The uncertainty with using risk prediction models to drive clinical decision making

A thesis submitted to the University of Manchester for the degree of

Doctor of Philosophy

in the Faculty of Biology, Medicine and Health

2020

Alexander D Pate

School of Health Sciences

Division of Informatics, Imaging and Data Science

Contents

List of Tables 6

List of Figures 9

List of statistical and clinical abbreviations 13

Abstract 14

Declaration 15

Copyright statement 15

Acknowledgements 16

Preface 17

1 General Introduction 18

1.1 How clinical prediction models are used to drive treatment 18

1.2 Sources of uncertainty associated with using risk prediction models to drive treatment 19

1.3 The use of risk prediction models for the primary prevention of CVD 24

1.4 Thesis structure and motivation for chapters 30

1.5 Thesis format and author contributions 33

2 Validation of data extraction 35

2.1 Introduction 35

2.2 Methods 35

2.3 Results 39

2.4 Discussion 73

2.5 Software, programs and code lists 75

3 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 77

3.1 Abstract 77

3.2 Background 78
3.3 Methods 80
3.4 Results 86
3.5 Discussion 98
4 Impact of sample size on the stability of risk scores from clinical prediction models: a case study in cardiovascular disease 103
4.1 Abstract 103
4.2 Background 104
4.3 Methods 105
4.4 Results 108
4.5 Discussion 115
5 An assessment of the potential miscalibration of cardiovascular disease risk predictions caused by a secular trend in cardiovascular disease in England 118
5.1 Abstract 118
5.2 Background 119
5.3 Methods 120
5.4 Results 128
5.5 Discussion 134
6 The impact of statin discontinuation and restarting rates on the optimal time to initiate statins and on the number of cardiovascular events prevented 138
6.1 Abstract 138
6.2 Background 139
6.3 Methods 140
6.4 Results 144
6.5 Discussion 151
7 Impact of lowering the risk threshold for initiating statin treatment on statin prescribing 155

7.1 Abstract 155
7.2 Background 156
7.3 Methods 157
7.4 Results 158
7.5 Discussion 164
8 General discussion 168
8.1 Chapter 3 168
8.2 Chapter 4 169
8.3 Chapter 5 171
8.4 Chapter 6 172
8.5 Chapter 7 174
8.6 Overall limitations 177
8.7 Main conclusions 181
9 Reference list 182
A.3 Chapter 3 Appendices 197
A.3.1 Density and convergence plots for imputed variables 197
A.3.2 Tripod statement 212
A.3.3 Supplementary tables and figures 214
A.4 Chapter 4 Appendices 219
A.4.1 Calculation of N_{min}, minimum required sample size following published criteria, and shrinkage factor of models meeting this sample size criteria 219
A.4.2 Supplementary tables and figures 223
A.5 Chapter 5 Appendices 228
A.5.1 Supplementary tables and figures 228
A.6 Chapter 6 appendices 233
A.6.1 Derivation of statins treatment periods 233

A.6.2 Derivation of CVD transition probabilities (p_c) and how they are used to run the simulation 234

A.6.3 Calibration plots of Cox models from which transition probabilities were derived 238

A.6.4 Age stratification and extrapolation of discontinuation rates 240

A.6.5 Discontinuation and restarting rates in the cohort of statin users where treatment periods with only one prescription are removed 247

A.6.6 Results from the female analysis and the sensitivity analyses 249

A.7 Chapter 7 appendices 252

A.7.1 Density and convergence plots of imputed variables 252

A.7.2 Supplementary tables and figures 259

Total word count (excluding references and appendices): 46,010

List of Tables

Table 1.1: Key features of CVD risk prediction models recommended for use in guidelines around the world 26

Table 2.1: Comparison of definition of variables from the QRISK3 manuscript, and how this definition was interpreted. 39

Table 2.2: List of variables and source for where the code list was obtained 45

Table 2.3: Demographics at cohort entry date for the QRISK3 development cohort and the first derivation of the CPRD cohort 46

Table 2.4: Comparison of prevalence of variables where two code lists were available for derivation 49

Table 2.5: Completeness of ethnicity recording for all acceptable patients in the database, those actively registered on July 2012, and those registered after April 1st 2006. 53

Table 2.6: Distribution of the usable ethnicity codes for CPRD cohort, Mathur cohort and 2011 UK census. 54

Table 2.7: Prevalence of severe mental illness after changes 56

Table 2.8: Prevalence of family history of coronary heart disease variable after changes 57

Table 2.9: Number of type 1 and type 2 diabetes cases in the cohort using original definitions 59

Table 2.10: Number of type 1 and type 2 diabetes cases in the cohort using updated definitions 59

Table 2.11: Prevalence of diabetes variables after changes 59

Table 2.12: Prevalence of CKD variables after changes 61

Table 2.13: Prevalence of atypical anti-psychotic medication use and corticosteroid use broken down 63

Table 2.14: Demographics at cohort entry date for the QRISK3 development cohort and the final derivation of the CPRD cohort 64

Table 2.15: Summary of code provided at the GitHub page 76

Table 3.1: CVD incidence and baseline characteristics of study population 87

Table 3.2: HRs (95% CI) of fixed and random effects introduced into models C, D and E. HRs reported are all from model E. 90

Table 3.3: Performance metrics for each of the models....91
Table 3.4.1: Distribution of risks (2.5th and 97.5th percentile) of patients in the test cohort according to each model, stratified by their risk in model A, and average 95% CI for risks in model A (female cohort)....92
Table 3.4.2: Distribution of risks (2.5th and 97.5th percentile) of patients in the test cohort according to each model, stratified by their risk in model A, and average 95% CI for risks in model A (male cohort)....92
Table 3.5: Numbers and percentages of patients who cross the treatment threshold (10 year CVD risk of 10%) when using models B – F instead of model A....95
Table 3.6: Relative rates (95% CI) associated with calendar year and risk at start of statin treatment period, in Poisson models modelling CVD incidence....98
Table 4.1: Baseline characteristics of each female cohort....109
Table 4.2: Quantiles of C statistics and calibration-in-the-large of the 1000 models, for each sample size....111
Table 4.3: Probability of being classified on the opposite side of the treatment threshold by a randomly selected model of a given sample size, stratified by population derived risk ...113
Table 5.1: Baseline variables in development and validation cohorts....130
Table 5.2: Average predicted CVD risk for patients in the validation cohort before and after secular trend was introduced, using an MSM and an interval censored Cox model....132
Table 5.3: Hazard ratios of the categorical variables in the marginal structural model with and without secular trend included as a predictor variable....133
Table 6.1: Baseline table for statin cohort and CVD primary prevention cohort....145
Table 7.1: Number of people initiated on statins each year, and number of those who had a coded QRISK score in their medical record....159
Table 8.1: Number of English practices actively registered in CPRD in each calendar year, stratified by region....178
Table A.3.1: Distribution of real data and imputed data (%) for smoking status in imputation of female cohort....203
Table A.3.2: Distribution of real data and imputed data (%) for ethnicity in imputation of female cohort....204
Table A.3.3: Distribution of real data and imputed data (%) for smoking status in imputation of male cohort....210

Table A.3.4: Distribution of real data and imputed data (%) for ethnicity in imputation of male cohort.....211
Table A.3.5: Incidence rates of CVD from the CPRD cohort and comparison with QRISK3 development cohort by age band (female cohort) 214
Table A.3.6: Incidence rates of CVD from CPRD cohort and comparison with QRISK3 development cohort by age band (male cohort) 215
Table A.3.7: Comparison of hazard ratios of categorical variables from model B with QRISK3 216
Table A.3.8: Numbers and percentages of patients registered on 1st Jan 2016 who cross the treatment threshold (10-year CVD risk of 10%) when using models B – F instead of model A 217
Table A.5.1: Prevalence of statin use each year in CVD primary prevention cohort.....228
Table A.5.2: Rate of initiation of statin treatment each year in CVD primary prevention ...229
Table A.7.1: Distribution of real data and imputed data (%) for smoking status in imputation of statin initiation cohort.....258
Table A.7.2: Baseline demographics of statin cohort 259

List of Figures

Figure 2.1: Flow chart of derivation of the CPRD cohort 37
Figure 2.2: Percentage of patients with an ethnicity code recorded, and a usable ethnicity code recorded, stratified by year of registration in the CPRD cohort 53
Figure 2.3: Histogram of age 67
Figure 2.4: Histogram of body mass index (BMI) 68
Figure 2.5: Histogram of systolic blood pressure (SBP) 69
Figure 2.6: Histogram of SBP variability 70
Figure 2.7: Histogram of cholesterol/high density lipoprotein (HDL) ratio 71
Figure 2.8: Histogram of eGFR values recorded directly in the database 72
Figure 2.9: Histogram of eGFR values derived from creatinine measurements 73
Figure 3.1: Calibration plots by 10th percentile of risk for model B 90
Figure 3.2: Distribution of risks according to each model for those with risks 9 – 10% in model A 94
Figure 3.3: Percentages of patients registered 1st Jan 2016 who cross the treatment threshold when using model F, compared to model A 96
Figure 3.4: The secular trend in CVD incidence in the model derivation cohort and the statin users cohort 97
Figure 4.1: A graphical representation of the sampling process 107
Figure 4.2: Boxplots of the percentile ranges of risk for individuals across the 1000 models (female cohort) 110
Figure 4.3: Percentile range of risk for individuals across subsets of the 1000 models, defined by the C-statistic of the models (female cohort, N = 10 000) 112
Figure 4.4: Percentile ranges of risk for individuals across subsets of the 1000 models, defined by the calibration-in-the-large of the models (female cohort, N = 10 000) 114
Figure 5.1: DAG-1 121
Figure 5.2: DAG-2 123
Figure 5.3: DAG-3 124
Figure 5.4: Calibration of development (pre 2010) and validation (post 2010) cohorts 128
Figure 5.5: Predicted risks in the development and validation cohort 129
Figure 5.6: Calibration of the validation cohort when adjusting for calendar time 131

Figure 6.1: Design of the health state transition model 141
Figure 6.2: Kaplan Meier plots of the time until discontinuation and restarting statins for the first, second, third and fourth discontinuation periods, and the first, second and third restarting periods. 146
Figure 6.3: Number of cardiovascular events prevented per 100 people over the duration of follow up with different time delays in starting statins, stratified by baseline age and 10 year risk of cardiovascular disease (male).....148
Figure 6.4: Number of cardiovascular events prevented per 100 people over the duration of follow up with different time delays in starting statins, stratified by baseline age and discontinuation rate (male) 150
Figure 7.1: Statin initiation rate in each calendar year 160
Figure 7.2: Average 10-year risk of people initiated on statins in each year for primary prevention of CVD. The EHR derived risks were calculated for all patients initiated on statins, coded risks are restricted to those with coded risks in their medical record 161
Figure 7.3: Proportion of people initiated on statins each year that belong to each risk category (< 10%, 10 – 20%, > 20%). Separate plots for EHR derived risk scores (all statin initiators) and coded risk scores (restricted to those with coded risks in their medical record). 162
Figure 7.4: EHR derived risks plotted against coded risks, stratified by year of statin initiation. Blue line = 1:1 correlation 163
Figure A.3.1: BMI convergence plot for imputation of female cohort 198
Figure A.3.2: BMI density plot for imputation of female cohort 198
Figure A.3.3: SBP convergence plot for imputation of female cohort.....199
Figure A.3.4: SBP density plot for imputation of female cohort.....199
Figure A.3.5: SBP variability convergence plot for imputation of female cohort.....200
Figure A.3.6: SBP variability density plot for imputation of female cohort.....200
Figure A.3.7: Cholesterol convergence plot for imputation of female cohort 201
Figure A.3.8: Cholesterol density plot for imputation of female cohort 201
Figure A.3.9: HDL convergence plot for imputation of female cohort 202
Figure A.3.10: HDL density plot for imputation of female cohort 202
Figure A.3.11: BMI convergence plot for imputation of male cohort 205
Figure A.3.12: BMI density plot for imputation of male cohort 205

Figure A.3.13: SBP convergence plot for imputation of male cohort 206
Figure A.3.14: SBP density plot for imputation of male cohort 206
Figure A.3.15: SBP variability convergence plot for imputation of male cohort 207
Figure A.3.16: SBP variability density plot for imputation of male cohort 207
Figure A.3.17: Cholesterol convergence plot for imputation of male cohort.....208
Figure A.3.18: Cholesterol density plot for imputation of male cohort.....208
Figure A.3.19: HDL convergence plot for imputation of male cohort 209
Figure A.3.20: HDL density plot for imputation of male cohort 209
Figure A.4.1: Boxplots of the 95% percentile range of risk for individuals across the 1000 models (male cohort).....223
Figure A.4.2: Boxplots of the 95% percentile range of risk for individuals across subsets of the 1000 models, defined by the C-statistic of the models (female cohort) 224
Figure A.4.3: Boxplots of the 95% percentile range of risk for individuals across subsets of the 1000 models, defined by the C-statistic of the models (male cohort).....225
Figure A.4.4: Boxplots of the 95% percentile range of risk for individuals across subsets of the 1000 models, defined by the calibration-in-the-large of the models (female cohort)...226
Figure A.4.5: Boxplots of the 95% percentile range of risk for individuals across subsets of the 1000 models, defined by the calibration-in-the-large of the models (male cohort).....227
Figure A.5.1: Calibration of the MSM in the validation cohort, with or without adjustment for calendar time (male cohort) 230
Figure A.5.2: Calibration of the MSM in the validation cohort, with or without adjustment for calendar time (female cohort) 231
Figure A.5.3: Calibration of the interval censored Cox model in the validation cohort, with or without adjustment for calendar time (male cohort) 231
Figure A.5.4: Calibration of the interval censored Cox model in the validation cohort, with or without adjustment for calendar time (female cohort) 232
Figure A.6.1: Calibration of female Cox model used to derive CVD transition probabilities 239
Figure A.6.2: Calibration of female Cox model used to derive CVD transition probabilities 240
Figure A.6.3: Kaplan Meier plots for patients during first treatment period, stratified by age 241
Figure A.6.4: Age stratified discontinuation rates derived from the Cox proportional hazards model 242

Figure A.6.5: Extrapolation of the discontinuation rates derived from CPRD for the first treatment period, stratified by age, extrapolation made from year 13 onwards.....244
Figure A.6.6: Extrapolation of the discontinuation rates derived from CPRD for the second and third treatment periods, extrapolation made from year 10 and year 8 onwards. 245
Figure A.6.7: Extrapolation of the restarting rates derived from CPRD for the first and second treatment period. Extrapolation made from 10 years onwards.....246
Figure A.6.8: Kaplan Meier plots of the time until discontinuing statins for the first, second, third and fourth discontinuation, single prescription treatment periods removed 247
Figure A.6.9: Kaplan Meier plots of the time until restarting statins for the first, second and third time, single prescription treatment periods removed 248
Figure A.6.10: Number of cardiovascular events prevented over the duration of follow up with different time delays in starting statins, stratified by baseline age and 10 year CVD risk using the discontinuation rates as observed in the statin cohort (female cohort).....250
Figure A.6.11: Number of cardiovascular events prevented over the duration of follow up with different time delays in starting statins, stratified by baseline age and discontinuation rate (female cohort).....251
Figure A.7.1: BMI convergence plot for imputation of statin initiation cohort.....253
Figure A.7.2: BMI density plot for imputation of statin initiation cohort.....253
Figure A.7.3: SBP convergence plot for imputation of statin initiation cohort 254
Figure A.7.4: SBP density plot for imputation of statin initiation cohort 254
Figure A.7.5: SBP variability convergence plot for imputation of statin initiation cohort....255
Figure A.7.6: SBP variability density plot for imputation of statin initiation cohort 255
Figure A.7.7: Cholesterol convergence plot for imputation of statin initiation cohort.....256
Figure A.7.8: Cholesterol density plot for imputation of statin initiation cohort.....256
Figure A.7.9: HDL convergence plot for imputation of statin initiation cohort.....257
Figure A.7.10: HDL density plot for imputation of statin initiation cohort.....257

List of statistical and clinical abbreviations

ACC	American College of Cardiology
AHA	American Heart Association
BMI	Body mass index
BNF	British National Formulary
CPRD	Clinical Practice Research Datalink
CG181	Guideline introduced by National Institute of Health and Care Excellence reducing the risk at which patients became eligible for initiation of statin treatment
CKD	Chronic Kidney Disease
CVD	Cardiovascular disease
DAG	Directed acyclic graph
EAS	European Atherosclerosis Society
EHR	Electronic health record
ESC	European Society of Cardiology
eGFR	Estimated Glomerular filtration rate
HDL	High-density lipoprotein
HES	Hospital Episode Statistics
HR	Hazard ratio
MSM	Marginal Structural Model
NICE	National Institute for Health and Care Excellence
ONS	Office for National Statistics
QOF	Quality and outcomes framework
SBP	Systolic blood pressure
SD	Standard deviation
SLE	Systemic lupus erythematosus
THIN	The Health Improvement network

Abstract

Risk prediction models have become embedded into the health system. They are used to guide clinical decision making in a variety of settings: risk of death following surgery (should we operate?), diagnostic models for cancer (should we screen?), or the probability of having a clinical event over a certain time period (should we take preventative measures?). Despite clear guidelines on the development and reporting of models, features of models developed for the same purpose often differ. Furthermore, in the field of cardiovascular disease (CVD), risk thresholds for initiating statin therapy vary across England, Scotland, the US and Europe, despite a large body of evidence on when treatment becomes cost effective. This results in uncertainty when using these models to guide treatment for a patient, as using different models or clinical guidelines may result in a different decision for an individual. This thesis focused on identifying sources of uncertainty associated with both parts of this process, generating risk predictions, and making clinical decisions based on these risk predictions. Case studies consider the primary prevention of CVD, which was chosen due to the high incidence of CVD, the saturation of CVD risk prediction models in the literature, and the fierce debate over the last 10 years about the best approach for the primary prevention of CVD.

Chapter 3 found the impact of covariate selection on the risks of individuals to be small, apart from a large secular trend. Chapter 4 identified high levels of instability in risk scores when using sample sizes of widely used models, and when derived from recently published sample size formula. Chapter 5 found that the secular trend in CVD (identified in Chapter 3) caused over prediction of risks for patients in the present day and was not driven by increasing statin use. Chapter 6 highlighted that a small number of extra CVD events could be prevented by delaying statin initiation to when patients are at risks higher than 10% (given the high statin discontinuation rates identified in practice). Chapter 7 showed that the reduction of the risk threshold for initiating statins for the primary prevention of CVD, from a 10-year CVD risk of 20% to 10%, had little impact on clinical practice in England. This finding is contrary to current evidence.

The findings in this thesis are a mix of methodological findings of interest to those developing models, and those that have a direct impact on the prevention of CVD in the UK.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright statement

I. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

II. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

III. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

IV. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=24420), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in The University's policy on Presentation of Theses

Acknowledgements

This PhD would not have been possible without the support of many others I would like to acknowledge.

Firstly thank you to my supervisors, Prof Tjeerd van Staa and Prof Richard Emsley. I could not have wished for more supportive supervisors, you always made time for me no matter the question. Your guidance has been essential to me completing this thesis, but also developing as a person and navigating the world of academia. It was a pleasure working with you both, and I hope we can do so again in the future.

Secondly, I'd like to thank my family, Alison, John, Sally and Clare, for continuing to support me wherever I go.

Thirdly, I would like to thank all the staff and students in the Centre for Health Informatics. Whether it was giving your time to offer me advice and tutelage for no gain of your own, or for keeping me sane on a day to day basis, it's been a pleasure working alongside all of you.

I would also like to thank the MRC for funding this project, and the MRC-DTP team for their support. A thank you to CPRD for collecting and collating the data I used in this project, Prof Darren Ashcroft and his team for making it available for use at the University, and all the people who contribute their data to this database. Finally, a thank you to all the team who work on and maintain the iCSF (incline), on which all my analyses were run. This really would not have been possible without you.

Alex

Preface

Candidate degrees

2009 – 2013: MSci Mathematics (First Class Honours), University of Bristol

Publications

Chapter 3 of this thesis was published in *BMC Medicine* on the 17th July 2019.[1] Chapter 6 was published in *Pharmacoepidemiology and Drug Safety* on the 11th May 2020.[2] Chapter 7 was accepted for publication in the British Journal of General Practice on the 17th May 2020. Chapter 4 and was under review in a journal at the time this thesis was submitted. Chapter 5 had not been submitted for publication at the time this thesis was submitted.

1. Pate A, Emsley R, Ashcroft DM, et al. 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: 134.

2. Pate A, Elliott RA, Thompson A, et al. The impact of statin discontinuation and restarting rates on the optimal time to initiate statins and on the number of cardiovascular events prevented. *Pharmacoepidemiol Drug Saf* 2020; 29: 644–652.

1 General Introduction

1.1 How clinical prediction models are used to drive treatment

Risk scores from clinical prediction models are used to drive treatment in many disease areas and settings.[3–11] This may be the risk of having a clinical event at some point in the future (prognostic) or the risk of currently having an undiagnosed medical condition (diagnostic). These models are often developed on data from cohort studies, where data may be collected in a pre-determined cohort of individuals, or from routinely collected data sources such as electronic health records or registry data. Depending on the outcome one is trying to predict, a range of statistical algorithms can be used to predict either the outcome itself, or probability of the outcome occurring (i.e. for binary outcomes). This thesis will focus on the latter, the probabilities known more commonly as risks. Risk scores can be generated through a wide range of statistical algorithms, ranging from regression modelling techniques to machine learning. Based on the risk score derived for a given patient, a certain course of preventive action or screening may be taken. There are therefore two key aspects of this process: 1) Calculating the risk score of a patient on which clinical decisions will be based, and 2) Deciding what action to take for a patient with a given risk score. This thesis will focus on identifying areas of uncertainty associated with these two parts of the treatment process.

Uncertainty when deriving a risk score for a patient is inevitable, as there is no true underlying risk to calculate. A patient will either have the event (or underlying health condition), or they won't. This means their 'true' underlying risk, if it exists, is either a 0 or a 1. At the prediction stage we do not know this underlying truth and assign probabilities that a patient will have the event or not, a risk score. There are often many equally valid methods that can be used to derive such risk scores, resulting in many possible risk scores for an individual patient. Depending on the level of agreement between these risk scores, a clinician may have more or less confidence in using a particular risk prediction model as the primary tool for guiding treatment. This thesis will use the agreement or discordance between risk scores for an individual patient derived from different models as a key metric, explore what the major drivers are behind this, and whether the discordance matters in practice.

The uncertainty when deciding what action to take for a given patient is arguably less clear cut. Historically, a clinician has always decided what action to take for a patient presenting with a given set of symptoms/comorbidities. Risk scores act as another 'symptom' which the clinician and patient can use to help make a decision. However with the current abundance of clinical prediction models, and studies to indicate at what risk threshold a treatment becomes cost-effective, guidelines are being created to advise clinicians as to the threshold at which they should initiate or offer treatment. Given the many components that go into creating these guidelines there are many potential sources of uncertainty associated with the development of these guidelines. However, uncertainty may also stem from patient-centered factors a clinician must also consider, and how they interpret and implement these guidelines, rather than in their development.

1.2 Sources of uncertainty associated with using risk prediction models to drive treatment

This section is split into two halves, focusing on the two different aspects of the treatment process, the calculation of risk scores, and the clinical decision making process based on the risk scores.

1.2.1 Sources of uncertainty when calculating risk scores

When talking about sources of uncertainty in calculating a risk score, I am referring to things in the model development process that may alter a risk score for a given individual. These can be broadly split into two categories: 1) model features (or modelling decisions), and 2) sampling variation. I make this distinction as both have the same effect of altering an individual's risk score, but through different guises. Different features of a model result in the estimation of a different quantity, or a different estimator of the same quantity. Ultimately a different process is undertaken to develop the model, and so a change in risk for a given individual is somewhat expected. Uncertainty driven by sampling variation is a purely statistical artefact, the exact same steps are taken to develop the model but random variability in the sampling process means the model, and resulting risk scores may change.

Model features

There are many decisions to made when developing a risk prediction model (how to define the cohort, statistical methods for analysing the data, how to deal with missing data, variable selection, etc). A cohort may be selected to best match the target population for the treatment of interest; a statistical method may be chosen based on the outcome of interest and the assumptions that must be made; an imputation method will be chosen to best suit the relationship between variables in the dataset; and predictor variables may be selected to include key causal predictors or to maximise model performance. All the key factors that should be considered are mentioned in the TRIPOD reporting guidelines.[12] When developing a model there may be multiple equally correct decisions with advantages and drawbacks, resulting in multiple viable models and multiple viable risk scores for an individual. From a statistical point of view, there is no issue with this variability, as one would expect different modelling decisions to affect the risk scores of an individual. From a patient's perspective this is more problematic as it could alter their treatment pathway.[13,14] Much of the literature surrounding this uncertainty induced by modelling decisions falls under the 'reference class problem'.

Reference class problem

The reference class problem is a well-documented issue caused by using group based risk scores for individuals.[15] The concept is that an individual belongs to many different groups, or reference classes, with each group having a different collective risk. When using group based risk scores for an individual, the individual therefore has multiple risk scores simultaneously depending on which reference class you assign them to. The issue was first noted by John Venn in 1866,[16] but has become more prominent since the increase in use of clinical prediction models, which naturally produce risk scores based on the subgroups defined by the variables included in the model.

A review of studies that elicited discordance in individual risk estimates was carried out by Stern in 2010[17] who discusses 9 different studies, including the somewhat seminal paper by Lemeshow et al.[13] Six of these analyse models predicting mortality in intensive care units or post-surgery settings, two for breast cancer screenings and one for cardiovascular disease (CVD). The authors stated that such discordance is rarely evaluated, and in many of these

studies the models that were compared did not have comparable performance, reducing their relevance. Other findings were that models with similar predictors suffer from less discordance, and that "*discordance of individual risk estimates does not weaken the economic rationale for their use in allocation of resources, but it does weaken the clinical rationale*". The problem is eloquently discussed by Stern[18] and Kent,[19] who both support the notion that disagreement between models does not undermine their use, but does mean the interpretation of risk scores is complex and caveated. However there has not been much primary research into quantifying the extent of this problem in various disease areas since then. Of the 10 citations of the Stern[18] paper, and 33 citations of the Kent[19] paper, only two explicitly compare the risk scores of individuals across models.[20,21] Since 2010 (the year of Stern's review) there have been 58 citations of the Lemeshow et al.[13] study, but only one that directly compared discordance in risk estimates from multiple models.[22] While other studies exist,[23] this phenomenon may be discussed without the term 'reference class problem' making them difficult to identify. There is a similar stream of work looking at the impact of reference class problem on heterogeneous treatment effects,[24–26] but this thesis focuses on risk predictions.

Strictly speaking the reference class problem refers to the variability in an individual's risk resulting from a patient belonging to multiple reference classes (defined by variable selection in a risk prediction model); but it could be extended to encompass how an individual's risk changes when other model features are altered. Consider altering the statistical model itself, the time period of the data used to derive the model, or the inclusion and exclusion criteria used to define the cohort. While a patient's reference class may remain the same (assuming the same variables are kept in the model), the estimator of their risk is inherently different. It is likely this drives a large portion of discordance between models referenced in the literature and ultimately the result is the same as the reference class problem; an individual is assigned competing risk scores with no clear way to choose between them.

Sampling variation

Quantifying sampling variation is the cornerstone of statistics. An estimate of any population level parameter should be provided with a confidence interval which quantifies the uncertainty associated with that estimate. This is because the estimate is only ever calculated in a sub sample of the overarching target population. Depending on the size of that sub

sample, there may be more or less uncertainty associated with that estimate. For example in the CONSORT[27] (randomised controlled trials), STROBE[28] (observational studies) and PRISMA[29] (systematic reviews) reporting guidelines, confidence intervals of effect estimates are expected. In scenarios where the variance of an estimator does not have a closed form solution, bootstrap resampling methods can be used to derive the confidence interval empirically.[30]

It is therefore surprising that most risk prediction calculators do not provide confidence intervals associated with risk scores. None of the QSCORES developed from the QResearch database[31] (e.g. QRisk, QKidney, QCancer, QFracture, QDiabetes) provide a confidence interval associated with the risk scores when using their online calculators, and this is common practice. There is nothing in the TRIPOD[12] statement that indicates confidence intervals should be provided alongside risk scores (despite confidence intervals for odds ratios and hazard ratios being expected). This may be because an individual's risk is not viewed as a population level parameter and therefore cannot have an associated confidence interval. However, the risk calculated is the risk of the subgroup which that patient belongs to, which is a population level parameter and is calculated with a degree of statistical uncertainty. This is not an issue for the above mentioned models as they are developed on large routinely collected datasets; however this is not the case for all models. Systematic reviews of CVD,[32] chronic kidney disease[33] and type 2 diabetes[34] risk prediction models show large variability in the number of patients included for model development. If models are developed on a small number of individuals, it may be important to report the statistical uncertainty associated with the risk scores.

Given that historically there has been very little research on sample sizes for risk prediction models, underpowered models may be more commonplace than expected. There has recently been some progress in this area.[35,36] However the work focuses on overfitting, with a small section on the precision of effect estimates, which is what would result in precise (low degree of statistical uncertainty) risk scores for individuals.

1.2.2 Sources of uncertainty in the clinical decision making process

There may be debate over what the correct threshold is to initiate treatment for a patient. This may be preventive treatment, or diagnostic tests to identify the presence of a disease. This has happened in breast cancer screening[37] where the frequency of screenings was questioned in the 40 – 49 age group, and recently work has been published suggesting that current blood pressure based thresholds for initiating blood pressure lowering treatment may be inadequate.[38] There has also been a large debate over the risk at which statins should be initiated in patients for primary prevention of CVD, which is discussed in detail in section 1.3.3.

The types of analysis carried out to determine a treatment threshold are often a benefit-harm analysis,[39] to see whether the benefit of a drug outweighs the negative effects, or a cost effectiveness analysis,[40] to assess the cost of the gain in health, possibly relative to another intervention. One area of uncertainty with both of these types of analysis is quantifying the possible adverse events caused by a drug. Adverse events may not be well recorded in trials, or be underpowered for formal hypothesis testing.[41,42] A Cochrane review[43] found that the type of questioning used may affect the ability to pick up adverse events, and another review[44] found that systematic reviews of adverse events may compound poor reporting of adverse events data in primary studies.

Beyond the uncertainty induced by the limitations of scientific evidence, there may also be social factors to consider in how developed guidelines are interpreted and used. This was the case in England when the National Institute for Health and Care Excellence (NICE) reduced the threshold for initiating statin treatment from a 10-year risk of incident CVD of 20% to 10%.[45,46] While they found it to be cost effective, this was largely driven by the low cost of statins, and there was opposition from clinicians. A group of leading doctors wrote a letter to NICE citing six major concerns: “the medicalisation of millions of healthy people; conflicting levels of adverse events; hidden data; industry bias; loss of professional confidence; and conflicts of interest”.[47] Factors such as these, and the opinion of each individual clinician may cause differential implementation of guidelines for similar subgroups of the population.

1.3 The use of risk prediction models for the primary prevention of CVD

CVD was chosen as the disease area in which to explore the themes outlined in section 1.2. The use of risk prediction models is common in the primary prevention of CVD. The disease itself presents a major health risk to most individuals and results in large burdens on health systems across the world. Also, access had been gained to the Clinical Practice Research Datalink (CPRD),[48] linked with Hospital Episode Statistics[49] (HES) and Office for National Statistics[50] (ONS) data. CVD outcomes are well recorded across all three databases, and most risk factors are well recorded in primary care enabling the development of high quality CVD risk prediction models.

1.3.1 Cardiovascular disease (CVD)

CVD is a collective term for coronary heart disease, strokes and transient ischaemic attacks, peripheral arterial disease and aortic disease, all conditions which affect the heart or blood stream.[51] The British Heart foundation estimated that 7.4 million people are living with CVD in the UK, and it accounts for 27% of all deaths.[52] Bhatnagar et al. summarised the trends in the epidemiology of CVD in the UK from 1979 to 2013.[53] The reported total CVD mortality declined by 68% from 1980 to 2013. Given that this thesis considers models for primary prevention of CVD, the CVD incidence rates were of most interest. They do not report directly on incidence rates, but state the finished consultant episodes (FCEs) outcome can be used as a proxy for incidence. Trends in FCEs for coronary heart disease and stroke both showed an overall decrease between 2005 and 2014. This is in agreement with other findings of a 31% (female) and 33% (male) reduction in incidence of myocardial infarction between 2002 and 2010,[54] a 30% decrease in the incidence of stroke between 1998 and 2008,[55] and a 29% decrease in the incidence of stroke from 1981 – 1984 to 2001 – 2004.[56]

Despite these reductions incidence rates remain unsatisfactory. In a report from Heart UK, after cross party discussions with NHS England, Public Health England and industry partners, they highlight that the high number of individuals living with CVD not only presents a major public health risk, but also a large burden on the National Health Service.[57] However, there is huge opportunity to reduce this number as most CVD cases are preventable, through lifestyle changes which impact blood pressure, smoking, cholesterol, exercise, diet and medical interventions. They recommend prioritising the prevention of CVD in primary, secondary and

tertiary actions. This thesis focuses on the use of risk prediction models to help guide treatment for primary prevention of CVD.

1.3.2 CVD risk prediction models

Risk prediction is extremely common in the field of CVD. One of the first cases of multivariate logistic regression was its use in analysing data from the Framingham heart study,[58] a cohort study designed to assess risk factors for coronary heart disease.[59] Since then many more models have been developed. A systematic review of risk prediction models for CVD in the general population carried out by Damen et al.[32] found 363 different models published by June 2013. It was of interest for this thesis to review models that are currently being used; to understand what modelling choices have been made and what the current best practice is. However, the authors of that systematic review found that the usefulness of the majority of developed models was unclear (methodological shortcomings, no external validation, incomplete presentation). Instead, I reviewed the models that are currently recommended in guidelines in various countries across the world and reported on the key model features of each.

In the US and worldwide, historically The Framingham Risk Score[6] has been the most commonly used algorithm. However recently the pooled cohort equations have been recommended by the American College of Cardiology and American Heart Association (ACC/AHA).[60,61] In England QRISK3[5] is currently recommended by NICE,[62] in Scotland ASSIGN[63] is suggested by the Scottish Intercollegiate Guidelines Network,[64] in Europe the SCORE equation[65] is recommended by the European Society of Cardiology and European Atherosclerosis Society (ESC/EAS),[66,67] and in New Zealand the PREDICT[68] equations are recommended by the Ministry of Health NZ.[69] Also, recently the GLOBORISK[70] equations have been released which are designed to be recalibrated for use in various countries across the world. While the primary difference in these algorithms is the country in which the data used to derive the algorithms comes from, there is also a wide range of different methodological features associated with these models. The key features of each model are presented in Table 1.1. This provides an updated version to the third table from the review of models for assessment of CVD risk by Cooney et al.,[71] from 2009. The Reynolds[72,73] and PROCAM[74] models have been omitted as they are not recommended in any national guidelines.

Table 1.1: *Key features of CVD risk prediction models recommended for use in guidelines around the world*

	Dates	Data, population and sample type	Data source for outcome	Sample size (F/M)	Statistical methods	Length of prediction	Age range	Variables	Competing risks
ASSIGN[63]	1984 – 1987, 1989, 1995	SHHEC prospective study: sample from general population in Scotland	Scottish record linkage scheme listed hospital admissions and deaths, ICD9/10 codes only	6757/ 6450	Cox	10 years	30 - 74	Sex, age, total cholesterol, HDL cholesterol, SBP, Smoking—no. of cigarettes, diabetes, area-based index of deprivation, family history of CHD	No
QRISK3[5]	1998	QResearch: Health records of general practice attendees	Primary care, secondary care (ICD9/10 codes), mortality data (ICD9/10 codes)	4.02 mil/ 3.87 mil	Cox	10 years	25 - 84	Age, Ethnicity, Deprivation status, SBP, BMI, Total cholesterol/HDL ratio, smoking status, premature family history of CVD, diabetes, treated hypertension, rheumatoid arthritis, atrial fibrillation, CKD, SBP variability, migraine, corticosteroid use, systematic lupus erythematosus, atypical anti-psychotic medication, severe mental illness, HIV or AIDS, erectile dysfunction	No
SCORE[65]	1972 - 1991	12 pooled prospective studies from 11 European countries: samples from general population and occupational cohorts	12 cohort studies, events identified through ICD9/10 codes. Does a fatal and fatal + non fatal analysis	88080/ 117098	Cox and Weibull	10 years	40 - 65	Sex, age, total cholesterol or total cholesterol/ HDL cholesterol ratio, SBP, smoking status. Different versions for use in high- and low-risk countries	No
Framingham[6]	1968 - 1971 1971 - 1975 1984 - 1987	Prospective studies: Framingham Heart Study and Framingham Offspring Study: General population, Framingham and U.S Volunteers	Patients followed up through physical examinations at study clinic, hospitalization records and communication with personal physicians	4522/ 3969	Cox	10 years	30 - 75	Sex, age, total cholesterol, HDL cholesterol, SBP, smoking status, diabetes, hypertensive treatment	No

Models	Dates	Data, population and sample type	Data source for outcome	Sample size (F/M)	Statistical methods	Length of prediction	Age range	Variables	Competing risks
Pooled cohort equations[60]	1968 - 1971 1971 - 1975 1984 - 1987 1985 - 1986 1987 - 1989	Several NHLBI-sponsored cohort studies, combined with applicable data from Framingham: samples from general population	Patients followed up through physical examinations at study clinic, hospitalization records and communication with personal physicians	White: 11240/ 9098 African American: 2641/ 1647	Cox	10 years	40 - 79	Age, sex, Ethnicity, SBP, total cholesterol and HDL cholesterol, smoking status, diabetes	No
Globorisk[70]	1987–1989, 1989–1993, 1948–1951, 1971, 1965-1968, 1973-1976, 1965, 1991	Eight prospective cohort studies: samples from general population	Event data as defined by each cohort's event adjudication committee ICD9/10 codes Does a fatal and fatal + non fatal analysis	44409/ 21151	Cox	10 years	40 - 80	Age, sex, SBP, total cholesterol, diabetes, smoking status Model recalibrated separately for: China, Czech Republic, Denmark, England, Iran, Japan, Malawi, Mexico, South Korea, Spain and USE	No
PREDICT (NZ)[68]	2002	PREDICT study: individuals automatically recruited when clinician completes cardiovascular risk assessment	ICD10 codes	175699/ 226053	Cox	5 years	F: 55-74 M: 45-74	Age, Ethnicity, Deprivation status, premature family history of CVD, smoking status, diabetes, atrial fibrillation, SBP, total cholesterol/HDL ratio, on blood pressure lowering medication?, on lipid lowering medication?, on antithrombotic medication?	No

*SBP, systolic blood pressure; HDL, high density lipoprotein; BMI, body mass index; hsCRP, high-sensitivity C-reactive protein; HbA1c, haemoglobin A1c.

There are some areas which are very consistent across the models. A Cox[75] model is by far the most common method for data analysis. None of the models implement a competing risks analysis, despite the fact these models will be used frequently in older populations where competing risks are most prevalent.[76–78] It is possible this is because clinicians are expected to consider the competing risk themselves, rather than explicitly trying to model it. There is a core set of variables included in all the models (age, gender, smoking status, systolic blood pressure and some form of cholesterol measurement). Family history of CVD, diabetic status and anti-hypertensive treatment are common but are not in all models. QRISK3 contains far more variables than any other calculator. All models produce 10-year risk scores except PREDICT in New Zealand, which produces a 5-year risk score.

The models differ significantly in their sample size, which considering the arguments outlined in section 1.2.1 indicates some risk scores will have a much higher level of statistical uncertainty than others. There are also significant differences in how the composite outcome of CVD is defined. The baseline event rate across studies may therefore be highly variable (on top of the fact they are developed in different populations). There is some variability in the limits of the age range considered, but all models consider individuals aged 40 – 65. There is also a lot of variability in the start of data collection. PREDICT and QRISK are the most recent (2002 and 1998 respectively), with many others using data from the 1970's and 1980's.

1.3.3 Statin treatment thresholds around the world and debate in the literature

There is a wide range of recommended thresholds at which patients become eligible for statin treatment in different countries for the primary prevention of CVD. England[45] and the US[60,61] have the lowest thresholds at a 10-year CVD risk of 10% and 7.5% respectively. The ESC/EAS[66,67] recommended threshold is a 10-year risk of a fatal CV-event of 5%, which is equates approximately to a 15% risk of any CV-event,[66] and in Scotland the threshold is 20% for asymptomatic patients.[64] In New Zealand a 5-year risk score is used instead and the threshold is 5%.[69] It should be noted that all guidelines recommend the use of risk scores alongside other contextual information such as the patient's cholesterol, blood pressure, diet and exercise levels in a conversation with patient, to help decide the best lifestyle alteration for that patient (one of which may be statin treatment).

Historically, statins were prescribed only to higher risk patients, and the lower thresholds in the US and England have only been brought in since 2013/2014. These decisions were made during a high profile debate over what the optimal risk threshold to initiate statins for the primary prevention of CVD should be. Below I summarise the arguments on each side of the debate.

Evidence for

The evidence for these thresholds is based mostly on trial data. The ACC/AHA[79,80] and ESC/EAS[66,67] reviewed a series of high profile statin trials and whether at specific risk thresholds the clinical benefits from statin therapy outweigh the potential harm from adverse events. A Cochrane review from 2013,[81] a systematic review for the US Preventive Services Task Force[82] and a high profile review in 2016[83] summarise the results from these trials and are consistent in their results. NICE ran their own cost effectiveness simulations[84] to work out at what threshold statins are deemed cost effective, using trial data from a clinical review to inform statin efficacy in the simulations. A variety of cost effectiveness thresholds have also been considered for a US population,[85] the results supporting the recommended threshold of the ACC/AHA. The Scottish guidelines[64] acknowledge the lowered threshold in NICE guidance from 2014. However, they do not believe the impact of increased workload on the healthcare system has been properly evaluated, and propose further research into age dependent thresholds before making any changes. The Ministry of Health in NZ state "evidence from a meta-analysis of RCT's that benefits are apparent in all risk groups, although the benefit is very small when five-year risk is below 5 percent", but do not reference the study.[69]

Evidence against

All the evidence quoted by the guideline developers is in general agreement; however there has been some conflicting evidence. Abramson et al.[86,87] argue that the results from the 2013 Cochrane review[81] are driven mostly by a meta-analysis published in 2012[88] (results from a previous Cochrane review[89] by the same authors two years prior without this meta-analysis included were consistent with a 20% threshold). They argue a large proportion of the outcomes from trials included in this meta-analysis are softer outcomes such as coronary revascularisation procedures, and when considering all-cause mortality statins do not have a significant effect. Furthermore, they state the meta-analysis did not consider the effect of

statins on serious adverse events properly. Redberg et al.[90] argued that results from the US Preventive Services Task Force's systematic review included studies with patients taking statins for secondary prevention who were likely to have an increased baseline risk (and bigger benefit from statins), despite this review being used to make recommendations for primary prevention.[91] They also note authors did not have access to patient level data, increasing potential for bias. Yebyo et al.[92] present data indicating that statins only provide a net benefit with respect to potential harm at higher risks than is recommended in most guidelines.

Impact assessment of lowering threshold in England

In England NICE used to recommend prescribing at a 20% threshold. In June 2014, the threshold was lowered to 10%.[45] This sparked a lot of debate in the UK and led to widespread media coverage,[93–96] as well as opposition from leading doctors.[47] Despite all the discussion in the British media and academic literature after NICE lowered the treatment threshold, to my knowledge only one study has evaluated the impact of this guideline change in practice.[97] They found a large drop in the average risk of patients initiated on statins after the guideline change. However the analysis was restricted to patients who had a coded QRISK score recorded in their medical record, of which they state only 72.9% of statin initiators since 2012 did.

1.4 Thesis structure and motivation for chapters

This thesis broadly focuses on two aims:

1) To quantify the main sources of uncertainty associated with the calculation of risk predictions

2) To quantify the main sources of uncertainty associated with the clinical decision making process based on the risk predictions

Each chapter helps answer one of those aims, and is motivated by a combination of output from the literature reviewed in sections 1.2 and 1.3, and findings from subsequent chapters. All chapters consider CVD risk prediction and statin treatment, and the generalisability of each to the wider risk prediction community beyond CVD is variable.

Chapter 2 is a data profile. It describes how the dataset for Chapter 3 was derived and reports on the steps taken to validate the extraction of this dataset. While slightly different cohorts were used in the subsequent chapters, the same variables (or subsets of) were derived for each of those cohorts and the code for variable extraction was re-used between studies. Chapter 2 therefore acts as a validation on the extraction of the datasets used in Chapters 4 – 7 as well. It has been included to provide confidence that a structured and careful approach was taken to derive the datasets used throughout this thesis, and therefore it doesn't focus on the aims outlined above. Chapters 3 – 5 all focus on different aspects of aim 1, while chapters 6 and 7 focus on different aspects of aim 2.

Chapter 3 compares the risk scores for individuals across a range of CVD risk prediction models which progressively include more information about the patient (variables). This acts as a way to measure the impact of the reference class problem on CVD risk scores. Section 1.2.1 highlighted that the reference class problem is a well understood issue, but there is little primary research summarising its actual impact in a variety of scenarios. Many of the published studies include models developed on small sample sizes, and the reference class problem reported may be conflated with sampling variation. Also, in a certain disease areas there may be some very strong causal predictors meaning there is less discordance between models. The only studies looking at this in CVD risk prediction compare risks from models that not only have different predictors, but are developed on different cohorts, where the outcome and predictors may have been defined differently.[23,98] This chapter considers a scenario where models are developed on the same cohort and variables are defined in the same way. This means only the impact of patient's risks being conditioned on different variables is assessed, without the impact of underlying differences in databases, variable definitions or sampling variation. If only a small effect of the reference class problem was found, then in the context of Lemeshow et al.,[13] this would support the 'clinical rationale' for using any CVD risk prediction model defined in a similar way.

Chapter 4 evaluates the impact of sampling variation on CVD risk scores in models developed using different sample sizes. None of the CVD risk prediction models reviewed in Table 1.1 provide a confidence interval or measure of statistical uncertainty alongside the risk score. Depending on the sample size used in model development, the level of uncertainty associated with these scores will be variable, and there is a large level of variation in the sample size of

these models. This is relevant in the current methodological landscape as sample size formula for risk prediction models are now being developed,[35,36] however the stability of risk scores from models that meet these criteria has not been assessed. This study evaluates the statistical uncertainty (precision) of risk scores generated from models that meet these sample size criteria, and also ones that match the sample size of widely used CVD risk prediction models. How sample size may affect model performance through the stability of risk scores is also considered.

Chapter 5 looks at the secular trend in CVD incidence that is present in England and evaluates whether this is being driven by increasing statin use during follow up. This secular trend was identified in the literature, however the idea to focus on this in a study arose from the results of Chapter 3, which finds calendar time to have a large impact on the predicted risks of individuals. The potential miscalibration caused by choosing not to model the secular trend is assessed, and also whether it should be modelled. In particular, if the secular trend is driven by an increase in statin use, it should not be incorporated into a risk score which is used to decide whether a patient should receive statins or not. A marginal structural model is used to determine the presence of the secular trend after adjusting for statin use during follow up.

Chapter 6 explores the relationship between statin discontinuation and the optimal time to initiate statin therapy. The work is based on the hypothesis that if discontinuation rates are high, it may be beneficial to delay initiation to ensure patients receive treatment when they are at a higher risk and will benefit most from it. Section 1.3.3 presented a lot of research that has gone into identifying the optimal threshold at which to initiate statin therapy, all of which is based around trial data, benefit-harm analyses and cost effectiveness studies. While effect estimates from trials do account for poor adherence and discontinuation, none of the research considers the effect that statin discontinuation may have on the optimal time at which to initiate treatment. In this chapter, our data is used to calculate discontinuation rates for the first, second and third time taking statins, which has not been done before. This data is then used in a simulation to answer when the optimal time to initiate statin therapy is for a patient with a given risk profile, given the discontinuation rates seen in practice.

Chapter 7 evaluates the impact of NICE reducing the recommended threshold for initiating statin therapy (for primary prevention of CVD) from 20% to 10% on prescribing behaviour in England. As outlined in section 1.3.3, there was a lot of discussion in the media as well as in

academic and clinical settings over the lowering of the risk threshold. However, since 2014 only one study has evaluated the impact this had on the people being initiated on statins.[97] It is imperative for NICE to know if their policies are being implemented in practice, given the time and resources that are put into developing them. This chapter builds on the analysis carried out by Finnikin et al.[97] and addresses some limitations in their work, whilst also providing a replication of their analysis in CPRD (original analysis carried out in THIN).

Chapter 8 brings together the findings from each study under a common framework and links them back to the original aims of the thesis, considers the broader implications of this work, expands on the further work suggestions where necessary and discusses limitations of the thesis as a whole.

1.5 Thesis format and author contributions

This section is recommended by the University of Manchester Presentation of Thesis Policy.

This thesis is presented in journal format (also known as alternative format), a series of papers in manuscript format, which have either been accepted in or are under review in peer reviewed journals. A number of appendices have been included, to provide extra clarity over methods used, and present results from a variety of extra analyses, all of which are referenced in the main body of the thesis.

Data acquisition was carried out by **AP**, the application was critically reviewed for important intellectual content by TVS, RE and DA (collaborator named in Chapter 3). The author contributions of chapters written for publication are as follows:

Chapter 3: **AP**, TVS, RE and DA designed the study, **AP** conducted the analysis and interpreted the results in discussion with TVS, RE, DA and BB. **AP** wrote the initial draft of the manuscript, which was then critically reviewed for important intellectual content by all authors.

Chapter 4: **AP** designed the study with support from TVS, RE, MS and GM, **AP** conducted the analysis and interpreted the results in discussion with TVS, RE, MS and GM. **AP** wrote the initial draft of the manuscript, which was then critically reviewed for important intellectual content by all authors.

Chapter 5: **AP** designed the study with support from TVS and RE. **AP** conducted the analysis and interpreted the results in discussion with TVS and RE. **AP** wrote the initial draft of the manuscript, which was then critically reviewed for important intellectual content by all authors.

Chapter 6: **AP** and TVS designed the study with support from RAE, AT, GG and RE. **AP** conducted the analysis and interpreted the results in discussion with TVS, RAE, AT, GG and RE. **AP** wrote the initial draft of the manuscript, which was then critically reviewed for important intellectual content by all authors.

Chapter 7: **AP** designed the study with support from TVS and RE. **AP** conducted the analysis and interpreted the results in discussion with TVS and RE. **AP** wrote the initial draft of the manuscript, which was then critically reviewed for important intellectual content by all authors.

2 Validation of data extraction

2.1 Introduction

Throughout this thesis many cardiovascular disease (CVD) risk prediction models were developed, and risk predictions were also generated using existing models. This required the development of cohorts that included CVD outcomes, CVD risk factors or predictors, and statin use. While the cohorts in each individual chapter varied on some inclusion/exclusion criteria, the set of variables derived was always a subset of these. Where possible, models were based around the QRISK series of models.[5,99] These could be viewed as the gold standard for CVD risk prediction in the UK (recommended in National Institute for Health and Care Excellence (NICE) guidelines,[45] externally validated with strong performance,[100,101] updated yearly, developed on large datasets and methods clearly reported). All of the predictor variables included in the QRISK3[5] model, CVD outcomes and statin use therefore had to be derived for the cohorts used in this thesis.

To do this, access was gained to the Clinical Practice Research Datalink (CPRD) database (ISAC:17_125RMn2A2), with linkage to Hospital Episode Statistics[49] (HES) and Office for National Statistics[50] (ONS) data. CPRD is a primary care database representative of the UK in terms of age, sex and ethnicity.[48] This is similar to QResearch (the database in which QRISK models are developed on) which is also a primary care database, and contains all the required information for prediction of CVD. HES contains secondary care (hospitalisation) data, and ONS contains mortality data. Extracting data from large routinely collected data sources such as CPRD and HES is not straightforward and involves a lot of programming. Therefore it was important to quality control the extraction process.

2.2 Methods

2.2.1 Overview

A CPRD cohort was defined using the same inclusion/exclusion criteria as the QRISK3[5] development cohort, and all the required variables were extracted. The distribution of each

covariate was then compared against those reported in the QRISK3 development cohort. Given the CPRD cohort was developed using the same criteria as the QRISK3 development cohort, the distribution of each variable should broadly match up. If this was the case, I was happy to re-use the programs when extracting the same variables for other cohorts. However, there were some key problems in making this comparison:

1) CPRD (practices use Vision software) and QResearch (practices use EMIS software), contain completely different sets of practices.

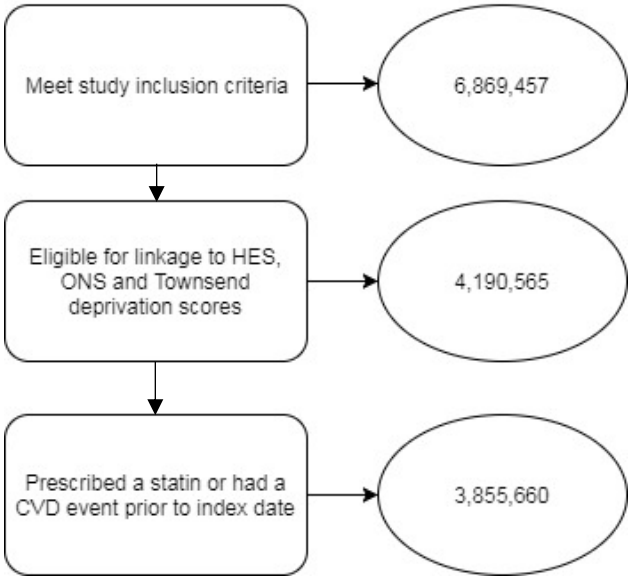
2) Code lists and detailed definitions of covariates used in QRISK3 were not made publicly available (apart from the outcome, CVD).

This meant when comparing the prevalence of certain covariates between the CPRD cohort and the QRISK3 cohort, it was hard to distinguish whether differences were due to underlying differences in the populations of the two databases, differences in data recording between the two software systems, differences in the methods used for derivation of the variables, or coding errors on my behalf. For each variable, steps were taken to rule out the possibility that the difference was due to coding errors. Differences caused by the other reasons would not affect the validity of the work in this thesis, as long as the method of extraction was not flawed.

2.2.2 Derivation of cohort

The CPRD cohort was developed on the January 2017 extract of CPRD, linked to HES, ONS and Townsend deprivation scores.[102] It was derived in the same way as the cohort used to develop the QRISK3 model. Patients met the initial study inclusion criteria if they had at least one day of up-to-standard follow up aged 25 – 84 and within the study follow up period (1st Jan 1998 – 31st December 2015), and 1 year up-to-standard registration prior to this day. The index date was defined as the first day which meets these criteria, alternatively specified as: latest of (i) date turned 25, (ii) year valid follow up in CPRD, (iii) 1st January 1998, the study start date. Patients who were not eligible for linkage to HES, ONS and Townsend deprivation scores were then excluded. Finally, patients were excluded if they had a history of CVD (identified through CPRD, HES or ONS) or had received a statin prior to their index date. Code lists for CVD outcomes and statin prescriptions are provided on the GitHub page for this thesis.[103] Figure 2.1 contains a flow chart for the derivation of the CPRD cohort.

Figure 2.1: *Flow chart of derivation of the CPRD cohort*



There were also a number of decisions made about how to process the data from HES and ONS. For both of these sources, only linked data on cardiovascular events was available, specified by code lists provided on the GitHub page associated with this thesis.[103] In HES, hospitalisations that overlapped were combined into a single hospitalisation, retaining the primary diagnosis of the first hospitalisation as the primary diagnosis of the hospitalisation. We only considered CVD events in HES or ONS where CVD was the primary cause of the hospitalisation or death. If a patient had a CVD related death 30 days or less after the date of censoring in CPRD, it was brought it forward to the date of censoring and used as an event. This was to account for possible delays in recording between the two databases.

2.2.3 Process for derivation of variables (outcomes and predictors)

A five stage process was carried out to derive the variables for the CPRD cohort.

1. Interpretation of the terminology used in QRISK3

A table was produced containing information provided in the QRISK3 paper about how each variable was defined, and how we interpreted and implemented this information to derive variables for the CPRD cohort.

2. Search for code lists

For each variable the literature was searched for code lists that could be used to derive each variable. Code lists were also available from within our research team from previous studies.

3. Initial comparison of variables between CPRD cohort and QRISK3 development cohort

Each variable was derived using information gathered in stage 1 and 2. The CPRD cohort was then compared to the QRISK3 development cohorts on the prevalence of comorbidities, the mean and standard deviation of continuous covariates, the distribution of categorical variables, and the proportion of missing data. For each predictor variable, univariate Cox[75] models were also fitted to the outcome (time until CVD event) to produce hazard ratios as a second way to check the validity of the variable. For example if increasing age was associated with lower CVD risk it would be clear something was wrong.

4. Exploration of the reason for differences in relevant variables, and modification of variable definitions for the final cohort when deemed necessary

Where results in the CPRD cohort seemed incorrect, potential reasons for this were assessed to help determine whether the variable had in fact been extracted incorrectly. The method used to explore each variable was unique to the problem at hand, and is outlined in the corresponding part of section 2.3.4. Depending on what was found, the algorithm or code list used to derive a given variable may have been changed. However, to avoid chasing the values in the QRISK3 development cohort, this was only done in situations that there was strong evidence that the initial variable derivation was wrong.

5. Summary of final cohort

Upon completion of stage 4, the final cohort was again compared to the QRISK3 development cohort. Histograms of continuous variables were also plotted to gain a better understanding of the distribution of these variables.

2.3 Results

2.3.1 Interpretation of terminology used in QRISK

Table 2.1 contains the available information in the QRISK3 manuscript[5] (or the online calculator[104]) about variable derivation, and how this information was interpreted and used for extraction of the CPRD cohort.

Table 2.1: *Comparison of definition of variables from the QRISK3 manuscript, and how this definition was interpreted.*

Variable	QRISK3 available information (from manuscript, or online calculator)	How this information was interpreted and implemented
Outcome variable		
Time until first CVD event	*"The primary outcome measure was the first recorded diagnosis of cardiovascular disease recorded on the general practice clinical computer system or their linked ONS death certificate during the study period. For this study, we included coronary heart disease (angina and myocardial infarction), stroke, or transient ischaemic attacks in the term cardiovascular disease but not peripheral vascular disease."*	If a patient died, was transferred out of their practice or the practice stopped contributing data for CPRD, then they were censored on this date. If the time until the first CVD event either in primary care (CPRD), secondary care (HES) or death (ONS) happened before a patients censoring date, this was recorded.
Demographics		
Age	*"Age is defined at the index date."* *"We determined an entry date to the cohort for each patient, which was the latest of the following: 25th birthday, date of registration with the practice plus one year, date on which the practice computer system was installed plus one year, or the study start date (1 January 1998)."*	Age was defined as (index date – date of birth)/365.25. Only the year of birth is provided in CPRD, therefore it was assumed everybody was born on the 1st July on the year of their birth.
Ethnicity	*"We used Read codes for self-assigned ethnicity. The codes were*	Any medical code prior to or after the index date was extracted. A

	grouped into the NHS standard 16+1 categories28 for the initial descriptive analysis. The 16+1 categories were then further grouped into the final nine reporting groups to ensure sufficient numbers of events to enable a meaningful analysis."	persons ethnicity cannot change therefore it didn't matter if it was recorded in the database after the index date.
Test data		
BMI	*"For clinical values (systolic blood pressure and body mass index) and smoking status we obtained the most recent values recorded before the baseline date."*	We took the most recent value before the index date, looking as far back as five years prior to the index date. The full algorithm, which deals with extreme values, different unit measurements, etc, is provided on the GitHub page.[103]
Cholesterol/HDL ratio	*"We selected the closest value to cohort entry for total cholesterol: high density lipoprotein cholesterol ratio, restricting values after the baseline date to those before the patient had a diagnosis of cardiovascular disease or was censored, and before any statin prescriptions."*	The value taken was the one closest to the index date, in between five years prior to the index date, and minimum of: first CVD event, first statin prescription, date censored, five years after index date. The full algorithm, which deals with extreme values, different unit measurements, etc, is provided on the GitHub page.[103]
Systolic blood pressure	*"For clinical values (systolic blood pressure and body mass index) and smoking status we obtained the most recent values recorded before the baseline date."*	We took the most recent value before index date, looking as far back as five years prior to the index date. The full algorithm, which deals with extreme values, different unit measurements, etc, is provided on the GitHub page.[103]

Systolic blood pressure variability	*"To assess variability in systolic blood pressure, we identified all systolic blood pressure values recorded in the five years before study entry and calculated the standard deviation where there were two or more recorded values."*	The standard deviation of all the values within the five years prior to the index date was calculated. At least two values were required otherwise this was set to missing. The full algorithm, which deals with extreme values, different unit measurements, etc, is provided on the GitHub page.[103]
Smoking status	*"For clinical values (systolic blood pressure and body mass index) and smoking status we obtained the most recent values recorded before the baseline date."*	The most recent before index date was taken, looking as far back as the start of valid follow up. If a patient had either a smoker or ex-smoker entry prior to a non-smoker entry, the non-smoker was changed to an ex-smoker. The full algorithm, which deals with extreme values, different unit measurements, etc, is provided on the GitHub page.[103]
Medical history		
Atrial fibrillation	*"Atrial fibrillation (including atrial fibrillation, atrial flutter, and paroxysmal atrial fibrillation"*	A medical code prior to the index date was required.
Atypical Antipsychotic use	*"Second generation 'atypical' antipsychotic use (including amisulpride, aripiprazole, clozapine, lurasidone, olanzapine, paliperidone, quetiapine, risperidone, sertindole, or zotepine)"* *"Use of drugs at baseline was defined as at least two prescriptions, with the most recent one no more than 28 days before the date of entry to the cohort."*	At least one prescription in the 28 days prior to index date, and at least two at any point prior to the index date was required.
Chronic kidney disease	*"Chronic kidney disease (stage 4 or 5) and major chronic renal disease*	A medical code prior to the index date was required.

