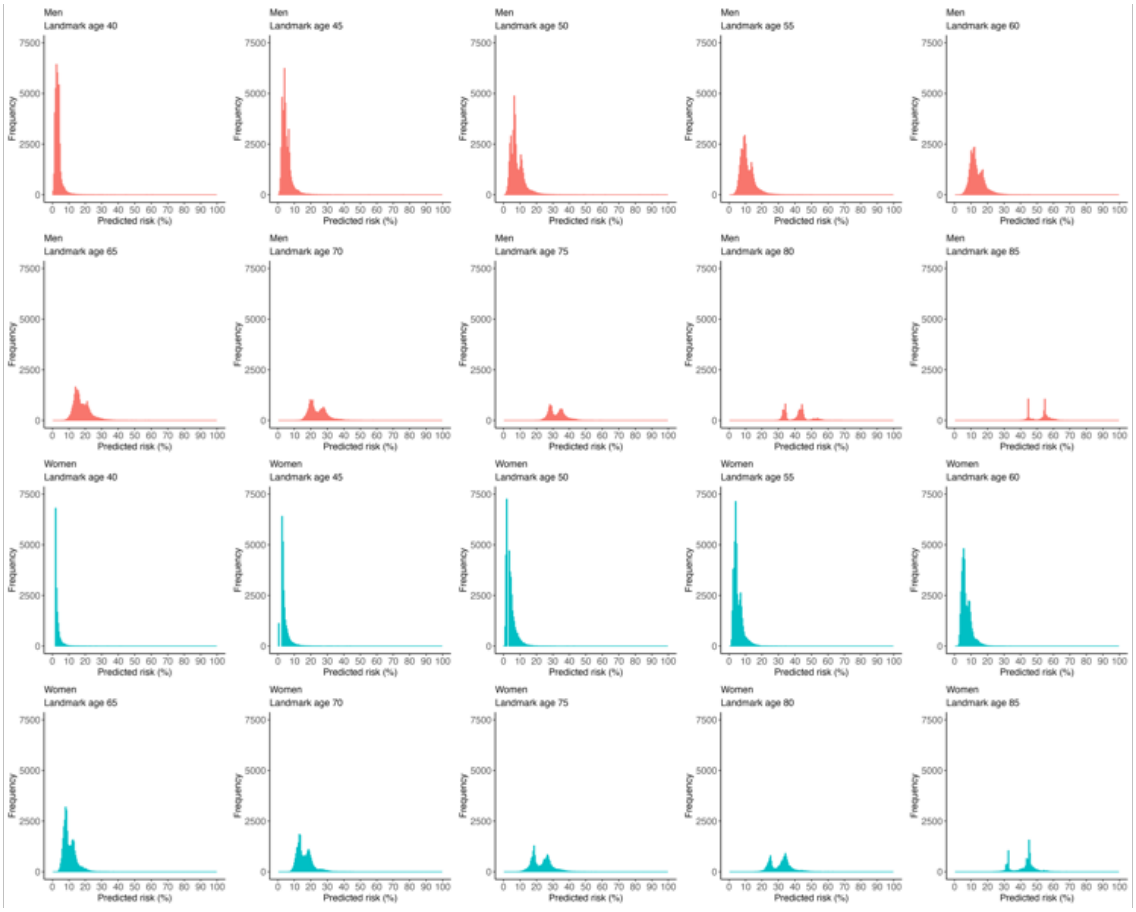


Figure 3.12: Histogram of predicted 10-year cardiovascular disease risk distributions by age group and sex in the validation dataset in CPRD

3.3.3 Internal validation of eHEART

In the validation cohort, the overall C-index for eHEART was 0.771 (95% CI: 0.769, 0.772) and was higher in women (C-index = 0.786, 95% CI:0.784, 0.788) than in men (C-index = 0.741, 95% CI: 0.739, 0.742) and higher in younger individuals (**Figure 3.13**). The Brier score was also lower (better) in women (0.2354, 95% CI: 0.2340, 0.2368) than in men (0.3085, 95% CI: 0.3068, 0.3102) (**Table 3.5**). Calibration plots by decile of eHEART risk showed good agreement with observed risk (**Figures 3.14-3.15**). The calibration slope was more deviatory from 1 for older ages (**Figure 3.16**).

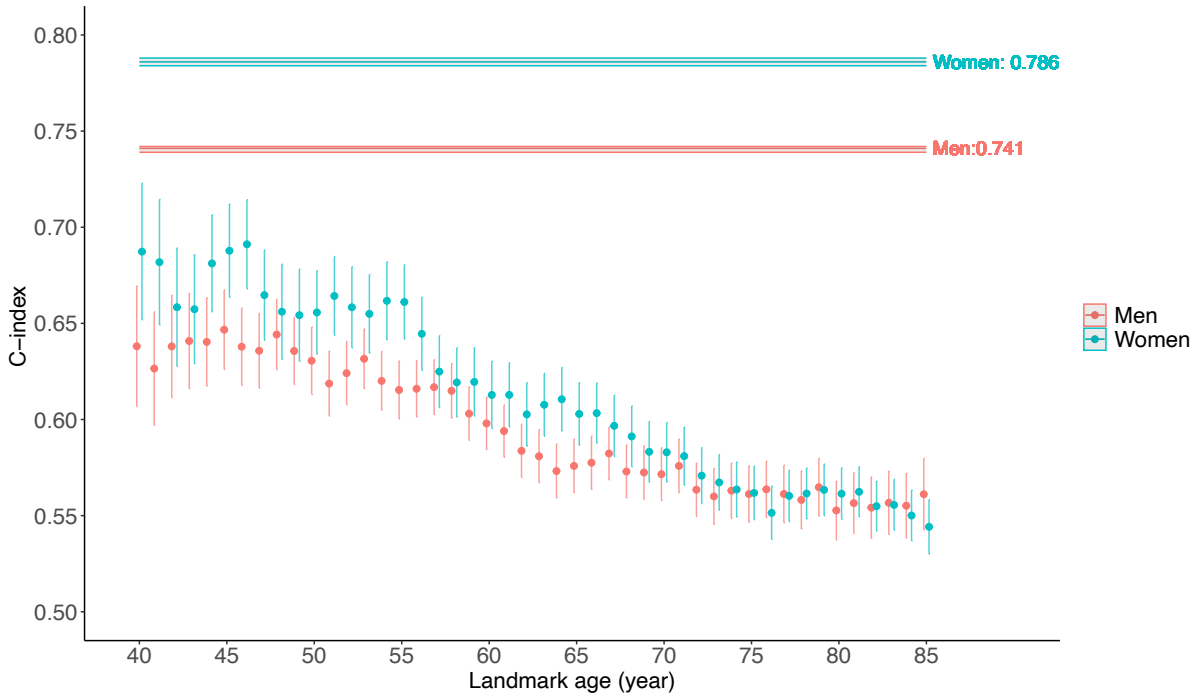


Figure 3.13: Age and sex-specific C indices of eHEART in validation cohort

C-indices and 95% confidence intervals (shown as vertical lines) from eHEART for the prediction of 10-year cardiovascular disease by age and for all ages, in men and women in the validation dataset, Clinical Practice Research Datalink, Hospital Episode Statistics, and the Office for National Statistics, England, United Kingdom, 2004–2019. The overall C-indices account for the discriminatory information in age, hence are higher than the age-specific C-indices.

Table 3.5: Brier score in the validation dataset.

Sex	Brier score (95% CI)
Overall	0.2684 (0.2671, 0.2697)
Men	0.3085 (0.3068, 0.3102)
Women	0.2354 (0.2340, 0.2368)

Validation dataset: Clinical Practice Research Datalink, Hospital Episode Statistics, and the Office for National Statistics, England, United Kingdom, 2004-2019

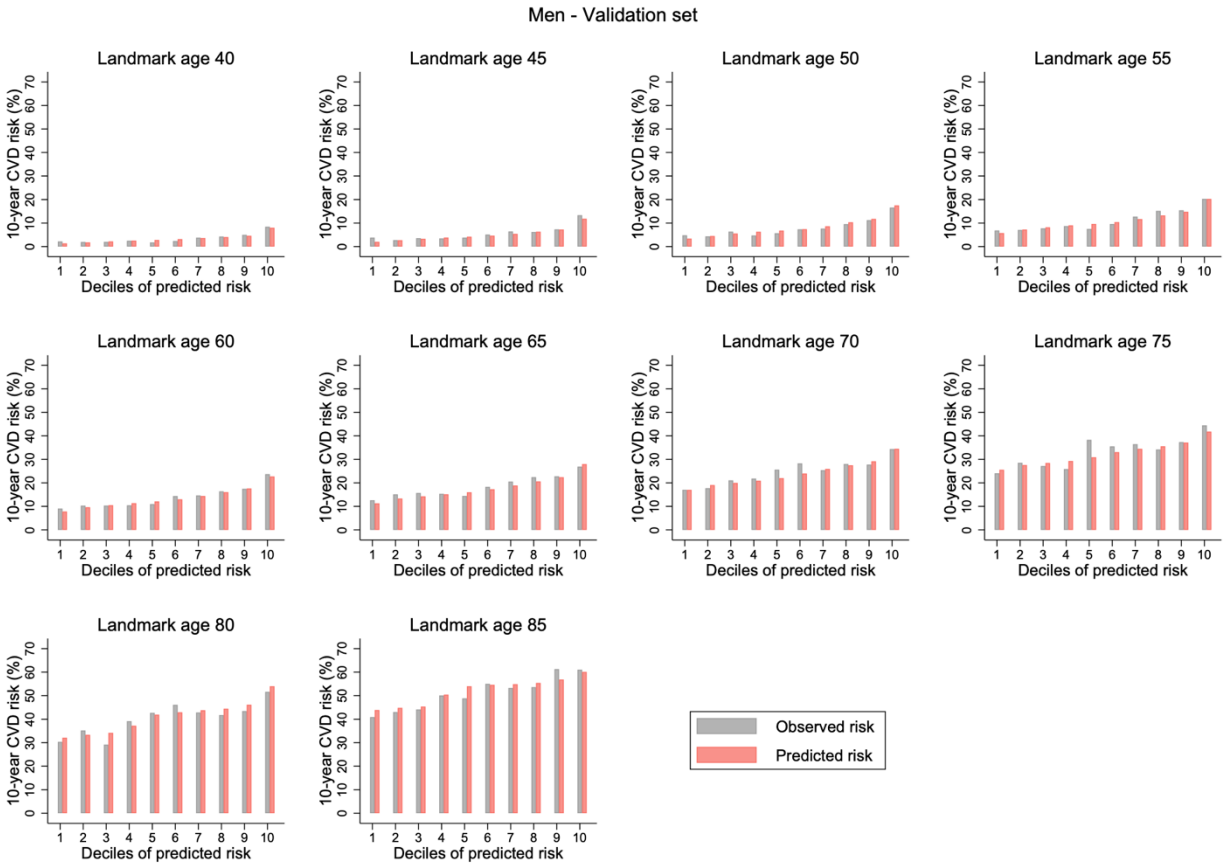


Figure 3.14: Calibration plots by deciles of predicted risk at 5-year age groups among men in the validation dataset in CPRD.

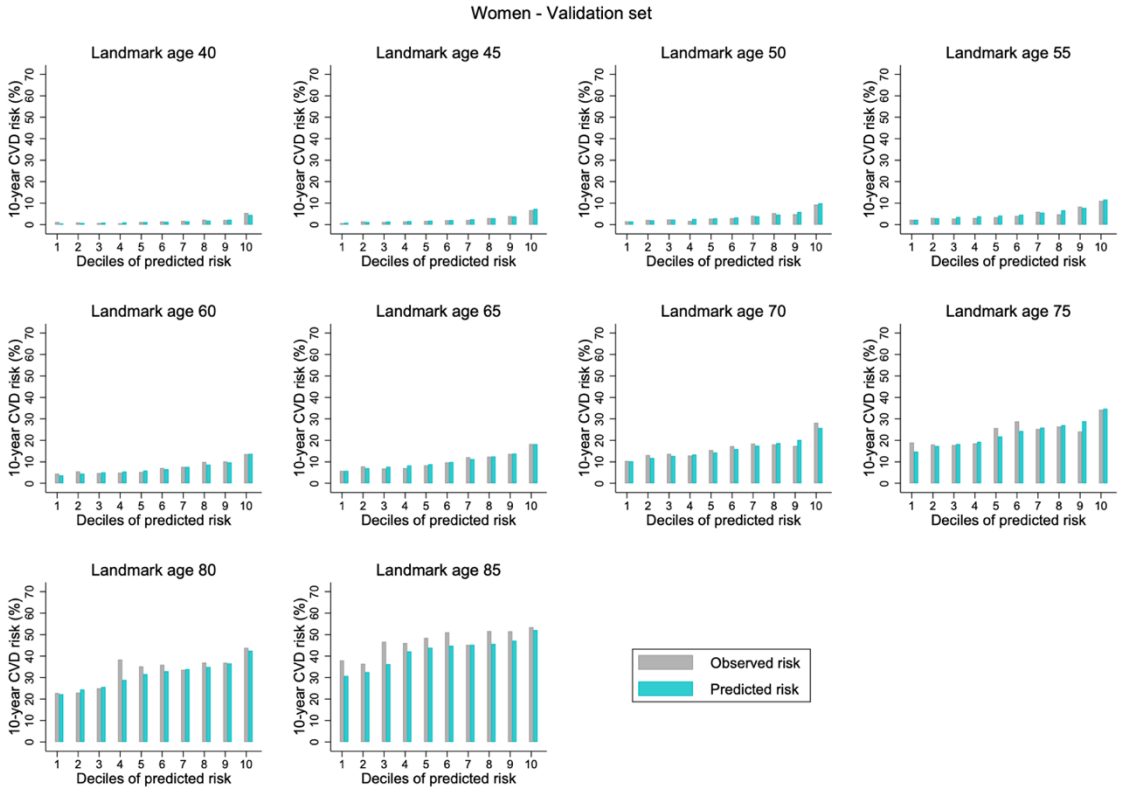


Figure 3.15: Calibration plots by deciles of predicted risk at 5-year age groups among women in the validation dataset in CPRD.

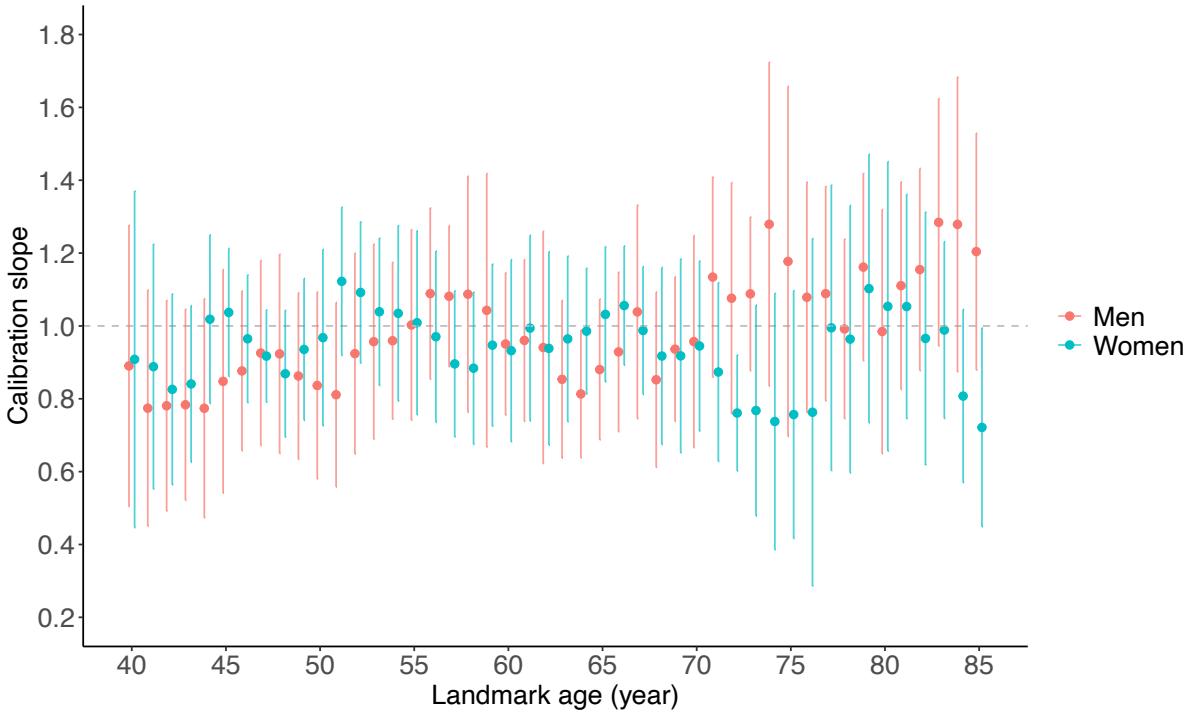


Figure 3.16: Calibration slopes by landmark age among men and women in the validation dataset in CPRD.

3.3.4 Population health modelling

Overall, 119,137 individuals from UKB, with primary care records available, were used to model a representative population of 50,000 men and 50,000 women aged 40-69 in the UK (**Figure 3.17 and Table 3.6**). We estimated that there would be 3,566 and 1,819 CVD events over the next 10 years in men and women respectively. If QRISK2 was used as a formal assessment on the whole population, then 2,654 (74.4%) men and 831 (45.7%) women with CVD events over the next 10 years would be classified at high risk, i.e., QRISK2 ≥10% (**Table 3.7**). These percentages of "events captured" varied considerably by age group; for example, only 19% of events amongst 40–49-year-old men but 98% of events amongst 60–69-year-old men would be captured. Assuming statin therapy would be initiated for persons with predicted QRISK2 ≥10%, the number needed to screen to prevent one CVD event in men and in women would be 188 and 602 respectively. Assuming 50% invitation uptake, the number needed to invite to prevent one CVD event in men and in women would be 377 and 1203 respectively (**Table 3.8**).

If the eHEART tool with a prioritisation threshold of 10% was used to prioritise QRISK2 assessment in the population (and assuming all individuals had at least one primary care record of either smoking, SBP, HDL or total cholesterol) then 2,209 (61.9%) men and 471 (25.9%) women with CVD events over the next 10 years would be classified at high risk (**Table 3.7)**. The number needed to screen to prevent one CVD event in men and in women would be 92 and 121 respectively. The number needed to invite to prevent one CVD event in men and in women would be 168 and 219 respectively (**Table 3.8**).

In contrast, using age- and sex-specific prioritisation thresholds corresponding to 5% false negative rates would classify 2,560 (71.8%) men and 805 (44.3%) women with CVD events over the next 10 years as high risk (**Table 3.7)**. The number needed to screen to prevent one CVD event in men and in women would be 152 and 346 respectively. Assuming 55% invitation uptake after prioritisation, the number need to invite to prevent one CVD event in men and in women would be 277 and 629 respectively (**Table 3.8**).

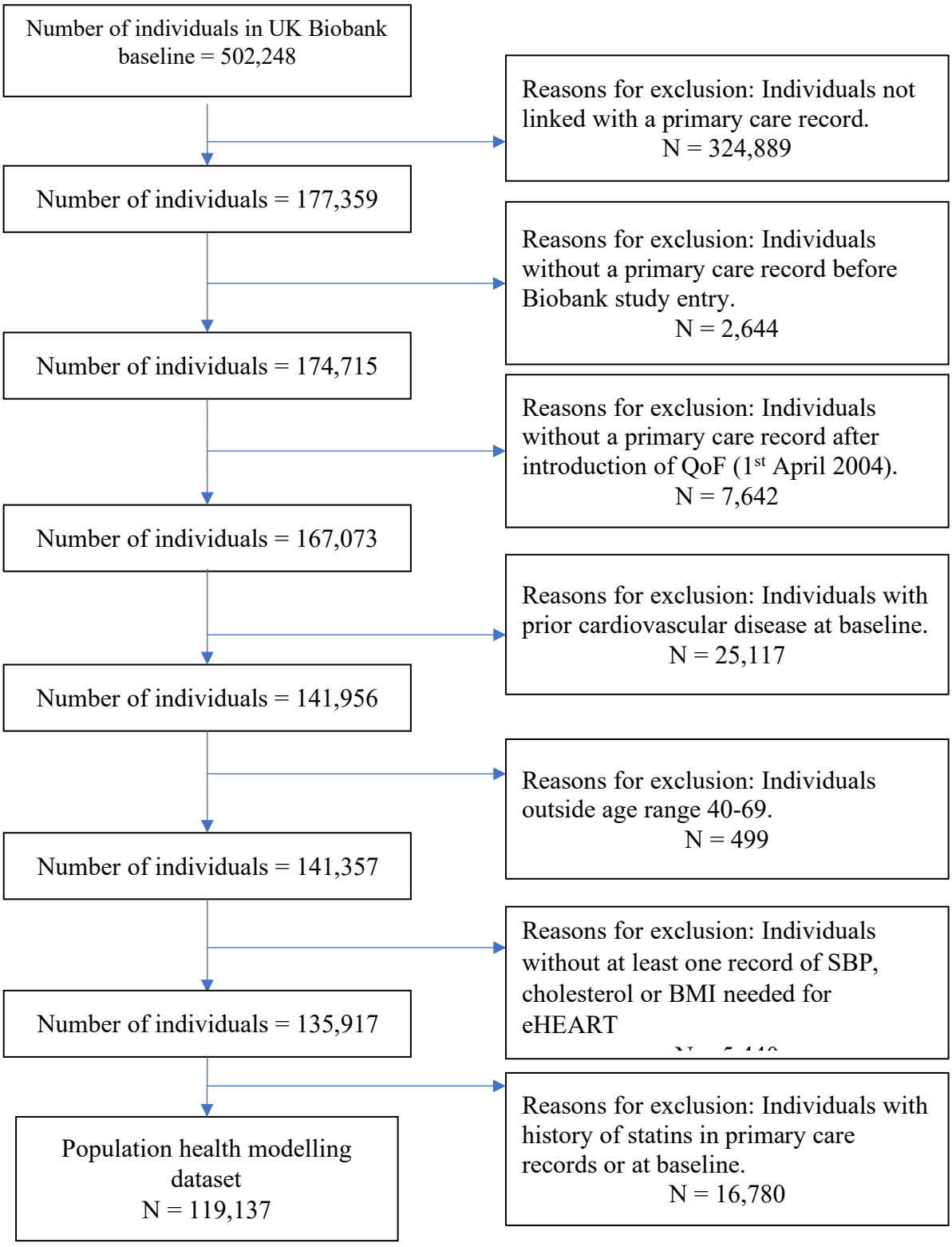


Figure 3.17: Flowchart showing selection of individuals for population health modelling in UK Biobank

Table 3.6: Key characteristics of individuals in the population health modelling in UK Biobank

	Primary care records				UK Biobank baseline			
	Men, N = 49,111		Women, N = 70,026		Men, N = 49,111		Women, N = 70,026	
Characteristics	**Mean (SD) or n (%)**	**No. (%) of persons with value**	**Mean (SD) or n (%)**	**No. (%) of persons with value**	**Mean (SD) or n (%)**	**No. (%) of persons with value**	**Mean (SD) or n (%)**	**No. (%) of persons with value**
Age at baseline (years)	55.9 (8.1)	49,111 (100%)	55.9 (7.9)	70,026 (100%)	55.9 (8.1)	49,111 (100%)	55.9 (7.9)	70,026 (100%)
History of diabetes	487 (1)	49,111 (100%)	436 (1)	70,026 (100%)	665 (1)	48,994 (100%)	484 (1)	69,904 (100%)
Blood pressure-lowering medication prescriptions	6,707 (14)	49,111 (100%)	10,198 (15)	70,026 (100%)	5,694 (12)	48,548 (99%)	7,920 (11)	69,557 (99%)
Current smoker	4,955 (10)	49,111 (100%)	5,106 (7)	70,026 (100%)	6,084 (12)	49,029 (100%)	6,116 (9)	69,939 (100%)
Systolic blood pressure mm Hg, mean (SD) [#]	131.6 (14.1)	39,994 (81%)	127.1 (15.0)	59,491 (85%)	140.8 (17.3)	49,025 (100%)	134.8 (19.1)	69,935 (100%)
Total cholesterol mmol/litre, mean (SD) [#]	5.4 (0.9)	24,844 (51%)	5.6 (0.9)	33,195 (47%)	5.8 (1.0)	48,565 (99%)	6.0 (1.1)	68,807 (98%)
HDL cholesterol mmol/litre, mean (SD) [#]	1.4 (0.4)	21,783 (44%)	1.7 (0.4)	29,154 (42%)	1.3 (0.3)	48,474 (99%)	1.6 (0.4)	68,691 (98%)
BMI kg/m2, mean (SD) [#]	27.1 (4.3)	29,829 (61%)	26.5 (5.3)	45,606 (65%)	27.5 (4.1)	48,975 (100%)	26.8 (5.0)	69,904 (100%)
Ethnicity – white[§]	46,853 (96)	49,014 (100%)	67,110 (96)	69,928 (100%)	46,853 (96)	49,014 (100%)	67,110 (96)	69,928 (100%)
Townsend score, mean (SD) [§]	-1.5 (3.0)	49,106 (100%)	-1.5 (2.9)	70,022 (100%)	-1.5 (3.0)	49,106 (100%)	-1.5 (2.9)	70,022 (100%)
Family history of CVD^	1,647 (3)	49,111 (100%)	2,615 (4)	70,026 (100%)	1,647 (3)	49,111 (100%)	2,615 (4)	70,026 (100%)
Chronic kidney disease^	60 (0.1)	49,111 (100%)	83 (0.1)	70,026 (100%)	60 (0.1)	49,111 (100%)	83 (0.1)	70,026 (100%)
Rheumatoid arthritis	175 (100.0)	49,111 (100%)	369 (100.0)	70,026 (100%)	414 (0.8)	49,111 (100%)	1,080 (2)	70,026 (100%)
Atrial fibrillation	635 (100.0)	49,111 (100%)	1,871 (100.0)	70,026 (100%)	137 (0.3)	49,111 (100%)	94 (0.1)	70,026 (100%)

[§]Risk factor was only available at baseline and was used when conducting population health modelling with primary care records.

^Risk factor was only available in primary care records and was used when conducting population health modelling with baseline data.

[#]Mean (standard deviation) of last observed measurement before UK Biobank baseline.

Table 3.7: Number needed to screen to prevent one event, and number of events captured when prioritising with eHEART in a hypothetical population of 100,000 individuals in England assuming all individuals invited for formal assessment attend.

Hypothetical population of 100,000				Full formal assessment only (QRISK2 threshold=10%)		Prioritisation tool (eHEART threshold=10%) followed by full formal assessment (QRISK2 threshold=10%)			Prioritisation tool (eHEART threshold corresponding to age- and sex-specific 5% false negative rates) followed by full formal assessment (QRISK2 threshold=10%)			
Sex	Age group	Expected N	Expected CVD events in 10 years	Events captured N (%)	Number needed to screen to prevent 1 CVD event	Individuals prioritised N	Events captured N (%)	Number needed to screen to prevent 1 CVD event	Individuals prioritised N	Prioritisation threshold for eHEART	Events captured N (%)	Number needed to screen to prevent 1 CVD event
Men	40-49	17673	516	96 (18.6%)	1832.8 (1335.2, 2168.8)	598	32 (6.3%)	183.2 (76.5, 236.5)	11054	3.0%	91 (17.7%)	1204.7 (849.1, 1435.1)
	50-59	18061	1294	844 (65.2%)	214.1 (204.3, 223.2)	6348	518 (40.0%)	122.7 (112.6, 131.0)	14576	6.2%	810 (62.6%)	180.0 (170.7, 188.1)
	60-69	14266	1756	1714 (97.6%)	83.2 (82.6, 83.8)	13391	1659 (94.5%)	80.8 (79.8, 81.6)	13391	10%	1659 (94.5%)	80.8 (79.8, 81.6)
	Total	**50000**	**3566**	**2654 (74.4%)**	**188.4 (185.2, 191.5)**	**20338**	**2209 (61.9%)**	**92.1 (90.4, 93.9)**	**39021**	**NA**	**2560 (71.8%)**	**152.4 (149.6, 155.1)**
Women	40-49	17488	277	17 (6.1%)	10605.6 (643.6, 14255.1)	78	7 (2.5%)	123.7 (0.0, 183.1)	5495	2.0%	16 (5.7%)	3610.2 (0.0, 4933.9)
	50-59	17986	607	116 (19.1%)	1552.2 (1254.2, 1787.3)	427	28 (4.5%)	156.4 (75.4, 198.2)	10519	3.3%	113 (18.7%)	930.5 (741.5, 1075.0)
	60-69	14526	936	699 (74.7%)	207.9 (201.1, 214.4)	5174	437 (46.7%)	118.4 (110.5, 124.6)	11850	5.9%	677 (72.4%)	175.0 (168.7, 180.5)
	Total	**50000**	**1819**	**831 (45.7%)**	**601.6 (578.8, 622.7)**	**5680**	**471 (25.9%)**	**120.6 (113.1, 127.2)**	**27864**	**NA**	**805 (44.3%)**	**346.0 (332.1, 358.4)**

Age structure of hypothetical population extrapolated from Office for National Statistics, England, United Kingdom 2017. Expected events at 10 years based on extrapolation of incidence rates from CPRD, 2014-2019. Age group and sex specific prioritisation thresholds were defined as the minimum of 10% and the level such that the expected false negative rate is controlled to be 5%. All individuals have at least one CVD risk factor (systolic blood pressure, smoking, total and/or HDL cholesterol) recorded for eHEART. Number needed to screen based on assuming that all individuals are formally assessed. Under each scenario, some individuals are not formally assessed and may go on to have events. Statin compliance assumed to be equal to 50%.

Table 3.8: Number needed to invite to prevent one event and number of events captured when prioritising with eHEART in a hypothetical population of 100,000 individuals in England assuming 55% of those invited for formal assessment attend.

Hypothetical population of 100,000				Full formal assessment only (QRISK2 threshold=10%)		Prioritisation tool (eHEART threshold=10%) followed by full formal assessment (QRISK2 threshold=10%)			Prioritisation tool (eHEART threshold corresponding to age- and sex-specific 5% false negative rates) followed by full formal assessment (QRISK2 threshold=10%)			
Sex	Age group	Expected N	Expected CVD events in 10 years	Events captured N (%)	Number needed to invite to prevent 1 CVD event	Individuals prioritised N	Events captured N (%)	Number needed to invite to prevent 1 CVD event	Individuals prioritised N	Prioritisation threshold for eHEART	Events captured N (%)	Number needed to invite to prevent 1 CVD event
Men	40-49	17673	516	96 (18.6%)	3665.6 (2670.4, 4337.6)	598	32 (6.3%)	333.1 (139.1, 430.0)	11054	3.0%	91 (17.7%)	2190.4 (1543.8, 2609.3)
	50-59	18061	1294	844 (65.2%)	428.2 (408.6, 446.4)	6348	518 (40.0%)	223.1 (204.7, 238.2)	14576	6.2%	810 (62.6%)	327.3 (310.4, 342.0)
	60-69	14266	1756	1714 (97.6%)	166.4 (165.2, 167.6)	13391	1659 (94.5%)	146.9 (145.1, 148.4)	13391	10%	1659 (94.5%)	146.9 (145.1, 148.4)
	Total	**50000**	**3566**	**2654 (74.4%)**	**376.8 (370.4, 383.0)**	**20338**	**2209 (61.9%)**	**167.5 (164.4, 170.7)**	**39021**	**NA**	**2560 (71.8%)**	**277.1 (272.0, 282.0)**
Women	40-49	17488	277	17 (6.1%)	21211.2 (1278.2, 28510.2)	78	7 (2.5%)	224.9 (0.0, 332.9)	5495	2.0%	16 (5.7%)	6564.0 (0.0, 8970.7)
	50-59	17986	607	116 (19.1%)	3104.4 (2508.4, 3574.6)	427	28 (4.5%)	284.4 (137.1, 360.4)	10519	3.3%	113 (18.7%)	1691.8 (1348.2, 1954.5)
	60-69	14526	936	699 (74.7%)	415.8 (402.2, 428.8)	5174	437 (46.7%)	215.3 (200.9, 226.5)	11850	5.9%	677 (72.4%)	318.2 (306.7, 328.2)
	Total	**50000**	**1819**	**831 (45.7%)**	**1203.2 (1157.6, 1245.4)**	**5680**	**471 (25.9%)**	**219.3 (205.6 231.3)**	**27864**	**NA**	**805 (44.3%)**	**629.1 (603.8, 651.6)**

Age structure of hypothetical population extrapolated from Office for National Statistics, England, United Kingdom 2017. Expected events at 10 years based on extrapolation of incidence rates from CPRD, 2014-2019. Age group and sex specific prioritisation thresholds were defined as the minimum of 10% and the level such that the expected false negative rate is controlled to be 5%. All individuals have at least one CVD risk factor (systolic blood pressure, smoking, total and/or HDL cholesterol) recorded for eHEART. Statin compliance assumed to equal 50%. Number needed to invite for full formal assessment only assumes that 50% of all individuals are formally assessed. Number needed to invite after prioritisation with eHEART assumes 55% of prioritised individuals are formally assessed. Under each prioritisation scenario, some individuals are not formally assessed and may go on to have events.

3.3.5 Sensitivity analyses

In comparison to eHEART being used as a prioritisation tool, if estimated QRISK2 with a prioritisation threshold of 10% was used to prioritise formal assessment in the population, then 2,391 (67.1%) men and 649 (35.7%) women with CVD events over the next 10 years would be classified at high risk. This equates to capturing approximately 87% of events when applying formal CVD risk assessment on the whole population (**Table 3.9**). The number needed to screen to prevent one CVD event in men and in women would be 97 and 151, a reduction of 48% and 75%, respectively, when applying formal CVD risk assessment to the whole population. Notably, the events captured reduced by at least half amongst men under 50 years and women under 60 years.

In contrast, using age- and sex-specific prioritisation thresholds corresponding to 5% false negative rates would classify 2,606 (73.1%) men and 803 (44.2%) women with CVD events over the next 10 years as high risk, around 98% events captured when applying formal CVD risk assessment on the whole population. The number needed to screen to prevent one CVD event in men and in women would be 149 (a reduction of 21%) and 320 (a reduction of 47%) respectively, with greatest reductions observed amongst younger individuals. The number need to invite to prevent one CVD event in men and in women would be 270 and 582 respectively (**Table 3.10**).

In sensitivity analyses including all individuals (i.e., all those without a primary care record for any one of SBP, HDL, total cholesterol and smoking status) (**Table 3.11**), we found comparable results for eHEART and estimated QRISK2, but as expected, we observed greater numbers needed to screen especially among younger individuals (**Table 3.12-3.13**).

Comparable results were observed in additional analyses assuming a 5% (rather than 10%) formal risk assessment threshold (**Table 3.14-3.15**). Using a fixed 5% prioritisation threshold resulted in a greater number needed to screen to prevent one CVD, than using a fixed 10% threshold for both prioritisation and formal assessment, however 88.5% of future CVD events in men and 63.3% in women being captured if prioritising with eHEART, and 91.3% of events in men and 71.8% of events in women being captured if prioritising with estimated QRISK2. Using age- and sex-specific prioritisation thresholds to limit the false negative rate to 2.5% resulted in comparable results for eHEART and estimated QRISK2.

Table 3.9: Number needed to screen to prevent one event, and number of events captured when prioritising with QRISK2 in a hypothetical population of 100,000 individuals in England.

Hypothetical population of 100,000				Full formal assessment only (QRISK2 threshold=10%)		Prioritisation tool (QRISK2 prioritisation threshold=10%) followed by full formal assessment (QRISK2 threshold=10%)			Prioritisation tool (QRISK2 prioritisation threshold corresponding to age- and sex-specific 5% false negative rates) followed by full formal assessment (QRISK2 threshold=10%)			
Sex	Age group	Expected N	Expected CVD events in 10 years	Events captured N (%)	Number needed to screen to prevent 1 CVD event	Individuals prioritised N	Events captured N (%)	Number needed to screen to prevent 1 CVD event	Individuals prioritised N	Prioritisation threshold for QRISK2	Events captured N (%)	Number needed to screen to prevent 1 CVD event
Men	40-49	17673	516	96 (18.6%)	1832.8 (1335.2, 2168.8)	608	31 (6.0%)	195.5 (75.8, 255.3)	10253	4.0%	93 (18.0%)	1098.8 (785.1, 1298.7)
	50-59	18061	1294	844 (65.2%)	214.1 (204.3, 223.2)	8372	647 (50.0%)	129.4 (121.4, 136.7)	14269	7.3%	800 (61.8%)	178.5 (169.3, 186.4)
	60-69	14266	1756	1714 (97.6%)	83.2 (82.6, 83.8)	14196	1713 (97.6%)	82.9 (82.2, 83.5)	14196	10.0%	1713 (97.6%)	82.9 (82.2, 83.5)
	Total	**50000**	**3566**	**2654 (74.4%)**	**188.4 (185.2, 191.5)**	**23176**	**2391 (67.1%)**	**96.9 (95.2, 98.7)**	**38719**	**NA**	**2606 (73.1%)**	**148.6 (145.8, 151.2)**
Women	40-49	17488	277	17 (6.1%)	10605.6 (643.6, 14255.1)	169	4 (1.5%)	444.6 (0.0, 690.1)	5628	2.2%	17 (6.1%)	3412.8 (183.0, 4597.9)
	50-59	17986	607	116 (19.1%)	1552.2 (1254.2, 1787.3)	1087	37 (6.1%)	295.8 (179.5, 363.8)	8789	4.9%	112 (18.5%)	784.0 (631.2, 902.6)
	60-69	14526	936	699 (74.7%)	207.9 (201.1, 214.4)	8510	609 (65.0%)	139.9 (133.4, 145.0)	11316	8.2%	675 (72.1%)	167.7 (161.6, 172.8)
	Total	**50000**	**1819**	**831 (45.7%)**	**601.6 (578.8, 622.7)**	**9766**	**649 (35.7%)**	**150.5 (144.0, 156.5)**	**25732**	**NA**	**803 (44.2%)**	**320.2 (306.8, 332.3)**

Age structure of hypothetical population extrapolated from Office for National Statistics, England, United Kingdom 2017. Expected events at 10 years based on extrapolation of incidence rates from CPRD, 2014-2019. Age group and sex specific prioritisation thresholds were defined as the minimum of 10% and the level such that the expected false negative rate is controlled to be 5%. All individuals eligible for prioritisation with QRISK2 were also eligible for prioritisation with eHEART. Statin compliance assumed to equal 50%. Number needed to screen based on assuming that all individuals are formally assessed. Under each scenario, some individuals are not formally assessed and may go on to have events.

Table 3.10: Number needed to invite to prevent one event and number of events captured when prioritising with QRISK2 in a hypothetical population of 100,000 individuals in England assuming 55% of those invited for formal assessment attend.

Hypothetical population of 100,000				Full formal assessment only (QRISK2 threshold=10%)		Prioritisation tool (QRISK2 prioritisation threshold=10%) followed by full formal assessment (QRISK2 threshold=10%)			Prioritisation tool (QRISK2 prioritisation threshold corresponding to age- and sex-specific 5% false negative rates) followed by full formal assessment (QRISK2 threshold=10%)			
Sex	Age group	Expected N	Expected CVD events in 10 years	Events captured N (%)	Number needed to invite to prevent 1 CVD event	Individuals prioritised N	Events captured N (%)	Number needed to invite to prevent 1 CVD event	Individuals prioritised N	Prioritisation threshold for QRISK2	Events captured N (%)	Number needed to invite to prevent 1 CVD event
Men	40-49	17673	516	96 (18.6%)	3665.6 (2670.4, 4337.6)	608	31 (6.0%)	355.5 (137.8, 464.2)	10253	4.0%	93 (18.0%)	1997.8 (1427.5, 12361.3)
	50-59	18061	1294	844 (65.2%)	428.2 (408.6, 446.4)	8372	647 (50.0%)	235.3 (220.7, 248.5)	14269	7.3%	800 (61.8%)	324.5 (307.8, 338.9)
	60-69	14266	1756	1714 (97.6%)	166.4 (165.2, 167.6)	14196	1713 (97.6%)	150.7 (149.5, 151.8)	14196	10.0%	1713 (97.6%)	150.7 (149.5, 151.8)
	Total	**50000**	**3566**	**2654 (74.4%)**	**376.8 (370.4, 383.0)**	**23176**	**2391 (67.1%)**	**176.2 (173.1, 179.5)**	**38719**	**NA**	**2606 (73.1%)**	**270.2 (265.1, 274.9)**
Women	40-49	17488	277	17 (6.1%)	21211.2 (1278.2, 28510.2)	169	4 (1.5%)	808.4 (0.0, 1254.7)	5628	2.2%	17 (6.1%)	6205.1 (332.7, 8359.8)
	50-59	17986	607	116 (19.1%)	3104.4 (2508.4, 3574.6)	1087	37 (6.1%)	537.8 (326.4, 661.5)	8789	4.9%	112 (18.5%)	1425.5 (1147.6, 1641.1)
	60-69	14526	936	699 (74.7%)	415.8 (402.2, 428.8)	8510	609 (65.0%)	254.4 (242.5, 263.6)	11316	8.2%	675 (72.1%)	304.9 (293.8, 314.2)
	Total	**50000**	**1819**	**831 (45.7%)**	**1203.2 (1157.6, 1245.4)**	**9766**	**649 (35.7%)**	**273.6 (261.8, 284.5)**	**25732**	**NA**	**803 (44.2%)**	**582.2 (557.8, 604.2)**

Age structure of hypothetical population extrapolated from Office for National Statistics, England, United Kingdom 2017. Expected events at 10 years based on extrapolation of incidence rates from CPRD, 2014-2019. Age group and sex specific prioritisation thresholds were defined as the minimum of 10% and the level such that the expected false negative rate is controlled to be 5%. All individuals have at least one CVD risk factor (systolic blood pressure, smoking, total and/or HDL cholesterol) recorded for eHEART. Statin compliance assumed to equal 50%. Number needed to invite for full formal assessment only assumes that 50% of all individuals are formally assessed. Number needed to invite after prioritising with QRISK2 assumes 55% of prioritised individuals are formally assessed. Under each prioritisation scenario, some individuals are not formally assessed and may go on to have events.

Table 3.11: Summary of number of individuals with no primary care records available in population health modelling dataset by age group and for men and women.

Sex	Age group	Individuals N	Individuals without at least one CVD risk factor in primary care records N (%)
Men	40-49	15959	3135 (19.6%)
	50-59	19422	2631 (13.5%)
	60-69	20035	1695 (8.46%)
	Total	**55416**	**7461 (13.5%)**
Women	40-44	19943	2692 (13.5%)
	45-49	27190	2644 (9.80%)
	65-69	27225	2058 (7.56%)
	Total	**74358**	**7414 (9.97%)**

Prioritisation with eHEART requires at least one CVD risk factor of: systolic blood pressure, smoking, total and/or HDL cholesterol. Population health modelling dataset: 129,774 individuals in UK Biobank

Table 3.12: Number needed to screen to prevent one event and number of events captured when prioritising with eHEART, including formally assessing individuals without a primary care record, in a hypothetical population of 100,000 individuals in England.

Hypothetical population of 100,000				Full formal assessment only (QRISK2 threshold=10%)		Prioritisation tool (eHEART threshold=10%) followed by full formal assessment (QRISK2 threshold=10%)			Prioritisation tool (eHEART threshold corresponding to age- and sex-specific 5% false negative rates) followed by full formal assessment (QRISK2 threshold=10%)			
Sex	Age group	Expected N	Expected CVD events in 10 years	Events captured N (%)	Number needed to screen to prevent 1 CVD event	Individuals prioritised N	Events captured N (%)	Number needed to screen to prevent 1 CVD event	Individuals prioritised N	Prioritisation threshold for eHEART	Events captured N (%)	Number needed to screen to prevent 1 CVD event
Men	40-49	17673	447	82 (18.4%)	2139.0 (1611.8, 2528.2)	1565	30 (6.7%)	513.0 (226.0, 657.1)	11429	3.0%	78 (17.5%)	1450.2 (1067.9, 1723.9)
	50-59	18061	1214	789 (64.9%)	228.8 (219.0, 238.5)	6888	497 (40.9%)	138.4 (128.2, 147.4)	14737	6.2%	759 (62.5%)	194.1 (185.0, 202.6)
	60-69	14266	1684	1646 (97.6%)	86.7 (86.09, 87.3)	13419	1594 (94.5%)	84.2 (83.3, 85.1)	13419	10.0%	1594 (94.5%)	84.2 (83.3, 85.1)
	Total	**50000**	**3345**	**2518 (75.2%)**	**198.6 (195.5, 201.8)**	**21872**	**2122 (63.4%)**	**103.1 (101.1, 105.1)**	**39584**	**NA**	**2432 (72.6%)**	**162.8 (159.9, 165.8)**
Women	40-49	17488	257	15 (6.0%)	11524.4 (1349.6, 15596.6)	1159	6 (2.3%)	1986.5 (0.0, 2949.0)	6240	2.0%	14 (5.5%)	4454.6 (0.0, 6067.7)
	50-59	17986	590	111 (18.8%)	1610.7 (1326.7, 1829.0)	1246	27 (4.6%)	450.0 (231.0, 565.9)	10867	3.3%	108 (18.4%)	997.1 (815.0, 1133.2)
	60-69	14526	918	684 (74.5%)	212.5 (205.4, 218.8)	5493	432 (47.1%)	127.0 (119.7, 133.9)	11942	5.9%	663 (72.2%)	180.1 (173.4, 186.0)
	Total	**50000**	**1765**	**810 (45.9%)**	**616.8 (594.8, 638.3)**	**7898**	**466 (26.4%)**	**169.5 (158.6, 179.0)**	**29049**	**NA**	**785 (44.5%)**	**369.6 (355.7, 382.7)**

Age structure of hypothetical population extrapolated from Office for National Statistics, England, United Kingdom 2017. Expected events at 10 years based on extrapolation of incidence rates from CPRD, 2014-2019. Age group and sex specific prioritisation thresholds were defined as the minimum of 10% and the level such that the expected false negative rate is controlled to be 5%. Individuals with at least one CVD risk factor (systolic blood pressure, smoking, total and/or HDL cholesterol) recorded in primary care electronic health records were assessed using eHEART. Individuals without at least one risk factor were invited for full formal risk assessment. Number needed to screen based on assuming that all individuals are formally assessed. Under each scenario, some individuals are not formally assessed and may go on to have events. Statin compliance assumed to equal 50%.

Table 3.13: Number needed to screen to prevent one event and number of events captured when prioritising with QRISK2 in all eligible individuals, in a hypothetical population of 100,000 individuals in England.

Hypothetical population of 100,000				Full formal assessment only (QRISK2 threshold=10%)		Prioritisation tool (QRISK2 prioritisation threshold=10%) followed by full formal assessment (QRISK2 threshold=10%)			Prioritisation tool (QRISK2 prioritisation threshold corresponding to age- and sex-specific 5% false negative rates) followed by full formal assessment (QRISK2 threshold=10%)			
Sex	Age group	Expected N	Expected CVD events in 10 years	Events captured N (%)	Number needed to screen to prevent 1 CVD event	Individuals prioritised N	Events captured N (%)	Number needed to screen to prevent 1 CVD event	Individuals prioritised N	Prioritisation threshold for QRISK2	Events captured N (%)	Number needed to screen to prevent 1 CVD event
Men	40-49	17673	447	82 (18.4%)	2139.0 (1611.8, 2528.2)	1574	29 (6.4%)	538.5 (199.5, 692.5)	10673	4.0%	80 (17.8%)	1332.8 (981.8, 1584.1)
	50-59	18061	1214	789 (64.9%)	228.8 (219.0, 238.5)	8818	613 (50.5%)	143.8 (135.3, 151.4)	14443	7.4%	750 (61.7%)	192.6 (183.5, 201.1)
	60-69	14266	1684	1646 (97.6%)	86.7 (86.09, 87.3)	14198	1645 (97.6%)	86.3 (85.7, 86.9)	14198	10.0%	1645 (97.6%)	86.3 (85.7, 86.9)
	Total	**50000**	**3345**	**2518 (75.2%)**	**198.6 (195.5, 201.8)**	**24591**	**2288 (68.3%)**	**107.5 (105.7, 109.3)**	**39316**	**NA**	**2475 (73.9%)**	**158.9 (156.2, 161.6)**
Women	40-49	17488	257	15 (6.0%)	11524.4 (1349.6, 15596.6)	1245	4 (1.4%)	3553.8 (0.0, 5562.4)	6364	2.2%	15 (6.0%)	4193.8 (548.0, 5674.7)
	50-59	17986	590	111 (18.8%)	1610.7 (1326.7, 1829.0)	1875	36 (6.1%)	512.0 (310.0, 636.4)	9128	4.9%	108 (18.2%)	852.8 (697.0, 967.4)
	60-69	14526	918	684 (74.5%)	212.5 (205.4, 218.8)	8715	597 (65.0%)	146.0 (139.7, 151.1)	11425	8.2%	661 (72.0%)	172.9 (166.7, 178.0)
	Total	**50000**	**1765**	**810 (45.9%)**	**616.8 (594.8, 638.3)**	**11835**	**637 (36.1%)**	**185.7 (177.2, 192.6)**	**27007**	**NA**	**784 (44.4%)**	**344.4 (331.5, 356.8)**

Age structure of hypothetical population extrapolated from Office for National Statistics, England, United Kingdom 2017. Expected events at 10 years based on extrapolation of incidence rates from CPRD, 2014-2019. Age group and sex specific prioritisation thresholds were defined as the minimum of 10% and the level such that the expected false negative rate is controlled to be 5%. All individuals were eligible for prioritisation with QRISK2. Number needed to screen based on assuming that all individuals are formally assessed. Under each scenario, some individuals are not formally assessed and may go on to have events. Statin compliance assumed to equal 50%.

Table 3.14: Number needed to screen to prevent one event and number of events captured when prioritising with eHEART in all eligible individuals, in a hypothetical population of 100,000 individuals in England, assuming a 5% formal risk assessment threshold.

Hypothetical population of 100,000				Full formal assessment only (QRISK2 threshold=5%)		Prioritisation tool (eHEART threshold=5%) followed by full formal assessment (QRISK2 threshold=5%)			Prioritisation tool (eHEART threshold corresponding to age- and sex-specific 2.5% false negative rates) followed by full formal assessment (QRISK2 threshold=5%)			
Sex	Age group	Expected N	Expected CVD events in 10 years	Events captured N (%)	Number needed to screen to prevent 1 CVD event	Individuals prioritised N	Events captured N (%)	Number needed to screen to prevent 1 CVD event	Individuals prioritised N	Prioritisation threshold for eHEART	Events captured N (%)	Number needed to screen to prevent 1 CVD event
Men	40-49	17673	516	320 (62.0%)	551.6 (502.1, 594.3)	4735	189 (36.7%)	249.6 (209.4, 281.9)	15506	2.1%	318 (61.7%)	486.4 (444.6, 524.2)
	50-59	18061	1294	1254 (96.9%)	144.1 (142.4, 145.7)	16506	1211 (93.6%)	136.3 (133.9, 138.5)	17296	4.3%	1237 (95.6%)	139.8 (137.9, 141.6)
	60-69	14266	1756	1756 (100.0%)	81.3 (81.3, 81.3)	14259	1755 (99.9%)	81.3 (81.2, 81.3)	14259	5.0%	1755 (99.9%)	81.3 (81.2, 81.3)
	Total	**50000**	**3566**	**3330 (93.4%)**	**150.2 (148.9, 151.5)**	**35500**	**3156 (88.5%)**	**112.5 (111.4, 113.7)**	**47062**	**NA**	**3311 (92.8%)**	**142.2 (140.8, 143.5)**
Women	40-49	17488	277	61 (22.2%)	2872.3 (1950.0, 3465.7)	711	26 (9.3%)	280.4 (102.3, 364.7)	10956	1.3%	59 (21.2%)	1877.8 (1245.2, 2293.0)
	50-59	17986	607	432 (71.1%)	416.9 (395.9, 436.4)	4595	226 (37.3%)	203.3 (180.9, 221.8)	15403	2.2%	422 (69.6%)	365.0 (345.2, 382.7)
	60-69	14526	936	930 (99.4%)	156.2 (155.5, 156.8)	13410	901 (96.3%)	148.9 (147.3, 150.3)	14074	4.4%	918 (98.1%)	153.4 (152.2, 154.4)
	Total	**50000**	**1819**	**1423 (78.2%)**	**351.5 (344.7, 357.9)**	**18717**	**1152 (63.3%)**	**162.4 (158.6, 166.1)**	**40433**	**NA**	**1398 (76.9%)**	**289.2 (283.5, 294.8)**

Age structure of hypothetical population extrapolated from Office for National Statistics, England, United Kingdom 2017. Expected events at 10 years based on extrapolation of incidence rates from CPRD, 2014-2019. Age group and sex specific prioritisation thresholds were defined as the minimum of 5% and the level such that the expected false negative rate is controlled to be 2.5%. All individuals have at least one CVD risk factor (systolic blood pressure, smoking, total and/or HDL cholesterol) recorded for eHEART.
Number needed to screen based on assuming that all individuals are formally assessed. Under each scenario, some individuals are not formally assessed and may go on to have events. Statin compliance assumed to be equal to 50%.

Table 3.15: Number to needed screen to prevent one event, and number of events captured when prioritising with QRISK2 in a hypothetical population of 100,000 individuals in England, assuming a 5% formal risk assessment threshold.

Hypothetical population of 100,000				Full formal assessment only (QRISK2 threshold=5%)		Prioritisation tool (QRISK2 prioritisation threshold=5%) followed by full formal assessment (QRISK2 threshold=5%)			Prioritisation tool (QRISK2 prioritisation threshold corresponding to age- and sex-specific 2.5% false negative rates) followed by full formal assessment (QRISK2 threshold=5%)			
Sex	Age group	Expected N	Expected CVD events in 10 years	Events captured N (%)	Number needed to screen to prevent 1 CVD event	Individuals prioritised N	Events captured N (%)	Number needed to screen to prevent 1 CVD event	Individuals prioritised N	Prioritisation threshold for QRISK2	Events captured N (%)	Number needed to screen to prevent 1 CVD event
Men	40-49	17673	516	320 (62.0%)	551.6 (502.1, 594.3)	6302	247 (47.9%)	254.9 (220.8, 282.0)	14500	3.0%	318 (61.7%)	454.8 (413.7, 490.8)
	50-59	18061	1294	1254 (96.9%)	144.1 (142.4, 145.7)	17877	1251 (96.7%)	142.9 (141.1, 144.5)	17877	5.0%	1251 (96.7%)	142.9 (141.1, 144.5)
	60-69	14266	1756	1756 (100.0%)	81.3 (81.3, 81.3)	14266	1756 (100.0%)	81.3 (81.3, 81.3)	14266	5.0%	1756 (100.0%)	81.3 (81.3, 81.3)
	Total	**50000**	**3566**	**3330 (93.4%)**	**150.2 (148.9, 151.5)**	**38444**	**3254 (91.3%)**	**118.1 (116.9, 119.3)**	**46643**	**NA**	**3326 (93.3%)**	**140.3 (139.0, 141.6)**
Women	40-49	17488	277	61 (22.2%)	2872.3 (1950.0, 3465.7)	957	28 (10.2%)	342.9 (157.8, 441.5)	9253	1.7%	59 (21.2%)	1585.77 (1043.7, 1937.6)
	50-59	17986	607	432 (71.1%)	416.9 (395.9, 436.4)	8307	348 (57.3%)	239.0 (221.9, 253.7)	14523	3.2%	428 (70.5%)	339.6 (321.3, 355.3)
	60-69	14526	936	930 (99.4%)	156.2 (155.5, 156.8)	14496	930 (99.4%)	155.8 (155.2, 156.4)	14496	5.0%	930 (99.4%)	155.8 (155.2, 156.4)
	Total	**50000**	**1819**	**1423 (78.2%)**	**351.5 (344.7, 357.9)**	**23761**	**1306 (71.8%)**	**182.0 (178.5, 185.4)**	**38272**	**NA**	**1416 (77.9%)**	**270.2 (264.9, 275.1)**

Age structure of hypothetical population extrapolated from Office for National Statistics, England, United Kingdom 2017. Expected events at 10 years based on extrapolation of incidence rates from CPRD, 2014-2019. Age group and sex specific prioritisation thresholds were defined as the minimum of 10% and the level such that the expected false negative rate is controlled to be 5%. All individuals eligible for prioritisation with QRISK2 were also eligible for prioritisation with eHEART. Statin compliance assumed to equal 50%. Number needed to screen based on assuming that all individuals are formally assessed. Under each scenario, some individuals are not formally assessed and may go on to have events.

3.4 Discussion

We have provided quantitative evidence for systematically prioritising individuals for invitation for full formal CVD risk assessment using existing longitudinal primary care records, by evaluating a novel prioritisation tool (eHEART) with the use of age- and sex-specific prioritisation thresholds. Compared to conducting formal assessments on all individuals, the prioritisation tool has the potential to reduce the number needed to screen to prevent one CVD event by approximately 50% for women and 20% for men whilst identifying 96-98% of high-risk individuals. We found no added value of using the repeat measures in the eHEART prioritisation tool in comparison to estimates using single measures in QRISK2 as a prioritisation tool. However, the use of a fixed 10% prioritisation threshold substantially reduces the number of high-risk individuals identified by 10-20%. Our study highlights the importance of using prioritisation thresholds well below formal risk assessment thresholds, and lower than the 10% prioritisation threshold currently recommended by the NICE guidelines in the UK.[7] We demonstrated the advantage of using a pragmatic strategy of selecting age- and sex-specific prioritisation thresholds corresponding to 5% false negative rates, to ensure a balance between efficiency and specificity across age and sex.

This is the first study, to our knowledge, to demonstrate the benefits of age- and sex-specific thresholds for prioritising individuals for full formal CVD risk assessment, although they have been shown to improve specificity when applied directly to the formal CVD risk assessment.[41,42] In our study we specified the age- and sex- specific thresholds to minimise the false negative rate to 5%. However, alternative thresholds that optimise for a different criterion, such as a 1% or 10% false negative rate, could be chosen to achieve varying specificity and efficiency gains. Whilst using age- and sex-specific thresholds have its benefits compared to a fixed threshold, the lower thresholds chosen may be a reflection of the uncertainty in the risks estimated, due to the low number of observed events in young men and women within the dataset used. Furthermore, age- and sex- specific thresholds may reduce equality in healthcare provision, especially in individuals who do not identify as cisgender.

We have illustrated the use of prioritisation tools within the current United Kingdom population-wide preventative programme, in which all individuals aged between 40 and 74 years are invited into their general practitioners for a National Health Service (NHS) health check to assess an individual's risk of CVD, diabetes, kidney disease and stroke and dementia

every five years.[43] A systematic review reported coverage (percentage eligible for health checks who received one) of 45.6%[33] and invitation uptake (percentage invited for health checks who received one) of 48.2% between 2009-2016 and suggested improved coverage and invitation uptake in populations targeted for prioritisation through general practices, including those with low socioeconomic status or ethnic minorities. As such, a complementary and automated systematic primary care records based prioritisation approach could be supportive to improve both the coverage and uptake of the programme over the next few years, helping to reduce health inequalities.[44]

In our study, we have chosen to focus on improvements in efficiency, whereby individuals not deemed at high risk during the prioritisation stage were deferred from the formal risk assessment stage. An alternative strategy is to rank individuals using existing information, and to continue inviting individuals for a formal risk assessment within the next 5 years. With the former strategy, the implementation will improve specificity at the cost of sensitivity. This may cause more harm to the patient, with some high-risk individuals being deferred. However, this approach would likely reduce the cost and resources used within the NHS. Conversely, the latter will improve sensitivity at the cost of specificity, benefit patients but will provide fewer efficiency benefits for the NHS.

This is also the first study to provide an independent evaluation of the use of the existing QRISK2 as a prioritisation tool compared with our novel eHEART tool. The derivation, validation and implementation of eHEART accounts for and leverages the sparse and sporadically observed longitudinal primary care records, by estimating current predictor values based on all past primary care records to make 10-year CVD predictions. By implementing a landmark model approach in deriving eHEART, the tool removes the need for complete risk predictor measurements in all individuals. Furthermore, eHEART uses only a small number of conventional risk predictors, thus making it transportable across different primary care electronic systems and country settings. In contrast, whilst the published derivation and validation methods for QRISK2 have used multiple imputation to handle non-recorded risk factors, its' clinical implementation is based on replacing missing non-recorded values with age, sex and ethnicity-specific population average values. However, QRISK2 contains many more risk factors than eHEART and is already integrated into the primary care electronic systems in England and used to prioritise patient invitation to formal CVD risk assessment. We conducted population health modelling utilising both primary care records and baseline measurements in UKB, and used recalibrated risk estimates and observed incidence rates for

estimating the population-health impact in England. Our results show only small differences between the tools, and that future focus should be on the selection and use of appropriate prioritisation thresholds.

Our study has potential limitations. First, thresholds were selected by age-group and sex; alternative thresholds such as ethnic-specific thresholds could also be considered. Second, optimism may exist due to possible overlap of a small proportion of individuals in UKB and in the databases used in deriving eHEART and QRISK2. Third, our population health modelling may be affected by healthy volunteer bias in UKB; however, this is largely overcome by the recalibration and rescaling with more representative risk factor levels and incidence rates. Fourth, our population health modelling was restricted to individuals aged between 40-69 years due to data availability; we expect even larger efficiency gains if such a prioritisation tool was used in individuals <40 years but not >69 years. Fifth, we assumed a fixed level of statin compliance and uptake for all individuals; whilst these are likely to vary by age and sex,[9,45,46] the levels of compliance and uptake are likely to change after prioritisation with age- and sex-specific thresholds is introduced. Sixth, our population health modelling focused on prioritising individuals for a formal CVD assessment at a single point in time, rather than every 5 or 10 years; however such prioritisation tools can be used in real-time with updated primary care records. This limitation will be addressed in **Chapter 5** where a dynamic population healhth modelling approach will be used. Finally, our population health modelling was restricted to the population of England assuming use of the QRISK2 formal CVD risk assessment tool; future work should investigate the generalisability of the tools across different countries and in combination with other formal risk assessment tools.

3.5 Conclusion

The use of a prioritisation tool, based on either utilising repeated measures or last observed values of primary care records, in conjunction with age- and sex-specific thresholds has the potential to systematically identify high-risk individuals for invitation for full formal CVD risk assessment. Our results suggest using age- and sex-specific thresholds can reduce the number needed to screen to prevent one CVD event whilst identifying the majority of high-risk individuals. Such a strategy could be particularly important as nations tackle the many COVID-19 related burdens and backlogs in primary care settings. Whilst minimal differences were observed between using a single or repeated measures of primary care records, **Chapter 4** will consider the use of polygenic risk scores in a prioritisation tool.

References

1. WHO/Europe | Cardiovascular diseases - Data and statistics. Accessed April 1, 2022. https://www.euro.who.int/en/health-topics/noncommunicable-diseases/cardiovascular-diseases/data-and-statistics
2. National Vascular Disease Prevention Alliance. Guidelines for the management of absolute CVD risk 2012. Published 2012. Accessed November 23, 2021. https://informme.org.au/en/Guidelines/Guidelines-for-the-assessment-and-management-of-absolute-CVD-risk
3. Visseren FLJ, MacH F, Smulders YM, et al. 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice. *Eur Heart J*. 2021;42(34):3227-3337. doi:10.1093/EURHEARTJ/EHAB484
4. New Zealand Ministry of Health. Cardiovascular Disease Risk Assessment and Management for Primary Care 2018.
5. Arnett DK, Blumenthal RS, Albert MA, et al. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation*. 2019;140(11):e596-e646. doi:10.1161/CIR.0000000000000678
6. Pearson GJ, Thanassoulis G, Anderson TJ, et al. 2021 Canadian Cardiovascular Society Guidelines for the Management of Dyslipidemia for the Prevention of Cardiovascular Disease in Adults. *Can J Cardiol*. 2021;37(8):1129-1150. doi:10.1016/J.CJCA.2021.03.016
7. National Institute for Health and Care Excellence (NICE). Lipid modification: Cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease (clinical guideline CG181). Published online 2014.
8. Stewart J, Addy K, Campbell S, Wilkinson P. Primary prevention of cardiovascular disease: Updated review of contemporary guidance and literature. *JRSM Cardiovasc Dis*. 2020;9:204800402094932. doi:10.1177/2048004020949326
9. Chamnan P, Simmons RK, Khaw KT, Wareham NJ, Griffin SJ. Estimating the population impact of screening strategies for identifying and treating people at high risk of cardiovascular disease: modelling study. *BMJ*. 2010;340(7754):1016. doi:10.1136/BMJ.C1693
10. Selvarajah S, Haniff J, Kaur G, et al. Identification of effective screening strategies for

cardiovascular disease prevention in a developing country: using cardiovascular risk-estimation and risk-reduction tools for policy recommendations. *BMC Cardiovasc Disord*. 2013;13(1):1-10. doi:10.1186/1471-2261-13-10/TABLES/5

11. Kypridemos C, Collins B, McHale P, et al. Future cost-effectiveness and equity of the NHS Health Check cardiovascular disease prevention programme: Microsimulation modelling using data from Liverpool, UK. Sheikh A, ed. *PLOS Med*. 2018;15(5):e1002573. doi:10.1371/journal.pmed.1002573
12. GOV.UK. Launch of the Clinical Practice Research Datalink. Accessed April 20, 2023. https://www.gov.uk/government/news/launch-of-the-clinical-practice-research-datalink
13. CPRD. Data | CPRD. Accessed April 20, 2023. https://cprd.com/data
14. Ghosh RE, Crellin E, Beatty S, Donegan K, Myles P, Williams R. How Clinical Practice Research Datalink data are used to support pharmacovigilance. *Ther Adv drug Saf*. 2019;10:204209861985401. doi:10.1177/2042098619854010
15. Herrett E, Gallagher AM, Bhaskaran K, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol*. 2015;44(3):827-836. doi:10.1093/ije/dyv098
16. Wolf A, Dedman D, Campbell J, et al. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int J Epidemiol*. 2019;48(6):1740-1740g. doi:10.1093/IJE/DYZ034
17. Kontopantelis E, Stevens RJ, Helms PJ, Edwards D, Doran T, Ashcroft DM. Spatial distribution of clinical computer systems in primary care in England in 2016 and implications for primary care electronic medical record databases: a cross-sectional population study. *BMJ Open*. 2018;8:20738. doi:10.1136/bmjopen-2017-020738
18. Ltd. IPS. The Health Improvement Network (THIN) | Vision. Accessed November 23, 2021. https://www.visionhealth.co.uk/portfolio-items/the-health-improvement-network-thin/
19. Tate AR, Dungey S, Glew S, Beloff N, Williams R, Williams T. Quality of recording of diabetes in the UK: how does the GP's method of coding clinical data affect incidence estimates? Cross-sectional study using the CPRD database. *BMJ Open*. 2017;7(1). doi:10.1136/BMJOPEN-2016-012905
20. NHS Digital. Quality and Outcomes Framework, 2020-21. Published 2021. Accessed November 24, 2021. https://digital.nhs.uk/data-and-information/publications/statistical/quality-and-outcomes-framework-achievement-prevalence-and-exceptions-data/2020-21
21. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk

in the general population: Systematic review. *BMJ*. 2016;353. doi:10.1136/bmj.i2416

22. Welch C, Petersen I, Walters K, et al. Two-stage method to remove population- and individual-level outliers from longitudinal data in a primary care database. *Pharmacoepidemiol Drug Saf*. 2012;21(7):725-732. doi:10.1002/PDS.2270
23. Paige E, Barrett J, Stevens D, et al. Landmark Models for Optimizing the Use of Repeated Measurements of Risk Factors in Electronic Health Records to Predict Future Disease Risk. *Am J Epidemiol*. 2018;187(7):1530-1538. doi:10.1093/AJE/KWY018
24. Laird NM, Ware JH. Random-Effects Models for Longitudinal Data. *Biometrics*. 1982;38(4):963. doi:10.2307/2529876
25. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128-138. doi:10.1097/EDE.0B013E3181C30FB2
26. Lloyd-Jones DM. Cardiovascular risk prediction: Basic concepts, current status, and future directions. *Circulation*. 2010;121(15):1768-1777. doi:10.1161/CIRCULATIONAHA.109.849166
27. Kaptoge S, Di Angelantonio E, Pennells L, et al. C-reactive protein, fibrinogen, and cardiovascular disease prediction. *N Engl J Med*. 2012;367(14):1310-1320. doi:10.1056/NEJMoa1107477
28. Pennells L, Kaptoge S, Wood A, et al. Equalization of four cardiovascular risk algorithms after systematic recalibration: Individual-participant meta-analysis of 86 prospective studies. *Eur Heart J*. 2019;40(7):621-631. doi:10.1093/eurheartj/ehy653
29. Pate A, Emsley R, Ashcroft DM, Brown B, Van Staa T. The uncertainty with using risk prediction models for individual decision making: An exemplar cohort study examining the prediction of cardiovascular disease in English primary care. *BMC Med*. 2019;17(1):134. doi:10.1186/s12916-019-1368-8
30. Office for National Statistics. Population estimates. Accessed November 26, 2021. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates#datasets
31. Collins R, Reith C, Emberson J, et al. Interpretation of the evidence for the efficacy and safety of statin therapy. *Lancet*. 2016;388(10059):2532-2561. doi:10.1016/S0140-6736(16)31357-5/ATTACHMENT/9C060D9A-F690-455D-942C-F1AA57FD59DC/MMC1.PDF
32. Patel R, Barnard S, Thompson K, et al. Evaluation of the uptake and delivery of the NHS Health Check programme in England, using primary care data from 9.5 million people: A cross-sectional study. *BMJ Open*. 2020;10(11):42963. doi:10.1136/bmjopen-2020-

042963

33. Martin A, Saunders CL, Harte E, et al. Delivery and impact of the NHS Health Check in the first 8 years: A systematic review. *Br J Gen Pract*. 2018;68(672):e449-e459. doi:10.3399/bjgp18X697649

34. Brown R, Lewsey J, Wild S, Logue J, Welsh P. Associations of statin adherence and lipid targets with adverse outcomes in myocardial infarction survivors: a retrospective cohort study. *BMJ Open*. 2021;11(9):e054893. doi:10.1136/BMJOPEN-2021-054893

35. Usher-Smith JA, Harvey-Kelly LLW, Rossi SH, Harrison H, Griffin SJ, Stewart GD. Acceptability and potential impact on uptake of using different risk stratification approaches to determine eligibility for screening: A population-based survey. *Heal Expect*. 2021;24(2):341-351. doi:10.1111/HEX.13175

36. Sallis A, Sherlock J, Bonus A, et al. Pre-notification and reminder SMS text messages with behaviourally informed invitation letters to improve uptake of NHS Health Checks: a factorial randomised controlled trial. *BMC Public Health*. 2019;19(1). doi:10.1186/S12889-019-7476-8

37. Gidlow CJ, Ellis NJ, Riley V, et al. Randomised controlled trial comparing uptake of NHS Health Check in response to standard letters, risk-personalised letters and telephone invitations. *BMC Public Health*. 2019;19(1):1-11. doi:10.1186/S12889-019-6540-8/FIGURES/2

38. 3.6.1. RDCT. A Language and Environment for Statistical Computing. *R Found Stat Comput*. 2018;2:https://www.R-project.org. http://www.r-project.org

39. Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLOS Med*. 2015;12(10):e1001885. doi:10.1371/JOURNAL.PMED.1001885

40. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. *Circulation*. 2015;131(2):211. doi:10.1161/CIRCULATIONAHA.114.014508

41. Yebyo HG, Zappacosta S, Aschmann HE, Haile SR, Puhan MA. Global variation of risk thresholds for initiating statins for primary prevention of cardiovascular disease: A benefit-harm balance modelling study. *BMC Cardiovasc Disord*. 2020;20(1):1-10. doi:10.1186/S12872-020-01697-6/FIGURES/3

42. Navar-Boggan AM, Peterson ED, D'Agostino RB, Pencina MJ, Sniderman AD. Using age- and sex-specific risk thresholds to guide statin therapy: One size may not fit all. *J Am Coll Cardiol*. 2015;65(16):1633-1639. doi:10.1016/j.jacc.2015.02.025

43. Public Health England. Guidance overview: NHS Health Checks: applying All Our Health - GOV.UK. Accessed November 23, 2021. https://www.gov.uk/government/publications/nhs-health-checks-applying-all-our-health
44. NHS Digital. Appointments in General Practice August 2021. Accessed November 24, 2021. https://digital.nhs.uk/data-and-information/publications/statistical/appointments-in-general-practice/august-2021
45. Colantonio LD, Rosenson RS, Deng L, et al. Adherence to statin therapy among US adults between 2007 and 2014. *J Am Heart Assoc*. 2019;8(1). doi:10.1161/JAHA.118.010376
46. Hope HF, Binkley GM, Fenton S, Kitas GD, Verstappen SMM, Symmons DPM. Systematic review of the predictors of statin adherence for the primary prevention of cardiovascular disease. *PLoS One*. 2019;14(1):e0201196. doi:10.1371/JOURNAL.PONE.0201196

Chapter 4

Supplementing primary care records with polygenic risk scores for prioritising individuals at greatest need of a CVD risk assessment

Chapter summary

Objective: To provide quantitative evidence of the use of polygenic risk scores (PRS) for systematically identifying individuals for invitation for full formal cardiovascular disease (CVD) risk assessment.

Methods: 108,685 participants aged 40-69, with measured biomarkers, linked primary care records and genetic data in UK Biobank were used for model derivation and population health modelling. Prioritisation tools using age, PRS for coronary artery disease and stroke, and conventional risk factors for CVD available within longitudinal primary care records were derived using sex-specific Cox models. Rescaling to account for the healthy cohort effect, we modelled the implications of initiating guideline-recommended statin therapy after prioritising individuals for invitation to a formal CVD risk assessment.

Results: 1,838 CVD events were observed over median follow up of 8.2 years. If primary care records were used to prioritise individuals for formal risk assessment using age- and sex-specific thresholds corresponding to 5% false negative rates then we would capture 65% and 43% events amongst men and women respectively. The numbers of men and women needed to be screened to prevent one CVD event (NNS) are 149 and 280 respectively. In contrast**,** adding PRS to both prioritisation and formal assessments, and selecting thresholds to capture the same number of events, resulted in a NNS of 116 for men and 180 for women.

Conclusion: The use of PRS together with primary care records to prioritise individuals at highest risk of a CVD event for a formal CVD risk assessment can more efficiently prioritise those who need interventions the most than using primary care records alone. This could lead to better allocation of resources by reducing the number of formal risk assessments in primary care while still preventing the same number CVD events. However, further work regarding the future collection and practical implementation of PRS will need to be conducted.

4.1 Introduction

Cardiovascular disease (CVD) remains a major cause of morbidity and mortality worldwide.[1] Identifying individuals at a high risk of CVD in order to manage and implement interventions to reduce risk of CVD remains an important aim.[2,3] Prediction tools utilising the risk factor levels of individuals to estimate a 5 or 10 year risk of CVD have been developed to aid clinical decision making and are recommended by healthcare guidelines across the world.[3–10] However, recent studies have debated the clinical value and cost effectiveness of national risk assessment programmes.[11–17] In line with this, recent guidelines have made recommendations to better utilise existing primary care records to improve the stratification of high-risk individuals prior to formal CVD risk assessments.[18] However, few strategies or tools to systematically identify such individuals have been recommended. Proposals have also been recommended to prioritise individuals using CVD-based polygenic risk scores (PRS); such PRS have been shown to be independent of other CVD risk factors, offering improved stratification with high concordance between categories of polygenic risk and future CVD risk across the life course, and to improve discriminatory performance when used to supplement existing CVD risk scores.[19–21] However, no studies have quantified the impact PRS would have for prioritisation.

Therefore, to investigate the benefits of PRS to systematically prioritise individuals at high risk of CVD, we compare systematically prioritising individuals using a PRS based prioritisation tool against current guidelines recommendations of using a prioritisation tool based on longitudinal primary care records.[22–26]

4.2 Methods

4.2.1 Data sources

4.2.1.1 UK Biobank

UKB is a prospective cohort study with detailed baseline information, genetic data and linked primary care record data available for 177,359 individuals in England recruited between 2006 and 2010.[27] Genetic data was sequenced using a genome wide array of approximately 826,000 markers with imputation to approximately 96 million markers.[27] Primary care data was provided from the The Phoenix Partnership, Egton Medical Information Systems and Vision GP system suppliers.[28] Data were linked with secondary care admissions from Hospital Episode Statistics (HES) and mortality records from the Office for National Statistics (ONS). For this study, primary care records were restricted to those measured between the 1st April 2004, the introduction of the Quality and Outcomes Framework (QOF) and UKB baseline survey. To assess the impact of PRS as a prioritisation tool and compare with primary care records, our primary analyses were restricted to individuals with complete genetic data necessary for calculating the PRS, at least one primary care record and without prior CVD or statin initiation before UKB baseline. Individuals contributing to the PRS derivation were also excluded. Data from UKB were used to derive CVD risk tools and to model the implications of prioritising individuals for formal assessment **(Figure 4.1)**.

4.2.1.2 CPRD

The Clinical Practice Research Datalink (CPRD) is a large UK primary care database containing primary care records with linked information from HES and mortality records from the ONS.[28] The most recent 5 years available primary care records were extracted for 870,486 individuals who were still alive and without prior CVD on the 1st January 2014 and had no statins throughout follow-up until 31st May 2019, the end of data availability **(Figure 4.2).** Data from CPRD were used to rescale estimated CVD risks in UKB participants to address the healthy cohort effect **(Figure 4.1).**

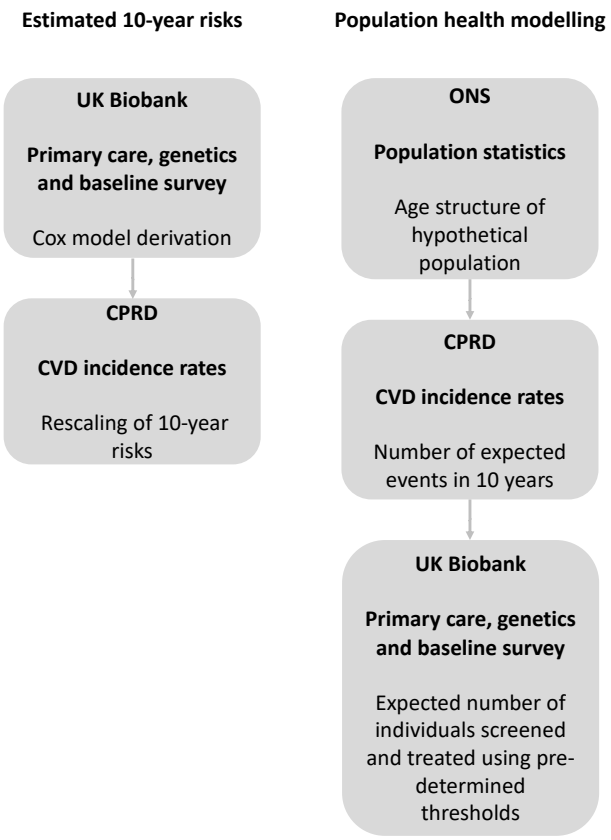


Figure 4.1: Flowchart showing data sources used for model derivation for estimated risks and population health modelling.

Abbreviations: CPRD, Clinical Practice Research Datalink; ONS, Office for National Statistics.

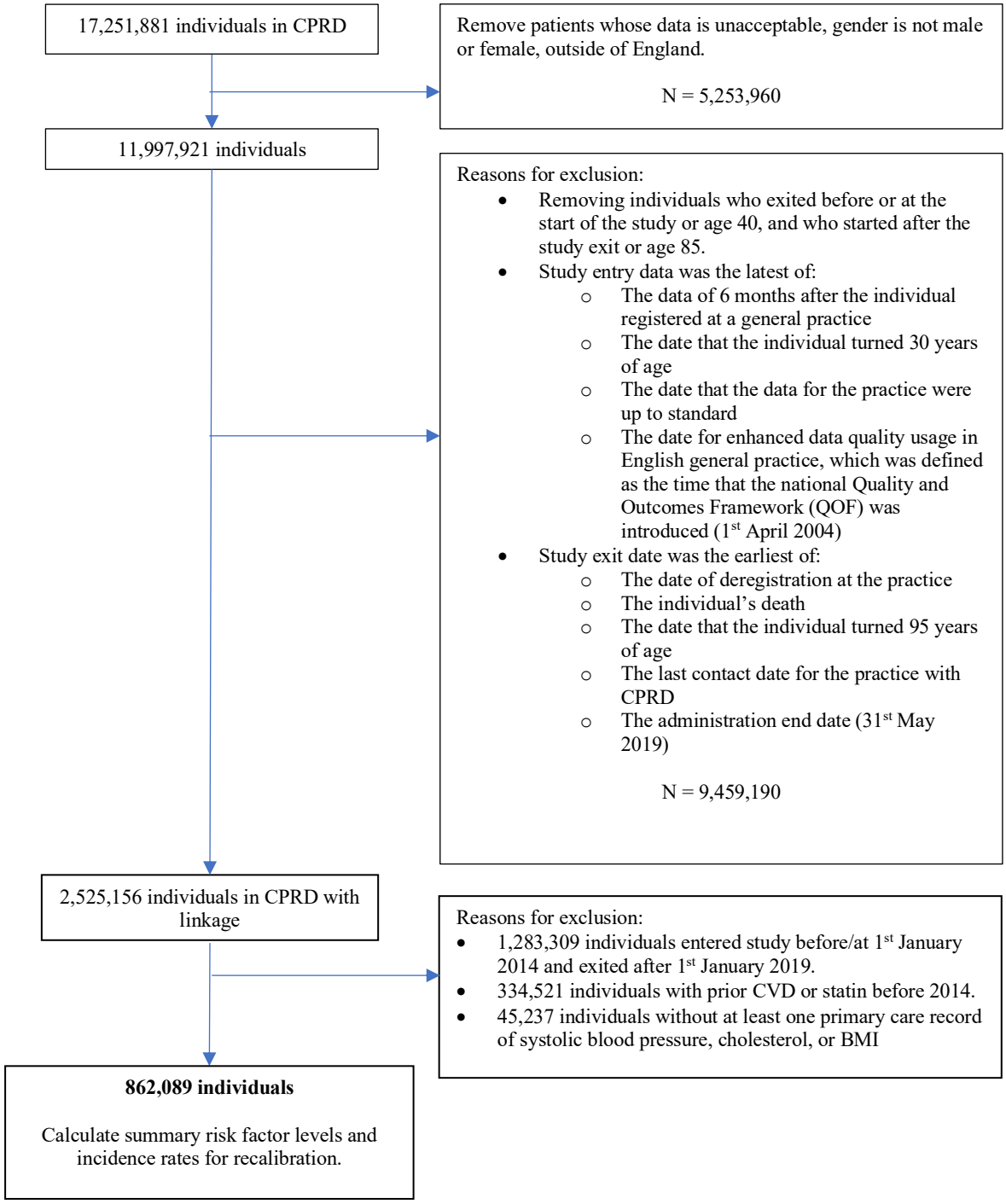


Figure 4.2: Flowchart showing selection of patient records for generating summary statistics from CPRD.