	(including nephrotic syndrome, chronic glomerulonephritis, chronic pyelonephritis, renal dialysis, and renal transplant)" *+ "Expanded definition of CKD (to include general practitioner recorded diagnosis of CKD stage 3)."*	
Corticosteroid Use	*"Corticosteroid use (British National Formulary (BNF) chapter 6.3.2 including oral or parenteral prednisolone, betamethasone, cortisone, depo-medrone, dexamethasone, deflazacort, efcortesol, hydrocortisone, methylprednisolone, or triamcinolone)"* *"Use of drugs at baseline was defined as at least two prescriptions, with the most recent one no more than 28 days before the date of entry to the cohort."*	At least one prescription in the 28 days prior to index date, and at least two at any point prior to the index date was required.
Erectile dysfunction	*"Diagnosis of erectile dysfunction or treatment for erectile dysfunction (BNF chapter 7.4.5 including alprostadil, phosphodiesterase type 5 inhibitors, papaverine, or phentolamine)"*	A medical code prior to index date, or at least one prescription in the 28 days prior to index date, and at least two at any point prior to the index date was required.
Family history of coronary heart disease	*"Family history of coronary heart disease in a first degree relative aged less than 60 years"* Online QRISK3 calculator: *"Angina or heart attack in a 1st degree relative < 60"*	A medical code prior to the index date was required.
HIV/AIDS	*"Diagnosis of HIV or AIDS"*	A medical code prior to the index date was required.

Migraine	*"Diagnosis of migraine (including classic migraine, atypical migraine, abdominal migraine, cluster headaches, basilar migraine, hemiplegic migraine, and migraine with or without aura)"*	A medical code prior to the index date was required.
Rheumatoid arthritis	*"Rheumatoid arthritis (diagnosis of rheumatoid arthritis, Felty's syndrome, Caplan's syndrome, adult onset Still's disease, or inflammatory polyarthropathy not otherwise specified"*	A medical code prior to the index date was required.
Severe Mental Illness	*"Diagnosis of severe mental illness (including psychosis, schizophrenia, or bipolar affective disease)"*	A medical code prior to the index date was required.
Systemic lupus erythematosus	*"Systemic lupus erythematosus (including diagnosis of SLE, disseminated lupus erythematosus, or Libman-Sacks disease"*	A medical code prior to the index date was required.
Treated hypertension	*"Diagnosis of hypertension and treatment with at least one antihypertensive drug."* Online QRISK3 calculator: *"On blood pressure treatment?"*	A medical code prior to the index date and at least one prescription of antihypertensive drug in the last 6 months was required.
Type 1 diabetes	*"Diabetes (type 1, type 2, or no diabetes)"*	A medical code prior to the index date was required.
Type 2 diabetes	*"Diabetes (type 1, type 2, or no diabetes)"*	A medical code prior to the index date was required.

2.3.2 Search for code lists

The final set of code lists used to derive the variables is provided on GitHub.[103] This section details how this set of code lists was determined.

Code lists for the outcome variable, CVD, were available amongst the supplementary material of the QRISK3 paper published online. For all covariates that were included in QRISK2,[99] code lists were available from the study by Van Staa et al.[23] I then also used the code lists available from QOF[105] (Quality and Outcomes Framework) as an alternative set of code lists, given I was not sure what had been used in the QRISK3 paper.

For variables not in QRISK2, or part of QOF, code lists were not available. These variables were atypical anti-psychotic medication, erectile dysfunction, HIV/AIDS, migraine and systemic lupus erythematosus. For these variables, codes were either generated through the CPRD code browser, or were taken from the Cambridge primary care unit website,[106] or on the University of Manchester clinical codes repository.[107]

For variables where two code lists were available, both were used to derive the variable of interest and the impact of using the different code lists was evaluated, results provided in section 2.3.4.1. A summary of the variables and the source of the code list used to derive each of them is given in Table 2.2. The different possible sources were:

(i) **QRISK**: code list was provided either in the main manuscript or supplementary material of the QRISK3 publication

(ii) **QOF**: code list was taken from the NHS digital website[105]

(iii) **Previous study:** code lists used were available from previously published work by Van Staa et al.[23], unless another study is explicitly referenced

(iv) **Cambridge**: code lists were taken from the Cambridge primary care unit website[106]

(v) **Clinical codes**: code lists were taken from the clinical codes repository[107]

(vi) **Custom**: code lists were generated from scratch using the CPRD code browser

Table 2.2: *List of variables and source for where the code list was obtained*

Variable	Code list source
Outcome variables	
Time until first CVD event	QRISK
Demographics	
Age	N/A
Ethnicity	Previous study
Test data	
BMI	Previous study
Cholesterol/HDL ratio	Previous study
Systolic blood pressure	Previous study
Systolic blood pressure variability	Previous study
Smoking status	Previous study
Medical history	
Atrial fibrillation	Previous study, QOF
Atypical Antipsychotic use	Custom. All drugs with the BNF code 04020102, which refers to second generation atypical antipsychotic drugs.
Chronic kidney disease	Previous study + QOF. Both the codes available from the previous study and the QOF website were combined for this definition, as codes for nephrotic syndrome, chronic glomerulonephritis, chronic pyelonephritis, renal dialysis, and renal transplant, were not included in the QOF code list.
Corticosteroid Use	QRISK (BNF chapters stated in QRISK3 paper)
Erectile dysfunction	QRISK and Custom. BNF chapter for prescriptions stated in QRISK3 paper, and the medical codes were found from searching 'erectile' and 'dysfunction' separately in the code browser.
Family history of coronary heart disease	Previous study
HIV/AIDS	Previous study[108]
Migraine	Custom, codes were identified through CPRD code browser.
Rheumatoid arthritis	Previous study, QOF

Severe Mental Illness	Cambridge, QOF
Systemic lupus erythematosus	Custom, codes were identified through CPRD code browser.
Treated hypertension	Previous study, QOF
Type 1 diabetes	Previous study, QOF
Type 2 diabetes	Previous study, QOF

2.3.3 Initial comparison of variables between CPRD cohort and QRISK3 development cohort

Table 2.3 contains a comparison of the CPRD cohorts (female and male) with the reported QRISK3 development cohorts by mean and standard of continuous variables, and distribution of categorical variables. Cohen's D between the two cohorts is also reported for continuous variables. After deriving the variables, a univariate Cox model predicting the outcome (CVD) was fitted to the data to provide a hazard ratio (HR) for categorical variables. A variable highlighted in red warranted further exploration (see section 2.3.4). Reasons why specific variables were chosen for further exploration are given in the appropriate sub-section of section 2.3.4. Variables highlighted in green indicate strong agreement and I was happy with the variable derivation.

Table 2.3: *Demographics at cohort entry date for the QRISK3 development cohort and the first derivation of the CPRD cohort*

QRISK3 FEMALE N=4,019,956	**QRISK3 MALE N=3,869,847**	**CPRD FEMALE N=1,965,078**	**CPRD MALE N=1,890,582**
Outcome variables			
Incidence of CVD (primary care + HES + ONS)			
Incident cases: 160,549 Person years: 25,943,236 Rate per 1000 person years: 6.19	Incident cases: 203,106 Person years: 24,821,632 Rate per 1000 person years: 8.18	Incident cases: 86,547 Person years: 13,801,919 Rate per 1000 person years: 6.27	Incident cases: 107,051 Person years: 12,977,234 Rate per 1000 person years: 8.25
Incidence of CVD (primary care only)			
NA	NA	Incident cases: 65,854 Person years: 13,843,035	Incident cases: 83,454 Person years: 13,020,717

		Rate per 1000 person years: 4.76	Rate per 1000 person years: 6.41
QRISK3 FEMALE	**QRISK3 MALE**	**CPRD FEMALE**	**CPRD MALE**
Demographics			
Age | mean(sd)			
43.3 (15.3)	42.6 (14.0)	43.07 (15.94) Cohen's D: 0.015	41.84 (14.57) Cohen's D: 0.054
Ethnicity			
Recorded:64.9%	59.7%	42.07%	38.21%
White or not recorded: 88.7%	88.8%	94.12%	94.48%
Indian: 1.9%	2.1%	1.14%	1.19%
Pakistani: 1.0%	1.2%	0.45%	0.49%
Bangladeshi: 0.8%	1.1%	0.14%	0.19%
Other Asian: 1.3%	1.2%	0.84%	0.78%
Black Caribbean: 0.9%	0.8%	Black = 1.73%	1.52%
Black African: 1.9%	1.8%		
Chinese: 0.8%	0.6%	0.33%	0.23%
Other: 2.6%	2.4%	1.27 % (includes mixed race)	1.12%
Test data			
BMI | mean(sd), %recorded			
25.4 (5.1), 72.8%	25.9 (4.2), 64%	25.60 (5.60), 68.83% Cohen's D: -0.040	26.13 (4.53), 53.62% Cohen's D: -0.053
Cholesterol/HDL ratio | mean(sd), %recorded			
3.7 (1.2), 39.8%	4.4 (1.4), 37.9%	3.72 (1.20), 38.48% Cohen's D: -0.017	4.48 (1.40), 35.71% Cohen's D: -0.057
Systolic blood pressure | mean(sd), %recorded			
123.2 (18.2), 82.8%	129.2 (16.3), 68.3%	123.91 (18.28), 81.01% Cohen's D: -0.039	130.02 (16.48), 59.22% Cohen's D: -0.050
Systolic blood pressure variability | mean(sd), % recorded			
9.3 (6.2), 77.7%	9.9 (6.8), 64.0%	9.45 (5.96), 50.39% Cohen's D: -0.032	10.12 (6.79), 20.94% Cohen's D: -0.025
Smoking status			
Recorded = 85% Never = 60% Ex = 17.3% Current = 22.7%	Recorded = 77.7% Never = 48.6% Ex = 19.8% Current = 31.5%	Recorded = 75.18% Never = 56.04% Ex = 16.97% (HR=1.23) Current = 26.99% (HR=1.09)	Recorded = 65.17% Never = 46.63% Ex = 17.38% (HR=1.83) Current = 35.99% (HR=1.34)

QRISK3 FEMALE	QRISK3 MALE	CPRD FEMALE	CPRD MALE
Medical history			
Atrial Fibrillation			
0.4%	0.5%	0.44% HR: 8.60	0.57% HR: 5.48
Atypical anti-psychotic use			
0.5%	0.5%	0.30% HR: 2.57	0.33% HR: 1.37
CKD (stage 3/4/5)			
0.5%	0.3%	0.32% HR: 4.04	0.25% HR: 4.27
CKD (stage4/5)			
0.2%	0.2%	0.11% HR: 4.39	0.14% HR: 4.47
Corticosteroid use			
2.4%	1.5%	0.47% HR: 3.64	0.29% HR: 4.14
Erectile dysfunction			
NA	2.3%	NA	1.45% HR: 1.93
Family history of coronary heart disease in first degree relative < 60			
12%	9.3%	3.12% HR: 0.88	2.49% HR: 1.14
HIV/AIDS			
0.1%	0.2%	0.06% HR: 0.25	0.09% HR: 1.07
Migraine			
6.4%	2.7%	7.27% HR: 0.87	2.94% HR: 0.99
Rheumatoid Arthritis			
1.1%	0.5%	0.69% HR: 3.15	0.26% HR: 3.45
Severe mental illness			
6.8%	4.3%	0.79% HR: 1.90	0.94% HR: 1.26
Systemic lupus erythematosus			
0.1%	0.0%	0.10% HR: 1.92	0.01% HR: 1.99
Treated hypertension			
5.6%	4.2%	6.18% HR: 4.41	4.50% HR: 3.79
Type 1 diabetes			
0.3%	0.3%	0.21% HR: 1.70	0.28% HR: 1.53
Type 2 diabetes			
1.2%	1.5%	1.26% HR: 4.85	1.56% HR: 4.22

**The hazard ratio (HR) associated with each variable is from a univariate Cox model predicting CVD*

2.3.4 Exploration of the reason for differences in relevant variables, and modification of variable definitions for the final cohort when deemed necessary

First there is an assessment of the code lists that were considered when there was multiple available for a single variable. Then there is a separate section for each variable where further steps were taken to explore the difference between the CPRD cohort and the QRISK3 development cohort. If changes were made to the definition of a variable for the final dataset, it is reported here.

2.3.4.1 Choice of code lists

The variables presented in Table 2.4 are those where two code lists were available. In the case of atrial fibrillation, rheumatoid arthritis, treated hypertension, and type 1 and 2 diabetes, these were available in QOF and from the work by Van Staa et al.[23] Severe mental illness and depression were available in QOF and on the Cambridge primary care website.[106] In general, differences in prevalence of variables were small. I decided to use the QOF code lists for all these variables as a consistent approach. Also there is a mention of using QOF code lists to derive the variable severe mental illness in QRISK3 (see section 2.3.4.3), therefore this may extend to other variables. The values reported in Table 2.3 are with respect to the code lists deduced from this analysis (i.e. the QOF code lists).

Table 2.4: *Comparison of prevalence of variables where two code lists were available for derivation*

QRISK3 female	QRISK3 male	Alternative code list female	Alternative code list male	QOF code list female	QOF code list male
Atrial Fibrillation					
0.4%	0.5%	0.44%	0.57%	0.44%	0.57%
Rheumatoid Arthritis					
1.1%	0.5%	0.73%	0.28%	0.69%	0.26%
Severe mental illness					
6.8%	4.3%	0.85%	0.90%	0.79%	0.94%
Treated hypertension					
5.6%	4.2%	6.29%	4.57%	6.18%	4.50%
Type 1 diabetes					
0.3%	0.3%	0.31%	0.41%	0.21%	0.28%
Type 2 diabetes					
1.2%	1.5%	0.95%	1.18%	1.26%	1.56%

2.3.4.2 Ethnicity

Excerpt from Table 2.3:

Category	QRISK3 FEMALE	QRISK3 MALE	CPRD FEMALE	CPRD MALE
Ethnicity				
Recorded	64.9%	59.7%	42.07%	38.21%
White or not recorded	88.7%	88.8%	94.12%	94.48%
Indian	1.9%	2.1%	1.14%	1.19%
Pakistani	1.0%	1.2%	0.45%	0.49%
Bangladeshi	0.8%	1.1%	0.14%	0.19%
Other Asian	1.3%	1.2%	0.84%	0.78%
Black Caribbean	0.9%	0.8%	Black = 1.73%	1.52%
Black African	1.9%	1.8%		
Chinese	0.8%	0.6%	0.33%	0.23%
Other	2.6%	2.4%	1.27 %	1.12%

Observations

There was quite a large difference in both the level of recording and also the distribution of ethnicities between the CPRD cohort and the QRISK3 development cohort. Therefore I wanted to ensure that this was due to the fact that CPRD and QResearch cover different cross sections of the UK population, or recording across the two databases is different, and not for some other reason.

Methods

To do this, I aimed to mimic the results found in this paper by Mathur et al.[109] This paper reports the prevalence and distribution of ethnicity recording in CPRD, and used the same code list for identifying ethnicity that was used in this data extraction. I created a cohort using the same steps from their paper and calculated key metrics for comparison. Good agreement between the two would indicate that the differences in Ethnicity found between the CPRD cohort and the QRISK3 development cohort were in fact due to differences in the databases, rather than an error in data extraction. There were two analyses (involving two different cohorts) carried out in the paper by Mathur et al. The methods to replicate each are outlined next.

A) Overall completeness of ethnicity recording

This section looks at how frequently ethnicity is recorded (completeness). The methods section in the Mathur paper reads:

"Hierarchies were extracted for all current and past patients contributing to the July 2012 build of the database."

Therefore I needed to replicate the July 2012 database using the Jan 2017 extract that I had access to. The paper also reads:

"For CPRD completeness was compared between:

(i) all patients including those who have left or died,

(ii) currently registered patients (that is all patients who have not died or transferred out of their general practice) and

(iii) patients registered after 1 April 2006 when incentivization of ethnicity recording was introduced to primary care."

Therefore I created three cohorts for comparison:

(i) A cohort of all registered patients with valid follow up prior to 1st July 2012, only using codes recorded prior to July 2012. This replicated the July 2012 extract as best as possible. This is referred to as the 'all acceptable patients' comparison.

(ii) Restricted cohort 1 to those patients who had a censoring date after 1st July 2012. This mimicked the currently registered requirement in July 2012. This is referred to as the 'actively registered' comparison.

(iii) Restricted cohort 2 to those patients who were first registered after 1st April 2006. The reasoning for applying this extra criteria to cohort 2 (as opposed to all patients), is that they explicitly state that the population of patients registered on or after 1st April 2006, and who are also still contributing to the database. This is referred to as the 'registered 1st April onwards' comparison.

A table was produced to look at the number of patients in each cohort, what proportion have an ethnicity code, and what proportion have a usable ethnicity code. They also produced a graph looking at the percentage of patients with an ethnicity code stratified by year of registration; I also replicated this graph.

B) Comparison of the CPRD population with the 2011 UK census population

In this section, I compare the CPRD cohort with both the 2011 UK census population, and the cohort derived in the Mathur paper for this same comparison. The methods section in the Mathur paper reads: *"The ethnic breakdown of the census population was compared with that of all CPRD patients who were actively registered on 27 March 2011"*. Therefore a cohort of patients that were actively registered on 27th March 2011 was created, and I calculated the distribution of the usable ethnicity codes in this cohort. This was compared with the UK census data and the cohort used in the Mathur paper. I only considered codes prior to 27th March 2011 when extracting the ethnicity variable, to mimic the fact that when the census was taken, information from the future could not be considered.

Results

A) Overall completeness of ethnicity recording

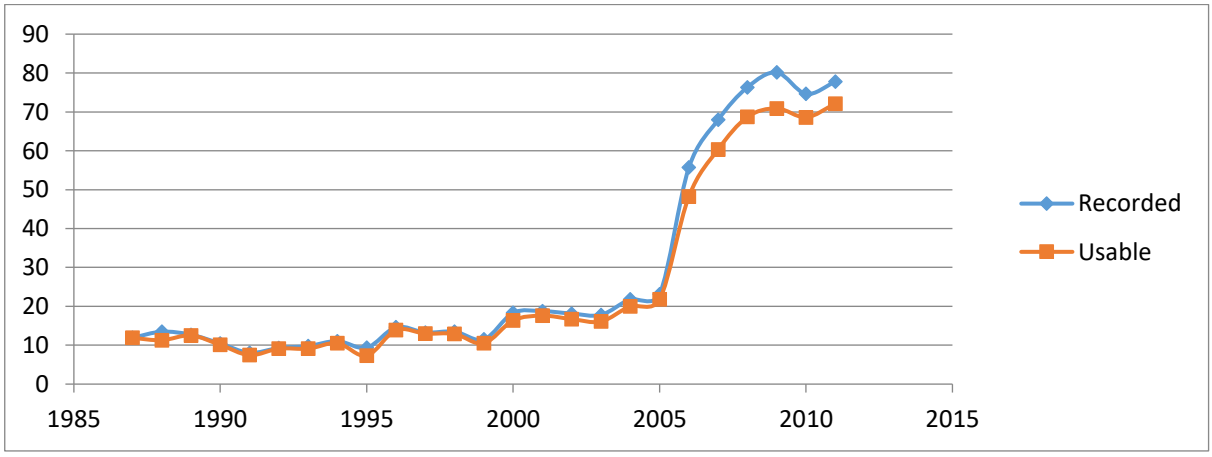
The level of ethnicity recording is reported in Table 2.5. For each comparison, there was a fairly significant difference in the number of patients in the cohort. However there was a close match in the proportion of patients with usable ethnicity codes, with 27.43% of all acceptable patients in the CPRD cohort, and 27.1% in the Mathur paper, 49.8% compared to 45.7% for those actively registered on July 2012, and 75.89% compared to 78.3% for those registered after 1st April 2006. Our data therefore closely matches the level of recording of ethnicity.

The graph of recorded ethnicity codes stratified by year produced from the CPRD cohort (Figure 2.2) is also a good match with the respective image from the Mathur paper (page 687).[109] Both have a sharp increase in 2006 and taper off at about 80%.

Table 2.5: *Completeness of ethnicity recording for all acceptable patients in the database, those actively registered on July 2012, and those registered after April 1st 2006.*

Cohort	Detailed description	N	% with any ethnicity code recorded	% with usable ethnicity code recorded
All acceptable patients (CPRD)	Patients registered prior to 1st July2012	12,620,406	29.79%	27.43%
All acceptable patients (Mathur et al.)	Patients in the July 2012 extract of CPRD	12,099,672	29.3%	27.1%
Actively registered (CPRD)	Patients registered prior to 1st July2012 AND not censored by 1st July 2012	4,702,098	53.34%	49.48%
Actively registered (Mathur et al.)	Patients actively registered in the July 2012 build	5,308,411	49.1%	45.7%
Registered 1st April onwards (CPRD)	Patients registered prior to 1st July2012 AND registered after 1st Apr 2006 AND not censored by 1st July 2012	1,752,826	82.74%	75.89%
Registered 1st April onwards (Mathur et al.)	Patients registered after 1st April 2006 in the July 2012	2,201,065	84.8%	78.3%

Figure 2.2: *Percentage of patients with an ethnicity code recorded, and a usable ethnicity code recorded, stratified by year of registration in the CPRD cohort*



B) Comparison of the CPRD population with the 2011 UK census population

Table 2.6 contains the distribution of ethnicities for the CPRD cohort, the cohort from the Mathur paper, and the UK census data. The distribution of ethnicities in the CPRD cohort and those in the Mathur paper matched quite closely. For example, for those active on census day, the biggest difference was in proportion of white people (1.07%), the next biggest difference was people in the other category (-0.79%), and then of all the other categories the largest difference was in black (-0.33%).

Table 2.6: *Distribution of the usable ethnicity codes for CPRD cohort, Mathur cohort and 2011 UK census.*

UK population 16 categories (condensed).	CPRD, active cohort on census day	2011 Census	Mathur paper, active cohort on census day
White	87.67%	87.25%	86.60%
Chinese	0.62%	0.66%	0.56%
Indian	2.13%	2.3%	2.27%
Pakistani	1.28%	1.8%	1.34%
Bangladeshi	0.49%	0.7%	0.48%
Other Asian	1.98%	1.3%	1.86%
Black Caribbean	3.4%	0.9%	0.77%
Black African		1.6%	1.88%
Black Other		0.4%	1.08%
Mixed	1.17%	2.0%	1.29%
Other	1.64%	0.9%	2.43%

Conclusions

The CPRD cohorts differed in size from those in the Mathur paper. This is likely due to the fact they were developed on the January 2017 build of CPRD and then restricted to patients who were actively registered in 2012. Patients will have been removed from the database, or patients added with their medical history backdated. More importantly, there was strong agreement in both the completeness of usable ethnicity codes and the distribution of these ethnicity codes. This supports the validity of this variable extraction. Differences in the level of recording between the CPRD and QRISK3 development cohorts could be due to the fact that recording of ethnicity sharply increased after 2006. If there are more practices in the

QResearch database contributing data after 2006 compared to CPRD, this could explain the difference.

2.3.4.3 Severe mental illness and depression

Excerpt from Table 2.3:

QRISK3 FEMALE	QRISK3 MALE	CPRD FEMALE	CPRD MALE
Severe mental illness			
6.8%	4.3%	0.79% HR: 1.90	0.94% HR: 1.26

Observations

The prevalence of severe mental illness in the CPRD cohort was significantly lower in both females (6.8% vs 0.85%) and males (4.3% vs 0.90%). There must have been an alternative way in which cases were being identified.

Methods

Upon further research, I came across an answer in the 'rapid responses' section of the BMJ where QRISK3 was published, with respect to the severe mental illness variable.[110] It reads:

"Our definition was based on a combination of the Quality and Outcomes Framework (QOF) definition of severe mental illness plus a subset of the codes from the QOF definition of depression (having excluded those codes indicating mild depression). We based our definition of depression on Read codes indicating moderate or severe depression, for example severe depression, major depression, recurrent depression, psychotic depression, depressive disorder, endogenous depression."

I therefore changed my definition to also include depression codes. The code list for depression was identified from QOF. However I augmented the code list to remove all terms that could be considered mild depression.

Results

The new variable is labelled 'severe mental illness qrisk', to differentiate it from severe mental illness as a standalone variable. After this change, there was much closer agreement for both men and women between the two cohorts (see Table 2.7).

Table 2.7: *Prevalence of severe mental illness after changes*

QRISK3 FEMALE	QRISK3 MALE	CPRD FEMALE	CPRD MALE
Severe mental illness			
NA	NA	0.79%	0.94%
Depression			
NA	NA	8.07%	3.83%
Severe mental illness qrisk (severe mental illness + depression)			
6.8%	4.3%	8.63%	4.59%

Conclusions

I used the variable defined in the same way it was outlined in the comment in the rapid responses on the BMJ website. This brought the prevalence closer to the prevalence in the QRISK3 development cohort.

2.3.4.4 Family history of coronary heart disease in first degree relative < 60 years

Excerpt from Table 2.3:

QRISK3 FEMALE	QRISK3 MALE	CPRD FEMALE	CPRD MALE
Family history of coronary heart diease in first degree relative < 60			
12.0%	9.3%	3.12% HR: 0.88	2.49% HR: 1.14

Observations

The prevalence of family history of coronary heart disease in first degree relative < 60 in the CPRD cohort was significantly lower in both females (12.0% vs 3.12%) and males (9.3% vs 2.49%). There must have been an alternative way in which cases were being identified. Also, the hazard ratios were small, with a protective effect of family history on women.

Methods

The initial code list used was a very specific code list, which required that in the read term it stated that the relative with history of coronary heart disease was < 60. I created another two code lists which were less specific, and compared the prevalence of the variable when using each. They were:

(i) The initial code list

(ii) Looking for any family history code where the read term stated that the family history was at < 65 (this is the highest number for which this happens)

(iii) Looking for any family history of coronary heart disease code, that doesn't require the age of the relative to be specified. We hypothesised this could be a reasonable code list to use, as GP's may be less likely to use these codes that specify the age of the relative in older patients. There may be many cases where a patient older patients just receive a non-specific code for family history of coronary heart disease, even if it was in a young relative.

Results

The prevalence of each of the above definitions is provided in Table 2.8. There was not a large difference between the first two definitions, as only a couple of extra codes were included. However there was a stark difference between definition (iii) and the first two. The prevalence for this difference was much more in line with that from the QRISK3 cohort (12% vs 15.08% and 9.3% vs 11.02% for females and males respectively).

Table 2.8: *Prevalence of family history of coronary heart disease variable after changes*

QRISK3 FEMALE	QRISK3 MALE	CPRD FEMALE	CPRD MALE
Family history of coronary heart disease			
12%	9.3%	*Most specific (definition i):* 3.12% HR: 0.88	*Most specific (definition i):* 2.49% HR: 1.14
		Medium specificity (definition ii): 3.16% HR: 0.88	*Medium specificity (definition ii):* 2.52% HR: 1.14
		Least specific (definition iii): 15.08% HR: 1.16	*Least specific (definition iii):* 11.02% HR: 1.37

Conclusions

The inclusion of the less specific but more sensitive codes brought the prevalence of this variable in line with the QRISK3 cohort. I was wary of chasing statistics purely to match those in the QRISK3 development cohort. However in this case, I feel that the definition used in the QRISK3 paper must also have been an unspecific definition alike to (iii) given the stark differences from definitions (i) and (ii). I believe the difference with the first two definitions were too large to be explained purely by differences in population or recording.

Furthermore, both definition (i) and (ii) had smaller HRs, and in some cases a protective effect over the risk of developing CVD. We know that CVD is hereditary, so the associations appear confounded in some way. It is possible that those with the specific codes are a younger groups of patients. If a patient was themselves over the age of 60, a GP may be less likely to ask whether they have a family history of heart disease in relatives under the age of 60. This means the codes are given selectively to a younger and therefore healthier subgroup of patients. Therefore I will use definition (iii).

2.3.4.5 Diabetes

Excerpt from Table 2.3:

QRISK3 FEMALE	QRISK3 MALE	CPRD FEMALE	CPRD MALE
Type 1 diabetes			
0.3%	0.3%	0.21%	0.28%
Type 2 diabetes			
1.2%	1.5%	1.26%	1.56%

Observations

The prevalence of both type 1 and type 2 diabetes were very close to the prevalence reported in the QRISK3 cohort. However I noticed when analysing the CPRD data that of all the patients with type 1 diabetes, 52.87% also had a type 2 diagnosis (Table 2.9). Although it is possible to have both types of diabetes, it is not this common.

Table 2.9: *Number of type 1 and type 2 diabetes cases in the cohort using original definitions*

All patients	Type 1 present	Type 1 absent
Type 2 present	4958	49379
Type 2 absent	4420	3796903

Methods

I explored what codes were being used to diagnose the diabetes in the cases where patients had a code for both type 1 and type 2 diabetes.

Results

Of the 4958 with both a type 1 and type 2 code, 4626 of the codes diagnosing type 2 had a non-specific read term such 'diabetes mellitus'. By non-specific, I refer to any read term without the phrase 'type 2 diabetes' in it. This is a non-specific read code, for which I initially made the assumption that it indicated type 2, given the prevalence of type 2 diabetes compared to type 1 in the general population. I therefore added an extra step to the algorithm, where if a patient had a type 1 code, and a non-specific type 2 code (i.e. the read term didn't explicitly state 'type 2 diabetes'), I removed the diagnosis of type 2. After making this change, the number of diagnoses and prevalence of diagnoses is presented in Table 2.10 and Table 2.11.

Table 2.10: *Number of type 1 and type 2 diabetes cases in the cohort using updated definitions*

All patients	Type 1 present	Type 1 absent
Type 2 present	332	49379
Type 2 absent	9046	3796903

Table 2.11: *Prevalence of diabetes variables after changes*

QRISK3 FEMALE	QRISK3 MALE	CPRD FEMALE	CPRD MALE
Type 1 diabetes			
0.3%	0.3%	0.21%	0.28%
Type 2 diabetes			
1.2%	1.5%	1.16%	1.42%

Conclusions

After the change, the prevalence of both types of diabetes in the CPRD cohort were still close to the ones in the QRISK3 cohort, and the proportion of patients with type 1 diabetes that also have type 2 diabetes was much more realistic than the initial 52.87%.

2.3.4.6 Chronic kidney disease (CKD)

Excerpt from Table 2.3:

QRISK3 FEMALE	QRISK3 MALE	CPRD FEMALE	CPRD MALE
CKD (stage 3/4/5)			
0.5%	0.3%	0.32%	0.25%
CKD (stage4/5)			
0.2%	0.2%	0.11%	0.14%

Observations

The prevalence of CKD in the CPRD cohort for stage 3/4/5 and 4/5 was smaller by around 40-50% for females, and 12.5-25% for males, than in the QRISK3 cohort. I hypothesised there might have been an alternative way in which cases were being identified.

Methods

After speaking with colleagues who specialise in chronic kidney disease, they informed me that there were published algorithms to:

(i) Calculate eGFR (Estimated Glomerular Filtration Rate) scores from creatinine levels

(ii) Diagnose chronic kidney disease from eGFR scores

This means cases could be identified through the test data, as well as the medical codes.

Calculate eGFR scores

eGFR scores were extracted using test data from patients' medical records. This was a combination of eGFR scores that were recorded directly, and ones that were calculated from creatinine measurements. Of all the eGFR measurements, 59% came from creatinine scores. To convert creatinine scores to eGFRs, the recommended equation to use is the CKD-EPI[111]

equation. This is recommended in the KDIGO guidelines,[112] and this recent comparison[113] comparing CKD-EPI to the MDRD equation. The comparison shows the CKD-EPI equation to be a better predictor of mortality and kidney failure, in a relatively broader population than other studies assessing the same question.

Diagnose chronic kidney disease from eGFR scores

eGFR scores can be used to diagnose CKD. The method for diagnosing chronic kidney disease from eGFR scores was taken from the paper by Jameson et al.[114] There must be at least two consecutive eGFR scores that are over 90 days apart, and are both below a certain threshold. For CKD stage 3, this is below 60, for CKD stages 4 and 5, this is below 30.

All instances of CKD calculated using test data were extracted and combined with the medical diagnoses to make a new definition of CKD stages 3, 4 and 5.

Results

The introduction of diagnosis via test data increased the prevalence of CKD stage 3/4/5 in the CPRD cohort from 0.32% to 0.45% and 0.14% to 0.32% for females and males respectively (Table 2.12). However the increase in CKD stage 4/5 was minimal, only increasing by 0.01% for both females and males.

Table 2.12: *Prevalence of CKD variables after changes*

QRISK3 FEMALE	QRISK3 MALE	CPRD FEMALE	CPRD MALE
CKD (a specific stage 3/4/5 code)			
0.5%	0.3%	0.45%	0.32%
CKD (a specific stage4/5 code)			
0.2%	0.2%	0.12%	0.15%

Conclusions

It was decided to use the updated algorithm that included diagnosis of CKD through eGFR scores as it uses a peer reviewed algorithm, is a common method to identify CKD and also brings the prevalence of CKD in the CPRD cohort closer the QRISK3 development cohort. However there is still a large difference in the prevalence of CKD stage 4/5.

One hypothesis is that the same definition of the outcome variable in QKIDNEY[115] is used for the CKD predictor variable in QRISK3. There is more information on this variable in that manuscript given is it the outcome in the QKIDNEY model. The algorithm to derive CKD cases from eGFR scores required only one eGFR score below a certain level to diagnose CKD, whereas the algorithm used here required two, which could explain the difference.

2.3.4.7 Prescription variables

Excerpt from Table 2.3:

QRISK3 FEMALE	QRISK3 MALE	CPRD FEMALE	CPRD MALE
Atypical anti-psychotic use			
0.5%	0.5%	0.30%	0.33%
Corticosteroid use			
2.4%	1.5%	0.47%	0.29%

Observations

The prevalence of both variables (particularly corticosteroid use) was smaller in the CPRD cohort than in the QRISK3 development cohort. It was suspected there may be a consistent issue in deriving prescription variables that was affecting both.

Methods

The variables were broken down to see what proportion of patients had had a prescription in the last 28 days, and what proportion had had at least two prior to their index date. Both criteria must be met for the definition used in QRISK3.

Results

For both variables, the proportion of patients that had received a prescription in the last 28 days was not high enough to match the number of patients that met both criteria in QRISK3 (Table 2.13). The proportion of patients with two prescriptions in their history is closer, but often still smaller than the proportion of patients that meet both criteria in QRISK3.

Table 2.13: *Prevalence of atypical anti-psychotic medication use and corticosteroid use broken down*

QRISK3 FEMALE	QRISK3 MALE	CPRD FEMALE	CPRD MALE
Atypical anti-psychotic medication use			
0.5%	0.5%	Both criteria: 0.30%	Both criteria: 0.33%
		28 days: 0.31%	28 days: 0.34%
		2 in history: 0.50%	2 in history: 0.57%
Corticosteroid use			
2.4%	1.5%	Both criteria: 0.48%	Both criteria: 0.29%
		28 days: 0.62%	28 days: 0.4%
		2 in history: 2.10%	2 in history: 1.42%

Conclusions

There should have been no problems with the code lists used as explicit BNF chapters are provided in the QRISK3 manuscript. Despite this, the proportion of patients classed as being on each medication was smaller for males and females in the CPRD cohort. There were nowhere near enough patients with a prescription in the last 28 days, which seems to be the main factor for the low prevalence's in the CPRD cohort. Despite this, no changes were made to the variable definition as it was not clear how this could be improved. It is more likely these are indeed differences in population or data recording, given there was a precise definition of how this variable was derived, and an exact code list was used.

2.3.5 Summary of final cohort and histograms of continuous variables

Table 2.14 compares the QRISK3 development cohort and finalised CPRD cohort with the adjusted variables. Variables reported in this table use the adjusted definitions based on the findings of section 2.3.4.

Table 2.14: *Demographics at cohort entry date for the QRISK3 development cohort and the final derivation of the CPRD cohort*

QRISK3 FEMALE **N=4,019,956**	**QRISK3 MALE** **N=3,869,847**	**CPRD FEMALE** **N=1,965,078**	**CPRD MALE** **N=1,890,582**
Outcome variables			
Incidence of CVD (primary care + HES + ONS)			
Incident cases: 160,549 Person years: 25,943,236 Rate per 1000 person years: 6.19	Incident cases: 203,106 Person years: 24,821,632 Rate per 1000 person years: 8.18	Incident cases: 86,547 Person years: 13,801,919 Rate per 1000 person years: 6.27	Incident cases: 107,051 Person years: 12,977,234 Rate per 1000 person years: 8.25
Incidence of CVD (primary care only)			
NA	NA	Incident cases: 65,854 Person years: 13,843,035 Rate per 1000 person years: 4.76	Incident cases: 83,454 Person years: 13,020,717 Rate per 1000 person years: 6.41
Demographics			
Age | mean(sd)			
43.3 (15.3)	42.6 (14.0)	43.07 (15.94) HR: 1.09	41.84 (14.57) HR: 1.08
Ethnicity			
Recorded:64.9%	59.7%	45.03%, or 42.07% when 'unclassified' counts towards not recorded	41.21%, or 38.21% when 'unclassified' counts towards not recorded
White or not recorded: 88.7%	White or not recorded	White = 36.19% White or missing = 91.16% White or missing or unclassified = 94.12%	White = 32.69% White or missing = 91.48% White or missing or unclassified = 94.48%
Indian: 1.9%	2.1%	1.14%	1.19%
Pakistani: 1.0%	1.2%	0.45%	0.49%

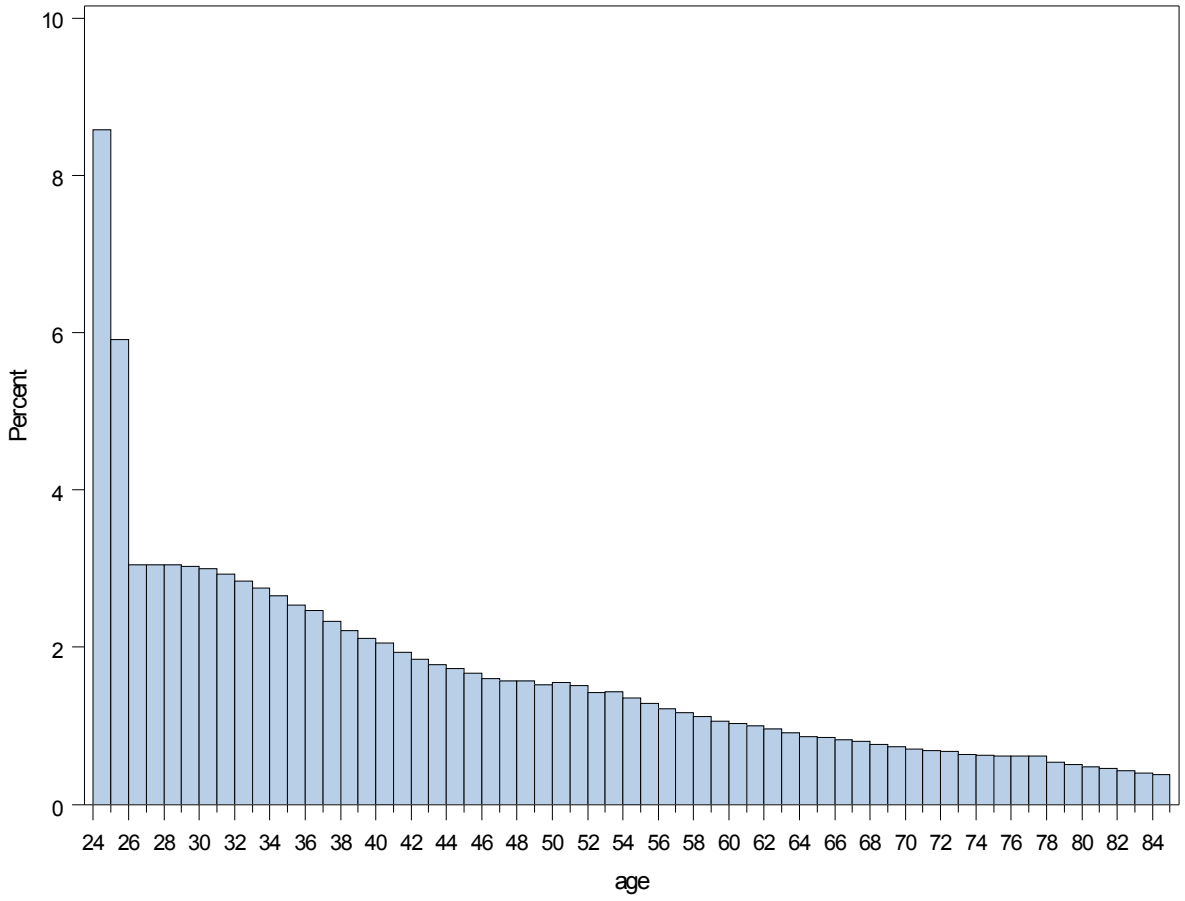
Bangladeshi: 0.8%	1.1%	0.14%	0.19%
Other Asain: 1.3%	1.2%	0.84%	0.78%
Black Carribean: 0.9%	0.8%	Black = 1.73%	1.52%
Black African: 1.9%	1.8%		
Chinese: 0.8%	0.6%	0.33%	0.23%
Other: 2.6%	2.4%	1.27 % (includes mixed race)	1.12%
Test data			
BMI | mean(sd), %recorded			
25.4 (5.1), 72.8%	25.9 (4.2), 64%	25.60 (5.60), 68.83% HR: 1.04	26.12 (4.54), 53.62% HR: 1.04
Cholesterol/HDL ratio | mean(sd), %recorded			
3.7 (1.2), 39.8%	4.4 (1.4), 37.9%	3.72 (1.20), 38.48% HR: 1.23	4.48 (1.40), 35.71% HR: 1.12
Systolic blood pressure | mean(sd), %recorded			
123.2 (18.2), 82.8%	129.2 (16.3), 68.3%	123.91 (18.28), 81.01% HR: 1.04	130.03 (16.48), 59.21% HR: 1.03
Systolic blood pressure variability | mean(sd), % recorded			
9.3 (6.2), 77.7%	9.9 (6.8), 64.0%	9.47 (5.98), 50.39% HR: 1.06	10.13 (6.80), 20.94% HR: 1.04
Smoking status			
Recorded = 85% Never = 60% Ex = 17.3% Current = 22.7%	Recorded = 77.7% Never = 48.6% Ex = 19.8% Current = 31.5%	Recorded = 75.18% Never = 56.04% Ex = 16.97% (HR=1.23) Current = 26.99% (HR=1.09)	Recorded = 65.17% Never = 46.63% Ex = 17.38% (HR=1.83) Current = 35.99% (HR=1.34)
Medical history			
Atrial Fibrilation			
0.4%	0.5%	0.44% HR: 8.60	0.57% HR: 5.48
Atypical anti psychotic use			
0.5%	0.5%	0.30% HR: 2.57	0.33% HR: 1.37
CKD (a specific stage 3/4/5 code)			
0.5%	0.3%	0.45% HR: 4.20	0.32% HR: 4.38
CKD (a specific stage4/5 code)			
0.2%	0.2%	0.12%	0.15%

		HR: 4.40	HR: 4.45
Corticosteroid use			
2.4%	1.5%	0.48% HR: 3.99	0.29% HR: 3.74
Erectile dysfunction			
NA	2.3%	NA	1.45% HR: 1.93
Family history of coronary heart disease			
12%	9.3%	15.08% HR: 1.16	11.02% HR: 1.37
HIV/AIDS			
0.1%	0.2%	0.06% HR: 0.25	0.09% HR: 1.07
Migraine			
6.4%	2.7%	7.27% HR: 0.87	2.94% HR: 0.99
Rheumatoid Arthritis			
1.1%	0.5%	0.69% HR: 3.15	0.26% HR: 3.45
Severe mental illness			
NA	NA	0.79% HR: 1.90	0.94% HR: 1.26
Depression			
NA	NA	8.07% HR: 1.39	3.83% HR: 1.51%
Severe mental illness qrisk (severe mental illness + depression)			
6.8%	4.3%	8.63% HR: 1.43	4.59% HR: 1.48
Systemic lupus erythematosus			
0.1%	0.0%	0.10% HR: 1.92	0.01% HR: 1.99
Treated hypertension			
5.6%	4.2%	6.18% HR: 4.41	4.50% HR: 3.79
Type 1 diabetes			
0.3%	0.3%	0.21% HR: 1.70	0.28% HR: 1.53
Type 2 diabetes			
1.2%	1.5%	1.16% HR: 5.19	1.42% HR: 4.54

**The hazard ratio (HR) associated with each variable is from a univariate Cox model predicting CVD*

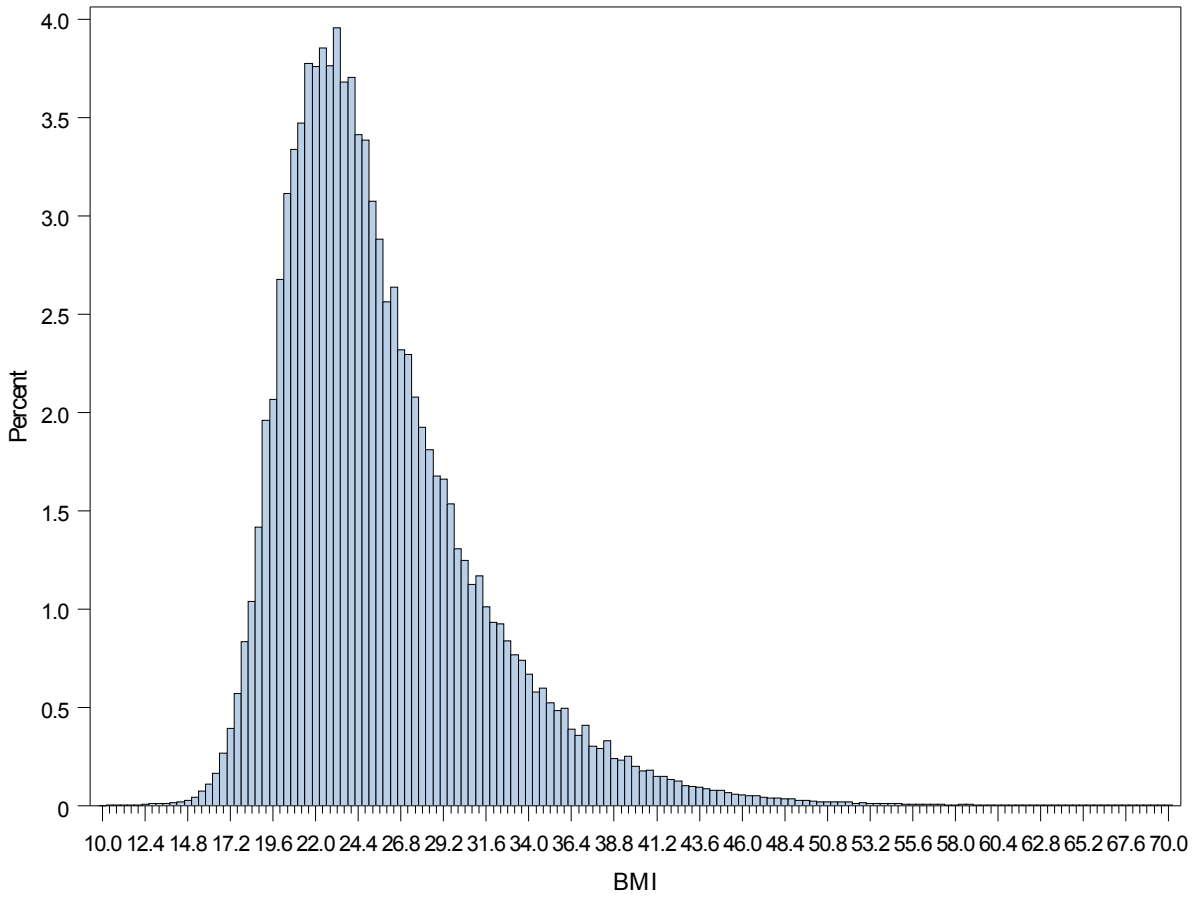
While the mean and standard deviation of the test data variables matched the published values in QRISK3 closely, there was a higher level of missing data across the board. Histograms of each continuous variable were plotted to ensure the data looked sensible. Only the histograms for the female cohort are presented as the histograms for the male cohort had the same shape (centred on a different mean) and did not provide any extra valuable information.

Figure 2.3: *Histogram of age*



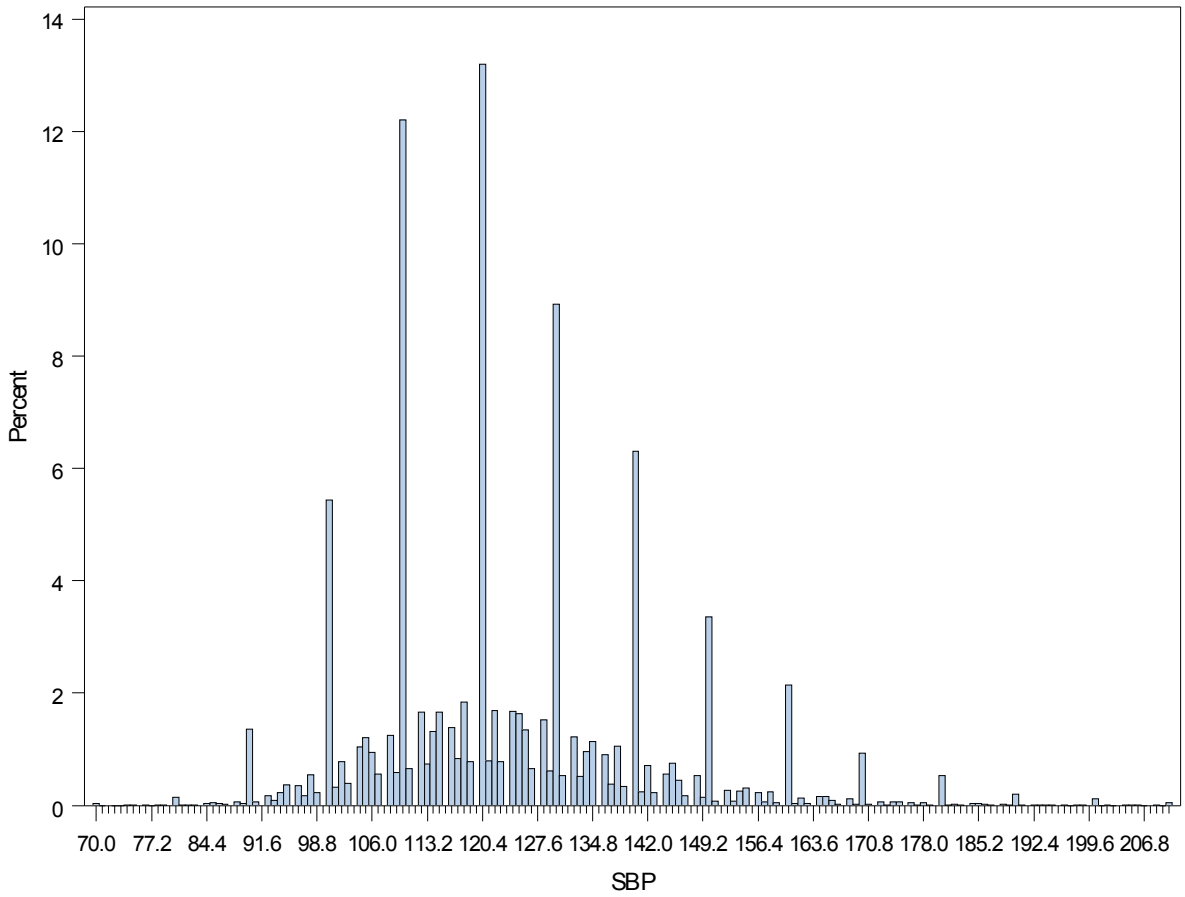
The range of age values looks normal except the spikes early on. This is a consequence of the index date definition, to be the max of date turned 25, year valid follow up in CPRD, and 1st January 1998. Therefore anyone that has one year of valid follow up prior to 1st Jan 1998, but turns 25 after 1st January 1998, will have an age on 25 on their index date.

Figure 2.4: *Histogram of body mass index (BMI)*



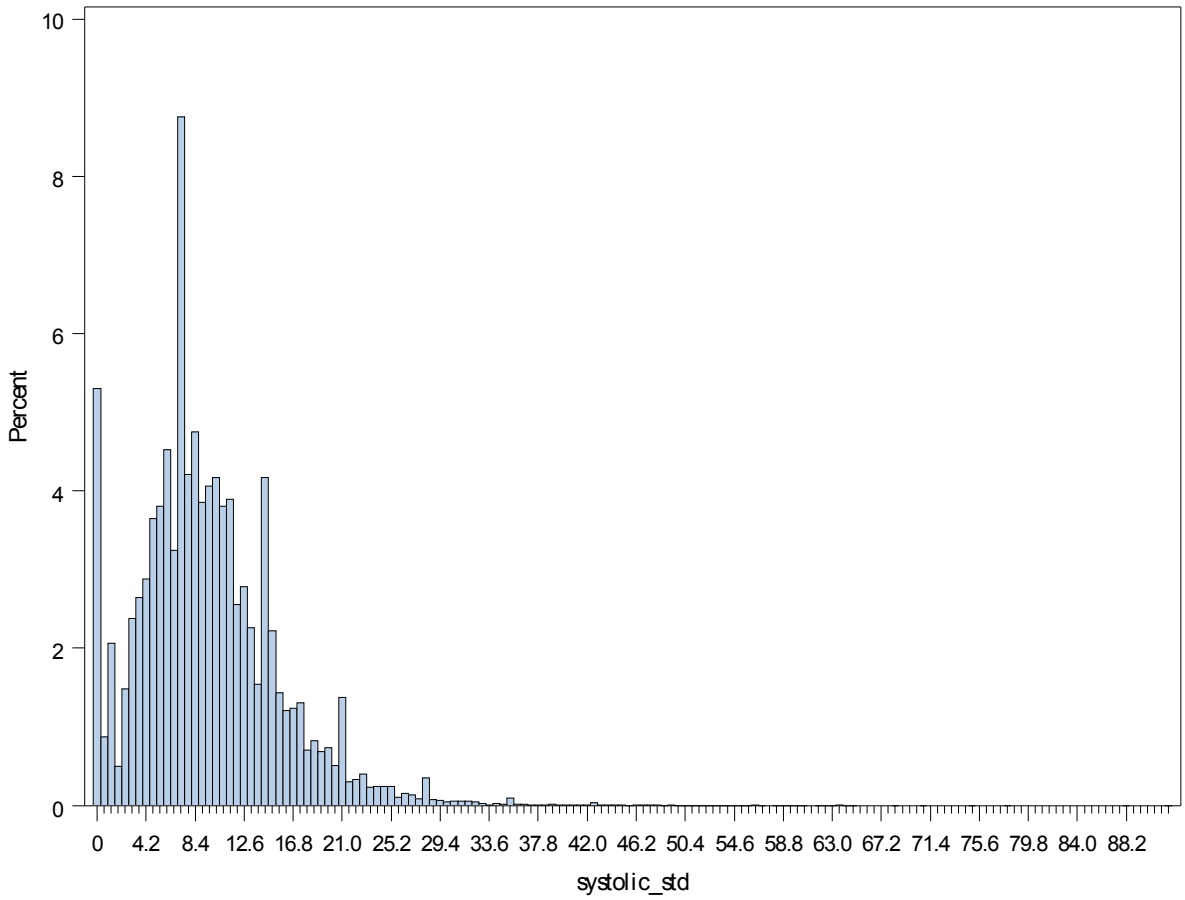
The range of BMI values looks normal.

Figure 2.5: *Histogram of systolic blood pressure (SBP)*



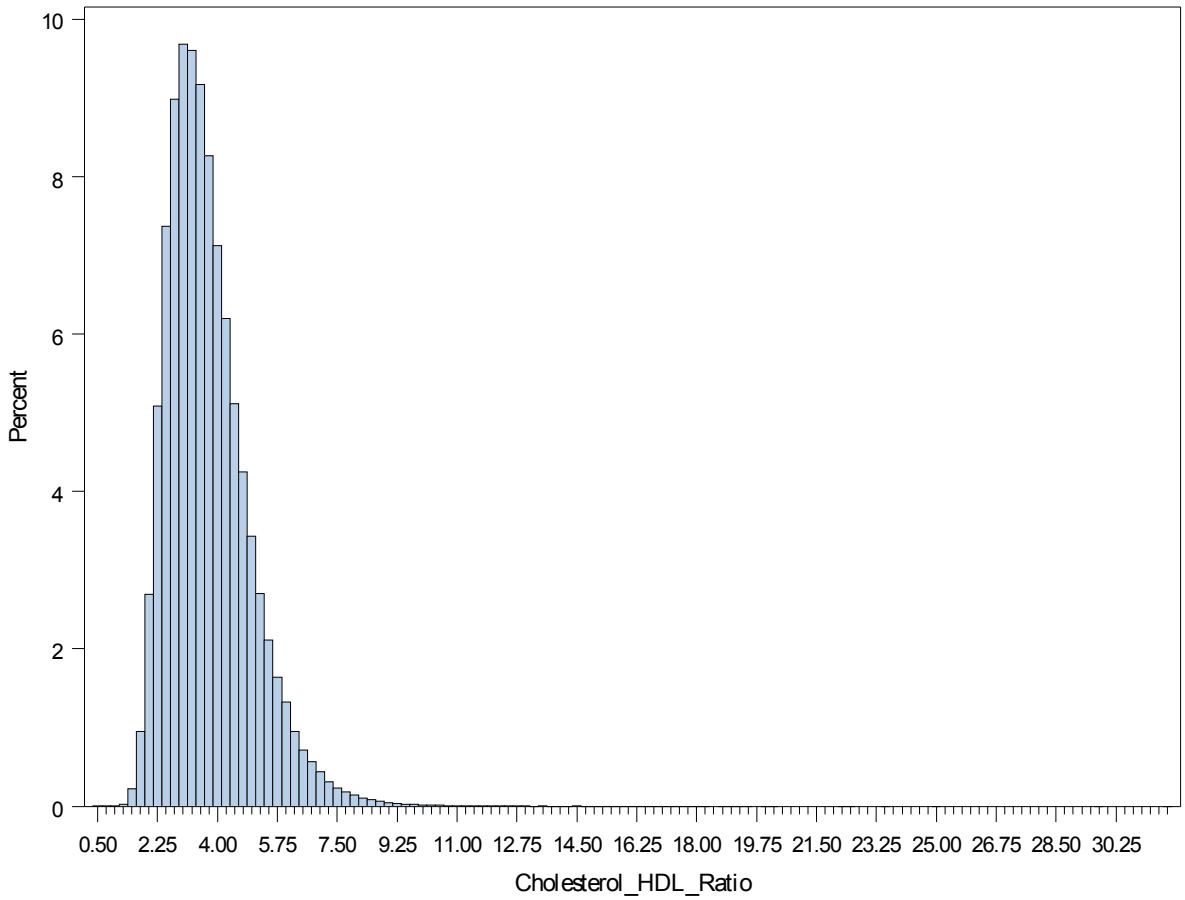
The distribution of systolic blood pressure values is odd as it has spikes at values rounded to the nearest 10.[116] This is because it is not a necessity to have more precise values and so it is often only recorded to the nearest 10, although this is not always the case as can be seen by the range of other values. I had considered adding a random effect centred on zero to each value that had been rounded to smooth out the distribution. However, doing so would add nothing to the relationship between systolic blood pressure and CVD. Furthermore, techniques now exist to impute non-normal data (i.e. predictive mean matching) so this will not be an issue either.

Figure 2.6: *Histogram of SBP variability*



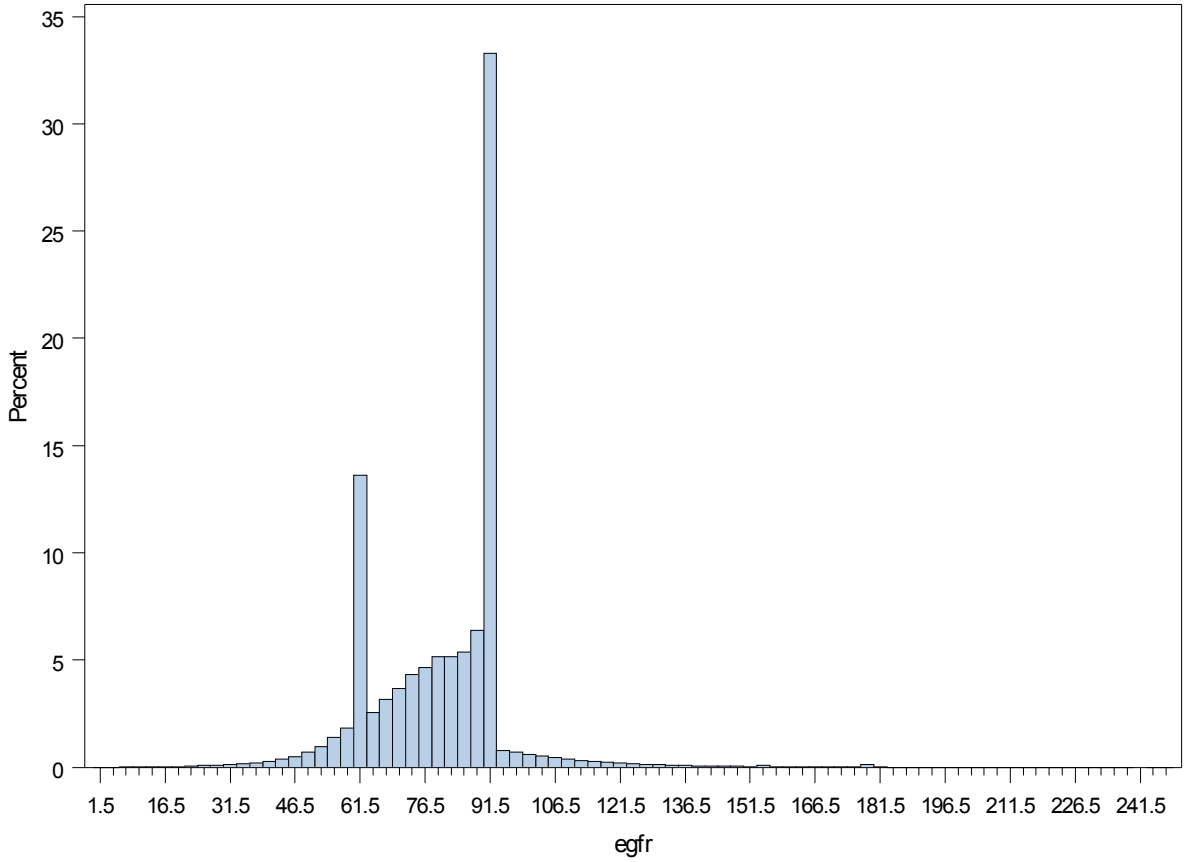
The standard deviation of systolic blood pressure is effected by the same rounding problem as systolic blood pressure is. A spike at zero is caused from patients that have had two (or more) readings that have all been rounded to the same value. Again I considered adding a random effect to the individual systolic blood pressure tests to stop this from happening. However, I feel any non-zero value must at least be close to zero, and so adding this variation in would unlikely change the relationship between this variable and the outcome.