Abbreviations: BMI, body mass index; CVD, cardiovascular disease.

Highlighted in red were individuals without necessary primary care records to calculate incidence rates for sensitivity analyses and were included for sensitivity analysis incidence rates calculation.

4.2.2 Outcomes and risk factors

CVD was defined as the first ever incident of fatal or non-fatal events of coronary heart disease (including angina and myocardial infarction), ischaemic heart disease and stroke appearing in the linked HES and ONS databases during follow up **(Table 4.1)**.

Two PRS for coronary artery disease (CAD) and stroke, constructed using a meta-score approach and external summary statistics from large genome wide association studies,[20,29] were used as independent variables. Conventional risk factors (as those in the QRISK2 scores[4]) were selected: age, sex, ethnicity, Townsend score, smoking status (current/ever smoker), history of diabetes (type 1 or type 2 or history of diabetes medication), family history of CVD, history of chronic kidney disease (stages 4 and 5), history of atrial fibrillation status, history of blood pressure treatment, history of rheumatoid arthritis, total and high density lipoprotein (HDL) cholesterol, systolic blood pressure (SBP), body mass index (BMI), and age interactions with Townsend score, history of diabetes, family history of CVD, history of atrial fibrillation, history of blood pressure treatment, SBP and BMI.

Table 4.1: Code list used to define cardiovascular disease.

Endpoint	ICD-10 code
Angina pectoris	I20
Myocardial infarction	I21, I22, I23
Coronary disease non- myocardial infarction	I24, I25
Ischemic stroke	I63
Unclassified stroke	I64

HES data available covered hospital admissions. Death registries provided data on deaths, with both primary and contributory causes of death coded in ICD-10

4.2.3 Statistical modelling

4.2.3.1 Prioritisation tool model derivation

Sex-specific Cox models were used to derive three different prioritisation tools for estimating 10-year CVD prioritisation risk using primary care and genetic data from UKB **(Figure 4.3)**. First, we derived a prioritisation tool with linear predictors of baseline age, CAD PRS[20] and stroke PRS[29]. Age interactions were considered but were not statistically significant at the 5% level. Second, we derived a prioritisation tool with predictors utilising longitudinal primary care records. Third, we derived a prioritisation tool with both PRS and longitudinal primary care records.

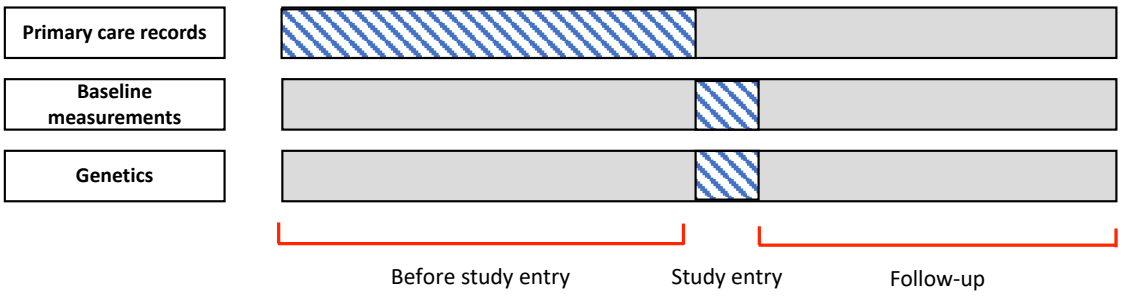


Figure 4.3: Illustration of UK Biobank data used in analysis

The UK Biobank baseline data was used to derive and calculate the formal CVD risk assessment. The genetics data was used to derive the polygenic risk scores and were taken at the same time (First UK Biobank survey). The retrospective primary care data were taken at different time points from 1st April 2004 until baseline survey.

4.2.3.2 Summarising repeated measures in longitudinal primary care records

For the second and third prioritisation tools, we used the historical repeated measures in an individual's primary care record to estimate expected risk factor values at the time of a formal risk assessment. Both tools were derived in two stages.

Stage 1:Multivariate mixed-effects model

In the first stage, we used sex-specific multivariate mixed effects regression models on longitudinal risk factor measurements for SBP, total and HDL cholesterol and BMI to estimate current risk factor values. The model was chosen as our aim was utilise all available longitudinal

data in primary care records. Consequently, the model needed to handle the repeated and sporadic structure of the data.

Using the primary care records before baseline survey in a population without prior CVD or diabetes, but including those with prior statin usage, sex-specific multivariate mixed effects models with a fixed slope, random intercept and an intra-correlation structure were used to estimate the risk factor levels at the same timepoint as when the individual attended the UK Biobank baseline assessment. Let SBP_{ij}, $Total\ cholesterol_{ij}$, $HDL\ cholesterol_{ij}$, BMI_{ij}, age_{ij}, age_{ij}^2, AHM_{ij} denote, respectively, the repeat measures of systolic blood pressure, total cholesterol, HDL cholesterol, BMI, age at visit in years, age at visit in years squared, an indicator for history of anti-hypertensive medication, and an indicator for history of statin medication individual i and measurement j. The sex-specific multivariate mixed models and its corresponding correlated covariance structure were of the following form:

$$SBP_{ij} = a_1 + b_1 age_{ij} + c_1 age_{ij}^2 + (d * AHM_{ij}) + u_{1i} + e_{1ij}$$
$$Total\ cholesterol_{ij} = a_2 + b_2 age_{ij} + c_2 age_{ij}^2 + (e * statin_{ij}) + u_{2i} + e_{2ij}$$
$$HDL\ cholesterol_{ij} = a_3 + b_3 age_{ij} + c_3 age_{ij}^2 + u_{3i} + e_{3ij}$$
$$BMI_{ij} = a_4 + b_4 age_{ij} + c_4 age_{ij}^2 + u_{4i} + e_{4ij}$$

Where $\begin{bmatrix} u_{1i} \\ u_{2i} \\ u_{3i} \\ u_{4i} \end{bmatrix} \sim multivariate\ normal \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{bmatrix} \right)$

And $\begin{bmatrix} e_{1ij} \\ e_{2ij} \\ e_{3ij} \\ e_{4ij} \end{bmatrix} \sim multivariate\ normal \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{e1}^2 & 0 & 0 & 0 \\ 0 & \sigma_{e2}^2 & 0 & 0 \\ 0 & 0 & \sigma_{e3}^2 & 0 \\ 0 & 0 & 0 & \sigma_{e4}^2 \end{bmatrix} \right)$

For $i = 1 \ldots N, j = 1 \ldots M_i, u_i \sim N(0, \sigma_u^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$ where N is defined as the number of individuals included in the model, M_i is defined as the total number of longitudinal primary care measurements observed for individual i, u_{1i} to u_{4i} represents the random intercepts which are correlated between risk factors, and e_{1ij} to e_{4ij} represents the uncorrelated residual errors for each risk factor.

A mixed effects model was chosen to take into account the sporadic nature of electronic health records, as well as being able to model the intra-correlations between each risk factor. In addition, the model only needs a minimum of one recorded measurement of any one risk factor to estimate all four of the risk factors. The model assumes that all risk factors jointly follow a multivariate normal distribution. Inference based from the multivariate normal distribution may often be reasonable even if the multivariate normality does not hold, especially in the context of imputation of missing data[30] and regression calibration[31,32].

Stage 2: Cox model derivation

In the second stage, we derived sex-specific Cox models with the estimated current risk factor values for SBP, total and HDL cholesterol and BMI from stage 1. For both the second and third prioritisation tool, the most recent primary care measurements for the remaining QRISK2 risk factors were also included in the Cox model. For the third prioritisation tool, in addition to the previously mentioned risk factors, linear predictors for the CAD PRS and stroke PRS were included in the Cox model.

4.2.3.3 Formal risk assessment model derivation

Sex-specific Cox models were used to derive two formal risk assessment models for predicting 10-year formal assessment CVD risk using risk factor measurements recorded at UKB baseline survey. First, we re-derived a model based on QRISK2 predictors and second, we derived a model based on QRISK2 predictors enhanced with the CAD PRS and stroke PRS.

4.2.3.4 Model performance in derivation cohort

All models were validated using 10-fold cross validation and prognostic ability was quantified using Harrell's C-index to measure discrimination and the net reclassification improvement (NRI).

4.2.4 Rescaling of estimated risks from prioritisation and formal risk assessment tools

As UKB consists of healthier individuals than the general primary care population in England, deriving and modelling the health impact of all prioritisation tools and formal risk tools in UK Biobank without adjustments would lead to biases. In particular, the distribution of the 10-year risks estimated, would be skewed to the right and be narrow relative to the distribution observed in the general population. Since a fixed 10% threshold was used to guide statin initiation, the biased distributions would lead to misleading population health modelling estimates. To more

accurately use UK Biobank for population health modelling, the distribution of 10-year risks estimated were rescaled.

Rescaling was completed for each tool and by sex, using methods similar to those previously described[33,34], and allowed the mean level of predicted risks based on UKB data to match what was observed in CPRD. We used sex-specific mean risk factor levels calculated from the Clinical Practice Research Datalink (CPRD) between the years 2014 and 2019 within 5-year age groups to estimate the predicted risk in the general population by fitting the average level risk factors into the published QRISK2 risk model **(Table 4.2).** This allows us to calculate scaling factors to rescale each prioritisation tool and formal risk assessment model to have a distribution similar to what would be expected in the general population.

A linear model was fit within each tool and by sex to relate the observed risk (θ_{obs}) and predicted risk (θ_{pred}) estimated for each 5-year age group (c_s):

$$log_e\left(-log_e(1-\theta_{obs,c_s})\right) = \beta_0 + \beta_1 \times log_e\left(-log_e(1-\theta_{pred,c_s})\right)$$

The estimated β_0 and β_1 were then used as scaling factors to rescale each individual's original 10-year risk ($\theta_{pred,i}$) to give a new rescaled estimate $\theta_{newpred,i}$:

$$\theta_{newpred,i} = 1 - \exp\left(-\exp\left(\beta_0 + \beta_1 \times log_e\left(-log_e(1-\theta_{pred,i})\right)\right)\right)$$

Table 4.2: Age- and sex-specific mean risk factor levels using records from 870,486 individuals in the CPRD database.

	Men						Women					
Age group	**40-44**	**45-49**	**50-54**	**55-59**	**60-64**	**65-69**	**40-44**	**45-49**	**50-54**	**55-59**	**60-64**	**65-69**
Ethnicity — White, (%)	89.4	91.8	93.6	94.7	96	97.3	88.7	90.8	92.9	94.1	95.6	96.9
Townsend, mean*	-1.5	-1.5	-1.5	-1.5	-1.5	-1.5	-1.5	-1.5	-1.5	-1.5	-1.5	-1.5
Systolic blood pressure — mmHg, mean	129.5	131.2	132.8	134.4	135.8	136.5	121.0	124.3	127.5	130.0	132.6	134.8
Total cholesterol — mmol/litre, mean	5.18	5.26	5.31	5.27	5.22	5.13	4.85	5.07	5.36	5.61	5.70	5.70
HDL cholesterol — mmol/litre, mean	1.27	1.29	1.32	1.34	1.38	1.4	1.53	1.57	1.64	1.68	1.69	1.71
BMI – kg/m^2, mean	28.2	28.6	28.5	28.3	27.9	27.4	27.8	28.2	28.2	28.1	27.7	27.2
Current/ever smoker, (%)	49.2	47.9	45.4	40.4	33.8	27.1	42.0	43.1	43.3	39.9	32.2	26.7
History of diabetes, (%)	6.0	7.0	8.3	10.2	11.6	13.4	10.1	10.5	11.0	11.3	11.7	12.6
Blood pressure-lowering medication prescriptions, (%)	6.9	9.9	13.7	18.7	24.2	30.5	12.3	16	20.5	25.6	30.0	35.9
Family history, (%)	6.1	7.0	7.5	7.6	7.0	6.7	7.4	0.9	9.7	10.4	10.4	9.8
Chronic kidney disease (4/5), (%)	0.2	0.1	0.2	0.2	0.2	0.3	0.2	0.2	0.2	0.2	0.3	0.3
Rheumatoid arthritis, (%)	0.4	0.6	0.7	0.8	1.1	1.2	1.1	1.4	1.6	2.0	2.5	2.7
Atrial fibrillation (%)	0.3	0.5	0.8	1.2	1.9	3.0	0.2	0.2	0.4	0.6	0.9	1.5
Coronary artery disease PRS**	-1.07	-1.10	-1.12	-1.14	-1.17	-1.17	-1.08	-1.10	-1.11	-1.13	-1.14	-1.15
Stroke PRS**	1.59	1.58	1.57	1.56	1.154	1.54	1.59	1.58	1.57	1.56	1.55	1.55

Abbreviations: BMI, body mass index; HDL cholesterol, high-density lipoprotein cholesterol; PRS, polygenic risk score.

*A mean Townsend score of -1.5 from UK Biobank due to insufficient data.

** Mean PRS values from UK Biobank were used due to lack of genetic data in CPRD.

4.2.5 Population health modelling

Population health modelling was conducted to compare the population health impact of 1) prioritising using a primary care records-based tool followed by a formal assessment with conventional risk factors, 2) prioritising using a PRS and age-based tool followed by a formal assessment with conventional risk factors and PRS and 3) prioritising using both PRS and primary care records, followed by a formal assessment with conventional risk factors and PRS.

A hypothetical population of 100,000 individuals (50,000 men and women) from the United Kingdom was created; the population age structure was obtained using data from the ONS in 2015[35] and the number of expected CVD events was calculated using age-group and sex-specific incidence rates from CPRD **(Table 4.3, Figure 4.4)**. A policy of statin initiation for individuals at ≥10% predicted 10-year formal assessment CVD risk as currently recommended by National Institute for Health and Care Excellence (NICE) guidelines and a 20% reduction in CVD risk were assumed.[36,37] The population health impact for each of the three prioritisation tools was modelled using age- and sex-specific prioritisation thresholds in two ways. First, we selected prioritisation thresholds to limit the formal risk assessment false negative rate to 5%. Second, we selected prioritisation thresholds for the tools utilising PRS, such that the same number of events identified would be equivalent to prioritising with primary care records (**Table 4.4)**.

Summary metrics were estimated for: the number needed to screen (NNS) to prevent one CVD event, the number of CVD events identified and the number needed to invite (NNI) to prevent one CVD event. We assumed 50% statin compliance and a 50% invitation uptake of a formal assessment if inviting all individuals.[38,39] We further assumed an increased invitation uptake of 55% if individuals were prioritised for an invitation to a formal assessment.

In sensitivity analyses, we repeated population-health analyses including all individuals, including those without a primary care record for any one of SBP, HDL, total cholesterol or BMI, where those without a record were all invited for formal assessment **(Table 4.3).** We also repeated analyses assuming a 5% formal risk assessment threshold, in addition to age- and sex-specific prioritisation thresholds selected to correspond to 2.5% false negative rates.

Analyses were conducted in R x64 3.6.1 (R Foundation for Statistical Computing, Vienna, Austria).

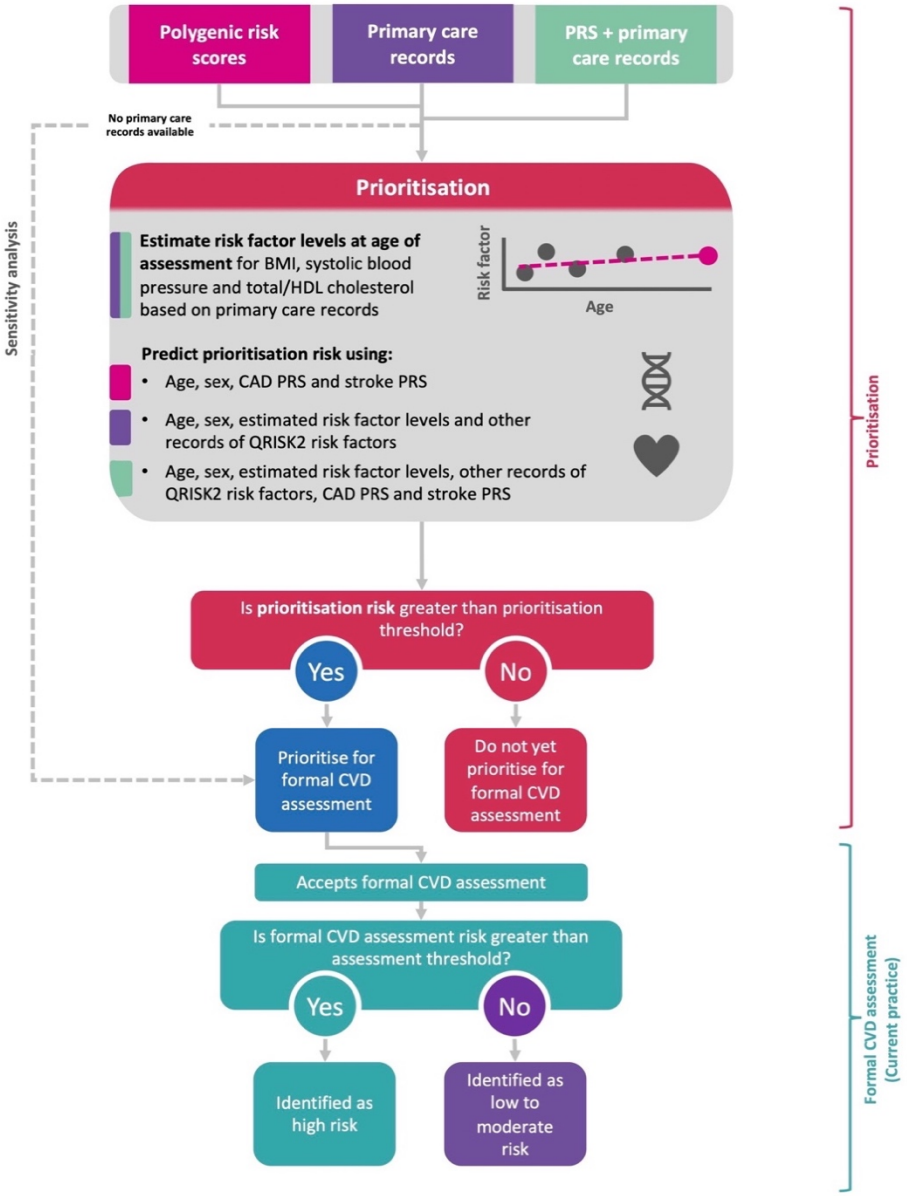


Figure 4.4: Flow chart of the implementation of a prioritisation tool for formal cardiovascular disease assessments

Abbreviations: BMI, body mass index; CAD, coronary artery disease; CVD; cardiovascular disease; HDL, high density lipoprotein; PRS, polygenic risk score.

Table 4.3: Age- and sex-specific crude 10-year cardiovascular disease incidence rates using records from 870,486 individuals in the CPRD database.

	Men						Women					
Age group	**40-44**	**45-49**	**50-54**	**55-59**	**60-64**	**65-69**	**40-44**	**45-49**	**50-54**	**55-59**	**60-64**	**65-69**
Crude incidence rate per 1,000 with at least at least one primary care record of systolic blood pressure, cholesterol, or BMI	18.2	34.4	58.1	86.6	112.3	144.6	11.5	18.0	28.6	39.0	58.1	73.3
Crude incidence rate per 1,000 including those without at least one primary care record of systolic blood pressure, cholesterol, or BMI	17.4	33.0	56.9	83.9	109.7	142.6	11.4	17.8	28.3	39.1	58.0	72.8

Abbreviations: BMI, body mass index.

Table 4.4: Age- and sex-specific prioritisation thresholds chosen for population health modelling

Age group	Prioritisation using primary care records	Prioritisation using PRS		Prioritisation using PRS and primary care records	
	5% FNR prioritisation threshold, %	5% FNR prioritisation threshold, %	Equivalent events prioritisation threshold, %	5% FNR prioritisation threshold, %	Equivalent events prioritisation threshold, %
Men					
40-49	3.4%	3.3%	3.5%	3.9%	4.5%
50-59	6.5%	6.3%	6.5%	6.6%	7.4%
60-69	10.0%	10.6%	10.8%	9.6%	10.4%
Women					
40-49	3.0%	1.8%	1.9%	2.9%	4.2%
50-59	4.3%	3.6%	3.8%	4.5%	6.0%
60-69	6.9%	7.6%	7.9%	7.0%	8.6%

Abbreviations: FNR, false negative rate; PRS, polygenic risk score

Age group and sex specific 5% FNR prioritisation thresholds were defined as the level such that the expected false negative rate of the formal risk assessment is controlled to be 5%. The prioritisation thresholds were chosen by first, ranking the estimated 10-year CVD risks from each prioritisation tool amongst individuals with a future CVD event. The FNR threshold was selected as the maximum estimated risk such that 5% of individuals with a future event would not be prioritised (i.e. were lower than the threshold).

Age group and sex specific 'equivalent events prioritisation thresholds', when prioritising using PRS or PRS and primary care records tool, were chosen such that the number of events identified would be similar to if prioritising with primary care records using a 5% FNR prioritisation threshold.

4.3 Results

4.3.1 Population characteristics in UK Biobank

For our primary analysis, we identified 108,685 individuals in UKB with genetic data and a primary care record for at least one of SBP, HDL, total cholesterol and BMI **(Figure 4.5).** All individuals had complete information for the conventional risk factors necessary to calculate a 10-year formal CVD risk at baseline survey. The mean age at baseline was 56.2 years (SD 8.0) for men and 56.1 years (SD 7.8) for women. During mean of follow-up of 8.2 years, there were 1,838 incident cardiovascular events (**Table 4.5).** Compared to the measurements observed at the UKB baseline survey, the measurements recorded in primary care records were lower for SBP and total cholesterol and although similar for current/ever smoking status and history of diabetes, were less concordant for the remaining disease statuses. The oldest primary care record was on average 3.8 years before baseline.

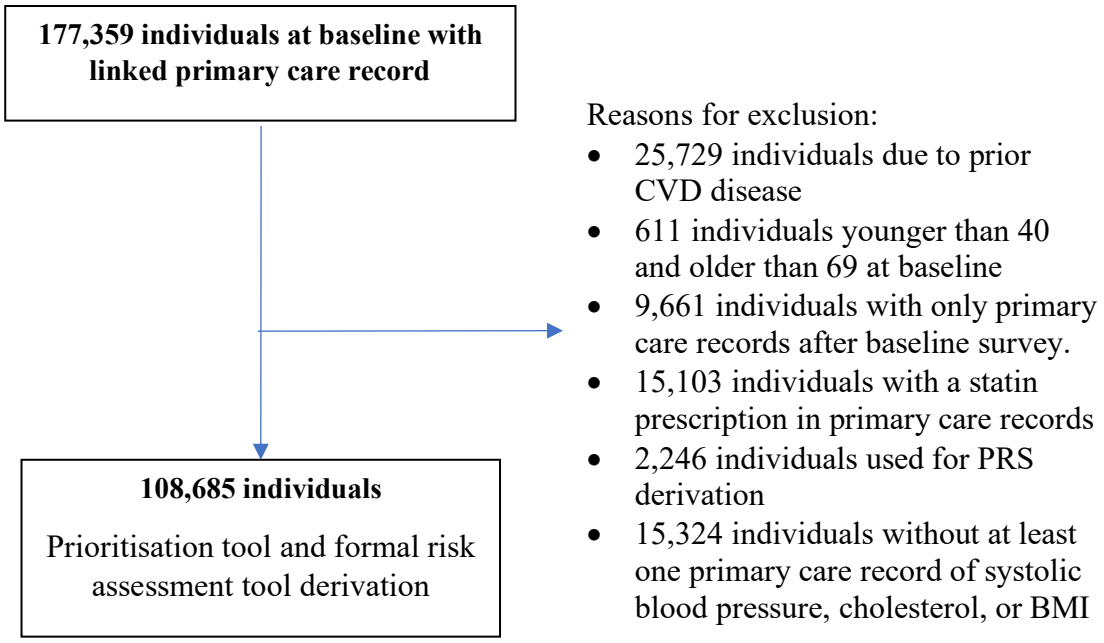


Figure 4.5: Flowchart showing selection of patient records for derivation and population health modelling in UK Biobank.

Abbreviations: BMI, body mass index; CVD, cardiovascular disease; PRS, polygenic risk score.
Highlighted in red are individuals without necessary primary care records for primary care based prioritisation tool that were formally assessed in sensitivity analysis.

Table 4.5: Key characteristics of individuals in UK Biobank baseline survey and linked primary care records

Characteristic	Men, N = 44,184 (41%)		Women, N = 64,501 (59%)	
CVD events, N	1230		608	
Follow up duration: years, median (5^{th}, 95^{th} percentile)	8.1 (6.1, 10.8)		8.2 (6.8, 10.9)	
Duration between first primary care record and baseline visit: years, median (5^{th}, 95^{th} percentile)	3.6 (0.9, 5.5)		3.9 (1.1, 5.6)	
	Primary care records	**Baseline**	**Primary care records**	**Baseline**
Age, mean (SD) [a]	-	56.2 (8.0)	-	56.1 (7.8)
Coronary artery disease PRS, mean (SD)	-	-1.15 (0.46)	-	-1.13 (0.46)
Stroke PRS, mean (SD)	-	1.55 (0.22)	-	1.56 (0.23)
Ethnicity — White, N (%) [a]	-	42,283 (95.7%)	-	61,977 (96.1%)
Townsend, mean (SD) [a]	-	-1.5 (3.0)	-	-1.5 (2.9)
Systolic blood pressure:				
mmHg, mean (SD) [b]	135.3 (7.82)	141.0 (17.3)	130.7 (9.62)	134.9 (19.1)
Number of historical records, mean	3.8	-	4.6	-
Total cholesterol:				
mmol/litre, mean (SD) [b]	5.48 (0.47)	5.79 (1.01)	5.71 (0.50)	6.03 (1.08)
Number of historical records, mean	2.0	-	2.0	-
HDL cholesterol:				
mmol/litre, mean (SD) [b]	1.35 (0.19)	1.30 (0.31)	1.68 (0.24)	1.61 (0.37)
Number of historical records, mean	1.8	-	1.9	-
BMI:				
kg/m^2, mean (SD) [b]	27.2 (3.1)	27.5 (4.1)	26.6 (4.1)	26.8 (5.0)
Number of historical records, mean	2.0	-	2.3	-
Current/ever smoker, N (%)	4,472 (10.1%)	5,233 (11.8%)	4,911 (7.61%)	5,511 (8.5%)
History of diabetes, N (%)	466 (1.05%)	630 (1.4%)	412 (0.64%)	459 (0.7%)
Blood pressure-lowering medication prescriptions, N (%)	6,396 (14.5%)	5,529 (12.51%)	9,737 (15.1%)	7,643 (11.85%)
Family history, N (%) [a]	1,568 (3.55%)	-	2,494 (3.87%)	-
Chronic kidney disease (4/5), N (%) [a]	57 (0.13%)	-	79 (0.12%)	-
Rheumatoid arthritis, N (%)	146 (0.33%)	381 (0.86%)	336 (0.52%)	989 (1.53%)
Atrial fibrillation, N (%)	336 (0.33%)	123 (0.28%)	1,749 (2.7%)	89 (0.14%)

Abbreviations: BMI, body mass index; CVD, cardiovascular disease; HDL cholesterol, high-density lipoprotein cholesterol; PRS, polygenic risk score; SD, standard deviation.

[a] Risk factor values in both baseline and primary care records if one was missing.

[b] Risk factor values for primary care records estimated using multivariate mixed effects model.

4.3.2 Model performance and comparison

Hazard ratios (HRs) in the prioritisation tools and formal assessment models, for the same predictors, were similar **(Tables 4.6-4.7).** All models had good discriminatory performance with higher performance in women **(Table 4.8)**. The greatest performance was observed in the model using conventional risk factors and PRS in men (C-index = 0.716, 95% CI: 0.702, 0.730) and in women (C-index = 0.742, 95% CI: 0.722, 0.762). Compared to using conventional risk factors, augmenting with PRS also improved the reclassification of high and low risk individuals in both men (NRI = 0.0262, 95% CI: 0.0072, 0.0458) and in women (NRI = 0.0265, 95% CI: 0.0065, 0.0502) (**Table 4.9).**

The estimated 10-year risks between the primary care records only prioritisation tool and the formal assessment model using conventional risk factors were highly correlated (correlation coefficient = 0.75 for men and 0.80 for women). In contrast, the estimated 10-year risks between the PRS + age prioritisation tool and the formal assessment model using conventional risk factors and PRS were less highly correlated (correlation coefficients = 0.67 for men and women), and the estimated 10-year risks between the PRS and primary care records based prioritisation tool and the formal assessment model using conventional risk factors with PRS were more highly correlated (correlation coefficients = 0.82 for men and women) **(Table 4.10).** Rescaled 10-year risk estimates between all models were similar **(Figure 4.6).**

Table 4.6: Hazard ratios (95% confidence intervals) for the prioritisation and formal risk assessment tools derived using 44.184 men in UK Biobank.

Risk factor	Primary care records prioritisation tool	PRS + age prioritisation tool	PRS + primary care records prioritisation tool	Conventional risk factor formal assessment tool	Conventional risk factor + PRS formal assessment tool
Age – per year increase	1.071 (1.060, 1.083)	1.070 (1.062, 1.078)	1.075 (1.063, 1.086)	1.072 (1.060, 1.083)	1.075 (1.064, 1.087)
Ethnicity – non-White	0.946 (0.680, 1.316)	NA	0.464 (0.324, 0.665)	0.910 (0.654, 1.266)	0.463 (0.324, 0.663)
Townsend	1.032 (1.012, 1.052)	NA	1.030 (1.010, 1.050)	1.029 (1.010, 1.049)	1.028 (1.008, 1.048)
Smoking status – current/ever smoker	2.060 (1.765, 2.404)	NA	2.028 (1.737, 2.367)	1.979 (1.704, 2.298)	1.957 (1.685, 2.272)
Diabetes status - Yes	1.066 (0.548, 2.074)	NA	1.040 (0.535, 2.020)	1.524 (0.968, 2.400)	1.486 (0.943, 2.341)
Chronic kidney disease (stages 4/5)	2.777 (1.149, 6.709)	NA	2.691 (1.113, 6.504)	2.699 (1.118, 6.516)	2.655 (1.100, 6.412)
History of atrial fibrillation - Yes	0.662 (0.320, 1.368)	NA	0.651 (0.316, 1.342)	1.740 (0.614, 4.929)	1.666 (0.599, 4.631)
Anti-hypertensive medication - Yes	1.087 (0.912, 1.294)	NA	1.068 (0.897, 1.272)	1.203 (1.006, 1.439)	1.169 (0.978, 1.398)
Rheumatoid arthritis – Yes	0.624 (0.201, 1.940)	NA	0.625 (0.201, 1.942)	1.614 (1.047, 2.488)	1.562 (1.013, 2.408)
Family history of CVD – Yes	1.252 (0.941, 1.666)	NA	1.169 (0.878, 1.556)	1.254 (0.942, 1.668)	1.186 (0.891, 1.579)
Total cholesterol – per mmol/litre increase	1.571 (1.386, 1.780)	NA	1.503 (1.326, 1.705)	1.303 (1.231, 1.380)	1.276 (1.205, 1.351)
HDL – per mmol/litre increase	0.373 (0.256, 0.543)	NA	0.398 (0.274, 0.579)	0.411 (0.330, 0.511)	0.424 (0.341, 0.527)
Systolic blood pressure – per mmHg increase	1.025 (1.017, 1.034)	NA	1.023 (1.014, 1.032)	1.013 (1.009, 1.016)	1.012 (1.008, 1.015)
BMI – per kg/m2 increase	1.008 (0.987, 1.030)	NA	1.007 (0.986, 1.029)	1.017 (1.002, 1.033)	1.015 (1.000, 1.031)
CAD PRS – per SD increase	NA	1.299 (1.221, 1.381)	1.312 (1.231, 1.397)	NA	1.299 (1.219, 1.384)
Stroke PRS – per SD increase	NA	1.123 (1.059, 1.192)	1.162 (1.090, 1.239)	NA	1.150 (1.078, 1.226)
Age * BMI	1.000 (0.997, 1.002)	NA	1.000 (0.997, 1.002)	1.000 (0.998, 1.002)	1.000 (0.998, 1.002)
Age * Townsend	0.998 (0.996, 1.001)	NA	0.998 (0.995, 1.000)	0.998 (0.996, 1.001)	0.998 (0.996, 1.001)
Age * Systolic blood pressure	0.999 (0.998, 1.000)	NA	0.999 (0.998, 1.000)	1.000 (0.999, 1.000)	1.000 (0.999, 1.000)
Age * Family history of CVD	0.991 (0.954, 1.030)	NA	0.993 (0.955, 1.032)	0.994 (0.957, 1.032)	0.996 (0.959, 1.035)
Age * Smoking status	0.984 (0.964, 1.005)	NA	0.985 (0.965, 1.006)	0.981 (0.962, 1.000)	0.982 (0.963, 1.002)
Age * Anti-hypertensive medication	0.994 (0.972, 1.017)	NA	0.993 (0.971, 1.016)	0.984 (0.962, 1.008)	0.983 (0.961, 1.007)
Age * Diabetes status	1.069 (0.988, 1.158)	NA	1.070 (0.988, 1.158)	1.016 (0.961, 1.075)	1.018 (0.962, 1.077)
Age * History of atrial fibrillation	1.060 (0.968, 1.161)	NA	1.058 (0.966, 1.158)	1.026 (0.905, 1.164)	1.024 (0.906, 1.157)
Baseline survival estimate at 10 years	0.9668333	0.9670151	0.9672866	0.9777007	0.977394

Abbreviations: BMI, body mass index; CAD, coronary artery disease; CVD, cardiovascular disease; HDL cholesterol, high-density lipoprotein cholesterol; PRS, polygenic risk score; SD, standard deviation

Table 4.7: Hazard ratios (95% confidence intervals) for the prioritisation and formal risk assessment tools derived using 64,501 women in UK Biobank.

Risk factor	Primary care records prioritisation tool	PRS + age prioritisation tool	PRS + primary care records prioritisation tool	Conventional risk factor formal assessment tool	Conventional risk factor + PRS formal assessment tool
Age – per year increase	1.065 (1.045, 1.085)	1.094 (1.081, 1.107)	1.067 (1.048, 1.087)	1.076 (1.059, 1.094)	1.078 (1.061, 1.096)
Ethnicity – non-White	0.936 (0.573, 1.530)	NA	0.578 (0.338, 0.988)	0.872 (0.533, 1.426)	0.562 (0.329, 0.962)
Townsend	1.068 (1.038, 1.100)	NA	1.067 (1.036, 1.098)	1.064 (1.034, 1.096)	1.063 (1.032, 1.094)
Smoking status – current/ever smoker	2.695 (2.134, 3.405)	NA	2.662 (2.107, 3.362)	2.528 (2.003, 3.190)	2.502 (1.983, 3.157)
Diabetes status – History or	1.770 (0.811, 3.861)	NA	1.771 (0.812, 3.865)	1.501 (0.670, 3.363)	1.479 (0.659, 3.319)
Chronic kidney disease (stages 4/5)	0.841 (0.118, 5.993)	NA	0.795 (0.111, 5.667)	0.735 (0.102, 5.290)	0.695 (0.096, 5.021)
History of atrial fibrillation - Yes	0.874 (0.445, 1.717)	NA	0.881 (0.45, 1.726)	0.062 (0.000, 19.868)	0.064 (0.000, 21.315)
Anti-hypertensive medication - Yes	0.995 (0.771, 1.286)	NA	0.982 (0.76, 1.269)	1.658 (1.296, 2.123)	1.622 (1.267, 2.076)
Rheumatoid arthritis – Yes	1.934 (0.917, 4.077)	NA	1.921 (0.911, 4.05)	1.479 (0.924, 2.367)	1.495 (0.934, 2.393)
Family history of CVD – Yes	1.124 (0.734, 1.722)	NA	1.095 (0.714, 1.677)	1.097 (0.716, 1.679)	1.074 (0.701, 1.645)
Total cholesterol – per mmol/litre increase	1.406 (1.181, 1.675)	NA	1.369 (1.149, 1.631)	1.227 (1.137, 1.324)	1.210 (1.121, 1.307)
HDL – per mmol/litre increase	0.365 (0.238, 0.559)	NA	0.375 (0.245, 0.575)	0.540 (0.420, 0.694)	0.547 (0.425, 0.703)
Systolic blood pressure – per mmHg increase	1.038 (1.026, 1.049)	NA	1.036 (1.025, 1.048)	1.015 (1.010, 1.020)	1.014 (1.009, 1.019)
BMI – per kg/m2 increase	0.990 (0.966, 1.014)	NA	0.990 (0.966, 1.014)	1.011 (0.994, 1.030)	1.011 (0.993, 1.029)
CAD PRS – per SD increase	NA	1.178 (1.079, 1.286)	1.165 (1.064, 1.274)	NA	1.152 (1.053, 1.260)
Stroke PRS – per SD increase	NA	1.117 (1.025, 1.217)	1.115 (1.016, 1.225)	NA	1.106 (1.008, 1.215)
Age * BMI	1.002 (0.999, 1.004)	NA	1.002 (0.999, 1.004)	1.001 (0.999, 1.003)	1.001 (0.999, 1.003)
Age * Townsend	0.998 (0.994, 1.002)	NA	0.998 (0.994, 1.002)	0.998 (0.994, 1.001)	0.998 (0.994, 1.001)
Age * Systolic blood pressure	0.999 (0.998, 1.000)	NA	0.999 (0.998, 1.000)	1.000 (0.999, 1.000)	1.000 (0.999, 1.001)
Age * Family history of CVD	0.974 (0.916, 1.036)	NA	0.974 (0.916, 1.036)	0.977 (0.920, 1.037)	0.977 (0.920, 1.037)
Age * Smoking status	1.009 (0.977, 1.042)	NA	1.009 (0.977, 1.042)	1.019 (0.987, 1.051)	1.018 (0.987, 1.051)
Age * Anti-hypertensive medication	1.005 (0.972, 1.040)	NA	1.006 (0.972, 1.040)	0.975 (0.944, 1.007)	0.975 (0.944, 1.008)
Age * Diabetes status	1.005 (0.904, 1.117)	NA	1.006 (0.905, 1.118)	1.030 (0.928, 1.143)	1.032 (0.930, 1.146)
Age * History of atrial fibrillation	1.066 (0.979, 1.162)	NA	1.065 (0.978, 1.160)	1.551 (0.906, 2.653)	1.540 (0.898, 2.643)
Baseline survival estimate at 10 years	0.9859755	0.9896236	0.9862068	0.9915961	0.9915758

Abbreviations: BMI, body mass index; CAD, coronary artery disease; CVD, cardiovascular disease; HDL cholesterol, high-density lipoprotein cholesterol; PRS, polygenic risk score; SD, standard deviation

Table 4.8: C indices of prioritisation tools and formal CVD risk assessment tools in UK Biobank

Model	C-index (95% confidence interval)		
	All individuals	Men	Women
Prioritisation tool			
Primary care records only	0.730 (0.719, 0.741)	0.684 (0.670, 0.699)	0.734 (0.715, 0.754)
PRS + age	0.663 (0.652, 0.675)	0.663 (0.649, 0.678)	0.686 (0.665, 0.707)
PRS + primary care records	0.740 (0.730, 0.751)	0.704 (0.691, 0.718)	0.738 (0.718, 0.758)
Formal risk assessment tool			
Conventional risk factors	0.730 (0.719, 0.740)	0.700 (0.686, 0.714)	0.739 (0.720, 0.759)
Conventional risk factors +PRS	0.738 (0.727, 0.749)	0.716 (0.702, 0.730)	0.742 (0.722, 0.762)

Abbreviations: PRS, polygenic risk score.

C indices and 95% confidence intervals from each model for the prediction of 10-year cardiovascular disease by sex and for the combined population in UK Biobank after 10-fold cross validation.

Table 4.9: Net reclassification improvement between prioritisation tools and formal risk assessment tools

Reference	Comparison	Net reclassification improvement (95% confidence interval)		
		Combined	Men	Women
Primary care records only prioritisation tool	PRS + age prioritisation tool	-0.0508 (-0.0787, -0.0299)	-0.0487 (-0.0814, -0.0131)	-0.0870 (-0.1257, -0.0537)
Primary care records only prioritisation tool	PRS + primary care records prioritisation tool	0.0220 (0.0074, 0.0392)	0.0230 (-0.0005, 0.0481)	0.0160 (-0.0036, 0.0327)
PRS + age prioritisation tool	PRS + primary care records prioritisation tool	0.0729 (0.0570, 0.0928)	0.0717 (0.0481, 0.1026)	0.1030 (0.0765, 0.1346)
Conventional risk factor formal assessment tool	Conventional risk factor + PRS formal assessment tool	0.0237 (0.0078, 0.0409)	0.0262 (0.0072, 0.0458)	0.0265 (0.0065, 0.0502)

Abbreviations: PRS, polygenic risk score.

Table 4.10: Correlation of predicted 10-year risks between prioritisation tools and formal assessment tools by sex in the derivation dataset

Men	**Primary care records only prioritisation tool**	**PRS + age prioritisation tool**	**PRS + primary care records prioritisation tool**
Conventional risk factor formal assessment tool	0.75	-	-
Conventional risk factor + PRS formal assessment tool	-	0.67	0.82
Women	**Primary care records only prioritisation tool**	**PRS + age prioritisation tool**	**PRS + primary care records prioritisation tool**
Conventional risk factor formal assessment tool	0.80	-	-
Conventional risk factor + PRS formal assessment tool	-	0.67	0.82

Abbreviations: PRS, polygenic risk score.

Correlations shown are for combinations of prioritisation tool and formal assessment tool as defined in each strategy.

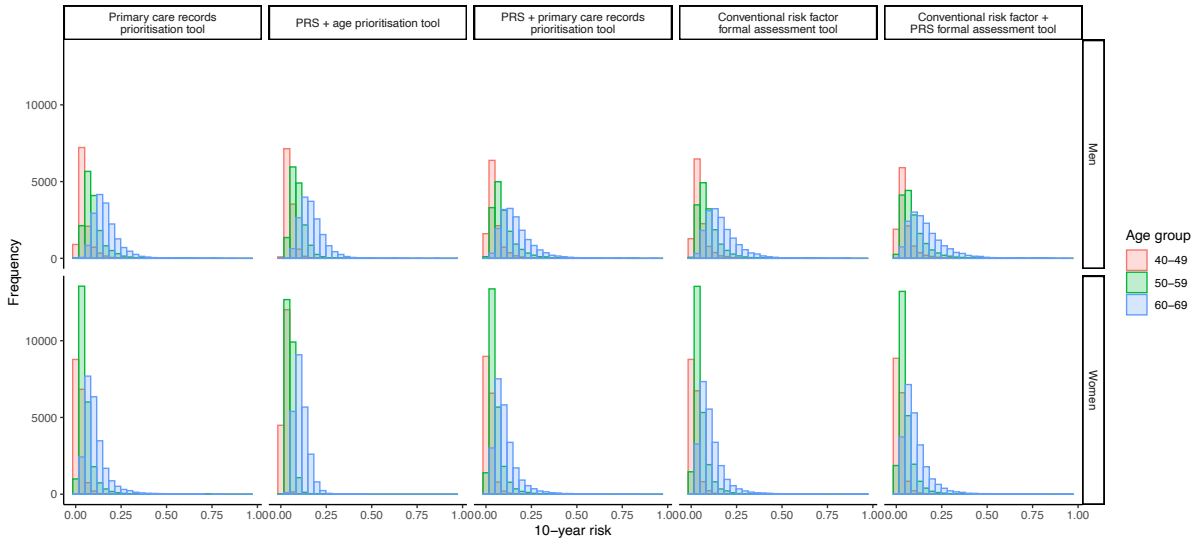


Figure 4.6: Age group and sex specific distributions of rescaled 10-year risks for each prioritisation tool and formal assessment tool.

Abbreviations: PRS, polygenic risk score.

4.3.3 Population health modelling

4.3.3.1 Prioritisation using primary care records

In our representative population of 100,000 individuals aged 40 to 69, 3,573 men and 1,808 women would experience a CVD event over the next 10 years. If conventional risk factors were used to formally assess the whole population, 2,426 (67.9%) men and 801 (44.3%) women would be identified at high risk (**Figure 4.7, Table 4.11**). Assuming statin therapy would be initiated on high-risk individuals, and no other preventive interventions implemented, the NNS to prevent one CVD event in men and women would be 206 (95% CI: 200, 213) and 624 (95% CI: 576, 668) respectively. If the primary care records-based prioritisation tool was first used to prioritise formal assessment in the population, then 2,335 (65.3%) men and 785 (43.4%) women would be identified at high risk. The NNS to prevent one event would reduce to 149 (95% CI: 143, 155) in men and 280 (95% CI: 259, 301) in women (27.7% and 55.1% reduction respectively). The greatest reduction in the NNS would be in men and women aged between 40-49 years, with a reduction of 55% and 82% respectively. The reduction was almost statistically significant at the 5% level in 40-49 year old men.

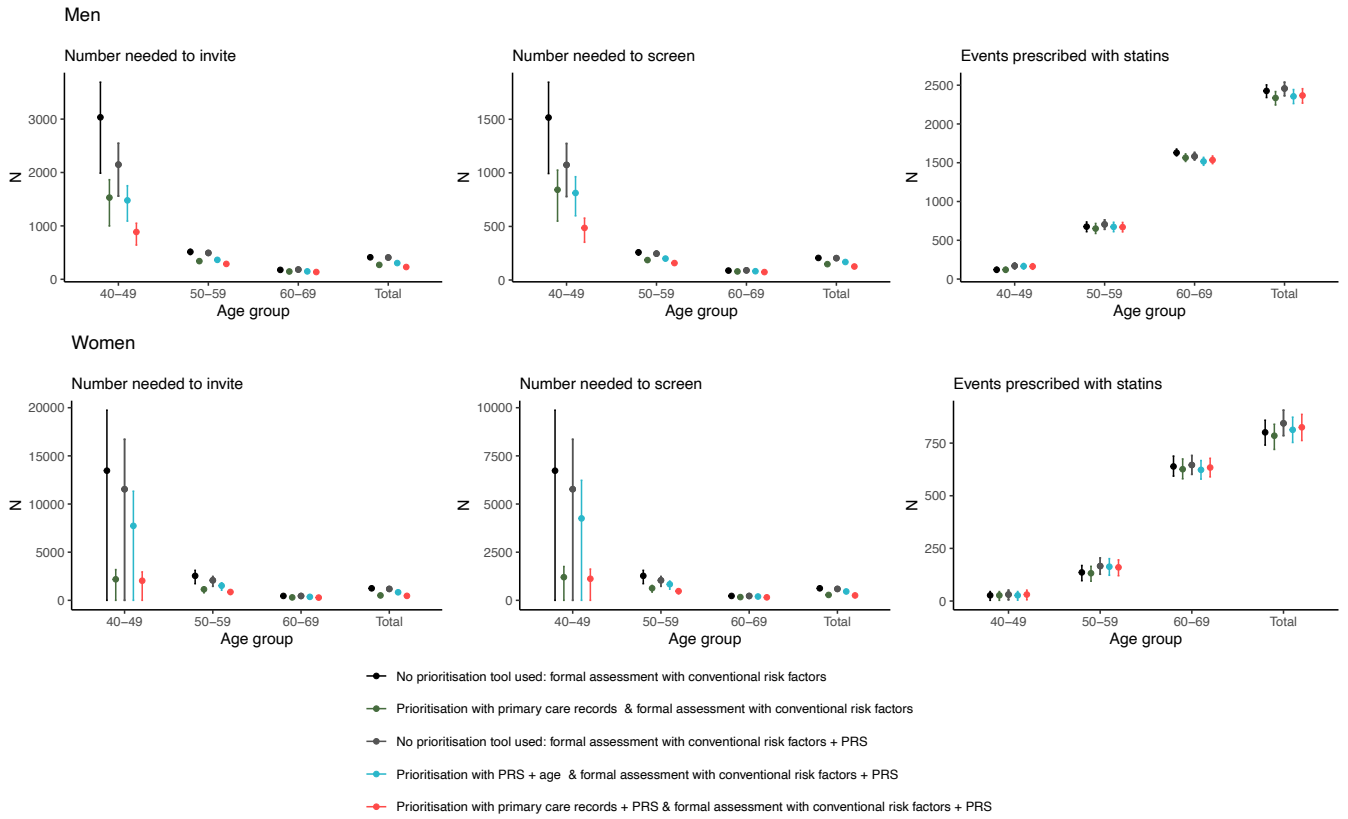


Figure 4.7: Number needed to invite, number needed to screen and number of events identified after prioritising for a formal CVD assessment, in a hypothetical population of 100,000 individuals in England.

Abbreviations: NNS, number needed to screen; NNI, number needed to invite; PRS, polygenic risk score.

95% confidence intervals are represented by vertical lines. Age group and sex specific prioritisation thresholds were defined as the level such that the expected false negative rate is controlled to be 5%. NNI and NNS assumes 50% statin compliance, and half of all individuals invited for formal assessment attend.

Table 4.11: Number needed to invite and screen to prevent one event, and number of events identified when prioritising with primary care records in a hypothetical population of 100,000 individuals in England.

			No prioritisation tool used: formal assessment with conventional risk factors used for all individuals			**Prioritisation using primary care records, followed by formal assessment with conventional risk factors**			
Age group	Participants	Expected number of events in 10 years	NNI (95% CI)	NNS (95% CI)	Number of events identified as high risk (%)	Participants prioritised (%)	NNI (95% CI)	NNS (95% CI)	Number of events identified as high risk (%)
Men									
40-49	18253	485	3033 (1986.4, 3692.0)	1517 (993.2, 1846.0)	120 (24.7%)	10126 (55.5%)	1530 (997.9, 1865.6)	841 (548.9, 1026.1)	120 (24.8%)
50-59	17391	1240	515 (463.8, 560.2)	257 (231.9, 280.1)	676 (54.5%)	12134 (69.8%)	339 (301.2, 368.3)	187 (165.7, 202.6)	651 (52.5%)
60-69	14356	1847	176 (171.0, 180.6)	88 (85.5, 90.3)	1629 (88.2%)	12517 (87.2%)	146 (140.8, 149.8)	80 (77.5, 82.4)	1564 (84.7%)
Total	**50000**	**3573**	**412 (398.4, 426.3)**	**206 (199.2, 213.1)**	**2426 (67.9%)**	**34777 (69.6%)**	**271 (260.7, 281.1)**	**149 (143.4, 154.6)**	**2335 (65.3%)**
Women									
40-49	18107	269	13462 (0.0, 19743.6)	6731 (0.0, 9871.8)	27 (10.0%)	3233 (17.9%)	2185 (0.0, 3192.3)	1202 (0.0, 1755.7)	27 (10.0%)
50-59	17282	577	2544 (1729.3, 3115.2)	1272 (864.6, 1557.6)	136 (23.6%)	8329 (48.2%)	1143 (778.9, 1403.4)	629 (428.4, 771.9)	132 (23.0%)
60-69	14611	962	457 (418.9, 488.1)	229 (209.4, 244.1)	639 (66.4%)	10459 (71.6%)	304 (277.4, 324.1)	167 (152.6, 178.3)	626 (65.1%)
Total	**50000**	**1808**	**1248 (1151.8, 1336.1)**	**624 (575.9, 668.0)**	**801 (44.3%)**	**22021 (44.0%)**	**510 (470.8, 547.0)**	**280 (258.9, 300.8)**	**785 (43.4%)**

Abbreviations: NNS, number needed to screen; NNI, number needed to invite.

Age structure of hypothetical population extrapolated from Office for National Statistics, England, United Kingdom 2015. Expected events at 10 years based on extrapolation of incidence rates from CPRD, 2014-2019. Age group and sex specific prioritisation thresholds were defined as the level such that the expected false negative rate was controlled to be 5%. NNI and NNS assumes 50% statin compliance. NNI assumes a 50% invitation uptake if assessing without using prioritisation tool, and a 55% invitation uptake if assessing with using prioritisation tool. Formal assessment threshold set at 10% and prioritisation threshold set at equivalents to 5% false negative rate.

4.3.3.2 Prioritisation using polygenic risk scores

If conventional risk factors enhanced with PRS was used to formally assess the whole population, then 2,457 (68.8%) men and 844 (46.7%) women would be identified as being at high risk **(Figure 4.7, Table 4.12)**. The NNS to prevent one CVD event in men and women would be 204 (95% CI: 197, 211) and 592 (95% CI: 545, 631) respectively.

If the PRS + age prioritisation tool was first used to prioritise formal assessment in the population, then 78.8% of men and 74.8% of women would be prioritised and, amongst them, 2,356 (65.9%) men and 813 (45.0%) women with CVD events over the next 10 years would be classified at high risk. The NNS to prevent one event would reduce to 167 (95% CI: 161, 174) in men and 460 (95% CI: 423, 491) in women (18.1% and 22.3% reduction respectively). The largest reduction in the NNS would be in the youngest men and women, with a reduction of 24% and 26% respectively.

If the PRS and primary care records-based prioritisation tool was first used to prioritise formal assessment in the population, then 2,367 (66.3%) men and 825 (45.6%) women would be classified at high risk. **(Figure 4.7, Table 4.13).** The NNS to prevent one event would reduce to 127 (95% CI: 122, 132) in men and 255 (95% CI: 234, 273) in women (37.7% and 56.9% reduction respectively). The largest reduction in the NNS would be in the youngest men and women, with a reduction of 55% and 80% respectively. The reduction in the youngest men was statistically significant at the 5% level. Comparing with other age groups shows that despite having the fewest CVD events, the large reductions in NNS in 40-49 year olds can have a positive impact.

Table 4.12: Number needed to invite and screen to prevent one event, and number of events identified when prioritising with PRS + age in a hypothetical population of 100,000 individuals in England.

			No prioritisation tool used: formal assessment with conventional risk factors + PRS used for all individuals			**Prioritisation using PRS + age, followed by formal assessment with conventional risk factors + PRS**			
Age group	Participants	Expected number of events in 10 years	NNI (95% CI)	NNS (95% CI)	Number of events identified as high risk (%)	Participants prioritised (%)	NNI (95% CI)	NNS (95% CI)	Number of events identified as high risk (%)
Men									
40-49	18253	485	2149 (1557.2, 2547.5)	1074 (778.6, 1273.8)	170 (35.1%)	13525 (74.1%)	1478 (1088.5, 1752.7)	813 (598.7, 964.0)	166 (34.3%)
50-59	17391	1240	494 (448.6, 534.4)	247 (224.3, 267.2)	705 (56.9%)	13456 (77.4%)	364 (328.9, 394.1)	200 (180.9, 216.7)	673 (54.3%)
60-69	14356	1847	181 (175.5, 186.7)	91 (87.7, 93.3)	1582 (85.7%)	12424 (86.5%)	149 (143.3, 153.9)	82 (78.8, 84.6)	1517 (82.1%)
Total	**50000**	**3573**	**407 (393.0, 422.0)**	**204 (196.5, 211.0)**	**2457 (68.8%)**	**39405 (78.8%)**	**304 (292.3, 315.5)**	**167 (160.8, 173.5)**	**2356 (65.9%)**
Women									
40-49	18107	269	11538 (0.0, 16718.6)	5769 (0.0, 8359.3)	31 (11.5%)	11442 (63.2%)	7733 (0.0, 11331.6)	4253 (0.0, 6232.4)	27 (10.0%)
50-59	17282	577	2077 (1455.0, 2472.3)	1038 (727.5, 1236.2)	166 (28.8%)	13590 (78.6%)	1516 (1050.2, 1820.7)	834 (577.6, 1001.4)	163 (28.3%)
60-69	14611	962	452 (418.5, 481.4)	226 (209.3, 240.7)	646 (67.2%)	12357 (84.6%)	360 (333.3, 383.7)	198 (183.3, 211.0)	623 (64.8%)
Total	**50000**	**1808**	**1185 (1090.5, 1261.7)**	**592 (545.2, 630.8)**	**844 (46.7%)**	**37389 (74.8%)**	**836 (768.7, 893.4)**	**460 (422.8, 491.4)**	**813 (45.0%)**

Abbreviations: NNS, number needed to screen; NNI, number needed to invite; PRS, polygenic risk score.

Age structure of hypothetical population extrapolated from Office for National Statistics, England, United Kingdom 2015. Expected events at 10 years based on extrapolation of incidence rates from CPRD, 2014-2019. Age group and sex specific prioritisation thresholds were defined as the level such that the expected false negative rate was controlled to be 5%. NNI and NNS assumes 50% statin compliance. NNI assumes a 50% invitation uptake if assessing without using prioritisation tool, and a 55% invitation uptake if assessing with using prioritisation tool. Formal assessment threshold set at 10% and prioritisation threshold set at equivalents to 5% false negative rate.

Table 4.13: Number needed to invite and screen to prevent one event, and number of events identified when prioritising with PRS and primary care records in a hypothetical population of 100,000 individuals in England.

			No prioritisation tool used: formal assessment with conventional risk factors + PRS used for all individuals			**Prioritisation using PRS and primary care records, followed by formal assessment with conventional risk factors + PRS**			
Age group	Participants	Expected number of events in 10 years	NNI (95% CI)	NNS (95% CI)	Number of events identified as high risk (%)	Participants prioritised (%)	NNI (95% CI)	NNS (95% CI)	Number of events identified as high risk (%)
Men									
40-49	18253	485	2149 (1557.2, 2547.5)	1074 (778.6, 1273.8)	170 (35.1%)	7930 (43.4%)	885 (640.1, 1051.0)	487 (352.1, 578.0)	163 (33.6%)
50-59	17391	1240	494 (448.6, 534.4)	247 (224.3, 267.2)	705 (56.9%)	10622 (61.1%)	288 (259.9, 312.3)	159 (143.0, 171.8)	670 (54.0%)
60-69	14356	1847	181 (175.5, 186.7)	91 (87.7, 93.3)	1582 (85.7%)	11436 (79.7%)	135 (130.4, 139.7)	75 (71.7, 76.9)	1535 (83.1%)
Total	**50000**	**3573**	**407 (393.0, 422.0)**	**204 (196.5, 211.0)**	**2457 (68.8%)**	**29988 (60.0%)**	**230 (221.3, 239.5)**	**127 (121.7, 131.7)**	**2367 (66.3%)**
Women									
40-49	18107	269	11538 (0.0, 16718.6)	5769 (0.0, 8359.3)	31 (11.5%)	3513 (19.4%)	2035 (0.0, 2953.9)	1119 (0.0, 1624.7)	31 (11.7%)
50-59	17282	577	2077 (1455.0, 2472.3)	1038 (727.5, 1236.2)	166 (28.8%)	7616 (44.1%)	867 (613.9, 1038.0)	477 (337.6, 570.9)	160 (27.7%)
60-69	14611	962	452 (418.5, 481.4)	226 (209.3, 240.7)	646 (67.2%)	9872 (67.6%)	283 (261.8, 301.7)	156 (144.0, 165.9)	634 (65.9%)
Total	**50000**	**1808**	**1185 (1090.5, 1261.7)**	**592 (545.2, 630.8)**	**844 (46.7%)**	**21001 (42.0%)**	**463 (426.0, 495.6)**	**255 (234.3, 272.6)**	**825 (45.6%)**

Abbreviations: NNS, number needed to screen; NNI, number needed to invite; PRS, polygenic risk score.

Age structure of hypothetical population extrapolated from Office for National Statistics, England, United Kingdom 2015. Expected events at 10 years based on extrapolation of incidence rates from CPRD, 2014-2019. Age group and sex specific prioritisation thresholds were defined as the level such that the expected false negative rate was controlled to be 5%. NNI and NNS assumes 50% statin compliance. NNI assumes a 50% invitation uptake if assessing without using prioritisation tool, and a 55% invitation uptake if assessing with using prioritisation tool. Formal assessment threshold set at 10% and prioritisation threshold set at equivalents to 5% false negative rate.

4.3.3.3 Prioritisation assuming strategies identify same number of events

Choosing prioritisation thresholds such that all strategies would identify the same number of events to if prioritising using primary care records **(Table 4.4)**, prioritising using PRS and age resulted in a NNS of 164 (95% CI: 157, 170) in men and 446 (95% CI: 410, 480) in women. Prioritising using PRS and primary care records resulted in a NNS of 116 (95% CI: 111, 121) in men and 180 (95% CI: 166, 193) in women. **(Table 4.14, Figure 4.8).** Compared to using primary care records, prioritising using PRS and primary care records led to statistically significant differences in the NNS at the 5% level for all except in women aged 40-49.

Table 4.14: Number needed to invite and screen to prevent one event and number of events identified after prioritisation and formal assessment in a hypothetical population of 100,000 individuals in England, with prioritisation thresholds selected to identify the same number of events if prioritising with primary care records with prioritisation thresholds controlling the false negative rate to 5%.

		Prioritisation using primary care records followed by conventional risk factors				**Prioritisation using PRS + age, followed by conventional risk factors + PRS**				**Prioritisation using PRS and primary care records, followed by conventional risk factors + PRS**			
Age group	Participants	Participants prioritised (%)	NNI (95% CI)	NNS (95% CI)	Number of events identified as high risk (%)	Participants prioritised (%)	NNI (95% CI)	NNS (95% CI)	Number of events identified as high risk (%)	Participants prioritised (%)	NNI (95% CI)	NNS (95% CI)	Number of events identified as high risk (%)
Men													
40-49	18253	10126 (55.5%)	1530 (997.9, 1865.6)	841 (548.9, 1026.1)	120 (24.8%)	13016 (71.3%)	1454 (1060.7, 1713.0)	800 (583.4, 942.2)	163 (33.6%)	7208 (39.5%)	738 (531.5, 876.8)	406 (292.3, 482.2)	163 (33.6%)
50-59	17391	12134 (69.8%)	339 (301.2, 368.3)	187 (165.7, 202.6)	651 (52.5%)	13100 (75.3%)	358 (325.1, 386.6)	196 (178.8, 212.6)	666 (53.7%)	9878 (56.8%)	266 (236.7, 288.8)	146 (130.2, 158.9)	654 (52.7%)
60-69	14356	12517 (87.2%)	146 (140.8, 149.8)	80 (77.5, 82.4)	1564 (84.7%)	12196 (84.9%)	148 (141.7, 152.2)	80 (77.9, 83.7)	1506 (81.5%)	11064 (77.1%)	130 (124.2, 133.7)	72 (68.3, 73.6)	1519 (82.3%)
Total	**50000**	**34777 (69.6%)**	**271 (260.7, 281.1)**	**149 (143.4, 154.6)**	**2335 (65.3%)**	**38313 (76.6%)**	**298 (286.0, 309.9)**	**164 (157.3, 170.4)**	**2335 (65.4%)**	**28150 (56.3%)**	**210 (200.9, 219.1)**	**116 (110.5, 120.5)**	**2336 (65.4%)**
Women													
40-49	18107	3233 (17.9%)	2185 (0.0, 3192.3)	1202 (0.0, 1755.7)	27 (10.0%)	10139 (56.0%)	7100 (0.0, 10397.2)	3904 (0.0, 5718.5)	27 (10.0%)	1748 (9.7%)	1012 (0.0, 1462.2)	558 (0.0, 804.2)	31 (11.7%)
50-59	17282	8329 (48.2%)	1143 (778.9, 1403.4)	629 (428.4, 771.9)	132 (23.0%)	12436 (72.0%)	1484 (1013.9, 1783.4)	816 (557.7, 980.9)	156 (27.1%)	4683 (27.1%)	612 (397.5, 746.1)	336 (218.6, 410.3)	139 (24.1%)
60-69	14611	10459 (71.6%)	304 (277.4, 324.1)	167 (152.6, 178.3)	626 (65.1%)	11357 (77.7%)	356 (326.4, 379.5)	196 (179.5, 208.7)	603 (62.7%)	7714 (52.8%)	228 (210.4, 242.5)	126 (115.7, 133.4)	616 (64.0%)
Total	**50000**	**22021 (44.0%)**	**510 (470.8, 547.0)**	**280 (258.9, 300.8)**	**785 (43.4%)**	**33932 (67.9%)**	**812 (744.7, 872.1)**	**446 (409.6, 479.6)**	**786 (43.5%)**	**14145 (28.3%)**	**326 (301.0, 351.0)**	**180 (165.6, 193.0)**	**786 (43.5%)**

Abbreviations: NNI, number needed to invite; NNS, number needed to screen; PRS, polygenic risk score.

Age structure of hypothetical population extrapolated from Office for National Statistics, England, United Kingdom 2015. Expected events at 10 years based on extrapolation of incidence rates from CPRD, 2014-2019. Age group and sex specific prioritisation thresholds when prioritising using primary care records were defined as the level such that the expected false negative rate is controlled to be 5%. Age group and sex specific prioritisation thresholds when prioritising using PRS or PRS and primary care records tool were selected to result in the same number of events identified if prioritising using primary care records. NNI and NNS assumes 100% statin compliance. NNI assumes a 50% invitation uptake if assessing without using prioritisation tool, and a 55% invitation uptake if assessing with using prioritisation tool.

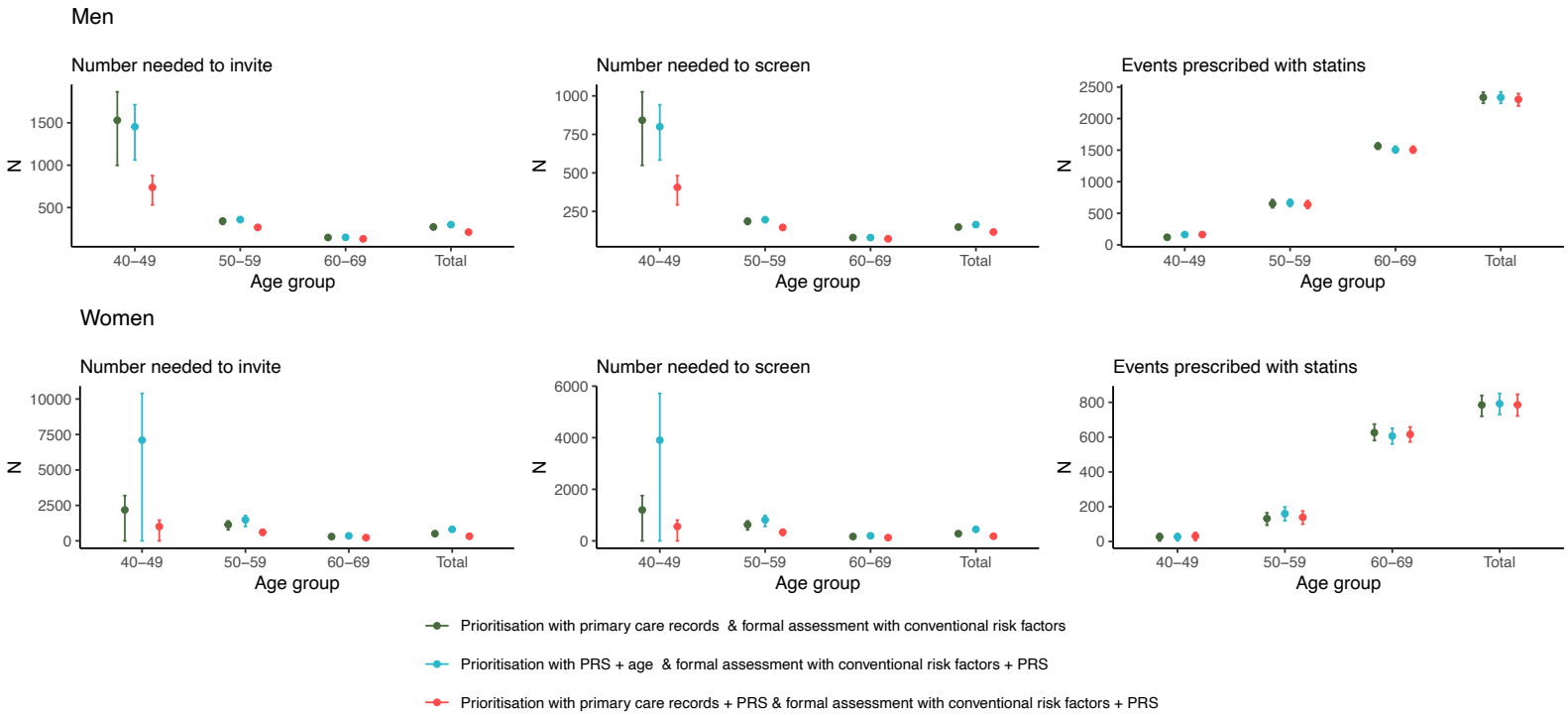


Figure 4.8: Number needed to invite and screen to prevent one event and number of events identified after prioritisation and formal assessment in a hypothetical population of 100,000 individuals in England, with prioritisation thresholds selected to identify the same number of events if prioritising with primary care records with prioritisation thresholds controlling the false negative rate to 5%.

Abbreviations: NNS, number needed to screen; NNI, number needed to invite; PRS, polygenic risk score.

95% confidence intervals are represented by vertical lines. Age group and sex specific prioritisation thresholds were defined as the level such that the expected false negative rate was controlled to be 5%. NNI and NNS assumes 50% statin compliance, and half of all individuals invited for formal assessment attend.

4.3.3.4 Sensitivity analyses

In sensitivity analyses including all individuals (i.e., including 15,324 individuals without a primary care record for any one of SBP, total cholesterol, HDL cholesterol or BMI) (**Table 4.15**), we found comparable results for the PRS-based prioritisation tool and the primary care-based prioritisation tool in men and women. As expected, we increased the NNS if prioritising with primary care records, especially amongst the youngest group (**Tables 4.16-4.18, Figure 4.9**).

In sensitivity analyses assuming a 5% formal risk assessment threshold, in addition to age- and sex-specific prioritisation thresholds selected to correspond to 2.5% false negative rates (**Table 4.19)**, we found significant improvements in the number of events identified, as expected. Consequently, this reduces the NNS when formally assessing all individuals, using either conventional risk factors or conventional risk factors enhanced with PRS **(Tables 4.20-4.22).** Whilst prioritisation can still reduce the overall NNS and the NNS amongst the youngest, the differences are smaller compared to when using a 10% formal risk assessment threshold.

Table 4.15: Summary of number of individuals without primary care records in UK Biobank.

Sex	Age group	Individuals without at least one CVD risk factor in primary care record N (%)
Men	40-49	3851 (25.0%)
	50-59	2736 (14.8%)
	60-69	1734 (9.3%)
Women	40-49	2714 (14.0%)
	50-59	2415 (9.2%)
	60-69	1874 (7.2%)

Abbreviations: CVD, cardiovascular disease.

Prioritisation with primary care records requires at least one CVD risk factor of: systolic blood pressure, total cholesterol, HDL cholesterol and/or BMI.

Table 4.16: Number needed to invite and screen to prevent one event and number of events identified when prioritising using primary care records, including all individuals without a primary care record for any one of SBP, HDL, total cholesterol or BMI, in a hypothetical population of 100,000 individuals in England.

			No prioritisation tool used: formal assessment with conventional risk factors used for all individuals			**Prioritisation using primary care records, followed by formal assessment with conventional risk factors**			
Age group	Participants	Expected number of events in 10 years	NNI (95% CI)	NNS (95% CI)	Number of events identified as high risk (%)	Participants prioritised (%)	NNI (95% CI)	NNS (95% CI)	Number of events identified as high risk (%)
Men									
40-49	18253	465	3042 (2126.4, 3710.8)	1521 (1063.2, 1855.4)	120 (25.8%)	12154 (66.6%)	1842 (1301.8, 2245.1)	1012 (716.0, 1234.8)	120 (25.8%)
50-59	17391	1207	536 (487.7, 577.3)	268 (243.9, 288.6)	650 (53.9%)	12914 (74.3%)	374 (340.0, 404.5)	206 (187.0, 222.5)	628 (52.0%)
60-69	14356	1814	178 (173.6, 182.4)	89 (86.8, 91.2)	1611 (88.8%)	12688 (88.4%)	148 (144.2, 152.4)	82 (79.3, 83.8)	1551 (85.5%)
Total	**50000**	**3486**	**420 (407.2, 433.8)**	**210 (203.6, 216.9)**	**2381 (68.3%)**	**37756 (75.5%)**	**298 (288.5, 308.9)**	**164 (158.7, 169.9)**	**2300 (66.0%)**
Women									
40-49	18107	267	14268 (0.0, 20446.9)	7134 (0.0, 10223.4)	25 (9.4%)	5316 (29.4%)	3808 (0.0, 5459.1)	2094 (0.0, 3002.5)	25 (9.5%)
50-59	17282	575	2526 (1502.0, 3023.8)	1263 (751.0, 1511.9)	137 (23.8%)	9153 (53.0%)	1244 (731.9, 1501.7)	684 (402.5, 825.9)	134 (23.3%)
60-69	14611	957	464 (434.0, 494.8)	232 (217.0, 247.4)	629 (65.7%)	10760 (73.6%)	316 (295.8, 338.7)	174 (162.7, 186.3)	617 (64.5%)
Total	**50000**	**1799**	**1264 (1167.6, 1350.6)**	**632 (583.8, 675.3)**	**792 (44.0%)**	**25229 (50.5%)**	**590 (543.1, 633.7)**	**324 (298.7, 348.5)**	**776 (43.2%)**

Abbreviations: HDL, high-density lipoprotein; NNI, number needed to invite; NNS, number needed to screen; SBP, systolic blood pressure.

Age structure of hypothetical population extrapolated from Office for National Statistics, England, United Kingdom 2015. Expected events at 10 years based on extrapolation of incidence rates from CPRD, 2014-2019. Age group and sex specific prioritisation thresholds were defined as the level such that the expected false negative rate was controlled to be 5%. NNI and NNS assumes 50% statin compliance. NNI assumes a 50% invitation uptake if assessing without using prioritisation tool, and a 55% invitation uptake if assessing with using prioritisation tool. Formal assessment threshold set at 10% and prioritisation threshold set at equivalents to 5% false negative rate.

Table 4.17: Number needed to invite and screen to prevent one event and number of events identified when prioritising using PRS and age, including all individuals without a primary care record for any one of SBP, HDL, total cholesterol or BMI, in a hypothetical population of 100,000 individuals in England.

			No prioritisation tool used: formal assessment with conventional risk factors + PRS used for all individuals			**Prioritisation using PRS + age, followed by formal assessment with conventional risk factors + PRS**			
Age group	Participants	Expected number of events in 10 years	NNI (95% CI)	NNS (95% CI)	Number of events identified as high risk (%)	Participants prioritised (%)	NNI (95% CI)	NNS (95% CI)	Number of events identified as high risk (%)
Men									
40-49	18253	465	2282 (1717.6, 2656.7)	1141 (858.8, 1328.4)	160 (34.4%)	13287 (72.8%)	1534 (1150.5, 1802.9)	844 (632.8, 991.6)	157 (33.9%)
50-59	17391	1207	506 (460.6, 543.5)	253 (230.3, 271.7)	688 (57.0%)	13338 (76.7%)	368 (335.9, 397.3)	202 (184.8, 218.5)	658 (54.5%)
60-69	14356	1814	184 (178.1, 188.8)	92 (89.0, 94.4)	1563 (86.2%)	12393 (86.3%)	150 (145.4, 156.0)	82 (80.0, 85.8)	1495 (82.4%)
Total	**50000**	**3486**	414 (400.2, 427.8)	207 (200.1, 213.9)	2411 (69.2%)	39018 (78.0%)	308 (296.0, 316.7)	168 (162.8, 174.2)	2310 (66.3%)
Women									
40-49	18107	267	12230 (0.0, 17211.5)	6115 (0.0, 8605.7)	30 (11.2%)	11224 (62.0%)	8040 (0.0, 11532.0)	4422 (0.0, 6342.6)	25 (9.5%)
50-59	17282	575	2098 (1490.7, 2462.6)	1049 (745.3, 1231.3)	165 (28.7%)	13545 (78.4%)	1552 (1081.7, 1836.9)	854 (594.9, 1010.3)	159 (27.6%)
60-69	14611	957	460 (430.5, 489.6)	230 (215.3, 244.8)	634 (66.2%)	12341 (84.5%)	368 (341.3, 393.0)	202 (187.7, 216.1)	610 (63.8%)
Total	**50000**	**1799**	1206 (1110.0, 1293.8)	603 (555.0, 646.9)	829 (46.1%)	37110 (74.2%)	850 (783.2, 914.0)	468 (430.8, 502.7)	794 (44.1%)

Abbreviations: HDL, high-density lipoprotein; NNI, number needed to invite; NNS, number needed to screen; PRS, polygenic risk score; SBP, systolic blood pressure.

Age structure of hypothetical population extrapolated from Office for National Statistics, England, United Kingdom 2015. Expected events at 10 years based on extrapolation of incidence rates from CPRD, 2014-2019. Age group and sex specific prioritisation thresholds were defined as the level such that the expected false negative rate was controlled to be 5%. NNI and NNS assumes 50% statin compliance. NNI assumes a 50% invitation uptake if assessing without using prioritisation tool, and a 55% invitation uptake if assessing with using prioritisation tool. Formal assessment threshold set at 10% and prioritisation threshold set at equivalents to 5% false negative rate.

Table 4.18: Number needed to invite and screen to prevent one event and number of events identified when prioritising using PRS and primary care records, including all individuals without a primary care record for any one of SBP, HDL, total cholesterol or BMI, in a hypothetical population of 100,000 individuals in England.

			No prioritisation tool used: formal assessment with conventional risk factors + PRS used for all individuals			**Prioritisation using PRS + age, followed by formal assessment with conventional risk factors + PRS**			
Age group	Participants	Expected number of events in 10 years	NNI (95% CI)	NNS (95% CI)	Number of events identified as high risk (%)	Participants prioritised (%)	NNI (95% CI)	NNS (95% CI)	Number of events identified as high risk (%)
Men									
40-49	18253	465	2282 (1717.6, 2656.7)	1141 (858.8, 1328.4)	160 (34.4%)	10505 (57.6%)	1232 (931.5, 1451.5)	678 (512.3, 798.3)	155 (33.3%)
50-59	17391	1207	506 (460.6, 543.5)	253 (230.3, 271.7)	688 (57.0%)	11626 (66.9%)	322 (290.4, 346.7)	176 (159.7, 190.7)	658 (54.5%)
60-69	14356	1814	184 (178.1, 188.8)	92 (89.0, 94.4)	1563 (86.2%)	11708 (81.6%)	140 (135.1, 144.4)	76 (74.3, 79.4)	1521 (83.8%)
Total	**50000**	**3486**	**414 (400.2, 427.8)**	**207 (200.1, 213.9)**	**2411 (69.2%)**	**33840 (67.7%)**	**264 (254.5, 273.2)**	**144 (140.0, 150.3)**	**2334 (66.9%)**
Women									
40-49	18107	267	12230 (0.0, 17211.5)	6115 (0.0, 8605.7)	30 (11.2%)	5557 (30.7%)	3412 (0.0, 4796.1)	1876 (0.0, 2637.9)	30 (11.1%)
50-59	17282	575	2098 (1490.7, 2462.6)	1049 (745.3, 1231.3)	165 (28.7%)	8506 (49.2%)	976 (674.4, 1143.8)	536 (370.9, 629.1)	159 (27.6%)
60-69	14611	957	460 (430.5, 489.6)	230 (215.3, 244.8)	634 (66.2%)	10215 (69.9%)	298 (278.5, 317.6)	164 (153.2, 174.7)	622 (65.0%)
Total	**50000**	**1799**	**1206 (1110.0, 1293.8)**	**603 (555.0, 646.9)**	**829 (46.1%)**	**24278 (48.6%)**	**544 (499.8, 582.7)**	**300 (274.9, 320.5)**	**810 (45.0%)**

Abbreviations: HDL, high-density lipoprotein; NNI, number needed to invite; NNS, number needed to screen; PRS, polygenic risk score; SBP, systolic blood pressure.

Age structure of hypothetical population extrapolated from Office for National Statistics, England, United Kingdom 2015. Expected events at 10 years based on extrapolation of incidence rates from CPRD, 2014-2019. Age group and sex specific prioritisation thresholds were defined as the level such that the expected false negative rate was controlled to be 5%. NNI and NNS assumes 50% statin compliance. NNI assumes a 50% invitation uptake if assessing without using prioritisation tool, and a 55% invitation uptake if assessing with using prioritisation tool. Formal assessment threshold set at 10% and prioritisation threshold set at equivalents to 5% false negative rate.