Figure 2.7: *Histogram of cholesterol/high density lipoprotein (HDL) ratio*



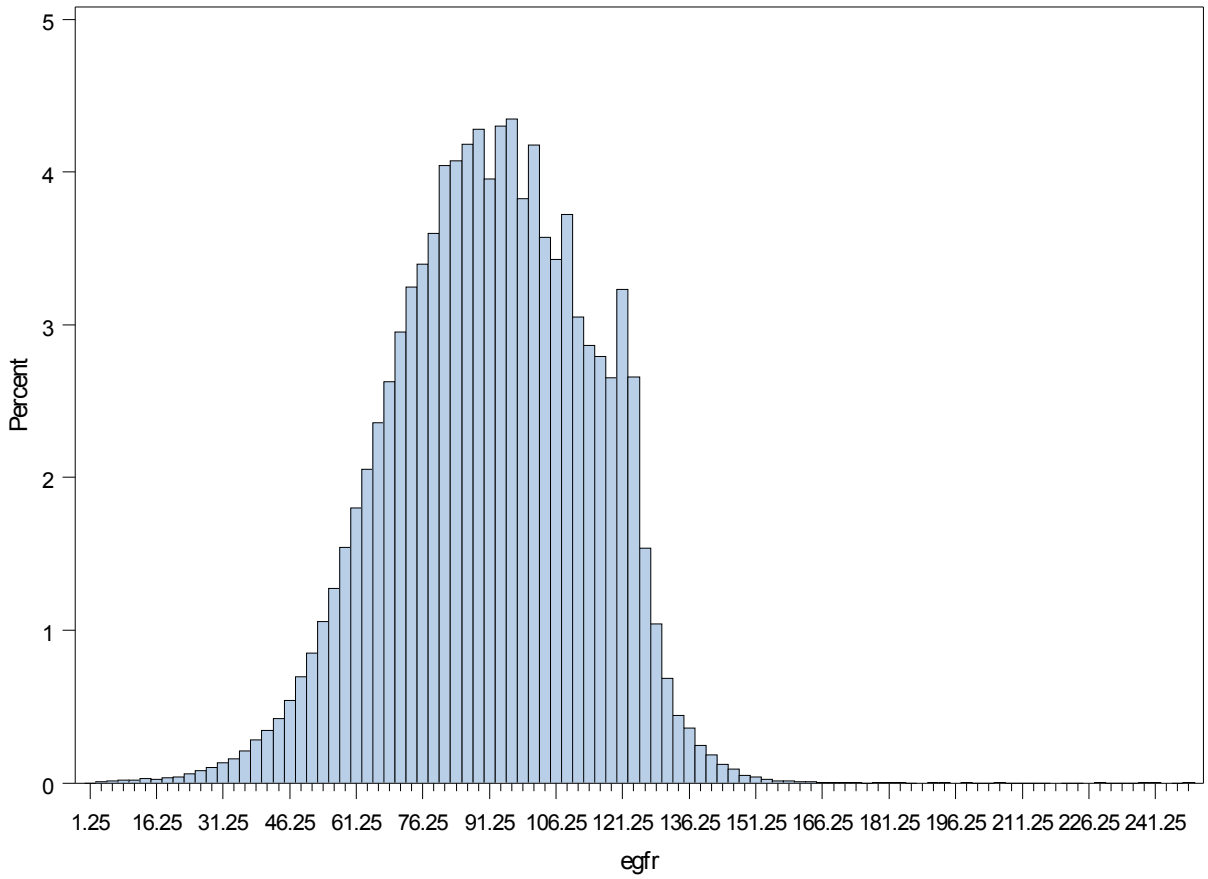
The range of cholesterol/HDL ratio values looks normal.

Figure 2.8: *Histogram of eGFR values recorded directly in the database*



There are huge spikes at 60 and 90. I was informed by clinical colleagues that this is because many labs will report anything > 90 as 90, and anything between 60 – 90 as 60. This is not problematic for me, given the cut offs for deriving CKD are at 60 and 30. If I was using this as a continuous variable, this would be problematic.

Figure 2.9: *Histogram of eGFR values derived from creatinine measurements*



The creatinine measurements give a much smoother distribution of eGFR values. Due to the way raw eGFR values are recorded, it is hard to compare the two. However, both have a decreasing trend in prevalence from 90 and below, giving some indication that the eGFR values derived from creatinine do in fact have a similar distribution to those raw eGFR measurements.

2.4 Discussion

2.4.1 Validation of cohort

For variables age, BMI, cholesterol/HDL ratio, smoking status, SBP, SBP variability, atrial fibrillation, treated hypertension, rheumatoid arthritis and systemic lupus erythematosus, the initial derivation appeared valid. The levels of missing data were higher in my cohort than the QRISK3 development cohort for all of the test data, however the means, standard

deviations and distribution of categories matched up very indicating the data was missing under the same mechanism as in the QRISK3 cohort.

For all the other variables, differences existed between the CPRD and QRISK3 cohorts. Tailored strategies were used to investigate was what driving these differences. For some variables (such as diabetes or severe mental illness) this resulted in changing the extraction method used for the final CPRD cohort. For other variables (such as ethnicity and prescription data), the extraction method remained unchanged. There are multiple potential reasons for differences between the two cohorts, which do not affect the validity of the work in this thesis:

The underlying samples are different

If the underlying samples are different, the prevalence of many variables could be different across the two cohorts. The samples could be different if patients in each database are not a representative distribution of patients across the UK. This possibility is highlighted by the difference in Ethnicity distributions. This variable extraction was validated extensively yet there were fairly significant differences between the CPRD cohort and the QRISK3 development cohort. Therefore the distribution of patients between the two databases may just be different.

The recording of patients' demographics is different across the databases

CPRD is based on practices that use VISION computer software, whereas QResearch practices use EMIS software. The way in which GPs enter information is different between the two systems. This could cause certain comorbidities to be recorded at different frequencies. Another possible reason for this is that the time in which patients are registered may be systematically different across the two databases. It is known that the levels of recording have changed over time, so if one database has more patients registered recently, there may be different levels of recording for certain variables.

Method for derivation of variables may be different

The description of how each predictor variable is calculated in QRISK3 is not completely reproducible. For example code lists are not provided for many of the variables and the period in which to look for codes is often not reported. Also, the algorithms for deriving test data

(how to deal with extreme values, different units, etc) are not given. Exactly how the information given to derive the variables was interpreted is detailed in section 2.3.1, and alterations to this interpretation throughout section 2.3.4. As long as the definition used for this work is valid, it does not matter if it is slightly different from the method used in QRISK3.

2.4.2 Lessons learnt

While the main aim was to validate the data extraction process, doing so has highlighted many important issues with regards to reproducibility. Without code lists and algorithms publicly available it makes going through what should be a straightforward comparison much more difficult and time consuming. For example differences in the outcome CVD, where QRISK3 provided a published code list and detailed algorithm, can instantly be attributed to either differences in underlying population or recording of data, neither of which affect the internal validity of this work. However with incomplete information about how some QRISK3 variables were derived, other possibilities had to be considered. The main conclusion I drew from this was to make all the code lists and programs used in this thesis publicly available.

2.5 Software, programs and code lists

The raw data extraction and creation of analysis datasets was carried out using SAS version 9.4.[117] Copyright © 2013 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA. The analyses of the extracted data throughout this thesis were carried out using R version 3.4.2.[118]

The code used to run the analyses in each chapter is provided on the GitHub page for this thesis, along with all the code lists used for data extraction.[103] Generally, the code for raw data extraction is not provided (with the exception of algorithms to derive test data), but the code for running the analyses on the extracted cohorts is. The re-usability of the code varies between chapters. For example, the code for some chapters is provided mostly for transparency. Whereas for some chapters, the code can all be run from within the chapter directory, simulated data is provided to run the code on, and a batch file to run all the programs. A README file is provided with the code for each chapter to explain what it

contains, and the extent of what can be done with the code. A brief summary of this is provided in Table 2.15.

Table 2.15: *Summary of code provided at the GitHub page*

Chapter	Extent of what can be done with the code provided
3	The code used to run all analyses is provided, but the coding structure is poor and so it would not be straightforward to provide data with which the code can be run. The code is provided mostly for transparency, rather than to be re-used.
4	The code and data is provided within a hierarchical structure of files, from which everything can be run relative to the root directory in which the files are placed. Simulated patient level data is provided on which the code can be run to produce dummy figures. The dummy figures that will be generated are also provided.
5	The code is provided within a hierarchical structure of files, from which everything can be run relative to the root directory in which the files are placed. However simulated data is not provided as it was not straightforward to produce data on which the code could run. The data structure on which code runs is referenced in the README file if it is of interest.
6	The code and data is provided within a hierarchical structure of files, from which everything can be run relative to the root directory in which the files are placed. Real data is provided on which the code can be run to produce the figures that appear in Chapter 6. This was possible as the simulation is based off population level discontinuation rates (i.e. no patient level data has to be provided in order to reproduce the results.
7	The code and data is provided within a hierarchical structure of files, from which everything can be run relative to the root directory in which the files are placed. Simulated patient level data is provided on which the code can be run to produce dummy figures. The dummy figures that will be generated are also provided.

3 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

Alexander Pate, Richard Emsley, Darren M Ashcroft, Benjamin Brown, Tjeerd van Staa

3.1 Abstract

Background: Risk prediction models are commonly used in practice to inform decisions on patients' treatment. Uncertainty around risk scores beyond the confidence interval is rarely explored. We conducted an uncertainty analysis of the QRISK prediction tool to evaluate the robustness of individual risk predictions with varying modelling decisions.

Methods: We derived a cohort of patients eligible for cardiovascular risk prediction from the Clinical Practice Research Datalink with linked hospitalisation and mortality records (N = 3,855,660). Risk prediction models were developed using the methods reported for QRISK2 and 3, before adjusting for additional risk factors, a secular trend, geographical variation in risk and the method for imputing missing data when generating a risk score (model A – model F). Ten year risk scores were compared across the different models alongside model performance metrics.

Results: We found substantial variation in risk on the individual level across the models. The 95 percentile range of risks in model F for patients with risks between 9-10% according to model A was 4.4% – 16.3% and 4.6% - 15.8% for females and males. Despite this the models were difficult to distinguish using common performance metrics (Harrell's C ranged from 0.86 to 0.87). The largest contributing factor to variation in risk was adjusting for a secular trend (HR per calendar year: 0.96 [0.95 – 0.96] and 0.96 [0.96 – 0.96]). When extrapolating to the UK population, we found that 3.8 million patients may be reclassified as eligible for statin prescription depending on the model used. A key limitation of this study was that we could not assess the variation in risk that may be caused by risk factors missing from the database (such as diet or physical activity).

Conclusions: Risk prediction models that use routinely collected data provide estimates strongly dependent on modelling decisions. Despite this large variability in patient risk, the models appear to perform similarly according to standard performance metrics. Decision making should be supplemented with clinical judgement and evidence of additional risk factors. The largest source of variability, secular trend in CVD incidence, can be accounted for and should be explored in more detail.

3.2 Background

Risk prediction models have become an important part of clinical decision making. They provide a quick and simple way to assess a patient's risk of a given disease or particular event which can then guide treatment. A recent review by Damen et al.[32] found 363 models for predicting a patient's risk of developing cardiovascular disease (CVD) and a review by Goldstein et al. found 107 models from 2009-2014 that use routinely collected data from electronic health records (EHRs).[119] As of 2018 in the United Kingdom, national guidelines recommend that clinicians use a risk prediction model (QRISK2[99]) to determine whether to prescribe a statin for primary prevention of CVD (if a patient's CVD risk is 10% or more[46]). The public availability of these algorithms contradicts the National Institute for Health and Care Excellence (NICE) guidance, which emphasises the approximate nature of these algorithms when applied to a specific patient and the need for interpreting the risk scores alongside informed clinical judgement.[46]

The validity and usefulness of risk prediction models are currently assessed using population-level statistics that measure calibration and discrimination. Calibration[120] is a measure of predictive accuracy assessing whether the average predicted risk is close to the observed risks in the overall population or in subgroups of that population. Discrimination is a relative measure of whether patients with higher risks are more likely to have an event (i.e. in a logistic regression model) or more likely to have an event sooner (i.e. in a survival analysis) than those with lower risks. In logistic regression the Area Under the Curve[120] can be calculated, whereas for survival models Harrell's C is a commonly used metric.[121] One characteristic of note of these measures is that they are population-based and derived from classifying larger groups of patients. They do not provide evidence of the level of uncertainty around a risk prediction

for an individual patient beyond the statistical confidence interval. Uncertainty on a patient level may occur if major risk factors are not considered, models are applied outside the setting in which they were developed or different EHR systems or coding dictionaries are being used with varying standards in data collection.[122,123] Furthermore, modelling decisions such as which variables to include or how to define the cohorts for the development of the models may also yield different risk predictions for the same patient. Variable selection is often based on prior/expert knowledge, risk factors identified from the literature, or data driven selection criteria, all of which may result in different models depending which researchers or methods are involved. Recent research found that well-established risk prediction models (such as Framingham and QRISK2) provided inconsistent predictions for individuals[23] despite these models having good population-level performance metrics. Uncertainty analyses have been proposed in order to establish whether models can be used for individual decisions.[124] These go beyond the classical statistical confidence interval which evaluates the uncertainty associated with the fitted values, a group mean for all patients with the same covariates. Instead they evaluate the uncertainty associated with other sources such as the modelling decisions that are made.

The objective of this study was to conduct an uncertainty analysis of the QRISK2 risk prediction model for CVD and to evaluate whether modelling decisions, in particular what patient data we choose to include in the model, had a meaningful impact on individual risk predictions (i.e., whether they substantially changed individual risk predictions). We focus in this study on the type of uncertainty which is known as 'epistemic' and caused by a lack of knowledge,[124] as opposed to aleatory uncertainty, which is inherent due to the complex processes going on in the human body. This study consisted of a comparison of alternative models, evaluating whether they changed individual risk predictions and population-level performance metrics. Clinicians could face substantial uncertainty if alternative models that perform equally well give different predictions for their patients.

3.3 Methods

3.3.1 Overview of development of QRISK risk prediction models

The models developed in this paper are based on the QRISK series of models. These CVD risk prediction models were built using routinely collected EHRs from primary care practices in the UK. Up until 2019 QRISK2[99] was the recommended model for assessment of risk in clinical practice.[46] However since 2019 the third version, QRISK3,[5] has been recommended by NICE.[62] All individuals aged 25-84 with no medical history of CVD or prior statin treatment are eligible for risk prediction using this model. We have chosen to base the current analysis around these because they are widely used in clinical practice, have been developed in very large populations (QRISK3 was developed in 4,019,956/3,869,847 females and males), and have been externally validated reporting strong performance (Harrell's C[121] was 0.880 and 0.858 for female and male models respectively, and the D statistic[125] was 2.49 and 2.26 respectively, and R^2 [126] was 59.6 and 55.0 respectively).[100,101] Variables proposed for inclusion in these models are those that are known or thought to affect CVD from literature and NICE guidelines.

3.3.2 Study population

This study used data from the Clinical Practice Research Data link[48] (CPRD) linked with Hospital Episode Statistics[49] (HES), mortality records from the Office for National Statistics[50] (ONS) and Townsend deprivation data. CPRD is a primary care database that is representative of the UK in terms age, sex and ethnicity.[48] It contains the anonymised EHRs from a large group of general practices and is comparable to The Health Improvement Network (THIN) database which was used in the external validation of QRISK2.[100] The study population was derived using the same definitions as specified in QRISK3,[5] the most recent version. Overall implementation could be followed closely, although code lists for predictor variables and algorithms for deriving test data were not available, therefore differences will exist here. It included patients aged 25-84 with no history of CVD or statin medication prior to the index date. The index date was the latest date of 25th birthday, one year of follow-up for a permanently registered patient or the 1st Jan 1998 (study start date). Follow up ended on the earliest date of patient's transfer out of the practice or death, last data collection for practice or study end date of 31st December 2015. The outcome of interest was defined as the time until the first CVD event (transient ischaemic attack, ischaemic stroke or coronary heart

disease) identified either through CPRD, HES or ONS records (code lists provided on GitHub[103]).

3.3.3 Definition of different risk prediction models

A series of different risk prediction models were developed in the study population with increasing amounts of information. Each model contained all the same covariates as the previous one, with some extra variables added to the model. Variables beyond those included in QRISK2 or 3 were identified in literature as thought to be predictive, similar to the method for identifying variables for inclusion in QRISK. We emphasise the point that by selecting variables in such a fashion, we are not trying to answer the question "what is the best variables to predict CVD with?", we are asking "how sensitive are individual risks to the addition of new variables?". The following models were fitted:

(i) Model A (same covariates as QRISK2[99]) including: Age, body mass index (BMI), atrial fibrillation, cholesterol/high-density lipoprotein (HDL) ratio, chronic kidney disease (CKD, stage 4/5), ethnicity, family history of CVD, treated hypertension, rheumatoid arthritis, systolic blood pressure (SBP), smoking status, type 1 diabetes, type 2 diabetes, Townsend deprivation score

(ii) Model B (same covariates as QRISK3[5]), covariates added: atypical antipsychotic use, corticosteroid use, CKD (stage 3/4/5 instead of 4/5), erectile dysfunction, HIV/AIDS, migraine, severe mental illness, SBP variability, systemic lupus erythematosus

(iii) Model C included covariates believed to be predictive of CVD risk as identified from literature, covariates added: alcohol abuse,[46] anxiety,[127] left ventricular hypertrophy,[23] number of days with a medical record in year prior to index date[23] and number of prescription items in one year prior to index date[23]

(iv) Model D added the calendar time at the patients index date to account for a secular trend in CVD[53]

(v) Model E added the region the patient resides in to account for regional variation in CVD incidence[128] (taken at the strategic health authority level, restructured in 2013[129] Strategic Health Authorities now represent 10 geographical locations across England

The same methods were used to derive variables as in QRISK3 when possible. Detailed information on the derivation of all covariates can be found in Chapter 2.

3.3.4 Development of risk prediction models

We used multiple imputation by chained equations to impute missing data for BMI, SBP and SBP variability, cholesterol, HDL, smoking status and ethnicity. All predictor variables from model E were included as predictors in the imputation procedure, as well as the Nelson Aalen estimate of the cumulative baseline hazard at the point of censoring or an event. The program used to impute the data was the R package 'mice'.[130] We imputed 20 datasets and carried out 20 iterations for each dataset. Full details about the imputation process can be found in Appendix A.3.1. The same randomly selected 200,000 patients were removed from each dataset, with the remaining patients making up the development cohort. All models were developed on the same set of 20 imputed datasets. For model development, Cox proportional hazards models were fitted, similar to QRISK, predicting the 10-year risks of developing CVD and estimating the hazard ratios (HRs) for each of the covariates. Models were developed separately for females and males. For model E, a random intercept model was fitted for region (strategic health authority level). Fractional polynomials for age and BMI were tested when developing model A using the R package mfp,[131] and these fractional polynomials were used in all subsequent models. Fractional polynomials were tested for secular trend in model D and were used in all subsequent models. When developing risk scores, survival estimates were combined using Rubin's rules on the log(-log) scale [132].

3.3.5 Validation of models

Key aspects of data and model B were compared with QRISK3 to highlight that the cohort used to develop the models were similar. We have chosen to make these comparisons with QRISK3 as the cohort is defined over the same time period. We compared incidence rates, distribution of covariates, HRs and predicted risks. This was done for model B, as this was developed using the same covariates as QRISK3, the comparator. The calibration of model B was also tested using internal validation with 200,000 randomly sampled patients to make up the test data and the remaining patients to develop the models (split sample approach). Average predicted risks were compared with the Kaplan Meier survival estimate at 10 years to assess calibration across groups defined by 10th percentile of risk.

Various model performance measures evaluated the performance of all our models, identified from the literature.[133–136] These included a variety of discrimination measures

(Harrell's C_H,[121] Uno's C_U,[137] Gonen and Hellers C_{GH}[138] and Royston and Saurbrei's D measure[125]), two measures of explained randomness (Kent and O Quigley's $\rho_{w,a}$[139] and ρ_k from O' Quigley et al.[140]), one measure of predictive accuracy (Integrated Brier Score (IBS)[141,142]), and four measures of explained variation (Kent and O'Quigleys R^2_{PM},[139] then Roystons R^2, Roystons R^2_D [126] and R^2_{IBS},[141] which are based on the measures ρ_k, D and IBS respectively). These were calculated to validate the models, but also as a key outcome in our study. We were interested in knowing to what extent the model performance metrics change between models if those models are predicting sizably different risks for individuals. While these metrics are not designed to assess model performance on an individual level, they are commonly used to evaluate models which are in turn used for individualised risk prediction. It is therefore important to know how sensitive they are to changes in risk on that individual level. We therefore report a range of metrics to help highlight which types of metric may best explain these changes in individual risk. When possible, performance metrics were calculated using a split sample approach (validation cohort size 200,000). ρ_k,R^2_K and C_{GH} are based on model features rather than event and censoring times, and therefore the split sample approach does not apply. C_{GH} was calculated on model developed on a sample of 200,000 patients as the algorithm used was unable to handle larger sample sizes.

The three concordance indexes estimate the probability that for a randomly selected pair of patients, the higher risk patient will have the event sooner. The range of values is 0.5 – 1, with a higher value indicating better performance. The D statistic, which calculates the log HR between two groups of patients split at the median of the linear predictor, does not have this restriction and may take values between 0 and infinity. Austin et al.[135] found that C_H and C_U were equally sensitive to the inclusion of new novel risk factors, and were more sensitive than C_{GH}. They also echo the sentiments of Harrell[143] and Uno,[137] that concordance statistics may not be sensitive when choosing between competing models, and measures of explained variation may be more sensitive in detecting differences in predictive ability. The measures of explained variation and explained randomness may take values between 0 and 1. Choodari-Oskooei et al.[133] recommended using explained variation measures R^2_{PM} and R^2_D for best meeting their criteria (independence from censoring, monotonicity, interpretability and robustness against outliers). For explained randomness $\rho_{w,a}$, is recommended by both Choodari-Oskooei et al.[136] and Austin et al.[135] despite their differing criteria of importance.

This measure is very similar to R^2_{PM}, where the variance error term $\sigma^2/6$ is replaced by 1. Finally the integrated brier score is included as it has a different aim, which is to calculate the probability of correctly predicting an event. The development and validation of models was checked against the recommendations for reporting in the TRIPOD statement (Appendix A.3.2).

3.3.6 Comparison of predicted risks between different models

After developing the models the next step was to produce risk scores, replicating the process of someone having their risk assessed in practice. In this situation, if a patient has missing data for specific covariates, the QRISK calculator will impute this using mean imputation based on age, sex and ethnicity[144]. This involved setting all originally missing values of BMI, Cholesterol HDL ratio, SBP and SBP variability, back to missing, and then imputing these using mean imputation based on age, sex and ethnicity, giving one mean imputed dataset. The same 200,000 patients were then extracted from the mean imputed dataset giving the test cohort. For each patient in the test cohort, a predicted risk according to each of model A – E was then generated. This is like a split sample approach, apart from the fact the imputation method for the development cohort and test cohort is different (as is the case in practice).

Finally, risks were also generated using model E, but for a test cohort made up of 200,000 patients from one of the multiply imputed datasets, rather than mean imputed. This represents a best estimate of the true values of each patient's missing data. The aim of this was to understand how much variation in patient risk may be masked by using mean imputation to generate a risk, as opposed to prospectively collecting their real values, as recommended by NICE. This will be referred to as model F.

The predicted CVD risks for each patient were compared between model A and models B – F. We started with model A as this model replicates the risk scores developed using QRISK2, which is the model currently used in practice. This evaluated the magnitude in which risks for an individual patient change dependent on what patient characteristics were introduced into the model. Patients were grouped into risk groups of width 1% according to their risk in model A. Then for models B – F, we provide histograms to illustrate the distribution of risks for patients from the same group, report the 2.5-97.5 percentile range for each group (average 95% CI according to model A also provided for comparison), and report the proportion of

patients from each group with a risk above or below 10%, which is the threshold for being eligible for a statin prescription in England.[46]

The final analysis consisted of the extrapolation of results to the population of England in order to assess what proportion of patients would have their treatment pathway altered depending on the model used. We extrapolated the proportion of patients eligible for CVD risk prediction in CPRD on January 1 2016 to the population in England[145] and then estimated the level of reclassification when using model F instead of Model A (QRISK2). Eligibility for patients on January 1st 2016 was the same as in the development cohorts, except index date was set to 1st Jan 2016 for all patients. This dataset was mean imputed when calculating risks according to model A - E, and one stochastically imputed dataset when calculating risks according to model F.

3.3.7 Sensitivity analyses

We found a large effect of a secular trend in CVD incidence, resulting in 56% of the patients from the 2016 cohort to be re-classified from above to below the statin treatment threshold of 10% (see results - extrapolation to English population). We therefore ran two sensitivity analyses to validate this finding. First, we verified the existence of the secular trend reporting crude incidence rates per calendar year among the model derivation cohort. For the second, we evaluated the existence of the secular trend in a cohort of statin users. For this cohort, all patients that were eligible for linkage and had more than one statin prescription between ages 25 – 85 and dates 1st Jan 1998 to 31st Dec 2015 were included. Follow up started on the first statin prescription date, and ended after a 6 month gap with no prescription. A patient could re-enter the cohort if they initiated statins again. A patient was not followed up after the event of interest (CVD). We check for the presence of this trend amongst the statin users cohort as the secular trend in CVD incidence could be explained by an increase in statin use.

To analyse this data, each patients' follow up was segmented into time followed up in each calendar year. It was also recorded whether a patient had an incident CVD event in that calendar year. We then fit a Poisson model to the data, outcome being the CVD event, adjusting for calendar year and using the time at risk in each year as an offset. This was done for the development cohort and the statin users cohort. Another model was also fit to the statin users cohort adjusting for the risk score at the start of the period of statin treatment as

well. This model attempts to find out whether the secular trend could also be explained by better prescribing of statins to those who are at high risk, through the use of models such as QRISK. The secular trend would only be of interest if it is still present in this model, which accounts for a potential change in the use of statins.

3.4 Results

3.4.1 Validation of the models

CPRD contained 6,869,457 patients with > 1 day follow up aged 25-84 during the study period. Of these, 3,855,660 (from 392 practices) were eligible for linkage to HES, ONS and Townsend quintiles and were without history of CVD or statin treatment at baseline. Table 3.1 contains the baseline characteristics for all patients who met the study eligibility criteria, which includes all patients which we generate risk scores for. There was 42.07% and 38.21% of data recorded for Ethnicity for the male and female cohorts respectively, 68.83% and 53.62% for BMI, 38.48% and 35.71% for Cholesterol/HDL ratio, 81.01% and 59.21% for SBP, 50.39% and 20.94% for SBP variability and 75.18% and 65.17% for smoking status. The mean ages were 43.07 and 41.81 for females and males, mean BMI was 25.60 and 26.12, while cholesterol/HDL ratio was 3.72 and 4.48 respectively. More importantly, we found these values to match closely with those from the derivation cohort of QRISK3 (age 43.3 and 42.6, BMI 25.4 and 25.6, Cholesterol/HDL ratio 3.7 and 4.4 respectively). A full comparison has already been provided in Table 2.14. Prevalence of medical history variables were broadly similar with those in QRISK3. Similarly, the incidence rate of CVD (see Appendix A.3.3) matched closely for both datasets (for females, there were 6.19 CVD cases per 1000 person-years in our study population compared to 6.27 in QRISK3; for males, these were 8.18 vs 8.24, respectively).

Table 3.1: *CVD incidence and baseline characteristics of study population*

	CPRD FEMALE **N=1 965 078**	**CPRD MALE** **N=1 890 582**
Outcome variables		
Incident CVD cases	86547	107051
Person years	13801919	12977235
Rate per 1000 person years	6.27	8.24
Demographics		
Age	43.07 (15.94)	41.84 (14.57)
Ethnicity: Recorded	42.07%	38.21%
White/not recorded	94.12%	94.48%
Indian	1.14%	1.19%
Pakistani	0.45%	0.49%
Bangladeshi	0.14%	0.19%
Other Asian	0.84%	0.78%
Black	1.73%	1.52%
Chinese	0.33%	0.23%
Other	1.27 %	1.12%
Test data		
BMI	25.60 (5.60)	26.12 (4.54)
Cholesterol/HDL ratio	3.72 (1.20)	4.48 (1.40)
SBP	123.91 (18.28)	130.03 (16.48)
SBP variability	9.47 (5.98)	10.13 (6.80)
Smoking status	Never = 56.04% Ex = 16.97% Current = 26.99%	Never = 46.63% Ex = 17.48% Current = 35.99%
Medical History		
Atrial Fibrillation	0.44%	0.57%
Atypical antipsychotic medication use	0.30%	0.33%
Chronic Kidney Disease stage 3/4/5	0.45%	0.32%
stage 4/5	0.12%	0.15%
Corticosteroid use	0.48%	0.30%
Erectile dysfunction	NA	1.45%

Family history of CVD	15.08%	11.02%
HIV/AIDS	0.06%	0.09%
Migraine	7.27%	2.94%
Rheumatoid arthritis	0.69%	0.26%
Severe Mental Illness	8.63%	4.59%
Systemic Lupus Erythematosus	0.10%	0.01%
Treated hypertension	6.18%	4.50%
Type 1 diabetes	0.21%	0.28%
Type 2 diabetes	1.16%	1.42%
Variables not in QRISK		
Number medical records in previous year	14.94 (13.97)	8.83 (11.45)
> 50 medical records in previous year	2.84%	1.37%
Number of prescription items in previous year	9.60 (19.87)	5.72 (16.00)
Number with > 50 prescription items in previous year	3.49%	2.04%
Alcohol abuse	0.65%	1.46%
Anxiety	13.44%	7.96%
Left Ventricular Hypertrophy	0.14%	0.18%
Region*: North East	1.89%	1.96%
North west	13.10%	13.38%
Yorkshire and the Humber	3.93%	3.85%
East Midlands	3.14%	3.23%
West Midlands	11.04%	11.28%
East of England	11.67%	11.68%
South west	11.99%	11.88%
South Central	12.84%	12.81%
London	17.52%	17.18%
South East Coast	12.88%	12.74%

The HRs for model B (Appendix A.3.3) were generally consistent with those reported in QRISK3. The HRs for covariates introduced for models C, D and E are reported in Table 3.2. All introduced covariates had a sizeable effect on risk. For example, the HRs for patients in the North West were 1.17 for females and 1.14 and males, compared to 0.92 and 0.94, respectively for patients from South Central. The HR associated with calendar time was also large, with a 0.95 and 0.96 reduction for females and males each year.

The calibration plots for model B showed overall good calibration (

Figure 3.1), which is expected considering these are optimistic calibration plots (internal validation only). The female model is very well calibrated with the calibration error no larger than 0.5% for any 10^{th} percentile group. The largest miscalibration for the male model is for group 9, an under prediction by 1.29%.

The overall performance metrics calculated for each of the models are given in Table 3.3. The largest increase is in D and R^2_D (which is derived from D), which increase from 2.39 to 2.55 and 0.58 to 0.61 (females) across the models respectively. There was little change in any of the three C statistics across the different models. While Uno's C, C_U, went from 0.85 to 0.88 for the female cohort, there was not a consistent upwards trend in the male models. Harrell's C, the most commonly reported metric was very insensitive to the model choice. Measures of explained variation and randomness showed an upward trend from model A to model F, while measures derived from the IBS were not sensitive to model choice.

Table 3.2: *HRs (95% CI) of fixed and random effects introduced into models C, D and E. HRs reported are all from model E.*

	Female	Male
Fixed effects		
Alcohol abuse	1.36 (1.25 – 1.48)	1.32 (1.25 – 1.39)
Anxiety	1.10 (1.08 – 1.13)	1.10 (1.07 – 1.12)
Left ventricular hypertrophy	1.65 (1.53 – 1.78)	1.67 (1.56 - 1.80)
> 50 medical records in year prior to index date	1.30 (1.25 – 1.36)	1.25 (1.18 – 1.31)
> 50 prescription items in year prior to index date	1.55 (1.51 – 1.59)	1.49 (1.44 – 1.54)
Calendar time (by year)	0.96 (0.95 - 0.96)	0.96 (0.96 – 0.96)
Region (random effect):		
North East	1.07 (1.00 - 1.14)	1.09 (1.08 - 1.09)
North west	1.17 (1.11 - 1.24)	1.14 (1.13 - 1.15)
Yorkshire and the Humber	1.11 (1.05 - 1.19)	1.09 (1.08 - 1.10)
East Midlands	1.00 (0.93 - 1.06)	0.99 (0.98 - 0.99)
West Midlands	0.99 (0.94 - 1.05)	0.99 (0.99 - 1.00)
East of England	0.94 (0.89 1.00)	0.94 (0.93 - 0.94)
South west	0.98 (0.92 1.04)	0.99 (0.99 - 0.99)
South Central	0.92 (0.87 - 0.98)	0.94 (0.94 - 0.95)
London	0.89 (0.84 - 0.95)	0.88 (0.88 - 0.89)
South East Coast	0.96 (0.90 - 1.02)	0.97 (0.97 - 0.97)

Figure 3.1: *Calibration plots by 10th percentile of risk for model B*

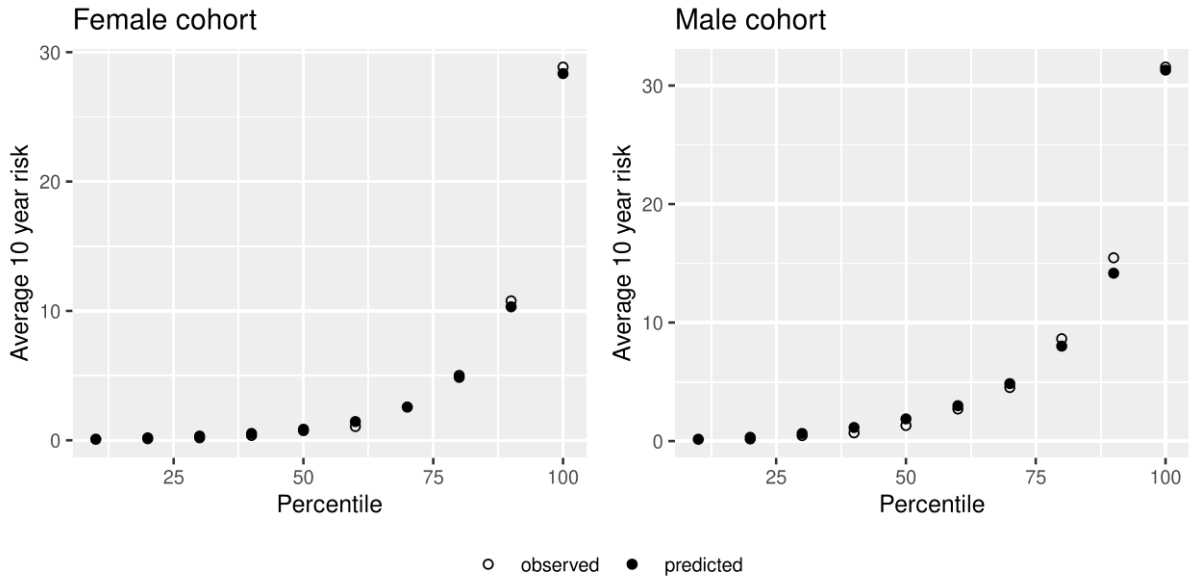


Table 3.3: *Performance metrics for each of the models*

Measure*	Model A	Model B	Model C	Model D	Model E	Model F
Female						
IBS	0.02	0.02	0.02	0.02	NA	NA
R^2_{IBS}	0.12	0.13	0.13	0.13	NA	NA
R^2_{PM}	0.65	0.65	0.66	0.67	0.67	0.67
P^2_k	0.85	0.86	0.86	0.86	0.86	NA
$P_{w,a}$	0.76	0.76	0.76	0.77	0.77	0.77
R^2	0.62	0.62	0.63	0.63	0.64	NA
D	2.39	2.42	2.49	2.52	2.52	2.55
R^2_D	0.58	0.58	0.60	0.60	0.60	0.61
C_H	0.86	0.87	0.87	0.87	0.87	0.87
C_U	0.85	0.86	0.86	0.86	0.86	0.88
C_{GH}	0.81	0.82	0.82	0.82	NA	NA
Male						
IBS	0.03	0.03	0.03	0.03	NA	NA
R^2_{IBS}	0.12	0.12	0.12	0.12	NA	NA
R^2_{PM}	0.62	0.63	0.63	0.63	0.64	0.64
P^2_k	0.78	0.79	0.79	0.79	NA	NA
$P_{w,a}$	0.73	0.73	0.73	0.74	0.74	0.75
R^2	0.49	0.49	0.50	0.50	NA	NA
D	2.12	2.12	2.18	2.21	2.21	2.24
R^2_D	0.52	0.52	0.53	0.54	0.54	0.55
C_H	0.84	0.84	0.84	0.84	0.84	0.85
C_U	0.75	0.74	0.74	0.74	0.74	0.77
C_{GH}	0.81	0.81	0.81	0.82	NA	NA

3.4.2 Analysis of risk scores

Table 3.4.1 and 3.4.2 show the distribution of changes in predicted CVD risks when using Models B-F instead of Model A. Females with a risk between 9-10% with Model A (QRISK2) were found to have risks with a 95% percentile range of 8.0 to 13.6 with model B (QRISK3) and range of 4.4 to 16.5% with Model F. The impact of the choice of model on the distribution of risks increased with higher CVD risks. For females with a risk of 19 to 20% with Model A, their risks were between 9.6 and 34.6 (95% percentile) when using Model F. These are shown graphically in Figure 3.2.

Table 3.4.1: *Distribution of risks (2.5th and 97.5th percentile) of patients in the test cohort according to each model, stratified by their risk in model A, and average 95% CI for risks in model A (female cohort)*

Female cohort						
10 year risk using model A	Average 95% CI in model A	Percentile range Model B	Percentile range Model C	Percentile range Model D	Percentile range Model E	Percentile range Model F
0-1%	0.3-0.4%	0.1-0.9%	0.1-0.9%	0.1-1.0%	0.1-1.0%	0.1-1.0%
1-2%	1.4-1.5%	1.0-2.2%	1.0-2.3%	0.6-2.3%	0.6-2.4%	0.6-2.6%
2-3%	2.3-2.6%	1.9-3.6%	1.9-3.9%	1.2-3.9%	1.1-4.0%	1.1-4.4%
3-4%	3.3-3.6%	2.8-5.0%	2.7-5.5%	1.7-5.4%	1.6-5.6%	1.5-6.1%
4-5%	4.3-4.7%	3.7-6.5%	3.6-7.4%	2.3-7.2%	2.2-7.3%	2.1-8.0%
5-6%	5.3-5.7%	4.5-7.8%	4.4-8.9%	2.8-8.5%	2.6-8.8%	2.5-9.5%
6-7%	6.2-6.8%	5.4-9.3%	5.3-10.6%	3.3-10.0%	3.1-10.1%	3.0-11.3%
7-8%	7.2-7.8%	6.3-10.4%	6.1-11.8%	3.8-11.5%	3.6-11.7%	3.5-12.6%
8-9%	8.2-8.8%	7.1-11.7%	6.8-14.3%	4.3-13.3%	4.1-13.3%	4.0-14.6%
9-10%	9.1-9.9%	8.0-13.5%	7.7-16.1%	4.9-15.0%	4.6-15.5%	4.4-16.3%
10-11%	10.1-10.9%	8.8-14.5%	8.5-16.8%	5.3-16.6%	5.1-16.8%	4.9-18.1%
11-12%	11.1-11.9%	9.8-16.3%	9.4-19.6%	5.9-19.3%	5.6-20.1%	5.4-21.1%
12-13%	12.1-12.9%	10.7-17.9%	10.1-21.3%	6.4-20.5%	6.0-21.5%	5.8-22.5%
13-14%	13.0-14.0%	11.4-18.5%	10.9-21.5%	7.1-21.1%	6.7-22.6%	6.6-23.3%
14-15%	14.0-15.0%	12.2-19.7%	11.6-23.3%	7.5-22.5%	7.3-22.7%	6.9-24.3%
15-16%	15.0-16.0%	13.1-22.0%	12.3-26.5%	8.2-26.3%	7.7-27.2%	7.6-28.1%
16-17%	15.9-17.1%	13.9-22.1%	13.0-27.4%	8.5-26.9%	8.0-27.5%	7.9-28.1%
17-18%	16.9-18.1%	14.9-23.9%	14.1-27.9%	9.2-28.9%	8.8-28.7%	8.7-29.5%
18-19%	17.9-19.2%	15.7-25.4%	14.8-30.0%	9.8-29.8%	9.6-29.7%	9.1-32.7%
19-20%	18.8-20.2%	16.6-25.8%	15.7-32.1%	10.5-32.6%	10.0-33.8%	9.7-34.4%

Table 3.4.2: *Distribution of risks (2.5th and 97.5th percentile) of patients in the test cohort according to each model, stratified by their risk in model A, and average 95% CI for risks in model A (male cohort)*

Male cohort						
10 year risk using model A	Average 95% CI in model A	Percentile range Model B	Percentile range Model C	Percentile range Model D	Percentile range Model E	Percentile range Model F
0-1%	0.3-0.4%	0.1-1.0%	0.1-1.0%	0.0-1.0%	0.0-1.0%	0.0-1.1%
1-2%	1.4-1.6%	1.0-2.0%	1.0-2.1%	0.7-2.2%	0.7-2.2%	0.6-2.7%
2-3%	2.4-2.6%	2.0-3.2%	2.0-3.3%	1.3-3.4%	1.2-3.5%	1.1-4.4%
3-4%	3.3-3.6%	2.9-4.5%	2.9-4.8%	1.9-4.7%	1.8-5.0%	1.6-6.0%
4-5%	4.3-4.7%	3.9-5.7%	3.8-6.2%	2.5-6.0%	2.4-6.3%	2.1-7.5%
5-6%	5.3-5.7%	4.8-6.9%	4.7-7.7%	3.1-7.6%	2.9-7.6%	2.5-9.3%
6-7%	6.2-6.7%	5.7-8.2%	5.6-8.9%	3.7-8.9%	3.5-9.0%	3.1-11.1%
7-8%	7.2-7.8%	6.6-9.5%	6.5-10.7%	4.3-10.4%	4.1-10.6%	3.7-12.9%
8-9%	8.2-8.8%	7.5-10.6%	7.5-11.8%	4.9-11.5%	4.6-11.9%	4.2-14.1%
9-10%	9.2-9.8%	8.4-11.8%	8.3-13.8%	5.5-13.2%	5.1-13.7%	4.6-15.8%
10-11%	10.1-10.8%	9.2-13.0%	9.0-15.3%	6.0-14.9%	5.6-15.1%	5.2-17.8%
11-12%	11.1-11.9%	10.1-14.2%	9.9-16.5%	6.7-15.8%	6.3-16.0%	5.9-18.9%
12-13%	12.1-12.9%	11.0-15.5%	10.8-17.8%	7.3-17.0%	6.8-17.8%	6.4-20.5%
13-14%	13.1-14.0%	12.0-16.8%	11.8-19.6%	8.0-19.2%	7.5-19.9%	6.9-22.6%
14-15%	14.0-15.0%	12.8-17.8%	12.6-21.7%	8.5-20.3%	8.0-20.8%	7.5-23.3%
15-16%	15.0-16.0%	13.5-19.2%	13.3-22.8%	9.0-21.7%	8.6-22.2%	8.1-23.8%
16-17%	16.0-17.0%	14.6-20.4%	14.2-24.0%	9.8-24.0%	9.2-24.7%	8.7-27.4%
17-18%	16.9-18.1%	15.3-22.8%	15.0-26.9%	10.3-25.5%	9.7-25.5%	9.2-28.6%
18-19%	17.9-19.1%	16.3-23.8%	15.8-26.7%	10.9-26.8%	10.4-26.7%	10.0-29.5%
19-20%	18.9-20.2%	17.2-25.2%	16.7-28.7%	11.5-29.5%	10.9-28.9%	10.4-32.1%

Table 3.5 summarises the number of patients in the study population who were reclassified with models B – F based on a treatment threshold of 10%. In the female cohort, 8% of those with a CVD risk between 7-8% with Model A were reclassified to a risk of ≥ 10% with Model F (for risks between 8-9% and 9-10%, this was 17% and 28% respectively). Substantially more patients were reclassified downward with predicted risks reduced. In the female cohort, 32% of those with a risk between 12-13% were reclassified to a risk of < 10% with Model F (for

risks between 11-12% and 10-11%, this was 43% and 57% respectively). Similar effects on the risk scores were found amongst the male cohort.

Figure 3.2: *Distribution of risks according to each model for those with risks 9 – 10% in model A*

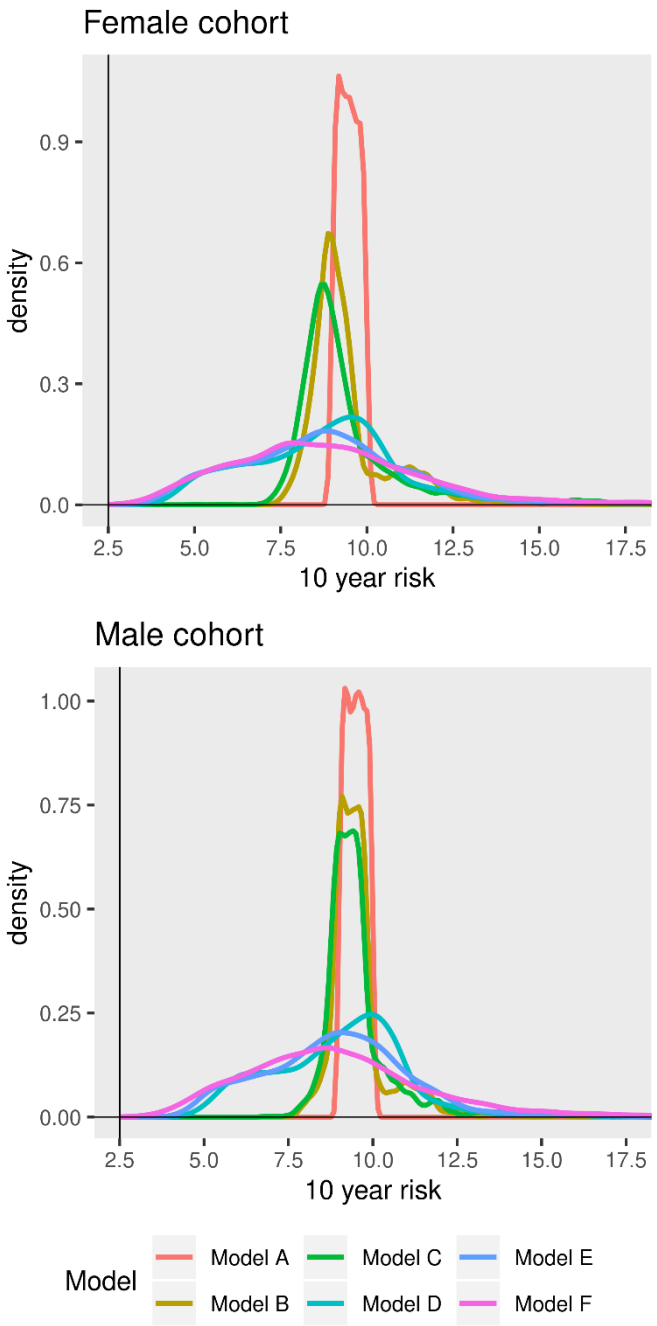


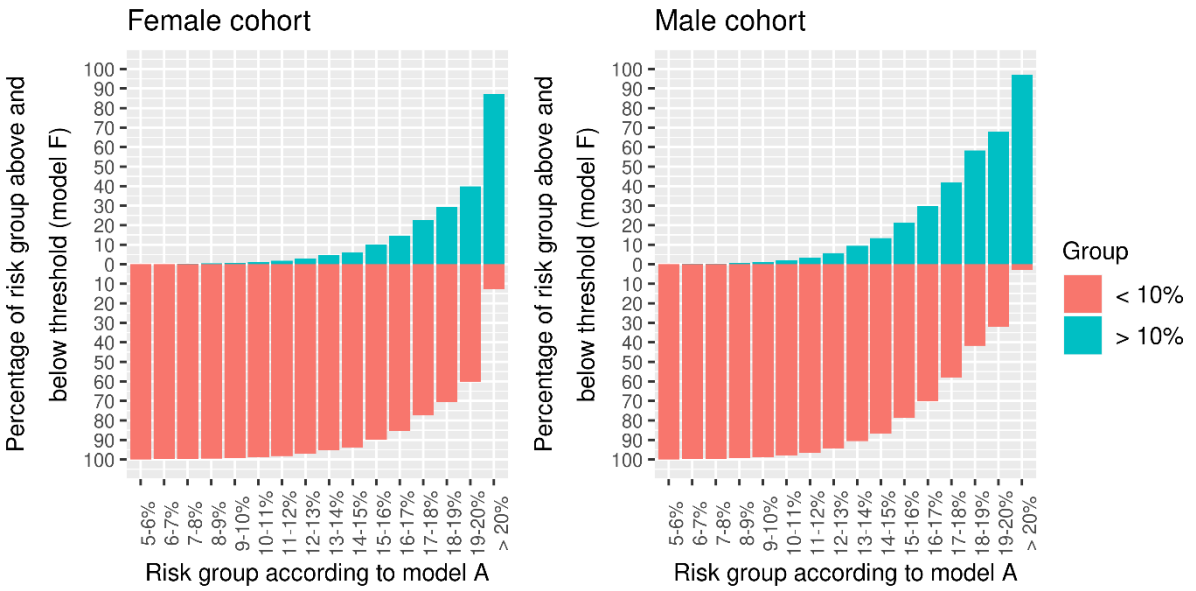
Table 3.5: *Numbers and percentages of patients who cross the treatment threshold (10 year CVD risk of 10%) when using models B – F instead of model A*

	Predicted 10 year CVD risk according to model A (QRISK2)						Predicted 10 year CVD risk according to model A (QRISK2)					
	5-6%	6-7%	7-8%	8-9%	9-10%	**< 10%**	10-11%	11-12%	12-13%	13-14%	14-15%	**> 10%**
Female (N = 200,000)												
N	5163	4356	3561	3080	2602	**169491**	2321	2086	1849	1617	1525	**30509**
Model B	21 (0%)	64 (1%)	105 (3%)	392 (13%)	547 (21%)	**1140 (1%)**	1000 (43%)	114 (5%)	3 (0%)	2 (0%)	4 (0%)	**1128 (4%)**
Model C	73 (1%)	142 (3%)	220 (6%)	374 (12%)	592 (23%)	**1446 (1%)**	1260 (54%)	264 (13%)	33 (2%)	7 (0%)	5 (0%)	**1575 (5%)**
model D	63 (1%)	108 (2%)	206 (6%)	388 (13%)	689 (26%)	**1498 (1%)**	1204 (52%)	757 (36%)	461 (25%)	286 (18%)	203 (13%)	**3252 (11%)**
model E	69 (1%)	116 (3%)	236 (7%)	466 (15%)	703 (27%)	**1649 (1%)**	1276 (55%)	833 (40%)	533 (29%)	320 (20%)	223 (15%)	**3645 (12%)**
model F	100 (2%)	184 (4%)	310 (9%)	540 (18%)	765 (29%)	**1978 (1%)**	1328 (57%)	906 (43%)	589 (32%)	367 (23%)	253 (17%)	**3953 (13%)**
Male (N = 200,000)												
N	7422	6077	5364	4483	3922	**158036**	3355	3108	2710	2421	2283	**41964**
Model B	17 (0%)	33 (1%)	91 (2%)	255 (6%)	583 (15%)	**983 (1%)**	733 (22%)	51 (2%)	0 (0%)	0 (0%)	0 (0%)	**784 (2%)**
Model C	37 (0%)	86 (1%)	200 (4%)	318 (7%)	728 (19%)	**1384 (1%)**	1042 (31%)	95 (3%)	4 (0%)	1 (0%)	1 (0%)	**1143 (3%)**
model D	25 (0%)	63 (1%)	164 (3%)	341 (8%)	1220 (31%)	**1823 (1%)**	1625 (48%)	1008 (32%)	627 (23%)	375 (15%)	266 (12%)	**4146 (10%)**
model E	30 (0%)	66 (1%)	191 (4%)	505 (11%)	1146 (29%)	**1953 (1%)**	1792 (53%)	1129 (36%)	705 (26%)	435 (18%)	302 (13%)	**4744 (11%)**
model F	126 (2%)	267 (4%)	483 (9%)	750 (17%)	1146 (29%)	**2850 (2%)**	1896 (57%)	1337 (43%)	888 (33%)	539 (22%)	375 (16%)	**5563 (13%)**

3.4.3 Extrapolation to English and population

Figure 3.3 shows the proportion of patients reclassified from each risk group when model F is used, applied to the cohort of patients eligible in CPRD for risk assessment on 1st January 2016. When using Model F, there was a substantive reclassification downwards across the higher risk categories, in which 64% of females and 52% of males with a risk >10% would no longer be eligible for statin treatment (Table A.3.8). This shift is caused by the introduction of the secular trend. When extrapolating results to the population of England, there were 37,273,200 people aged 25 – 84 in England[145] in 2016 and 29,382,463 would have been eligible for risk assessment using QRISK2 (79% of patients registered on 1st Jan 2016 were eligible). 6,652,920of these patients would be classified as high CVD risk (≥ 10%) using Model A (QRISK2). If model F was used, 3,792,474(57%) of them would be reclassified downwards and cross the treatment threshold. The 57% is calculated as average of the 64% of females and 52% of males, weighted by the female/male ratio. A full breakdown of these calculations and data used to derive Figure 3.3 is in Appendix A.3.3.1.

Figure 3.3: *Percentages of patients registered 1st Jan 2016 who cross the treatment threshold when using model F, compared to model A*



3.4.4 Post HOC and analyses of secular trend

There was a strong secular trend in CVD incidence in both the female and male derivation cohorts as can be seen in Figure 3.4. The RR was 0.96 (0.96-0.96) and 0.97 (0.97-0.97) annually for females and males respectively (Table 3.6). A stronger trend was found in the cohort of statin users, with a RR of 0.94 (0.94-0.94) for both cohorts. Adjusting for baseline QRISK2 score, the annual reduction in CVD incidence was unchanged from 0.94 (0.94-0.94) for the female cohort, and changed slightly to 0.94 (0.94 – 0.95) for the male cohort.

Figure 3.4: *The secular trend in CVD incidence in the model derivation cohort and the statin users cohort*

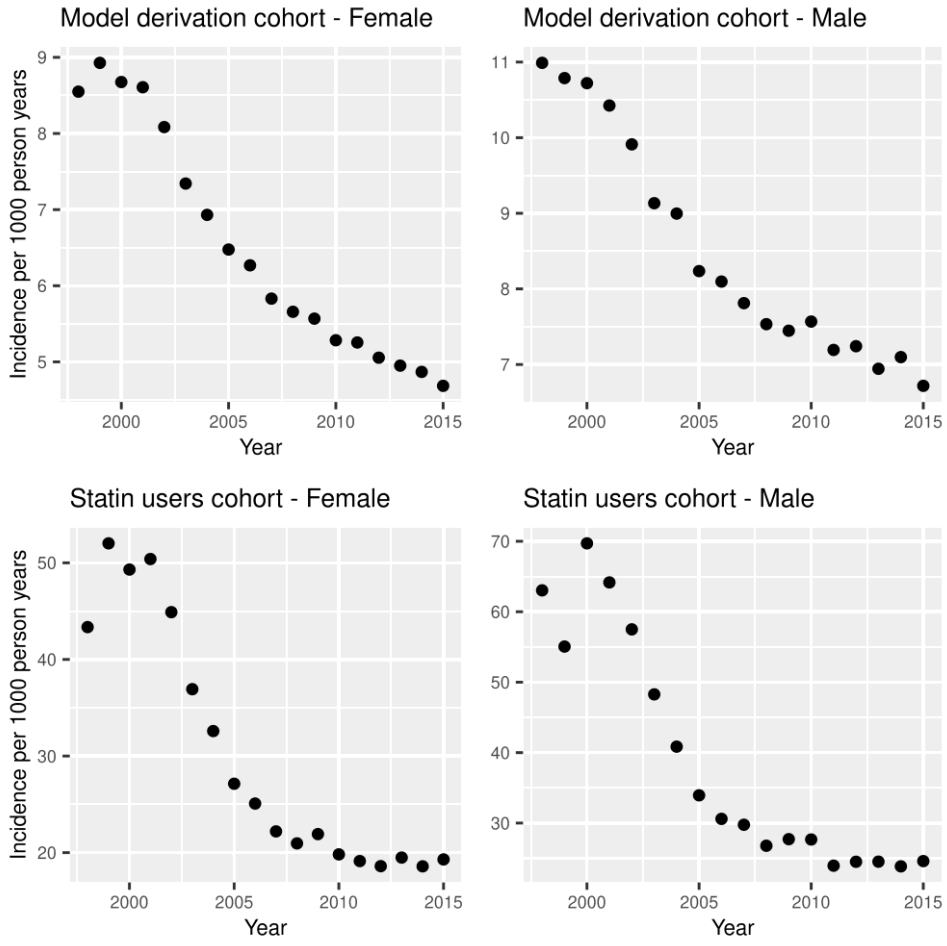


Table 3.6: *Relative rates (95% CI) associated with calendar year and risk at start of statin treatment period, in Poisson models modelling CVD incidence*

	Female		Male	
Model	**Calendar year**	**Risk at start of treatment**	**Calendar year**	**Risk at start of treatment**
Development cohort	0.96 (0.96 – 0.96)	NA	0.97 (0.97 – 0.97)	NA
Statin users cohort	0.94 (0.93 – 0.94)	NA	0.94 (0.94 – 0.94)	NA
Statin users cohort (also adjusting for 10 year risk at inception into cohort)	0.94 (0.94 - 0.94)	1.02 (1.02 – 1.03)	0.94 (0.94 – 0.94)	1.02 (1.02 – 1.02)

3.5 Discussion

In this study, we assessed the uncertainty in individual risk predictions by using different modelling approaches. A large amount of variability in individual risk predictions was found when taking into account different information about the patient. The introduction of secular trend substantially changed individual risk predictions. The largest uncertainty in individual risk prediction occurred in patients with higher risks (i.e., those who are considered for statin treatment) with a large number of patients being reclassified as no longer requiring statin treatment.

The QRISK models did not consider the secular trend and their follow-up was also restricted to more historic data (starting in 1998[5,146]). In the present study, the largest contributing factor to the within-person variability in the CVD estimates was the secular trend. After introducing the secular trend into the modelling, 62% of females and 51% of males in 2016 would be classified down from a CVD risk ≥ 10% to less than 10% risk and thus no longer be eligible for statin treatment according to guidelines. When extrapolating to the population in England, this could affect almost 4 million individuals. Other studies have also reported a reduction in the CVD incidence over time[53–55]. A nation-wide study in England reported that the rate of hospitalisations for acute myocardial infarction reduced by 5% annual between 2002 and 2010, which is similar to our estimates.[54] Better CVD prevention may have contributed to this decline, which could include an increase in statin use.[147] Given the use of these models is mandated in NICE guidelines, it is quite likely this is caused by QRISK resulting

in a prediction paradox,[148] and the increase in statin use could explain this secular trend. However our analyses found that the cohort of statin users also showed a decreased CVD risk over time, suggesting that other factors may have contributed to the decline in CVD incidence. It is important that clinicians and patients are made aware of this as inclusion of the secular trend into the QRISK models could massively reduce the number of patients who were eligible to receive treatment with statin therapy. There are many ways to address a secular trend in predictive models. The first is to re-calibrate the model to the time period of interest,[120,149] which is effectively what QRISK developers do by updating the time period in which they derive the model each year. However this still allows for a large un-modelled secular trend occurring between the study start and end date. This can also be done on a continuous scale using continuous model/Bayesian updating, and can be used with a forgetting factor to down weight historical data.[149] However this also constitutes developing a model in some data, and updating it in light of new data, and therefore suffers the same problems. Varying coefficient models are also available which allow the relationship between predictors and outcomes to vary over time.[149] Our approach is equivalent to a special case of these models, where only the intercept is allowed to vary over time. The use of varying coefficient models to model the secular trend should be considered in future work, although a more detailed assessment of whether the secular trend in associated with changes in database usage, and the role of statin use on the secular trend would have to be carried out.

Other factors also contributed to non-negligible levels of variability in risk prediction, for example the effect of using mean imputation to impute patient data. This is relevant because we found there to be missing data among the statin users cohort at statin initiation, which is the group of patients who should be having their risk assessed. For these patients, using mean imputation adds an avoidable level of uncertainty to the risk score. It is therefore important to measure all risk factors and include the measurements rather than relying on mean imputed values. Beyond this we highlighted the variability in risk scores caused by introducing a variety of risk factors into the models. All factors that were introduced into the models have been shown in the literature to be risk factors of CVD[23,46,53,127]. However there are many other factors that we could not evaluate, such as diet[150,151], level of physical inactivity,[152] an accurate measure of alcohol consumption, transaminase levels,[153] C-reactive protein levels[154] or biomarkers and genetic information.[155,156] This means the level of uncertainty associated

with a risk score is likely to be far higher than what we have been able to highlight in this paper. Despite this, there is no feasible way for these risk factors to be incorporated into a model used at point of care in routine practice, as they are not routinely recorded. We are not trying to recommend the collection and inclusion of such factors to improve the current models used in practice. Rather, we have highlighted that the introduction of new risk factors that could be measured has a sizeable effect on individual risk, and this effect would be bigger if one were able to collect such risk factors and incorporate them also.