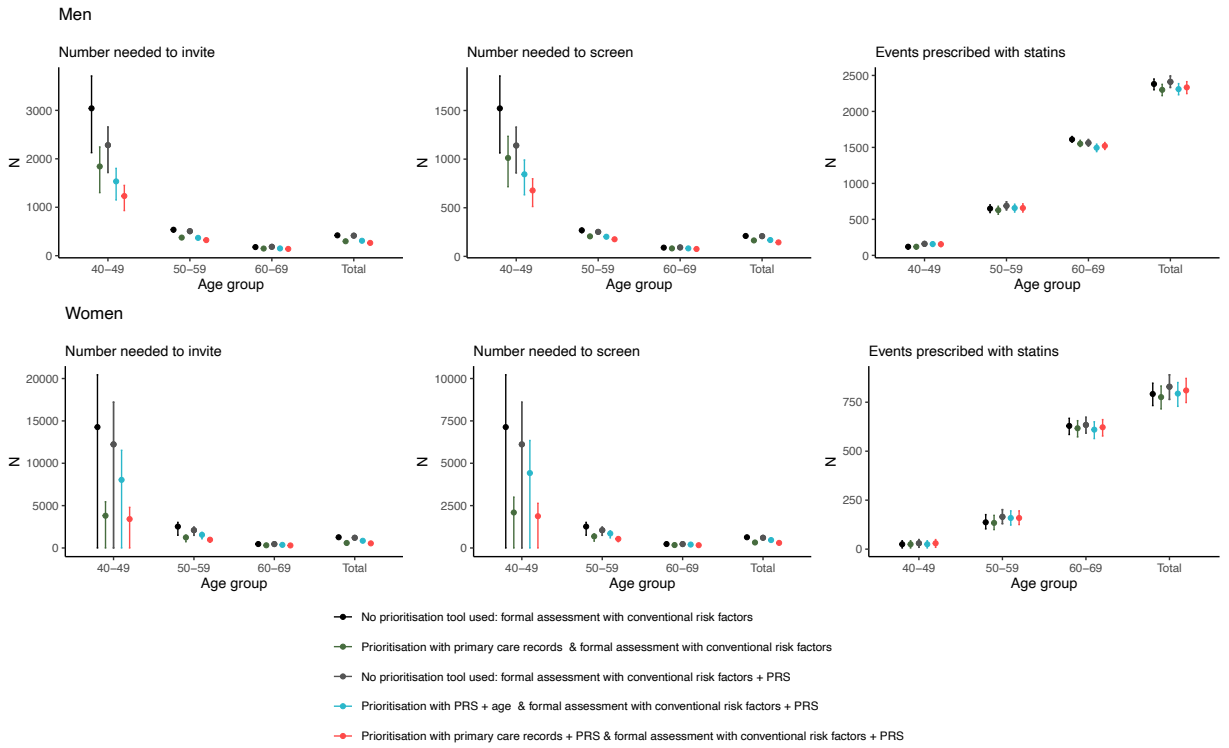


Figure 4.9: Number needed to invite, number needed to screen and number of events identified after prioritising for a formal CVD assessment, including all individuals without a primary care record for any one of SBP, HDL, total cholesterol or BMI, in a hypothetical population of 100,000 individuals in England.

Abbreviations: HDL, high-density lipoprotein; NNI, number needed to invite; NNS, number needed to screen; PRS, polygenic risk score; SBP, systolic blood pressure.

95% confidence intervals are represented by vertical lines. Age group and sex specific prioritisation thresholds were defined as the level such that the expected false negative rate was controlled to be 5%. NNI and NNS assumes 50% statin compliance, and half of all individuals invited for formal assessment attend.

Table 4.19: Age- and sex-specific prioritisation thresholds chosen for population health modelling with 2.5% false negative rate prioritisation thresholds.

	Prioritisation using primary care records	**Prioritisation using PRS**	**Prioritisation using PRS and primary care records**
Age group	2.5% FNR prioritisation threshold, %	2.5% FNR prioritisation threshold, %	2.5% FNR prioritisation threshold, %
Men			
40-49	2.2%	2.7%	2.3%
50-59	4.3%	4.9%	4.1%
60-69	7.5%	8.3%	6.2%
Women			
40-49	1.7%	1.4%	1.6%
50-59	2.8%	3.0%	2.9%
60-69	4.5%	6.5%	4.3%

Abbreviations: FNR, false negative rate; PRS, polygenic risk score

Age group and sex specific 2.5% FNR prioritisation thresholds were defined as the level such that the expected false negative rate of the formal risk assessment is controlled to be 2.5%. The prioritisation thresholds were chosen by first, ranking the estimated 10-year CVD risks from each prioritisation tool amongst individuals with a future CVD event. The FNR threshold was selected as the maximum estimated risk such that 2.5% of individuals with a future event would not be prioritised (i.e. were lower than the threshold).

Table 4.20: Number needed to invite and screen to prevent one event, and number of events identified when prioritising with primary care records in a hypothetical population of 100,000 individuals in England, assuming a 5% formal risk assessment threshold.

			No prioritisation tool used: formal assessment with conventional risk factors used for all individuals			**Prioritisation using primary care records, followed by formal assessment with conventional risk factors**			
Age group	Participants	Expected number of events in 10 years	NNI (95% CI)	NNS (95% CI)	Number of events identified as high risk (%)	Participants prioritised (%)	NNI (95% CI)	NNS (95% CI)	Number of events identified as high risk (%)
Men									
40-49	18253	485	1063 (937.6, 1160.5)	532 (468.8, 580.3)	343 (70.7%)	14972 (82.0%)	810 (708.0, 888.0)	445 (389.4, 488.4)	336 (69.3%)
50-59	17391	1240	310 (300.4, 319.7)	155 (150.2, 159.8)	1122 (90.5%)	15891 (91.4%)	260 (251.2, 268.0)	143 (138.2, 147.4)	1113 (89.7%)
60-69	14356	1847	157 (155.5, 157.7)	78 (77.8, 78.8)	1831 (99.1%)	13873 (96.6%)	140 (137.7, 140.9)	77 (75.7, 77.5)	1808 (97.9%)
Total	**50000**	**3573**	**303 (299.2, 308.0)**	**152 (149.6, 154.0)**	**3297 (92.3%)**	**44735 (89.5%)**	**250 (245.8, 254.1)**	**137 (135.2, 139.7)**	**3257 (91.2%)**
Women									
40-49	18107	269	4251 (2235.2, 5417.1)	2126 (1117.6, 2708.6)	85 (31.6%)	8316 (45.9%)	1775 (936.4, 2263.8)	976 (515.0, 1245.1)	85 (31.7%)
50-59	17282	577	901 (798.7, 984.6)	450 (399.3, 492.3)	384 (66.6%)	13140 (76.0%)	634 (559.1, 694.9)	349 (307.5, 382.2)	377 (65.3%)
60-69	14611	962	321 (313.2, 327.7)	160 (156.6, 163.9)	911 (94.7%)	13626 (93.3%)	276 (267.9, 282.5)	152 (147.3, 155.4)	898 (93.4%)
Total	**50000**	**1808**	**725 (697.2, 752.1)**	**362 (348.6, 376.0)**	**1380 (76.3%)**	**35082 (70.2%)**	**469 (449.4, 487.1)**	**258 (247.2, 267.9)**	**1361 (75.2%)**

Abbreviations: NNS, number needed to screen; NNI, number needed to invite.

Age structure of hypothetical population extrapolated from Office for National Statistics, England, United Kingdom 2015. Expected events at 10 years based on extrapolation of incidence rates from CPRD, 2014-2019. Age group and sex specific prioritisation thresholds were defined as the level such that the expected false negative rate was controlled to be 5%. NNI and NNS assumes 50% statin compliance. NNI assumes a 50% invitation uptake if assessing without using prioritisation tool, and a 55% invitation uptake if assessing with using prioritisation tool. Formal assessment threshold set at 5% and prioritisation thresholds set at equivalent to 2.5% false negative rate.

Table 4.21: Number needed to invite and screen to prevent one event, and number of events identified when prioritising with PRS + age in a hypothetical population of 100,000 individuals in England, assuming a 5% formal risk assessment threshold.

			No prioritisation tool used: formal assessment with conventional risk factors + PRS used for all individuals			**Prioritisation using PRS + age, followed by formal assessment with conventional risk factors + PRS**			
Age group	Participants	Expected number of events in 10 years	NNI (95% CI)	NNS (95% CI)	Number of events identified as high risk (%)	Participants prioritised (%)	NNI (95% CI)	NNS (95% CI)	Number of events identified as high risk (%)
Men									
40-49	18253	485	1063 (938.9, 1169.9)	532 (469.4, 585.0)	343 (70.7%)	15963 (87.5%)	863 (747.8, 955.3)	475 (411.3, 525.4)	336 (69.3%)
50-59	17391	1240	317 (306.4, 327.8)	159 (153.2, 163.9)	1097 (88.5%)	15932 (91.6%)	266 (256.9, 275.9)	147 (141.3, 151.7)	1087 (87.7%)
60-69	14356	1847	157 (155.7, 158.1)	79 (77.9, 79.1)	1829 (99.0%)	13800 (96.1%)	139 (137.6, 140.8)	77 (75.7, 77.5)	1800 (97.5%)
Total	**50000**	**3573**	**306 (301.0, 311.5)**	**153 (150.5, 155.7)**	**3269 (91.5%)**	**45695 (91.4%)**	**258 (253.1, 262.5)**	**142 (139.2, 144.4)**	**3224 (90.2%)**
Women									
40-49	18107	269	4487 (2184.7, 5762.6)	2244 (1092.3, 2881.3)	81 (30.1%)	15735 (86.9%)	3545 (1720.9, 4546.4)	1950 (946.5, 2500.5)	81 (30.0%)
50-59	17282	577	893 (788.5, 971.7)	446 (394.3, 485.9)	387 (67.1%)	15653 (90.6%)	755 (668.7, 826.2)	415 (367.8, 454.4)	377 (65.3%)
60-69	14611	962	326 (317.3, 334.4)	163 (158.7, 167.2)	896 (93.1%)	13869 (94.9%)	286 (276.5, 294.2)	158 (152.1, 161.8)	880 (91.5%)
Total	**50000**	**1808**	**733 (704.6, 759.3)**	**367 (352.3, 379.6)**	**1364 (75.4%)**	**45257 (90.5%)**	**615 (591.0, 637.9)**	**338 (325.0, 350.9)**	**1338 (74.0%)**

Abbreviations: NNS, number needed to screen; NNI, number needed to invite; PRS, polygenic risk score.

Age structure of hypothetical population extrapolated from Office for National Statistics, England, United Kingdom 2015. Expected events at 10 years based on extrapolation of incidence rates from CPRD, 2014-2019. Age group and sex specific prioritisation thresholds were defined as the level such that the expected false negative rate was controlled to be 5%. NNI and NNS assumes 50% statin compliance. NNI assumes a 50% invitation uptake if assessing without using prioritisation tool, and a 55% invitation uptake if assessing with using prioritisation tool. Formal assessment threshold set at 5% and prioritisation thresholds set at equivalent to 2.5% false negative rate.

Table 4.22: Number needed to invite and screen to prevent one event, and number of events identified when prioritising with PRS and primary care records in a hypothetical population of 100,000 individuals in England, assuming a 5% formal risk assessment threshold.

			No prioritisation tool used: formal assessment with conventional risk factors + PRS used for all individuals			**Prioritisation using PRS and primary care records, followed by formal assessment with conventional risk factors + PRS**			
Age group	Participants	Expected number of events in 10 years	NNI (95% CI)	NNS (95% CI)	Number of events identified as high risk (%)	Participants prioritised (%)	NNI (95% CI)	NNS (95% CI)	Number of events identified as high risk (%)
Men									
40-49	18253	485	1063 (938.9, 1169.9)	532 (469.4, 585.0)	343 (70.7%)	13310 (72.9%)	705 (624.0, 777.2)	388 (343.2, 427.5)	343 (70.8%)
50-59	17391	1240	317 (306.4, 327.8)	159 (153.2, 163.9)	1097 (88.5%)	14961 (86.0%)	251 (241.4, 259.8)	138 (132.8, 142.9)	1084 (87.4%)
60-69	14356	1847	157 (155.7, 158.1)	79 (77.9, 79.1)	1829 (99.0%)	13629 (94.9%)	136 (135.1, 137.7)	75 (74.3, 75.8)	1816 (98.3%)
Total	**50000**	**3573**	**306 (301.0, 311.5)**	**153 (150.5, 155.7)**	**3269 (91.5%)**	**41899 (83.8%)**	**235 (230.7, 239.3)**	**129 (126.9, 131.6)**	**3243 (90.8%)**
Women									
40-49	18107	269	4487 (2184.7, 5762.6)	2244 (1092.3, 2881.3)	81 (30.1%)	8708 (48.1%)	1962 (960.2, 2525.6)	1079 (528.1, 1389.1)	81 (30.0%)
50-59	17282	577	893 (788.5, 971.7)	446 (394.3, 485.9)	387 (67.1%)	12540 (72.6%)	610 (538.4, 665.9)	336 (296.1, 366.3)	374 (64.8%)
60-69	14611	962	326 (317.3, 334.4)	163 (158.7, 167.2)	896 (93.1%)	13376 (91.5%)	276 (267.5, 283.9)	152 (147.1, 156.2)	880 (91.5%)
Total	**50000**	**1808**	**733 (704.6, 759.3)**	**367 (352.3, 379.6)**	**1364 (75.4%)**	**34623 (69.2%)**	**472 (451.1, 489.4)**	**259 (248.1, 269.2)**	**1335 (73.8%)**

Abbreviations: NNS, number needed to screen; NNI, number needed to invite; PRS, polygenic risk score.

Age structure of hypothetical population extrapolated from Office for National Statistics, England, United Kingdom 2015. Expected events at 10 years based on extrapolation of incidence rates from CPRD, 2014-2019. Age group and sex specific prioritisation thresholds were defined as the level such that the expected false negative rate was controlled to be 5%. NNI and NNS assumes 50% statin compliance. NNI assumes a 50% invitation uptake if assessing without using prioritisation tool, and a 55% invitation uptake if assessing with using prioritisation tool. Formal assessment threshold set at 5% and prioritisation threshold set at equivalents to 2.5% false negative rate

4.4 Discussion

This study has rigorously assessed the impact of using PRS both alone and in combination with traditional risk factors for systematically prioritising individuals for a formal CVD risk assessment. Comparing against current recommendations of using existing primary care records, we found that adding PRS to both prioritisation and formal assessment improves their correlation. This subsequently leads to higher efficiency and effectiveness, especially amongst younger individuals. Consequently, augmenting primary care records with PRS reduces the NNS by around 20% and 35% in men and women respectively, relative to using primary care records alone and identifying the same number of events. In contrast, using only PRS and age in a prioritisation tool leads to a larger NNS. These results support the addition of PRS with primary care records to prioritise individuals at highest risk for a formal CVD risk assessment, which could lead to better allocation of resources by reducing the number of assessments.

This study has provided a comparison of prioritisation tools using longitudinal primary care records and/or PRS within a population in England aged between 40 and 69 years who are currently invited for a National Health Service (NHS) Health Check to assess their individual risk of CVD. We have demonstrated the benefits of PRS not only by measuring model discrimination, but also by evaluating the health impact if implemented within this population. Compared with previous studies which have generally focussed on the role of PRS in a formal CVD risk assessment model,[20,21,29,40] our study has uniquely assessed its role in a prioritisation tool, in conjunction with a CVD risk model. We have also shown that if PRS were widely available, the inclusion of PRS in a prioritisation tool could improve the effectiveness of a prioritisation tool especially in younger individuals by reducing the reliance on primary care records.

The benefits in prioritising a subgroup of those individuals at low absolute risk to increase efficiency echoes other studies, which have also shown that selecting a smaller proportion of younger, low-risk individuals can lead to dramatically reduced costs whilst resulting in more Quality Adjusted Life Years (QALYs) gained.[41] Whilst the addition of PRS has the potential to prioritise individuals earlier for a formal CVD assessment, further extensions include using PRS to identify individuals at high risk of other common chronic diseases, including diabetes, dementia and kidney disease.[42–46]

Future healthcare systems will however need to focus on the implementation of PRS and subsequent prioritisation. In our study, we assumed all individuals had PRS necessary for a direct comparison of the different prioritisation tools and formal risk assessment models. A more realistic implementation of PRS could exist whereby individuals are prioritised using primary care records and undergo a formal risk assessment using conventional CVD risk factors. Genetic data may then be collected afterwards, resulting in a more seamless approach to collecting the data required future PRS use.

In addition, our study focussed on supplementing conventional CVD risk factors with a CAD and stroke PRS. However, recent research has led to development of PRS for conventional CVD risk factors, including blood pressure and cholesterol levels.[47–50] These have shown potential and offer similar predictive performance in place of measured risk factors. As such, a future risk model may also include PRS for these risk factors to further enhance performance.

Policymakers should also focus on the costs of implementing PRS and whether its utility remains cost effective. In particular, health economics analysis should focus not only on the benefits of PRS for CVD, but also taking into account its utility for other chronic diseases and cancers.

4.4.1 Strengths

Our study has several strengths. This study is the first of our knowledge to directly compare how using different data types for a prioritisation tool can impact on the CVD risk assessment programme in England. This was possible due to the unique data linkage of primary care records along with a baseline survey in UK Biobank. We derived the PRS based prioritisation tools using two current and well documented PRS that have been shown to improve model performance independent of traditional CVD risk factors. We also took advantage of the sporadically observed longitudinal primary care records when deriving the primary care records-based prioritisation tools, by estimating current risk factor values using a multivariate mixed model. Whilst QRISK2, which replaces missing values with age, sex and ethnicity-specific population average values, could have been used as a prioritisation tool, we chose to optimise the available data in primary care records to reduce possible over-inflation of the information from PRS. Noteworthy, we would expect greater improvements if augmenting PRS in CVD risk scores with fewer risk factors. Another strength of this study is the use of 10-fold

cross validation to correct for over optimism that may exist in our analyses as we derived and conducted the population health modelling in the same individuals. Further, we used rescaling methods to adjust the 10-year risk estimates for all of the models to minimise the healthy selection bias when deriving models in UK Biobank, and to ensure results were representative to the general population of England. Such rescaling methods could be adapted towards other countries.

4.4.2 Limitations

However, several potential limitations exist. First, whilst we used primary care records that were no more than six and a half years old before baseline, the mean risk factor levels between primary care records and at the UKB baseline differed within the same individuals which could lead to a different distribution of 10-year risk estimates. This may also weaken the correlations between the prioritisation tool and formal risk assessment models reported. Second, we determined the number of events identified in the population health modelling by calculating the model's sensitivity in UKB and translating to a hypothetical population; due to the low number of events in UKB, the sensitivity of each model may be limited in accuracy, especially in younger age groups with fewer events. Third, PRS for cardiovascular disease are still under active development and, while we utilise two extensively studied and validated PRS, there are likely more powerful PRS soon to be available.[51] Fourth, the age-range of the population health modelling was limited to between 40 and 69 years old due to the use of UK Biobank. This restricts the population health modelling and in particular limits the ability to investigate the early prioritisation capabilities of PRS (which are fixed at conception). Fifth, we have focussed on estimating the differences in a primary care population in England. Further work should generalise the findings to other countries and their respective healthcare systems. Sixth, we assumed a constant 20% reduction in risk due to statins, which is unlikely in practice, where reductions may be greater in those with a greater genetic risk.[52] Seventh, we did not model the combined effects of other preventative interventions, such as lifestyle advice. It is likely that the benefits of communicating polygenic risk may lead to beneficial lifestyle changes, which may impact health outcomes.[53] Finally, we acknowledge our use of ICD-10 codes to identify CVD outcomes may have missed some events, although this is unlikely to affect our between model comparisons.

4.5 Conclusion

Population health guidelines in England recommend individuals at higher estimated risk of CVD be prioritised for formal risk assessment. Our results show that incorporating PRS improves the correlation between prioritisation tools and formal CVD risk assessment models. In particular, the use of PRS together with primary care records to prioritise individuals at highest risk of a CVD event for a formal CVD risk assessment has the ability to efficiently prioritise those who need interventions the most, which could lead to better allocation of resources by reducing the number of formal risk assessments in primary care.

Chapter 5 will expand on the work shown in **Chapter 4** but address some of its limitations. In particular, **Chapter 5** will aim to improve the population health modelling by generalising the risk assessment process over a wider range of ages (25-75 years versus 40-69 years). The work will also take into account the impact of competing risks, the number of risk assessments needed across an individual's lifetime and personalised invitation strategies using PRS.

References

1. Roth GA, Johnson C, Abajobir A, et al. Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. *J Am Coll Cardiol*. 2017;70(1):1-25. doi:10.1016/j.jacc.2017.04.052
2. Wallace ML, Ricco JA, Barrett B. Screening strategies for cardiovascular disease in asymptomatic adults. *Prim Care - Clin Off Pract*. 2014;41(2):371-397. doi:10.1016/j.pop.2014.02.010
3. Siontis GCM, Tzoulaki I, Siontis KC, Ioannidis JPA. Comparisons of established risk prediction models for cardiovascular disease: Systematic review. *BMJ*. 2012;344(7859). doi:10.1136/bmj.e3318
4. Hippisley-Cox J, Coupland C, Vinogradova Y, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ*. 2008;336(7659):1475-1482. doi:10.1136/BMJ.39609.449676.25
5. 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*. 2017;357. doi:10.1136/bmj.j2099
6. Piepoli MF, Hoes AW, Agewall S, et al. 2016 European guidelines on cardiovascular disease prevention in clinical practice. The Sixth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts. Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation. *G Ital Cardiol (Rome)*. 2017;18(7):547-612. doi:10.1714/2729.27821
7. Anderson TJ, Grégoire J, Pearson GJ, et al. 2016 Canadian Cardiovascular Society Guidelines for the Management of Dyslipidemia for the Prevention of Cardiovascular Disease in the Adult. *Can J Cardiol*. 2016;32(11):1263-1282. doi:10.1016/J.CJCA.2016.07.510
8. Arnett DK, Blumenthal RS, Albert MA, et al. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation*. 2019;140(11):e596-e646. doi:10.1161/CIR.0000000000000678
9. D'Agostino RB, Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: The Framingham heart study. *Circulation*. 2008;117(6):743-753. doi:10.1161/CIRCULATIONAHA.107.699579

10. Deanfield J, Sattar N, Simpson I, et al. Joint British Societies' consensus recommendations for the prevention of cardiovascular disease (JBS3). *Heart*. 2014;100(SUPPL. 2):ii1-ii67. doi:10.1136/heartjnl-2014-305693
11. Larsen LB, Sondergaard J, Thomsen JL, et al. Step-wise approach to prevention of chronic diseases in the Danish primary care sector with the use of a personal digital health profile and targeted follow-up- A n assessment of attendance. *BMC Public Health*. 2019;19(1):1092. doi:10.1186/s12889-019-7419-4
12. Krogsbøll LT, Jørgensen KJ, Grønhøj Larsen C, Gøtzsche PC. General health checks in adults for reducing morbidity and mortality from disease: Cochrane systematic review and meta-analysis. *BMJ*. 2012;345(7884). doi:10.1136/bmj.e7191
13. Kypridemos C, Allen K, Hickey GL, et al. Cardiovascular screening to reduce the burden from cardiovascular disease: Microsimulation study to quantify policy options. *BMJ*. 2016;353. doi:10.1136/bmj.i2793
14. Si S, Moss JR, Sullivan TR, Newton SS, Stocks NP. Effectiveness of general practice-based health checks: A systematic review and meta-analysis. *Br J Gen Pract*. 2014;64(618):e47-e53. doi:10.3399/bjgp14X676456
15. Capewell S, McCartney M, Holland W. Invited debate: NHS Health Checks--a naked emperor? *J Public Health (Bangkok)*. 2015;37(2):187-192. doi:10.1093/pubmed/fdv063
16. Forster AS, Burgess C, Dodhia H, et al. Do health checks improve risk factor detection in primary care? Matched cohort study using electronic health records. *J Public Health (Bangkok)*. 2016;38(3):552-559. doi:10.1093/pubmed/fdv119
17. Robson J, Dostal I, Madurasinghe V, et al. NHS Health Check comorbidity and management: An observational matched study in primary care. *Br J Gen Pract*. 2017;67(655):e86-e93. doi:10.3399/bjgp16X688837
18. National Institute for Health and Care Excellence (NICE). Lipid modification: Cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease (clinical guideline CG181). Published online 2014.
19. Abraham G, Havulinna AS, Bhalala OG, et al. Genomic prediction of coronary heart disease. *Eur Heart J*. 2016;37(43):3267-3278. doi:10.1093/eurheartj/ehw450
20. Inouye M, Abraham G, Nelson CP, et al. Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. *J Am Coll Cardiol*. 2018;72(16):1883-1893. doi:10.1016/j.jacc.2018.07.079
21. Sun L, Pennells L, Kaptoge S, et al. Polygenic risk scores in cardiovascular risk prediction: A cohort study and modelling analyses. Hindy G, ed. *PLOS Med*.

2021;18(1):e1003498. doi:10.1371/journal.pmed.1003498

22. Paige E, Barrett J, Pennells L, et al. Use of Repeated Blood Pressure and Cholesterol Measurements to Improve Cardiovascular Disease Risk Prediction: An Individual-Participant-Data Meta-Analysis. *Am J Epidemiol*. 2017;186(8):899. doi:10.1093/aje/kwx149
23. Paige E, Barrett J, Stevens D, et al. Landmark Models for Optimizing the Use of Repeated Measurements of Risk Factors in Electronic Health Records to Predict Future Disease Risk. *Am J Epidemiol*. 2018;187(7):1530-1538. doi:10.1093/AJE/KWY018
24. Arruda-Olson AM, Afzal N, Priya Mallipeddi V, et al. Leveraging the Electronic Health Record to Create an Automated Real-Time Prognostic Tool for Peripheral Arterial Disease. *J Am Heart Assoc*. 2018;7(23):e009680. doi:10.1161/JAHA.118.009680
25. Barrett JK, Sweeting MJ, Wood AM. Dynamic Risk Prediction for Cardiovascular Disease: An Illustration Using the ARIC Study. In: *Handbook of Statistics*. Vol 36. Elsevier B.V.; 2017:47-65. doi:10.1016/bs.host.2017.05.004
26. Sweeting MJ, Barrett JK, Thompson SG, Wood AM. The use of repeated blood pressure measures for cardiovascular risk prediction: a comparison of statistical models in the ARIC study. *Stat Med*. 2017;36(28):4514-4528. doi:10.1002/sim.7144
27. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203-209. doi:10.1038/S41586-018-0579-Z
28. Herrett E, Gallagher AM, Bhaskaran K, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol*. 2015;44(3):827-836. doi:10.1093/ije/dyv098
29. Abraham G, Malik R, Yonova-Doing E, et al. Genomic risk score offers predictive performance comparable to clinical risk factors for ischaemic stroke. *Nat Commun 2019 101*. 2019;10(1):1-10. doi:10.1038/s41467-019-13848-1
30. Schafer JL. Analysis of Incomplete Multivariate Data, 1st edition , New York: Chapman and Hall/CRC. Chapman and Hall/CRC, ed. Published online 1997:444.
31. Wood AM, Thompson SG, Kostis JB, et al. Correcting for multivariate measurement error by regression calibration in meta-analyses of epidemiological studies. *Stat Med*. 2009;28(7):1067-1092. doi:10.1002/SIM.3530
32. White I, Frost C, Tokunaga S. Correcting for measurement error in binary and continuous variables using replicates. *Stat Med*. 2001;20(22):3441-3457. doi:10.1002/SIM.908
33. Pennells L, Kaptoge S, Wood A, et al. Equalization of four cardiovascular risk algorithms after systematic recalibration: Individual-participant meta-analysis of 86

prospective studies. *Eur Heart J*. 2019;40(7):621-631. doi:10.1093/eurheartj/ehy653

34. Kaptoge S, Pennells L, De Bacquer D, et al. World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions. *Lancet Glob Heal*. 2019;7(10):e1332-e1345. doi:10.1016/S2214-109X(19)30318-3
35. Office for National Statistics. Estimates of the population for the UK, England and Wales, Scotland and Northern Ireland. Accessed January 17, 2022. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland
36. Mihaylova B, Emberson J, Blackwell L, et al. The effects of lowering LDL cholesterol with statin therapy in people at low risk of vascular disease: Meta-analysis of individual data from 27 randomised trials. *Lancet*. 2012;380(9841):581-590. doi:10.1016/S0140-6736(12)60367-5/ATTACHMENT/8DC81FE3-D592-47AE-BD9D-E06A7C9E9817/MMC1.PDF
37. Cook NR, Ridker P. Further insight into the cardiovascular risk calculator: the roles of statins, revascularizations, and underascertainment in the Women's Health Study. *JAMA Intern Med*. 2014;174(12):1964-1971. doi:10.1001/JAMAINTERNMED.2014.5336
38. Patel R, Barnard S, Thompson K, et al. Evaluation of the uptake and delivery of the NHS Health Check programme in England, using primary care data from 9.5 million people: A cross-sectional study. *BMJ Open*. 2020;10(11):42963. doi:10.1136/bmjopen-2020-042963
39. Martin A, Saunders CL, Harte E, et al. Delivery and impact of the NHS Health Check in the first 8 years: A systematic review. *Br J Gen Pract*. 2018;68(672):e449-e459. doi:10.3399/bjgp18X697649
40. Riveros-Mckay F, Weale ME, Moore R, et al. Integrated Polygenic Tool Substantially Enhances Coronary Artery Disease Prediction. *Circ Genomic Precis Med*. 2021;14(2):E003304. doi:10.1161/CIRCGEN.120.003304
41. Crossan C, Lord J, Ryan R, Nherera L, Marshall T. Cost effectiveness of case-finding strategies for primary prevention of cardiovascular disease: A modelling study. *Br J Gen Pract*. 2017;67(654):e67-e77. doi:10.3399/bjgp16X687973
42. Padilla-Martínez F, Collin F, Kwasniewski M, Kretowski A. Systematic Review of Polygenic Risk Scores for Type 1 and Type 2 Diabetes. *Int J Mol Sci*. 2020;21(5). doi:10.3390/IJMS21051703
43. Najar J, van der Lee SJ, Joas E, et al. Polygenic risk scores for Alzheimer's disease are related to dementia risk in APOE ε4 negatives. *Alzheimer's Dement Diagnosis, Assess*

Dis Monit. 2021;13(1). doi:10.1002/DAD2.12142

44. Chaudhury S, Brookes KJ, Patel T, et al. Alzheimer's disease polygenic risk score as a predictor of conversion from mild-cognitive impairment. *Transl Psychiatry 2019 91*. 2019;9(1):1-7. doi:10.1038/s41398-019-0485-7
45. Leonenko G, Baker E, Stevenson-Hoare J, et al. Identifying individuals with high risk of Alzheimer's disease using polygenic risk scores. *Nat Commun 2021 121*. 2021;12(1):1-10. doi:10.1038/s41467-021-24082-z
46. Hirohama D, Susztak K. From mapping kidney function to mechanism and prediction. *Nat Rev Nephrol 2021 182*. 2021;18(2):76-77. doi:10.1038/s41581-021-00512-5
47. Wu H, Forgetta V, Zhou S, Bhatnagar SR, Paré G, Richards JB. Polygenic Risk Score for Low-Density Lipoprotein Cholesterol Is Associated With Risk of Ischemic Heart Disease and Enriches for Individuals With Familial Hypercholesterolemia. *Circ Genomic Precis Med*. 2021;14(1):E003106. doi:10.1161/CIRCGEN.120.003106
48. Vaura F, Kauko A, Suvila K, et al. Polygenic Risk Scores Predict Hypertension Onset and Cardiovascular Risk. *Hypertension*. 2021;77(4):1119-1127. doi:10.1161/HYPERTENSIONAHA.120.16471
49. Parcha V, Pampana A, Shetty NS, et al. Association of a Multiancestry Genome-Wide Blood Pressure Polygenic Risk Score With Adverse Cardiovascular Events. *Circ Genomic Precis Med*. 2022;15(6):E003946. doi:10.1161/CIRCGEN.122.003946
50. Meisner A, Kundu P, Zhang YD, et al. Combined Utility of 25 Disease and Risk Factor Polygenic Risk Scores for Stratifying Risk of All-Cause Mortality. *Am J Hum Genet*. 2020;107(3):418-431. doi:10.1016/J.AJHG.2020.07.002
51. Mishra A, Malik R, Hachiya T, et al. Stroke genetics informs drug discovery and risk prediction across ancestries. *Nat 2022*. Published online September 30, 2022:1-15. doi:10.1038/s41586-022-05165-3
52. Natarajan P, Young R, Stitziel NO, et al. Polygenic Risk Score Identifies Subgroup With Higher Burden of Atherosclerosis and Greater Relative Benefit From Statin Therapy in the Primary Prevention Setting. *Circulation*. 2017;135(22):2091-2101. doi:10.1161/CIRCULATIONAHA.116.024436
53. Widén E, Junna N, Ruotsalainen S, et al. How Communicating Polygenic and Clinical Risk for Atherosclerotic Cardiovascular Disease Impacts Health Behavior: an Observational Follow-up Study. *Circ Genomic Precis Med*. 2022;15(2):E003459. doi:10.1161/CIRCGEN.121.003459

Chapter 5

The lifetime population health impact of utilising polygenic risk scores for determining the age at which to make formal cardiovascular risk assessments

Chapter summary

Objective: Expanding on the results found in **Chapter 4**, this chapter aims to investigate how polygenic risk scores (PRS) could be used to determine the optimal age at which to invite an individual for a formal cardiovascular disease (CVD) risk assessment, and estimate the lifetime benefits of such an approach.

Methods: 300,088 participants, aged 40-69 with measured biomarkers, genetic data and without a history of CVD, diabetes and lipid lowering medication in UK Biobank were used to derive three risk models. Sex-specific Cox models, using a combination of conventional CVD risk factors, and PRS for coronary artery disease and stroke, were used to inform both the invitation strategies and the formal CVD risk assessment. We used PRS to personalise invitations to a formal CVD risk assessment to begin between the ages of 25 and 74 years, and compared this to current clinical guideline recommendations for population-wide invitations to begin at age 40. We also investigated using only genetically predicted risk factor levels to guide treatment decisions and remove the formal risk assessment process as currently recommended. We modelled the implications of initiating statin therapy according to each strategy using a lifetime-risk population health modelling approach accounting for competing risks.

Results: Compared to a population-wide invitation strategy followed by assessment using conventional CVD risk factors and PRS, a strategy using PRS to personalise the age of first invitation prior to an assessment led to a 43% and 39% reduction in the NNS in men and women respectively whilst saving a similar number of events over a lifetime.

A strategy that removes the need for multiple invitations to a formal assessment, using only genetically predicted risk factor levels to guide treatment decisions, however led to an 81% and

92% reduction in the NNS in men and women respectively and saved a greater number of events, especially in women. However, if the risk threshold to determine statin initiation changed from 10% to 5%, a population-wide invitation strategy can be efficient in men but women may still benefit from the personalised invitation strategy.

Conclusion: By extending the work from **Chapter 4**, we have improved the population health modelling by estimating the lifetime impact of statins. We have created strategies that implement PRS to personalise the first age of invitation. Our results suggest that PRS can be effectively used to personalise the invitation process by inviting high-risk individuals earlier, and low-risk individuals later. These results highlight the future potential of PRS and its implementation in the healthcare system.

5.1 Introduction

Complementing existing cardiovascular disease (CVD) risk scores with PRS has been seen as a priority for the healthcare system, as stated by the UK government's Department of Health and Social Care green paper on disease prevention (see **Chapter 1, Section 1.4.2**).[1] Previous research into PRS has predominantly focussed on the improved discrimination and stratification when incorporating disease based PRS (e.g., coronary artery disease and stroke) into a risk tool, with some research demonstrating its population health utility and its clinical implications (see **Chapter 4, Section 4.1)**.[2–5] However, quantitative evidence investigating the lifetime population health impact is limited. As PRS is fixed from birth, the PRS presents a unique opportunity to personalise risk assessment programmes to individuals, by prioritising those with a high genetic risk for a risk assessment at an earlier age. Consequently, the opportunity to identify younger individuals at high-risk of CVD may be able to assist in reducing future CVD burden.[6,7]

Therefore, to add to the growing evidence base of the benefits implementing PRS can have for CVD risk assessments, we aim to evaluate how using PRS to create a strategy that personalises the first age of invitation prior to a CVD risk assessment, could potentially impact the number of invitations needed prior to treatment and the number of events saved across a lifetime in the general population.

The research in **Chapter 5** expands on the work shown in **Chapter 4**. In **Chapter 4**, population health modelling was restricted to a single time point and for individuals between the ages of 40 to 69 years, as measurements recorded at baseline were used to simulate the formal risk assessment. This limits the understanding of PRS in individuals younger than 40 years old, where younger individuals with a greater genetic risk may benefit more from earlier intervention. This also limits the long-term view of a formal assessment programme; whilst **Chapter 4** presents a snapshot of a formal assessment programme, the long-term population health impact of repeated formal assessments over a lifetime could not evaluated. **Chapter 5** will address this limitation by using the linked primary care records to estimate the expected risk factor levels across an individual's lifetime, from the ages of 25 to 75 years.

With research on PRS continuing to advance, recent research has provided evidence of the predictive utility of using PRS to predict conventional CVD risk factors (i.e. genetically predicted risk factor levels), including SBP, low-density lipoprotein (LDL) cholesterol and high-density lipoprotein (HDL) cholesterol.[8–11] As such, **Chapter 5** will investigate the utility of risk-factor based PRS in addition to the CVD-based PRS previously used (see **Chapter 4)**.

Unlike **Chapter 4,** where age- and sex-specific prioritisation thresholds were investigated, **Chapter 5** will present findings under a framework of existing clinical guidelines, and use the fixed 10% CVD risk threshold currently recommended for prioritisation and statin initiation. This was chosen to reflect the challenges in updating key aspects of the guidelines, and allows us to focus primarily on how personalising the age at which to invite people to CVD risk assessments using polygenic risk scores impacts the lifetime population health.

5.2 Methods

5.2.1 Data sources

As with **Chapter 4,** data from UK Biobank (UKB), a prospective cohort study with detailed baseline information, genetic data and with linked information on hospital episodes from Hospital Episode Statistics (HES), and death registrations from the Office of National Statistics (ONS) were used for model derivation and population health modelling. Unlike the data used in **Chapters 3 and 4**, we do not rely on primary care records for model derivation. We therefore used a larger subset of UKB individuals, aged 40-69, with measured biomarkers and genetic for model derivation.

For the population health modelling, we used a subset of UKB individuals with linked primary care records. Primary care records were used to model the implications of using a PRS-based risk model to personalise the first age of invitation on the number of events saved up to the age of 75 in the general population. Primary care records after the 1st April 2004 (the date of introduction of the Quality and Outcomes Framework) were used in the analysis.

To translate the findings to a more representative population in the United Kingdom, average CVD incidence rates from the Clinical Practice Research Datalink (CPRD) were used to recalibrate risk models using methods previously described (see **Chapter 1, Section 1.5.2.2**).

5.2.2 Outcomes and risk factors

We ascertained individuals with a first ever incident CVD as those with a relevant Read-code or ICD-10 code appearing in hospital episodes or death registry (underlying or contributing cause of death) during follow-up **(Table 5.1)**.

Conventional CVD risk predictors recorded at UKB's baseline survey were included in the risk model: age (in years), sex (men or women), SBP (mmHg), LDL cholesterol (mmol/litre), HDL cholesterol and smoking status (current smoker or not).

In addition, PRS for coronary artery disease (CAD), stroke, and risk factor levels of LDL, HDL and SBP were constructed. All PRS were constructed using a meta-score approach and external summary statistics from large genome wide association studies. The PRS for CAD and stroke used were previously constructed using external summary statistics from large genome wide association studies and a meta-score approach consisting of 1.7 million and 3.2 million variants respectively.[2,4] The LDL, HDL and SBP PRS were constructed using non-UKB participants from previously published GWAS. The variants were filtered to a set of 2.3 million LD-thinned ($r^2 < 0.9$) variants present in UKB. The LDL and HDL PRS consisted of 515,000 variants, and the SBP PRS consisted of 2.1 million variants.[12,13]

Table 5.1: Code list used to define cardiovascular disease.

Endpoint	ICD-10 code
Vascular dementia	F01
Angina pectoris	I20
Acute myocardial infarction	I21
Subsequent myocardial infarction	I22
Complications after myocardial infarction	I23
Other acute ischaemic heart disease	I24
Chronic ischaemic heart disease	I25
Cerebral infarction	I63
Stroke not specified as haemorrhage or infarction	I64
occlusion and stenosis of precerebral arteries, not resulting in cerebral infarction	I65
Occlusion and stenosis of cerebral arteries, not resulting in cerebral infarction	I66
other cerebrovascular diseases	I67
Cerebrovascular disorders in diseases classified elsewhere	I68
Sequelae of cerebrovascular disease	I69

Cardiovascular disease was defined as a combination of newly diagnoses of nonfatal or fatal events of coronary heart disease (CHD) (including myocardial infarction and angina) and stroke. Diagnoses are coded using the linked HES and ONS datasets where the International Classification of Disease 10th revision (ICD-10) codes were used.[14]

5.2.3 Statistical methods to derive CVD risk assessment models

We derived three CVD risk assessment models using sex-specific Cox regression, with follow-up since study entry as the underlying time scale to estimate a 10-year CVD risk. The outcome was incident CVD and censoring occurred for individuals at end of their follow-up, or date of death (from non-CVD causes). Models were derived in 300,088 individuals with no prior history of CVD, no history of diabetes and no history of lipid lowering medication at UKB's baseline.

The three CVD risk models derived were:

1) a 'conventional risk model', included measures of conventional CVD risk factors: age, SBP, LDL cholesterol, HDL cholesterol and smoking status.

2) a 'conventional + PRS risk model', included the same CVD risk factors as the 'conventional risk model' and also included a CAD PRS and stroke PRS.

3) a ‘PRS based risk model’ which including the same risk factors as the ‘conventional + PRS risk model’, but replaces the SBP, LDL cholesterol and HDL cholesterol with sex-specific genetically predicted values using PRS for SBP, LDL and HDL. The age- and sex-specific genetically predicted values were estimated from sex specific linear models, derived by regressing the observed, cross-sectional SBP, LDL and HDL values in UKB against their respective PRS, adjusted for baseline age and BMI, with the predicted values assuming mean levels of BMI.

Prognostic ability of models in the derivation sample was quantified using Harrell’s C-index to measure discrimination. The estimated 10-year risks from each model were rescaled to sex-specific CPRD incidence rates in order to accurately estimate the number of individuals that were deemed at high risk in the general population (see **Chapter 1**, **Section 1.5.2.2**).[15]

5.2.4 Assessment of potential clinical impact

5.2.4.1 Defining invitation and treatment strategies

We devised four strategies for determining the first age of invitation, followed by a formal risk assessment at which treatment was allocated if the individual had a predicted risk above the threshold (i.e., was deemed at high risk) **(Table 5.2).** Each was chosen to represent increasing levels of PRS implementation.

Strategy 1: Population-wide invitation

The first strategy follows the recommendations of NICE guidelines in England, with all individuals invited for a formal assessment at age 40 years (population-wide invitation) and treatment offered if CVD risk ≥10% at the formal assessment.[16] The ‘conventional risk model’ was used to estimate the 10-year risk at formal assessment using measured risk factor levels. Statins were offered if the 10-year formal assessment risk exceeds 10%. Otherwise, the individual was invited for another assessment five years later.

Strategy 2: Population-wide invitation enhanced with PRS

Similar to the first strategy, the second strategy follows the population-wide invitation strategy. However, risk assessments were conducted using measured risk factor levels and the 'conventional + PRS risk model'. Similar to the first strategy, statins were offered if the 10-year formal assessment risk exceeds 10% or invited 5 years later otherwise.

Strategy 3: Personalised invitation using PRS

Unlike the first two strategies, the third strategy personalises the first age of invitation prior to a formal risk assessment. The first age of invitation was determined using the 'conventional + PRS risk model'. To estimate the projected 10-year CVD risk, in addition to an individual's PRS for CAD and stroke, we used fixed "least favourable LDL, HDL and SBP levels" in the broader age and sex specific population, and fixed age- and sex-specific mean smoking prevalence. We defined the least favourable risk factor values as the 90th percentile of observed, cross-sectional values observed at UKB baseline survey by age and sex **(Table 5.3).** We summarised the risk factor levels over a 4 year window (2 years before and after the chosen age); for example, the SBP value for a 60-year old man was summarised from SBP levels in men who attended the UKB baseline survey visit between 58 years and 62 years. Risk factor values for those aged between 25-41 and 69-74 (ages at which a 4-year window in UKB could not be achieved) used the risk factor values obtained for those aged at 42 and 68 respectively.

An individual was first invited at the age at which their projected 10-year CVD risk exceeded 10%. By using "least favourable risk factor levels", most individuals would have an over-estimated risk to ensure the early identification of high-risk individuals **(Figure 5.1)**. This approach was designed under a framework of existing clinical guidelines, and was an alternative to using age- and sex specific thresholds (see **Chapter 4)**.

Afterwards, a formal risk assessment was conducted using measured risk factor levels and the 'conventional + PRS risk model'. Statins were offered if the 10-year formal assessment risk exceeds 10% or invited 5 years later otherwise. For this strategy, the first age of invitation ranged from 25 to 55 years, where all individuals yet to be invited for a formal risk assessment were invited at 55 years.

Strategy 4: Treatment with genetically predicted risk factors

A fourth strategy was created to use only genetically predicted risk factor levels. As the PRS has potential to estimate an individual's lifetime risk of CVD, this strategy was devised to represent a possible scenario where only genetic data could be used to determine treatment allocation.[17,18] By using the 'PRS based risk model', individuals were offered statins at the age when their genetically predicted risk exceeds 10%. In addition, the strategy eliminated the assumption of multiple invitations every five years as used in the previous three strategies.

Table 5.2: Matrix of strategies defined to determine the first age of invitation followed by a formal CVD risk assessment to at which treatment was allocated if the individual was at high risk

	Strategy 1: Population-wide invitation	Strategy 2: Population-wide invitation enhanced with PRS	Strategy 3: Personalised invitation using PRS	Strategy 4: Treatment with genetically predicted risk factors
	Population-wide invitation		**Personalised invitation**	
Invitation process	Invite all individuals for formal CVD assessment from age 40. Invite every subsequent 5 years.		First age of invitation (25-74 years*) determined using conventional CVD risk factors, CAD PRS and stroke PRS. Invite at age when estimated risk exceeds 10% and invite every subsequent 5 years.	
Invitation risk factors used	N/A		Age- and sex-specific least favourable population averages conventional CVD risk factors + individual-level CAD and stroke PRS	Genetically predicted values for CVD risk factors + individual-level CAD and stroke PRS
Formal CVD risk assessment	10-year CVD risk estimated using conventional CVD risk factors	10-year CVD risk estimated using conventional CVD risk factors, CAD PRS and stroke PRS		N/A
Statin initiation	Offer lipid lowering medication if 10-year CVD risk greater than 10% (current guidelines)			Offer lipid lowering medication when genetically predicted 10-year CVD risk exceeds 10%.

Abbreviations: CAD, coronary artery disease; CVD, cardiovascular disease; PRS, polygenic risk score

* For Strategy 3, the first age of invitation was from 25 to 55 years, where all individuals yet to be invited for a formal risk assessment would be invited automatically at 55 years.

Table 5.3: Summary risk factor levels from UK Biobank used in the strategy "personalised invitation using PRS"

	Men				Women			
Age	**SBP (mmHg)**	**LDL cholesterol (mmol/L)**	**HDL cholesterol (mmol/L)**	**Current smoker (%)**	**SBP (mmHg)**	**LDL cholesterol (mmol/L)**	**HDL cholesterol (mmol/L)**	**Current smoker (%)**
25	141	4.93	0.94	16.1	141	4.44	1.12	12.4
26	141	4.93	0.94	16.1	141	4.44	1.12	12.4
27	141	4.93	0.94	16.1	141	4.44	1.12	12.4
28	141	4.93	0.94	16.1	141	4.44	1.12	12.4
29	141	4.93	0.94	16.1	141	4.44	1.12	12.4
30	141	4.93	0.94	16.1	141	4.44	1.12	12.4
31	141	4.93	0.94	16.1	141	4.44	1.12	12.4
32	141	4.93	0.94	16.1	141	4.44	1.12	12.4
33	141	4.93	0.94	16.1	141	4.44	1.12	12.4
34	141	4.93	0.94	16.1	141	4.44	1.12	12.4
35	141	4.93	0.94	16.1	141	4.44	1.12	12.4
36	141	4.93	0.94	16.1	141	4.44	1.12	12.4
37	141	4.93	0.94	16.1	141	4.44	1.12	12.4
38	141	4.93	0.94	16.1	141	4.44	1.12	12.4
39	141	4.93	0.94	16.1	141	4.44	1.12	12.4
40	141	4.93	0.94	16.1	141	4.44	1.12	12.4
41	141	4.93	0.94	16.1	141	4.44	1.12	12.4
42	141	4.93	0.94	16.1	141	4.44	1.12	12.4
43	142	4.97	0.94	15.8	142	4.45	1.13	12.0
44	143	4.99	0.95	15.7	143	4.49	1.13	11.4
45	144	4.99	0.95	15.4	144	4.54	1.14	11.4
46	146	4.99	0.95	15.2	146	4.59	1.14	11.3
47	147	5.00	0.95	14.7	147	4.66	1.15	11.2
48	148	5.01	0.96	14.1	148	4.74	1.17	11.0
49	150	5.03	0.96	13.6	150	4.80	1.17	10.8
50	151	5.02	0.96	13.5	151	4.86	1.18	10.4

51	153	5.01	0.96	13.3	153	4.92	1.18	10.2
52	154	5.04	0.96	13.1	154	4.97	1.19	10.2
53	155	5.03	0.97	12.9	155	5.04	1.20	9.8
54	156	5.05	0.97	12.5	156	5.10	1.20	9.6
55	157	5.06	0.97	12.2	157	5.15	1.21	9.4
56	158	5.05	0.97	11.9	158	5.20	1.21	8.9
57	159	5.03	0.97	11.5	159	5.23	1.21	8.6
58	160	4.99	0.98	11.2	160	5.24	1.21	8.2
59	161	5.00	0.98	11.0	161	5.27	1.21	7.8
60	163	4.97	0.98	10.7	163	5.29	1.21	7.5
61	164	4.96	0.98	10.5	164	5.32	1.21	7.2
62	165	4.97	0.98	10.3	165	5.33	1.20	6.9
63	166	4.95	0.98	9.9	166	5.33	1.21	6.6
64	167	4.94	0.98	9.7	167	5.32	1.21	6.6
65	168	4.92	0.98	9.4	168	5.31	1.21	6.5
66	169	4.87	0.98	9.1	169	5.29	1.21	6.4
67	171	4.85	0.98	8.8	171	5.30	1.21	6.1
68	172	4.83	0.98	8.4	172	5.30	1.20	5.6
69	172	4.83	0.98	8.4	172	5.30	1.20	5.6
70	172	4.83	0.98	8.4	172	5.30	1.20	5.6
71	172	4.83	0.98	8.4	172	5.30	1.20	5.6
72	172	4.83	0.98	8.4	172	5.30	1.20	5.6
73	172	4.83	0.98	8.4	172	5.30	1.20	5.6
74	172	4.83	0.98	8.4	172	5.30	1.20	5.6

Abbreviations: HDL cholesterol, high-density lipoprotein cholesterol; LDL cholesterol, low-density lipoprotein cholesterol; PRS, polygenic risk score; SBP, systolic blood pressure.

Least favourable SBP, LDL and HDL levels defined as the 90th percentile of observed, cross-sectional values in UK Biobank's baseline survey, by age and sex. Current smoking status was defined by age- and sex-specific mean prevalence at baseline. Risk factor levels were summarised over a 4 year window (2 years before and after the chosen age);. Risk factor values for those aged between 25-41 and 69-74 used the risk factor values obtained for those aged at 42 and 68 respectively.

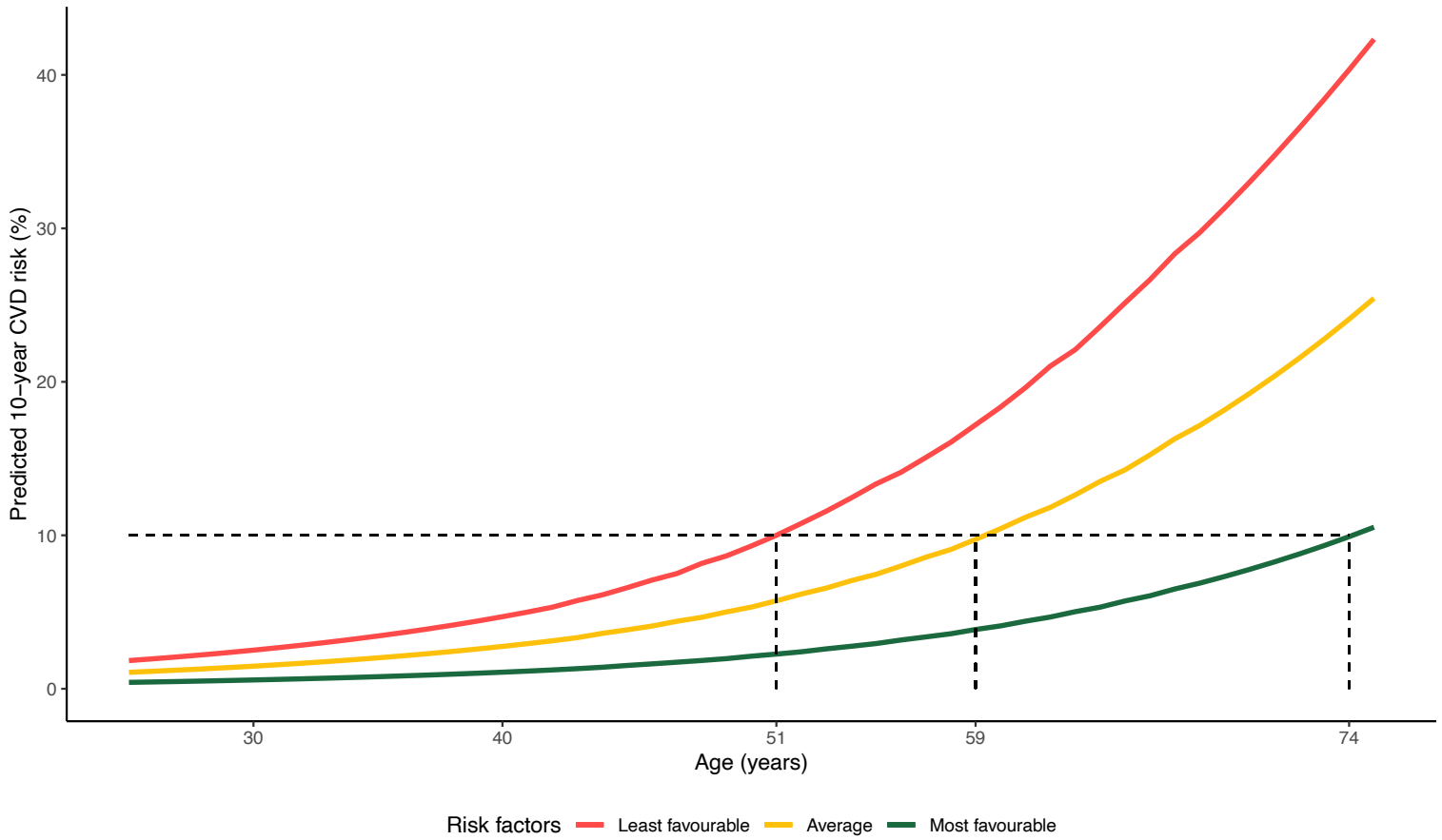


Figure 5.1: Example of an individual's projected 10-year risk to determine first age of invitation using the individual's CAD and stroke PRS values along with fixed risk factor levels for blood pressure, LDL and HDL cholesterol and smoking status in the broader age and sex specific population in UK Biobank.

Abbreviations: CAD, coronary artery disease; CVD, cardiovascular disease; HDL, high density lipoprotein; LDL, low density lipoprotein; PRS, polygenic risk score

5.2.4.2 Population health modelling

A subset of individuals in UKB with linked primary care records before and after the study baseline was used to estimate the potential benefit of applying each invitation and treatment strategy in terms of the number of events identified and prevented by the age of 75. The analysis involved four steps:

Step 1: Determine first age of invitation for each strategy

For strategies 1 and 2, the population-wide invitation strategy was used and the first age of invitation age was 40 years for all individuals. For strategy 3, the first age of invitation (25-55 years) was determined using conventional CVD risk factors, CAD PRS and stroke PRS. The first age of invitation was determined as when the estimated risk exceeds 10%. All individuals yet to be invited before 55 years were automatically invited at 55 years. For strategy 4, the first age of invitation was determined at the age which their genetically predicted risk exceeds 10%**.**

Step 2: Estimate 10-year CVD formal assessment risks

Unlike in previous analysis where the risk factor measurements at UKB's baseline survey was used to represent the formal risk assessment at a single time point (see **Chapters 3 and 4**), the work in **Chapter 5** was complicated by the lack of detailed baseline measurements at every potential age of invitation. As such, we used the linked primary care records in UKB to estimate individual-level risk factor levels at all ages. By using methods previously discussed (see **Chapter 4),** we fit sex-specific multivariate mixed-effects model on individual-level longitudinal risk factor measurements in the linked primary care records to estimate individual-level expected risk factor levels for SBP, LDL and HDL cholesterol and smoking prevalence extrapolated for ages 25-74. Backwards extrapolation was performed for risk factor data between the ages of 25-35 years due to the limited primary care records before UKB baseline survey visit. The expected risk factor levels from the mixed-effects model were then recalibrated to correct for the differences observed between the primary care records and the baseline measurements (see **Chapter 2**). We used the same methods previously described (see **Chapter 1, Section 1.5.2.2**) however in this case, we regressed the recorded baseline measurements as the reference values against the expected levels at the age at which the individual attended the UKB baseline-survey.

For strategies 1-3, the expected risk factor levels at the time of the invitation age were used to calculate a 10-year formal CVD risk estimated at the formal risk assessment. If the 10-year

formal CVD risk was greater than 10%, the individual was treated. If the 10-year formal CVD risk was lower than 10%, the individual was deferred and invited for another assessment 5 years later. The process was then repeated until the individual was either treated with statins or they reached the age of 75. For strategy 4, the individual was treated upon their first age of invitation and was not invited for another assessment.

Step 3: Estimate events saved due to statin initiation in a representative population

We grouped individuals by their statin-initiation age in 5-year intervals (25-29, 30-34,…, 70-74) for each of the proposed strategies. We estimated the potential public health impact of statin initiation in terms of the number of CVD events saved by the age of 75 accounting for non-CVD deaths. For each strategy, in each age-group, we estimated the cumulative risk of a CVD event by the age of 75, using a cause-specific hazards framework to adjust for non-CVD deaths. To ensure the survival estimates are representative of the general population, and to account for UKB being a healthier cohort with a lower CVD incidence rate than the general population, during the population health modelling we upweighted UKB individuals with a CVD event with weights equal to the sex- and age-group specific ratio of CVD incidence in CPRD to UKB. The same upweighting method was conducted for non-CVD deaths.

The number of CVD events saved by the age of 75 in each age group was calculated using the difference in cumulative incidence rates with and without statin initiation, assuming a 20% reduction in CVD risk upon statin initiation.

Step 4: Summarise population health metrics

Similar to **Chapters 3 and 4**, the number needed to screen (NNS) was estimated. However, in **Chapter 5** this was defined as the total number of invitations, across all individuals' lifetime, needed to save one event due to statin initiation. In addition, we estimated the number needed to treat (NNT), defined as the total number of risk assessments needed to save one event over a lifetime. The NNT was used to better understand differences between the first three strategies, which rely on a separate formal risk assessment, and the fourth strategy, which treats upon invitation.

We assumed all invited individuals attend a formal assessment, i.e., 100% invitation compliance. We also assumed 50% compliance for those prescribed statins.[19,20] We performed analyses by assuming 100% statin compliance, and applied a crude adjustment by halving the number of events saved, which consequently doubles the NNS and NNT. We then standardised

the results based on the age distribution of a population of 100,000 individuals (50,000 men and women) from the United Kingdom.

5.2.4.3 Sensitivity analysis

Population health modelling was conducted assuming a 5% CVD risk assessment threshold (see **Chapter 1**, **Section 1.3.3)**. This replaces the fixed 10% risk threshold used to determine both the first age of invitation and whether statins were offered.

5.3 Results

5.3.1 Study population and baseline characteristics in UK Biobank

For the model derivation, of the initial 502,536 individuals in UK Biobank, we identified a subset of 300,088 individuals without prior CVD, diabetes or statins at baseline and with complete data on smoking status, SBP, LDL cholesterol, HDL cholesterol and genetic data required to calculate the CAD and stroke PRS, as well as polygenic risk scores for SBP, LDL and HDL levels **(Figure 5.2).** For the population health modelling, we identified a subset of 138,317 individuals without diabetes and who had the necessary primary care records required for the mixed-effects model. **(Figure 5.3).**

The baseline characteristics of individuals **(Table 5.4)** show 43% of the derivation cohort were men. 3,568 men and 1,877 women had an incident CVD event over a median follow up period of 8.1 years (IQR: 6.7, 9.3) and 8.1 years (IQR: 6.8, 9.3) respectively.

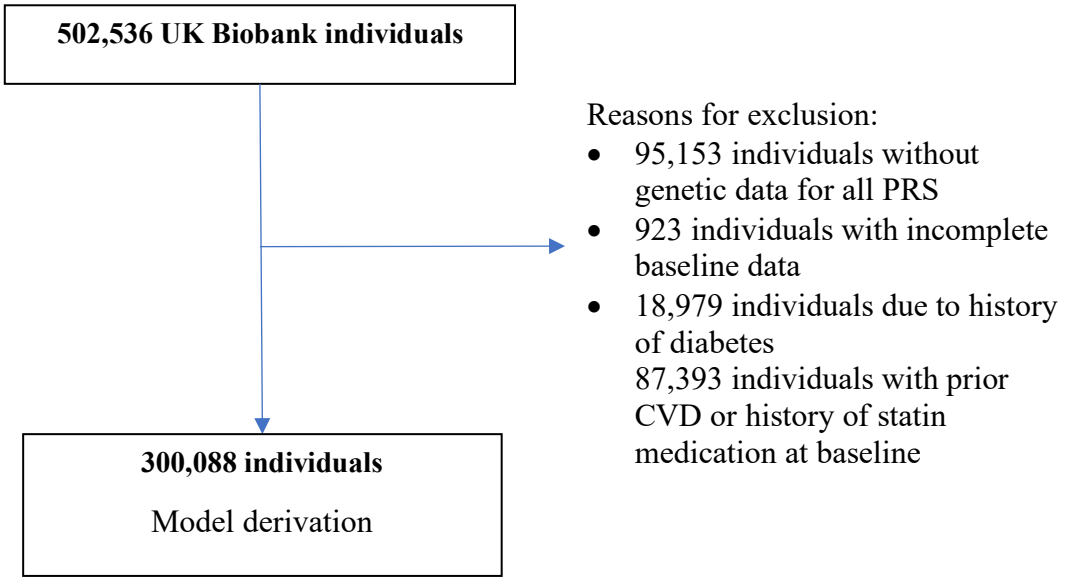


Figure 5.2: Flowchart of individuals included for model derivation in UK Biobank

Abbreviations: CVD, cardiovascular disease; PRS, polygenic risk score

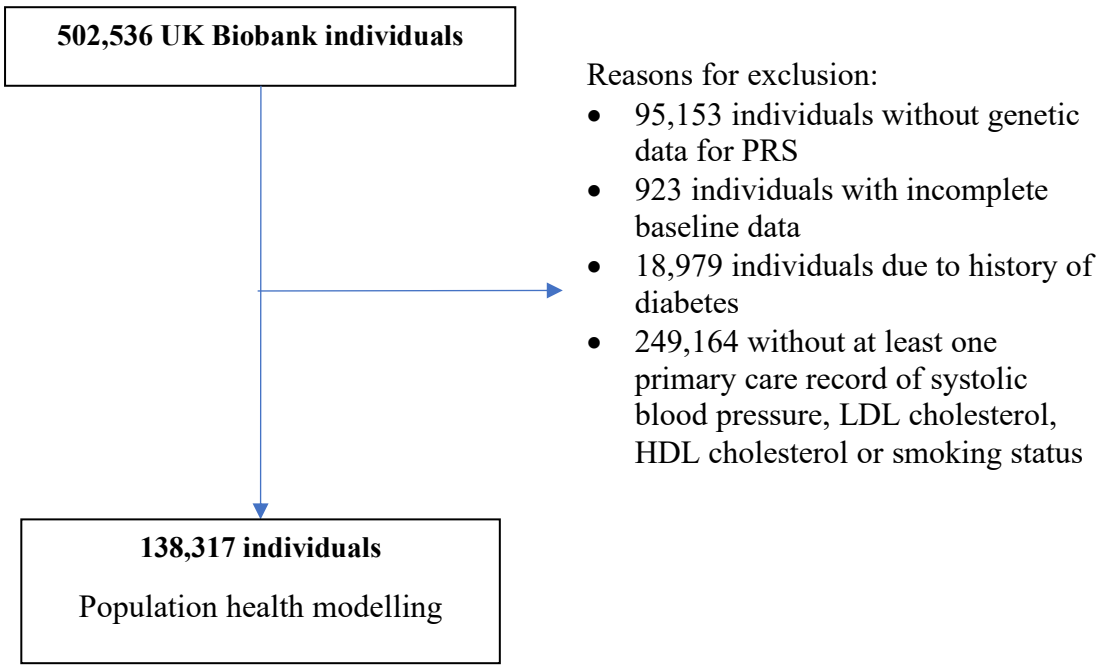


Figure 5.3: Flowchart of individuals included for population health modelling in UK Biobank

Abbreviations: HDL, high density lipoprotein; LDL, low density lipoprotein; PRS, polygenic risk score

Table 5.4: Key characteristics of individuals in UK Biobank

Characteristic	Men		Women	
	Model derivation	Population health modelling	Model derivation	Population health modelling
Individuals, N	128,322 (43%)	61,994 (45%)	171,766 (57%)	76,323 (55%)
CVD events*, N	3,568	3,407	1,877	1,518
Follow up duration: years, median (5^{th}, 95^{th} percentile)	8.1 (6.7, 9.3)	8.0 (5.9, 9.0)	8.1 (6.8, 9.3)	8.1 (6.8, 9.0)
Age, mean (SD)	56.0 (8.1)	57.6 (8.0)	56.2 (7.9)	57.2 (7.9)
Systolic blood pressure: mmHg, mean (SD)	140.7 (17.3)	141.8 (17.6)	134.6 (19.3)	136.2 (19.3)
LDL cholesterol: mmol/litre, mean (SD)	3.93 (1.18)	3.76 (1.25)	3.95 (1.21)	3.90 (1.25)
HDL cholesterol: mmol/litre, mean (SD)	1.57 (0.76)	1.56 (0.78)	1.87 (0.73)	1.86 (0.74)
Current smoker, N (%)	15,123 (11.8%)	7,241 (11.7%)	14,593 (8.5%)	6,589 (8.6%)
Coronary artery disease PRS, mean (SD)	-1.67 (0.42)	-1.64 (0.42)	-1.66 (0.42)	-1.64 (0.43)
Stroke PRS, mean (SD)	589.6 (54.1)	590.8 (54.4)	590.7 (53.9)	592.0 (54.1)

Abbreviations: CVD, cardiovascular disease; HDL cholesterol, high-density lipoprotein cholesterol; LDL cholesterol, low-density lipoprotein cholesterol; PRS, polygenic risk score; SD, standard deviation.

*For model derivation, CVD events after baseline were used. For population health modelling, all events prior to baseline were also included.

5.3.2 Model performance

All CVD risk assessment models had good discriminatory performance **(Table 5.5)**. The C-index for the standard CVD risk model with conventional CVD risk factors in men (C-index = 0.665, 95% CI: 0.659, 0.671) was lower than in women (C-index = 0.715, 95% CI: 0.706, 0.724). The addition of the CAD and stroke PRS (PRS-based risk model) improved the C-index in both men (C-index = 0.690, 95% CI: 0.684, 0.696) and women (C-index = 0.724, 95% CI: 0.715, 0.733), with a greater improvement in men. Substituting the conventional risk factors with genetically predicted risk factors performed as well as the PRS-based risk model in men (C-index = 0.690, 95% CI: 0.683, 0.696) but was more similar to the standard risk model in women (C-index = 0.717, 95% CI: 0.707, 0.726).

Table 5.5: C-index of risk models derived in UK Biobank

Model	C-index (95% CI)	
	Men	Women
Conventional risk model	0.665 (0.659, 0.671)	0.715 (0.706, 0.724)
Conventional + PRS risk model	0.690 (0.684, 0.696)	0.724 (0.715, 0.733)
PRS based risk model	0.690 (0.683, 0.696)	0.717 (0.707, 0.726)

Abbreviations: PRS, polygenic risk scores

5.3.3 Population health modelling of invitation and treatment strategies

CVD incidence rates in UKB were consistently lower than incidence rates calculated in the more general population in CPRD for both men and women **(Table 5.6).** As such, weights were greater than one for all age groups and increased as age increased. This resulted in the upweighting of events in UKB during the calculation of cumulative incidence. The weights estimated within non-CVD deaths however varied more by age group and were generally smaller than for the weights estimated for CVD events **(Table 5.7)**.

Table 5.6: Age- and sex-specific incidence rates per 1000 person years in UK Biobank and in CPRD used to reweight CVD events in population health modelling.

	Men			Women		
Age group	**UK Biobank CVD incidence rate**	**CPRD CVD incidence rate**	**CVD event weights**	**UK Biobank CVD incidence rate**	**CPRD CVD incidence rate**	**CVD event weights**
40-44	1.040	1.161	1.117	0.279	0.504	1.805
45-49	1.488	1.890	1.270	0.410	0.984	2.403
50-54	2.234	3.415	1.528	0.685	1.240	1.810
55-59	3.092	5.131	1.659	1.061	1.804	1.700
60-64	4.386	6.198	1.413	1.469	2.533	1.725
65-69	6.381	7.733	1.212	2.437	4.285	1.759
70-74	7.896	10.08	1.277	3.894	7.557	1.941

Abbreviations: CPRD, Clinical Practice Research Datalink; CVD, cardiovascular disease

A weight greater than one indicates an event in UK Biobank will be upweighted during the calculation of cumulative incidence.

* Incidence rates in UK Biobank defined using study entry and follow up for incident CVD events. CVD event weights estimated for 40–44-year-olds were substituted for those younger than 40 years in population health modelling.

*CPRD incidence rates was calculated in those without diabetes/prior CVD, for both CVD and non-CVD death in 2016.

Table 5.7: Age- and sex-specific incidence rates in UK Biobank and in CPRD used to reweight non-CVD deaths in population health modelling.

	Men			Women		
Age group	**UK Biobank non-CVD mortality rate per 1,000**	**CPRD non-CVD mortality per 1,000 (2016)**	**Non-CVD mortality weights**	**UK Biobank non-CVD mortality rate per 1,000**	**CPRD non-CVD mortality per 1,000 (2016)**	**Non-CVD mortality weights**
40-44	0.320	0.353	1.104	0.419	0.353	0.842
45-49	0.875	0.779	0.890	0.534	0.726	1.359
50-54	0.917	1.350	1.472	1.101	1.312	1.192
55-59	2.152	1.900	0.883	1.364	1.977	1.449
60-64	2.889	3.934	1.362	1.913	2.175	1.137
65-69	4.678	5.081	1.089	3.004	4.812	1.603
70-74	8.679	9.851	1.135	5.006	7.099	1.418

Abbreviations: CPRD, Clinical Practice Research Datalink; CVD, cardiovascular disease

A weight greater than one indicates an event in UK Biobank will be upweighted during the calculation of cumulative incidence

* Incidence rates in UK Biobank defined using study entry and follow up for non-CVD deaths. Weights estimated for 40–44-year-olds were substituted for those younger than 40 years in population health modelling.

*CPRD incidence rates was calculated in those without diabetes/prior CVD, for both CVD and non-CVD death in 2016.

*Due to data availability in CPRD, non-CVD events were available after 40 years old only. The weights estimated for 40–44-year-olds were used in substitution for younger age groups.

5.3.3.1 Strategy 1: Population-wide invitation

In our representative population of 50,000 men and 50,000 women, following current practice of a population-wide invitation strategy followed by assessment using a risk model with conventional CVD risk predictors, we estimated a cumulative total of 301,704 invitations in men and 332,267 invitations in women between the ages of 40 and 75 **(Table 5.8, Figure 5.4)**. All men and women were invited at the age of 40 (**Figure 5.5)**. 50% of men had initiated treatment by age 65 whilst 12% of women had initiated treatment by the same age **(Figure 5.6)**. The earliest age of treatment was 40 years in men and 45 years in women. We estimated the treatment of 44,229 (88%) men and 29,364 (59%) women resulting in 186 and 102 CVD events saved respectively due to statin initiation. The corresponding NNS were 238 and 3,254 and corresponding NNT were 238 and 288 respectively (**Table 5.8)**.

5.3.3.2 Strategy 2: Population-wide invitation enhanced with PRS

Using the second strategy of using a population-wide invitation strategy followed by assessment using conventional risk factors and CAD and stroke PRS, we estimated a cumulative total of 297,354 invitations in men and 330,155 invitations in women between the ages of 40 and 75 **(Table 5.8, Figure 5.4)**. Whilst all men and women were invited at the age of 40, overall fewer invitations were required as more individuals were treated earlier due to the inclusion of PRS in the risk model (**Figure 5.5)**. 50% of men had initiated treatment by age 65 whilst 7% of women had initiated treatment by the same age **(Figure 5.6)**. We estimated the treatment of 36,753 (74%) men and 27,149 (54%) women resulted in 183 and 94 events saved respectively due to statin initiation. The corresponding NNS were 1,622 and 3,512 and corresponding NNT were 201 and 289 respectively (**Table 5.8)**.

5.3.3.3 Strategy 3: Personalised invitation using PRS

When using the third strategy of a personalised first invitation age by CAD and stroke PRS followed by assessment using conventional risk factors and CAD and stroke PRS, we estimated a cumulative total of 164,084 invitations in men and 186,123 invitations in women between the ages of 25 and 75 **(Table 5.8, Figure 5.4)**. The first invitations were in men aged 42 and in women aged 55 (**Figure 5.5)**. 50% of men had initiated treatment by the age of 70, whilst 36% of women had initiated treatment by the same age **(Figure 5.6)**. We estimated the treatment of 31,845 (64%) men and 18,035 (36%) women resulted in 178 and 87 events saved respectively due to statin initiation. The corresponding NNS were 924 and 2,140 and corresponding NNT were 179 and 207 respectively **(Table 5.8)**. Compared to Strategy 2, the NNS was reduced by 43% in men and 39% in women.

5.3.3.4 Strategy 4: Treatment with genetically predicted risk factors

Using the fourth strategy and treating individuals based off a single invitation with genetically predicted risk factor levels, we estimated a higher total of 40,380 (81%) men and 35,558 (71%) women treated between the ages of 25 and 75 **(Table 5.8, Figure 5.4)**. 50% of men had initiated treatment by age 69 whilst 15% of women had initiated treatment by the same age **(Figure 5.6)**. We estimated that initiation of treatment resulted in 198 and 127 events saved respectively due to statin initiation. The corresponding NNS (and therefore NNT) was 204 for men and 280 for women respectively **(Table 5.8).** Compared to Strategy 2, the NNS was reduced by 87% in men and 92% in women.

Table 5.8: Number needed to screen and treat to prevent one CVD event by the age of 75 in a hypothetical population of 100,000 individuals using each proposed invitation and treatment strategy using a 10% risk threshold to determine statin initiation.

			Men		
Invitation and treatment strategy	**Total invitations**	**Total treated**	**Events saved**	**NNS**	**NNT**
Strategy 1: Population-wide invitation	301,704	44,229	186	1,620	238
Strategy 2: Population-wide invitation enhanced with PRS	297,354	36,753	183	1,622	201
Strategy 3: Personalised invitation using PRS	164,084	31,845	178	924	179
Strategy 4: Treatment with genetically predicted risk factor levels	40,380	40,380	198	204	204
			Women		
Invitation and treatment strategy	**Total invitations**	**Total treated**	**Events saved**	**NNS**	**NNT**
Strategy 1: Population-wide invitation	332,267	29,364	102	3,254	288
Strategy 2: Population-wide invitation enhanced with PRS	330,155	27,149	94	3,512	289
Strategy 3: Personalised invitation using PRS	186,123	18,035	87	2,140	207
Strategy 4: Treatment with genetically predicted risk factor levels	35,558	35,558	127	280	280

Abbreviations: CVD, cardiovascular disease; NNS, number needed to screen; NNT, number needed to treat; PRS, polygenic risk score

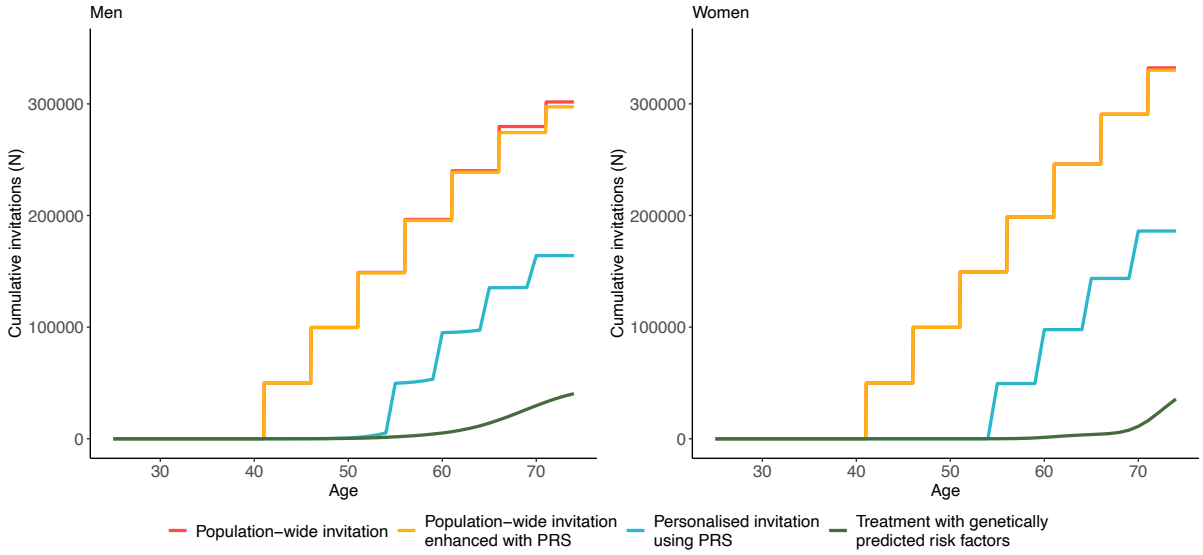


Figure 5.4: Cumulative number of invitations using each proposed strategy after accounting for treated individuals and a 10% risk threshold to determine statin initiation.

Abbreviations: PRS, polygenic risk score.

Vertical jumps in cumulative invitations occur due to repeated invitations every five years for individuals yet to be prescribed statins.

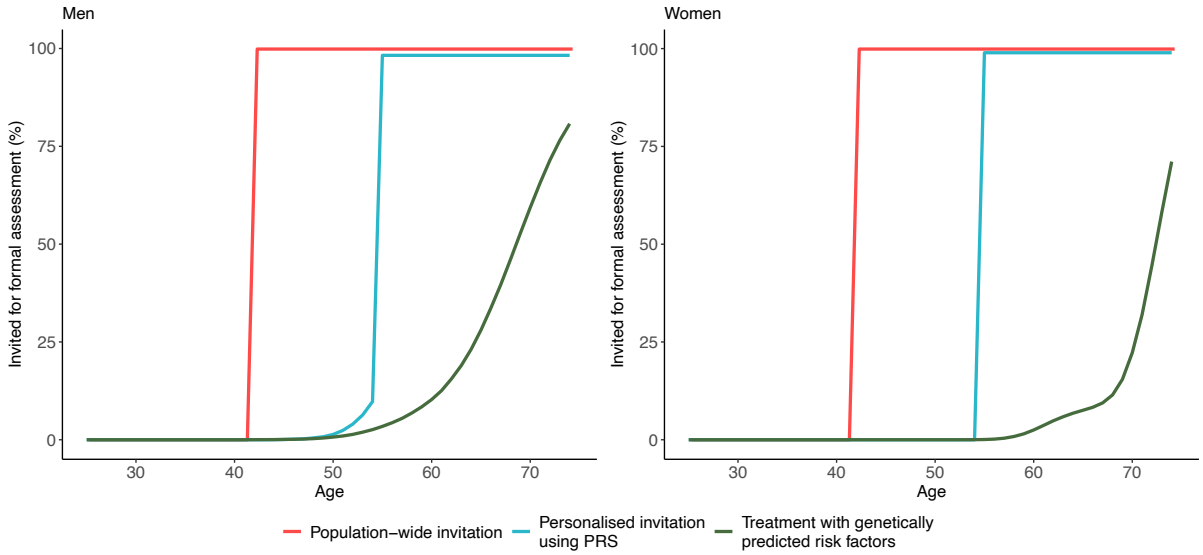


Figure 5.5: Cumulative percentage of individuals invited at least once using each proposed strategy and a 10% risk threshold to determine statin initiation.

Abbreviations: PRS, polygenic risk score.