This study found that widely used population-level performance metrics of risk predictions were not very sensitive with varying modelling approaches in contrast to the individual risk predictions. Harrell's C statistic[121] is the most commonly used performance metric but the comparisons between models showed marginal change. This finding is consistent with literature that reported that in well performing models, C statistics are not sensitive to the introduction of new covariates.[135,143] The measures of explained variation and randomness were more sensitive to the modelling decisions, mostly increasing by 0.02 across all the models. The D statistic showed the largest absolute increase, although this is unsurprising given it is not bounded by 0 and 1. While none of these metrics were developed to assess variability on the individual level, the large variability in individual risk, but lack of variability in population level performance metrics is of importance to the patient being treated. It should also be noted that there was a general trend of improved performance as variables were added to the models, potentially leading to the conclusion that adding any variable that may be associated with CVD will improve risk prediction. We do not believe this to be the case, and think the trend is likely explained by increasing amounts of overfitting as more variables are added to the model. Although split sample techniques were used to derive the performance metrics, the sample is very large and the test data is likely to be representative of the development cohort. You therefore would expect improved performance as more variables were added when carrying out internal validation. National treatment guidelines in the UK state that *"all CVD risk assessment tools can provide only an approximate value for CVD risk"* and that *"interpretation of CVD risk scores should always reflect informed clinical judgement"*.[46] Our results highlight the importance of this, considering clinical judgement and supplementing these model estimates with evidence on additional risk factors. Despite this recommendation, our experience is that output from QRISK is regularly used to guide

treatment decisions, while confusion remains around its interpretation.[157] Furthermore, there has been a recent push by Public Health England[158,159] for self-assessment by the public of risk using a tool JBS3[160] which is based on the lifetime QRISK model.[161] Arguably, patients will need to be informed about the approximate estimates of these tools and the need for clinical judgement. This is very much an issue about communication of the limitations of such estimates, rather than an issue with the models themselves. It may be important not to communicate a single value which does not take into account important risk factors such as diet, exercise and life style[162], the severity of presenting comorbidities or the uncertainty underlying the modelling decisions.

3.5.1 Limitations

There are several limitations in this study. While the dataset used to derive the models is similar to that used to derive QRISK3 in terms of demographics, there may be many other hidden differences between the datasets, for example geographical coverage or coding practices between the databases. This means our models do not directly represent the ones used in practice in the England. One limitation was that a crude disease classification was used to derive many of the predictor variables. A combination of medical and/or prescription codes were used which may be sensitive to the choice of the code lists. Another limitation of this study was that important information on other risk factors was missing (such as diet or exercise), which could explain a large amount of unexplained variation in risk. Frailty models were considered to quantify the level of unexplained variation in patient risk due to missing covariates.[163] However we were unable to fit these models in a consistent fashion to the data, while also finding strong arguments against this methodology.[164] We also did not consider the variability in coding between practices, or between databases. Models may perform erroneously when used in a database in which it was not developed an issue which has caused issues in recent history.[123] For example how will a model perform in a database that uses a different coding system? This was not considered in this study as data from two databases with different coding systems was not available; however is an important area for future research. Finally, this paper focused on uncertainty induced by considering different information about the patient. However there may also be uncertainty associated with the risk scores caused by various modelling decisions. For example in models developed in this way the target population is not well defined. The association of covariates with the outcome

may change with age, and although interaction terms are included it is difficult to truly model these relationships. Given these models are used to generate risk scores for patients over a wide age range, this could also induce uncertainty on the patient level. There are many other methodological choices which induce uncertainty, which should be explored in their own right. This paper focuses primarily on the choice of what information about the patients to include in the models.

3.5.2 Conclusions

In conclusion, we found sizeable levels of uncertainty in the prediction of individual CVD risks for patients, although this was mostly driven by the introduction of secular trend. This high level of instability was not detected with conventional population-level model performance metrics (in particular Harrell's C, the most commonly used measure of discrimination). Extrapolating to the population in England, 3.8 million patients could be misclassified as requiring statin treatment depending on the model used. Clinical judgement, as recommended in national treatment guidelines,[46] supplemented with evidence of additional risk factors, should be an essential part of individual decision making. Uncertainty analyses with varying of modelling choices and quantification of incomplete evidence should routinely be conducted to assess uncertainty beyond the confidence interval.

4 Impact of sample size on the stability of risk scores from clinical prediction models: a case study in cardiovascular disease

Alexander Pate, Richard Emsley, Matthew Sperrin, Glen P. Martin, Tjeerd van Staa

4.1 Abstract

Background: Stability of risk estimates from prediction models may be highly dependent on the sample size of the dataset available for model derivation. In this paper, we evaluate the stability of cardiovascular disease risk scores for individual patients when using different sample sizes for model derivation; such sample sizes include those similar to models recommended in national guidelines, and those based on recently published sample size formula for prediction models.

Methods: We mimicked the process of sampling N patients from a population to develop a risk prediction model by sampling patients from the Clinical Practice Research Datalink. A cardiovascular disease risk prediction model was developed on this sample and used to generate risk scores for an independent cohort of patients. This process was repeated 1000 times, giving a distribution of risks for each patient. N = 100 000, 50 000, 10 000 and N_{min} (derived from sample size formula) were considered. The 2.5 – 97.5 percentile range of risks across these models was used to evaluate instability. Patients were grouped by a risk derived from a model developed on the entire population (population derived risk) to summarise results.

Results: For a sample size of 10 000, the median 2.5 – 97.5 percentile range of risks for patients across the 1000 models was approximately 60% of their population derived risk. For example, for patients with a population derived risk of 9 - 10% or 19 - 20%, the median percentile range was 6.25% and 12.59% respectively. Using the formula derived sample size, the range was approximately 170% of their average risk score. Restricting this analysis to models with high discrimination or good calibration reduced the percentile range, but high levels of instability remained.

Conclusions: Widely used cardiovascular disease risk prediction models suffer from high levels of instability induced by sampling variation. Stability of risk estimates should be a criterion when determining the minimum sample size to develop models.

4.2 Background

Risk prediction models are used to guide clinical decision-making in a variety of disease areas and settings, ranging from the prevention of cardiovascular disease (CVD) in primary care to intensive care unit based models such as APACHE or SOFA.[3,4,32–34] As such, developing risk prediction models appropriately is vital. One aspect of appropriate derivation of prediction models is ensuring sufficient sample size in the development dataset; unfortunately, sample size calculations for models are often not made, or at best are based on the simplistic "10 events per predictor" rule.[165] Risk prediction models that are recommended in treatment guidelines for routine use by clinicians often vary in sample sizes. As an example, QRISK3[5] (recommended by the National Institute for Health and Care Excellence to guide CVD primary prevention in England[46]) was developed on a cohort of 4 019 956 females and 3 869 847 males, whereas the pooled cohort equations (recommended by American College of Cardiology and American Heart Association to guide CVD prevention in the US[60]) were based on 9098 females and 11 240 males for white ethnicity, and 2641 females and 1647 males for African-American ethnicity.

If the sample size is too small, the most commonly cited issue is that of overfitting, which may cause extreme predictions outside of the development data set. Another potential issue, of which the implications are less clear, is that small sample sizes could lead to instability in the risk scores of individuals depending on which sample of the population is used for model development. By stability, we mean how risk scores for a given individual vary when generated from different prediction models. It is well known that differently defined prediction models may produce different risks for individuals, even if the models perform similarly on the population level.[13–15,18,19] However, if a patient's risk score, and therefore treatment decision, is highly unstable due to sample size, this is undesirable. In this scenario, the instability of a patient's risk score is driven by statistical uncertainty around the risk estimate of the subgroup which that patient belongs to, distinguishing this from the reference class problem.[15] Therefore it is important to minimise this instability if wanting to base clinical decisions on risk scores generated from such models.

The aim of this study was to evaluate the stability of CVD risk predictions for individual patients when using different sample sizes in the development of the risk prediction models

(including a recommended minimum sample size from work focusing on the issue of overfitting, representing state of the art techniques for sample size calculations in risk prediction models).[35]

4.3 Methods

4.3.1 Data source

We defined two cohorts from a Clinical Practice Research Datalink (CPRD)[48] dataset, which comprised primary care data linked with Hospital Episode Statistics (HES),[49] and mortality data provided by the Office for National Statistics (ONS).[50] For the first cohort (referred to as historical cohort) the cohort entry date was the latest of: attaining age 25 years; attaining one year follow up as a permanently registered patient in CPRD; or 1st Jan 1998. The end of follow up was the earliest date of: patient's transfer out of the practice or death; last data collection for practice; or 31st Dec 2015. Patients were excluded if they had a CVD event (identified through CPRD, HES or ONS) or statin prescription prior to their cohort entry date (code lists available on GitHub[103]). The second cohort comprised patients actively registered on 1st Jan 2016 (referred to as contemporary cohort). This cohort of patients represents a contemporary population, for whom a risk prediction model would subsequently be applied to estimate their CVD risks. To be eligible for this second cohort, a patient had to be aged 25 – 85 years on 1st Jan 2016, and be actively registered in CPRD with one year prior follow up with no history of CVD or statin treatment. For both cohorts, all predictor variables included in the QRISK3[5] risk prediction model were extracted at cohort entry date. Details on variable definitions is provided in Chapter 2.

4.3.2 Overview of methods

We mimicked the process of sampling an overarching target population for the development of a risk prediction model by randomly sampling N patients from the historical cohort (CPRD is representative of the UK in terms of age, sex and ethnicity[48]). A risk prediction model was developed on this sample and used to generate risk scores for the contemporary cohort. This process was repeated 1000 times, giving 1000 risk scores for each patient, for each sample size. The sample sizes considered were N = 10 000, 50 000, 100 000 and N_{min} (minimum

sample size required to meet criteria outlined by Riley et al.[35]). We chose 10 000 as it is similar to the number of females and males used to develop ASSIGN[63] (6540 and 6757), Framingham[6] (3969 and 4522) and Pooled Cohort Equations[60] (9098 and 11 240). The upper limit of 100 000 was chosen to match the SCORE[65] equations, which were developed on 117 098 and 88 080 females and males respectively. Derivation of N_{min} = 1434 (female) and 1405 (male) is described in Appendix A.4.1.

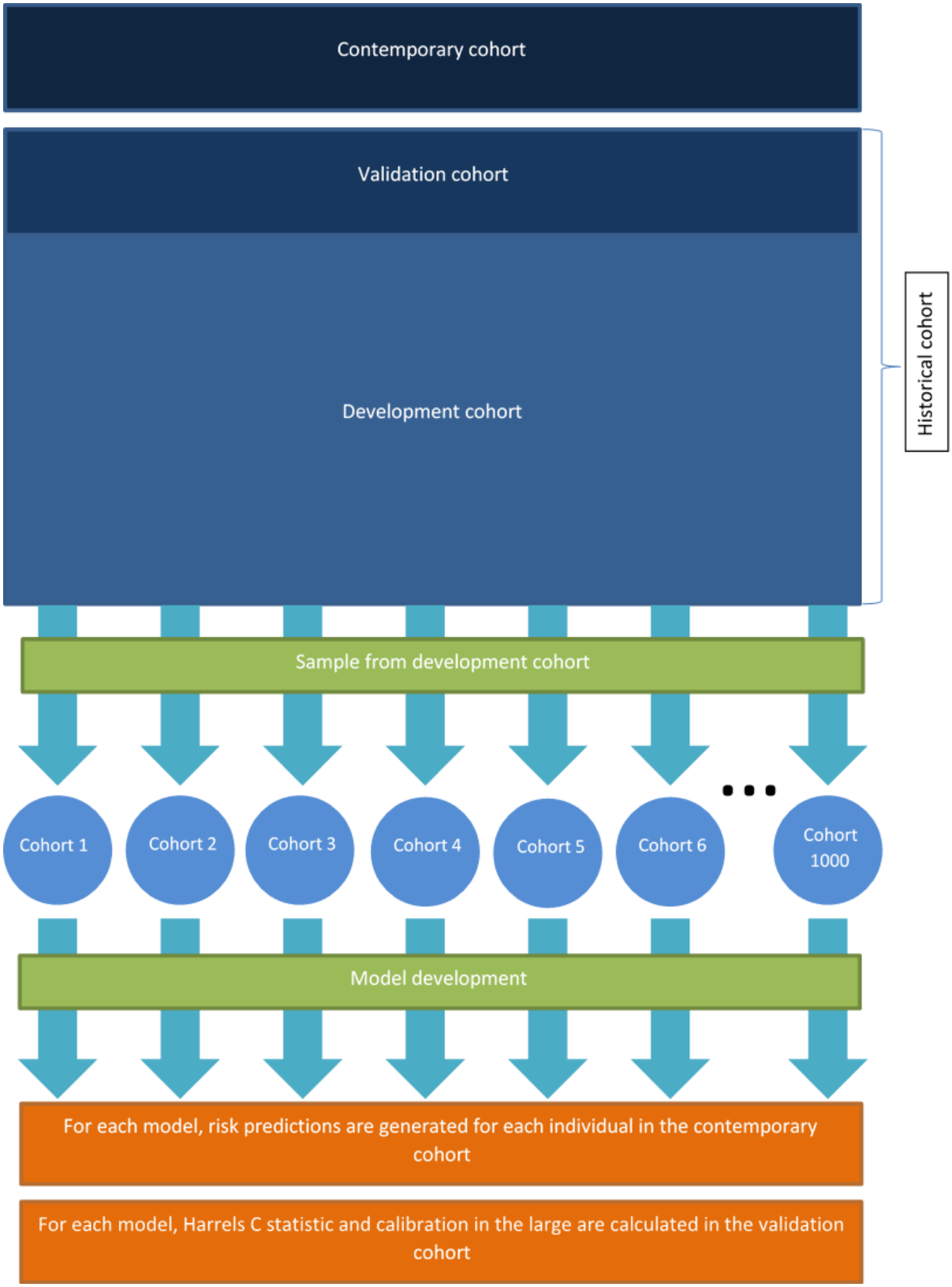
4.3.3 Generation of risk scores

The historical cohort and contemporary cohort were both split into female and male cohorts and imputed using one stochastic imputation using the mice package.[130] All variables included in QRISK3,[5] including the Nelson Aalen estimate of the baseline cumulative hazard at the event time and the outcome indicator, were included in the imputation process. The following process was then carried out separately for females and males: 100 000 individuals were chosen at random from the historical cohort to form an internal validation cohort, the remaining individuals formed the development cohort. The development cohort was then viewed as the population. For each value of N, we sampled N patients from this population without replacement, 1000 times.

The following process was repeated within each sample. A Cox model was fit to the sampled data, where the outcome was defined as the time until the first CVD event. Predictor variables included in the model were continuous variables, and categorical variables with > 1% prevalence in all categories calculated from the entire development cohort (age, body mass index (BMI), cholesterol/high density lipoprotein (HDL) ratio, family history of CVD, treated hypertension, smoking status, systolic blood pressure (SBP), Townsend deprivation index and type 2 diabetes). This set of variables reflects the smaller number of variables used in models with lower sample sizes in practice.[6,60,63] The developed model was used to generate 10 year risk scores for the contemporary cohort. Harrell's C^{121} statistic for this model, and the calibration-in-the-large (mean predicted risk – observed/Kaplan Meier risk) were calculated in the validation cohort. A graphical representation of this process is given in Figure 4.1.

Finally, we calculated a 10 year risk for each patient in the contemporary cohort using a model developed on the entire development cohort, called the population derived risk, and also calculated the Harrell's C and calibration-in-the-large of this model in the validation cohort.

Figure 4.1: *A graphical representation of the sampling process*



4.3.4 Analysis of stability of risk scores

For each sample size, four different analyses were carried out to summarise the stability of risks across the 1000 models. First, the 2.5 – 97.5 percentile range of risks was calculated for each patient across the 1000 models. The distribution of these ranges was then plotted in box plots stratified by the population derived risk. Second, we split the models into three groups of equal size that had the lowest, medium or highest C statistics. We then calculated the 2.5 - 97.5 percentile range of risks within these subsets of models, and presented in box plots stratified by population derived risk. This allowed us to explore whether models with high C statistics had more stability than those with lower C statistics. Third, we split the models into groups defined by their calibration-in-the-large, and presented boxplots of the 2.5 - 97.5 percentile range of risks within these subsets of models. Here, the groups were defined as models with calibration-in-the-large deviating from the population derived model by less than 0.1%, 0.25%, 0.5%, and then all models. This allowed us to explore how much of the instability of the risk scores was driven by variation in overall calibration. Finally, we grouped patients into risk groups of width 1% as defined by their population derived risk. The proportion of the 1000 models that classified a patient above/below the 10% risk threshold (threshold for statin eligibility according to the recommended guidelines in the UK[46]) was then calculated.

Also, the shrinkage factor for each model generated using N = N_{min} was calculated and is provided in Appendix A.4.1.

4.4 Results

The baseline characteristics for the female development cohort, validation cohort, and the contemporary cohort are provided in Table 4.1. See Appendix 0 for the equivalent table for the male cohort.

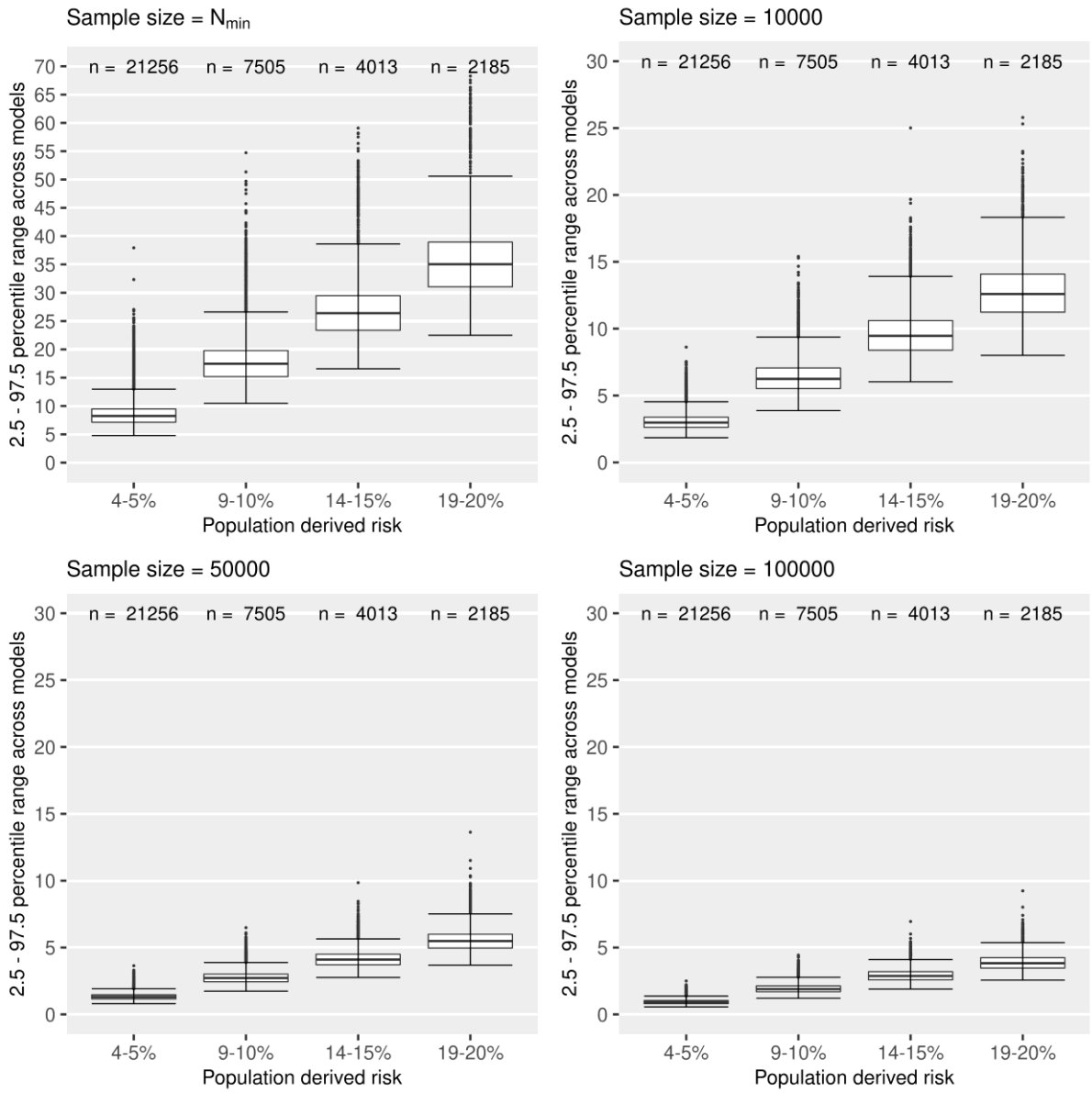
Figure 4.2 plots the 2.5 – 97.5 percentile range in risks for patients across the 1000 models, grouped by population derived risk (female cohort). For N = 100 000, the median 2.5 - 97.5 percentile range was 0.91%, 1.90%, 2.87% and 3.83% for patients in the 4-5%, 9-10%, 14-15% and 19-20% risk groups respectively. For N = 50 000, the median percentile range was 1.31%, 2.72%, 4.10% and 5.49% in the respective groups, for N = 10 000 it was 2.98%, 6.25%, 9.46% and 12.59%, and for N = N_{min} it was 8.25%, 17.46%, 26.40% and 35.05%. For each sample size,

the median percentile ranges were approximately constant on the relative scale compared to the population derived risk of the group, i.e, for a sample size of 10 000 the median percentile range was approximately 60% of the population derived risk, and for N_{min} it was approximately 170%. Results for the male cohort followed a similar pattern, but the level of instability was slightly lower (Appendix 0).

Table 4.1: *Baseline characteristics of each female cohort*

Variable	Category	Development (n=1 865 078)	Validation (n = 100 000)	Contemporary (n = 387 557)
Outcome	CVD events	82 065	4482	NA
	Follow up (years)	13 098 449	703 471	NA
Age		43.07 (15.94)	43.14 (15.96)	48.38 (14.43)
SBP		123.91 (18.28)	124 (18.22)	123.97 (15.17)
BMI		25.6 (5.60)	25.56 (5.56)	27.1 (6.31)
Cholesterol/HDL ratio		3.72 (1.20)	3.72 (1.21)	3.46 (1.04)
Smoking status	Never	56.04	56.15	46.05
	Ex	16.97	16.98	31.66
	Current	27.00	26.87	22.29
Townsend	1 (least deprived)	21.96%	21.96%	24.95%
	2	21.99%	21.81%	22.35%
	3	21.17%	21.46%	21.56%
	4	20.46%	20.36%	18.70%
	5 (most deprived)	14.42%	14.41%	12.44%
Treated hypertension		6.18%	6.19%	8.45%
Family history of CVD		15.08%	15.13%	20.86%
Type 2 diabetes		1.16%	1.19%	1.15%

Figure 4.2: *Boxplots of the percentile ranges of risk for individuals across the 1000 models (female cohort)*



**Each data point represents the 2.5 = 97.5 percentile range in risk for an individual across the 1000 models*

The distribution of the C statistic and the calibration-in-the-large of the 1000 models are given in Table 4.2. The 97.5th percentile of C statistics was similar for each sample size, but as the sample size decreased, the 2.5th percentile got smaller (0.802 vs 0.868 female and 0.805 vs 0.843 male). All C statistics in the 2.5 – 97.5 percentile range were > 0.8. The variation in the calibration-in-the-large decreased as the sample size increased. The 2.5 – 97.5 percentile

ranges of the calibration-in-the-large values was 2.61% (female) and 3.12% (male) for N = N_{min}, decreasing to 0.32% (female) and 0.36% (male) for N = 100 000.

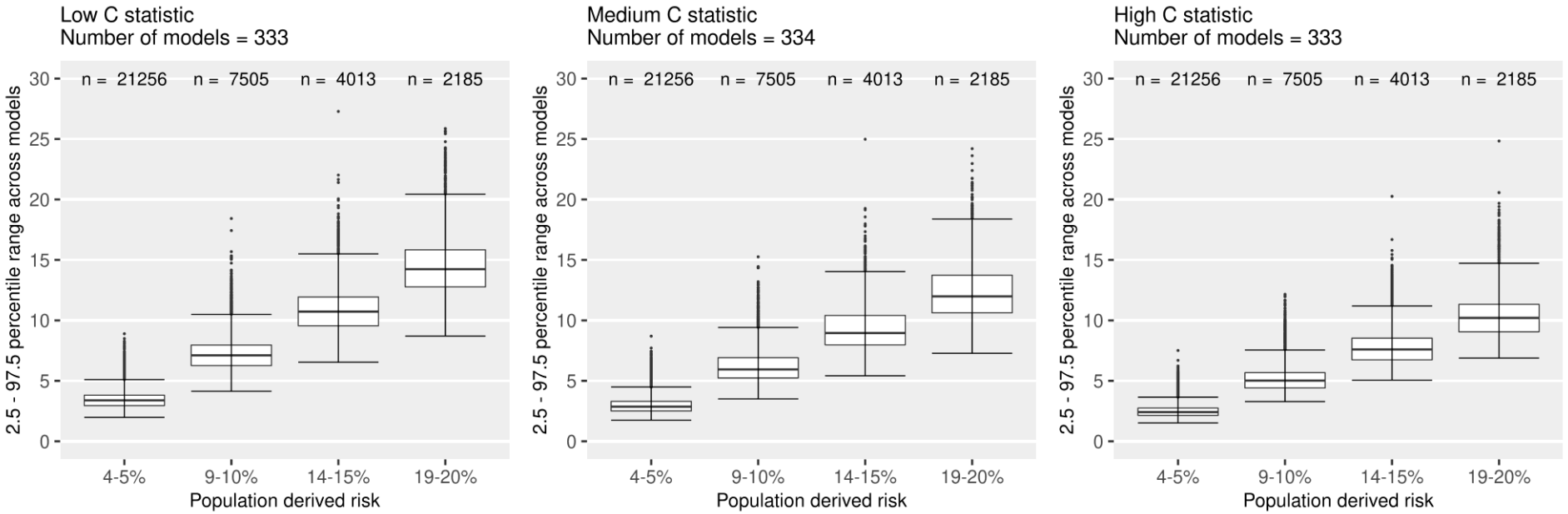
Table 4.2: *Quantiles of C statistics and calibration-in-the-large of the 1000 models, for each sample size*

		Quantiles of C statistic					**Quantiles of calibration-in-the-large (as a %)**				
	Sample size	2.5%	25%	50%	75%	97.5%	2.5%	25%	50%	75%	97.5%
Female	**N_{min}**	0.802	0.852	0.857	0.861	0.864	-2.22	-1.43	-0.95	-0.47	0.39
	10000	0.865	0.866	0.867	0.867	0.868	-1.45	-1.13	-0.95	-0.78	-0.44
	50000	0.867	0.868	0.868	0.868	0.868	-1.18	-1.03	-0.95	-0.87	-0.73
	100000	0.868	0.868	0.868	0.868	0.868	-1.11	-1.01	-0.96	-0.90	-0.79
Male	**N_{min}**	0.805	0.827	0.831	0.835	0.839	-2.56	-1.49	-1.01	-0.45	0.56
	10000	0.840	0.841	0.842	0.843	0.843	-1.61	-1.20	-1.01	-0.80	-0.39
	50000	0.843	0.843	0.843	0.843	0.844	-1.28	-1.11	-1.02	-0.93	-0.77
	100000	0.843	0.843	0.843	0.844	0.844	-1.21	-1.08	-1.02	-0.95	-0.85

*C statistics of population derived models in the validation dataset are 0.868 (female) and 0.844 (male). Calibration-in-the-large of the population derived models in the validation dataset are -0.95% (female) and -1.02% (male).

Figure 4.3 plots the 2.5 – 97.5 percentile range in risks for patients across models stratified by the C statistic of the models (female cohort, N = 10 000). The median 2.5 - 97.5 percentile range for models with high C statistics was 2.42%, 5.02%, 7.60% and 10.20% for patients in the respective risk groups. This equates to an 18 – 20% reduction in the median percentile range when using well discriminating models compared to all models (2.98%, 6.25%, 9.46% and 12.59%). Results for other sample sizes presented in Appendix 0.

Figure 4.3: *Percentile range of risk for individuals across subsets of the 1000 models, defined by the C-statistic of the models (female cohort, N = 10 000)*



**Each data point represents the 2.5-97.5 percentile range in risk for an individual across a group of models defined by their C-statistics.*

Figure 4.4 plots the 2.5 – 97.5 percentile range in risks for patients across models stratified by the calibration-in-the-large of the models (female cohort, N = 10 000). The median 2.5 - 97.5 percentile range across models with the best calibration-in-the-large was 2.72%, 5.70%, 8.72% and 11.69%, for the respective risk groups. This equates to a 7-9% reduction in the median percentile range compared to when using all models (2.98%, 6.25%, 9.46% and 12.59%). Results for other sample sizes presented in Appendix 0.

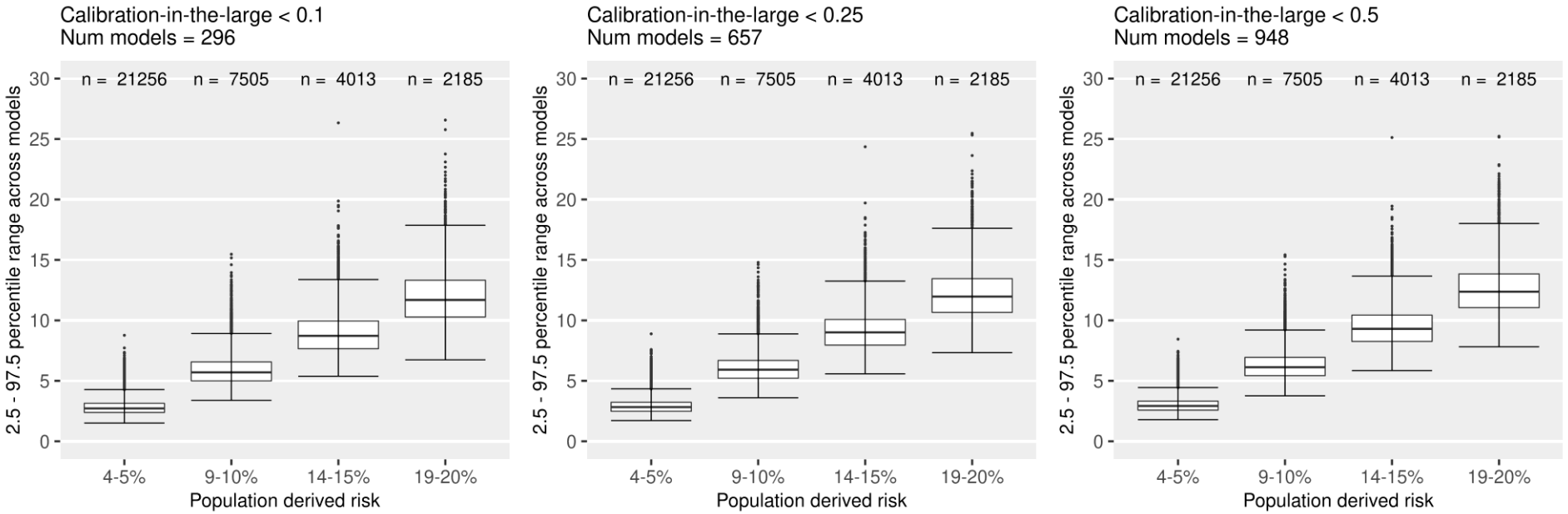
Table 4.3 shows the probability that a patient from a given risk group (according to population derived model) may be classified on the opposite side of the 10% threshold by a randomly chosen model. For example when using a sample size of N_{min}, 26.91% of patients with a population derived risk between 14-15% would be classified as having a risk below 10%, whereas this is only 2.50% for N = 10 000, 0.01% for 50 000 and < 0.01% for 100 000.

Table 4.3: *Probability of being classified on the opposite side of the treatment threshold by a randomly selected model of a given sample size, stratified by population derived risk*

		Population derived risk									
	Sample size	5-6%	6-7%	7-8%	8-9%	9-10%	10-11%	11-12%	12-13%	13-14%	14-15%
Female	**N_{min}**	6.46	12.55	20.49	29.63	38.69	52.48	44.46	37.72	31.95	26.91
	10,000	0.08	0.74	4.25	15.24	35.07	41.17	22.55	11.40	5.50	2.50
	50,000	0.00	0.00	0.08	2.29	24.49	27.67	4.44	0.46	0.06	0.01
	100,000	0.00	0.00	0.00	0.50	18.56	21.50	1.09	0.04	0.00	0.00
Male	**N_{min}**	4.32	9.98	18.18	28.54	39.14	50.87	41.97	34.37	28.13	22.97
	10,000	0.03	0.33	2.51	12.40	34.43	38.89	18.84	8.13	3.33	1.34
	50,000	0.00	0.00	0.02	1.28	21.80	26.51	3.07	0.26	0.03	0.00
	100,000	0.00	0.00	0.00	0.23	16.02	19.79	0.63	0.01	0.00	0.00

**For patients with a population derived risk < 10%, the probabilities represent the chance of being classified above the threshold, for patients with a population derived risk > 10%, the probabilities represent the chance of being classified below the threshold.*

Figure 4.4: *Percentile ranges of risk for individuals across subsets of the 1000 models, defined by the calibration-in-the-large of the models (female cohort, N = 10 000)*



**Each data point represents the 2.5 – 97.5 percentile range in risk for an individual across a group of models defined by their calibration in the large*

4.5 Discussion

This study found that at sample sizes typically used for developing risk models (e.g. in the CVD domain, the pooled cohort equations[60] and ASSIGN[63] were based on approximately 10 000 individuals or less), there is substantial instability in risk estimates attributable to sampling error. Furthermore, when restricting the analysis to models with high discrimination or good calibration, high levels of instability remained.

This variability in individual risk is especially relevant if using the model to make clinical decisions based on whether a risk score is above or below a fixed threshold (a common use for risk prediction models). From an individual's and clinician's perspective, it is unsatisfactory that a different treatment decision may be made dependent on the model used. However this is also an issue at the population level. Consider statin therapy in the UK. Initiating statins in patients who have a 10-year risk of CVD > 10% has been shown to be cost effective.[84] This intervention becomes more cost effective the better the performance (calibration and discrimination) of the model used to calculate the risk scores. Sample size is strongly correlated with model performance, and a small sample size will likely lead to a poorly performing model, and less events prevented. However, it is difficult to assess when increasing sample size will improve model performance, given that model performance is affected by many other factors (prevalence of outcome, inclusion of important predictors, strength of association between predictors and outcome). Sample size affects model performance through the precision of coefficients, and imprecise estimates will cause the risk of fixed subgroups in the population to be miss-calculated (the central theme of this paper). Therefore, if the coefficients are precise, and risk estimates are stable, one will unlikely be able to improve model performance by increasing the sample size unless doing so allows for incorporating more predictors. The stability of risk scores (and ultimately precision of coefficients) could therefore be used as a proxy to determine whether increasing sample size will improve model performance. When N = 10 000 we see levels of instability that indicate the performance of the model could be improved by increasing sample size, resulting in fewer CVD events.

At the sample size suggested by Riley et al.[35] the instability in risk is even higher and the issues are heightened. However, there are no CVD risk prediction models used in practice with such small sample sizes, so the implications are more general. There is often ample data to produce

CVD risk prediction models; however this may not be the case for other disease areas, where the outcomes are not well recorded in routinely collected datasets. In this scenario one may have to actively recruit patients into a cohort and the work by Riley et al.[35] could be used in order to derive a sample size. We propose that if risk scores from a model are going to be used to drive clinical decision making above or below a fixed threshold, section 6 of Riley et al.[35] "*Potential additional criterion: precise estimates of predictor effects*", should be properly considered. It is imprecise estimates of the predictor effects that leads to instability of risk scores. If this criterion is not met, as is the case for $N = N_{min}$ in this paper, risks scores have high levels of instability and models poorer performance. The number of patients required to ensure stable risk scores will depend on the prevalence of the outcome, the number of predictors and the strength of the association between outcomes and predictors among other things, and therefore will vary for each model.

In practice, to ascertain whether a given development cohort has a sufficient sample size, the process undertaken in this manuscript could be replicated using bootstrap resampling methods. Instead of sampling the population without replacement (not possible in practice), sampling the development cohort with replacement (i.e. bootstrapping) can replicate this process and one could obtain a similar range of risks for each patient. The stability of the risk scores could then be assessed, and a decision made on whether more patients should be recruited. One proposal on how to use this information to determine a sufficient sample size could be to ensure the bootstrapped 2.5 - 97.5 percentile range for all patients must be smaller than x% of their estimated risk. Another proposal may be to ensure that for patients whose estimates are a certain distance away from a treatment threshold, that there is a less than an x% chance of deriving a risk on the other side of the treatment threshold if one resampled.

4.5.1 Limitations

There are some limitations that warrant discussion. The first is that the calibration-in-the-large of the population derived model was poor. We don't believe this is a problem as a similar miss calibration-in-the-large is found in QRISK3,[5] despite the model being well calibrated within risk deciles. It is likely caused by incompatible assumptions under how the observed risks (Kaplan Meier assumes unconditional independent censoring) and predicted risks (Cox

model assumes independent censoring only after conditioning on the covariates) are estimated. When looking within risk deciles, the difference in assumptions is not as large and good calibration was found. Centring these measurements thus allowed the evaluation of whether the instability in risk was being driven by over and under predicting models. A second limitation was that one may argue that variation in predicted risk was observed because the proper process for deriving risk prediction models wasn't followed. We didn't do this as it would have resulted in different variables and non-linear terms being selected across the models, and we believe this would have increased the variation in risks across the models, rather than reduce it. Finally, this study concerned the outcome CVD and used a specific set of variables for prediction. However the results are likely to be generalizable to other disease areas as the study evaluated the effects of random variability in sampling.

4.5.2 Conclusions

In conclusion, CVD risk prediction models developed on randomly sampled cohorts of size 10 000 or less suffer from high levels of instability in individual risk predictions. There are multiple models used in practice that are developed on sample sizes this small. To avoid this, models should be developed on larger cohorts such as the QRISK3[5] and SCORE[65] models. More generally, if developing a risk prediction model to guide treatment for patients above a fixed threshold, consideration should be given to the stability of risks scores and precision of effect estimates when choosing a sample size.

5 An assessment of the potential miscalibration of cardiovascular disease risk predictions caused by a secular trend in cardiovascular disease in England

Alexander Pate, Tjeerd van Staa, Richard Emsley

5.1 Abstract

Background: A downwards secular trend in the incidence of cardiovascular disease (CVD) in England was identified through previous work and the literature. Risk prediction models for primary prevention of CVD do not model this secular trend, this could result in over prediction of risk for individuals in the present day. We evaluate the effects of modelling this secular trend, and also assess whether it is driven by an increase in statin use during follow up.

Methods: We derived a cohort of patients (1998 – 2015) eligible for cardiovascular risk prediction from the Clinical Practice Research Datalink with linked hospitalisation and mortality records (N = 3,855,660). Patients were split into development and validation cohort based on their cohort entry date (before/after 2010). The calibration of a CVD risk prediction model developed in the development cohort was tested in the validation cohort. The calibration was also assessed after modelling the secular trend. Finally, the presence of the secular trend was evaluated under a marginal structural model framework, where the effect of statin treatment during follow up is adjusted for.

Results: Substantial over prediction of risks in the validation cohort was found when not modelling the secular trend. This miscalibration could be minimised if one was to explicitly model the secular trend. The secular trend was still present under the marginal structural model framework, indicating increasing statin use during follow up is not the cause.

Conclusions: Inclusion of the secular trend into the model substantially changed the CVD risk predictions. Models that are being used in clinical practice in the UK do not model secular trend and may thus overestimate the risks, possibly leading to patients being treated unnecessarily.

5.2 Background

Cardiovascular disease (CVD) risk prediction models such as QRISK are developed on longitudinal data spanning a long period of time (QRISK3 runs from 1998 – 2015[5]). These models are updated each year to include the most recent data and at times remove old data. However, any secular trend in the outcome itself occurring within the time span of the development dataset is not modelled. Pate et al.[1] found a large downwards secular trend in CVD incidence over this time period in England. Downwards secular trends in the incidence of coronary heart disease, myocardial infarction, and stroke have also been reported in the literature.[53–56] Not including this trend in the prediction modelling could be resulting in the miscalibration of risk scores for patients in the present day, while including it would cause a large reduction in the predicted risks of these patients. Further research around this is needed, to quantify the impact of modelling this secular trend, and identify what is driving it and whether it should be modelled or not. One important possible cause is if the secular trend is being driven by an increase in statin use over time. In this scenario it should not be modelled, as it would result in risks predictions becoming lower and patients would be subsequently advised not to initiate statin treatment, despite this being the cause for the drop in risk.

In this paper we evaluate the effects of developing a model using the same methodology as QRISK3 (in the presence of the secular trend) and producing risk scores for patients in a time period after that of model development. We then propose an approach to incorporate secular trends in prediction models from longitudinal data, accounting for changes in treatment during follow up. This is formalised in four sequential analyses: A) quantifying the miscalibration in risk predictions of patients in the present day caused by this secular trend, B) assessing the sensitivity of the risk prediction model created to changes in patient characteristics, which could explain any miscalibration, C) an attempt to model the secular trend to remove miscalibration, D) developing a marginal structural model (MSM) to assess secular trend after adjusting for statin use during follow up.

5.3 Methods

All analyses are carried out separately for male and female cohorts, as they have separate CVD risk prediction models in practice.

5.3.1 Data source

A 'CVD primary prevention cohort' was defined from a Clinical Practice Research Datalink (CPRD)[48] dataset linked with Hospital Episode Statistics[49] (HES) and Office for National Statistics[50] (ONS) using the same criteria as QRISK3.[5] The study period was 1st Jan 1998 to 31st Dec 2015 and the cohort entry date defined as the latest of: date turned 25; one year follow up as a permanently registered patient in CPRD; or 1st Jan 1998. Patients were excluded if they had a CVD event (identified through CPRD, HES or ONS) or statin prescription prior to their cohort entry date. The end of follow up was: the earliest date of patient's transfer out of the practice or death; last data collection for practice; 31st Dec 2015 or five years follow up. Patients were censored after five years as five year risk predictions are used throughout this chapter. All predictor variables included in the QRISK3[5] risk prediction model were extracted at cohort entry date. Code lists and detailed information on how variables were defined is provided in Chapter 2.

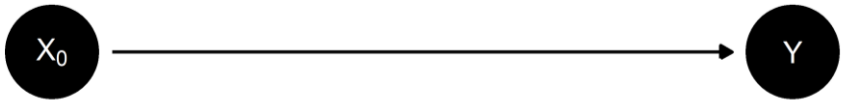
5.3.2 Quantifying the miscalibration in risk predictions of patients in the present day

The first step was to quantify the miscalibration induced by developing a model over a time period in which a secular trend in CVD was present, and using it to calculate risk predictions for patients after this time period. Missing data for body mass index (BMI), systolic blood pressure (SBP), SBP variability, cholesterol, high density lipoprotein (HDL), smoking status and ethnicity in the CVD primary prevention cohort was imputed using multiple imputation by chained equations. The imputation model included all predictor variables from QRISK3, the Nelson Aalen estimation of the cumulative baseline hazard at the point of censoring or an event, and the outcome indicator. Only one imputed dataset was produced, as running the analysis across multiple datasets and combining estimates was not essential to answering our hypotheses, and the computational time to do so was significant. This is particularly relevant to section 5.3.5 when developing the MSM, and the decision was made across all analyses for consistency. The package used to do this was mice.[130]

Patients were then split into two cohorts defined by their cohort entry date. Those with a cohort entry date prior to 1st Jan 2010 were put into the development cohort, with the remaining patients making up the validation cohort. Patients in the development cohort were then censored at 1st Jan 2010 if their follow up extended beyond this point. The data was split like this because if QRISK3 was replicated exactly using data from 1998 – 2015 for model development, it would not have been possible to assess the calibration of risk scores for patients after 2015, as they would have no follow up.

A Cox proportional hazards model using the same predictor variables as QRISK3 was then fit to the development cohort. Fractional polynomials of age, BMI and SBP were tested for using the mfp package.[131] Five year risk predictions were then generated for both the development and validation cohort using this model, and the calibration of these risks was assessed. For consistency throughout this manuscript, the Directed Acyclic Graph (DAG) and equation is stated for each model used. All DAGs were generated using the dagitty software.[166] Figure 5.1 (DAG-1) and equation (1) correspond to this model, where $h(t)$ denotes the hazard function, $h_0(t)$ the baseline hazard at time t, X_0 the vector of predictors at cohort entry date and β_X a vector of the associated coefficients . Unmeasured confounding is left off the DAGs to reduce the number of arrows and maintain clarity (particularly for DAG-3), however it may be present. The implications of unmeasured confounding are discussed in the limitations section.

Figure 5.1: *DAG-1*



$$h(t) = h_0(t) * \exp(\beta_X . X_0) \quad (1)$$

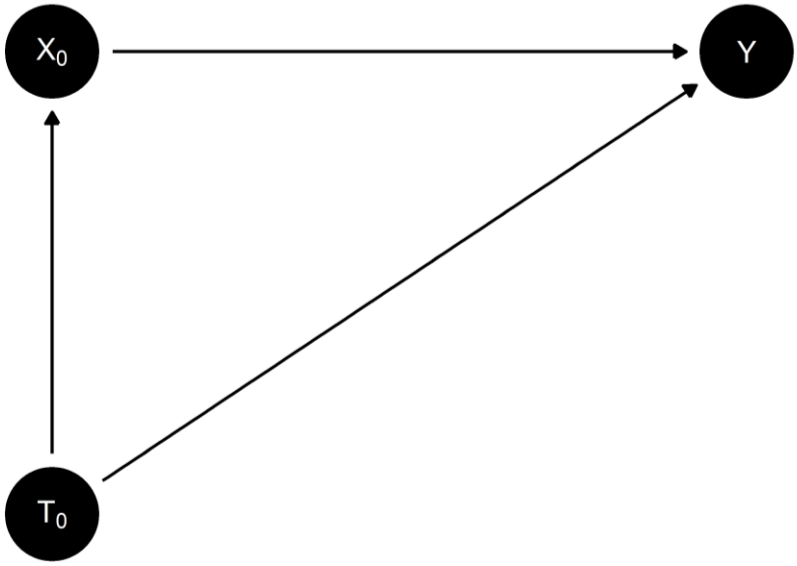
5.3.3 Assessing the sensitivity of the risk prediction model created to changes in patient characteristics

The next step was to assess whether the miscalibration in the validation cohort was driven by a poor model which did not reflect differences between the cohorts, i.e., if the characteristics of the validation cohort were different from the development cohort and explained the reduction in risk, but the model was not reflecting this. The characteristics of each cohort were compared, and also the predicted risks of the development and validation cohorts, to assess whether the changes in predicted risk reflected the changes in the patient characteristics. This is not an exact test with a clear outcome, and the results were interpreted by the authors.

5.3.4 Attempt to model the secular trend to remove miscalibration in validation cohort

Given the miscalibration in the validation dataset, and evidence indicating that the model was reflecting changes in patient characteristics, this indicated that the secular trend could not be explained by changes in predictor variables alone. This provided support for modelling the secular trend in the development cohort, to try and remove the miscalibration in the validation cohort. The same Cox model defined by equation (1) was fitted to the development cohort, but with cohort entry date included as a variable, referred to as calendar time. This is denoted by T_0 in Figure 5.2 (DAG-2) and equation (2). Fractional polynomials for this variable were tested using the mfp package.[131] Five year risks were generated for validation cohort and the calibration of the models was assessed.

Figure 5.2: *DAG-2*



$$h(t) = h_0(t) * \exp(\beta_T.T_0 + \beta_X.X_0) \quad (2)$$

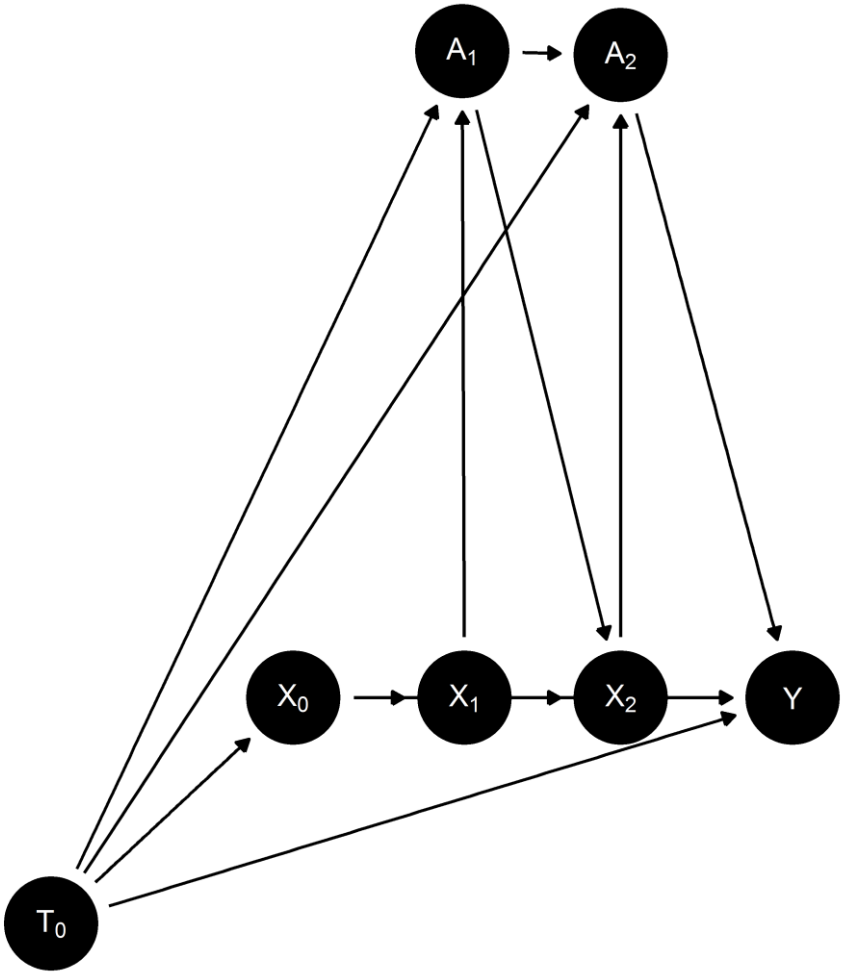
5.3.5 Developing an MSM to assess secular trend after adjusting for statin use during follow up.

MSM – overview

A major concern was that an increase in statin use over time may have caused some of the reduction in CVD incidence. If the secular trend was driven by statin use, then modelling it (which would result in lower predicted risks) would make lots of patients whose risk if they remained untreated was > 10%, ineligible for treatment. Statin use at baseline could not have been driving this secular trend as the development cohort only considered patients who were statin free at baseline, however patients could initiate statins during follow up. The aim of this section was therefore to assess the presence of the secular trend when adjusting for statin use during follow up.

Consider Figure 5.3, where $k = 0$ denotes baseline, and $k = 1, 2$ two time points during follow up (this could be extended to any number of time points). A_k denotes the statin treatment status at time k, X_k covariate information prior to time k, and T_k calendar time at time k. Note A_0 is not included in DAG-3 as $A_0 = 0$ by definition of the CVD primary prevention cohort. It is possible to adjust for changes in X_k and A_k post baseline using standard regression techniques (such as an interval censored Cox model). This would result

Figure 5.3: *DAG-3*



in an estimate of the direct effect of calendar time on CVD incidence, the portion of which is not explained through changes in X_k and A_k during follow up. This would be sufficient for assessing our aim of whether the secular trend remained after adjusting for statin use during follow up. However it would be useless in a risk prediction setting, as there is no way of knowing a patients future set of predictors. Therefore the proposed method to answer our question was an MSM.

MSMs were developed to calculate the causal effect of a time dependent exposure on an outcome in an observational setting, where the treatment and outcome are confounded by time varying covariates.[167,168] Sperrin et al.[169] have shown how MSMs can be used to adjust for 'treatment drop in', the issue of patients starting treatment during follow up in a dataset being used for risk prediction. In the absence of unmeasured confounding, they allow for the estimation of $E[Y(\underline{A} = \underline{0})|X_0]$, where $\underline{A}$ denotes the entire treatment course during follow up, as opposed to $E[Y(A_0 = 0)|X_0]$. The strategy involves adjusting for variables at baseline

as normal and then re-weighting the population by variables that may be on the treatment causal pathway, breaking the links from X_k to A_k. In the resulting pseudo population the allocation of treatment during follow up happens at random (within the levels of the variables defined at baseline). This allows the generation of risk scores using data at baseline only, but also accounting for statin use during follow up. Importantly for this study, if calendar time only effected the outcome Y through increasing statin use in follow up, when using an MSM the direct effect of T_0 on Y would be zero, and adjusting for calendar time at baseline would not result in a drop in the average risk score of patients in the validation cohort.

The estimator of $E[Y(\underline{A} = \underline{0})|X_0]$ is only valid under the three identifiability assumptions of causal inference (exchangeability, consistency and positivity) and correct specification of the marginal structural model, and the model used to calculate the weights. The viability of these assumptions in this study is discussed in the limitations (section 5.5.1).

MSM - data derivation

The CVD primary prevention cohort was used as a starting point. However in order to derive the MSM, patient information was extracted at 10 time points, at 6 month intervals from the cohort entry date, denoted as X_k and A_k for $k = 0, 1, 2,\ldots, 9$. The variable X_k contained all the QRISK3 predictors evaluated at time k (for test data this was the most recent value prior to time k). $A_k = 1$ if a patient had initiated statin treatment prior to k, and $A_k = 0$ otherwise. As patients were excluded from the cohort if they have had a statin prescription prior to their cohort entry date, $A_0 = 0$ for all patients. If a CVD event happened within 6 months of a statin initiation, the statin initiation was ignored. This was to stop any effects of poorly recorded data (start of statins may have been triggered by the CVD event).

A key issue in deriving the dataset was missing data. A combination of imputation techniques were implemented to maintain consistency in variable information within each patient across the 10 time points. First, where possible, last observation carried forward imputation was implemented within each patient. Then, where possible, next observation carried backwards imputation was used to impute the remaining missing data. However, there was still missing data for patients who had no entries across all 10 time points for a given variable. The data at baseline was then extracted and missing values were imputed using one stochastic imputation. All predictor variables, Nelson Aalen estimate of baseline hazard and the

outcome indicator were included in the imputation model (same process used as the imputation in section 5.3.2). These imputed baseline values were then used at each following time point (last observation carried forward imputation).

MSM - Calculation of weights and specification of model

The MSM was fitted as a weighted interval censored Cox model using the coxph function from the survival package.[170] The weights themselves were calculated using the IPW package.[171] Stabilised weights were calculated as is common practice to provide more precise estimation of the weights. For individual i, the formula for the weight of interval/time period K was defined as:

$$sw_i = \prod_{k=0}^{K} (\hat{p}_{ki}^*)^{A_{ki}} (1 - \hat{p}_{ki}^*)^{1-A_{ki}} \Big/ \prod_{k=0}^{K} (\hat{p}_{ki})^{A_{ki}} (1 - \hat{p}_{ki})^{1-A_{ki}} \quad (3)$$

where $\hat{p}_{ki}^* = P[A_k = 1|\underline{A}_{k-1}, X_0]$ and $\hat{p}_{ki} = P[A_k = 1|\underline{A}_{k-1}, \underline{X}_k, X_0]$, and $\underline{A}_k$ and $\underline{X}_k$ denote treatment history and covariate history respectively up time point k for individual i. More simply put, the denominator is the probability that the individual received the treatment they did, based on time varying predictors and predictors at baseline. The numerator is the probability that the individual received the treatment they did, based on predictors at baseline only. The models used to estimate the probability of treatment when deriving the weights were interval censored Cox models. If calendar time at baseline, T_0, was being included in the MSM, it was also included as a stabilising factor in the calculation of the weights as part of X_0. Detailed information on how to calculate weights is also given in the literature[168,171,172] and the formula for calculating weights (and notation for variables) matches that from the work by Sperrin et al.[169]

Two MSM's were created, one that adjusted for calendar time at baseline and one that did not:

$$h(t) = h_0(t) * \exp(\beta_A . A_t + \beta_X . X_0) \quad (4)$$

$$h(t) = h_0(t) * \exp(\beta_A . A_t + \beta_X . X_0 + \beta_T T_0) \quad (5)$$

The same fractional polynomials of age, BMI, SBP and calendar time that were found to be optimal in the Cox models from sections 5.3.2 and 5.3.4 were used in the MSM, and in the models used to calculate the weights. Ideally we would have re-calculated the optimal fractional polynomials for the weighted model fitted to the interval censored data, however software was not available to do this. Using the same fractional polynomials from the standard Cox analysis was preferred to having no fractional polynomials, as removing them led to poorly calibrated models. The coefficient β_A is the average causal effect of initiating statin treatment after adjusting for all other variables. It is quite common to allow the effect of statin treatment to be modified by baseline variables, which could be achieved by including interaction terms $A_t X_0$. However the primary aim was to account for statin use in follow up, rather than calculate the effect of statin treatment in different subgroups, so we did not feel this was necessary.

As a comparison, unweighted interval censored Cox models using only data at baseline (i.e. equation (1) from section 5.3.2 and equation (2) from section 5.3.4 were fitted to the same data as the MSM. The effect of modelling the secular trend could then be assessed when using normal Cox regression, as well as under the MSM framework. This was preferred to re-using the models directly from sections 5.3.2 and 5.3.4, as the data they were fitted to underwent a different imputation process.

MSM – analysis of interest

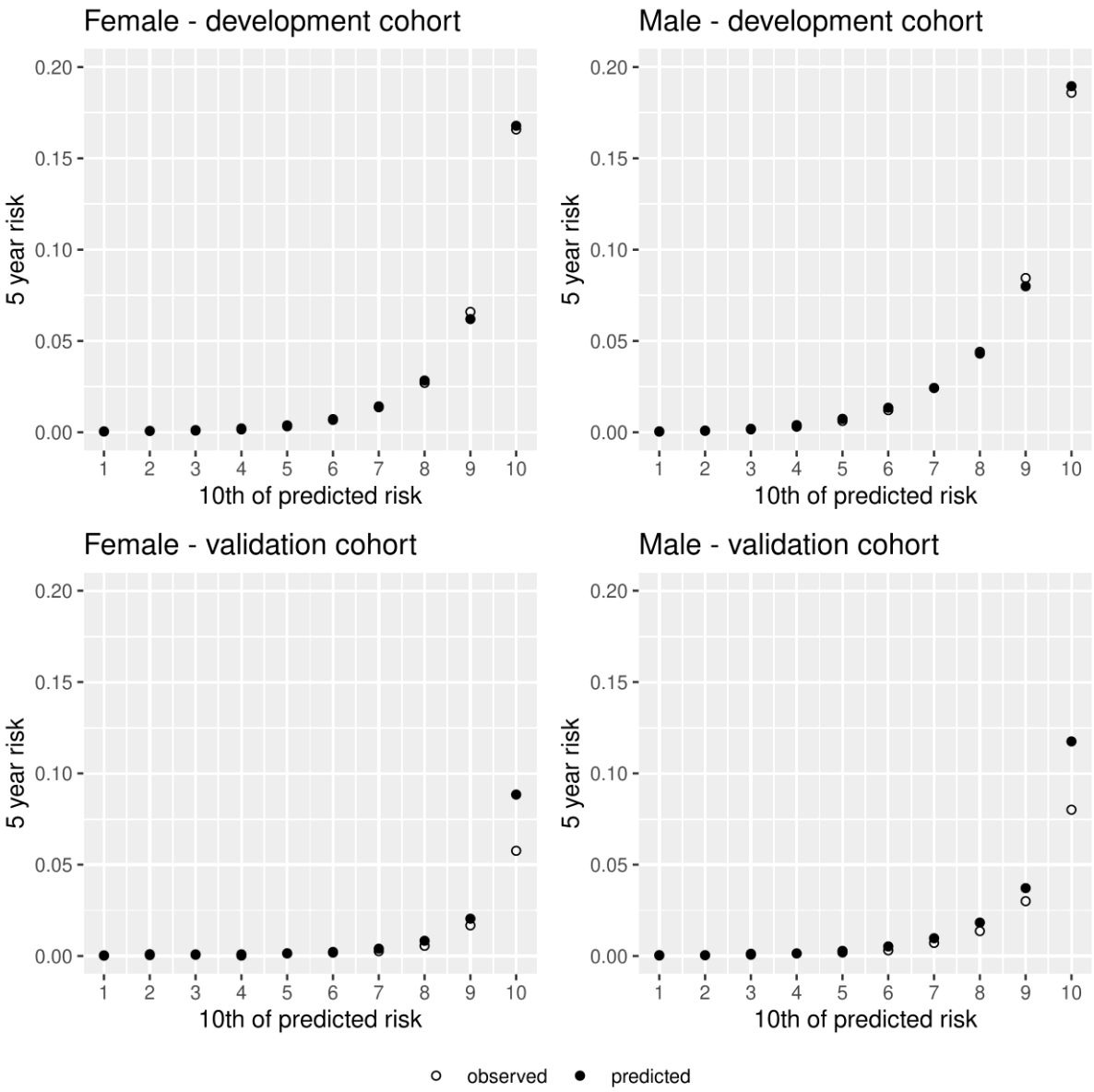
The MSM was used to generate risk predictions assuming no statin treatment at baseline or during follow up, $E[Y|X_0, \underline{A} = \underline{0}]$, the estimator of $E[Y(\underline{A} = \underline{0})|X_0]$. The interval censored Cox model only produced risk predictions based on no statin treatment at baseline, $E[Y|X_0, A_0 = 0]$, the estimator of $E[Y(A_0 = 0)|X_0,]$. The outcome of interest was the risk ratio of the average predicted risk of patients in the validation cohort, before and after adjusting for calendar time at baseline in the MSM framework, $E[Y(\underline{A} = \underline{0})|X_0, T_0]/E[Y(\underline{A} = \underline{0})|X_0]$. This was compared to the risk ratio after adjusting for calendar time at baseline in the unweighted interval censored Cox models, $(E[Y(A_0 = 0)|X_0, T_0]/E[Y(A_0 = 0)|X_0])$.

5.4 Results

5.4.1 Quantifying the miscalibration in risk predictions of patients in the present day

Figure 5.4 shows the calibration of the model in the development and validation cohorts. While the model was well calibrated in the development cohort, as expected, there was a large under prediction of risks in the validation cohort. Statin prevalence and incidence rates in the primary prevention cohort are provided in Table A.5.1 and Table A.5.2 in Appendix A.5.1.

Figure 5.4: *Calibration of development (pre 2010) and validation (post 2010) cohorts*



5.4.2 Assessing the sensitivity of the risk prediction model created to changes in patient characteristics

Differences between the development and validation cohorts are shown in Table 5.1. In the validation cohort, patients were generally younger and healthier. As shown in Figure 5.5, the predicted risks in the validation cohort were significantly smaller than those in the development cohort. This indicates that the model did appropriately reflect the differences in baseline predictors between the cohorts, and the secular trend in CVD incidence could not be explained by this.

Figure 5.5: *Predicted risks in the development and validation cohort*

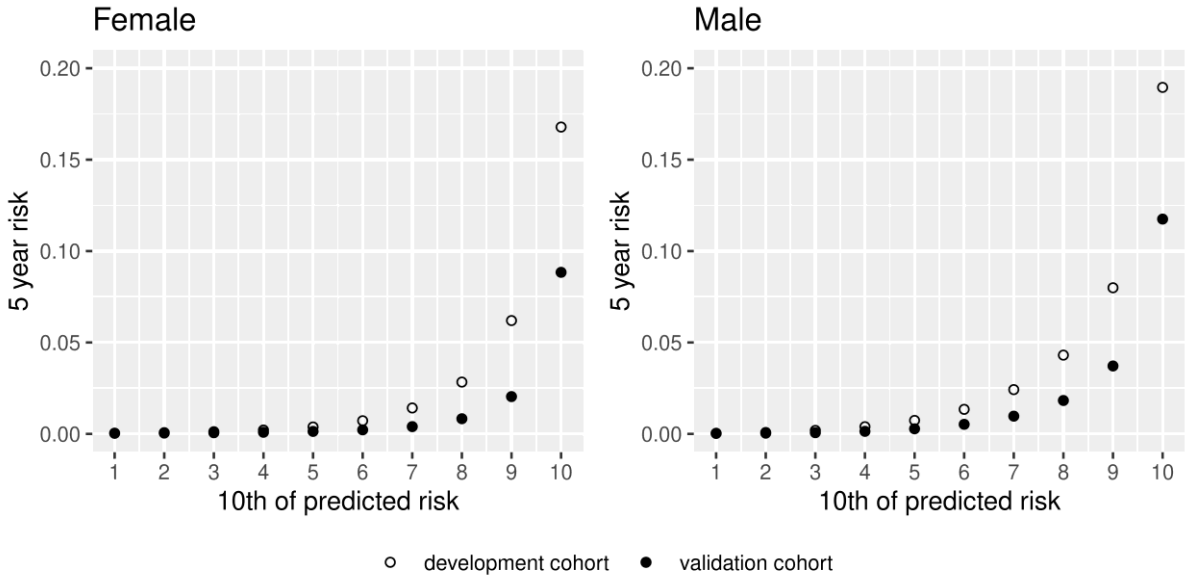


Table 5.1: *Baseline variables in development and validation cohorts*

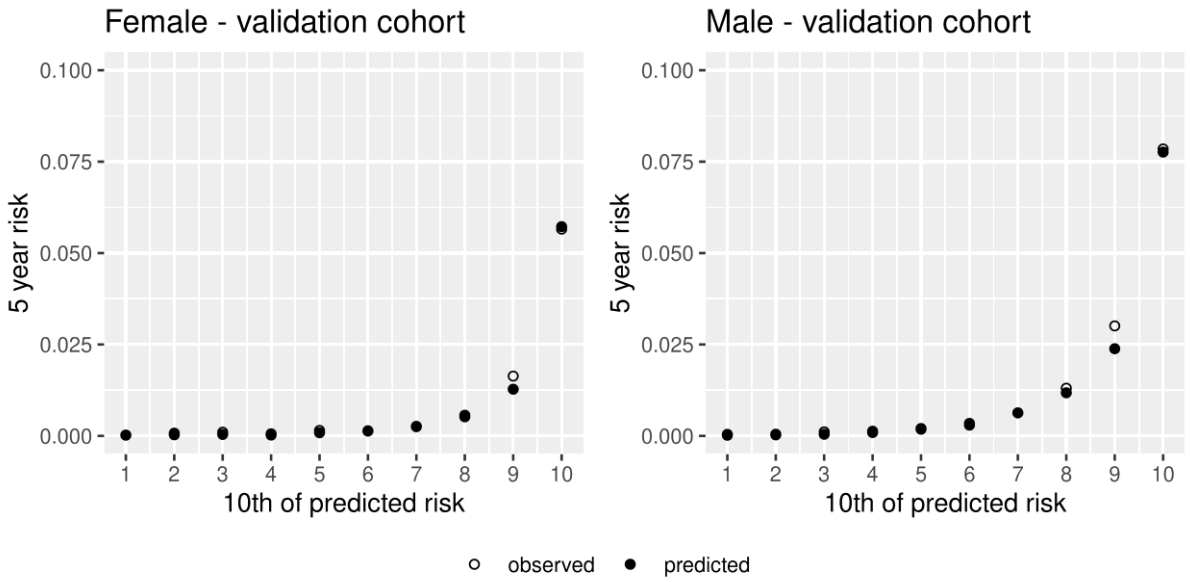
	Male development	Male validation	Female development	Female validation
N	1,497,511	393,071	1,555,010	410,068
Age	43.07 (14.84)	37.18 (12.42)	44.56 (16.22)	37.4 (13.41)
BMI	26.07 (4.43)	26.3 (4.8)	25.54 (5.47)	25.78 (5.96)
Cholesterol/HDL ratio	4.51 (1.4)	4.32 (1.37)	3.76 (1.21)	3.52 (1.1)
SBP	130.67 (17.04)	127.71 (14.07)	125.15 (19.04)	119.53 (14.43)
SBP variability	10.37 (6.92)	9.39 (6.37)	9.66 (6.21)	8.87 (5.17)
Atrial fibrillation	0.61	0.44	0.48	0.28
Atypical anti-psychotic medication	0.25	0.62	0.23	0.58
Corticosteroid use	0.31	0.22	0.51	0.36
CKD stage 3/4/5	0.25	0.57	0.33	0.95
Diabetes (type 1)	0.26	0.36	0.19	0.27
Diabetes (type 2)	1.56	0.93	1.26	0.78
Ethnicity = Asian other	1.56	2.84	1.49	2.88
Bangladesh	0.34	0.79	0.24	0.48
Black	2.93	5.80	3.12	5.90
Chinese	0.45	0.87	0.56	1.17
Indian	2.49	4.18	2.21	3.63
Mixed	0.69	1.47	0.75	1.64
Other	1.53	2.72	1.45	2.84
Pakistan	0.92	1.94	0.76	1.64
White	89.09	79.39	89.42	79.81
Family history of CHD	10.67	12.36	14.89	15.80
HIV/AIDS	0.06	0.19	0.04	0.13
Migraine	2.71	3.85	6.73	9.30
Rheumatoid arthritis	0.28	0.17	0.74	0.47
Severe mental illness	4.59	4.55	9.07	6.95
SLE	0.01	0.01	0.09	0.11
Smoking = Never	47.37	44.77	57.03	53.30
Smoking = Ex	16.09	20.59	14.97	22.49
Smoking = Yes	36.53	34.63	28.00	24.21
Townsend = 1 (least deprived)	22.79	17.30	23.08	17.70
Townsend = 2	22.32	18.38	22.76	19.03
Townsend = 3	20.77	20.82	21.19	21.17
Townsend = 4	20.23	22.85	19.91	22.53
Townsend = 5	13.89	20.65	13.06	19.57
Treated hypertension	4.82	3.28	6.81	3.81

**BMI, body mass index; CKD, chronic kidney disease; HDL, high-density lipoprotein; SBP, systolic blood pressure; SLE, systemic lupus erythematosus.*

5.4.3 Attempt to model the secular trend to remove miscalibration in validation cohort

The calibration in the validation cohort after including secular trend into the model is shown in Figure 5.6. There was still an under-prediction in the second highest risk group in the second highest risk group for both the female and male cohorts, but overall there was a substantive improvement in calibration compared to not modelling the secular trend.

Figure 5.6: *Calibration of the validation cohort when adjusting for calendar time*



5.4.4 Developing an MSM to assess secular trend after adjusting for statin use during follow up.

The average predicted risks of patients in the validation cohort before and after adjusting for calendar time, in the interval censored Cox and MSM setting, are presented in Table 5.2. The risk reduction caused by accounting for secular trend was marginally smaller under the MSM framework compared to the standard Cox. This means the effect of secular trend was slightly smaller when adjusting for statin use during follow up. However the difference would not be clinically significant, and there was still a large drop in risks. The hazard ratios from the two MSM's are provided in Table 5.3, the coefficient of statin initiation is a causal estimate and can be used to help verify if the model has been derived correctly. Calibration of the interval censored Cox model and the MSM are presented in Appendix A.5.1, both are well calibrated.

Table 5.2: *Average predicted CVD risk for patients in the validation cohort before and after secular trend was introduced, using an MSM and an interval censored Cox model*

	Predicted CVD risk (average)		**Relative reduction in risk**
	Not adjusted for secular trend	Adjusted for secular trend	
Interval censored Cox			
Female	1.284%	0.826%	35.68%
Male	1.911%	1.274%	33.31%
Marginal structural model			
Female	1.287%	0.859%	33.24%
Male	1.941%	1.307%	32.67%

Table 5.3: *Hazard ratios of the categorical variables in the marginal structural model with and without secular trend included as a predictor variable*

	Female		**Male**	
	Secular trend not accounted	Secular trend accounted	Secular trend not accounted	Secular trend accounted
Statin initiation	0.71	0.77	0.75	0.81
Ethnicity: Asian other	0.95	1.07	0.99	1.11
Bangladeshi	1.27	1.42	2.03	2.22
Black	0.90	0.99	0.53	0.57
Chinese	0.81	0.88	0.42	0.46
Indian	1.27	1.36	1.22	1.29
Other ethnic group	0.58	0.73	0.82	0.90
Pakistani	1.24	1.39	1.93	2.12
Townsend = 2	1.10	1.10	1.01	1.01
Townsend = 3	1.13	1.13	1.08	1.08
Townsend = 4	1.20	1.20	1.15	1.16
Townsend = 5 (most deprived)	1.37	1.35	1.27	1.26
Atrial fibrillation	1.97	1.97	1.69	1.70
Atypical antipsychotic medication	1.47	1.69	1.50	1.73
CKD stage 3/4/5	1.02	1.15	1.30	1.39
Corticosteroid use	1.62	1.63	1.55	1.52
Type 1 diabetes	2.31	2.31	1.51	1.49
Type 2 diabetes	1.91	1.87	1.83	1.79
Erectile dysfunction			1.17	1.26
Family history CVD	1.16	1.16	1.28	1.28
HIV	1.22	1.32	2.72	2.95
Hypertension	1.20	1.23	1.22	1.25
Migraine	1.19	1.19	1.21	1.21
Rheumatoid arthritis	1.32	1.32	1.28	1.28
Severe mental illness	1.43	1.39	1.32	1.29
Smoking = Ex	1.12	1.14	1.10	1.12
Smoking = Current	1.55	1.55	1.57	1.58
SLE	1.49	1.51	1.29	1.26

**CKD, chronic kidney disease; SLE, systemic lupus erythematosus.*

5.5 Discussion

This results in this paper show that not modelling the secular trend in CVD incidence in England causes over prediction of risks for patients in the present day. Also, the secular trend in CVD incidence cannot be explained by changes in statin use over time, because when adjusting for calendar time in the MSM framework the risk predictions of patients in the validation cohort still dropped substantially.

These findings support the need to adjust for calendar time in prediction models used to drive clinical decision making in England. However the drop in risk caused by accounting for this secular trend is drastic and changes should not be made in practice without the generation of more evidence. Most importantly, these findings should be reproduced in a different dataset. This should not be difficult as QRISK3 has been developed in the QResearch database, and QRISK2 has been externally validated in the Health Improvement Network database.[100] This means analysis ready datasets exist and could be tested for secular trends in CVD with minimal extra work.

The next step would then be to try and identify what is causing this drop in CVD incidence. In this study, we ruled out one potential cause, the use of statins during follow up. If it is driven by changing recording practices, this would be another reason not to model it. Primary care records in particular may be susceptible to differential recording over time as monetary incentives are given for recording specific things. However, a large portion of the events are identified in HES and ONS which will not have suffered from the same level of differential recording. This is backed up by the trends reported in the literature, which are also not based on primary care codes.[53–56] Further work in a causal framework to establish what is causing this drop would be really valuable and could provide a much stronger argument for modelling the secular trend (e.g. if its driven by lifestyle changes). However, given the current evidence, there is still not a strong argument against modelling it.

Risk scores should be based on current data; this is why the series of QRISK models have used a rolling window for their development datasets. If there was a much higher incidence of CVD in the 1990s due to various differences in healthcare management, we would not want to incorporate this into current risk scores as it would inflate the risks. Therefore, there is also no reason to assume the incidence of CVD has been the same throughout the time window

of data we are using. In this sense, current approaches to risk prediction are contradictory. We are happy to omit old data from our cohort periodically to reflect changes in the population; but we are not willing to model changes in the population over the time period in which we have defined our cohort. If wanting to do so, dynamic models are what should be used to model changes over time.

With respect to the dynamic modelling methods outlined by Jenkins et al.,[149] the current approach in England implemented by QRISK series is discrete model updating (models are re-calculated in a more recent dataset each year). In this chapter, we modelled the secular trend by including a calendar time variable at baseline. This effectively allowed the intercept (or baseline hazard) to vary by calendar time, and is a special case of a varying coefficient model. However, there are more complex methods such as Bayesian model updating and varying coefficient models that allow changes in predictor coefficients over time, and could give more control over how the secular trend is modelled. If a dynamic model was to be developed for use in practice, these methods should be considered, alongside how to how to use these methods within an MSM framework. Arguably the use of an MSM should be standard procedure in the presence of 'treatment drop in' during follow up, as a normal Cox model under predicts the risk of patients if they were to remain untreated, which is what treatment decisions should be based on.[169] If modelling a secular trend in the outcome that was being partially driven by this treatment drop in (which was not the case in this study), it would be even more important to work under an MSM framework. However, currently it is not clear how the more complex dynamic modelling approaches would be handled in an MSM framework. This is therefore a key area for future research.

5.5.1 Limitations

There are several limitations to the study. The first is that the estimate of $E[Y(\underline{A} = \underline{0})|X_0]$ is only valid if the assumptions of exchangeability, consistency, positivity (identifiability assumptions) and correct model specification are all met. The untestable assumption of exchangeability, or no unmeasured confounding, represents the fundamental problem with deriving causal estimates from observational data. If violated the estimate of statin treatment will be biased (and subsequently the risk scores conditional on no statin treatment during follow up will be biased too). Given the large number of predictors available we hope that the

unmeasured confounding is not too extensive. The consistency assumption, that a subject's counterfactual outcome under their observed exposure history is precisely their observed outcome, is generally considered a reasonable assumption when estimating the effects of medical treatments.[172] This is maybe less true in our data as a patient could initiate statins any time over a 6 month period and be assigned the same exposure value. However we did not believe that initiating within a 6 month interval would have a significant impact on the outcome, and reducing the size of the intervals would have been impractical. The positivity assumption, that there were unexposed and exposed individuals at every level of the confounders, was reasonable given the large size of the development dataset and the resulting number of statin initiations.

The assumption of correct model specification, as is the case with all models, will have been violated to some extent in this study. For example, the fractional polynomials of continuous variables calculated from the standard Cox models were used in the MSM. It was not clear how to estimate optimal functional forms under the MSM framework, but re-using the functional forms from the Cox models provided better model performance than just having linear terms. Also, not all variables and interaction terms from the MSM were used in the model to calculate the weights. Doing so produced extreme values weights, and therefore variables in the weighting models were chosen to minimise this. This follows the advice of Cole and Hernan, who state "*one may wish to omit control for weak confounders that cause severe non-positivity bias because of a strong association with exposure*".[172] There is no clear-cut way to do this, and therefore a more appropriate set of predictors in the weighting model may have existed. Finally, we only considered the effect of initiating statin treatment. A more detailed MSM which also modelled discontinuation from treatment would allow the calculation of a patients risk if they were to initiate treatment at baseline and not discontinue (or discontinue after a fixed period of time), as opposed to just the risk if they initiate treatment at baseline. However, the density of data available in CPRD, or any other primary care electronic health record is probably not sufficient for this. To model statin initiation and discontinuation at that granularity, more regular updates on predictor variables would be required.

The second limitation was that the results are not directly applicable to the models used in practice in the UK, which are based on 10-year risk scores. However, we have no reason to

think the results would not be generalizable because a similar secular trend was found in previous work when dealing with 10-year risks.[1] The third limitation was the level of missing data. Changes in the time varying predictor variables is what drives the weighting in the MSM in order to calculate the effect of statin initiation. Therefore not having predictor information at each time point, and re-using predictor information from previous time points may have led to a biased estimate of statin initiation.

One way to assess the potential impact of limitations 1 (violating assumptions) and 3 (missing data) was to check the hazard ratio for initiating statin treatment (ranging between 0.71 – 0.81) was in a sensible range. We compared this to the effect estimates of statins from trials reported in the appendices of the NICE guidelines (see section L.2.3.4),[84] and there is reasonable agreement. It should be noted that they report relative rates for specific CVD outcomes which are not directly comparable to our composite definition. However, the similarities that exist still ease concerns over limitations 1 and 3, and that the model was well specified despite these limitations.

5.5.2 Conclusions

In conclusion, inclusion of the secular trend into the model substantially changed the CVD risk predictions. Models that are being used in clinical practice in the UK do not model secular trend and may thus overestimate the risks, possibly leading to patients being treated unnecessarily.

6 The impact of statin discontinuation and restarting rates on the optimal time to initiate statins and on the number of cardiovascular events prevented

Alexander Pate, Rachel A Elliott, Georgios Gkountouras, Alexander Thompson, Richard Emsley, Tjeerd van Staa

6.1 Abstract

Introduction: A patient is eligible for statins in the United Kingdom if they have a 10-year risk of cardiovascular disease of 10% or more. We hypothesise that if statin discontinuation rates are high it may be better to delay statin initiation until patients are at a higher risk, to maximise the benefit of the drug.

Methods: A four-state health state transition model was used to assess the optimal time to initiate statins after a risk assessment, in order to prevent the highest number of cardiovascular events, for a given risk profile (age, gender, risk) and adherence rate. A CPRD dataset linked to HES and ONS was used to inform the transition probabilities in this model, taking into account observed statin discontinuation and re-continuation patterns.

Results: Our results suggest, if statins are initiated in a cohort of 50 year old men with a 10% 10-year risk, we prevent 4.78 events per 100 individuals. If we wait 10 years to prescribe, at which point 10 year risk scores are at 20%, we prevent 5.45 events per 100 individuals. If the observed discontinuation rate was reduced by a sixth, third or half in the same cohort, we would prevent 7.29, 9.01 or 10.22 events per 100 individuals.

Conclusions: Based on discontinuation rates in England, evidence suggested there could be benefit to delaying statin past the 10% threshold in certain scenarios, but this approach has ethical concerns. Furthermore, the optimal time to initiate statins was driven by age, not cardiovascular risk.

6.2 Background

Cardiovascular disease is the number one cause of death globally accounting for 31% of all deaths in 2017,[173] and contributes more than any other disease to the total disease burden around the globe.[174] Treatment for primary prevention of cardiovascular disease is centred around lifestyle modifications such as changes to diet and exercise, and cholesterol-lowering medication such as hydroxymethylglutaryl-coenzyme A reductase inhibitors (statins). There has recently been a lot of debate in the literature over what the risk threshold to be eligible for statins in primary prevention of cardiovascular disease should be. Both England[45] (National Institute for Health and Care Excellence guidelines) and the US[79] (American College of Cardiology/American Heart Association guidelines) have recently dropped their thresholds to a 10-year risk of 10% and 7.5%, respectively. However, the European Society of Cardiology still recommends a 10-year risk of a fatal cardiovascular event of 5%, which equates to about a 15% risk of any cardiovascular event,[66] while in Scotland the recommended threshold is 20% for asymptomatic individuals.[64] In support of higher thresholds, a recent study found that statins only provide a net benefit over possible harms at higher 10-year risks than the thresholds in current guidelines, and the benefits vary considerably by age and sex.[92]

One factor that will affect the real-world impact of these guidelines is the widely reported suboptimal long-term adherence to and discontinuation from statins.[175–177] Studies examining factors affecting adherence to statins report consistent relationships between non-adherence and female gender, ethnic minority status, reduced income, lower number of concurrent cardiovascular medications, new statin users, use of statins for primary prevention, smoking, depression, reduced follow-up and increased copayments,[178–181] while a recent high profile meta-analysis concluded that exaggerated claims about side-effect rates with statin therapy may be responsible for its under-use among individuals at increased risk of cardiovascular disease.[83] The analyses underpinning the treatment thresholds do not incorporate the effects of non-adherence or discontinuation directly. We suggest that policy decisions around lowering of treatment thresholds may need to take account of real-world statin discontinuation rates in primary prevention. The reason could be that patients are initiating statins at a low risk and then discontinuing the drug when at a higher risk (risk increases with age), not maximising the benefit of the drug.

The overall aim of this study was to assess the optimal time to initiate statins after a risk assessment in order to prevent the highest number of cardiovascular events, given a patient's risk profile, and long-term adherence levels derived from real life data. We refer to adherence throughout this study specifically in relation to the combination of discontinuation and restarting rates. We also developed a range of scenarios where discontinuation rates were artificially decreased, allowing us to evaluate the effect that improving adherence would have on the number of cardiovascular events prevented.

6.3 Methods

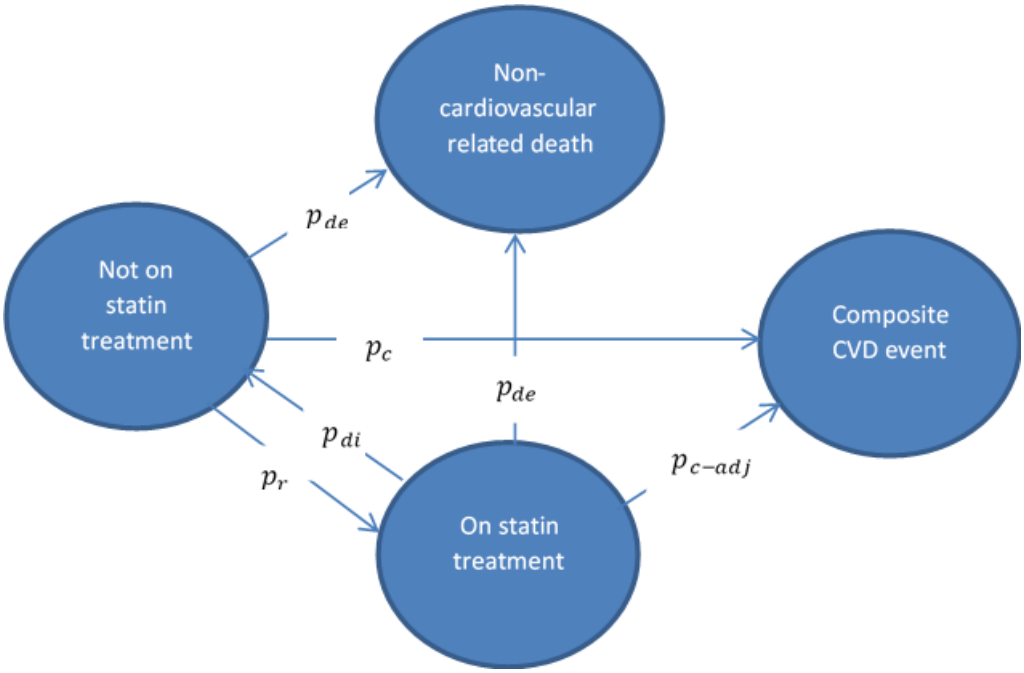
6.3.1 Overview of simulation model design

A four-state health state transition model with cycle length of one year was created (Figure 6.1) to answer our primary aim. Each scenario (age, gender, 10-year cardiovascular risk score and assumed adherence rate) represented a patient having their 10 year risk assessment, which is when a clinician would decide whether to initiate statin treatment. We varied the year of follow-up in which statins were initiated, and calculated the total number of cardiovascular events expected. For the main analysis the discontinuation and restarting rates were derived directly from the CPRD cohort, then for subsequent analyses the discontinuation rate were artificially decreased. The cost effectiveness of statins at various risk thresholds has already been extensively covered).[84] Instead, this model is set up to calculate the number of incident cardiovascular events prevented by initiating statins at different times, assuming real life risk profiles and adherence rates, and is what makes this study unique.

6.3.2 Data source

This project used data from the Clinical Practice Research Datalink (CPRD) linked with Hospital Episodes Statistics (HES) and Office for National Statistics (ONS). CPRD is a primary care database representative of the UK in terms of age, sex and ethnicity,[48] although linkage to HES restricts this dataset to England only. The data were used to create two cohorts, a cohort of statin users (statin cohort) and a cohort of patients at risk of cardiovascular disease (primary prevention cohort).

Figure 6.1: *Design of the health state transition model*



*p_c *is the probability of a cardiovascular event;* p_{c-adj} *is the probability of a cardiovascular event, while receiving statin treatment;* p_{de} *is the probability of death (mortality);* p_{di} *is the probability of discontinuing statin treatment;* p_r *is the probability of restarting statin treatment.*

The primary prevention cohort was defined in the same way as the QRISK3 development cohort.[5] To be eligible for the cohort, a patient must have had one day of follow up in CPRD that met the following inclusion criteria: 1) Aged 25 – 84, 2) Within study period of 1st Jan 1998 to 31st Dec 2015, 3) at least one year prior follow up. The cohort entry date for a patient was defined as the first date that met all these criteria. Patients were excluded if they met the following exclusion criteria: 1) Cardiovascular event (identified through CPRD, HES or ONS) or statin prescription prior to cohort entry date (code lists provided on GitHub[103]). Patients were censored at the earliest date of transferred out of practice, last data collection for practice, death, or 31st Dec 2015.

Inclusion criteria for the statin cohort was: 1) One or more statin prescriptions between 1st Jan 1998 and 31st Dec 2015 (code list for statins on GitHub[103]), 2) Aged 25 or over on date of first statin prescription. Exclusion criteria were: 1) Cardiovascular event prior to first statin, 2) Less than one year follow up prior to first statin prescription. Exclusion criteria 2 is to ensure all patients are first time users of statins, rather than current users who have transferred from

another practice. A patient entered the cohort on the date of their first statin prescription and exited the cohort at the end of that statin treatment period (detailed definition in Appendix A.6.1). A patient could leave and re-join the cohort (at the start of their next treatment period) multiple times before their censoring date. A patient was censored if transferred out of practice, at the end of data collection, death or occurrence of a cardiovascular event.

6.3.3 Estimation of transition probabilities

Cardiovascular event transition probabilities were calculated from the primary prevention cohort. A lifetime risk model was fitted using standard techniques for developing lifetime risk models.[161,182,183] This involved fitting a Cox model with age as the time scale, the outcome was time until first cardiovascular event, and the same predictor variables as QRISK3[5] (atrial fibrillation, atypical antipsychotic use, body mass index (BMI), cholesterol/high-density lipoprotein (HDL) ratio, chronic kidney disease (CKD, stage 3/4/5), corticosteroid use, erectile dysfunction (male model only), ethnicity, family history of CVD, HIV/AIDS, hypertension (treated), migraine, rheumatoid arthritis, severe mental illness, systolic blood pressure (SBP), SBP variability, smoking status, systemic lupus erythematosus, type 1 diabetes, type 2 diabetes and Townsend deprivation score; code lists and information about variable derivation provided in Chapter 2). Using the baseline hazard from this model, for a given age the hazard ratio could be adjusted to obtain a specific 10 year risk (for each scenario), and from this the corresponding lifetime risk could be derived. After deriving this, the conditional probability of having a cardiovascular event in each year of follow up was calculated (conditional on not having had an event prior to that year), giving the transition probabilities p_c. Full details on derivation provided in Appendix A.6.2, and calibration of the Cox models used in Appendix A.6.3.

The transition probabilities of a cardiovascular event while on statin treatment were calculated as $p_{c-adj} = 0.7 * p_c$. This estimate of statin effectiveness (relative rate: 0.7) was taken from the National Institute for Health and Care Excellence economic model for cost effectiveness of statins[45,84], based on using high intensity statin regimens. Given the varying incidence of each component of the outcome across different age categories and sexes, any single estimate of the relative rate on the composite outcome would be somewhat arbitrary.

We therefore chose 0.7 as a conservative estimate for the effect of statins, given that the estimated risk ratio for high-intensity statins on Myocardial Infarction and Angina was 0.46, and on transient ischaemic attack and stroke was 0.8.

The probabilities of discontinuing and restarting statins were calculated using the statin cohort. The data were split into different groups: first treatment period, off treatment for first time, second treatment period, off treatment for second time, etc. Kaplan Meier curves were then fit to each group and the probability of a patient discontinuing/restarting during each day of follow up was calculated. As the duration of follow-up in the simulation was longer than in our data, the discontinuation/restarting rates were extrapolated. If a patient discontinued for a third time we made the assumption they did not restart treatment because the discontinuation rate in the fourth treatment period was high (76%/90% after 1/2 years), and only 314 patients remained in this cohort after 5 years (see results section 6.4.2). For the first treatment period discontinuation rates were stratified by age (this was not possible for subsequent periods as sample size was deemed too small for some subgroups). A Cox model was fit to the discontinuation data from the first treatment period with age as a predictor variable, considering fractional polynomials of age using the mfp package.[131] This allowed the discontinuation rate to be a function of age. Full details of the stratification and extrapolation of the discontinuation rates is provided in Appendix A.6.4.

The transition probabilities of non-cardiovascular related death were calculated using the primary prevention cohort. The date of death was based on the data as recorded in primary care, shown to have 92% concordance with ONS within two weeks.[184] These data were combined with ONS, for which we had linkage to cardiovascular disease related deaths. Deaths identified in primary care that were cardiovascular disease related were then excluded. Incidence rates of death across each age category were then calculated.

6.3.4 Implementation of the simulation

Different scenarios were simulated based on a patient having a risk assessment (start of the simulation), and the decision of whether to initiate statins straight away, or delay. Variables that made up the different scenarios were: age, gender and 10-year cardiovascular disease risk at the start of the simulation, the statin initiation date, and an assumed adherence rate. The ages considered were 40, 50 and 60. For each age, we considered all 10 year risks within

the 1 – 99th percentile range of risk scores calculated for patients in that age group from our primary prevention cohort. The statin initiation date was varied in yearly intervals from the start of simulation. Given the discontinuation rate for the first treatment period was stratified by age, this meant the age at statin initiation time impacted the discontinuation rate used in each scenario. Duration of follow up was from the age at start of the simulation (risk assessment), until 90 years of age, and therefore varied depending on the age specified for the scenario. Cycle lengths were one year. For each scenario we simulated 10,000 patients and calculated the number of cardiovascular events over the course of the entire duration of follow up, which was compared with the number of events if no statins were given, providing the number of events prevented per 100 people.

This process was repeated using four different adherence rates. The discontinuation rate from the first, second and third treatment periods were altered so that the probability of discontinuation was 5/6th, 2/3rd, ½ or 0th (100% adherence) of the rate derived from CPRD.

6.3.5 Sensitivity analyses

The simulation was also run assuming a treatment effect of 0.65 and 0.6, given the uncertain nature of the estimate used in the primary simulation. Also, simulations were run using discontinuation and restarting rates from a cohort of statin users where any single prescriptions were removed, a step often taken to identify cohorts of long term statin users.

6.4 Results

6.4.1 Cohort characteristics

Table 6.1 contains baseline characteristics of the two study cohorts, stratified by gender. Patients were older in the statin cohort, (63.50/59.80 vs 42.07/41.84 for females and males respectively), had higher BMI, cholesterol/HDL ratio, systolic blood pressure and fewer never smokers compared with the primary prevention cohort. Comorbidities were also more common.

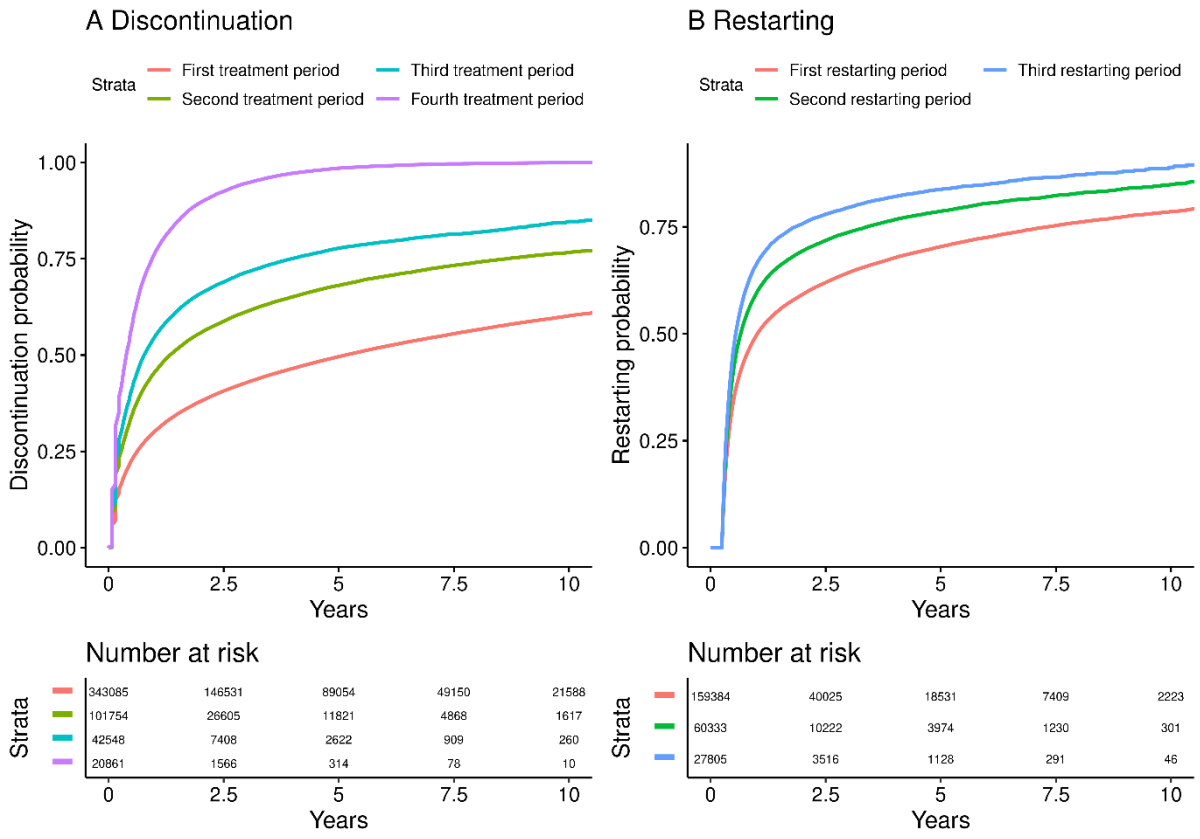
Table 6.1: *Baseline table for statin cohort and CVD primary prevention cohort*

	Statin users cohort: female	Statin users cohort: male	Primary prevention cohort: female	Primary prevention cohort: male
N	161,995	181,090	1,965,078	1,890,582
Demographics				
Age [mean, (sd)]	63.49 (11.05)	60.07 (11.09)	43.07 (15.94)	41.84 (14.57)
Townsend: 1	22.84%	24.91%	21.96%	21.65%
2	23.13%	23.77%	21.98%	21.50%
3	20.66%	20.40%	21.18%	20.78%
4	20.06%	18.80%	20.46%	20.78%
5	13.30%	12.12%	14.42%	15.29%
Test data				
Body mass index [mean, (sd)]	29.26 (6.33)	28.95 (5.04)	25.60 (5.60)	26.12 (4.54)
Cholesterol/ high density lipoprotein ratio [mean, (sd)]	4.36 (1.48)	4.88 (1.61)	3.72 (1.20)	4.48 (1.40)
Systolic blood pressure [mean, (sd)]	140.52 (18.39)	140.78 (17.25)	123.91 (18.28)	130.03 (16.48)
Systolic blood pressure variability [mean, (sd)]	13.10 (5.80)	12.15 (5.89)	9.47 (5.98)	10.13 (6.80)
Smoking status	Never = 46.79% Ex = 30.33% Current = 22.87%	Never = 32.35% Ex = 40.42% Current = 27.23%	Never = 56.04% Ex = 16.97% Current = 26.99%	Never = 46.63% Ex = 17.48% Current = 35.99%
Medical History				
Atrial Fibrillation	2.85%	3.61%	0.44%	0.57%
Chronic Kidney Disease stage 3/4/5	7.13%	4.00%	0.45%	0.32%
Family history of coronary heart disease	29.17%	23.02%	15.08%	11.02%
Rheumatoid arthritis	2.08%	0.87%	0.69%	0.26%
Treated hypertension	49.03%	43.84%	6.18%	4.50%
Type 1 diabetes	1.33%	1.72%	0.21%	0.28%
Type 2 diabetes	21.28%	22.33%	1.16%	1.42%

6.4.2 Discontinuation and restarting of statins

Figure 6.2 presents the discontinuation (A) and restarting (B) rates over the first 10 years of each treatment and restarting period. This demonstrates that 30% patients have stopped taking statins by the end of the first year of follow-up during the first treatment period, 38% have stopped after 2 years, and by 10 years 60% have stopped. Of all the patients that discontinue, 50% have restarted a year after the initial discontinuation, 59% after 2 years, and 79% after 10 years. The second discontinuation and restarting rates suggest patients are more likely to discontinue/restart during the subsequent treatment periods. Graphs for the discontinuation rate in the first treatment period stratified by age, extrapolated discontinuation and restarting rates beyond our period of data, and discontinuation and restarting rates for the cohort of long term statin users (no single prescriptions) are all presented in Appendix A.6.4 and A.6.5.

Figure 6.2: *Kaplan Meier plots of the time until discontinuation and restarting statins for the first, second, third and fourth discontinuation periods, and the first, second and third restarting periods.*

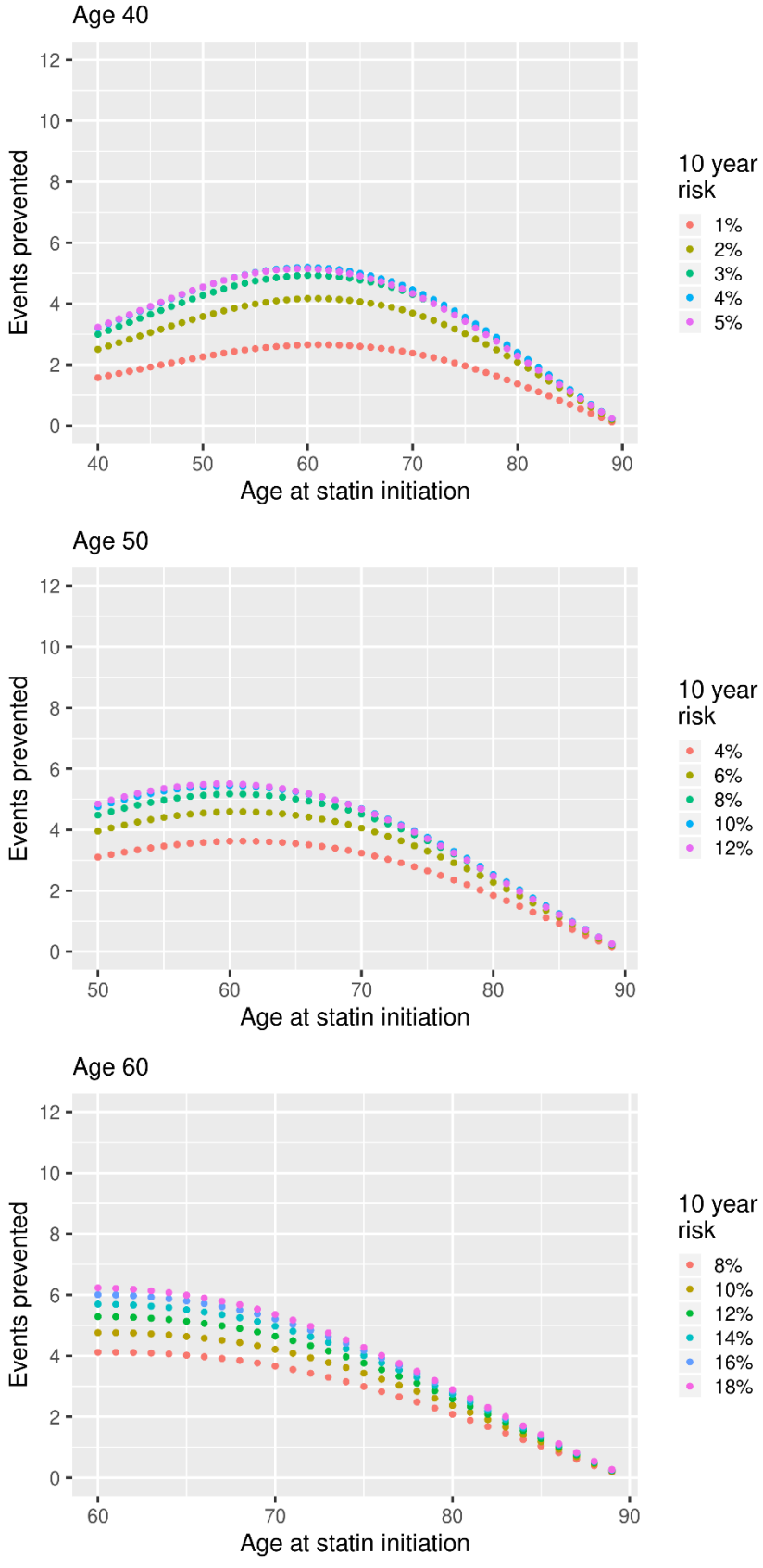


6.4.3 Effect of delaying initiation on cardiovascular events

Figure 6.3 shows the number of cardiovascular events prevented compared to no statin treatment when delaying statin initiation by different amounts (for males; females results are in Appendix A.6.6). Each data point in the graphs represents a different scenario. We present separate graphs defined by the age at the start of the simulation. Within each graph, we have a separate trajectory for each risk group (10-year risk at risk assessment). Within each trajectory the cohort of individuals is the same for each data point, the only difference is the year of follow up in which we initiated statin treatment (and therefore the risk level of the individuals at statin initiation also). We are interested of the maxima of each trajectory, which represents the optimal time to initiate statins for this group. For males aged 40, a delay of 15 years in starting statins resulted in a marginally higher number of cardiovascular events prevented. In contrast, for males aged 60, a delay in starting statins resulted in fewer cardiovascular events prevented due to competing effects of mortality. Results were similar for the female cohort, although the trajectories were shifted by around five years, with it being optimal to prescribe slightly later (Appendix A.6.6).

Illustrative example: Consider prescribing statins to a cohort of 50-year old men with a 10% 10-year risk of cardiovascular disease, we prevent 4.78 events per 100 individuals over the 40 year follow up. If we took this same cohort of men, but instead waited 10 years before initiating statins, at which point their 10-year risk of cardiovascular disease would be approximately 20%, then we prevent 5.45 events per 100 individuals over the 40 years follow up.

Figure 6.3: *Number of cardiovascular events prevented per 100 people over the duration of follow up with different time delays in starting statins, stratified by baseline age and 10 year risk of cardiovascular disease (male)*



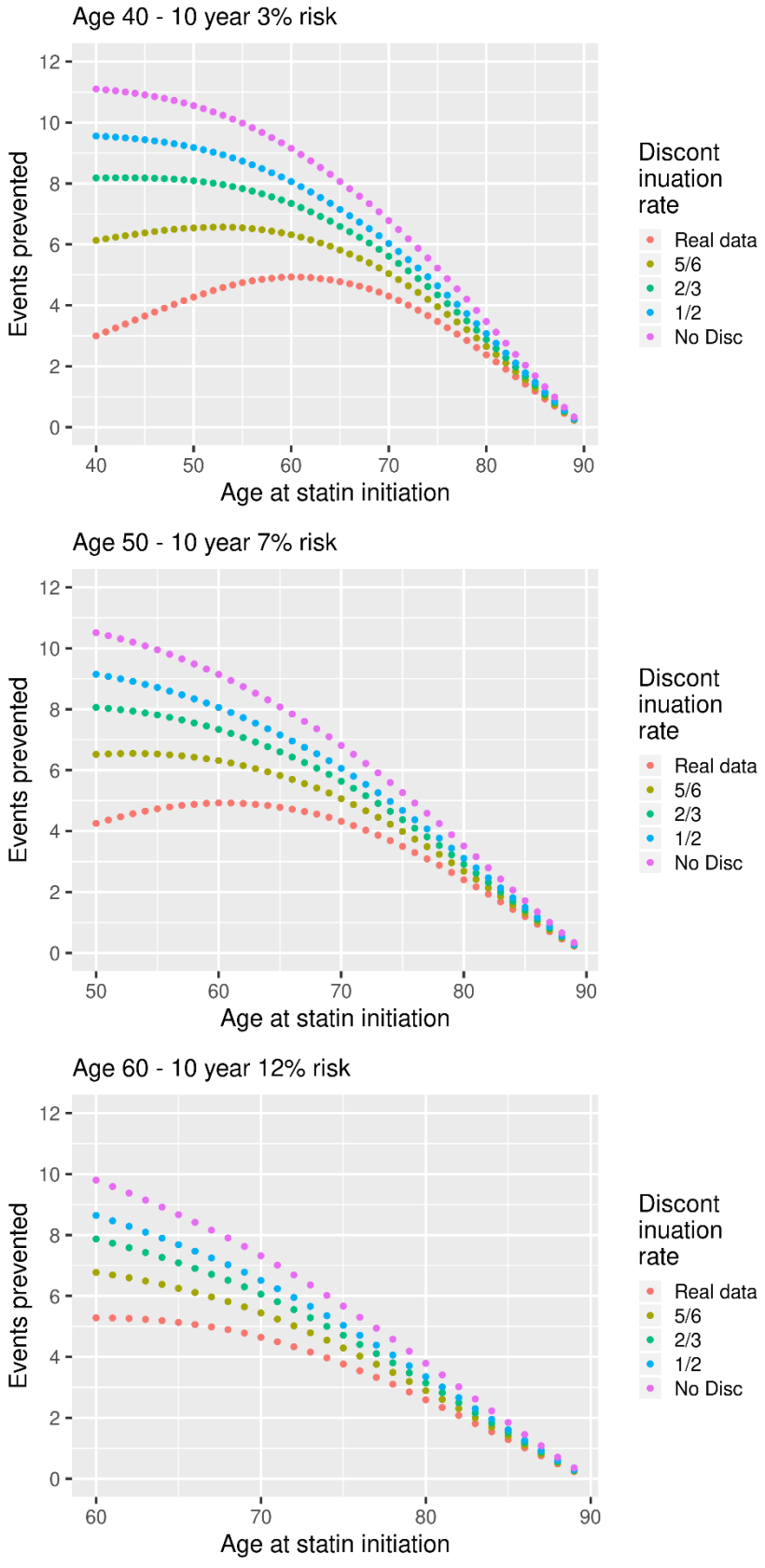
6.4.4 Effect of increasing statin adherence on cardiovascular events

Figure 6.4 shows the effect of reducing the discontinuation rate to 5/6, 2/3, 1/2 of the rate we found in practice, and no discontinuation. For each age group, a single 10 year risk (close to the median of that age group) was selected to showcase the effects, so all trajectories within a plot consider the same group of patients. It's shown the more adherent to statins people are, the more benefit they receive, and this benefit is increased the earlier prescribing is initiated (for males; females results are in Appendix A.6.6). This is in contrast to the trajectory derived from real-life discontinuation rates, which suggests little difference between initiating statins at age 50 or 60.

Illustrative example: Consider prescribing statins to a cohort of 50-year old men with a 7% 10-year risk of cardiovascular disease. Per 100 individuals, 4.25 events are prevented if discontinuation rates remain as normal, 6.52 if discontinuation is reduced to 5/6, 8.06 events if discontinuation is reduced to 2/3, 9.15 events if discontinuation reduced to 1/2, and 10.51 events if there is no discontinuation. The equivalent number of events prevented for a cohort with 10% 10-year cardiovascular risk are 4.76, 7.29, 9.01, 110.22 and 11.77.

Results from all sensitivity analyses outlined in the methods are provided in Appendix A.6.6. A small discussion is also provided, the results echoing those from the primary analysis, except there were slightly larger gains to be made by delaying statin initiation in women by the same amount.

Figure 6.4: *Number of cardiovascular events prevented per 100 people over the duration of follow up with different time delays in starting statins, stratified by baseline age and discontinuation rate (male)*



6.5 Discussion

There are three key findings from this study. The first is that between the ages of 40 - 70, the statin initiation time had a meaningful effect on the number of events prevented. Furthermore, the risk score of a patient had a negligible effect on the optimal time to initiate statins, which was driven by age. The second is that discontinuation and restarting rates get higher with consecutive treatment periods, underlining a complex pattern of statin usage over time. The third is that large gains could be made by improving adherence.

We see fairly large differences in the number of events prevented when statins were initiated between the ages of 40 – 70 with a peak around age 59 (male) or 63 (female), regardless of the risk scores of the patients. Initiating statins below the age of 50 was associated with far fewer events prevented, however it is unlikely for patients this young to have a risk > 10% (the threshold for cost effectiveness), and so this is unlikely to happen in practice. However it is not uncommon for a 50 year old to have a risk of 10% or more. Our data indicates that delaying statin initiation by 10 years could prevent an extra 0.67 events per 100 men treated, and 0.96 events per 100 women treated. These gains are small but not insignificant, and are likely driven by the fact that adherence improves with age (until around age 70, see Figure A.6.3), but also patients will not take the drug forever, and secondary or tertiary users are less likely to continue with treatment (Figure 6.2). There is therefore an optimal spot to be found which ensures patients are offered the drug when they are most adherent, at a high enough risk to gain benefit, but also that the risk of death or having a CVD event prior to receiving treatment is small enough.

Interestingly, for a given adherence level, the optimal time to prescribe is driven primarily by age rather than the 10-year risk (motivation for carrying out this study was that it may be best to initiate statins when patients reach a certain risk threshold). In Figure 6.3 the maxima of each trajectory are at the same age despite differing risk levels. This suggests that given the adherence levels we see in practice, in order to prevent the most events in the population, the optimal time to initiate statins for men is around 59 (women 63), irrespective of the risk score of the patient. While the risk score drives whether taking statins is cost effective or not, it does not drive when the optimal time in a patient's life to take statins is, which our work suggests is driven by age. The distinction can be highlighted by if a patient has perfect

adherence (Figure 6.4), the optimal time to initiate statins is as early as possible, but the treatment may not be cost effective at this point.

The potential to prevent more events in the population using such an approach brings up some important ethical concerns. Gains would be made from ensuring all patients will receive the drug when it will have most benefit (not too early, not too late). However, alongside any gains made by delaying statin initiation to a certain age, there will be a cost to adherent patients who would have continued treatment if starting at a younger age. Arguably it is unethical to improve the health of the population in this manner. In an ideal world we would know the adherence of a patient before initiating them on treatment, and could then initiate at the most appropriate time for that patient. Unfortunately this is not possible, and we would be forced to use population level discontinuation rates, which has these concerns.

We found inconsistent use of statins by patients in primary prevention. We also found higher discontinuation and restarting rates during the later treatment periods. This provides extra information beyond the current literature, which reports the initial discontinuation and restarting rates.[177] Figure 6.4 highlighted improving adherence could have a larger impact than adjusting when we initiate patients. This is not unsurprising, given this results in more time on treatment, however could be difficult to achieve. The most recent Cochrane review of 35 studies of statin adherence improving interventions suggested that only intensified patient care interventions (electronic reminders, pharmacist-led interventions) improved adherence when compared with usual care.[185] Like other studies,[177,186] this study suggests that people are likely to discontinue their statin when it is newly prescribed. Targeting a patient-centred, theory-based low-cost intervention which focuses on patients' concerns during this key initial period has been shown to improve adherence by 11% in a range of chronic illnesses,[187,188] and forms the basis of a National Health Service commissioned service in England (New Medicines Service[189]). This service is not currently provided to people starting statins, however, a randomised controlled trial of delivery of the same intervention in long term statins users demonstrated improved adherence.[190] This suggests that extension of the New Medicine Service into statin users could demonstrate effectiveness.

6.5.1 Limitations

There were three key limitations we identified in this study. 1) We used prescription data as a proxy for patients taking statins. This is a limitation as we only know a patient was given a prescription by their GP, we do not know if they picked the drug up, or took the drug. Therefore there is a possibility discontinuation rates are higher in practice, which would push the optimal time to prescribe further back. However there is currently no better way to measure adherence in the UK on a large scale, until prescribing and dispensing data are linked. Secondly, we only consider patients on treatment if they continually pick up their prescriptions (i.e. our algorithm). We think it is unlikely patients will have discontinued treatment but continue to pick it up. 2) We extrapolated the statin discontinuation and restarting rates for the length of the simulation. Data on statin usage over an individual patient's lifetime would be highly valuable to inform work such as this, but is not available. 3) We did not stratify the second and third discontinuation rates or first and second restarting rates based on age, despite age being a predictor of statin adherence.[191] Our reasoning is that this would have significantly reduced the cohort size available to calculate discontinuation rates, a particular issue for the second and third treatment periods at 10 years follow up. Given we were extrapolating data from this point, this was undesirable. Given the impact of discontinuation rates on the optimal time to initiate therapy, further work could be done to explore the impact of changes in statin intensity and dose on discontinuation rates, and subsequently the best time to implement these changes.

6.5.2 Conclusions

In certain scenarios, a small but not insignificant number of extra CVD events could be prevented by delaying statin initiation beyond a risk of 10% until reaching a certain age (59 for men, 63 for women). These findings are based on the discontinuation and restarting rates in England. Currently all thresholds are based around a patient's risk score, which drives cost effectiveness. However a combination of age and adherence levels are the most important factors in determining the optimal point in a patient's life to initiate statins. However, the clinical benefit must be weighed up against ethical concerns of such a strategy, which may disadvantage the most adherent patients. A less controversial strategy which could result in

preventing more events would be to focus on improving adherence, although this may be harder to achieve.

7 Impact of lowering the risk threshold for initiating statin treatment on statin prescribing

Mr Alexander Pate, Richard Emsley, Tjeerd van Staa

7.1 Abstract

Background: In 2014 the National Institute for Health and Care Excellence changed the recommended threshold for initiating statins from a 10-year risk of cardiovascular disease (CVD) of 20% to 10% (CG181), making 4.5 million extra people eligible for treatment.

Aim: To evaluate the impact of this guideline change on statin prescribing behaviour.

Design and Setting: A descriptive study using data from Clinical Practice Research Datalink (primary care database in England).

Method: We identified people aged 25–84 being initiated on statins for the primary prevention of CVD. CVD risk predictions were calculated for every person using data in their medical record (calculated risks), and were extracted directly from their medical record if a QRISK score was recorded (coded risks). The 10-year CVD risks of people initiated on statins in each calendar year was compared.

Results: The average 'calculated risk' of all people being initiated on statins was 20.65% in the year before the guideline change, and 20.27% after. When considering only the 'coded risks', the average risk was 21.85% before the guideline change, and 18.65% after. The proportion of people initiating statins that had a coded risk score in their medical record increased significantly from 2010 – 2017.

Conclusion: Currently available evidence, which only considers people with coded risk scores in their medical record, indicates the guideline change had a large impact on statin prescribing. However, that analysis likely suffers from selection bias. Our new evidence indicates only a modest impact of the guideline change. Further qualitative research about the lack of response to the guideline change is needed.

7.2 Background

In July 2014 the National Institute for Health and Care Excellence (NICE) changed the recommended threshold for initiating statin treatment for primary prevention of cardiovascular disease (CVD) from a 10-year CVD risk of 20% to 10% (CG181).[45] This decision came alongside huge debate in academic and clinical literature as lowering thresholds could have a huge impact on clinical practice.[81,83,86] It was estimated that the guideline change would make a total of 11.8 million people in England (37% of adults aged 30 – 84) eligible for statins,[192] and was met with opposition by a group of leading doctors.[47] NICE estimated that an additional 4.5 million people would be eligible for statins, preventing up to 28,000 heart attacks and 16,000 strokes each year.[193] Without an increase in statin prescribing in people with risks between 10 – 20%, this number of extra CVD events would not be prevented.

To our knowledge, only one study has assessed the impact of this major guideline change in practice (see section "*Impact of NICE guidance*").[97] In England, the QRISK3[5] (previously QRISK2[99]) risk prediction model is recommended by NICE for calculating the 10-year risk of a CVD event to guide treatment decisions for the primary prevention of CVD. The study analysed people that were initiated on statin treatment and had a QRISK2 score recorded in their electronic health record. They found that the average risk score of people receiving statins dropped from 23.06% before the guideline change to 19.28% after.[97] This provides evidence the guideline change was impactful and the results are quoted in the NICE impact report for cardiovascular disease prevention.[194] However the same study also reports that since 2012, 72.9% of people initiated on statins did not have a QRISK2 score recorded.

The aim of the present study was to evaluate the impact of reducing the risk threshold from 20% to 10% by analysing the risks of all people being initiated on statins for primary prevention of CVD. We also replicate the analysis carried out by Finnikin et al.,[97] considering only people with a QRISK score in their medical record.

7.3 Methods

7.3.1 Cohort definition

This project used data from the Clinical Practice Research Datalink (CPRD)[48]. This data was linked with Hospital Episodes Statistics[49] and Office for National Statistics[50] for identifying CVD events. Linkage to HES restricts this dataset to England only. Two cohorts were defined, a primary prevention cohort and a statin initiation cohort. The primary prevention cohort consisted of people aged 25 – 84 with no history of CVD (composite outcome of coronary heart disease, ischaemic stroke or transient ischaemic attack) or statin use. The cohort entry date was defined as the last of 25th birthday, one year permanently registered in CPRD, or 1st Jan 1998. People were excluded if they had a CVD event or statin prescription prior to their cohort entry date (code lists in Appendix 1). People were censored at the earliest date of transferred out of practice, last data collection for practice, CVD event, death, or 31st Dec 2017.

An individual from the primary prevention cohort was included in the statin initiation cohort on the date of their first statin prescription if this first statin prescription was issued at least one year after the start of follow up

7.3.2 Statin initiation rate

The primary prevention cohort was used to calculate the statin initiation rate each year. For each calendar year we calculated the total number of statin initiations and the total number of days follow up. Follow up for each person stopped either when they were censored or initiated on statin treatment. Calendar years ran from the 1st July each year, to match the date at which the guidelines were changed (July 2014). The final period (2017 – 2018) finished on 31st Dec 2017.

7.3.3 Comparisons of risks of people initiated on statins each year

For each individual in the statin initiation cohort, we extracted all the predictors required to generate a QRISK3[5] score from their electronic health record (EHR). A full list of variables, code lists, information on variable derivation, the amount of missing data and details of the imputation process are in Supplementary Boxes S1 and S2. The 10-year CVD risk of each

person at statin initiation was then calculated using QRISK3,[5] an R package was used for this[195]. We refer to these as the 'EHR derived risks'.

Where recorded, we extracted coded QRISK scores directly from the electronic health record if they were within 180 days prior to or 30 days after the first statin prescription (code list in Supplementary Box S1). These risk scores are referred to as 'coded risks', and are used to replicate the analysis by Finnikin et al.[97] The coded risks will have been calculated using a mix of iterations of the original QRISK[196] algorithm and the QRISK2[99] algorithm.

The following analyses were carried out using both the EHR derived and coded risks. The average risk of people initiated on statins in each calendar year was calculated. Intervals ran from 1st July, as to match the date of the threshold change, which was July 2014. We calculated the proportion of people initiated on statins each year that were classified as low risk (< 10%), intermediate risk (10 – 20%) or high risk (> 20%).

The agreement between the EHR derived risks and coded risks was evaluated using scatter plots. This was done to check agreement between the EHR derived and coded risk scores. A higher level of agreement would provide support that the analysis based on the EHR derived risks is valid, given the coded risks can be viewed as the gold standard. By using scatter plots, we compared agreement on the most granular level possible (i.e., does the EHR derived risk match the coded risk in the database for each individual person?). The intraclass correlation coefficient[197] was also calculated for agreement between the EHR derived and coded risks within each calendar year.

7.4 Results

The primary prevention cohort included 3,892,603 individuals (51% female). The statin cohort consisted of 351,553 individuals (47% female). The demographics of the statin cohort are provided in Supplementary Table S1.

The statin initiation rate per 1000 person years by calendar time is presented in Figure 7.1. We see a peak of 21.79 in 2005 and a drop until 2010 – 2011 when the incidence rate flattens out at around 12.5. The number of people initiated on statins each year is provided in Table 7.1, as well as the proportion of those people that had a coded risk in their medical record.

Prior to 2010 - 2011 less than 5% of the statin initiations had an associated coded risk score. After this the proportion increases to 66.29% by 2017 - 2018.