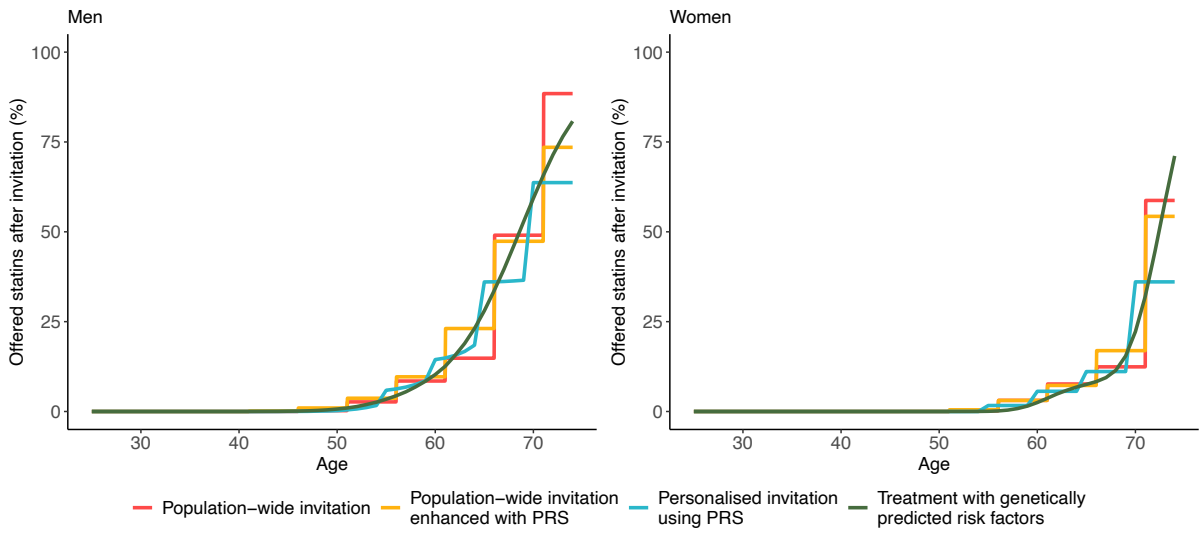


Figure 5.6: Cumulative percentage of individuals offered statins after being invited using each proposed strategy and a 10% risk threshold to determine statin initiation.

Abbreviations: PRS, polygenic risk score.

Vertical jumps in cumulative treatments occur due to repeated invitations every five years for individuals yet to be prescribed statins.

5.3.3.5 Earlier statin initiation in UK Biobank cohort

In UKB, we observed 3,407 men and 1,518 women who had a CVD event by the age of 75. Enhancing the risk model with PRS results in treating a greater number of individuals before an event occurred, and would also identify them as high risk earlier due to improved discriminatory performance and risk stratification **(Table 5.9).**

Using the first strategy 825 men (24.2%) men and 269 women (17.7%) were treated before experiencing their first event. Using the second strategy (Population-wide invitation enhanced with PRS), resulted in 981 men (28.8%) and 304 women (20.0%) being treated before experiencing their first event. Using the third strategy, 992 men (29.1%) and 311 women (20.5%) were treated before experiencing their first event. Using the fourth strategy, 1,123 men (33.0%) and 279 women (18.4%) were treated before experiencing their first event.

Enhancing a risk assessment model with PRS (e.g., from strategy 1 to 2) also identified high-risk individuals earlier. Using the first strategy, the median years of treatment before an event occurred was 4.2 years in men and 3.6 years in women. Using the second strategy increased

this to 5.3 years in men but decreased to 4.2 years in women. Using the third strategy and implementing a personalised invitation process further increased this to 5.4 years in men and increased to 4.1 years in women. Finally, using the fourth strategy led to the median years of treatment of 4.7 years in men and 3.2 years in women.

Table 5.9: Number of UK Biobank individuals who were treated with statins before an incident CVD event, and the median number of years of treatment, by sex and invitation strategy.

	Men	
Invitation and treatment strategy	**Treated before event, N (%)**	**Years of treatment before event, median (IQR)**
Strategy 1: Population-wide invitation	825 (24.2%)	4.2 (1,7, 8.5)
Strategy 2: Population-wide invitation enhanced with PRS	981 (28.8%)	5.3 (2.2, 9.8)
Strategy 3: Personalised invitation using PRS	992 (29.1%)	5.4 (2.4, 9.5)
Strategy 4: Treatment with genetically predicted risk factor levels	1,123 (33.0%)	4.7 (2.2, 8.6)
	Women	
Invitation and treatment strategy	**Treated before event, N (%)**	**Years of treatment before event, median (IQR)**
Strategy 1: Population-wide invitation	269 (17.7%)	3.6 (1.2, 8.4)
Strategy 2: Population-wide invitation enhanced with PRS	304 (20.0%)	3.5 (1.4, 8.5)
Strategy 3: Personalised invitation using PRS	311 (20.5%)	4.1 (2.1, 8.5)
Strategy 4: Treatment with genetically predicted risk factor levels	279 (18.4%)	3.2 (1.5, 7.0)

Abbreviations: CVD, cardiovascular disease; NNS, number needed to screen; NNT, number needed to treat; PRS, polygenic risk score

5.3.3.6 Sensitivity analysis assuming 5% threshold

In additional analyses assuming a 5% (rather than 10%) formal risk assessment threshold, each invitation and treatment strategy became more comparable **(Table 5.10).** The reduced threshold resulted in statins being prescribed to more than 98% of all men and women by age 75 across all strategies. The first and second strategy both showed similar results in men and women, with marginally higher NNS and NNT in men when enhancing the risk score with PRS.

However, comparing the second and the third strategy, the NNS increased in men. One possible explanation is due the strategy's ability to invite and treat between the ages of 25 and 40. Whilst 64% of men were invited between 25 and 40 years, a total of only 23% were treated by the age of 40 **(Figures 5.7-5.9).** This suggests the invitation process could be further optimised to reduce invitations if needed. However, whilst the NNS in men increased when using the third strategy, the NNS was reduced in women when compared to the second strategy. We observed that the first invitation began around 43 years old, with 94% of invitations occurring between 50 and 55 years. Consequently, this reduced the total number of invitations needed, further reducing the NNS.

Whilst the fourth strategy further highlights its ability to save a greater number of events and reduce its NNS, the improvements are diminished when using a fixed 5% risk assessment threshold.

Table 5.10: Number needed to screen and treat to prevent one CVD event by the age of 75 in a hypothetical population of 100,000 individuals using each proposed invitation and treatment strategy using a 5% risk threshold to determine statin initiation.

Men					
Invitation and treatment strategy	**Total invitations**	**Total treated**	**Events saved**	**NNS**	**NNT**
Strategy 1: Population-wide invitation, 5% threshold	78,492	49,807	351	224	142
Strategy 2: Population-wide invitation enhanced with PRS, 5% threshold	91,142	49,779	344	265	145
Strategy 3: Personalised invitation using PRS, 5% threshold	120,685	49,811	292	413	170
Strategy 4: Genetically predicted risk factor levels, 5% threshold	49,833	49,833	294	170	170
Women					
Invitation and treatment strategy	**Total invitations**	**Total treated**	**Events saved**	**NNS**	**NNT**
Strategy 1: Population-wide invitation, 5% threshold	202,733	49,339	262	773	188
Strategy 2: Population-wide invitation enhanced with PRS, 5% threshold	205,538	49,335	249	825	198
Strategy 3: Personalised invitation using PRS, 5% threshold	105,762	49,278	243	434	202
Strategy 4: Genetically predicted risk factor levels, 5% threshold	49,333	49,333	259	190	190

Abbreviations: CVD, cardiovascular disease; NNS, number needed to screen; NNT, number needed to treat; PRS, polygenic risk score

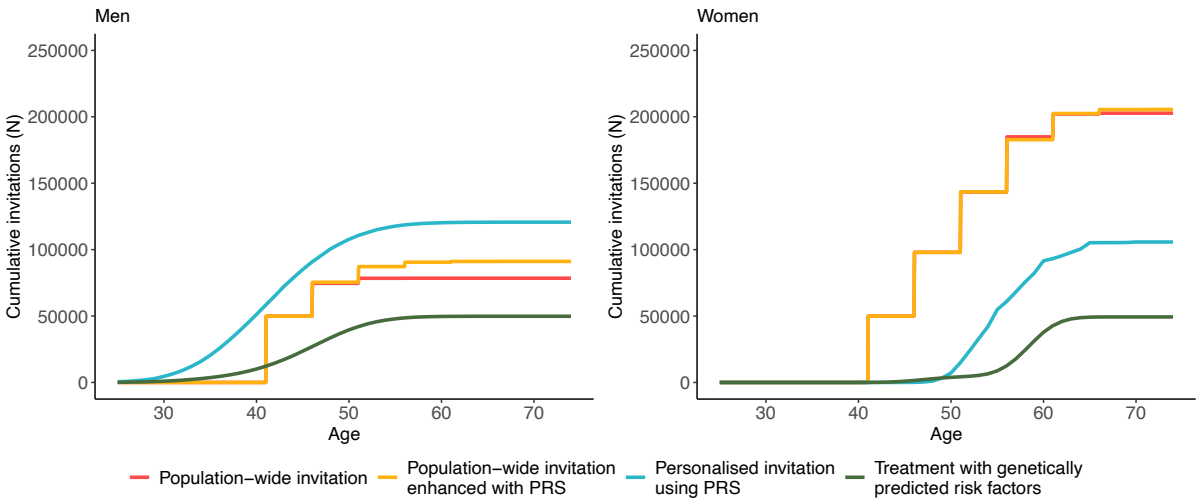


Figure 5.7: Cumulative number of invitations using each proposed strategy after accounting for treated individuals and a 5% risk threshold to determine statin initiation.

Abbreviations: PRS, polygenic risk score

Vertical jumps in cumulative invitations occur due to repeated invitations every five years for individuals yet to be prescribed statins.

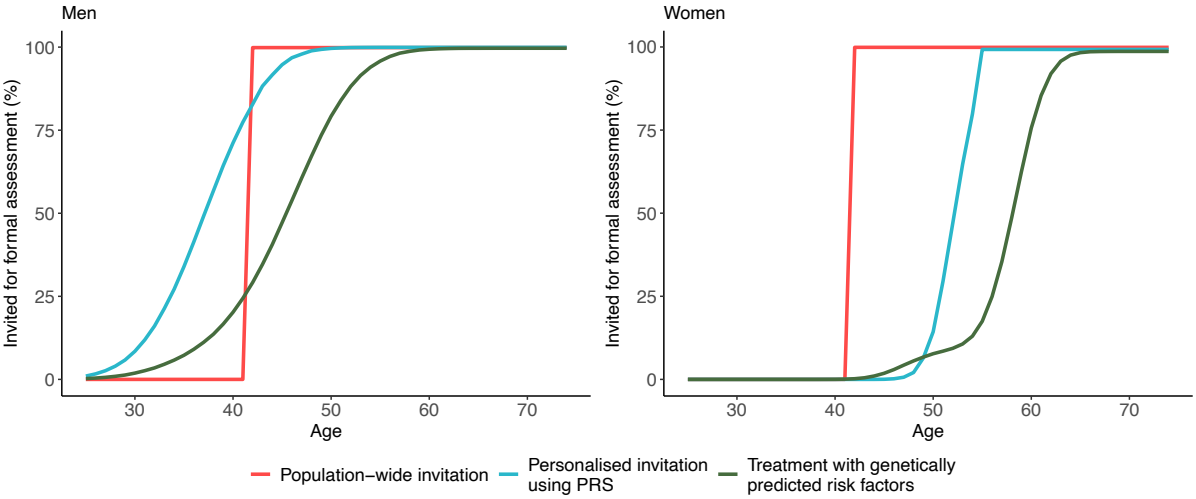


Figure 5.8: Cumulative percentage of individuals invited at least once using each proposed strategy and a 5% risk threshold to determine statin initiation.

Abbreviations: PRS, polygenic risk score

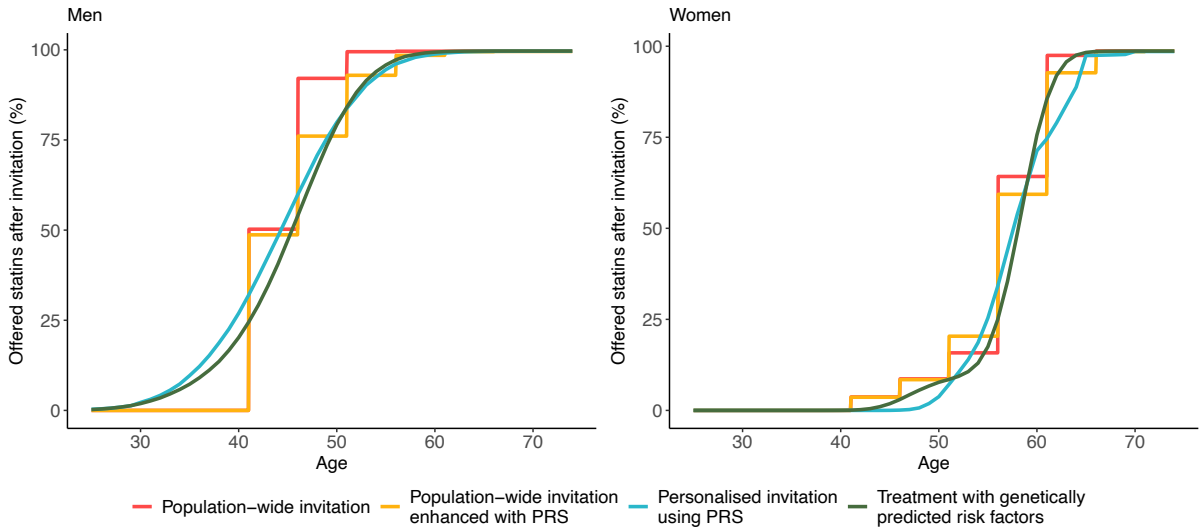


Figure 5.9: Cumulative percentage of individuals offered statins after being invited using each proposed strategy and a 5% risk threshold to determine statin initiation.

Abbreviations: PRS, polygenic risk score.

Vertical jumps in cumulative treatments occur due to repeated invitations every five years for individuals yet to be prescribed statins.

5.4 Discussion

This study has expanded on the work shown in **Chapters 3 and 4,** and assessed the impact of how PRS could be used to personalise invitation strategies for CVD risk assessment. We modelled the population health impact of implementing a personalised invitation strategy by estimating the total number of invitations needed, and the number of events saved due to statin initiation by the age of 75 years. Compared against current recommendations of a population-wide invitation approach followed by an assessment using a risk model with conventional risk factors (Strategy 1), enhancing the risk model with PRS (Strategy 2) resulted in improvements in the early identification of high-risk individuals, with the number of years of treatment increasing in those identified with an event in the UKB cohort. However**,** this did not translate to improvements across a lifetime in men in **Chapter 5**, with similar results in the NNS and NNT between Strategies 1 and 2.

The study also investigated a personalised invitation strategy that increased the level of PRS implementation. By using CVD-based PRS and least favourable risk factors across the population (Strategy 3), we showed the invitation process could be personalised by only inviting for a formal risk assessment when needed. Compared to Strategy 2, the personalised invitation strategy led to a reduction in the NNS and NNT whilst saving a similar number of events for both sexes. The results show similarities to the approach of using age- and sex-specific thresholds, which showed higher efficiency and improved effectiveness, especially amongst younger individuals (see **Chapter 4).** In this study, we also investigated a potential future scenario to highlight the utility of PRS, by using genetically predicted values to guide treatment decisions and forgoes the formal risk assessment process (Strategy 4), where we observed significantly lower NNS in both sexes.

The results aim to address the limited evidence relating to how PRS could be implemented into a national risk assessment programme. Whereas the majority of research has shown improvements in risk model performance, we have highlighted that PRS can be implemented in ways that could be more effective and more equitable, by allocating resources to individuals at highest risk first.[21] In addition to the results shown in **Chapter 3** and **Chapter 4**, we have shown the benefits of optimising the prioritisation of individuals, using either optimised risk thresholds or harnessing PRS for the same purpose.

Similar to our work in **Chapter 4,** the future implementation of PRS within the healthcare system to enable our findings is key. In particular, policymakers should focus on it is feasible and optimal to collect PRS. In the context of **Chapter 5**, this is especially important if focussing on the long-term benefits of PRS at targeting younger individuals who are at a potentially high lifetime risk of CVD. Health economics can also be used to evaluate the effectiveness of the proposed prioritisation schemes.

Like in **Chapters 3 and 4**, we have chosen to highlight efficiency improvements. In particular, we present a personalised invitation strategy using PRS which defers individuals for their first formal risk assessment until they are deemed at a high enough risk. This approach however will decrease sensitivity in favour of specificity and an alternative strategy, such as Strategy 1 or Strategy 2, may be preferred.

5.4.1 Strengths

Compared to our previous work in **Chapter 4,** this chapter demonstrates several strengths. First, we extended our population health modelling to estimate the lifetime benefits of statin initiation accounting for invitation and formal assessments over time rather than at a single time point. By modelling over a longer period of time, the long-term benefits of using a PRS for personalised invitations to CVD risk assessment was investigated. Second, we investigated different invitation strategies, each chosen to represent increasing levels of PRS implementation. Third, we utilised the full extent of the linked primary care records to estimate individual-level risk factor profiles for all ages. By doing so, we were able to estimate the expected 10-year risks at a formal risk assessment at any point, which allowed us to model each invitation strategy. This approach helped generalise the population health modelling across a wider range of ages. Fourth**,** we generalised our findings to the more general population of the UK. Whilst we used recalibration methods to better approximate the 10-year risks expected in the general population, as seen in **Chapters 3 and 4**, **Chapter 5** also adjusted for competing risks and incorporated methods to upweight both CVD and non-CVD events to adjust for UKB's low incidence rates.

5.4.2 Limitations

Our work has some limitations. First, we did not model other potential impacts to CVD incidence, including antihypertensive medication or other health interventions. One possible

approach would be to conduct a microsimulation model, which may be able to model these changes with greater precision. Second, we did not investigate the costs of sequencing and any reductions in costs due to fewer invitations and formal assessments. As introducing statins earlier will produce improvements in CVD events averted, it is therefore important to understand the trade-off in terms of costs and any other harms associated with this, including any clinical side effects. These could be assessed using a microsimulation model and be used to provide a more comprehensive comparison of the strategies shown in this study. Third, we assumed a constant reduction in risk due to statins, however this is unlikely to be true and the reduction is likely to be dependent on genetic risk.[22] Finally, whilst representative incidence rates from CPRD were used to recalibrate models, the population may not be completely representative of the general population in the UK (see **Chapter 3, Section 3.2.1.1)**.

5.4.3 Future work

Potential extensions to this work could investigate dynamic invitation intervals; in this work, the interval between invitations was assumed to remain at five years, chosen to be in line with current guidelines.[16] Future work could explore varied invitation intervals such as a reduced invitation interval for higher-risk individuals or invitation intervals based off PRS.[23] Future work could also investigate using PRS for invitations to risk assessments for other chronic diseases assessed in the NHS Health Check, including diabetes and kidney disease.

5.5 Conclusion

The use of PRS to personalise invitations to formal CVD risk assessments compared to a population-wide invitation strategy has the potential to substantially reduce the number needed to screen across a lifetime. Our results suggest that not only can PRS be used to improve risk model performance, it can also be effectively used to personalise the invitation process by inviting high-risk individuals earlier and low-risk individuals later. This invitation strategy highlights the future potential of PRS and its implementation in the healthcare system.

Chapter 6 will further discuss the results of this chapter, alongside the previous chapters, and its potential public health implications.

References

1. Secretary PU, Health P, Care P, Majesty H, July P. Advancing our health: prevention in the 2020s Printed on paper containing 75% recycled fibre content minimum. 2019;(July). Accessed January 29, 2023. https://www.gov.uk/government/consultations/advancing-our-health-prevention-in-the-2020s
2. Abraham G, Malik R, Yonova-Doing E, et al. Genomic risk score offers predictive performance comparable to clinical risk factors for ischaemic stroke. *Nat Commun 2019 101*. 2019;10(1):1-10. doi:10.1038/s41467-019-13848-1
3. Abraham G, Havulinna AS, Bhalala OG, et al. Genomic prediction of coronary heart disease. *Eur Heart J*. 2016;37(43):3267-3278. doi:10.1093/eurheartj/ehw450
4. Inouye M, Abraham G, Nelson CP, et al. Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. *J Am Coll Cardiol*. 2018;72(16):1883-1893. doi:10.1016/j.jacc.2018.07.079
5. Sun L, Pennells L, Kaptoge S, et al. Polygenic risk scores in cardiovascular risk prediction: A cohort study and modelling analyses. Hindy G, ed. *PLOS Med*. 2021;18(1):e1003498. doi:10.1371/journal.pmed.1003498
6. Gooding HC, Gidding SS, Moran AE, et al. Challenges and opportunities for the prevention and treatment of cardiovascular disease among young adults: Report from a national heart, lung, and blood institute working group. *J Am Heart Assoc*. 2020;9(19). doi:10.1161/JAHA.120.016115
7. Singh A, Collins BL, Gupta A, et al. Cardiovascular Risk and Statin Eligibility of Young Adults After an MI: Partners YOUNG-MI Registry. *J Am Coll Cardiol*. 2018;71(3):292-302. doi:10.1016/J.JACC.2017.11.007
8. Wu H, Forgetta V, Zhou S, Bhatnagar SR, Paré G, Richards JB. Polygenic Risk Score for Low-Density Lipoprotein Cholesterol Is Associated With Risk of Ischemic Heart Disease and Enriches for Individuals With Familial Hypercholesterolemia. *Circ Genomic Precis Med*. 2021;14(1):E003106. doi:10.1161/CIRCGEN.120.003106
9. Vaura F, Kauko A, Suvila K, et al. Polygenic Risk Scores Predict Hypertension Onset and Cardiovascular Risk. *Hypertension*. 2021;77(4):1119-1127. doi:10.1161/HYPERTENSIONAHA.120.16471
10. Parcha V, Pampana A, Shetty NS, et al. Association of a Multiancestry Genome-Wide Blood Pressure Polygenic Risk Score With Adverse Cardiovascular Events. *Circ Genomic Precis Med*. 2022;15(6):E003946. doi:10.1161/CIRCGEN.122.003946
11. Meisner A, Kundu P, Zhang YD, et al. Combined Utility of 25 Disease and Risk Factor

Polygenic Risk Scores for Stratifying Risk of All-Cause Mortality. *Am J Hum Genet*. 2020;107(3):418-431. doi:10.1016/J.AJHG.2020.07.002

12. Willer CJ, Schmidt EM, Sengupta S, et al. Discovery and refinement of loci associated with lipid levels. *Nat Genet*. 2013;45(11):1274-1285. doi:10.1038/NG.2797
13. Hoffmann TJ, Ehret GB, Nandakumar P, et al. Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. *Nat Genet*. 2017;49(1):54-64. doi:10.1038/NG.3715
14. Herrett E, Shah AD, Boggon R, et al. Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. *BMJ*. 2013;346(7909). doi:10.1136/BMJ.F2350
15. Pennells L, Kaptoge S, Wood A, et al. Equalization of four cardiovascular risk algorithms after systematic recalibration: Individual-participant meta-analysis of 86 prospective studies. *Eur Heart J*. 2019;40(7):621-631. doi:10.1093/eurheartj/ehy653
16. National Institute for Health and Care Excellence (NICE). Lipid modification: Cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease (clinical guideline CG181). Published online 2014.
17. Isgut M, Sun J, Quyyumi AA, Gibson G. Highly elevated polygenic risk scores are better predictors of myocardial infarction risk early in life than later. *Genome Med*. 2021;13(1):1-16. doi:10.1186/S13073-021-00828-8/FIGURES/6
18. Lewis CM, Vassos E. Polygenic risk scores: From research tools to clinical instruments. *Genome Med*. 2020;12(1):1-11. doi:10.1186/S13073-020-00742-5/TABLES/2
19. Patel R, Barnard S, Thompson K, et al. Evaluation of the uptake and delivery of the NHS Health Check programme in England, using primary care data from 9.5 million people: A cross-sectional study. *BMJ Open*. 2020;10(11):42963. doi:10.1136/bmjopen-2020-042963
20. Martin A, Saunders CL, Harte E, et al. Delivery and impact of the NHS Health Check in the first 8 years: A systematic review. *Br J Gen Pract*. 2018;68(672):e449-e459. doi:10.3399/bjgp18X697649
21. Kypridemos C, Collins B, McHale P, et al. Future cost-effectiveness and equity of the NHS Health Check cardiovascular disease prevention programme: Microsimulation modelling using data from Liverpool, UK. Sheikh A, ed. *PLOS Med*. 2018;15(5):e1002573. doi:10.1371/journal.pmed.1002573
22. Natarajan P, Young R, Stitziel NO, et al. Polygenic Risk Score Identifies Subgroup With Higher Burden of Atherosclerosis and Greater Relative Benefit From Statin Therapy in

the Primary Prevention Setting. *Circulation*. 2017;135(22):2091-2101. doi:10.1161/CIRCULATIONAHA.116.024436

23. Lindbohm J V, Sipilä PN, Mars NJ, et al. 5-year versus risk-category-specific screening intervals for cardiovascular disease prevention: a cohort study. *Lancet Public Heal*. 2019;4(4):e189-e199. doi:10.1016/S2468-2667(19)30023-4

Chapter 6

Discussion

Thesis summary

The overall aim of this thesis was to investigate and evaluate the potential health impact of using primary care records and genetics to improve the risk stratification and prioritisation of individuals at high-risk of cardiovascular disease (CVD) for risk assessments within a primary care setting.

The motivation behind this thesis was based off the clinical guidelines on the primary prevention of CVD in England by the National Institute for Health and Care Excellence (NICE). Current guidelines recommend the identification and prioritisation of individuals who may benefit the most from a full formal risk assessment, and is to be performed systematically using CVD risk factors already recorded in primary care electronic medical records. However, no dedicated prioritisation tool is currently recommended. In addition, genetic research has continued to develop with the creation of polygenic risk scores (PRS) for CVD and conventional CVD risk factors. With large foundations envisioning a future healthcare system that incorporates widespread genetic profiling, this thesis also explores its use within the primary care system and compares its potential with primary care records.

Epidemiological and statistical analysis was conducted using participants and their data from UK Biobank. Using the cohort's unique data structure, consisting of detailed baseline measurements, linked primary care records, genetic data and follow-up data, we modelled the population health benefits of implementing different prioritisation tools. Characteristics in risk factor measurements within the linked primary care records in UK Biobank and the measured data at baseline was first assessed to understand this unique dataset and potential challenges needed to be addressed for the population health modelling **(Chapter 2)**. A novel prioritisation tool using primary care records alone was created and then assessed in a representative population **(Chapter 3)**. Motivated by findings from **Chapter 3**, PRS were evaluated to understand how enhancing models that use conventional risk factors with PRS affect risk assessments with prioritisations. **(Chapter 4)**. Finally, PRS were used to create a strategy that

personalises the first age of invitation prior to a risk assessment, where the lifetime impact of statins was estimated **(Chapter 5).**

This chapter summarises the main findings, discusses the thesis's strengths and weaknesses, its potential health implications and outlines future work.

6.1 Summary of findings

6.1.1 Development of novel primary-care based prioritisation model with age- and sex- specific thresholds

Formal CVD risk assessments are a fundamental part of the NHS Health Check. Current clinical guidelines in England by NICE and policy makers have advocated using primary care records for systematic prioritisation before an assessment to reduce programme running costs and address health inequalities. Specifically, the current guidelines recommend systematically prioritising using existing records, and that a fixed 10-year CVD risk threshold of 10% should be used. However, a specific risk tool for use with prioritisation has not been recommended and quantitative evidence of the health impact of prioritisation when using a fixed 10% threshold in clinical practice is limited. **Chapter 3** aimed to develop a novel prioritisation tool and evaluate its potential impact, and compared against current practice.

The novel prioritisation tool (eHEART) was derived in 1,642,498 individuals from the Clinical Practice Research Datalink (CPRD). A two-stage landmark modelling approach was applied to repeated measures of conventional CVD risk predictors. In the first stage, landmark age- and sex- specific multivariate mixed-effects linear regression models, using random intercepts and fixed slopes, were applied to repeated measures for systolic blood pressure (SBP), total cholesterol, high-density lipoprotein (HDL) cholesterol and smoking status. In the second stage, sex-specific Cox regression models were fitted to predict 10-year CVD risk using the estimated risk factor values from the first stage, diabetes status and treatment for hypertension. The tool was derived in two-thirds of the dataset and internally validated in the remaining third.

Population health modelling of using eHEART in a primary-care setting was then evaluated in 119,137 individuals with linked primary care records from the UK Biobank cohort. A representative population of 100,000 individuals in England was created, where the age structure was estimated using data from the Office for National Statistics (ONS) and CVD incidence rates from CPRD, and estimated 10-year risks were recalibrated and rescaled for

validation in UK Biobank. Scenarios were created to compare formally assessing all individuals with a current CVD risk tool, QRISK2, with prioritisation with eHEART before assessment with QRISK2. A fixed 10% risk prioritisation threshold was then compared with age-group and sex-specific prioritisation thresholds that limited the false negative rate to 5%.

First, the results suggest that compared to formally assessing all individuals, prioritising with eHEART and choosing optimised, age- and sex-specific prioritisation thresholds can reduce the number needed to screen to prevent one CVD event (NNS) by approximately 50% for women and 20% for men whilst identifying 96-98% of high-risk individuals. Second, the results suggest prioritising using QRISK2, an existing risk score, can perform as well or even better than eHEART. However, by implementing a landmark model approach, eHEART removes the need for complete risk predictor measurements in all individuals. In contrast, the implementation of QRISK2 is based on replacing missing non-recorded values with age, sex and ethnicity-specific population average values. Third, the use of a fixed 10% prioritisation threshold may not be appropriate for younger individuals and women. These groups of individuals are on average likely to have a low 10-year CVD risk, but may lead to individuals with relatively poor risk factor levels to not be prioritised. The results demonstrated a pragmatic approach to selecting age- and sex-specific prioritisation thresholds to improve the number of individuals prioritised, subsequently reducing the NNS across the population and ensuring a balance between efficiency and specificity.

6.1.2 Supplementing primary care records with PRS for CVD risk prioritisation

The UK government has stated that as research into genetics continues to develop, complementing existing CVD risk scores with genetic data, for example in the form of PRS, is of importance before consideration for integration into the wider healthcare system.[1] However, the majority of research investigating the benefits of PRS have largely focussed on the improved performance of risk models. To date, no studies have quantified the impact implementing PRS would have within prioritisation. The work in **Chapter 4** aimed to assess the population heath impact of using prioritisation tools derived using either primary care records only, PRS only, or the combination of both.

Building on the population health modelling approach of Chapter 3, a direct within-person comparison of the prioritisation tools was conducted in participants of UK Biobank. To allow for a comparison between each of the proposed prioritisation tools, 108,685 participants aged

40-69, with measured biomarkers at study entry, linked primary care records and genetic data were used for both model derivation and population health modelling.

The three prioritisation tools, derived as sex-specific Cox models, used: 1) repeated measurements from primary care records summarised using sex- specific multivariate mixed-effects linear regression models, as motivated by the findings from **Chapter 2**, 2) age, CAD PRS and stroke PRS or 3) the combination of primary care records and CAD and stroke PRS. Two formal risk assessment models for predicting 10-year formal assessment CVD risk using risk factor measurements observed at UKB baseline survey were also derived. The first uses QRISK2 predictors and the second uses QRISK2 predictors enhanced with the CAD and stroke PRS.

Like in **Chapter 3,** population health modelling was conducted in a representative population of 100,000 individuals in England to quantify the wider benefits of using each of the proposed prioritisation tools. We compared three strategies: 1) prioritising using a primary care records-based tool followed by a formal assessment with conventional risk factors, 2) prioritising using a PRS and age-based tool followed by a formal assessment with conventional risk factors and PRS and 3) prioritising using both PRS and primary care records, followed by a formal assessment with conventional risk factors and PRS.

Results indicated that compared to prioritisation with primary care records followed by a formal risk assessment with conventional CVD risk factors, the addition of PRS to both prioritisation and formal assessment improved the correlation between the two tools, leading to a reduced NNS and improved the ability to identify individuals at high risk of a future CVD event. If the goal was to identify the same number of events that would be previously identified when prioritising with primary care records, then the addition of PRS to both stages can reduce the NNS by around 20% and 35% in men and women respectively. Results also suggests that as primary care records are readily available in current healthcare systems, prioritisation using only primary care records could be a viable approach, as the added benefits of PRS may not outweigh the challenges involved with the data collection. . In addition, prioritisation using only age, and the CAD PRS and stroke PRS is less effective than prioritisation with primary care records. This is to be expected as PRS offer a long-term indication of lifetime risk and may not be representative of current health status, which is more relevant for a risk model calculating 10-year CVD risk.

6.1.3 Lifetime impact of using PRS to personalise invitations to formal risk assessments

The evidence investigating the lifetime population health impact of PRS is limited. As PRS is fixed from birth, PRS presents a unique opportunity to tailor risk assessment programmes to individuals. Consequently, the opportunity to target younger individuals at high-risk of CVD may be able to assist in reducing future CVD burden. Compared to the population health modelling approaches presented in **Chapters 3 and 4**, where 10-year CVD risk was assessed at a single time point, **Chapter 5** expanded on this work, by investigating how polygenic risk scores (PRS) could be used to determine the optimal age at which to invite an individual for a cardiovascular disease (CVD) risk assessment, and estimate the lifetime benefits of such an approach, taking into account reassessments of CVD risk over time. In addition, **Chapter 5** worked under a framework of existing clinical guidelines, using fixed CVD risk thresholds instead of age-and sex specific thresholds. This was chosen to reflect the challenges in updating key aspects of guidelines.

300,088 participants, aged 40-70, with measured biomarkers, genetic data and without a history of CVD, diabetes and lipid lowering medication in UK Biobank were used to derive three risk models to estimate 10-year CVD risk. First, a 'conventional risk model' with conventional CVD risk factors, second a 'conventional + PRS risk model', which includes a CAD and stroke PRS, and third a 'PRS based risk model' using genetically predicted risk factor levels.

Four strategies for determining the first age of invitation, followed by a formal risk assessment at which treatment would be allocated, were devised: 1) *Population-wide invitation* where all individuals were invited for a formal assessment at age 40 years, and treatment offered if CVD risk ≥10% at the formal assessment using conventional risk factors. 2) *Population-wide invitation enhanced with PRS*, where all individuals are invited for a formal assessment at age 40 years, and treatment offered if CVD ≥10% at the formal assessment using conventional risk factors and PRS. 3) *Personalised invitation using PRS,* with the first age of invitation calculated using the individual's CAD and stroke PRS and "least favourable LDL, HDL and SBP levels", followed by formal assessment using conventional risk factors and PRS. 4) *Treatment with genetically predicted risk factors,* where individuals would be invited and offered statins at the age at which their genetically predicted risk exceeds 10%.

Unlike **Chapters 3 and 4,** population health modelling took into account multiple invitations across an individual's lifetime and adjusted for competing risks. Results indicated that compared to using a risk model using conventional risk factors, enhancing it with PRS resulted in improvements within women. However, the improved discriminatory performance in men (shown in **Chapter 4),** did not translate to improvements across a lifetime in men in **Chapter 5.** Second, we showed that a personalised strategy to personalise the first age of invitation can lead to a reduction in the NNS, similar to using age-and sex specific thresholds shown in **Chapters 3 and 4**. We also investigated a potential future scenario of using genetically predicted values to highlight the utility of PRS, where we observed significantly lower NNS in both sexes. We also observed that incorporating PRS into a risk model can improve the early identification of high-risk individuals.

6.2 Strengths and limitations

In addition to the specific strengths and limitations for each results chapter, the overall thesis has some overarching strengths and weaknesses.

Generally, this work has provided a comprehensive investigation in enhancing prioritisation for CVD risk assessments by using existing primary care records found in electronic health records and evaluated the future potential of incorporating genetic data, in the form of polygenic risk scores. This thesis has made novel findings; One major advantage of the thesis is the **use of UK Biobank to guide the population health modelling**. By leveraging the resources UK Biobank has to offer, in particular the detailed measurements taken at baseline, the linked primary care records for a subset of participants, and the genetic data captured at baseline, we were able to provide between-person comparisons between the different types of data. To the author's knowledge, UK Biobank is one of the largest available datasets in the UK with this combination of resource, with over 170,000 individuals, and allowed for a direct in-person comparison of the different prioritisation tools proposed.

This work also integrated **statistical methods to enhance the generalisability of the results**. By using simple and transparent methods, including the recalibration of estimated risks and using population health data to generate a hypothetical population of representative individuals, the results aimed to interpret the statistical models for use in a primary care setting in England.

Another key advantage shared between the results is the use of multivariate mixed-effects linear regression **models to leverage the repeated measurements found within the primary care records**. Taking advantage of its ability to impute any missing risk factor values using the available information for each individual and the intra-correlation matrix, allowed for a greater number of individuals to be included in the analysis.[2] The model allowed us to estimate the current expected risk factor levels for each individual, in **Chapters 3 and 4.** In addition, we used these methods in **Chapter 5** to offer individual-level projected risk factor levels over a lifetime. This enabled us to model counterfactual scenarios of whether an individual would be deemed at high risk.

This work also has several limitations. First, whilst UK Biobank is one of the largest available datasets that allowed for the comparison of primary care records and genetics for prioritisation, characteristics of those in UK Biobank differ to those in the general population, with lower incidence of disease and healthier risk factor levels. In addition, by using only those with a linked primary care record, further selection biases may exist, such as including fewer younger individuals and individuals in an ethnic minority group. Second, although statistical methods were employed to translate the results from UK Biobank to the general population, including the recalibration of estimated risks changes the distribution of estimated risks by applying a linear transformation to the non-recalibrated 10-year risks. However, **recalibration does not change the underlying risk factors and CVD incidence of the individuals** in UK Biobank. This may affect the generalisability of the results in **Chapter 5**, which relies on the events observed in UK Biobank to estimate total cumulative incidence. Third, whilst data from CPRD, used to estimate representative risk factor levels and incidence rates, are generally representative of the primary care attending population in the United Kingdom, **the CPRD data used does not have comprehensive coverage in the North and East of England**.[3,4] Fourth, **the results do not take into account other potential health impacts**. Whilst it was assumed that prioritisation increased uptake for a formal risk assessment due to potential behavioural changes due to prioritisation, we did not model the combined effects of hypertensive medication and statin initiation over time. In addition, the population modelling assumed a constant uptake and continuation of statin treatment. Fifth, the definition of CVD outcomes was chosen to reflect the outcomes used in the QRISK family of risk scores. With the variation in outcomes used for other published risk scores, **additional recalibration and adjustments are necessary to translate the analyses to other healthcare systems**.[5,6] Sixth, whilst genomics may offer the potential to improve risk stratification, **the results in this work assumes all individuals to**

have genetic data. Whilst this is not currently feasible, the potential may be realised in the near future due to continually decreasing costs associated with sequencing.

6.3 Public health implications

6.3.1 Utilising primary care records within primary prevention strategies

This thesis focusses on the implementation and impact of the prioritisation of individuals for a formal CVD risk assessment using existing information. In particular, analyses in **Chapter 3** highlights the potential benefits in harnessing the available longitudinal data within existing primary care records to prioritise with an estimate of an individual's 10-year CVD risk prior to a CVD risk assessment. The results found that two alternative approaches could be used. The first is to use an existing CVD risk score, for example QRISK2, or a dedicated tool derived using primary care records, such as the eHEART tool. The chapter observed that whilst using a risk score with a greater number of risk factors, and consequently a tool with greater discriminatory performance, can be used to effectively prioritise, a dedicated prioritisation tool that is designed to take advantage of existing repeated measurements may be feasible and inexpensive to implement.

With the eHEART tool, whilst sophisticated statistical methods (i.e., landmark-age specific mixed-effects linear regression models) were used to handle the longitudinal data from 1,642,498 individuals from CPRD, the tool can be easily transported once developed. As such, code for the eHEART tool was developed and shared onto GitHub. The accessibility and simplicity of the code could therefore allow for easy recalibration to other populations and for more primary care providers to offer a systematic and complementary approach for CVD risk assessments.

6.3.2 Personalised prediction to inform statin initiation

Currently, clinical guidelines recommend using a fixed 10% risk threshold for both prioritisation and for formal CVD risk assessments.[7] As age remains the largest contributor to CVD risk, younger individuals with a high relative risk will not be deemed at high risk of CVD until later in life.[8–10] This thesis presented a couple of alternative approaches to personalising risk prediction at both the group and individual level. In **Chapter 3**, the population health modelling of the eHEART prioritisation tool compared several scenarios: 1) Full formal assessment for all individuals, 2) prioritisation with a fixed 10% threshold followed by full

formal assessment and 3) prioritisation with age- and sex-specific corresponding to 5% false negative rates followed by full formal assessment. The results highlighted the benefits of the landmark modelling framework, and may allow for personalised prediction using routine data to be feasible and affect a large group of individuals. In particular, younger individuals who may have very few existing primary care records could be prioritised effectively using the eHEART prioritisation tool with its ability to impute missing values.[2] In addition, the results highlighted that the use of a fixed 10% threshold could result in a significant number of missed events, especially in younger individuals and women. We believe that this approach can be achieved easily and quickly, and can be implemented before other solutions, such as the improved integration of primary care records, and the implementation of PRS. As such, new guidelines should first focus on optimising prioritisation risk thresholds by age and sex.

Another potential implication of PRS is its role in improving risk communication. In **Chapter 5**, we investigated the use of PRS to not only improve model performance, but to personalise the invitation process and to potentially streamline the assessment process as a whole. PRS could be used as a communication tool, by showing patients their future CVD risk trajectories across their lifetime, using their genetic information and showing hypothetical scenarios using worst-case values for conventional CVD risk factors, average risk factor value and best-case scenario values. It has been shown that communicating polygenic risk to middle-aged individuals could motivate health behaviour changes, and enhance disease prevention.[11] As such, a future system may then be able to incorporate both PRS, to estimate lifetime risk, and primary care records to estimate current risk.

During the preparation of this thesis, NICE published new draft guidance for risk assessment and reduction of CVD.[12] Since 2014, individuals with a 10-year CVD risk greater than 10% were recommended statins. This is to change in the near future with a reduction in the fixed 10-year risk threshold from 10% to 5%. The decrease in threshold was motivated by new evidence on the safety of statins. In **Chapters 3-5**, additional results from sensitivity analyses showed that the change will lead to a greater number of statin prescriptions at younger ages, thus reducing the total number of assessments needed across a lifetime. It also suggests that prioritisation becomes less effective as there are fewer opportunities to prioritise before an individual is deemed at high risk. However, prioritisation still has its merits amongst women and amongst the youngest men as the majority have low absolute 10-year CVD risks. Prioritisation can therefore still be useful, especially when used with age- and sex-specific thresholds, and the use of PRS to communicate risk and personalise risk prediction are still

valid, however further work will need to be conducted. Furthermore, with the majority of healthcare systems across the world recommending 10-year CVD risk thresholds greater than 10%, opportunities remain in making healthcare systems more efficient globally.

6.4 Future work

6.4.1 Extension of current work

6.4.1.1 Extension to upcoming UK Biobank dataset

UK Biobank was used throughout the thesis and the analyses took advantage of its unique combination of linked primary care records, baseline measurements and genetic data. At the time of analysis, a subset of 177,361 individuals were linked with primary care record data. During the COVID-19 pandemic, UK Biobank further released primary care records for an additional 230,000 individuals, increasing the total number to 409,000 individuals.[13] However, it is currently unavailable for general research purposes and it is expected that this will change in the near future. The increase in data availability will allow for more accurate population health estimates, especially when updating work in **Chapter 5**.

6.4.1.2 Extension to other populations

Whilst the eHEART prioritisation tool developed using existing primary care records in CPRD (**Chapter 3)** provided good discriminatory performance and offered a unique opportunity to improve risk assessments in England, external validation was not conducted in other populations. Although internal validation using a split-sample derivation and validation approach was used, external validation in other populations, for example other European countries, will validate the tool's effectiveness in other healthcare systems and in populations with different risk characteristics. Potential cohorts for external validation include the Finnish CVD register, the Swedish national inpatient register, and the Estonian Biobank.[14–16]

Furthermore, population health modelling could easily be generalised to other countries. In **Chapters 3-4**, the modelling relied on the proportion of individuals with future events identified as high risk (sensitivity) to infer the total number of events identified in a hypothetical population. Future work could recalibrate the estimated risks to be specific for the population of interest and easily modify the population distribution and CVD incidence rates. Similar modifications could be performed if adapting the work presented in **Chapter 5,** including using incidence rates from the target population to calculate new weightings for individuals with

events, and using a mixed model to estimate person-level risk factor levels across the individual's lifetime.

6.4.1.3 Additional investigation of 5% risk threshold in England

As discussed in **Chapter 6.3.2,** new draft guidance for risk assessment and reduction of CVD in England proposed reducing the 10-year CVD risk threshold from 10% to 5%. Future work should aim to further understand the benefits and limitations of prioritisation after the change, and to also consider the future cost-effectiveness of PRS when integrated into a healthcare system where more statins will be offered to more individuals.

6.4.2 Exploration of PRS to enhance risk communication and guide treatment decisions

Although recent evidence has suggested the provision of genetic information may not importantly affect health-related behaviours, the genetic information presented consisted of a 10-year risk of coronary heart disease calculated using genetic risk factors only.[17] Future research could instead explore the utility of PRS in improving the CVD risk assessment process. This could include improving risk communication. For example, a future risk tool that presents an estimated risk trajectory using a combination of genetic data and measured risk factors, similar to Figure 2 in **Chapter 5**, could be used to further enhance risk communication between the clinician and patient about their future CVD risk over a lifetime. Interviews and panel groups could be used to better understand how such a resource could be used within primary care.

In addition to enhancing risk communication, a personalised risk trajectory could be used to inform treatment decisions. For example, if the expected 10-year risk using genetic information is greater than a set threshold by a predetermined age, then treatment should be initiated.

6.4.3 Microsimulation models to enhance population health modelling and evaluate cost-effectiveness

Chapter 5 presented a modelling approach to measure CVD risk across a lifetime and took into account competing risks. However, our current approach uses the cumulative CVD incidence within a group of individuals and relies on multiple assumptions, including a fixed reduction in risk due to statin initiation for all individuals. Microsimulation models may provide an alternative approach in evaluating decision making by simulating the impact of interventions

or strategies on individual trajectories, rather than the deterministic mean response of homogeneous cohorts.[18] Microsimulation models can also be implemented to assess other measures of interest, including quality adjusted life years (QALYs), net benefit and health economics.[19–21]

6.4.4 Running a trial to test predictive prevention policies against standard care

Throughout the thesis, we assumed that uptake of formal risk assessments increased due to the use of a prioritisation tool, and that compliance to interventions remained fixed over time. A future randomised controlled trial could be run to test the uptake and effectiveness. For example, the intervention arm could consist of personalised risk predictions using primary care records and/or genetic data, where tailored messages could be tested based on the estimated risks. To ensure patient safety, the trial would continue with standard care including a formal risk assessment every five years. This would then be compared with the placebo arm of standard care only. The trial could be assessed for invitation uptake, delivery of statins, statin compliance and a health economics analysis could be conducted.

References

1. GOV.UK. Genome UK: 2022 to 2025 implementation plan for England - GOV.UK. Accessed May 3, 2023. https://www.gov.uk/government/publications/genome-uk-2022-to-2025-implementation-plan-for-england/genome-uk-2022-to-2025-implementation-plan-for-england
2. Paige E, Barrett J, Stevens D, et al. Landmark Models for Optimizing the Use of Repeated Measurements of Risk Factors in Electronic Health Records to Predict Future Disease Risk. *Am J Epidemiol*. 2018;187(7):1530-1538. doi:10.1093/AJE/KWY018
3. Kontopantelis E, Stevens RJ, Helms PJ, Edwards D, Doran T, Ashcroft DM. Spatial distribution of clinical computer systems in primary care in England in 2016 and implications for primary care electronic medical record databases: a cross-sectional population study. *BMJ Open*. 2018;8:20738. doi:10.1136/bmjopen-2017-020738
4. Herrett E, Gallagher AM, Bhaskaran K, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol*. 2015;44(3):827-836. doi:10.1093/ije/dyv098
5. Pennells L, Kaptoge S, Wood A, et al. Equalization of four cardiovascular risk algorithms after systematic recalibration: Individual-participant meta-analysis of 86 prospective studies. *Eur Heart J*. 2019;40(7):621-631. doi:10.1093/eurheartj/ehy653
6. collaboration S-O working group and EC risk, de Vries TI, Cooney MT, et al. SCORE2-OP risk prediction algorithms: estimating incident cardiovascular event risk in older persons in four geographical risk regions. *Eur Heart J*. 2021;42(25):2455-2467. doi:10.1093/EURHEARTJ/EHAB312
7. National Institute for Health and Care Excellence (NICE). Lipid modification: Cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease (clinical guideline CG181). Published online 2014.
8. Lloyd-Jones DM, Leip EP, Larson MG, et al. Prediction of lifetime risk for cardiovascular disease by risk factor burden at 50 years of age. *Circulation*. 2006;113(6):791-798. doi:10.1161/CIRCULATIONAHA.105.548206
9. Dhingra R, Vasan RS. Age As a Risk Factor. *Med Clin North Am*. 2012;96(1):87-91. doi:10.1016/j.mcna.2011.11.003
10. Deanfield J, Sattar N, Simpson I, et al. Joint British Societies' consensus recommendations for the prevention of cardiovascular disease (JBS3). *Heart*. 2014;100(SUPPL. 2):ii1-ii67. doi:10.1136/heartjnl-2014-305693

11. Widén E, Junna N, Ruotsalainen S, et al. How Communicating Polygenic and Clinical Risk for Atherosclerotic Cardiovascular Disease Impacts Health Behavior: an Observational Follow-up Study. *Circ Genomic Precis Med*. 2022;15(2):E003459. doi:10.1161/CIRCGEN.121.003459
12. National Institute for Health and Care Excellence (NICE). Statins could be a choice for more people to reduce their risk of heart attacks and strokes, says NICE | News | News | NICE. Accessed March 29, 2023. https://www.nice.org.uk/news/article/statins-could-be-a-choice-for-more-people-to-reduce-their-risk-of-heart-attacks-and-strokes-says-nice
13. UK Biobank. UK Biobank Primary Care Data for COVID-19 Research. Published online 2021. Accessed March 29, 2023. www.ukbiobank.ac.uk
14. Sund R. Quality of the Finnish Hospital Discharge Register: a systematic review. *Scand J Public Health*. 2012;40(6):505-515. doi:10.1177/1403494812456637
15. Ludvigsson JF, Andersson E, Ekbom A, et al. External review and validation of the Swedish national inpatient register. *BMC Public Health*. 2011;11(1):1-16. doi:10.1186/1471-2458-11-450/COMMENTS
16. Leitsalu L, Haller T, Esko T, et al. Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int J Epidemiol*. 2015;44(4):1137-1147. doi:10.1093/IJE/DYT268
17. Silarova B, Sharp S, Usher-Smith JA, et al. Effect of communicating phenotypic and genetic risk of coronary heart disease alongside web-based lifestyle advice: the INFORM Randomised Controlled Trial. *Heart*. 2019;105(13):982-989. doi:10.1136/HEARTJNL-2018-314211
18. Krijkamp EM, Alarid-Escudero F, Enns EA, Jalal HJ, Hunink MGM, Pechlivanoglou P. Microsimulation modeling for health decision sciences using R: a tutorial. *Med Decis Making*. 2018;38(3):400. doi:10.1177/0272989X18754513
19. Kypridemos C, Collins B, McHale P, et al. Future cost-effectiveness and equity of the NHS Health Check cardiovascular disease prevention programme: Microsimulation modelling using data from Liverpool, UK. Sheikh A, ed. *PLOS Med*. 2018;15(5):e1002573. doi:10.1371/journal.pmed.1002573
20. Pasricha SR, Gheorghe A, Sakr-Ashour F, et al. Net benefit and cost-effectiveness of universal iron-containing multiple micronutrient powders for young children in 78 countries: a microsimulation study. *Lancet Glob Heal*. 2020;8(8):e1071-e1080. doi:10.1016/S2214-109X(20)30240-0
21. Si L, Eisman JA, Winzenberg T, et al. Microsimulation model for the health economic

evaluation of osteoporosis interventions: study protocol. *BMJ Open*. 2019;9(2):e028365. doi:10.1136/BMJOPEN-2018-028365

Appendix

Appendix 1: Code list of cardiovascular disease for eHEART derivation

Cardiovascular disease was defined as a combination of newly diagnoses of nonfatal or fatal events of coronary heart disease (CHD) (including myocardial infarction and angina), stroke, and transient ischemic attack (TIA), in line with the definition used in the QRISK3 CVD risk score[1]. In Clinical Practice Research Datalink (CPRD), diagnoses are coded using the hierarchical Read code system[1] and in the linked HES and ONS datasets, the International Classification of Disease 10th revision (ICD-10) codes were used[2].

Read code for CPRD data	
Read code	Description
G3...00	Ischaemic heart disease
G31..00	Arteriosclerotic heart disease
G32..00	Atherosclerotic heart disease
G33..00	IHD - Ischaemic heart disease
G30..00	Acute myocardial infarction
G301.00	Attack - heart
G302.00	Coronary thrombosis
G303.00	Cardiac rupture following myocardial infarction (MI)
G304.00	Heart attack
G305.00	MI - acute myocardial infarction
G306.00	Thrombosis - coronary
G307.00	Silent myocardial infarction
G309800	Coronary thrombosis
G309900	Myocardial Infarction
G300.00	Acute anterolateral infarction
G301.00	Other specified anterior myocardial infarction
G301000	Acute anteroapical infarction
G301100	Acute anteroseptal infarction
G301z00	Anterior myocardial infarction NOS
G302.00	Acute inferolateral infarction
G303.00	Acute inferoposterior infarction
G304.00	Posterior myocardial infarction NOS
G305.00	Lateral myocardial infarction NOS

G306.00	True posterior myocardial infarction
G307.00	Acute subendocardial infarction
G307000	Acute non-Q wave infarction
G307100	Acute non-ST segment elevation myocardial infarction
G308.00	Inferior myocardial infarction NOS
G309.00	Acute Q-wave infarct
G30A.00	Mural thrombosis
G30B.00	Acute posterolateral myocardial infarction
G30X.00	Acute transmural myocardial infarction of unspecif site
G30X000	Acute ST segment elevation myocardial infarction
G30y.00	Other acute myocardial infarction
G30y000	Acute atrial infarction
G30y100	Acute papillary muscle infarction
G30y200	Acute septal infarction
G30yz00	Other acute myocardial infarction NOS
G30z.00	Acute myocardial infarction NOS
G31..00	Other acute and subacute ischaemic heart disease
G319900	Acute/subacute IHD NOS
G310.00	Postmyocardial infarction syndrome
G310100	Dressler's syndrome
G311.00	Preinfarction syndrome
G311100	Crescendo angina
G311200	Impending infarction
G311300	Unstable angina
G311400	Angina at rest
G311000	Myocardial infarction aborted
G311010	MI - myocardial infarction aborted
G311100	Unstable angina
G311200	Angina at rest
G311300	Refractory angina
G311400	Worsening angina
G311500	Acute coronary syndrome
G311z00	Preinfarction syndrome NOS

G312.00	Coronary thrombosis not resulting in myocardial infarction
G31y.00	Other acute and subacute ischaemic heart disease
G31y000	Acute coronary insufficiency
G31y099	Acute coronary syndrome
G31y100	Microinfarction of heart
G31y200	Subendocardial ischaemia
G31y300	Transient myocardial ischaemia
G31yz00	Other acute and subacute ischaemic heart disease NOS
G32..00	Old myocardial infarction
G321.00	Healed myocardial infarction
G322.00	Personal history of myocardial infarction
G33..00	Angina pectoris
G330.00	Angina decubitus
G330000	Nocturnal angina
G330z00	Angina decubitus NOS
G331.00	Prinzmetal's angina
G331100	Variant angina pectoris
G332.00	Coronary artery spasm
G33z.00	Angina pectoris NOS
G33z000	Status anginosus
G33z100	Stenocardia
G33z200	Syncope anginosa
G33z300	Angina on effort
G33z400	Ischaemic chest pain
G33z500	Post infarct angina
G33z600	New onset angina
G33z700	Stable angina
G33zz00	Angina pectoris NOS
G34..00	Other chronic ischaemic heart disease
G349900	Chr. ischaemic heart dis. NOS
G340.00	Coronary atherosclerosis
G340100	Triple vessel disease of the heart
G340200	Coronary artery disease

G340000	Single coronary vessel disease
G340100	Double coronary vessel disease
G342.00	Atherosclerotic cardiovascular disease
G343.00	Ischaemic cardiomyopathy
G344.00	Silent myocardial ischaemia
G34y.00	Other specified chronic ischaemic heart disease
G34y000	Chronic coronary insufficiency
G34y100	Chronic myocardial ischaemia
G34yz00	Other specified chronic ischaemic heart disease NOS
G34z.00	Other chronic ischaemic heart disease NOS
G34z000	Asymptomatic coronary heart disease
G35..00	Subsequent myocardial infarction
G350.00	Subsequent myocardial infarction of anterior wall
G351.00	Subsequent myocardial infarction of inferior wall
G353.00	Subsequent myocardial infarction of other sites
G35X.00	Subsequent myocardial infarction of unspecified site
G36..00	Certain current complication follow acute myocardial infarct
G360.00	Haemopericardium/current comp follow acute myocardial infarct
G361.00	Atrial septal defect/curr comp follow acute myocardial infarct
G362.00	Ventricular septal defect/curr comp follow acute myocardial infarction
G363.00	Ruptur cardiac wall w'out haemopericard/cur comp follow ac MI
G364.00	Ruptur chordae tendinae/curr comp follow acute myocardial infarct
G365.00	Rupture papillary muscle/curr comp follow acute myocardial infarct
G366.00	Thrombosis atrium, auric append&vent/curr comp follow acute MI
G38..00	Postoperative myocardial infarction
G380.00	Postoperative transmural myocardial infarction anterior wall
G381.00	Postoperative transmural myocardial infarction inferior wall
G382.00	Postoperative transmural myocardial infarction other sites
G383.00	Postoperative transmural myocardial infarction unspec site
G384.00	Postoperative subendocardial myocardial infarction
G38z.00	Postoperative myocardial infarction, unspecified
G3y..00	Other specified ischaemic heart disease
G3z..00	Ischaemic heart disease NOS

G501.00	Post infarction pericarditis
Gyu3400	[X]Acute transmural myocardial infarction of unspecif site
F423600	Amaurosis fugax
Fyu5500	[X]Other transnt cerebral ischaemic attacks+related syndromes
G63y000	Cerebral infarct due to thrombosis of precerebral arteries
G63y100	Cerebral infarction due to embolism of precerebral arteries
G64..00	Cerebral arterial occlusion
G641.00	CVA - cerebral artery occlusion
G642.00	Infarction - cerebral
G643.00	Stroke due to cerebral arterial occlusion
G640.00	Cerebral thrombosis
G640000	Cerebral infarction due to thrombosis of cerebral arteries
G641.00	Cerebral embolism
G641100	Cerebral embolus
G641000	Cerebral infarction due to embolism of cerebral arteries
G64z.00	Cerebral infarction NOS
G64z100	Brainstem infarction NOS
G64z200	Cerebellar infarction
G64z990	Cerebral A. occlusion NOS
G64z000	Brainstem infarction
G64z100	Wallenberg syndrome
G64z110	Lateral medullary syndrome
G64z200	Left sided cerebral infarction
G64z300	Right sided cerebral infarction
G64z400	Infarction of basal ganglia
G65..00	Transient cerebral ischaemia
G651.00	Drop attack
G652.00	Transient ischaemic attack
G653.00	Vertebro-basilar insufficiency
G659900	Transient Ischaemic Attacks
G650.00	Basilar artery syndrome
G650100	Insufficiency - basilar artery
G652.00	Subclavian steal syndrome

G653.00	Carotid artery syndrome hemispheric
G654.00	Multiple and bilateral precerebral artery syndromes
G656.00	Vertebrobasilar insufficiency
G65y.00	Other transient cerebral ischaemia
G65z.00	Transient cerebral ischaemia NOS
G65z990	Transient Ischaemic Attacks
G65z000	Impending cerebral ischaemia
G65z100	Intermittent cerebral ischaemia
G65zz00	Transient cerebral ischaemia NOS
G66..00	Stroke and cerebrovascular accident unspecified
G661.00	CVA unspecified
G662.00	Stroke unspecified
G663.00	CVA - Cerebrovascular accident unspecified
G669800	Stroke/CVA - undefined
G669900	Stroke
G667.00	Left sided CVA
G668.00	Right sided CVA
G676000	Cereb infarct due cerebral venous thrombosis, nonpyogenic
G6W..00	Cereb infarct due unspcf occlus/stenos precerebr arteries
G6X..00	Cerebrl infarctn due/unspcf occlusn or sten/cerebrl artrs
Gyu6300	[X]Cerebrl infarctn due/unspcf occlusn or sten/cerebrl artrs
Gyu6400	[X]Other cerebral infarction
Gyu6500	[X]Occlusion and stenosis of other precerebral arteries
Gyu6600	[X]Occlusion and stenosis of other cerebral arteries
ZV12D00	[V]Personal history of transient ischaemic attack

ICD10 code for HES and ONS data	
ICD10-code	description
G45	transient ischaemic attack and related syndromes
G45.0	transient ischaemic attack and related syndromes
G45.1	transient ischaemic attack and related syndromes
G45.2	transient ischaemic attack and related syndromes

G45.3	transient ischaemic attack and related syndromes
G45.4	transient ischaemic attack and related syndromes
G45.8	transient ischaemic attack and related syndromes
G45.9	transient ischaemic attack and related syndromes
I20	angina pectoris
I20.0	angina pectoris
I20.1	angina pectoris
I20.8	angina pectoris
I20.9	angina pectoris
I21	acute myocardial infarction
I21.0	acute myocardial infarction
I21.1	acute myocardial infarction
I21.2	acute myocardial infarction
I21.3	acute myocardial infarction
I21.4	acute myocardial infarction
I21.9	acute myocardial infarction
I22	subsequent myocardial infarction
I22.0	subsequent myocardial infarction
I22.1	subsequent myocardial infarction
I22.8	subsequent myocardial infarction
I22.9	subsequent myocardial infarction
I23	complications after myocardial infarction
I23.0	complications after myocardial infarction
I23.1	complications after myocardial infarction
I23.2	complications after myocardial infarction
I23.3	complications after myocardial infarction
I23.4	complications after myocardial infarction
I23.5	complications after myocardial infarction
I23.6	complications after myocardial infarction
I23.8	complications after myocardial infarction
I24	other acute ischaemic heart disease
I24.0	other acute ischaemic heart disease
I24.1	other acute ischaemic heart disease

I24.8	other acute ischaemic heart disease
I24.9	other acute ischaemic heart disease
I25	chronic ischaemic heart disease
I25.0	chronic ischaemic heart disease
I25.1	chronic ischaemic heart disease
I25.2	chronic ischaemic heart disease
I25.3	chronic ischaemic heart disease
I25.4	chronic ischaemic heart disease
I25.5	chronic ischaemic heart disease
I25.6	chronic ischaemic heart disease
I25.8	chronic ischaemic heart disease
I25.9	chronic ischaemic heart disease
I63	cerebral infarction
I63.0	cerebral infarction
I63.1	cerebral infarction
I63.2	cerebral infarction
I63.3	cerebral infarction
I63.4	cerebral infarction
I63.5	cerebral infarction
I63.6	cerebral infarction
I63.8	cerebral infarction
I63.9	cerebral infarction
I64	stroke not specified as haemorrhage or infarction

Appendix 2: Two-stage dynamic landmark age model for risk prediction

Dynamic landmark age modelling

To optimize the use of repeated measurements of risk factors in electronic health records to predict future CVD risk, we used sliding landmark approach as described in our previous study[3] to construct 10-year CVD risk prediction models. The schematic of landmark age approach is presented in Web Figure 2. A landmark age is a reference point at which we use risk factor values collected prior to that age and from which to predict future risk[3]. In the derivation dataset, we derived a series of ninety two age- and sex-specific predictions models (i.e., for men and women and at ages 40, 41, 42, ...,85, denoted as "landmark ages"). Participants contributed to the models if they have 1) registered with a general practice at the landmark age, 2) no CVD diagnoses prior to the landmark age, and 3) no statin prescription prior to the landmark age.

We selected the key cardiovascular risk factors as those used in the validated 2013 ACC/AHA Pooled Cohort Equations[4]: age, sex, total cholesterol, high-density lipoprotein (HDL) cholesterol, systolic blood pressure (SBP), use of antihypertensive therapy, diabetes mellitus status, and smoking status. Values of SBP, total cholesterol and HDL cholesterol were standardised by centering on sex- specific means and dividing by the standard deviation (using means and standard deviations calculated from the first measurement from each individual). Age and sex were known for all participants. Values for diabetes mellitus status, anti-hypertensive therapy usage and statin therapy usage were set to zero until the first available health record indicated otherwise (i.e. for diabetes mellitus: at least one diabetes diagnostic code [Read code or diabetes test] plus either an additional diagnostic code or diabetes drug prescription[5]; first prescription of a blood-pressure or cholesterol medication) from which time the values were set to one. Repeat measurements of smoking status, systolic blood pressure, total cholesterol and HDL cholesterol were first summarised using age- and sex-specific multivariate mixed models[6] and entered the prediction model as single summary measures as described below.

The landmark age approach comprises of two stages:

Stage 1: Summarising repeated measures of risk factors using multivariate mixed- effects linear regression models

Let $\text{Smoking\_status}_{ij}$, SBP_{ij}, $\text{Total\_cholesterol}_{ij}$, $\text{HDL\_cholesterol}_{ij}$, BP_med_{ij} and Statin_{ij} denote all the repeat measurements of smoking status, SBP, total cholesterol, HDL cholesterol, indication of blood pressure-lowering medication, and indication of statin initiation for individual i recorded at measurement j. For males and females separately, for each landmark age $La = 40, 41, 42, ...,85$, we fit a multivariate mixed-effect model with a correlated covariance structure:

$$\text{SBP}_{ij} = \alpha_1 + \beta_1 * t_{ij} + \gamma * \text{BP\_med}_{ij} + u_{1i} + \varepsilon_{1ij}$$

$$\text{Total\_cholesterol}_{ij} = \alpha_2 + \beta_2 * t_{ij} + \delta * \text{Statin}_{ij} + u_{2i} + \varepsilon_{2ij}$$

$$\text{HDL cholesterol}_{ij} = \alpha_3 + \beta_3 * t_{ij} + u_{3i} + \varepsilon_{3ij}$$

$$\text{Smoking\_status}_{ij} = \alpha_4 + \beta_4 * t_{ij} + u_{4i} + \varepsilon_{4ij}$$

$$\text{Where } \begin{bmatrix} u_{1i} \\ u_{2i} \\ u_{3i} \\ u_{4i} \end{bmatrix} \sim \text{multivariate normal} \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{bmatrix} \right)$$

$$\text{And } \begin{bmatrix} \varepsilon_{1ij} \\ \varepsilon_{2ij} \\ \varepsilon_{3ij} \\ \varepsilon_{4ij} \end{bmatrix} \sim \text{multivariate normal} \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{e1}^2 & 0 & 0 & 0 \\ 0 & \sigma_{e2}^2 & 0 & 0 \\ 0 & 0 & \sigma_{e3}^2 & 0 \\ 0 & 0 & 0 & \sigma_{e4}^2 \end{bmatrix} \right)$$

Here α_1, α_2, α_3, α_4 represent fixed intercepts for each risk factor, β_1, β_2, β_3, $\beta_4$represent fixed slopes for each risk factor, γ represents an adjustment factor in systolic blood pressure levels for those with an indication of blood pressure-lowering medication and δ represents an adjustment factor in total cholesterol for those with an indication of statin medication.

Terms $u1i$, $u2i$, $u3i$ and $u4i$ represent random intercepts for each risk factor and are correlated between risk factors. These random intercepts are interpreted as the difference in the average level of the predictor for this individual compared to the population average level.

Finally, $\varepsilon 1ij$, $\varepsilon 2ij$, $\varepsilon 3ij$ and $\varepsilon 4ij$ represent uncorrelated residual errors for each risk factor.

This model allows incomplete records of the risk factors and includes all individuals with at least one measurement from at least one risk factor (see Web Figure 3). The correlation structure between the risk factors is estimated from individuals with observed data on more than one risk factor. Thus, the model assumes that, for each landmark age, risk factor values from individuals

with incomplete data are from the same multivariate normal distribution for risk factor values for individuals with observed data (that is, assuming "missing at random").

Our model assumes that all risk factors jointly follow a multivariate normal distribution, which is plausible for SBP, total cholesterol, HDL cholesterol but less plausible for smoking status which is defined as a binary variable (yes for current/ever smoker; no for never smoker). However, inference based from the multivariate normal distribution may often be reasonable even if the multivariate normality does not hold, especially in the context of imputation of missing data[7] and regression calibration[8,9].

In our previous work[3], we restricted the model derivation to repeat measurements recorded **before** the landmark age, i.e. $j \leq La$. However, we found slight improvements in sensitivity analyses when we used all available repeated measurements recorded **before** and **after** the landmark age, due to extra precision on parameter estimates. We accept a limitation is that it ignores informative censoring of individuals due to death or CVD events, however, our previous work shows informative censoring has little effect on the long-term usual levels of the included risk factors.

Best linear unbiased predictors (BLUPS)[10] are estimated for each risk factor for the random intercepts u_{1i}, u_{2i}, u_{3i} and u_{4i} using observed data for $j \leq La$ [11]. Note the restriction to only repeat measurements before the landmark age is important here, as the prediction model is intended for use in clinical practice where only past data will be available. The BLUPs are estimated as the mean of the empirical Bayes posterior distribution of the random intercepts conditional on observed risk factor measurements. Using the properties of multivariate normal distributions, this is also a multivariate normal distribution, and an exact formula for the mean can be calculated[12].

Specifically, for individual i

$$\begin{bmatrix} \hat{u}_{1i} \\ \hat{u}_{2i} \\ \hat{u}_{3i} \\ \hat{u}_{4i} \end{bmatrix} = GZ^T(ZGZ^T + \Sigma)^{-1}(Y\text{-}X\beta)$$

Here Y is the vector of risk factor observations, G is the covariance matrix of the random

effects = $\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{bmatrix}$, Z is the design matrix which selects the corresponding random effect for each risk factor, Z^T is the matrix transpose of Z and $\sum$ is a diagonal matrix containing the corresponding residual variance for each risk factor. Importantly, due to the correlations structure between the random intercepts, BLUPS can be estimated for all individuals with at least one repeat measurement for at least one risk factor.

Stage 2: Estimating 10-year CVD risk using landmark age- and sex-specific Weibull proportional hazards models, accounting for future statin initiation effect.

In the second stage, ten-year CVD risk was modelled using landmark age- and sex- specific Weibull models, with time since landmark age as the time scale. Landmark age datasets were constructed comprising of participants with no CVD diagnoses and/or statin prescription prior to that landmark age and included the following variables: (i) landmark-age-dependent outcome time-to-CVD-event and censoring indicator; (ii) sex; (iii) landmark- age-dependent estimated error-free risk factor values for SBP, total cholesterol, HDL cholesterol, smoking status and the most recent observed records for diabetes status and history of blood pressure-lowering medication prescriptions, denoted together as $X(La)$ and (iv) landmark-age- dependent time-to-statin-initiation during follow-up (set to the "time-to-CVD-event" if not observed) and statin-initiation indicator. We then split the time-to-CVD-event records at the time-to-statin-initiation, so that individuals who had an indication of statin initiation during follow-up had two records, one covering the landmark age before statin initiation, and the second from statin initiation to the CVD event or censoring. To each landmark age data set, we fit the following sex-stratified Weibull model:

$$h_S(t|X(La),La) = h_{0S}(t)\exp[\beta x^T X(La) + B \times Statin(t)]$$

where $h_{0S}(t) = \lambda \nu t^{\nu-1}$, with scale and shape parameters λ and ν, and the scale parameter λ is parameterized as exp ($\beta 0$); $Statin(t)$ is the time dependent indicator which equals 0 before an indication of statin-initiation, and equals 1 at and after the first indication of statin-initiation; and B represents the effect of statin-initiation on the risk of CVD, which is constrained to B = ln (0.75) to represent a 25% risk reduction as reported from published meta-analyses of trials.[13,14] This is done by using offset option in the Weibull survival model in Stata. The code sample for each landmark age by gender for estimating 10-year CVD risk accounting for statin-initiation is as follows:

```
stset ft, failure (cvd_ind ==1) id(patid)
stsplit new_statin_ind, after(time=statin_time) at(0)
replace new_statin_ind=new_statin_ind+1
*convert -1, 0 to 0,1 for on statins
replace new_statin_ind=0 if statins_prscd==. |
(statins_prscd!=. & statins_prscd>=exit_date)
```

```
*for never statin before exit_date
gen beta_x=new_statin_ind*ln(0.75)
streg sbp bp_medication tchol hdl smoke diabetes if
derivation==1, offset(beta_x) dist(weibull)
```

where ft is follow-up time; cvd_ind is the incident CVD indicator; statin_time is the time-to-statin-initiation during follow-up (set to the "time-to-CVD-event" if not observed); new_statin_ind is the time-varying indicator for statin initiation (1 for on statins, 0 for no statins); ln(0.75) is the 25% risk reduction as reported from published meta-analyses of trials; sbp, bp_medication, tchol, hdl, smoke, diabetes are the risk factor values estimated from Stage 1.

Predicted 10-year CVD risk is estimated for participants at each landmark age from the equation:

$$1 - P(T > La + 10 | T > La, X(La)) = 1 - S_{0S}(La + 10|La)^{\exp[\beta_X{}^T X(La) + B \times Statin(t)]}$$

where $S_{0S}(La + 10|La) = \exp(-\lambda t^{\nu})$ represents the sex-stratified 10-year baseline survival from landmark age La.

Other survival models, including the non-parametric Cox model, and more flexible parametric forms, could be used in place of the Weibull model. We selected the Weibull model due to a reasonable fit, and to enable a closed form solution to the calculation of counterfactual survival times in the absence of statin initiation (see Appendix 3). In our analysis, the fitted survival probability curves from the Weibull models were consistent with the Kaplan—Meier curves, indicating a reasonable fit (Web Figure 4). The fitted Weibull model shape parameters ranged between 1.07 to 1.34 across landmark ages (Web Figures 5 and 6) the Weibull model was more sufficient than exponential model of which the shape parameter is defined as 1.

Proportional hazard assumption in Weibull model:

The Weibull hazard function is $h_0(t) = \lambda \nu t^{\nu - 1}$ where the scale parameter λ is parametrized as $\exp(\beta 0)$. The hazard ratio for statin treatment is obtained as

$$HR = \frac{\exp(\beta_0 + \beta_s)\nu t^{\nu-1}}{\exp(\beta_0)\nu t^{\nu-1}} = \exp(\beta_s), \text{ which is the statin treatment effect.}$$

This result depends on the shape parameter ν having the same value for treatment vs. non-treatment to be cancelled out to get HR, and so that the proportional hazard assumption is satisfied. In our analysis, the proportional hazard assumption was satisfied since the shape parameters are generally same for models ignoring vs. accounting for statin effect.

Appendix 3: Recalibration and rescaling of eHEART risks for population health modelling

Our aim was to externally validate the eHEART tool to estimate the health impact in a general, English population. We used UK Biobank chosen due to its availability of detailed measurement at baseline, which was used to estimate a 10-year CVD risk using QRISK2 and represents a formal risk assessment, but also linked historical primary care electronic medical records necessary for prioritisation using eHEART.

However, UK Biobank participants has been shown to be healthier than the general population both in terms of risk factor levels and CVD incidence rates. Using eHEART and QRISK2 in UK Biobank without adjustments would lead to a biased distribution of 10-year risk estimated, with the distribution of risks being skewed to the right and be narrow relative to the distribution observed in the general population. To more accurately use UK Biobank for population health modelling, the distribution of 10-year risks estimated were adjusted using recalibration and rescaling.

Stage 1: recalibration

We first recalibrated both eHEART and QRISK2 to the UK Biobank population to ensure both tools were well calibrated to the population. This was done due to the different populations used to derive both risk tools and because of the different data types used in UK Biobank, with eHEART using the historical primary care records and QRISK2 using the baseline values. The methods used have been previously described.[15]

Recalibration was completed within each tool and sex, allowing the mean level of predicted risks to match what was observed. We used average risk factor levels calculated by 5-year age groups to estimate the predicted risk. For eHEART, the mean risk factor levels for continuous risk factors were calculated using the mean of the last observed values across all individuals with a primary care record before baseline, and for binary variables, the mean number of individuals with at least one positive record was used. For QRISK2, the mean value or prevalence of each risk factor at baseline was used. The observed risk was calculated using the CVD incidence rate of UK Biobank. A linear model was fit within each tool and sex to relate the observed risk (θ_{obs}) and predicted risk (θ_{pred}) estimated for each 5-year age group. (c_s):

$$\log_e(-\log_e(1-\theta_{obs,c_s})) = \beta_0 + \beta_1 \times \log_e(-\log_e(1-\theta_{pred,c_s}))$$

The estimated β_0 and β_1 were then used as scaling factors to recalibrate each individual's original 10-year risk ($\theta_{pred,i}$) to give a new recalibrated estimate $\theta_{newpred,i}$:

$$\theta_{newpred,i} = 1 - \exp(-\exp(\beta_0 + \beta_1 \times \log_e(-\log_e(1-\theta_{pred,i}))))$$

Stage 2: rescaling

Recalibrating both models using average risk factor values and incidence rates from UK Biobank resulted in the recalibrated 10-year risks to be heavily right skewed. To correct for this, we rescaled the recalibrated estimates to spread out the estimated risks to become more representative of what should be expected in the general population.

Rescaling was completed within each tool and sex. We calculated the mean recalibrated 10-year risk by age-group of UK Biobank to estimate the predicted risk. We then used mean risk factor levels calculated from the Clinical Practice Research Datalink (CPRD) between the years 2014 and 2019 within 5-year age groups to estimate the observed risk of using either tool in the general population. We then fit a linear model between the observed risk and the predicted risk using the same methods described in stage 1. The new scaling factors calculated were used to rescale each individual's recalibrated 10-year risk to give a new rescaled and recalibrated estimate.

Stage 3: Population health modelling

We created a hypothetical population of 100,000 individuals (50,000 men and women. We used data from the ONS to approximate the age structure of the general English population. We estimated the expected number of events observed in the hypothetical population using CVD incidence rates estimated from individuals with at least one primary care record of systolic blood pressure, total or HDL cholesterol or smoking status in CPRD between the years 2014 and 2019.

To estimate the number needed to screen (NNS) to prevent one CVD event, we first calculated the proportion of cases in UK Biobank that were deemed high risk (10-year risk >= 10%) by QRISK2 within 5-year age groups and sex. We used this proportion to estimate the number of CVD events identified in the hypothetical population, and applied a HR of 0.8 to model the benefits of statin initiation, with the number of events saved was defined as the difference in the number of events after statin initiation and the initial expected number of events. The NNS was calculated by taking the number of individuals assessed and dividing by the number of events saved. The NNS after using eHEART to prioritise individuals for a formal assessment was calculated in a similar manner, where the number of individuals formally assessed included

only those prioritised using eHEART. 95% confidence intervals for the NNS were estimated using the empirical bootstrap method with 1000 iterations for each age-group and sex. Age and sex specific prioritisation thresholds were chosen to optimise the false negative rate of QRISK2 in UK Biobank.

References

1. 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. 2017;357. doi:10.1136/bmj.j2099
2. Herrett E, Shah AD, Boggon R, et al. Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. BMJ. 2013;346(7909). doi:10.1136/BMJ.F2350
3. Paige E, Barrett J, Stevens D, et al. Landmark Models for Optimizing the Use of Repeated Measurements of Risk Factors in Electronic Health Records to Predict Future Disease Risk. Am J Epidemiol. 2018;187(7):1530-1538. doi:10.1093/AJE/KWY018
4. Goff DC, Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: A report of the American college of cardiology/American heart association task force on practice guidelines. Circulation. 2014;129(25 SUPPL. 1):49-73. doi:10.1161/01.CIR.0000437741.48606.98/-/DC1
5. Sharma M, Petersen I, Nazareth I, Coton SJ. An algorithm for identification and classification of individuals with type 1 and type 2 diabetes mellitus in a large primary care database. Clin Epidemiol. 2016;8:373-380. doi:10.2147/CLEP.S113415
6. Verbeke G, Fieuws S, Molenberghs G, Davidian M. The analysis of multivariate longitudinal data: a review. Stat Methods Med Res. 2014;23(1):42-49. doi:10.1177/0962280212445834
7. Schafer JL. Analysis of Incomplete Multivariate Data, 1st edition , New York: Chapman and Hall/CRC. Chapman and Hall/CRC, ed. Published online 1997:444.
8. Wood AM, Thompson SG, Kostis JB, et al. Correcting for multivariate measurement error by regression calibration in meta-analyses of epidemiological studies. Stat Med. 2009;28(7):1067-1092. doi:10.1002/SIM.3530
9. White I, Frost C, Tokunaga S. Correcting for measurement error in binary and continuous variables using replicates. Stat Med. 2001;20(22):3441-3457. doi:10.1002/SIM.908
10. Goldberger AS. Best Linear Unbiased Prediction in the Generalized Linear Regression Model. J Am Stat Assoc. 1962;57(298):369. doi:10.2307/2281645
11. Goldstein H. Multilevel Statistical Models. Published online October 29, 2010. doi:10.1002/9780470973394
12. Diggle P, Diggle P. Analysis of longitudinal data. Published online 2002:379.

13. Cook NR, Ridker P. Further insight into the cardiovascular risk calculator: the roles of statins, revascularizations, and underascertainment in the Women's Health Study. JAMA Intern Med. 2014;174(12):1964-1971. doi:10.1001/JAMAINTERNMED.2014.5336
14. Mihaylova B, Emberson J, Blackwell L, et al. The effects of lowering LDL cholesterol with statin therapy in people at low risk of vascular disease: Meta-analysis of individual data from 27 randomised trials. Lancet. 2012;380(9841):581-590. doi:10.1016/S0140-6736(12)60367-5/ATTACHMENT/8DC81FE3-D592-47AE-BD9D-E06A7C9E9817/MMC1.PDF
15. Pennells L, Kaptoge S, Wood A, et al. Equalization of four cardiovascular risk algorithms after systematic recalibration: Individual-participant meta-analysis of 86 prospective studies. Eur Heart J. 2019;40(7):621-631. doi:10.1093/eurheartj/ehy653