Table 7.1: *Number of people initiated on statins each year, and number of those who had a coded QRISK score in their medical record*

Date	Follow up (years)	Number initiated	Number with coded score	Proportion with coded score
98-99	1090072.9	3510	26	0.74%
99-00	887549.2	4240	66	1.56%
00-01	1141713.0	7498	232	3.09%
01-02	1318576.6	12335	450	3.65%
02-03	1449309.0	17908	457	2.55%
03-04	1547360.7	26959	322	1.19%
04-05	1563126.1	30529	272	0.89%
05-06	1588051.6	34604	390	1.13%
06-07	1591314.4	32967	316	0.96%
07-08	1598293.4	27432	211	0.77%
08-09	1601472.2	29554	501	1.70%
09-10	1569415.3	24883	1053	4.23%
10-11	1513887.1	18972	1156	6.09%
11-12	1453955.3	18622	2314	12.43%
12-13	1402210.8	18181	3219	17.71%
13-14	1245691.7	14689	3831	26.08%
14-15	1021942.4	10938	4677	42.76%
15-16	749647.7	8572	5012	58.47%
16-17	540323.3	6511	4188	64.32%
17-18	233582.6	2649	1756	66.29%

Figure 7.1: *Statin initiation rate in each calendar year*

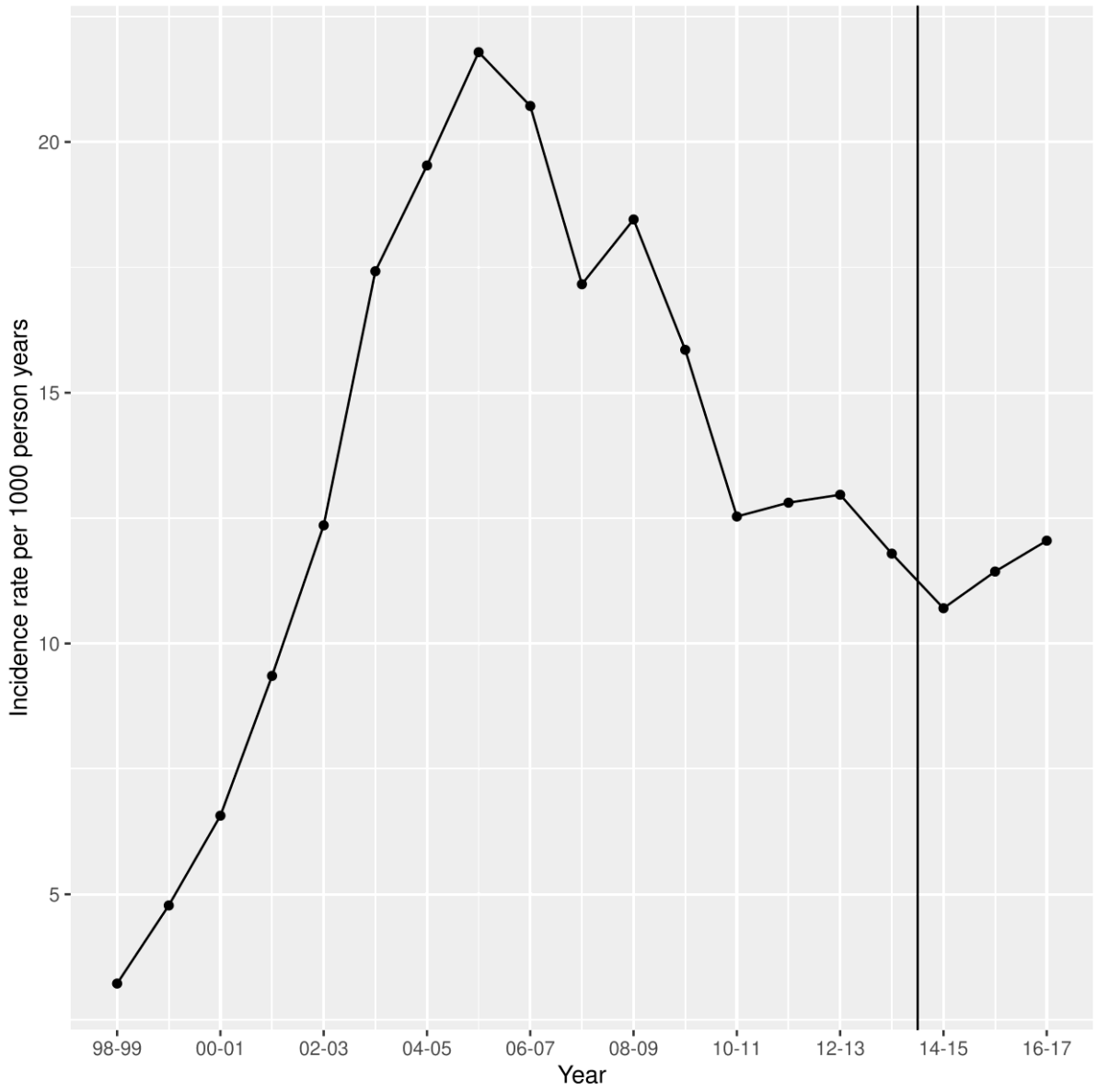


Figure 7.2 plots the average EHR derived risk and average coded risk of people being initiated on statins each year. The latter is restricted to those who had a coded risk score available. There is no clear change to the average EHR derived risk of people being initiated on statins from 2013 - 2014 (20.65%) to 2014 – 2015 (20.27%), the year of the guideline change. However, there is a drop in the average coded risk from 21.85% to 18.65%.

Figure 7.2: *Average 10-year risk of people initiated on statins in each year for primary prevention of CVD. The EHR derived risks were calculated for all patients initiated on statins, coded risks are restricted to those with coded risks in their medical record*

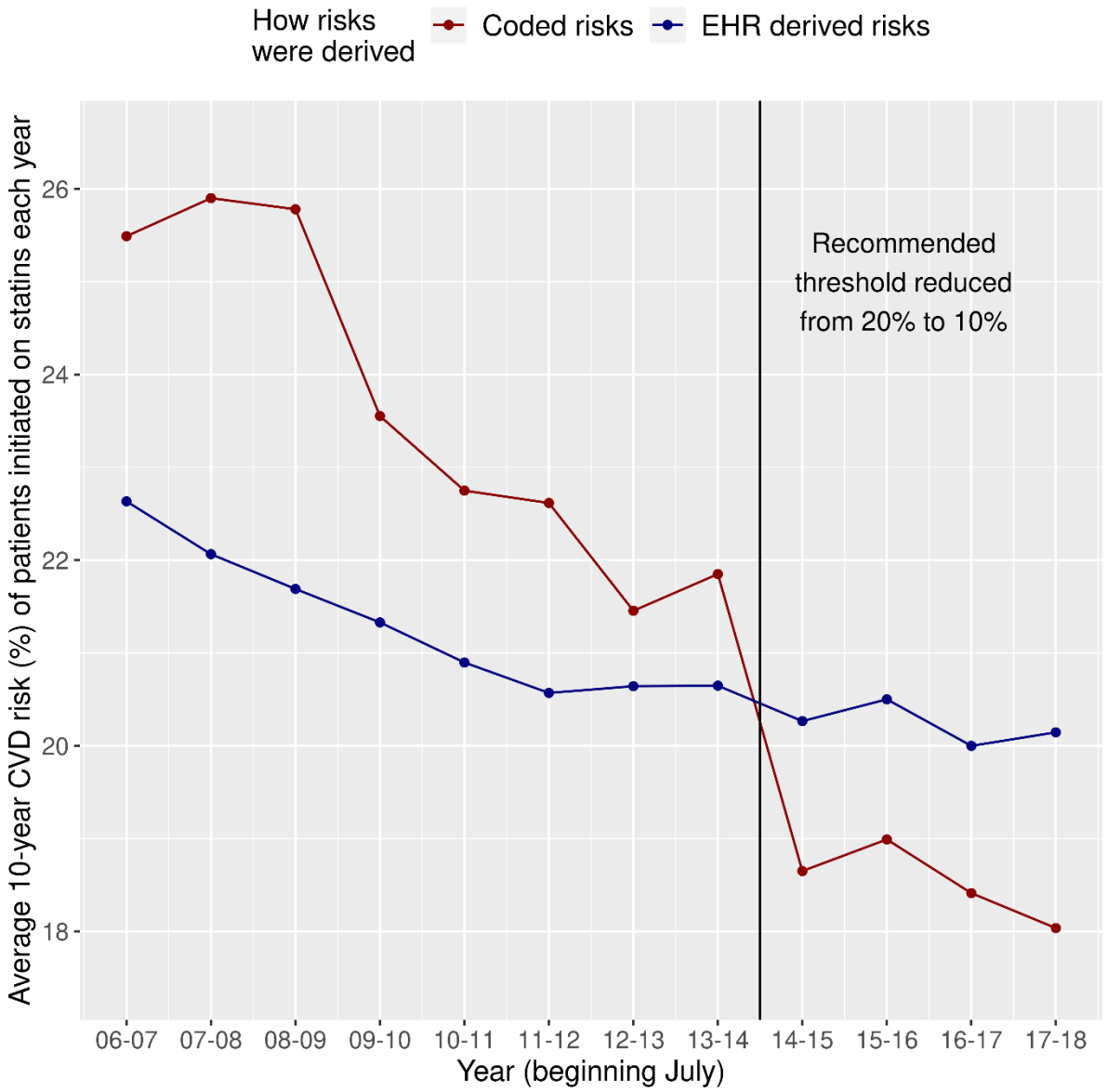


Figure 7.3 shows the proportion of people initiated on statins each year that belong to each risk category. For the EHR derived risk scores there is a steady increase in the proportion of people in the 10 – 20% risk group from 2013 – 2014 onwards. However this happens mostly at the expense of people from the < 10% group, as well as some from the > 20% group. For the average coded risk score, there is a sharp increase in the proportion of people in the 10 – 20% risk group, which comes at the expense of people in the > 20% group.

Figure 7.3: *Proportion of people initiated on statins each year that belong to each risk category (< 10%, 10 – 20%, > 20%). Separate plots for EHR derived risk scores (all statin initiators) and coded risk scores (restricted to those with coded risks in their medical record).*

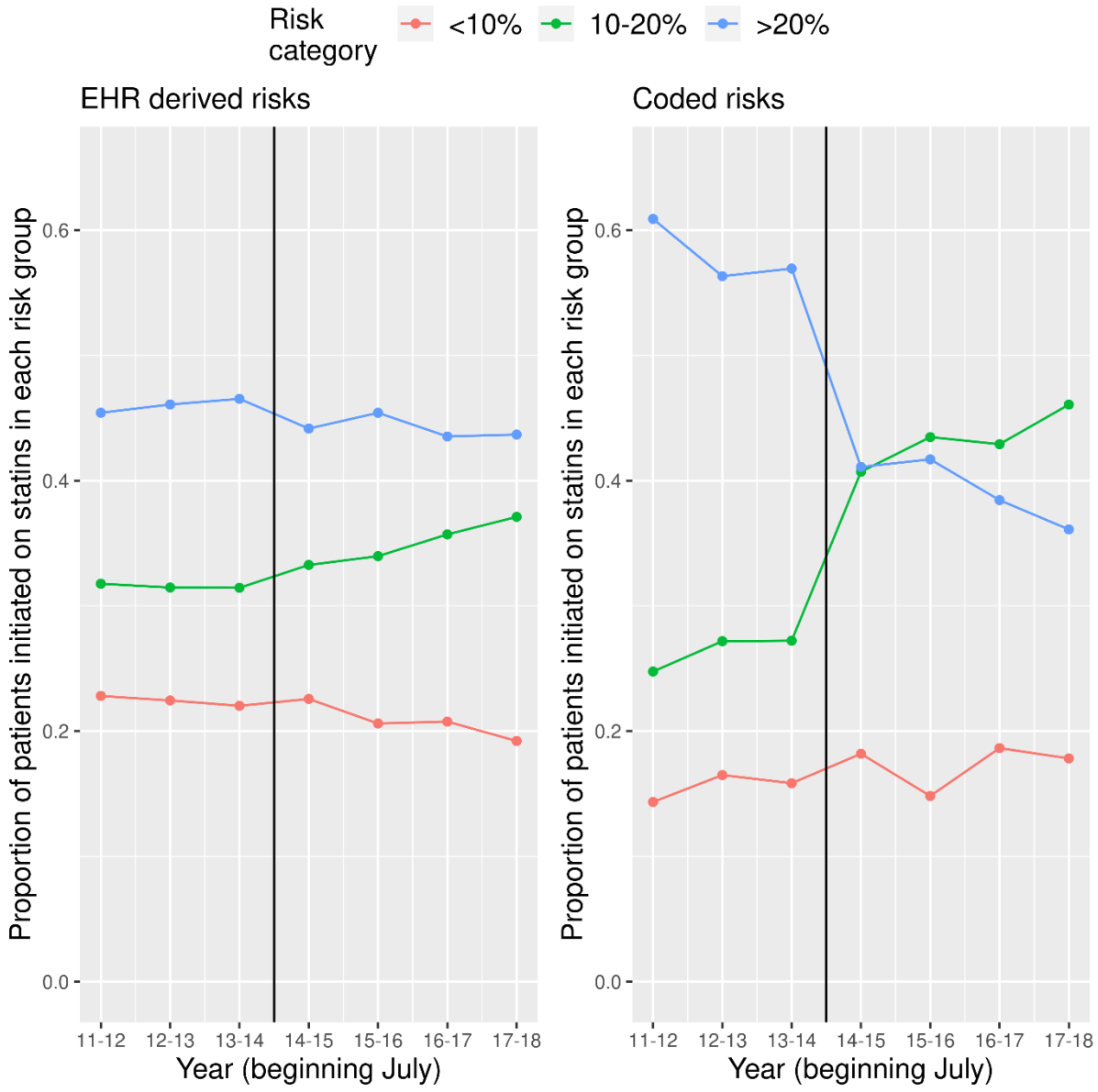


Figure 7.4 plots the EHR derived risks against the coded risk scores for each individual stratified by year, with a blue line added to illustrate perfect correlation. Overall, we see a strong positive relationship between the two, although there are quite large levels of variation either side of perfect agreement. Also, from 2014 onwards we see more consistent over prediction of the EHR derived algorithm compared to the coded risk scores. However,

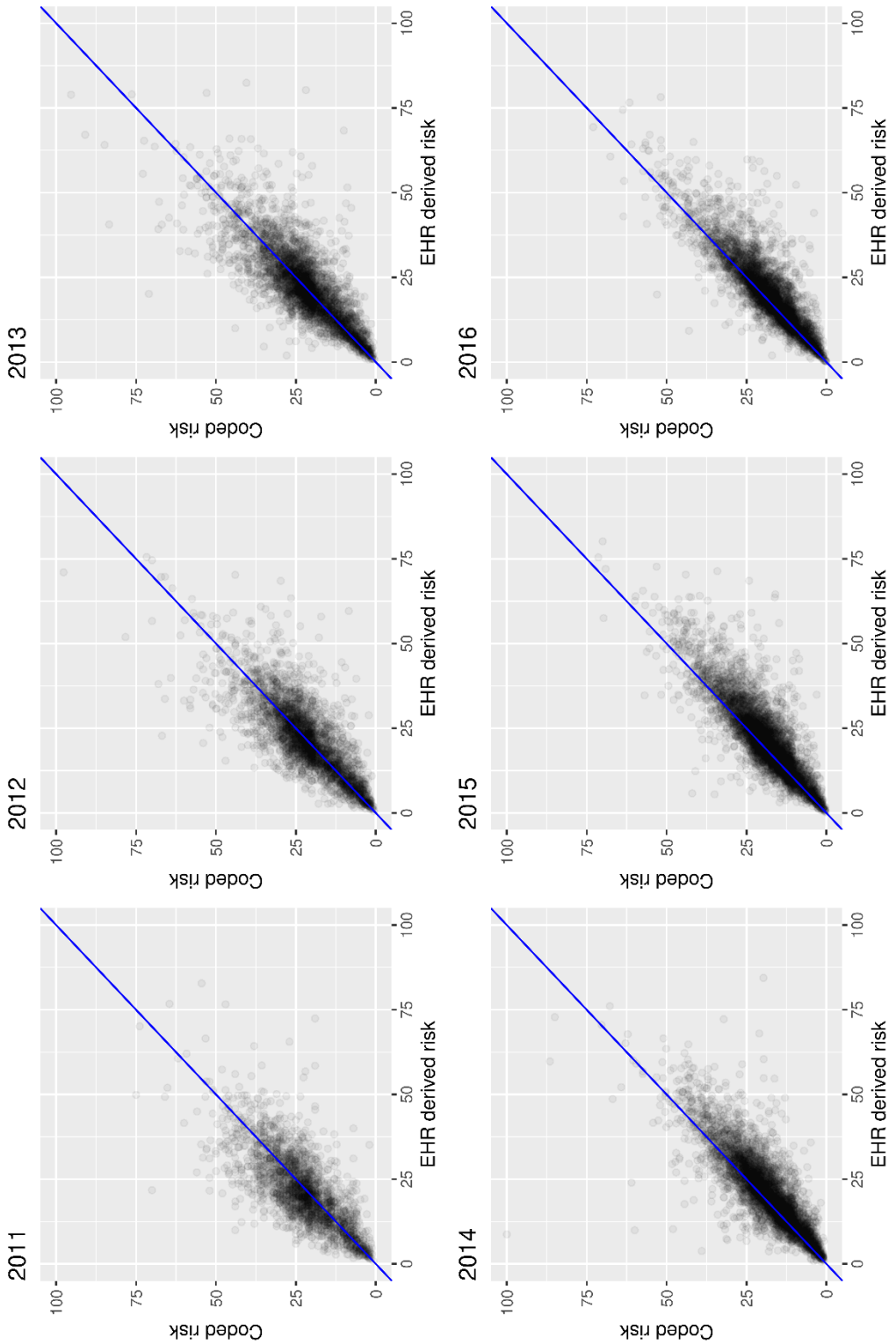


Figure 7.4: *EHR derived risks plotted against coded risks, stratified by year of statin initiation. Blue line = 1:1 correlation*

the intraclass correlation coefficient improves between 2011 and 2016 (0.75 in 2011, 0.79 in 2012, 0.81 in 2013, 0.81 in 2014, 0.82 in 2015, 0.85 in 2016).

7.5 Discussion

7.5.1 Summary

There was a large reduction in the average coded risk of people initiated on statins, which closely matches the currently available evidence.[97] When viewed in isolation, the reduction in the average coded risk score (Figure 7.2) and the change of proportion in each risk category (Figure 7.3) indicate a significant change in clinical practice. NICE have quoted this evidence in their impact report.[194] However, because the coded risk analyses only consider the subgroup of people with a coded risk score, this analysis is at risk of cohort selection bias as the subgroup may not be representative of all people initiated on statins. This risk is exacerbated by the increasing proportion of people initiated on statins that have a coded risk score (Table 7.1). As this subgroup increases in size, unless risk scores are recorded at random, this will have a significant impact on the average risk of this subgroup. Importantly, the changes in risk are driven by changes in who GPs are recording risk scores for, rather than a change in who is receiving statins.

We found no change in the average EHR derived risk of people being initiated on statins after the guideline change, and a small increase in the proportion of people that belonged to the 10 – 20% risk group. This analysis is not affected by the same selection bias, as it considers all people initiated on statins each year. Therefore with the extra data presented in this paper we believe the response to the guideline change is not as impactful as first thought.

7.5.2 Comparison with existing literature

To the authors knowledge, only one other study has measured the impact of CG181 on clinical practice.[97] We enhance this research with an expanded analysis considering all people initiated on statins, and validate their findings by replicating their analysis in CPRD. Our data indicates recording practices of GPs had a significant impact on the average coded risk. The proportion of people with a coded risk is small and increasing rapidly at this time (26.08% in 2013 – 2014 and 42.76% in 2014 – 2015). It is highly likely that the subgroup of people receiving a coded risk score was changing (it is unreasonable to assume this increase in

recording was happening at random), but the typical patient being initiated on statins was not. One hypothesis is that GPs became far more likely to calculate the risk of someone in the 10 – 20% range using a QRISK tool after the guideline change, but their prescribing behaviour remained the same.

We found no reduction in the average EHR derived risk after the guideline change (Figure 7.2). While this indicates the guideline change had no impact, considering all the results leads to a slightly different conclusion. This constant average risk appears to be caused by a combination of a small increase in 10 – 20% risk people initiated on statins, and a drop in low risk (< 10%) people. In Figure 7.3, we see a steady increase in the proportion of people in the 10 – 20% risk group, and a decrease in the other two groups (a larger decrease in the < 10% risk group). It is possible that the guideline change has had an equal effect on preventing statin initiation in low risk (< 10%) people, as it has on increasing statin initiation in the target 10 – 20% risk group, resulting in no change to the average EHR derived risk.

The data agrees with a modest impact of the guideline change, but the changes are far more subtle than would be concluded if looking at the currently available evidence. These findings are important as the numbers in the widely quoted statistic[93–96] "prevent up to 28000 heart attacks and 16000 strokes each year",[193] are likely far from being achieved.

7.5.3 Strengths and limitations

This is the first study to evaluate the impact of the NICE guidance CG181 on the risks of all people receiving statins in England. The study cohort is large and results are likely generalizable to the English population as CPRD is representative of the UK in terms of age, sex and ethnicity[48].

There are two key limitations in this work. The first is the imperfect agreement between the EHR derived risks and the coded risks, because the EHR derived risks should represent the risks of individuals as closely as possible. Potential reasons for the disagreement between these and the coded risks are that the EHR derived risks use the QRISK3[5] algorithm, whereas QRISK2[99] will have been used in practice over those years; this study used multiple imputation to impute missing data, whereas missing data is imputed using mean imputation when coded QRISK scores are generated by GPs; we have identified variables using code lists which may not perfectly match those used by the algorithm in practice; and this study considers coded

risks within a window of six months prior to statin initiation, while patient data could have changed in that time. Despite some disagreement, the relationship was strong enough that if the people with a coded risk score were a random subset of all people initiated on statins, there would have been a large drop in the average EHR derived risk after the guideline change (like there was in the average coded risk). This was not the case, indicating the likelihood of selection bias in the coded risks analyses. This necessitates the analysis using the EHR derived risks, even if the estimated risks are not perfect. The second limitation is that many practices left CPRD towards the end of this study, resulting in a risk of selection bias in our cohort if the drop out was not at random. However we have no reason to believe that people from practices that dropped out were more or less likely to be initiated on statins. Furthermore, our results considering the coded risk scores were comparable to those of Finnikin et al.,[97] a study carried out in The Health Improvement Network database[198] which has not suffered from this limitation.

7.5.4 Implications for research and practice

The change in NICE guidance appears to have had a small effect on statin prescribing by GPs. Given NICE invests time and resources into developing these guidelines, it would be worthwhile for them to understand why there has been such little response. We propose a qualitative study with GPs and patients to assess the barriers to statin initiation for the primary prevention of CVD in 10 – 20% risk people. A recent scoping review[199] of the current literature regarding the use of statins to prevent CVD found only three studies specifically considering primary prevention of CVD, and that "it was difficult to interpret how doctors' or patients' attitudes would vary according to the risk profile of the individual patients". A systematic review provided a comprehensive review on patient attitudes towards taking statins,[200] however the majority of studies were looking at long term adherence, as opposed to statin initiation. No studies had investigated specifically the willingness to initiate at a 10% or 20% threshold for primary prevention of CVD. A debate article published in 2016[201] discusses patient attitudes to taking statins in light of the NICE guidance change, attributing the lack of uptake in lower risk patients to transferability of evidence from research to practice and the potential for side effects. However, the evidence base[202–205] for their findings pre-dates the large amount of pro-statin research that came about in 2013 that has fuelled the statin debate. The authors also noted "there is sparse literature regarding the views of

GPs". Some qualitative research does exist in this area,[203,204,206–208] but again no studies have been carried out in the wake of the NICE guidance, or on prescribing specifically at 10% compared to 20% risks.

8 General discussion

The aim of this project was to explore two key aspects of the process of using risk prediction models to drive treatment for a patient: 1) Calculating the risk prediction of a patient on which clinical decisions will be based, and 2) Deciding what clinical action to take for a patient with a given risk prediction. These are two very broad areas, and the research in this thesis has focused on very specific aspects of them. Chapter by chapter, this discussion will link the findings back to the original aims, talk about the broader implications of this research (particularly for the clinicians who use these models), and discuss future work ideas in more detail than was possible in the chapters which were formatted for publication. The limitations of the thesis as a whole are also discussed.

8.1 Chapter 3

This chapter quantified the extent that cardiovascular disease (CVD) risk predictions for individuals varied when extra variables were added to the model. The discussion in the Chapter focused on the clinical implications of this uncertainty, here I will focus on its context with respect to the reference class problem. The models considered were developed on the same cohort, using consistent variable definitions for common variables across the models. This reduces the reference class problem to its purest sense, where differences in probabilities (risks) are down to conditioning on a different set of variables only. The most prominent finding, which was discussed heavily in the chapter, was the extent with which the calendar time variable affected the risk scores of individuals. However without the introduction of this variable, there were not large levels of variation in risk scores for individuals. This means beyond the secular trend, the reference class problem was not having a big effect.

The discordance between individual risk estimates calculated from ASSIGN, QRISK2 and Framingham shown in the literature,[23] is therefore probably due to the models being developed in different populations, using different outcome and predictor variable definitions and different cohort selection criteria. These differences are not an issue, as long

as when using a model to generate a risk score for an individual, that the individual belongs to the population in which the model was developed, the outcome of interest matches exactly the outcome in the model, and data on the individual has been collected in the same way as the data was for model development. If these criteria are met, and the major CVD risk predictors are included in the model (as was the case in our study), our research shows that risk scores for an individual will not be overly sensitive to which extra predictor variables are conditioned on (the reference class problem is minimised). In this case a clinician may be confident in the use of that risk prediction model. Comments outlined in section 1.2.1 stated that the reference class problem limits the *"clinical utility of risk prediction models"*.[13,17] I argue that in the case of CVD risk prediction models with all the major predictors included, it does not. To refer back to the original aim, this is not a major source of uncertainty in the generation of risk scores. This is of course ignoring the issue of calendar time, which was explored further in chapter 5.

The future work from this chapter has been largely carried out in Chapter 5 of this thesis, and so is not discussed here.

8.2 Chapter 4

In this chapter a simulation was carried out to quantify the stability of risk scores generated from models with various sample sizes. High levels of instability were found at sample sizes similar to models used in practice, and very high levels of instability when following recently published sample size formula. With respect to the original aims outlined in this thesis, this is a major source of uncertainty when deriving risk scores for individuals. The findings also provide an alternative explanation for the discordance in risk scores that has been attributed to the reference class problem that was reported in section 1.2.1. It is likely a large amount of this discordance was driven by sampling variation. The discussion in Chapter 4 argues why this uncertainty is an issue, not just on an individual level (patient's treatment decision dependent on a random process), but on a population level as well (imprecise estimation of the risk of subgroups of the population will lead to poorer model performance). The broader implications are relevant for both those developing models and the clinicians using them.

In terms of model development, it is important to develop more clarity over what an acceptable level of discrimination is. This is particularly pertinent in survival models. The sensitivity and specificity of a logistic model at a certain cut-off is a clear way of assessing whether the models performance is clinically acceptable or not. However due to censoring in survival data, sensitivity and specificity cannot be calculated. Instead relative measures of discrimination are used, such as Harrell's C,[121] which calculates the proportion of pairs of patients that both have events, where the patient with the higher risk has the event first. Therefore the impact of discrimination on the clinical applicability of a model is not as clear, and there is no clear threshold as to what is acceptable. These measures are a good way to compare different models, but are they sufficient for assessing the clinical usability of a given model?

I suggest that a requirement for a risk prediction models is that the risk scores are stable, and not heavily dependent on sampling variation. This ensures the risks of subgroups defined by the predictors included in the model are calculated precisely. If not, there is a high probability that the risks of these subgroups in the population (the desired estimate) will be incorrectly estimated by any given model. This leads to poorer model performance, and ultimately the wrong subgroups of patients receiving treatment. Given the stability of the risk scores from models meeting the sample size criteria, it appears this would be a stronger condition than minimising overfitting. The two are closely linked, as minimising overfitting ensures the optimism of the model performance in the development dataset is below a certain level, while minimising instability ensures the model performance in the population reaches its upper bound for a given set of predictors.

The implications for clinicians are that if they are using a CVD risk prediction model developed on cohorts of 10 000 or less, they should be concerned. This is the case in America[61] and Scotland[64]. The evidence from this study indicates that risk scores for individuals are highly dependent on the random sample of patients used for model development. This problem could be prevalent across many disease areas, particularly in models predicting rare outcomes or where associations between the predictors and the outcome are not as strong.

8.2.1 Future work

In order to implement a strategy which ensures stable risk scores in practice, the main hurdle to overcome is how to ensure stability as part of the sample size calculations. The methods outlined in this chapter only measure the stability of risks for an existing model (access to the development data is necessary). This is useful for assessing the performance of an existing model, but not useful for ensuring a certain level of performance from the data collection stage. Some techniques for ensuring precise estimates of coefficients in the model are discussed in section 6 of the work by Riley et al.,[35] however stronger requirements are needed to ensure stability of the risk scores themselves, as the linear predictor is a function of all these coefficients. Research on how to ensure risk scores will have a pre-defined level of stability before collecting the data would be the next step.

8.3 Chapter 5

The primary aim of this study was to assess the potential miscalibration of risks scores for present day patients caused by the drop in incidence of CVD found in Chapter 3, and assess whether it was being driven by increasing statin use. If it was, there would be a strong case that the drop in CVD incidence should not be modelled, as the treatment decision should be based on the risk of a patient if they were not to take statins. However, the results from this chapter indicated this was not the case. The cause of the secular trend remained unclear, meaning this is a major source of uncertainty associated with the generation of risks scores in England, and has major implications for clinicians prescribing statins.

The drop in risk caused by accounting for this secular trend was drastic. It was estimated in Chapter 3 that 3.8 million patients in England would no longer be eligible for statins at a 10% threshold. There is already lots of debate over whether to initiate intermediate risk (10 – 20%) patients on statin treatment (see section 1.3.3), and the results of this study indicate that people in this risk category may have even lower risks than thought. Given the current guidelines, this means many patients will be receiving statins unnecessarily. This has a negative impact on both patients and the National Health Service. The treatment of healthy patients, who may go on to experience side effects is clearly not good. Furthermore, it will be

costing the National Health Service money to do so, as they are treating patients at risks which are not cost effective to do so, and this money could be spent better elsewhere.

8.3.1 Future work

I would first like to stress the desire for this analysis to be repeated in another database. In earnest, the results were surprising. The drop in incidence in CVD was very large and effect of statin use during follow up on the secular trend was minimal. While it is not obvious why, there could be some level of differential recording over time in CPRD practices which is driving the secular trend. We discussed in Chapter 5 that this is unlikely given the number of CVD events identified through hospital data and mortality records. However, reproducing these results in another database would still be valuable. Second, the cause of this secular trend remains a mystery. There is no evidence in the literature as to what is causing this drop. This is most likely because it is a very difficult question to answer, and the data to do so does not really exist. While trends over time can be identified through electronic health records (EHRs), establishing a causal effect between two variables is difficult. Potential causes for the drop in CVD incidence such as changes in lifestyle factors like diet or exercise are not well recorded in EHRs. Furthermore, the estimated causal effect of most variables on the outcome would be unverifiable due to a lack of trial data, which was not the case for the effect of statins. A body of specialised causal inference work is needed here to identify what are the driving factors behind this drop in CVD incidence, although the data may not exist to do so.

8.4 Chapter 6

Chapter 6 focused on the development of guidelines used to drive the allocation of treatment, specifically about the use of statins in England. In some scenarios a benefit was found to delaying statin initiation beyond a 10% risk until patients reached a certain age. Surprisingly, for a given level of adherence, the risk of an individual was not a major factor in determining when a patient should be initiated on statins, it was age. The findings shift the focus of the conversation away from what the optimal risk threshold is (section 1.3.3), to whether statin initiation should be based on age rather than risk.

Age based strategies have been proposed before.[209,210] The idea is to reduce the level of certain CVD risk factors in the population, rather than treating only the high risk individuals. This strategy is viewed as quite radical, but given the strength of age as a predictor in CVD risk prediction models, age and sex specific thresholds effectively already exist.[211] Given opposition already exists in the medical community against these indirect age based thresholds,[47] explicit use of this type of approach would undoubtedly be met with fierce resistance. However, our results provide a different reason for wanting to use an age based strategy. Current arguments are that it is cheaper and simpler to use a strategy based only on age compared to using risk factor based strategies, and they only perform marginally worse. The time and money saved could then be invested elsewhere, so this is a cost effectiveness argument. Doing so would result in more patients being treated for the same number of events prevented, but for less cost. The reasoning from this chapter puts patient outcomes first. If a patient is only going to take the drug for a specific amount of time (due to discontinuation), it is best to prescribe it to them at the time at which they will receive most benefit from it. This optimal time is driven by their age. Therefore a rule based on adherence and age could be applied on top of whatever cost effectiveness threshold has been decided (which would be a risk based threshold). If such a rule was implemented, it would pan out as follows.

Based on the discontinuation rates we saw, the optimal time to initiate statins was 59 (male) and 63 (female). A white asymptomatic individual at these ages with average body mass index (BMI), Cholesterol/high-density lipoprotein (HDL) ratio, systolic blood pressure (SBP) and SBP variability would have a 10-year risk of 8.8% (male) or 6.5% (female), calculated using the online QRISK3 calculator.[104] It would not be cost effective to initiate statin treatment in these patients. Therefore by the time it is cost-effective to initiate treatment in these patients, it would also be optimal to initiate treatment based on the age threshold, and their treatment pathway would be unaffected. The individuals this would affect are symptomatic individuals, or individuals from higher risk ethnicities, who will have risks > 10% prior to age 59 or 63.

For these individuals, this leaves the clinician and patient with a tricky decision. On average, given the discontinuation rates of the population, it is better to delay statins until those patients are older and will receive more benefit from the drug. However for an individual that would be adherent this would not be the best decision. Strategies to initiate treatment at

different times in patients we expect to be non-adherent or adherent would therefore provide the optimal solution. However it is not possible to know exactly which patients will be adherent and which won't, and it would be unethical to have different rules for different subgroups of the population. Therefore using the thresholds based on a population discontinuation rate is the only option, yet unideal due to the trade off in outcomes for adherent and non-adherent patients. One may argue symptomatic individuals should be taking statins with extra emphasis placed on adherence, and with better adherence it becomes optimal to initiate statins at a younger age anyway. However, as discussed in Chapter 6, improving adherence without intensive patient-centered interventions is difficult.

The practical implications of these findings are therefore not straightforward. From the policy maker's perspective, there is a chance to prevent extra events in certain subgroups, however the ethics of this approach need to be investigated first. From the patient's perspective, this information could aid their decision making. If a patient knows that they do not want to for a pro-longed period of time, they may opt to delay their own statin initiation and take the drug when it is most beneficial to them. Patients who believe they will be adherent over a long period of time could opt to initiate straight away. While similar conversations to this are already happening between clinician and patient, there is currently no information on when statins are most beneficial to take for an individual. I therefore doubt the findings and thought processes from this study are being included in those conversations.

8.4.1 Future work

The aim of this work was to assess whether it would be possible to prevent extra CVD events. The feasibility of the approach suggested must be discussed by the people actually involved with the decision (National Institute for Health and Care Excellence (NICE), the clinicians and the patients). Qualitative research to highlight the opinions of these parties in light of the findings from this chapter is the next required step.

8.5 Chapter 7

Chapter 7 focused on how the guidelines developed to drive the allocation of treatment are implemented in practice, specifically looking at the use of statins in England. Despite evidence

on the whole supporting the use of statins in 10 – 20% risk patients, there was resistance from the medical community to adopt this strategy.[47] This study provides empirical evidence of this resistance in practice. NICE have quoted the results of an analysis restricted to patients with QRISK codes,[97] in an impact assessment published in 2018.[194] They are therefore likely under the assumption that the guideline change has had a higher impact than it has. The broader implications of this research are not for clinicians, but for NICE.

First, I would like to put aside the conversation over statin efficacy at lower risk thresholds. Let's assume that statins are cost effective in 10 -20% risk individuals and that their uptake in this risk group would cost effectively prevent extra events, which is what the majority of evidence suggests. Based on this evidence, NICE made statements saying that this guideline change could prevent an extra 28 000 heart attacks and 14 000 strokes a year. To attain anywhere near these numbers a drastic change in prescribing habits was required, which did not happen. This highlights a shortcoming in NICE's operations. An extremely detailed document was produced providing the evidence behind the new threshold.[45,84] This indicates a large amount of time and resources was put into developing this guideline. However, for the recommendations to have any impact on patient outcomes, it's equally important for this guidance to be properly implemented by the medical community. It is not obvious what steps NICE took to assess the feasibility of this guideline being adopted in practice before developing it. A better process may have been to first assess whether such a guideline change would be accepted. If not, more resources could have been put into figuring out how to engage with the medical community to promote uptake of the guideline.

This case study also points towards a larger issue. It was noted in Chapter 1 (section 1.3.3) that the risk thresholds are different in every country, yet every country has access to the same evidence. This indicates that the uncertainty in the allocation of treatment may not lie in the evidence itself, but in how this evidence is conveyed and how it is interpreted. Maybe less quantitative work is needed in this area, with a shift to qualitative work, to better understand why this evidence is being interpreted in different ways, and how to bring all parties together under a common strategy.

For example, the main arguments against the efficacy of statins in the 10 – 20% risk group[86] come from the choice of end point used to elicit findings, the way in which adverse events were handled, the funding source of the trials that make up the main body of evidence, and

the failure to make the patient level data available for re-analysis. All these factors, in my opinion, could be easily addressed if data was made available for re-analysis, and the two opposing parties[83,86] came together in discussion. I do not believe the validity of the trials themselves is under question, the disagreement stems from how the evidence is interpreted. It is not surprising there is discordance in the guidelines around the world, and variable implementation of guidelines themselves, given that leading academics and clinicians do not agree over the interpretation of the same evidence.

One final comment of interest from this chapter is that it provides evidence to support the lack of impact of statin use on the secular trend found in Chapter 5. The study by Finnikin et al.,[97] showed that between 2012 to 2015, only 35% of patients who had a QRISK score recorded > 20% were subsequently initiated on statins. This is quite a small proportion of those who were classified as high risk. If this proportion was also fairly constant over time, we would not expect the reduction in cardiovascular incidence over time to have been caused by this.

8.5.1 Future work

The need for further qualitative work to understand why the guideline change has had a small impact on clinical practice was discussed in detail in Chapter 7. While not technically future research, another area of development in the future could be structural changes which allow NICE to have direct access to data to enable them to carry out studies such as these. The guideline change was implemented in July 2014, and in the six years since then this is the second study to assess the impact of this in practice. I am not aware of NICE having direct access to routinely collected datasets in order to assess the uptake of their guidelines, instead the onus is placed on independent research groups to allocate their resources to do this. The impact this has on the speed at which NICE will get access to results is significant, considering the need to: A) identify the gap in the literature, B) apply for access to the data, C) undertake the study, and D) get the study published. If NICE had access to the data required to assess the impact of their guidelines in practice, once the data streams were set up, they could be continuously monitored for a relatively small amount of resources. A significant change in data sharing processes would be required to make this a possibility, but waiting 6 years to find out a major guideline change has had minimal clinical impact is far from ideal.

8.6 Overall limitations

Limitations of each specific project have been discussed within each chapter. I discuss overarching limitations of the thesis here, which I believe revolve around the use of CPRD data throughout this project.

The first is that in recent years there has been a drop in practices using the VISION computer system, which are the practices that contribute to the dataset I used (see Table 8.1). Aside from the loss in power, this would be particularly problematic if the drop out was not at random. This would cause a selection bias when assessing key metrics over time. Chapters 3 – 6 all use data from 1998 – 2015, and the level of drop out by this point is not too bad. In the next three years there is quite a significant drop out, so it is possible chapter 7 was affected the most by this. We made sure to verify models and trends in the data in comparable studies carried out in different databases where possible. For example, comparison of our model with QRISK3 in Chapter 3; comparison of the causal effect of statin initiation with trial data in Chapter 5; comparison of discontinuation rates with published data in Chapter 6; and reproducing a published analysis in Chapter 7 as part of the main analysis. All results provided good agreement indicating the drop out from practices was unlikely to be biasing our results in any way.

Table 8.1: *Number of English practices actively registered in CPRD in each calendar year, stratified by region*

Year	NE	NW	YORK	EM	WM	E	SW	SC	L	SEC	Total (N)
1998	7	41	20	17	31	22	28	15	22	24	227
1999	9	46	23	19	34	30	31	21	26	25	264
2000	9	53	24	21	37	35	39	32	36	33	319
2001	10	65	24	21	44	38	43	39	43	37	364
2002	9	72	25	23	48	43	46	45	51	43	405
2003	10	76	26	23	49	46	49	46	54	51	430
2004	10	76	24	23	53	48	51	48	56	55	444
2005	10	78	25	22	55	49	53	48	60	57	457
2006	10	79	25	22	55	49	54	48	64	59	465
2007	10	78	23	21	56	47	55	48	65	60	463
2008	11	78	21	20	56	45	55	51	67	60	464
2009	11	79	17	18	56	43	55	51	66	61	457
2010	10	77	15	15	54	41	55	51	68	61	447
2011	10	75	14	11	53	37	51	51	67	60	429
2012	9	71	9	9	52	35	49	51	70	59	414
2013	8	69	6	5	49	31	48	50	75	57	398
2014	6	65	4	1	44	25	40	47	72	55	359
2015	4	50	4	0	36	21	30	43	53	54	295
2016	3	32	3	0	24	10	19	28	38	49	206
2017	3	24	2	0	20	9	13	13	36	35	155
2018	0	20	1	0	15	6	7	9	33	28	119

Another issue was the high levels of missing data in the primary prevention datasets derived from CPRD (see Table 2.14). Particularly Ethnicity (57.95% and 61.79% for female and male), Cholesterol/HDL ratio (61.52% and 64.29%) and SBP variability (49.61% and 79.06%). Even for variables with lower levels of missing data, SBP (18.99% and 40.78%), Smoking (24.82% and 34.83%) and BMI (31.17% and 46.38%), the degree was still fairly substantial. The process of imputation will potentially introduce bias into estimation of predictor effects (and resulting risk predictions) if the untestable assumption of missing at random does not hold. However given the nature of the work, this was not always a problem.

In Chapters 3 and 4, where the interest was in the variation in individual risk scores when following standard processes for developing models, this was not an issue, rather than part of the process to be captured. Nor was it a problem in Chapter 5 when quantifying the potential miscalibration of models used in practice. It could have had an effect on the precision and accuracy of the causal effect estimate of statin initiation, but this estimate appeared reasonable. In Chapter 6 when trying to accurately model risk profiles (transition probabilities) over time through Cox models, this was more of an issue. The implications of the imputation process were not clear as only the trajectory of the baseline hazard was required, which was then manually adjusted in order to obtain given risk profiles. However by design of the study, the absolute risk over the first 10 years was always correct, all that could be wrong was how that risk was distributed (i.e. the shape of the trajectory). Despite the missing data, I felt modelling the risk trajectory based on the routinely collected data was better than assuming a constant relative rate increase each year, which is what is done in the NICE simulations (see section L.2.3.1.3[84]). Missing data was most concerning in Chapter 7 where the aim was to accurately calculate risks for patients initiated on statins. Bias in the imputed values may have resulted in consistently over or under predicted risks of patients each year. However, the levels of missing data were much lower here as it is was a cohort of statin users (Cholesterol/HDL ratio [17.56% and 16.91%], SBP [1.60% and 2.26%], SBP variability [6.26% and 9.77%], Smoking [9.71% and 8.50%] and BMI [18.44% and 20.65%]), which minimised the potential bias from the imputation procedure.

I believe multiple imputation was the best approach to handle the missing data when possible (Chapter 5 required a custom imputation process). Multiple imputation has been shown to be preferable to using complete case datasets even when data is missing not at random,[212] and also when the percentage of missing data is high.[212,213] Given the high number of variables included in the imputation models, the missing at random assumption may not be strongly violated. Combined with the fact the datasets used were very large resulting in highly powered imputation models, I am confident in the imputation process. Checks of the imputation process in Chapters 3 and 7 (see Appendices A.3.1 and A.7.1) show that the distribution of imputed variables were sensible, and there was strong mixing in the Markov chain by the time the algorithm was halted. No new datasets were required to be imputed

for Chapters 4 – 6 (using imputed datasets from Chapter 3 where necessary), so imputation results were not presented for these.

One final point to discuss about the imputation was that in Chapters 4 – 6, just one imputed dataset was used. For Chapters 4 and 6, the same process as in multiple imputation was used, but only one dataset was imputed. Specifically, one of the 20 imputed datasets from Chapter 3 was used. This choice was made because averaging the analyses across all 20 imputed datasets was not necessary to answer the aims in these chapters, and would have complicated the analyses. In Chapter 4 the quantity of interest was the uncertainty induced by sampling variation, so it made sense to re-sample one complete dataset representing the population. The uncertainty associated with those imputed values was not of interest. In Chapter 6, an imputed dataset was used to develop the risk trajectories for individuals, and risk scores were not calculated (i.e. just the baseline hazard was used). Given the size of the dataset, models developed on different datasets from a multiple imputation process should have similar baseline hazards. The majority of the variation in risk scores for individuals comes from predicting point values for their predictors. Therefore it was not necessary to average the baseline hazard across multiple imputed datasets. For Chapter 5, a custom imputation procedure was applied to maintain relationships for variables across different time points. This mostly involved last observation carried forwards, and next observation carried backwards imputation. The random element of multiple imputation, necessitating the creation of multiple imputed datasets, was therefore not relevant.

The final limitation I would like to discuss is the impact of using an EHR as a data source on the generalisability of the results, as most CVD risk prediction models are developed using data from trials or cohort studies in which data is collected prospectively, UK models being the exception (see Table 1.1). This is specific to Chapters 3 and 4, which evaluate the impact of modelling decisions on the generation of risk scores. Chapter 3 considers the reference class problem, which by nature is not unique to models developed on EHRs. I would expect to find similar results in models developed on data from prospective cohort studies or trials, but this may not be the case. Chapter 4 looked at the impact of sample size on the precision of risk predictions. I think the generalisability of these results to models developed in prospective cohort studies is strong. There is no reason why the precision would be different

for a model developed on data from such studies, unless there are much stronger associations between the predictors and outcomes.

8.7 Main conclusions

This thesis explored areas of uncertainty associated with using risk prediction models to drive clinical decision making for individuals, with a focus on CVD risk prediction.

The results indicated that the effects of the reference class problem on CVD risk prediction were small, and the reported discordance in risk scores was driven primarily by sampling variation and using different populations for model development. Clinicians using models built on large sample sizes (> 100 000) and including major CVD risk factors can be confident in the stability of risks generated. However, risk scores from many models used in practice will suffer from high levels of instability induced by their sample size, and this extends to all disease areas. A large drop in the incidence of CVD over time in England was also identified suggesting the risks of individuals in the present day may be overestimated.

Results also indicated extra CVD events could be prevented by initiating at higher risks than 10% for certain subgroups of the population. However, there are ethical implications to consider as realistically this must be based on population level discontinuation rates, rather than how long an individual would adhere for. It was also shown that the introduction of CG181 (risk threshold for initiating statins reduced from 20% to 10%) has only had a small impact on clinical practice in England, contrary to current evidence.

9 Reference list

1. Pate A, Emsley R, Ashcroft DM, et al. 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: 134.

2. Pate A, Elliott RA, Thompson A, et al. The impact of statin discontinuation and restarting rates on the optimal time to initiate statins and on the number of cardiovascular events prevented. *Pharmacoepidemiol Drug Saf* 2020; 29: 644–652.

3. Knaus WA, Wagner DP, Draper EA, et al. The APACHE III prognostic system: Risk prediction of hospital mortality for critically III hospitalized adults. *Chest* 1991; 100: 1619–1636.

4. Jentzer JC, Bennett C, Wiley BM, et al. Predictive value of the Sequential Organ Failure Assessment score for mortality in a contemporary cardiac intensive care unit population. *J Am Heart Assoc* 2018; 7: e008169.

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: j2099.

6. 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: 743–753.

7. Lim WS, Van Der Eerden MM, Laing R, et al. Defining community acquired pneumonia severity on presentation to hospital: An international derivation and validation study. *Thorax* 2003; 58: 377–382.

8. McAllister KSL, Ludman PF, Hulme W, et al. A contemporary risk model for predicting 30-day mortality following percutaneous coronary intervention in England and Wales. *Int J Cardiol* 2016; 210: 125–132.

9. Kanis JA, Johnell O, Oden A, et al. FRAX™ and the assessment of fracture probability in men and women from the UK. *Osteoporos Int* 2008; 19: 385–397.

10. Caldas C, Greenberg DC, Kearins O, et al. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Res* 2010; 12: R1.

11. Nashef SAM, Roques F, Sharples LD, et al. Euroscore II. *Eur J Cardio-thoracic Surg* 2012; 41: 734–745.

12. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Ann Intern Med* 2015; 162: W1–W73.

13. Lemeshow S, Klar J TD. Outcome prediction for individual intensive care patients: useful, misused, or abused? *Intensive Care Med* 1995; 21: 770–776.

14. Steyerberg EW, Eijkemans MJC, Boersma E, et al. Equally valid models gave divergent predictions for mortality in acute myocardial infarction patients in a comparison of

logical regression models. *J Clin Epidemiol* 2005; 58: 383–390.

15. Hájek A. The reference class problem is your problem too. *Synthese* 2007; 156: 563–585.

16. Venn J. *The Logic of Chance*. London: Macmillan, 1866.

17. Stern RH. The Discordance of Individual Risk Estimates and the Reference Class Problem. *arXiv*, https://arxiv.org/abs/1001.2499 (2010).

18. Stern RH. Individual Risk. *J Clin Hypertens* 2012; 14: 261–264.

19. Kent DM, Shah ND. Risk Models and Patient-Centered Evidence: Should Physicians Expect One Right Answer? *JAMA* 2012; 307: 1585–1586.

20. Lagerweij GR, Moons KGM, De Wit GA, et al. Interpretation of CVD risk predictions in clinical practice: Mission impossible? *PLoS One* 2019; 14: e0209314.

21. Allen LA, Matlock DD, Shetterly SM, et al. Use of risk models to predict death in the next year among individual ambulatory patients with heart failure. *JAMA Cardiol* 2017; 2: 435–441.

22. Saver BG, Hargraves JL, Mazor KM. Are population-based diabetes models useful for individual risk estimation? *J Am Board Fam Med* 2011; 24: 399–406.

23. Van Staa TP, Gulliford M, Ng ESW, et al. Prediction of cardiovascular risk using framingham, ASSIGN and QRISK2: How well do they predict individual rather than population risk? *PLoS One* 2014; 9: e106455.

24. Kent DM, Steyerberg E, Van Klaveren D. Personalized evidence based medicine: Predictive approaches to heterogeneous treatment effects. *BMJ* 2018; 363: k4245.

25. Dahabreh IJ, Hayward R, Kent DM. Using group data to treat individuals: Understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. *Int J Epidemiol* 2016; 45: 2184–2193.

26. Kent DM, Van Klaveren D, Paulus JK, et al. The Predictive Approaches to Treatment effect Heterogeneity (PATH) statement: Explanation and elaboration. *Ann Intern Med* 2020; 172: W1–W25.

27. Schulz KF, Altman DG, Moher D, et al. CONSORT 2010 statement: Updated guidelines for reporting parallel group randomized trials. *BMJ* 2010; 4: c332.

28. Elm E von, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting. *Ann Intern Med* 2015; 147: 573–578.

29. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med* 2009; 6: e1000097.

30. Efron B, Tibshirani RJ. *An introduction to the bootstrap*. New York: Chapman and Hall, 1994.

31. QResearch, http://www.qresearch.org/ (accessed 3 May 2018).

32. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* 2016; 353: i2416.

33. Tangri N, Kitsios GD, Inker LA, et al. Risk Prediction Models for Patients With Chronic Kidney Disease. *Ann Intern Med* 2014; 158: 596–603.

34. Abbasi A, Peelen LM, Corpeleijn E, et al. Prediction models for risk of developing type 2 diabetes: Systematic literature search and independent external validation study. *BMJ* 2012; 345: 1–16.

35. Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med* 2019; 38: 1276–1296.

36. Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: Part I – Continuous outcomes. *Stat Med* 2019; 38: 1262–1275.

37. Quanstrum KH, Hayward RA. Lessons from the Mammography Wars. *N Engl J Med* 2011; 362: 567–571.

38. Herrett E, Gadd S, Jackson R, et al. Eligibility and subsequent burden of cardiovascular disease of four strategies for blood pressure-lowering treatment: a retrospective cohort study. *Lancet* 2019; 394: 663–671.

39. Puhan MA, Singh S, Weiss CO, et al. A framework for organizing and selecting quantitative approaches for benefit-harm assessment. *BMC Med Res Methodol* 2012; 12: 173.

40. Jamison DT, Breman JG, Measham AR, et al. Cost-Effectiveness Analysis. In: *Priorities in Health*. Washington (DC): The International Bank for Reconstruction and Development/The World Bank, 2006.

41. Schroll JB, Penninga EI, Gøtzsche PC. Assessment of Adverse Events in Protocols, Clinical Study Reports, and Published Papers of Trials of Orlistat: A Document Analysis. *PLoS Med* 2016; 13: 1–14.

42. Phillips R, Hazell L, Sauzet O, et al. Analysis and reporting of adverse events in randomised controlled trials: a review. *BMJ Open* 2019; 9: e024537.

43. Allen EN, Chandler CIR, Mandimika N, et al. Eliciting adverse effects data from participants in clinical trials. *Cochrane Database Syst Rev* 2018; 16: MR000039.

44. Zorzela L, Golder S, Liu Y, et al. Quality of reporting in systematic reviews of adverse events: Systematic review. *BMJ* 2014; 348: 1–12.

45. NICE. *CG181 Lipid modification - Cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease*, https://www.nice.org.uk/guidance/cg181/evidence/lipid-modification-update-full-guideline-pdf-243786637 (2014).

46. NICE. *Cardiovascular disease: risk assessment and reduction, including lipid modification*, https://www.nice.org.uk/guidance/cg181/chapter/1-recommendations (2014, accessed 3 May 2018).

47. Wise J. Open letter raises concerns about NICE guidance on statins. *BMJ* 2014; 348: 10–11.

48. Herrett E, Gallagher AM, Bhaskaran K, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015; 44: 827–836.

49. NHS Digital. Hospital Episode Statistics, https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics (accessed 3 May 2018).

50. Office for National Statistics, https://www.ons.gov.uk/ (accessed 3 May 2018).

51. National Health Service. Cardiovascular disease, https://www.nhs.uk/conditions/cardiovascular-disease/.

52. British Heart Foundation. *BHF Statistics Factsheet - UK*, https://www.bhf.org.uk/what-we-do/our-research/heart-statistics (2020).

53. Bhatnagar P, Wickramasinghe K, Williams J, et al. Trends in the epidemiology of cardiovascular disease in the UK. *Heart* 2016; 102: 1945–1952.

54. Smolina K, Wright FL, Rayner M, et al. Determinants of the decline in mortality from acute myocardial infarction in England between 2002 and 2010: linked national database study. *BMJ* 2012; 344: d8059–d8059.

55. Lee S, Shafe ACE, Cowie MR. UK stroke incidence, mortality and cardiovascular risk management 1999-2008: time-trend analysis from the General Practice Research Database. *BMJ Open* 2011; 1: e000269–e000269.

56. Rothwell PM, Coull AJ, Giles MF, et al. Change in stroke incidence, mortality, case-fatality, severity, and risk factors in Oxfordshire, UK from 1981 to 2004 (Oxford Vascular Study). *Lancet* 2004; 363: 1925–1933.

57. HEART UK. *Prioritising the prevention of cardiovascular disease (CVD)*, https://www.heartuk.org.uk/downloads/health-professionals/heart-uk-cvd-prevention-policy-paper---july-2019.pdf (2019).

58. Truett J, Cornfield J, Kannel W. A multivariate analysis of the risk of coronary heart disease in Framingham. *J Chronic Dis* 1967; 20: 511–524.

59. Dawber TR, Kannel WB, Lye LP. An approach to longitudinal studies in a community: The Framingham study. *Ann New York Acad Sci* 1963; 539–556.

60. Goff DC, Lloyd-jones DM, Bennett G, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk. *Circulation* 2014; 129: S49-73.

61. 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: e596–e646.

62. NICE. CVD risk assessment and management, https://cks.nice.org.uk/cvd-risk-assessment-and-management#!scenario (2019, accessed 19 July 2019).

63. 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: 172–176.

64. Scottish Intercollegiate Guidelines Network (SIGN). *SIGN 149 - Risk estimation and the prevention of cardiovascular disease*, http://www.sign.ac.uk (2017).

65. Conroy RM, Pyörälä K, Fitzgerald AP, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: The SCORE project. *Eur Heart J* 2003; 24: 987–1003.

66. Catapano AL, Graham I, De Backer G, et al. 2016 ESC/EAS Guidelines for the Management of Dyslipidaemias. *Eur Heart J* 2016; 37: 2999–3058l.

67. Mach F, Baigent C, Catapano AL, et al. 2019 ESC/EAS Guidelines for the management of dyslipidaemias: lipid modification to reduce cardiovascular risk. *Eur Heart J* 2020; 41: 111–188.

68. 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: 1897–1907.

69. Ministry of Health NZ. *What's new in cardiovascular disease risk assessment and management for primary care clinicians*, https://bpac.org.nz/2018/cvd.aspx (2018).

70. Hajifathalian K, Ueda P, Lu Y, et al. A novel risk score to predict cardiovascular disease risk in national populations (Globorisk): A pooled analysis of prospective cohorts and health examination surveys. *Lancet Diabetes Endocrinol* 2015; 3: 339–355.

71. Cooney MT, Dudina AL, Graham IM. Value and Limitations of Existing Scores for the Assessment of Cardiovascular Risk. A Review for Clinicians. *J Am Coll Cardiol* 2009; 54: 1209–1227.

72. Ridker PM, Buring JE, Rifai N, et al. Algorithms for the Assessment of Global Cardiovascular Risk in Women. *JAMA* 2007; 297: 611–620.

73. Ridker PM, Cook NR. Statins: New American guidelines for prevention of cardiovascular disease. *Lancet* 2013; 382: 1762–1765.

74. Assmann G, Cullen P, Schulte H. Simple Scoring Scheme for Calculating the Risk of Acute Coronary Events Based on the 10-Year Follow-Up of the Prospective Cardiovascular Münster (PROCAM) Study. *Circulation* 2002; 105: 310–315.

75. Cox DR. Regression Models and Life-Tables. *J R Stat Soc Ser B Stat Methodol* 1972; 34: 187–220.

76. Wolbers M, Koller MT, Witteman JCM, et al. Prognostic Models With Competing Risks. *Epidemiology* 2009; 20: 555–561.

77. Austin PC, Lee DS, Fine JP. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. *Circulation* 2016; 133: 601–609.

78. Abdel-Qadir H, Fang J, Lee DS, et al. Importance of considering competing risks in time-to-event analyses: Application to stroke risk in a retrospective cohort study of elderly patients with atrial fibrillation. *Circ Cardiovasc Qual Outcomes* 2018; 11: 1–11.

79. Stone NJ, Robinson JG, Lichtenstei AH, et al. 2013 ACC/AHA Guideline on the Treatment of Blood Cholesterol to Reduce Atherosclerotic Cardiovascular Risk in Adults. *Circulation* 2014; 129: S1–S45.

80. Grundy SM, Stone NJ, Bailey AL, et al. 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *J Am Coll Cardiol* 2019; 73: e285–e350.

81. Taylor F, Huffman M, Macedo A, et al. Statins for the primary prevention of cardiovascular disease. *Cochrane database Syst Rev* 2013; 31: CD004816.

82. Chou R, Dana T, Blazina I, et al. Statins for prevention of cardiovascular disease in adults: Evidence report and systematic review for the US preventive services task force. *JAMA* 2016; 316: 2008–2024.

83. Collins R, Reith C, Emberson J, et al. Interpretation of the evidence for the efficacy and safety of statin therapy. *Lancet* 2016; 388: 2532–2561.

84. NICE. *CG181 Lipid modification Appendicies - Cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease*, https://www.nice.org.uk/guidance/cg181/evidence/lipid-modification-update-appendices-pdf-243786638 (2014).

85. Pandya A, Sy S, Cho S, et al. Cost-effectiveness of 10-year risk thresholds for initiation of statin therapy for primary prevention of cardiovascular disease. *JAMA* 2015; 314: 142–150.

86. Abramson JD, Rosenberg HG, Jewell N. Should people at low risk of cardiovascular disease take a statin? *BMJ* 2013; 347: f6123.

87. Abramson J, Rosenberg HG, Jewell N, et al. Safety and efficacy of statins. *Lancet* 2017; 389: 1097.

88. Cholesterol Treatment Trialists. 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: 581–590.

89. Taylor F, Ward K, Moore T, et al. Statins for the primary prevention of cardiovascular disease. *Cochrane database Syst Rev* 2011; 19: CD004816.

90. Redberg R, Katz M. Statins for Primary Prevention. The Debate Is Intense, but the Data Are Weak. *JAMA* 2016; 316: 1979–1981.

91. Bibbins-Domingo K, Grossman DC, Curry SJ, et al. Statin use for the primary prevention of cardiovascular disease in adults: US preventive services task force recommendation statement. *JAMA* 2016; 316: 1997–2007.

92. Yebyo HG, Aschmann HE, Puhan MA. Finding the Balance between Benefits and Harms When Using Statins for the Primary Prevention of Cardiovascular Disease: A Modeling Study. *Ann Intern Med* 2019; 170: 1–10.

93. New Scientist. Everyday drugs: The great statins debate,

https://www.newscientist.com/article/mg22630213-400-everyday-drugs-the-great-statins-debate/ (2015, accessed 1 April 2019).

94. The Guardian. Statins prevent 80,000 heart attacks and strokes a year in UK, study finds, https://www.theguardian.com/society/2016/sep/08/statins-prevent-80000-heart-attacks-and-strokes-a-year-in-uk-study-finds (2016, accessed 1 April 2019).

95. Guidelines in practice. Low-cost statins offer clear benefits and low levels of risk, https://www.guidelinesinpractice.co.uk/cardiovascular/low-cost-statins-offer-clear-benefits-and-low-levels-of-risk/352523.article (2014, accessed 1 April 2019).

96. BJCardio Staff. New NICE guidance published. *Br J Cardiol* 2014; 21: 98.

97. Finnikin S, Ryan R, Marshall T. Statin initiations and QRISK2 scoring in UK general practice: a THIN database study. *Br J Gen Pract* 2017; 67: e881–e887.

98. Reynolds TM, Twomey PJ, Wierzbicki AS. Concordance evaluation of coronary risk scores: Implications for cardiovascular risk screening. *Curr Med Res Opin* 2004; 20: 811–818.

99. 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: 1475–1482.

100. Collins GS, Altman DG. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. *BMJ* 2010; 340: c2442.

101. Collins GS, Altman DG. Predicting the adverse risk of statin treatment: an independent and external validation of an updated version of QRISK2. *BMJ* 2012; 98: 1091–7.

102. Townsend P, Davidson N. *The Black Report*. London: Penguin, 1982.

103. Pate A. The uncertainty with using risk predictions models to drive clinical decision making. *GitHub respository*, https://github.com/alexpate30/AlexPate-PhD (2020).

104. QRISK3 calculator, https://qrisk.org/three/index.php (accessed 3 May 2018).

105. NHS Digital. Quality and Outcomes Framework QOF business rules v36.0, https://www.digital.nhs.uk/qof/qofbrv36 (accessed 24 January 2018).

106. Primary Care Unit CU. CPRD @ Cambridge - Code Lists, http://www.phpc.cam.ac.uk/pcu/cprd_cam/codelists/ (accessed 24 January 2018).

107. Institute of Population Health U of M. ClinicalCodes.org, https://clinicalcodes.rss.mhs.man.ac.uk/ (accessed 24 January 2018).

108. Pate A, Barrowman M, Webb D, et al. Study investigating the generalisability of a COPD trial based in primary care (Salford Lung Study) and the presence of a Hawthorne effect. *BMJ Open Respir Res* 2018; 5: e000339.

109. Mathur R, Bhaskaran K, Chaturvedi N, et al. Completeness and usability of ethnicity data in UK-based primary care and hospital databases. *J Public Health (Oxf)* 2014; 36:

684–692.

110. Hippisley-Cox J. Authors' response to letters on QRISK3 regarding the definition of severe mental illness and analysis of systolic blood pressure variability. *BMJ*, https://www.bmj.com/content/357/bmj.j2099/rr-2 (2017, accessed 13 March 2020).

111. Levey AS, Stevens LA, Frcp C, et al. A New Equation to Estimate Glomerular Filtration Rate Andrew. *Ann Intern Med* 2009; 150: 604–612.

112. Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney Int Suppl* 2013; 3: 1–150.

113. Matsushita K, Woodward M, Jafar TH, et al. Comparison of Risk Prediction Using the CKD-EPI Equation and the MDRD Study Equation for Estimated Glomerular Filtration Rate. *JAMA* 2012; 307: 1941–1951.

114. Jameson K, Jick S, Hagberg KW, et al. Prevalence and management of chronic kidney disease in primary care patients in the UK. *Int J Clin Pract* 2014; 68: 1110–1121.

115. Hippisley-Cox J, Coupland C. Predicting the risk of Chronic Kidney Disease in men and women in England and Wales: prospective derivation and external validation of the QKidney ® Scores. *BMC Fam Pract* 2010; 11: 49.

116. Broad J, Wells S, Marshall R, et al. Zero end-digit preference in recorded blood pressure and its impact on classification of patients for pharmacologic management in primary care — PREDICT-CVD. *Br J Gen Pract* 2007; 57: 897–903.

117. SAS Institute Inc 2013. SAS/ACCESS® 9.4 Interface to ADABAS: Reference. Cary, NC: SAS Institute Inc.

118. R Core Team. R: A language and environment for statistical computing. R Foundation for statistical computing, Vienna, Austria., http://www.r-project.org/ (2017).

119. Goldstein BA, Navar AM, Pencina MJ, et al. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Informatics Assoc* 2017; 24: 198–208.

120. Steyerberg EW. *Clinical Prediction Models - A Practical Approach to Development, Validation, and Updating*. Springer, 2009.

121. Harrell, Jr 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: 361–387.

122. BBC Health. Statins alert over IT glitch in heart risk tool, http://www.bbc.co.uk/news/health-36274791 (2016, accessed 1 December 2016).

123. Digitalhealth.net. QRisk2 in TPP 'fixed' but up to 270,000 patients affected, http://www.digitalhealth.net/analytics/47759/qrisk2- (2016, accessed 1 December 2016).

124. Hofer E. *The Uncertainty Analysis of Model Results: A practical guide*. Springer International Publishing, 2018.

125. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. 2004; 748: 723–748.

126. Royston P. Explained variation for survival models. *Stata J* 2006; 6: 83–96.

127. Allgulander C. Anxiety as a risk factor in cardiovascular disease. *Curr Opin Psychiatry* 2016; 29: 13–17.

128. British Heart Foundation. *Cardiovascular Disease Statistics 2018*, https://www.bhf.org.uk/research/heart-statistics (2018).

129. NHS. NHS authorities and trusts, https://www.nhs.uk/using-the-nhs/about-the-nhs/nhs-authorities-and-trusts/.

130. van Buuren S, Groothuis-oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *J Stat Softw*; 45, https://www.jstatsoft.org/article/view/v045i03 (2011).

131. Benner A. Multivariable Fractional Polynomials, https://cran.r-project.org/web/packages/mfp/vignettes/mfp_vignette.pdf (accessed 24 July 2018).

132. Marshall A, Altman DG, Holder RL, et al. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Med Res Methodol* 2009; 9: 57.

133. Choodari-Oskooei B, Royston P, Parmar MKB. A simulation study of predictive ability measures in a survival model I: Explained variation measures. *Stat Med* 2012; 31: 2627–2643.

134. Rahman MS, Ambler G, Choodari-Oskooei B, et al. Review and evaluation of performance measures for survival prediction models in external validation settings. *BMC Med Res Methodol* 2017; 17: 60.

135. Austin PC, Pencinca MJ, Steyerberg EW. Predictive accuracy of novel risk factors and markers: A simulation study of the sensitivity of different performance measures for the Cox proportional hazards regression model. *Stat Methods Med Res* 2017; 26: 1053–1077.

136. Choodari-Oskooei B, Royston P, Parmar MKB. A simulation study of predictive ability measures in a survival model II: Explained randomness and predictive accuracy. *Stat Med* 2012; 31: 2644–2659.

137. Hajime Uno, Tianxi Cai, Micae J Pencina, Ralph B D'Agostino LJW. On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data. *Stat Med* 2011; 30: 1105–1117.

138. Gönen M, Heller G. Concordance Probability and Discriminatory Power in Proportional Hazards Regression. *Biometrika* 2005; 92: 965–970.

139. Kent JT, O'Quigley J. Measures of Dependence for Censored Survival Data. *Biometrika* 1988; 75: 525–534.

140. O'Quigley J, Xu R, Stare J. Explained randomness in proportional hazards models. *Stat Med* 2005; 24: 479–489.

141. Graf E, Schmoor C, Sauerbrei W, et al. Assessment and Comparison of Prognostic Classification Schemes for Survival Data. *Stat Med Stat Med* 1999; 18: 2529–2545.

142. Gerds TA, Schumacher M. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical J* 2006; 48: 1029–1040.

143. Harrell, Jr FE. *Regression Modeling Strategies*. Ney York: NY: Springer-Verlag, 2001.

144. ClinRisk. QRISK2 online calculator, Information, What is the difference between an 'estimated' QRISK®2 CVD score and an 'actual' QRISK®2 CVD score?, https://qrisk.org/2017/index.php (accessed 7 December 2018).

145. Office for National Statistics. Estimates of the population for the UK, England and Wales, Scotland and Northern Ireland, https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland (2018).

146. ClinRisk. *QRisk 2016 Annual Update Information*, https://qrisk.org/2017/QRISK2-2016-Annual-Update-Information.pdf (2016).

147. Pajouheshnia R, Damen JAAG, Groenwold RHH, et al. Treatment use in prognostic model research: a systematic review of cardiovascular prognostic studies. *Diagnostic Progn Res* 2017; 1: 15.

148. Peek N, Sperrin M, Mamas M, et al. Hari Seldon, QRISK3, and the Prediction Paradox. *BMJ responses*, https://www.bmj.com/content/357/bmj.j2099/rr-0 (2017).

149. Jenkins DA, Sperrin M, Martin GP, et al. Dynamic models to predict health outcomes: current status and methodological challenges. *Diagnostic Progn Res* 2018; 2: 1–9.

150. Schulze MB, Martínez-González MA, Fung TT, et al. Food based dietary patterns and chronic disease prevention. *BMJ* 2018; k2396.

151. Hooper L, Summerbell CD, Higgins JPT, et al. Dietary fat intake and prevention of cardiovascular disease: systematic review. *BMJ* 2001; 322: 757–763.

152. Brown WJ, Pavey T, Bauman AE. Comparing population attributable risks for heart disease across the adult lifespan in women. *Br J Sports Med* 2015; 49: 1069–1076.

153. Weng SF, Kai J, Guha IN, et al. The value of aspartate aminotransferase and alanine aminotransferase in cardiovascular disease risk assessment. *Open Hear* 2015; 2: e000272.

154. Libby P, Ridker PM, Maseri A. Inflammation and atherosclerosis. *Circulation* 2002; 105: 1135–1143.

155. Berezin AE. Biomarkers for cardiovascular risk in patients with diabetes. *Heart* 2016; 102: 1939–1941.

156. Dhingra R, Vasan R. Trends in Cardiovascular Medicine. *Trends Cardiovasc Med* 2017; 27: 123–133.

157. Liew SM, Blacklock C, Hislop J, et al. Cardiovascular risk scores: Qualitative study of how primary care practitioners understand and use them. *Br J Gen Pract* 2013; 63: 401–407.

158. BBC NEWS. How your heart age is key to heart attack or stroke risk, https://www.bbc.co.uk/news/health-45395576# (accessed 11 September 2018).

159. Petter O. How to check your heart age: three ways to determine the likelihood of a stroke or heart attack, https://www.independent.co.uk/life-style/health-and-families/heart-age-check-health-public-health-england-stroke-a8522706.html (accessed 11 September 2018).

160. Deanfield J, Sattar N, Simpson I, et al. Joint British Societies' consensus recommendations for the prevention of cardiovascular disease (JBS3). *Heart* 2014; 100: ii1-ii67.

161. Hippisley-Cox J, Coupland C, Robson J, et al. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: Cohort study using QResearch database. *BMJ* 2010; 342: 93.

162. Stewart J, Manmathan G, Wilkinson P. Primary prevention of cardiovascular disease: A review of contemporary guidance and literature. *JRSM Cardiovasc Dis* 2017; 6: 204800401668721.

163. Austin PC. A Tutorial on Multilevel Survival Analysis: Methods, Models and Applications. *Int Stat Rev* 2017; 85: 185–203.

164. O'Quigley J, Stare J. Proportional hazards models with frailties and random effects. *Stat Med* 2002; 21: 3219–3233.

165. Van Smeden M, De Groot JAH, Moons KGM, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Med Res Methodol* 2016; 16: 163.

166. Textor J, Zander B Van Der, Gilthorpe MS, et al. Robust causal inference using directed acyclic graphs: the R package ' dagitty '. *Int J Epidemiol* 2016; 45: 1887–1894.

167. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; 11: 550–560.

168. Hernán M, Brumback B, Robins J. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* 2000; 11: 561–70.

169. Sperrin M, Martin GP, Pate A, et al. Using marginal structural models to adjust for treatment drop-in when developing clinical prediction models. *Stat Med* 2018; 37: 4142–4154.

170. Therneau TM. A package for Survival Analysis in S_. version 2.38, https://cran.r-project.org/package=survival (2015).

171. van der Wal WM, Geskus RB. ipw: An R Package for Inverse Probability Weighting. *J Stat Softw* 2011; 43: 13.

172. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol* 2008; 168: 656–664.

173. Benjamin EJ, Blaha MJ, Chiuve SE, et al. Heart Disease and Stroke Statistics—2017 Update. *Circulation* 2017; 135: e146–e603.

174. Institute for Health Metrics and Evaluation. Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2016 (GBD 2016), http://ghdx.healthdata.org/gbd-results-tool (2017, accessed 10 April 2019).

175. Maningat P, Gordon BR, Breslow JL. How do we improve patient compliance and adherence to long-term statin therapy? *Curr Atheroscler Rep* 2013; 15: 1–12.

176. Wei MY, Ito MK, Cohen JD, et al. Predictors of statin adherence, switching, and discontinuation in the USAGE survey: Understanding the use of statins in America and gaps in patient education. *J Clin Lipidol* 2013; 7: 472–483.

177. Vinogradova Y, Coupland C, Brindle P, et al. Discontinuation and restarting in patients on statin treatment: prospective open cohort study using a primary care database. *BMJ* 2016; 353: i3305.

178. Mauskop A, Borden WB. Predictors of statin adherence. *Curr Cardiol Rep* 2011; 13: 553–558.

179. Simoens S, Sinnaeve PR. Patient co-payment and adherence to statins: A review and case studies. *Cardiovasc Drugs Ther* 2014; 28: 99–109.

180. Mann DM, Woodward M, Muntner P, et al. Predictors of non-adherence to statins: A systematic review and meta-analysis. *Ann Pharmacother* 2010; 44: 1410–1421.

181. Ofori-Asenso R, Jakhu A, Curtis AJ, et al. A Systematic Review and Meta-analysis of the Factors Associated with Nonadherence and Discontinuation of Statins among People Aged ≥65 Years. *Journals Gerontol - Ser A Biol Sci Med Sci* 2018; 73: 798–805.

182. Lloyd-Jones DM, Larson MG, Beiser A, et al. Lifetime risk of developing coronary heart disease. *Lancet* 1999; 353: 89–92.

183. Beiser A, D'Agostino RB, Seshadri S, et al. Epidemiology: Computing Estimates of Incidence, Including Lifetime Risk: Alzheimer's Disease in the Framingham Study. The Practical Incidence Estimators (PIE) Macro. *Tutorials Biostat Stat Methods Clin Stud* 2000; 1: 1–30.

184. Harshfield A, Abel GA, Barclay S, et al. Do GPs accurately record date of death? A UK observational analysis. *BMJ Support Palliat Care* 2018; 1–8.

185. van Driel ML, Morledge MD, Ulep R, et al. Interventions to improve adherence to lipid-lowering medication. *Cochrane Collab* 2016; 12: CD004371.

186. Barber N, Parsons J, Clifford S, et al. Patients' problems with new medication for chronic conditions. *Qual Saf Heal Care* 2004; 13: 172–175.

187. Clifford S, Barber N, Elliott R, et al. Patient-centred advice is effective in improving adherence to medicines. *Pharm World Sci* 2006; 28: 165–170.

188. Elliot RA, Boyd MJ, Salema N-E, et al. Supporting adherence for people starting a new medication for a long-term condition through community pharmacies: A pragmatic randomised controlled trial of the New Medicine Service. *BMJ Qual Saf* 2016; 25: 747–758.

189. Department of Health. Pharmacy in England: Building on strengths - delivering the future, https://www.gov.uk/government/publications/pharmacy-in-england-building-on-strengths-delivering-the-future (accessed 22 February 2019).

190. Lyons I, Barber N, Raynor DK, et al. The Medicines Advice Service Evaluation (MASE): A randomised controlled trial of a pharmacist-led telephone based intervention designed to improve medication adherence. *BMJ Qual Saf* 2016; 25: 759–769.

191. Hope HF, Binkley GM, Fenton S, et al. Systematic review of the predictors of statin adherence for the primary prevention of cardiovascular disease. *PLoS One* 2019; 14: e0201196.

192. Hawkes N. NICE guidelines could put 12 million UK adults on statins. *BMJ* 2017; 358: j3674.

193. NICE. Wider use of statins could cut deaths form heart disease, https://www.nice.org.uk/news/article/wider-use-of-statins-could-cut-deaths-from-heart-disease (2014, accessed 15 January 2020).

194. NICE. NICEimpact cardiovascular disease prevention, https://www.nice.org.uk/media/default/about/what-we-do/into-practice/measuring-uptake/nice-impact-cardiovascular-disease-prevention.pdf (2018).

195. Li Y, Sperrin M, van Staa T. R package 'QRISK3': an unofficial research purposed implementation of ClinRisk's QRISK3 algorithm into R. *F1000Research* 2019; 8: 2139.

196. Hippisley-Cox J, Coupland C, Vinogradova Y, et al. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: Prospective open cohort study. *Br Med J* 2007; 335: 136–141.

197. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med* 2016; 15: 155–163.

198. The Health Improvement Network, https://www.ucl.ac.uk/pcph/research-groups-themes/thin-pub/database (accessed 3 May 2018).

199. Byrne P, Donovan ÓO, Smith SM, et al. Medicalisation, risk and the use of statins for primary prevention of cardiovascular disease: a scoping review of the literature. *Health Risk Soc* 2019; 21: 390–406.

200. Ju A, Hanson CS, Banks E, et al. Patient beliefs and attitudes to taking statins: Systematic review of qualitative studies. *Br J Gen Pract* 2018; 68: e408–er419.

201. Hassan Y, Ford J, Steel N, et al. Debate & analysis why are statin prescribing guidelines for primary prevention not always followed in primary care? *Br J Gen Pract* 2016; 66: 536–537.

202. Wu J, Zhu S, Yao GL, et al. Patient Factors Influencing the Prescribing of Lipid Lowering Drugs for Primary Prevention of Cardiovascular Disease in UK General

Practice: A National Retrospective Cohort Study. *PLoS One* 2013; 8: e67611.

203. Kedward J, Dakin L. A qualitative study of barriers to the use of statins and the implementation of coronary heart disease prevention in primary care. *Br J Gen Pract* 2003; 53: 684–689.

204. Fairhurst K, Huby G. From trial data to practical knowledge: Qualitative study of how general practitioners have accessed and used evidence about statin drugs in their management of hypercholesterolaemia. *BMJ* 1998; 317: 1130–1134.

205. Gale NK, Greenfield S, Gill P, et al. Patient and general practitioner attitudes to taking medication to prevent cardiovascular disease after receiving detailed information on risks and benefits of treatment: A qualitative study. *BMC Fam Pract* 2011; 12: 59.

206. Krüger K, Leppkes N, Gehrke-Beck S, et al. Improving long-term adherence to statin therapy: A qualitative study of GPS' experiences in primary care. *Br J Gen Pract* 2018; 68: e401–e407.

207. Tanner RM, Safford MM, Monda KL, et al. Primary Care Physician Perspectives on Barriers to Statin Treatment. *Cardiovasc Drugs Ther* 2017; 31: 303–309.

208. Elisabeth A, Denig P, van Vliet T, et al. Reasons of general practitioners for not prescribing lipid-lowering medication to patients with diabetes: A qualitative study. *BMC Fam Pract* 2009; 10: 24.

209. Wald NJ, Law MR. A strategy to reduce cardiovascular disease by more than 80%. *BMJ* 2003; 326: 1419–1423.

210. Wald NJ, Simmonds M, Morris JK. Screening for future cardiovascular disease using age alone compared with multiple risk factors and age. *PLoS One* 2011; 6: e18742.

211. Leening MJG, Cook NR, Ridker PM. Should we reconsider the role of age in treatment allocation for primary prevention of cardiovascular disease? *Eur Heart J* 2017; 38: 1542–1547.

212. Kontopantelis E, White IR, Sperrin M, et al. Outcome-sensitive multiple imputation: a simulation study. *BMC Med Res Methodol* 2017; 17: 2.

213. Janssen KJM, Donders ART, Harrell FE, et al. Missing covariate data in medical research: To impute is better than to ignore. *J Clin Epidemiol* 2010; 63: 721–727.

214. Van Houwelingen J, Le Cessie S. Predictive value of statistical models. *Stat Med* 1990; 9: 1303–1325.

215. Nagelkerke NJD. A note on a general definition of the coefficient of determination. *Biometrika* 1991; 78: 691–692.

216. Mittlboeck M, Heinzl H. Pseudo R-squared measures of generalized linear models. In: *Proceedings of the 1st European Workshop on the Assessment of Diagnostic Performance*. Milan, Italy, 2004.

217. Cox DR, Snell EJ. *Analysis of Binary Data*. 2 edition. Chapman and Hall/CRC, 1989.

218. Magee L. R2 measures based on wald and likelihood ratio joint significance tests. *Am*

Stat 1990; 44: 250–253.

219. Daskalopoulou SS, Delaney JAC, Filion KB, et al. Discontinuation of statin therapy following an acute myocardial infarction: A population-based study. *Eur Heart J* 2008; 29: 2083–2091.

220. Length of Prescriptions - Recommendations (Herts Valley Clinical Commissioning Group), https://hertsvalleysccg.nhs.uk/application/files/1915/3633/0676/HVCCG_Recommendations_on_Length_of_Repeat_Prescriptions.pdf.

A.3 Chapter 3 Appendices

A.3.1 Density and convergence plots for imputed variables

A.3.1.1 Methods

Methods for running the imputation

The same imputation process was carried out for female and male cohorts separately. Multiple imputation by chained equations was used to impute missing data for body mass index (BMI), systolic blood pressure (SBP) and SBP variability, cholesterol, high-density lipoprotein (HDL), smoking status and ethnicity. The program used to impute the data was the R package 'mice'.[130] There were 20 imputation procedures carried out, with 20 iterations for each one. Variables included in the imputation model were all predictor variables from the final model (including interaction terms and fractional polynomials), Nelson Aalen estimate of the cumulative hazard at the time of event/censored and the censoring indicator. All continuous variables were imputed using predictive mean matching, and polytomous regression for categorical variables[130]. Interactions terms were imputed empirically from the two component variables (i.e. not stochastically), and interactions terms were not used to impute their component variables.

Methods for assessment performance of imputation process

For continuous data the density plots shown assess whether there were any systematic differences in for the non-missing data and the imputed data. This also enabled us to check that the distribution of imputed values was reasonable (i.e. no extreme values, or a distribution shape which clearly indicates an issue with the imputation procedure). In the plots, each red line is a density plot of the imputed data in one of the imputed datasets, and the blue line is the density plot of the non-missing data.

The convergence plots assess whether the Markov chain in the imputation process had reached a steady state by the final iteration. The x-axis is the iteration number, y-axis the mean or standard deviation of the imputed values, and each coloured line a different

imputation process. For categorical variables, the distribution of the variable from each imputation stream are presented, as well as the distribution of non-missing values.

A.3.1.2 Results for imputation of female cohort

All convergence plots reached a steady state very quickly, well before the 20th. All density plots had reasonable distributions with no extreme values. All plots presented below.

BMI

Figure A.3.1: *BMI convergence plot for imputation of female cohort*

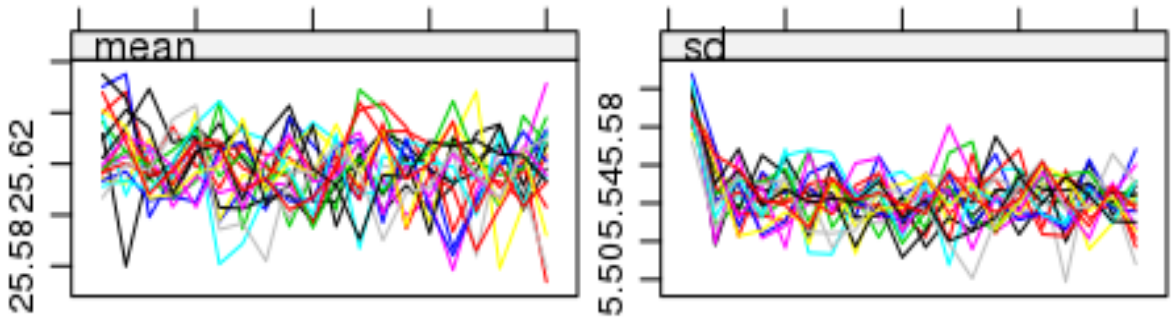
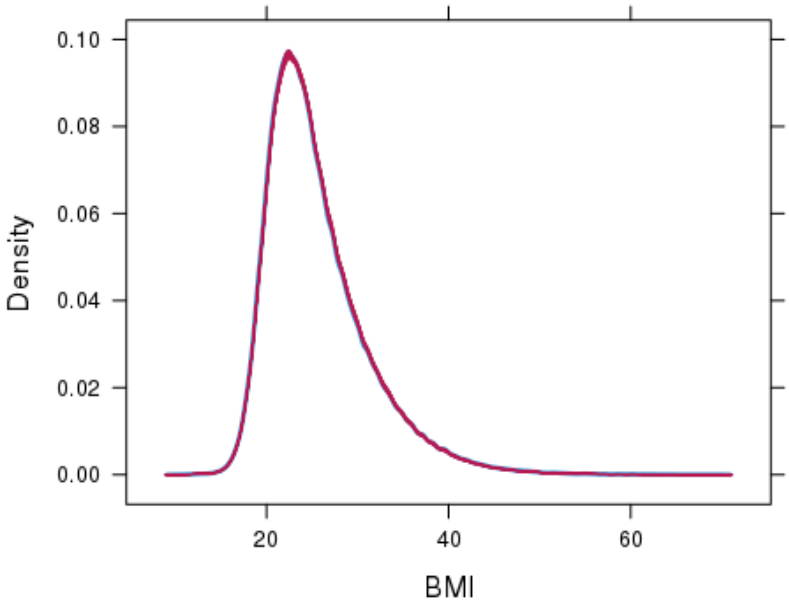


Figure A.3.2: *BMI density plot for imputation of female cohort*



SBP

Figure A.3.3: *SBP convergence plot for imputation of female cohort*

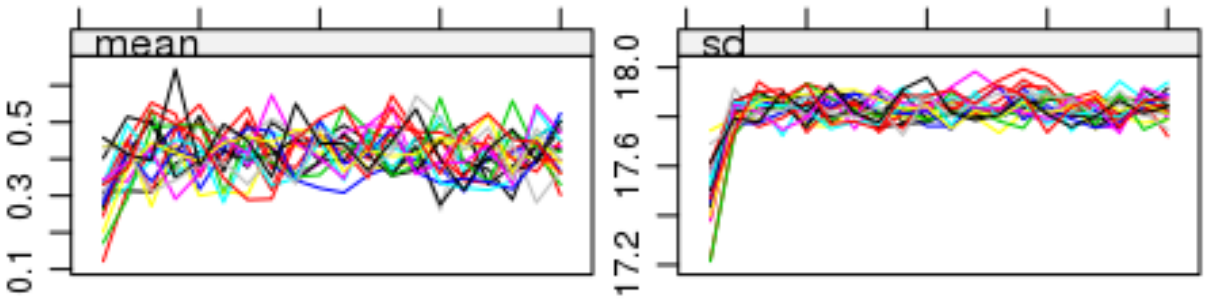
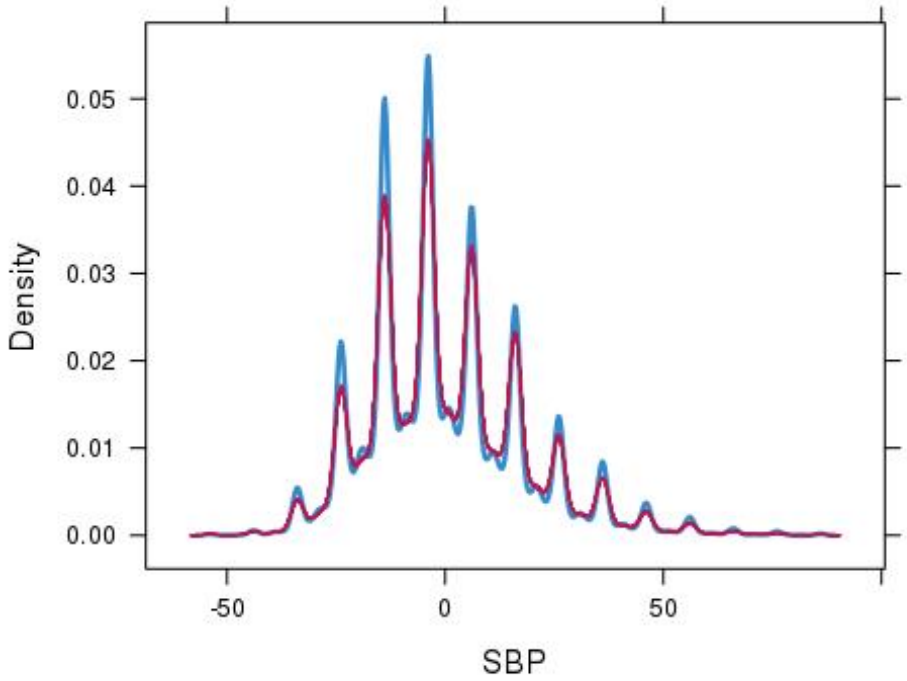


Figure A.3.4: *SBP density plot for imputation of female cohort*



SBP variability

Figure A.3.5: *SBP variability convergence plot for imputation of female cohort*

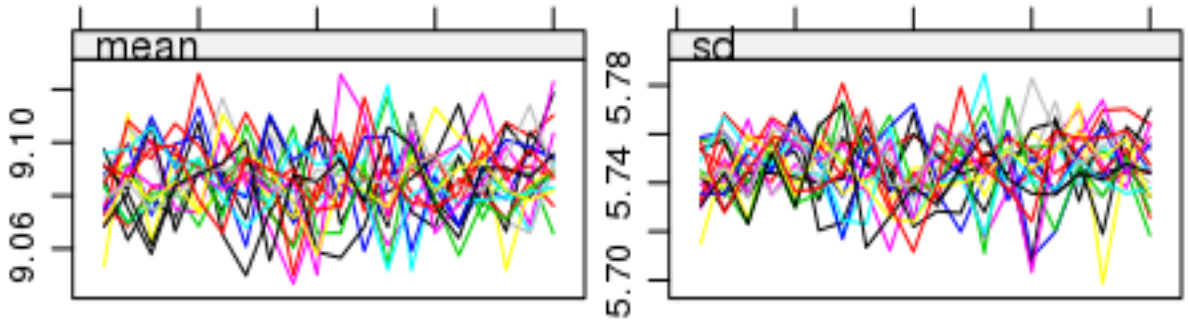
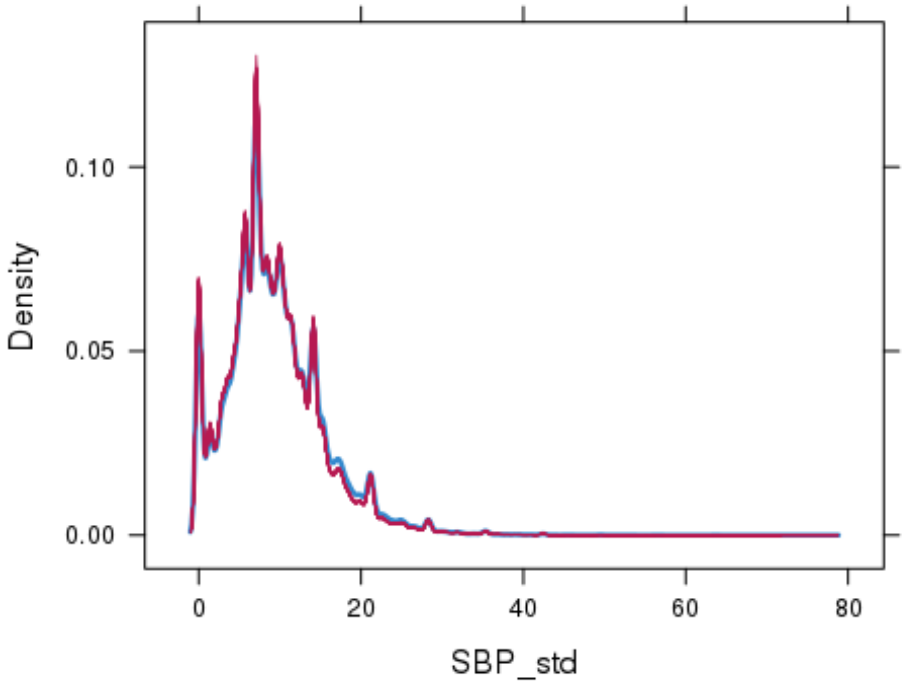


Figure A.3.6: *SBP variability density plot for imputation of female cohort*



Cholesterol

Figure A.3.7: *Cholesterol convergence plot for imputation of female cohort*

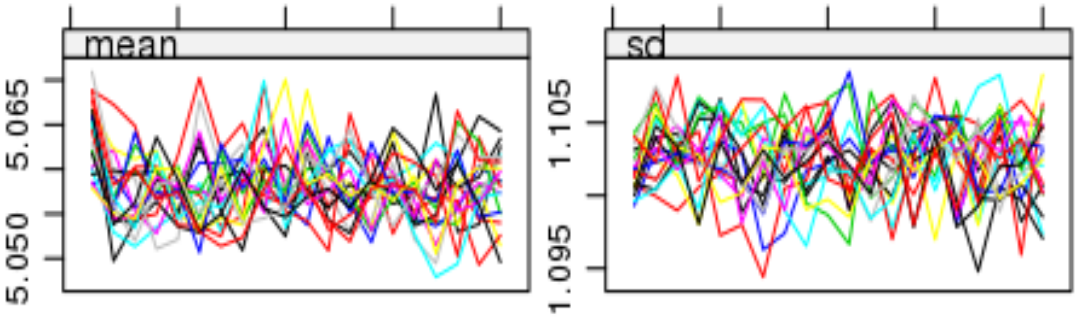
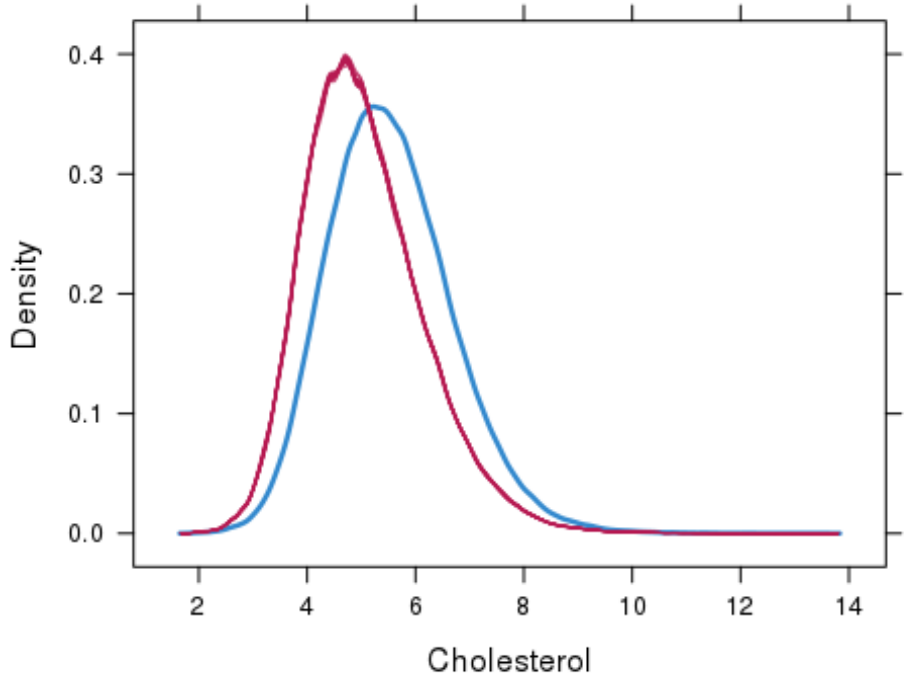


Figure A.3.8: *Cholesterol density plot for imputation of female cohort*



HDL

Figure A.3.9: *HDL convergence plot for imputation of female cohort*

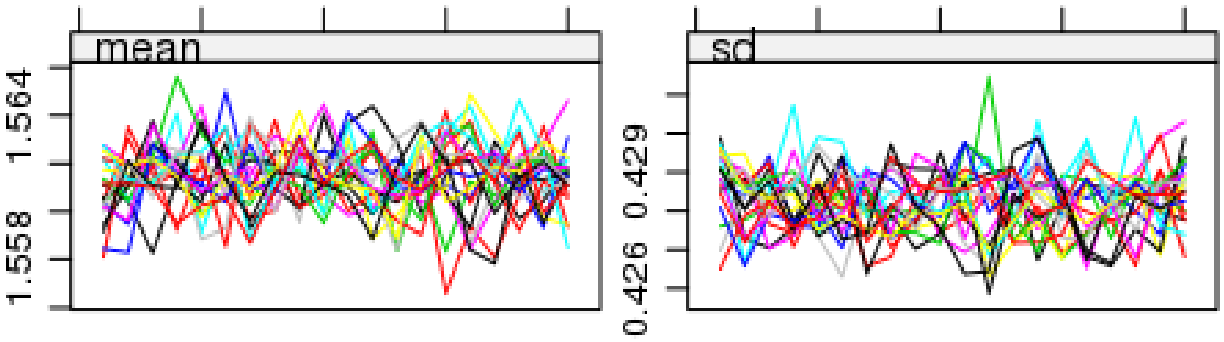
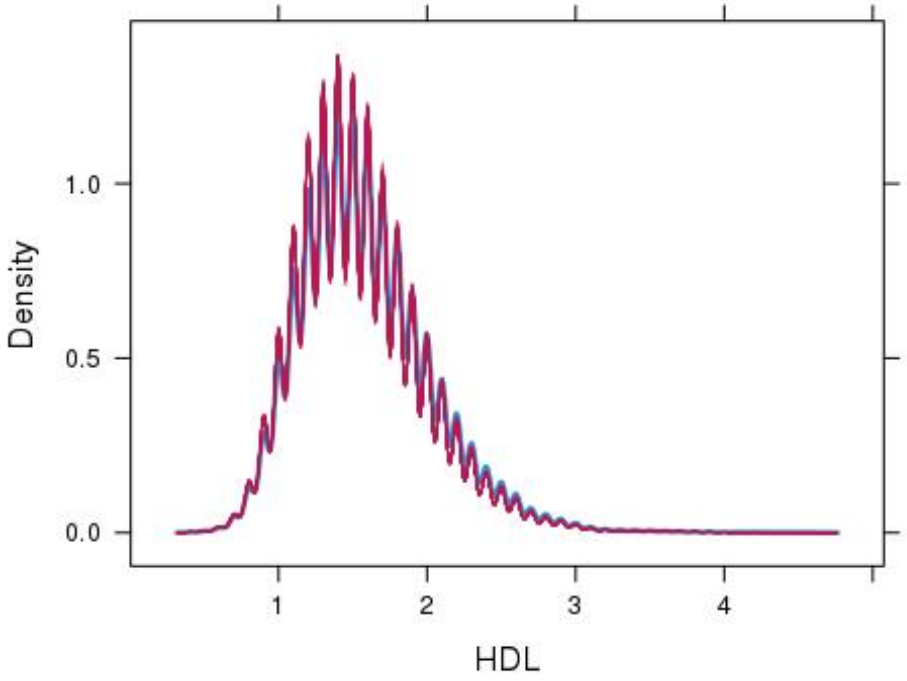


Figure A.3.10: *HDL density plot for imputation of female cohort*



Smoking status

Table A.3.1: *Distribution of real data and imputed data (%) for smoking status in imputation of female cohort*

	Smoking status (%)		
Imputation	Never	Ex	Current
Real data	56.03	16.98	27.00
1	57.92	16.15	25.93
2	58.33	16.27	25.40
3	58.78	15.76	25.47
4	57.50	15.96	26.55
5	59.48	15.92	24.61
6	58.98	16.12	24.90
7	57.65	15.90	26.45
8	57.94	15.77	26.29
9	58.33	15.71	25.96
10	59.18	15.70	25.12
11	58.90	16.05	25.05
12	59.06	15.45	25.49
13	58.35	15.77	25.88
14	57.90	16.19	25.91
15	58.62	15.78	25.60
16	58.56	16.00	25.44
17	58.84	15.69	25.46
18	59.23	15.11	25.66
19	58.50	15.65	25.85
20	59.45	15.56	25.00

Ethnicity

Table A.3.2: *Distribution of real data and imputed data (%) for ethnicity in imputation of female cohort*

	Ethnicity (%)									
Imputa-tion #	Asian other	Bangla-deshi	Black	Chinese	Indian	Mixed	Other Asian	Other	Pakistani	White
Real data	1.55	0.33	4.13	0.77	2.70	1.06	0.42	1.94	1.07	86.02
1	1.30	1.00	3.67	1.06	3.20	1.12	1.03	2.19	1.88	83.56
2	1.30	2.11	3.92	1.79	3.91	2.67	0.28	2.09	1.74	80.19
3	1.33	0.93	5.07	1.57	4.49	1.87	0.44	2.76	2.00	79.54
4	1.80	1.58	5.58	0.69	2.56	1.45	0.86	2.81	1.85	80.82
5	1.26	1.06	4.70	3.32	2.01	0.86	1.52	0.94	1.18	83.16
6	1.98	0.66	5.78	2.78	2.50	1.48	0.82	1.68	2.67	79.66
7	1.45	0.62	6.16	0.36	2.28	1.86	0.54	1.02	2.53	83.18
8	0.89	1.41	4.20	1.58	5.21	1.60	0.47	2.76	2.45	79.43
9	0.96	0.95	8.73	0.45	2.68	1.25	0.71	1.29	1.35	81.64
10	0.94	1.59	6.07	0.84	3.33	1.77	0.95	4.29	0.99	79.23
11	0.55	0.59	5.31	0.69	1.66	0.96	1.34	1.16	1.19	86.56
12	1.01	1.03	4.60	0.49	2.62	0.45	1.54	0.90	1.19	86.17
13	1.04	1.28	4.40	0.86	3.04	1.29	1.24	1.13	1.64	84.07
14	1.33	0.57	4.69	0.40	2.16	0.48	0.73	1.65	1.32	86.68
15	1.99	1.03	4.15	1.01	2.98	0.51	1.52	1.56	1.39	83.85
16	1.32	1.05	5.78	0.91	3.61	1.28	1.33	1.99	2.78	79.94
17	1.46	0.79	5.01	0.73	3.04	1.03	1.26	3.00	2.43	81.24
18	0.81	1.51	5.93	0.72	3.11	2.07	1.13	1.53	1.56	81.63
19	1.76	0.47	3.74	0.74	3.38	1.14	0.76	1.44	0.99	85.59
20	2.00	0.50	4.01	1.36	2.24	1.82	1.49	2.32	7.29	76.97

A.3.1.3 Results for imputation of male cohort

BMI

Figure A.3.11: *BMI convergence plot for imputation of male cohort*

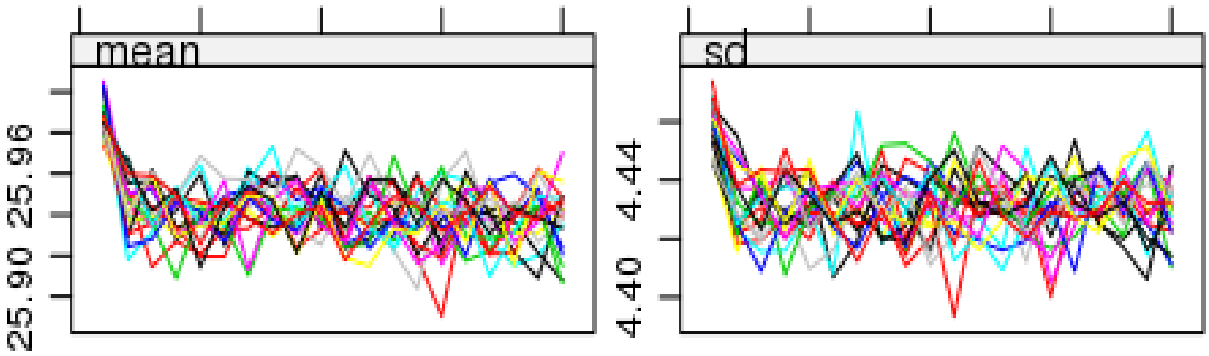
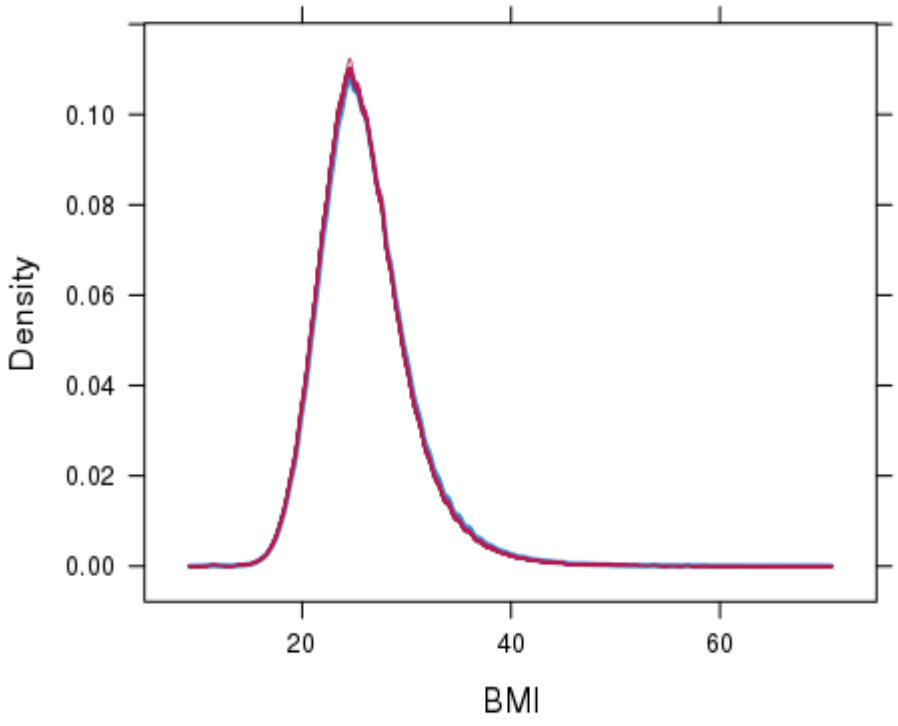


Figure A.3.12: *BMI density plot for imputation of male cohort*



SBP

Figure A.3.13: *SBP convergence plot for imputation of male cohort*

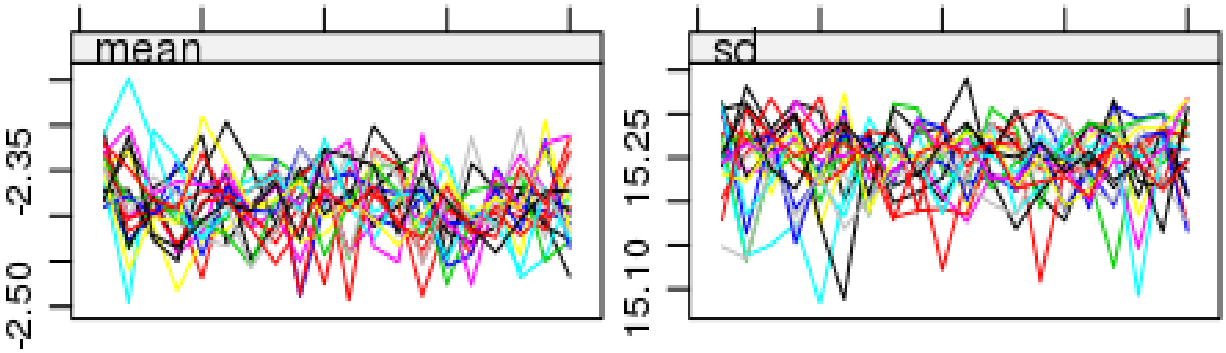
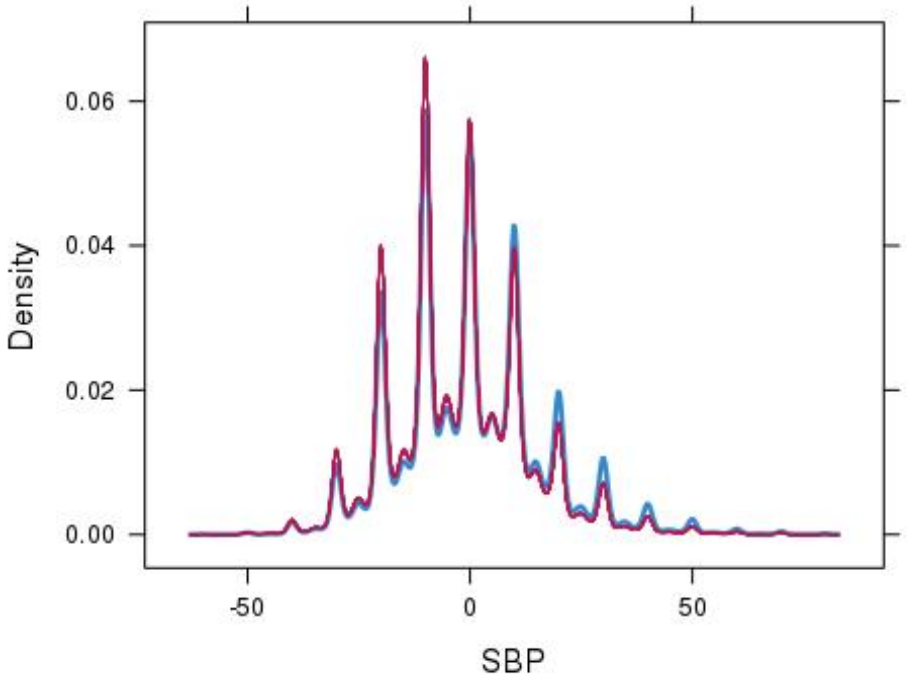


Figure A.3.14: *SBP density plot for imputation of male cohort*



SBP variability

Figure A.3.15: *SBP variability convergence plot for imputation of male cohort*

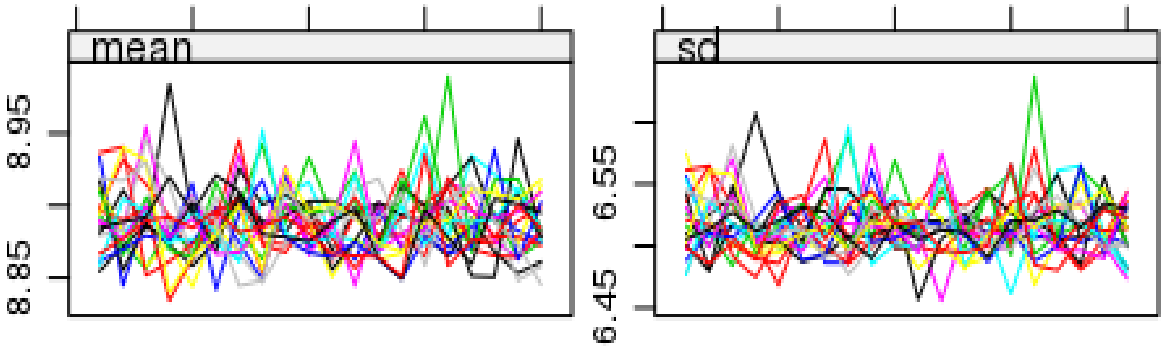
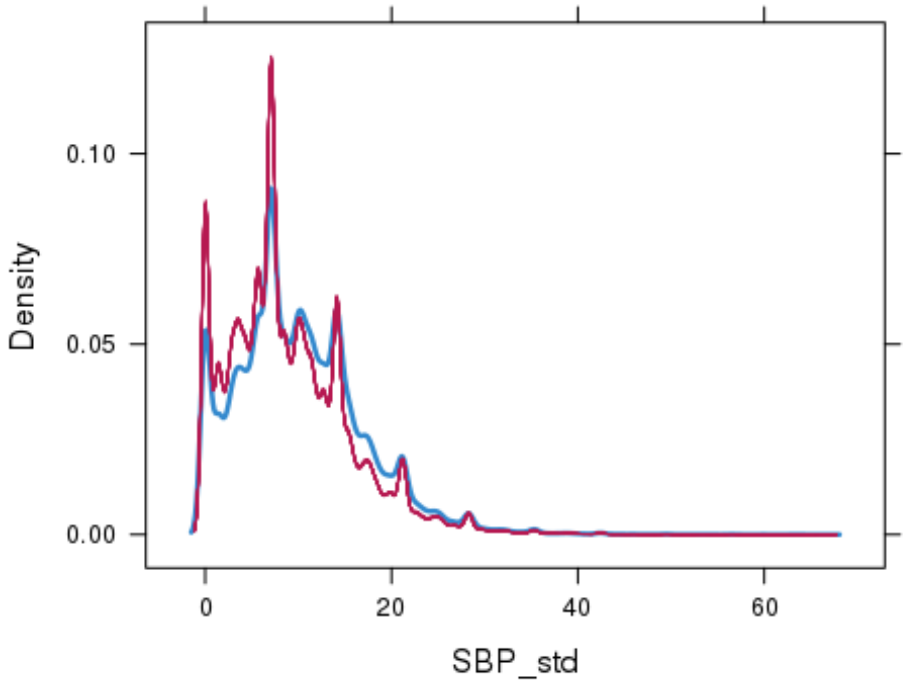


Figure A.3.16: *SBP variability density plot for imputation of male cohort*



Cholesterol

Figure A.3.17: *Cholesterol convergence plot for imputation of male cohort*

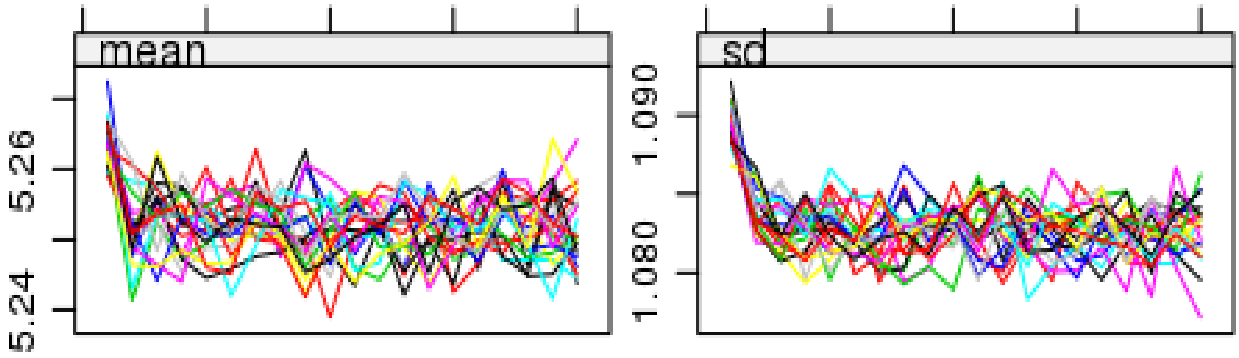
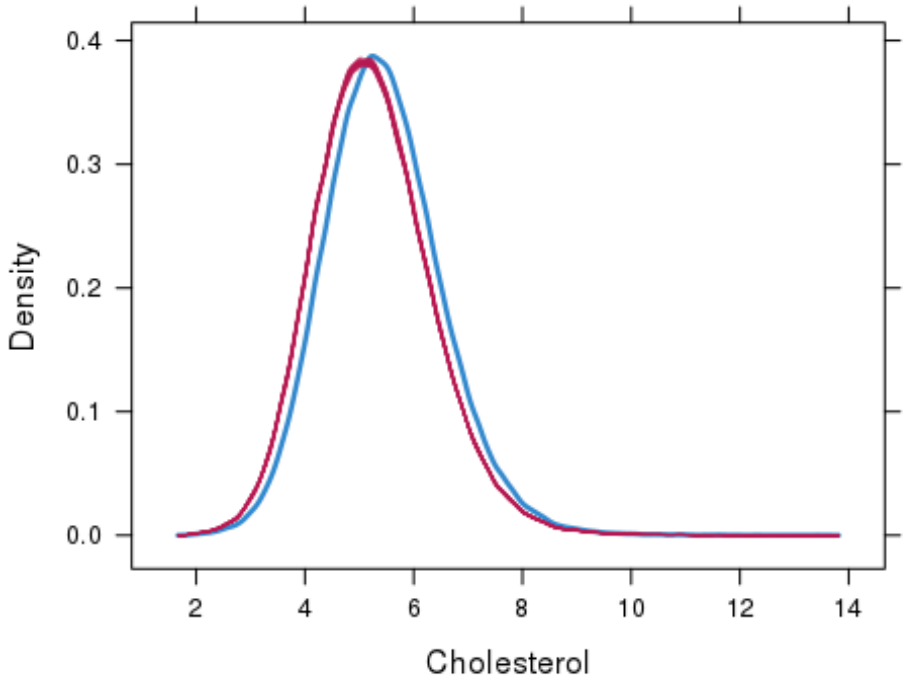


Figure A.3.18: *Cholesterol density plot for imputation of male cohort*



HDL

Figure A.3.19: *HDL convergence plot for imputation of male cohort*

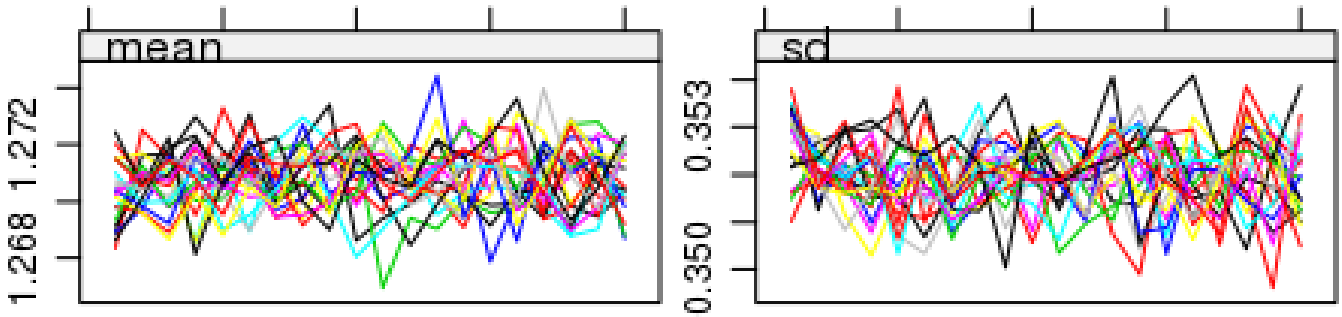
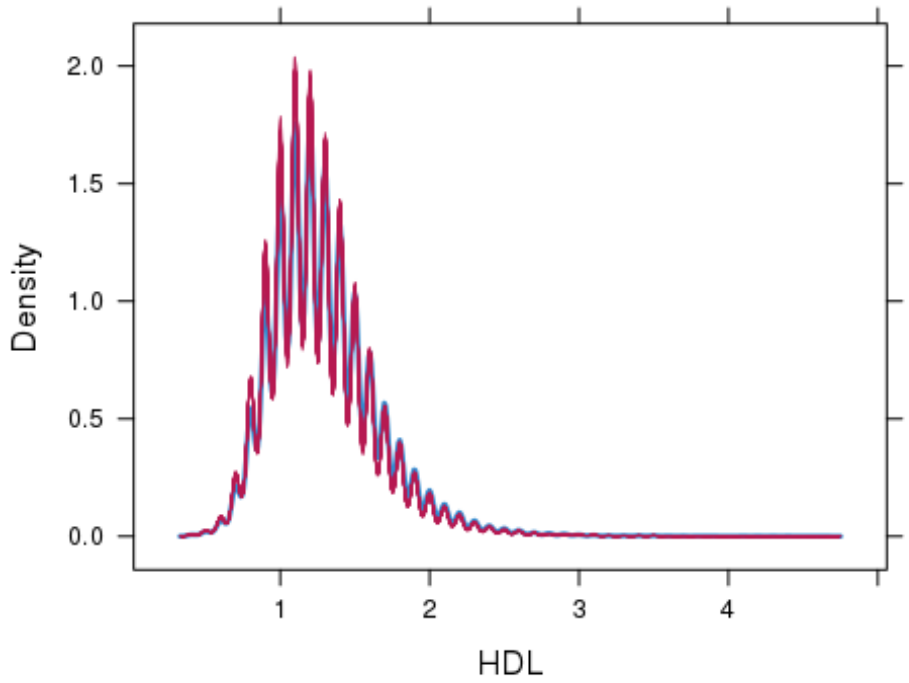


Figure A.3.20: *HDL density plot for imputation of male cohort*



Smoking

Table A.3.3: *Distribution of real data and imputed data (%) for smoking status in imputation of male cohort*

	Smoking status (%)		
Imputation	Never	Ex	Current
Real data	46.68	17.39	35.93
1	48.44	17.29	34.27
2	48.65	17.20	34.15
3	48.83	17.05	34.12
4	48.20	17.51	34.29
5	48.35	17.24	34.40
6	48.31	17.34	34.35
7	48.23	17.27	34.50
8	48.28	17.34	34.39
9	48.20	17.36	34.44
10	48.76	17.03	34.22
11	48.43	17.18	34.39
12	48.18	17.30	34.51
13	47.90	17.35	34.75
14	48.32	17.45	34.23
15	48.37	17.36	34.27
16	48.62	17.21	34.18
17	48.43	17.21	34.35
18	48.21	17.42	34.37
19	48.16	17.24	34.60
20	49.21	17.06	33.70

Ethnicity

Table A.3.4: *Distribution of real data and imputed data (%) for ethnicity in imputation of male cohort*

	Ethnicity (%)									
Imputa-tion #	Asian other	Bangla-deshi	Black	Chinese	Indian	Mixed	Other Asian	Other	Pakistani	White
Real data	1.55	0.5	4.00	0.61	3.11	0.99	0.49	1.95	1.26	85.54
1	2.62	1.24	3.97	1.64	4.49	2.44	2.39	2.63	3.48	75.11
2	2.36	1.86	3.58	1.75	1.88	1.68	2.27	1.53	3.39	79.70
3	2.16	1.20	4.93	0.86	3.54	1.53	1.52	2.69	2.72	78.84
4	1.17	0.61	6.83	0.17	2.30	1.41	1.12	1.61	1.74	83.04
5	2.53	2.19	2.39	2.12	2.81	2.59	2.11	1.99	4.74	76.52
6	2.88	1.35	2.94	2.26	4.70	1.55	1.89	1.81	3.40	77.23
7	0.82	0.57	2.44	1.35	1.72	1.27	0.79	0.73	2.13	88.19
8	1.99	1.09	2.02	1.01	1.85	1.45	0.92	1.63	2.72	85.35
9	1.67	1.10	3.44	0.75	2.56	1.10	1.18	1.02	2.54	84.65
10	2.69	2.31	3.05	1.06	2.82	2.50	2.00	2.41	4.49	76.66
11	1.44	0.99	2.19	0.80	1.40	1.50	1.59	0.59	2.07	87.44
12	1.84	0.91	2.06	1.15	1.74	1.36	0.89	1.94	2.35	85.75
13	1.42	0.97	1.95	1.31	1.03	1.10	1.65	1.08	1.98	87.51
14	0.97	0.82	1.87	0.67	3.07	0.83	0.97	1.97	1.66	87.16
15	1.16	1.17	1.67	1.52	1.14	1.33	1.68	1.38	3.00	85.96
16	1.54	1.33	2.88	0.93	1.36	1.18	1.68	0.88	2.42	85.81
17	1.27	0.65	3.94	0.70	3.75	2.64	0.60	1.72	1.56	83.17
18	1.33	1.08	2.16	0.97	2.72	0.77	0.86	1.51	1.96	86.65
19	1.59	0.73	2.42	0.71	2.68	0.74	0.68	0.45	1.33	88.66
20	4.76	1.87	15.10	1.64	4.89	2.42	1.66	7.36	4.60	55.71

A.3.2 Tripod statement

Section/Topic			Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted. The main objectives for this manuscript are not a model development + validation; therefore the title does not follow this format. The manuscript does however involve the development and internal validation of a risk prediction model, and we report on the guidelines below for this process, where relevant.	74
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	74
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	75
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both. The main objectives for this manuscript are not a model development + validation, therefore the objectives differ from what is requested here.	76
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	77
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	77
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	77
	5b	D;V	Describe eligibility criteria for participants.	77
	5c	D;V	Give details of treatments received, if relevant.	NA
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	77/78
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted. None – routinely collected data	NA
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	78/Chapter 2 .
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors. None – routinely collected data	NA
Sample size	8	D;V	Explain how the study size was arrived at. This was all patients that met the eligibility criteria	77
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	79/197
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	78/79
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation. Classical predictor selection was not carried out (i.e. backwards selection) as we pre-specified different sets of predictors for each model, this was a key part of the study.	79
	10c	V	For validation, describe how the predictions were calculated.	79
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	79/80
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	NA
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	NA
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors. *Rather than comparing the development and validation cohorts (which were chosen at random from the same cohort), we chose to compare the development cohort with the one from the published QRISK algorithm. This is more important for this study as it helps validate the generalisability of the results to the models used in practice. This is not meant to be a classical model development project.* The table was however very large given we had separate male and female cohorts, and therefore only the CPRD cohort is presented in the main text, with the comparison presented in Chapter 2.	Chapter 2
Results				

Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	83
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	, Table 3.1
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome). Validation was not the main aim of this paper, all comparisons have been done with QRISK, which is the model used in practice across the UK. See point 12.	NA
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	Table 3.1
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	NA
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	Table 3.2/Supplementary Table A.3.7
	15b	D	Explain how to the use the prediction model.	NA
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model. Given the size of the cohort performance metrics took a long time to derive. Confidence intervals for majority of these metrics can only be obtained by bootstrapping. This would involve deriving the metrics hundreds of times, which could take a lot of computational time. Given the size of the cohort I expect the confidence interval to be small and therefore I have not done this.	Table 3.3
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	NA
Discussion **The discussion has a very different structure as the main aim of this paper was not development and validation of a model to be used in practice**				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	98
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	NA
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	95/96/97
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	NA
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	Code provided on GitHub
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	Given in published manuscript, not thesis

Table A.3.5: *Incidence rates of CVD from the CPRD cohort and comparison with QRISK3 development cohort by age band (female cohort)*

CPRD cohort					QRISK3 cohort			
Age	Incident cases	Person years	Rate per 1000 person years		Age	Incident cases	Person years	Rate per 1000 person years
25-29	626	2499863	0.25		25-29	832	3455662	0.24
30-34	1026	1777476	0.58		30-34	1878	3802577	0.49
35-39	1973	1739617	1.13		35-39	3636	3551460	1.02
40-44	3098	1507321	2.06		40-44	5651	2971995	1.9
45-49	4304	1342797	3.21		45-49	8272	2581104	3.2
50-54	6481	1281258	5.06		50-54	12022	2490263	4.83
55-59	7845	1013648	7.74		55-59	14524	1944140	7.47
60-64	9941	823913	12.07		60-64	18471	1625795	11.4
65-69	12374	678773	18.23		65-69	22510	1314303	17.1
70-74	13833	525681	26.31		70-74	25462	1015263	25.1
75-79	14134	390288	36.21		75-79	26883	765681	35.1
80-84	10912	221282	49.31		80-84	20408	424994	48.0
Total	86547	13801919	6.27		Total	160549	25943236	6.19
Average observed risk			6.0%		Average observed risk			5.8%
Average predicted risk			5.0%		Average predicted risk			4.7%

Table A.3.6: *Incidence rates of CVD from CPRD cohort and comparison with QRISK3 development cohort by age band (male cohort)*

CPRD cohort			
Age	Incident cases	Person years	Rate per 1000 person years
25-29	906	2548923	0.36
30-34	2012	1748443	1.15
35-39	4245	1788096	2.37
40-44	6811	1560256	4.37
45-49	9826	1346772	7.30
50-54	13036	1225972	10.63
55-59	13637	911484	14.96
60-64	14097	685171	20.57
65-69	14111	507205	27.82
70-74	12573	344401	36.51
75-79	9976	213435	46.74
80-84	5821	97078	59.96
Total	107051	12977235	8.24
Average observed risk			7.7%
Average predicted risk			6.5%

QRISK3 cohort			
Age	Incident cases	Person years	Rate per 1000 person years
25-29	1351	3379716	0.4
30-34	3823	3880890	0.99
35-39	7963	3748285	2.12
40-44	12750	3192048	3.99
45-49	17763	2972642	6.65
50-54	24040	2437106	9.86
55-59	25464	1796342	14.18
60-64	27021	1372104	19.69
65-69	26903	1013291	26.55
70-74	24549	691866	35.48
75-79	19820	438864	45.16
80-84	11569	198481	58.29
Total	203016	24821632	8.18
Average observed risk			7.5%
Average predicted risk			6.4%

Table A.3.7: *Comparison of hazard ratios of categorical variables from model B with QRISK3*

	Female		Male	
	CPRD cohort (model B)	QRISK3	CPRD cohort (model B)	QRISK3
Atrial fibrillation	3.29	4.92	2.16	2.42
Atypical antipsychotic medication use	1.33	1.29	1.17	1.14
Corticosteroid Use	2.22	1.81	1.93	1.58
CKD (stage 3/4/5)	1.96	1.93	2.02	2.05
Erectile dysfunction	NA	NA	1.08	1.25
Ethnicity:asianother	1.37	1.08	1.13	1.04
Ethnicity:bangladeshi	1.30	1.34	1.16	1.70
Ethnicity:black (African/Caribbean)	1.35	0.84/0.67	0.95	0.70/0.67
Ethnicity:chinese	1.32	0.722	1.01	0.66
Ethnicity:indian	1.29	1.32	1.18	1.32
Ethnicity:mixed	1.51	NA	1.09	NA
Ethnicity:other	1.38	0.84	1.09	0.76
Ethnicity:pakistani	1.46	1.76	1.19	1.61
Family history of CVD	1.41	1.58	1.45	1.72
Hypertension (treated)	1.35	1.66	1.32	1.68
Migraine	1.23	1.35	1.19	1.29
Rheumatoid arthritis	1.42	1.24	1.35	1.23
Severe mental illness	1.30	1.13	1.22	1.13
Smoker (Ex)	1.28	1.14	1.19	1.21
Smoker (current | light/moderate/heavy)	1.96	1.75/1.95/ 2.34	1.84	1.74/1.89/ 2.20
Systemic lupus erythematosus	1.55	2.14	1.10	1.55
Townsend = 2	1.10	NA	1.03	NA
Townsend = 3	1.27	NA	1.13	NA
Townsend = 4	1.47	NA	1.26	NA
Townsend = 5 (most deprived)	1.83	NA	1.40	NA
Type 1 diabetes	4.29	5.62	2.88	3.44
Type 2 diabetes	2.92	2.91	2.30	2.36

Table A.3.8: *Numbers and percentages of patients registered on 1st Jan 2016 who cross the treatment threshold (10-year CVD risk of 10%) when using models B – F instead of model A*

	Predicted CVD risk according to model A (QRISK2)						Predicted CVD risk according to model A (QRISK2)					
	5-6%	6-7%	7-8%	8-9%	9-10%	**(<10%)**	10-11%	11-12%	12-13%	13-14%	14-15%	**(>10%)**
Female (N = 387547)												
N	16148	13183	10816	9405	8180	**317387**	7387	6682	5935	5170	4684	**70160**
Model B	144 (1%)	391 (3%)	724 (7%)	1582 (17%)	2435 (30%)	**5362 (2%)**	3129 (42%)	373 (6%)	29 (0%)	15 (0%)	9 (0%)	**3575 (5%)**
Model C	118 (1%)	263 (2%)	582 (5%)	1198 (13%)	2159 (26%)	**4363 (1%)**	3863 (52%)	975 (15%)	134 (2%)	26 (1%)	14 (0%)	**5064 (7%)**
model D	4 (0%)	6 (0%)	18 (0%)	27 (0%)	43 (1%)	**98 (0%)**	7322 (99%)	6583 (99%)	5804 (98%)	4970 (96%)	4440 (95%)	**43822 (62%)**
model E	5 (0%)	10 (0%)	19 (0%)	36 (0%)	42 (1%)	**112 (0%)**	7329 (99%)	6581 (98%)	5789 (98%)	4965 (96%)	4432 (95%)	**44826 (64%)**
model F	7 (0%)	13 (0%)	22 (0%)	45 (0%)	58 (1%)	**147 (0%)**	7313 (99%)	6561 (98%)	5762 (97%)	4928 (95%)	4404 (94%)	**44761 (64%)**
Male (N = 352014)												
N	18974	16897	14601	13211	11208	**254714**	9690	8668	7462	6737	6004	**97300**
Model B	64 (0%)	152 (1%)	333 (2%)	988 (7%)	2300 (21%)	**3859 (2%)**	2696 (28%)	154 (2%)	5 (0%)	1 (0%)	0 (0%)	**2856 (3%)**
Model C	55 (0%)	202 (1%)	432 (3%)	1055 (8%)	2489 (22%)	**4250 (2%)**	3305 (34%)	319 (4%)	17 (0%)	6 (0%)	3 (0%)	**3652 (4%)**
model D	2 (0%)	5 (0%)	8 (0%)	20 (0%)	55 (0%)	**91 (0%)**	9610 (99%)	8507 (98%)	7224 (97%)	6349 (94%)	5481 (91%)	**50014 (51%)**
model E	2 (0%)	6 (0%)	11 (0%)	22 (0%)	70 (1%)	**112 (0%)**	9596 (99%)	8499 (98%)	7209 (97%)	6310 (94%)	5461 (91%)	**51368 (53%)**
model F	7 (0%)	30 (0%)	36 (0%)	72 (1%)	127 (1%)	**283 (0%)**	9505 (98%)	8367 (97%)	7036 (94%)	6104 (91%)	5202 (87%)	**50699 (52%)**

A.3.3.1 Breakdown of calculations for extrapolation to UK population

Given the secular trend there are very few patients whose risk increases when comparing model A to model F. Therefore we solely focus on the proportion of patients who are initially classified as high risk, that cross the threshold to low risk. Lots of the data used in the below calculations was taken from Table A.3.8.

1) Number of patients aged 25-84 in England (37,273,000)

This was taken directly from the reference[145] given = 37,273,000

2) Proportion (number) of patients aged 25-84 eligible for risk assessment = 79% (29,382,463)

The number of patients aged 25-84 that were registered on 1st Jan 2016 and had linked data = 938,150. The number of this group that had not had a CVD event or statin treatment prior to 1st Jan 2016 = 739,561. We took the ratio of these to be the proportion of patients aged 25-84 that would be eligible for risk assessment = 79%.

3) Proportion of patients that would be classified as high risk (> 10%) = 22.64% (6,652,920)

Of the 739,583 patients, 167,460 patients were classified as high risk = 22.64%. Therefore we assumed 22.64% of the English population aged 25-84 would be classified as high risk = 6,652,920

4) Proportion (number) of high risk patients that would be reclassified as low risk according to model F = 57.00% (3,792,474)

Of the 167,460 patients classified as high risk, 95,460 are reclassified to low risk = 57.00%. Therefore we assumed 57.00% of the high risk group in the English population would also be reclassified = 3,792,474.

A.4 Chapter 4 Appendices

A.4.1 Calculation of N_{min}, minimum required sample size following published criteria, and shrinkage factor of models meeting this sample size criteria

A.4.1.1 Methods

We followed the criteria outlined by Riley et al.[35] for calculating the minimum required sample size for a risk prediction model. These are (i) ensure a global shrinkage factor $S_{VH} > 0.9$; (ii) ensuring a small absolute difference in the apparent and adjusted $R^2_{Nagelkerke}$; (iii) ensure precise estimate of overall risk (model intercept). The recommended estimate of the global shrinkage factor is the shrinkage factor of Van Houwelingen and Le Cessie[214], S_{VH}. The entity $R^2_{Nagelkerke}$[215] is an estimate of the proportion of variance explained, that always lies between 0 and 1. When estimating this quantity, the apparent estimate will be optimistic, and can be adjusted to get an unbiased estimate, which are the two values of interest here.

The main challenge in following these recommendations is the need to calculate $R^2_{CS\_ADJ}$[216] (an unbiased estimate of the Cox-Snell[217] R^2) to calculate the sample size, which can only be calculated after fitting the model. It is recommended to use metrics provided by previous prediction models developed on similar populations to estimate $R^2_{CS\_ADJ}$. In this study, we can use the model developed on the whole development cohort to calculate $R^2_{CS\_ADJ}$ directly. This value of $R^2_{CS\_ADJ}$ allows us to calculate the minimum required sample size for a model developed in this population.

A.4.1.2 Results

Criteria (i)

We start by calculating:

$$R^2_{CS\_APP} = 1 - \exp\left(\frac{LR}{n}\right)$$
$$= 0.0780$$

Where $R^2_{CS\_APP}$ is a biased estimate of the Cox-Snell[217] R^2 (based on the work by Magee[218]), LR = likelihood ratio of the model developed on the entire population, and n = 1,865,078 is the size of the cohort used in that model. Next we calculate:

$$S_{VH} = 1 + \left(\frac{p}{n * \ln(1 - R^2_{CS_{APP}})}\right)$$
$$= 0.99991$$

Where S_{VH} is the global shrinkage factor of Van Houwelingen and Le Cassie[214], and p = 13 is the number of predictor variables. There are 9 variables, and but Smoking contributes two dummy variables (categories = yes/ex/never) and Townsend contributes 4 dummy variables (5 deprivation categories). Then we can calculate:

$$R^2_{CS\_ADJ} = S_{VH} * R^2_{CS\_APP}$$
$$= 0.0780$$

To get a model which has a shrinkage of at least S_{VH} = 0.9, as is recommended in the guidelines, we use the following formula:

$$N_{min} = \frac{p}{(S_{VH} - 1) * ln\left(1 - {R^2_{CS\_ADJ}}/{S_{VH}}\right)}$$
$$= \frac{13}{(0.9 - 1) * ln\left(1 - {0.0781}/{0.9}\right)}$$
$$= 1434$$

Criteria (ii)

In order for the difference between the apparent and adjusted $R^2_{Nagelkerke}$[215] to be suitable, the following equation must be satisfied:

$$S_{VH} \geq \frac{R^2_{CS\_ADJ}}{R^2_{CS\_ADJ} + \delta * max\left(R^2_{CS\_APP}\right)}$$

Where $S_{VH} = 0.9$ is the desired shrinkage, δ is the acceptable difference between the apparent and adjusted $R^2_{Nagelkerke}$, and

$$max\left(R^2_{CS\_APP}\right) = 1 - exp\left(\frac{2 * \ln(L_{null})}{n}\right)$$
$$= 0.6987$$

where L_{null} is the log likelihood of the null model with no covariates, and was calculated directly from the population derived model. The recommended $\delta = 0.05$ and therefore:

$$\frac{R^2_{CS\_ADJ}}{R^2_{CS\_ADJ} + \delta * max\left(R^2_{CS\_APP}\right)} = \frac{0.0780}{0.0780 + 0.05 * 0.6987}$$
$$= 0.6906$$
$$\leq 0.9$$

and the criteria is satisfied.

Criteria (iii)

This requires that the confidence interval around the cumulative incidence at t, time point of interest, to be smaller than 0.05. We will assume an exponential distribution which is the simplest approach to this.

Let T = total follow up time in years if N_{min} is the sample size (average follow up multiplied by sample size), $\hat{\lambda}$ be the estimated number of events per person year, and t = 10 years is the point of interest (time at which we are making risk predictions). Then the confidence interval is then calculated as:

$$CI = 1 - exp\left(-\left(\hat{\lambda} \pm 1.96 * \sqrt{\frac{\hat{\lambda}}{T}}\right) * t\right)$$
$$= 1 - exp\left(-\left(0.0063 \pm 1.96 * \sqrt{\frac{0.0063}{7.0230 * 1434}}\right) * 10\right)$$
$$= (0.0290, 0.0751)$$

The size of the confidence interval is $0.0290 < 0.05$.

Therefore the value of N_{min} = 1434 satisfies all the criteria and is included as a sample size in our main analysis.

The exact same process was followed for the male cohort, and the value of N_{min} = 1405 was found to satisfy all the criteria.

A.4.1.3 Shrinkage factor of models that have N = N_{min}

For the female cohort, the shrinkage factors of models generated using N = N_{min} had a mean of 0.898, 2.5th percentile = 0.854, 50th percentile = 0.901, and 97.5th percentile = 0.927.

For the male cohort, the shrinkage factors of models generated using N = N_{min} had a mean of 0.898, 2.5th percentile = 0.850, 50th percentile = 0.900, and 97.5th percentile = 0.927.

A.4.2 Supplementary tables and figures

Figure A.4.1: *Boxplots of the 95% percentile range of risk for individuals across the 1000 models (male cohort)*

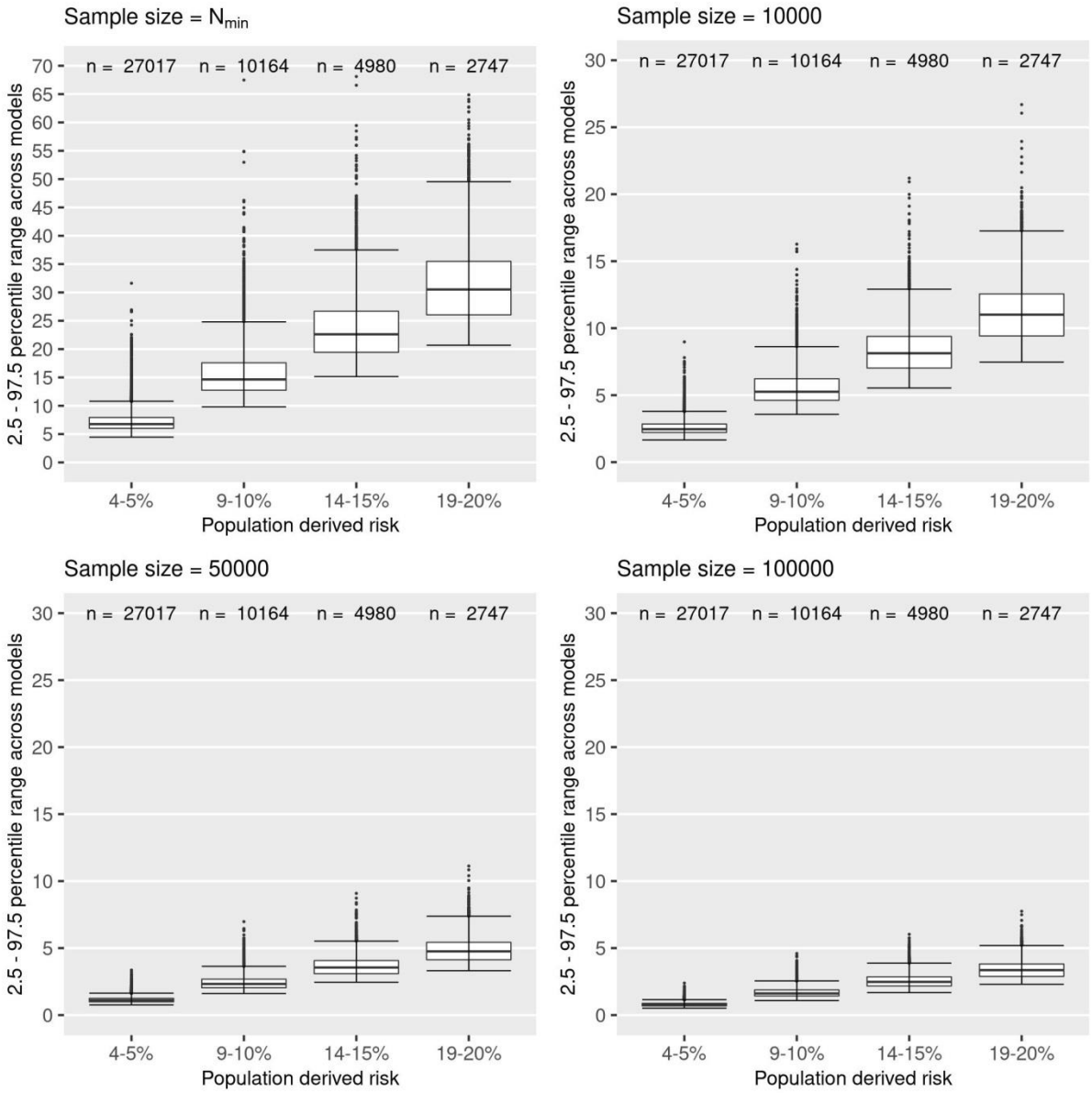
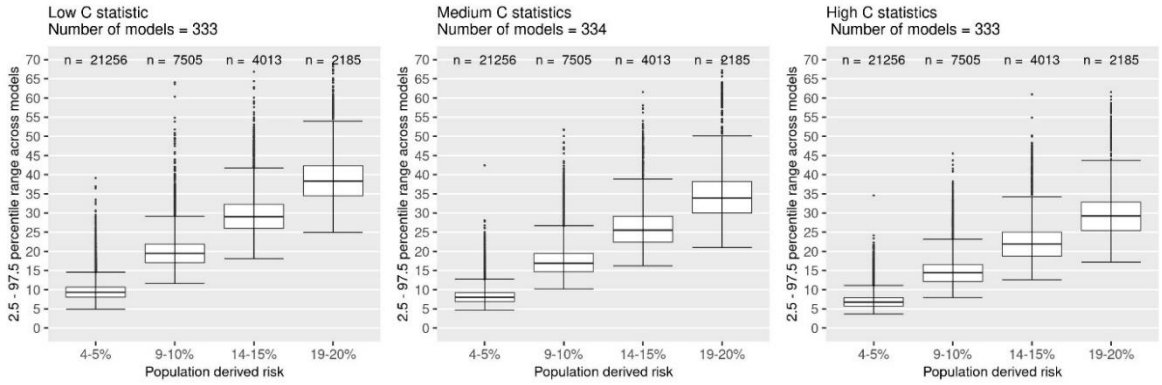
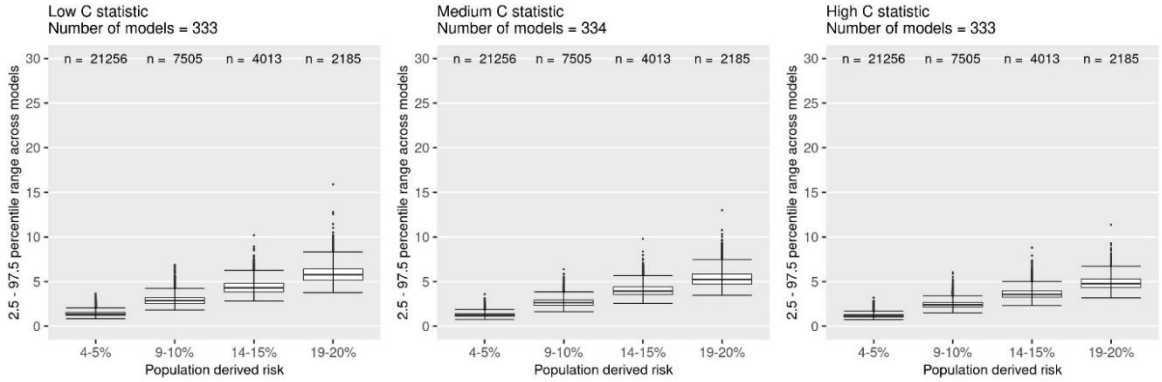


Figure A.4.2: *Boxplots of the 95% percentile range of risk for individuals across subsets of the 1000 models, defined by the C-statistic of the models (female cohort)*

Sample size = N_{min}



Sample size = 50 000



Sample size = 100 000

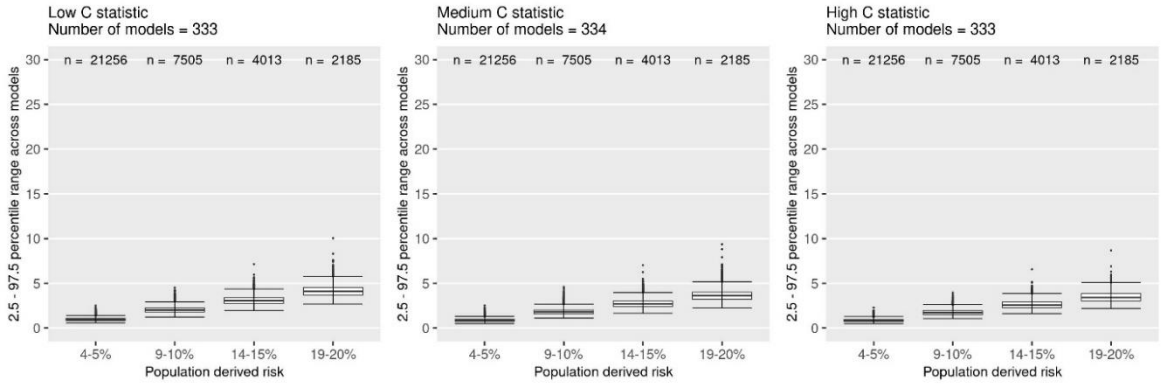
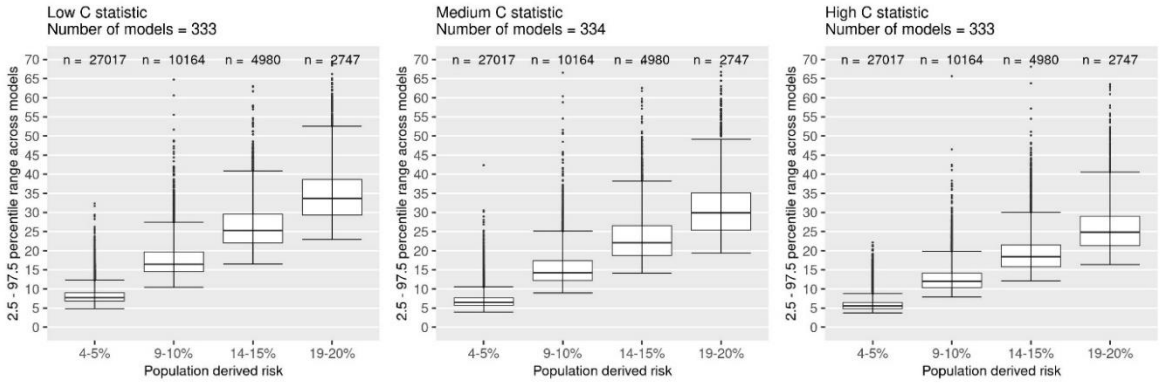
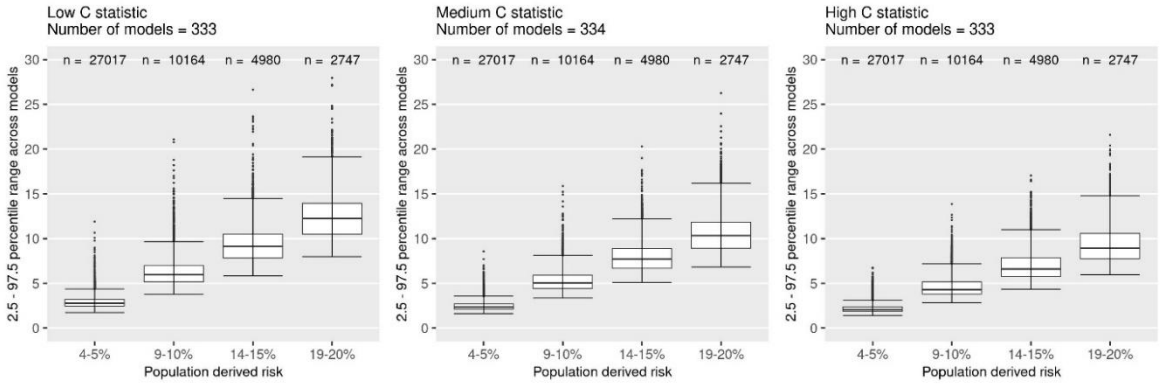


Figure A.4.3: *Boxplots of the 95% percentile range of risk for individuals across subsets of the 1000 models, defined by the C-statistic of the models (male cohort)*

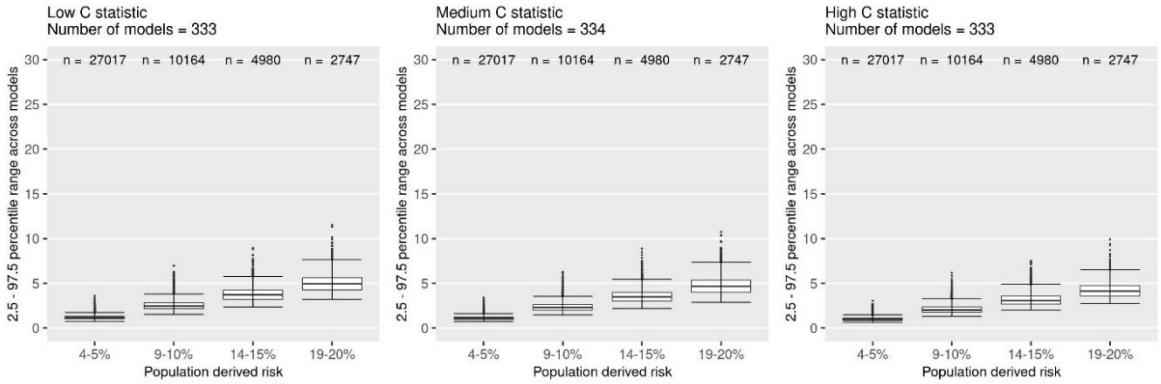
Sample size = N_{min}



Sample size = 10 000



Sample size = 50 000



Sample size = 100 000

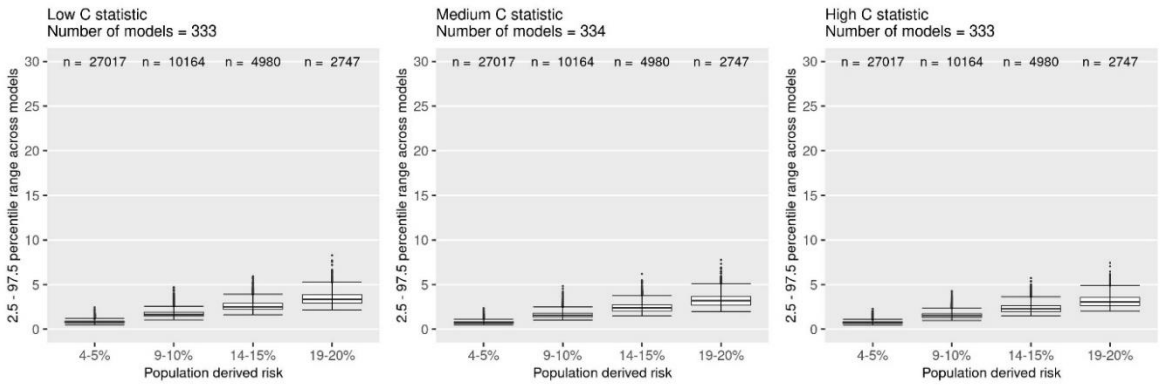
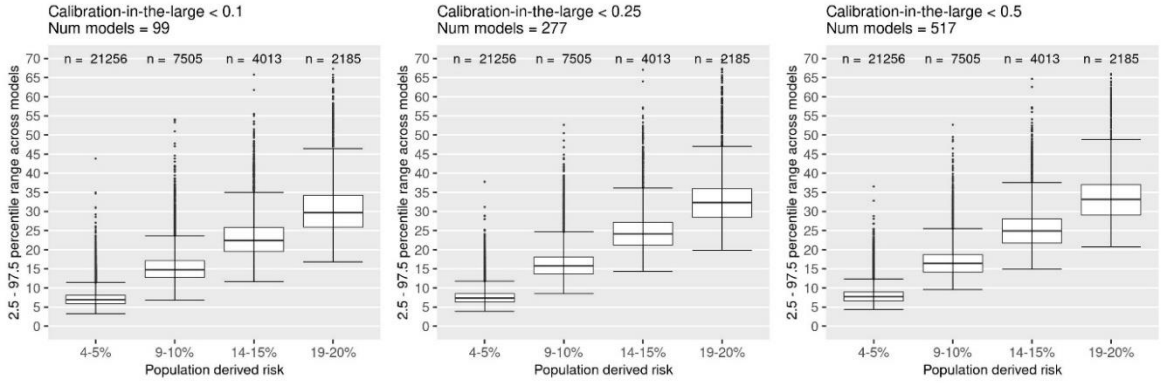
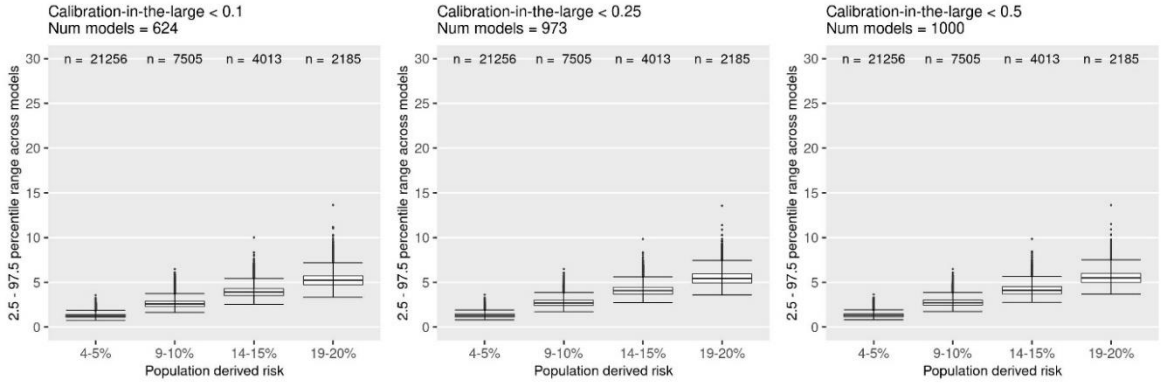


Figure A.4.4: *Boxplots of the 95% percentile range of risk for individuals across subsets of the 1000 models, defined by the calibration-in-the-large of the models (female cohort)*

Sample size = N_{min}



Sample size = 50 000



Female Figure 3 Sample size = 100 000

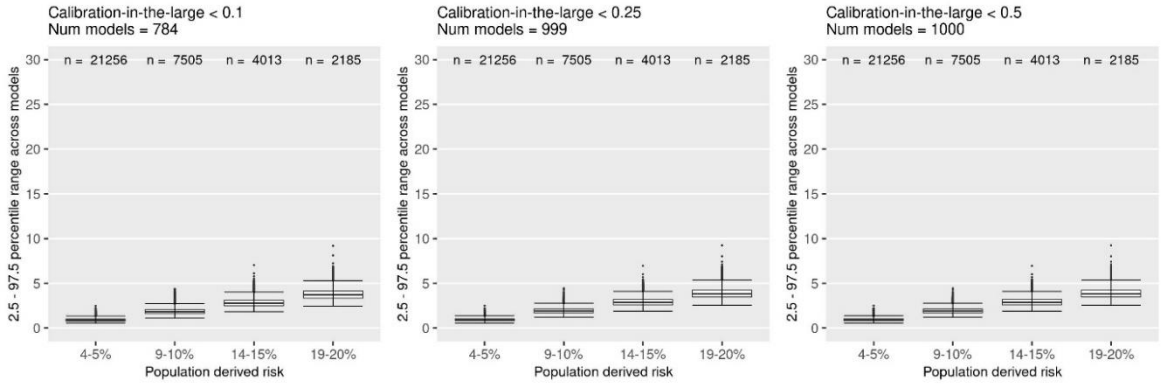
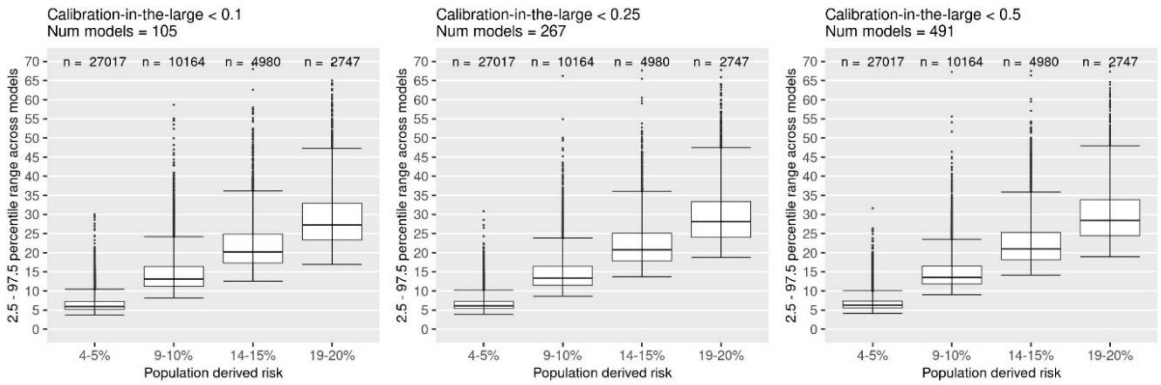
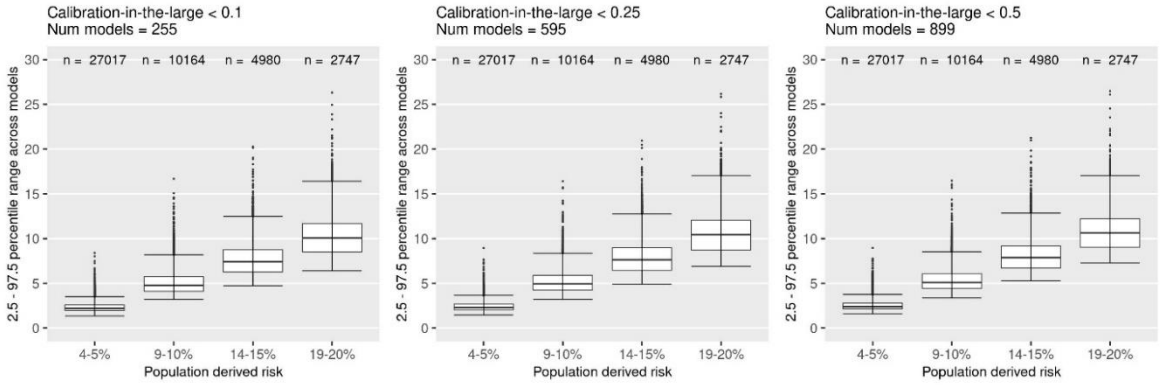


Figure A.4.5: *Boxplots of the 95% percentile range of risk for individuals across subsets of the 1000 models, defined by the calibration-in-the-large of the models (male cohort)*

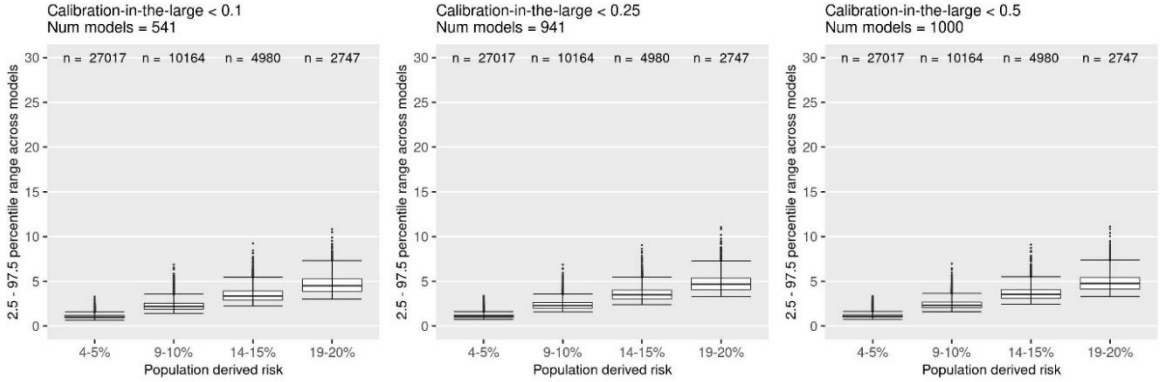
Sample size = N_{min}



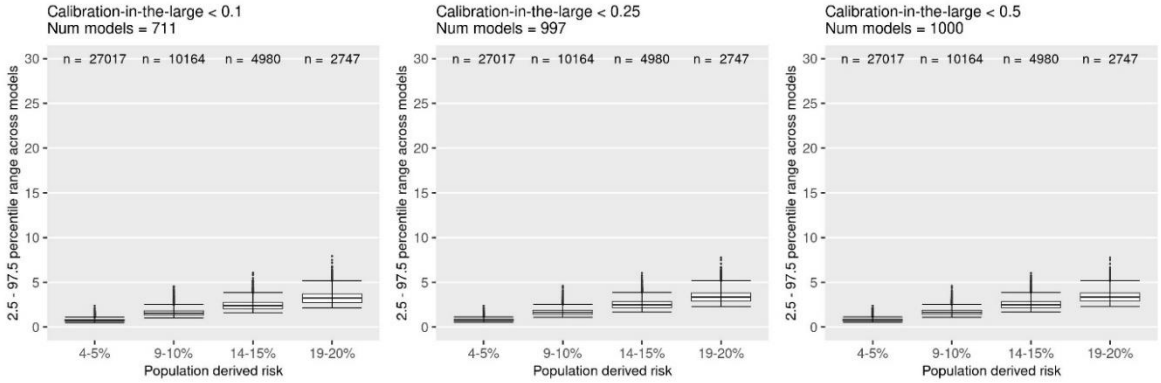
Sample size = 10 000



Sample size = 50 000



Sample size = 100 000



A.5 Chapter 5 Appendices

A.5.1 Supplementary tables and figures

Table A.5.1: *Prevalence of statin use each year in CVD primary prevention cohort*

Year	Total follow up in years	Prescriptions	Prescriptions per 1000 person years
1998	738031	8585	11.63
1999	837959	27044	32.27
2000	1040735	56827	54.60
2001	1295439	106348	82.09
2002	1429880	186787	130.63
2003	1591167	310417	195.09
2004	1648158	488761	296.55
2005	1700595	661788	389.15
2006	1744492	870761	499.15
2007	1775518	1030767	580.54
2008	1801027	1142698	634.47
2009	1799115	1240413	689.46
2010	1777610	1263314	710.68
2011	1712269	1217875	711.26
2012	1669433	1216451	728.66
2013	1546937	1152425	744.97
2014	1325230	1000030	754.61
2015	1051748	806999	767.29
2016	204290	154993	758.69
Total	26689633	12943283	484.96

Table A.5.2: *Rate of initiation of statin treatment each year in CVD primary prevention*

Year	Total follow up in years	Number initiated	Incidence rate per 1000 person years
1998	738031	2133	2.89
1999	837959	3391	4.05
2000	1040735	5664	5.44
2001	1295439	9587	7.40
2002	1429880	15370	10.75
2003	1591167	22554	14.17
2004	1648158	31251	18.96
2005	1700595	30791	18.11
2006	1744492	37520	21.51
2007	1775518	29573	16.66
2008	1801027	29384	16.32
2009	1799115	28322	15.74
2010	1777610	22398	12.60
2011	1712269	18499	10.80
2012	1669433	19326	11.58
2013	1546937	16623	10.75
2014	1325230	13255	10.00
2015	1051748	10295	9.79
2016	204290	2330	11.41
Total	26689633	348266	13.05

A.5.1.1 Calibration of the marginal structural model and interval censored Cox model

Calibration of the models from section 5.4.4 (the marginal structural model (MSM) and the interval censored model) are presented here. When assessing the calibration of the MSM (in either the development or validation cohorts), it was done on the subset of patients who received no statin treatment during follow up. This is because the risk scores generated were conditional on having no statin treatment during follow up. The calibration of the interval censored Cox model was carried out on the entire development/validation cohorts. The interval censored Cox model would under predict the risk of patients who do not receive statins during follow up.

Figure A.5.1 and Figure A.5.2 show that when secular trend was not adjusted for under the MSM setting, there was a significant under prediction of risks in the validation cohort, which could be accounted for by modelling calendar time. This is very similar to the non-MSM setting, for which the calibration of the interval censored Cox model is presented in Figure A.5.3 and Figure A.5.4.

Figure A.5.1: *Calibration of the MSM in the validation cohort, with or without adjustment for calendar time (male cohort)*

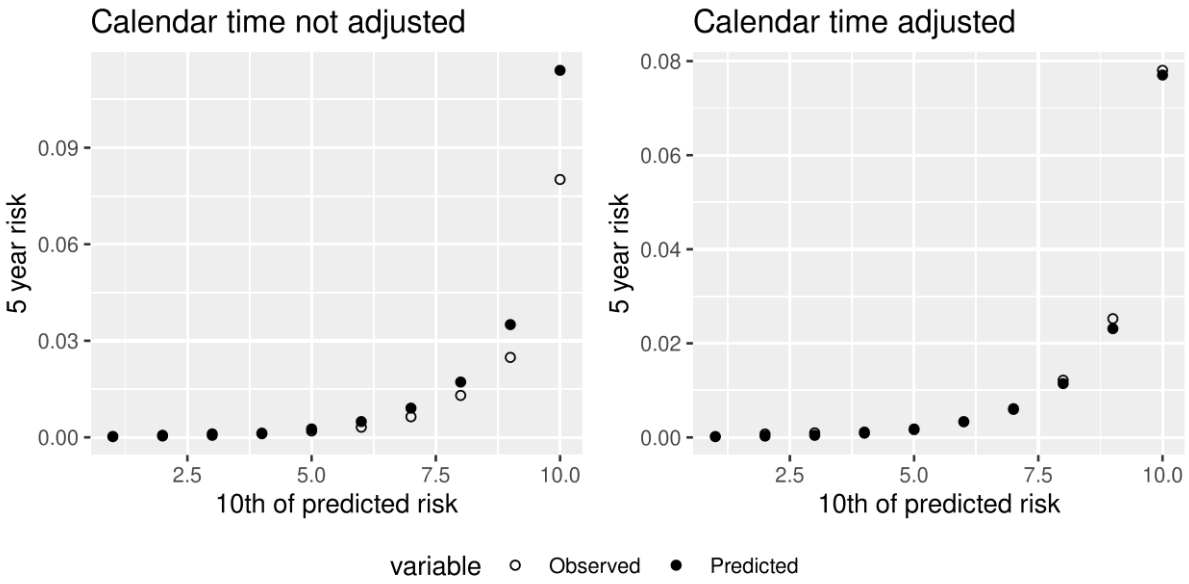


Figure A.5.2: *Calibration of the MSM in the validation cohort, with or without adjustment for calendar time (female cohort)*

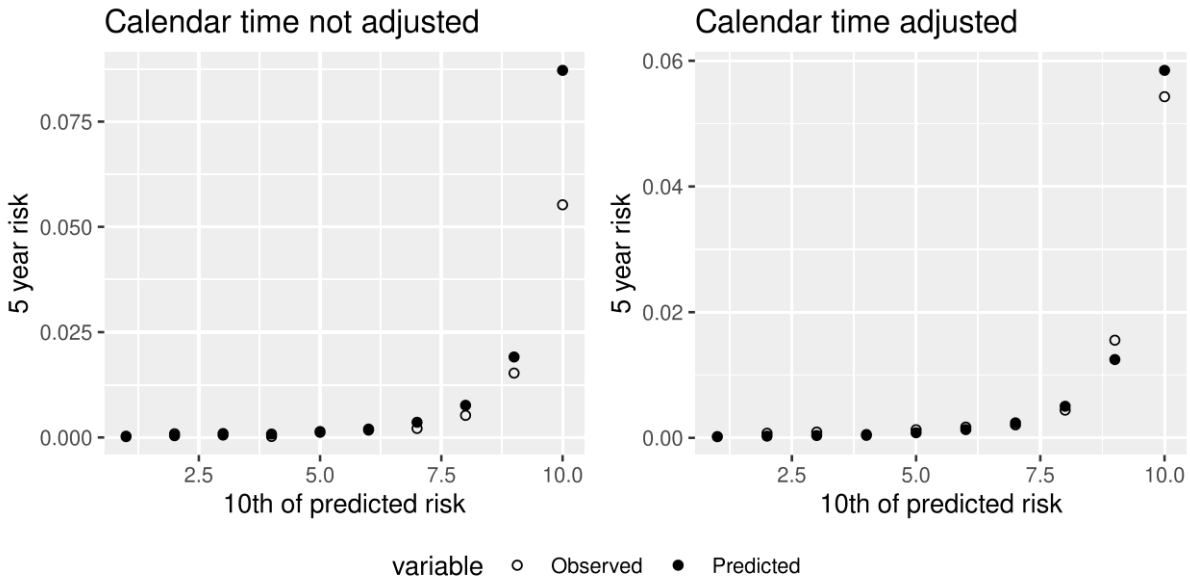


Figure A.5.3: *Calibration of the interval censored Cox model in the validation cohort, with or without adjustment for calendar time (male cohort)*

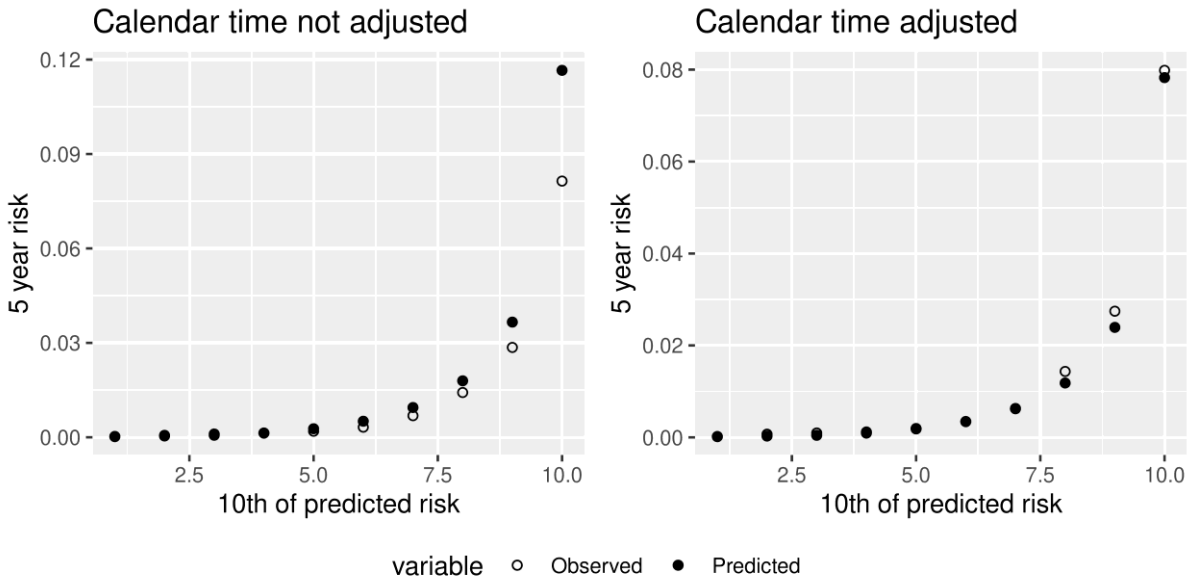
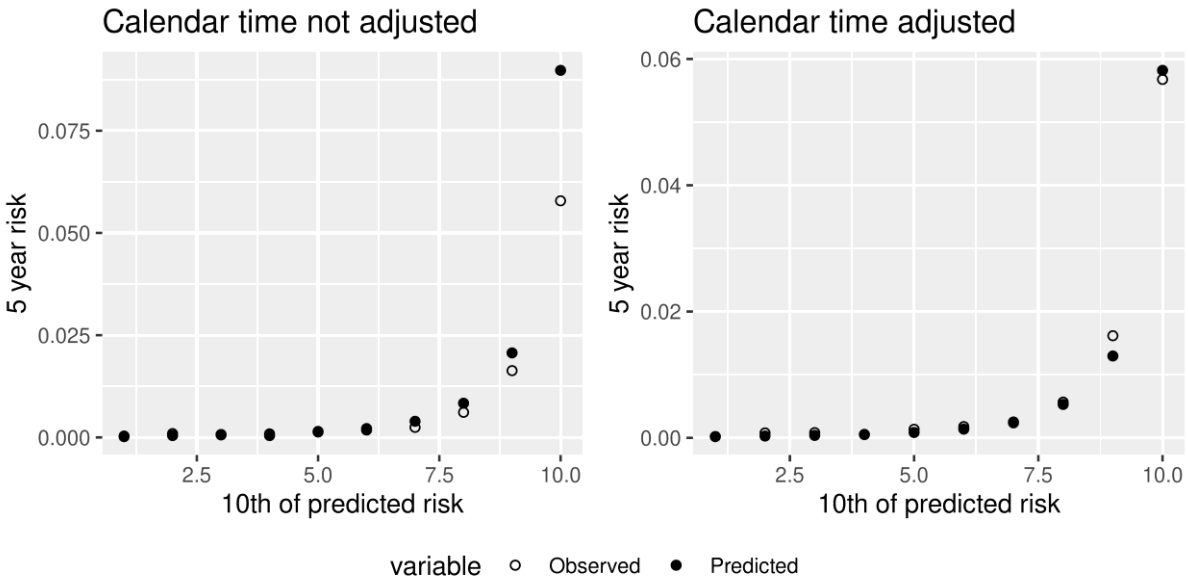


Figure A.5.4: *Calibration of the interval censored Cox model in the validation cohort, with or without adjustment for calendar time (female cohort)*



It is worth noting here that after adjusting for calendar time, the calibration is slightly poorer than in section 5.4.3, where standard cox models were used on data at baseline. I hypothesise this is because we were unable to test for fractional polynomials of the continuous variables when running an interval censored Cox model, meaning we re-used the same fractional polynomials from the standard Cox regression carried out in section 5.4.3. Given that the data was imputed differently, these may not be the best fractional polynomials to use. We did compare them to using no fractional polynomials, and the models were better calibrated when including them.

A.6 Chapter 6 appendices

A.6.1 Derivation of statins treatment periods

A.6.1.1 Definition of treatment period

The length of a statin treatment period was calculated using data from the prescription records in CPRD. First the length of each prescription was estimated using information on the quantity prescribed. The quantity variable was present in 99.4% of the recorded prescriptions. Of these, 8% were 7 days long, 57% were 28 days long, 29% were 56 days long and 3% were 84 days long. Given this, the quantity variable was set to missing if the recorded value was > 84 days. All missing values were imputed with the mode, 28 days. This quantity was then divided by the daily dose variable, to give the prescription length. The daily dose variable was present in 93.2% of records. Of these, 99% were 1, therefore we set the daily dose to one for all entries.

When the prescription lengths had been calculated, these were combined into treatment periods. Each patient's first treatment period began on the date of their first prescription. At the end of this prescription, we looked to see if there was another prescription in the following 90 days (defined as the washout period). The treatment period was assumed to be over when the gap between the end of the current prescription and the start of the next prescription was larger than 90 days. When this condition was met, the current treatment period was assumed to be over at the end of the prescription duration. The next treatment period, if it occurred, started on the date of the first prescription after the washout period. A period of 90 days was chosen due to its usage in other studies,[177,219] which allowed us to verify our results, and prescription length in the UK is very rarely longer than 56 days for long-term condition medicines.[220]

A.6.2 Derivation of CVD transition probabilities (p_c) and how they are used to run the simulation

The transition probabilities to CVD represent the probability of having a CVD event each year (these are adjusted when on statin treatment). These are calculated using similar methods that are used in the QRISK lifetime risk models,[161] and are standard methods for lifetime risk models. Using such methods, the corresponding life time risk could be calculated for a patient of any given age and specified 10-year risk.

A.6.2.1 Data used to derive transition probabilities

The primary prevention cohort was used to derive these transition probabilities. This is a cohort of patients who were at risk of CVD and not currently undergoing treatment. Patients were aged 25 – 84, and were excluded from the cohort if they have had a statin prescription or a CVD event prior to their cohort entry date. Cohort entry date was defined as the latest of turning aged 25, and 1 year of valid follow up in CPRD. Patients were censored at last data collection or death.

A.6.2.2 Steps for calculation of transition probabilities, for a patient of a given age, gender and 10 year risk score, r.

Note these probabilities ignore the competing risk of death, as this was factored in separately when running the simulation. The below steps were carried out separately for males and females in order to calculate respective transition probabilities.

Step 1 – Fit a Cox model to the data where age is used as the time scale.

Using age as the time scale would normally mean a patient enters the cohort at birth and is followed up until they either have a CVD event or are censored. However, as no patients in the cohort were younger than age 25, we used age 25 as the start of follow up for each patient. The data was left truncated for patients that began follow up in CPRD at ages > 25. The data was then right censored at the end of follow up in CPRD. The outcome was defined using the same code lists as QRISK3.[5] This means the CVD event probabilities used in the simulation are with respect to the type of event the QRISK3 calculator predicts. Predictor variables used in this model were the same as those used in QRISK3,[5] except that all the age interaction terms were removed, as the relationship between age and CVD was now being

modelled in a non-linear fashion through the baseline hazard function (which is a step function).

A standard Cox model can be fitted to data in this form. The only change we made to the data was to round all age values to the nearest 0.01, to make the results easier to handle, and to remove random variation in the size of steps from each age category.

Step 2 – Derive the baseline cumulative hazard function and reduce to by year

The coxph function in R was used to fit this model and derive the baseline cumulative hazard function. We were only interested in the probability that an individual has an event in a given year therefore we extracted the baseline cumulative hazard at yearly intervals. These values are referred to as $haz_i, for\ i\ \in\ 25 \leq i \leq 89$. The nature of a Cox models mean if a patient is at a higher or lower risk, the baseline hazard is multiplied by the hazard ratio for that patient, to get their individual cumulative hazard. We utilised this property to generate risk profiles for patients of the same age, with different risks.

Step 3 – Calculating the baseline risk of an event between age1 and age2, given a patient has reached age1

From this point onwards we use the terms conditional and marginal probabilities. The marginal probability, p_m, refers to the risk of a having an event at given age, whereas the conditional probability, p_c, refers to the risk of having an event at a given age, assuming the patient has lived to this age without a CVD event. The marginal risk of someone aged 89 is therefore much lower than the conditional risk, as there is a high chance they will have an event before the age of 89.

The survival probabilities for the cohort, $surv_i, for\ i\ \in\ 26 \leq i \leq 90$, represent the probability of surviving to age i without having had a CVD event, $P(A > i), for\ i\ \in\ 26 \leq i \leq 90$. These can be calculated as:

$$surv_i = \exp(-haz_i)\,, for\ i\ \in\ 26 \leq i \leq 90$$

where haz_i is the baseline cumulative hazard. The baseline risk, probability of having an event before reaching age i, can be calculated as,

$$risk_i = 1 - surv_i, for\ i\ \in\ 26 \leq i \leq 90$$

These are standard survival analysis relationships.

The marginal risk of having an event at a given age, i, is $p_{m,i} = risk_i - risk_{i-1}, for\ i\ \in\ 26 \leq i \leq 90$

The conditional risk of having an event in each year is then calculated as:

$$p_{c,i} = p_{m,i}\ for\ i = 26$$

$$p_{c,i} = \frac{p_{m,i}}{\prod_{j=1}^{i-1}(1 - p_{c,j})}\ for\ i \in\ 27 \leq i \leq 90$$

This comes from the fact that the marginal probability of having an event at a given age, is equal to the product of not having an event in any of the subsequent years, multiplied by the probability of having an event in that year of interest.

Finally, we can use these conditional probabilities to calculate the cumulative risk of an event between any two ages, $risk_{age1-age2}$, conditional on a patient living to that age. To do this, we want to calculate the probability of a 50 year old having an event in each year of follow up, conditional on the patient having lived to 50 years of age. Assuming a patient is alive at age1, we can calculate the marginal probability of an event in each subsequent year, $p_{m,i|age1}$

$$p_{m,i|age1} = p_{c,i} \prod_{j=age1}^{i-1} (1 - p_{c,j}), for\ i\ \in\ age1 \leq i \leq age2$$

For each year, this is taking the product the conditional probabilities of not having an event in each year, starting at age1, up to the age of interest – 1, and multiplying this with the conditional probability of having an event in that year of interest.

The risk of having an event between age1 and age2, given a patient has reached age1, is the sum of these quantities:

$$risk_{age1,age2} = \sum_{k=age1+1}^{age2} p_{m,k|age1}$$

However note that this is the risk between any two ages, for someone with the baseline cumulative hazard.

Step 4 – Calculating the hazard ratio to give a specific 10 year risk for a patient of a given age

For this simulation, we needed to calculate the lifetime risk for someone with a given age, and a specific 10 year risk. To do this, we had to calculate the hazard ratio which gave us the required 10 year risk, for someone of a given age.

Suppose we wanted to consider the scenario where a patient of is of age = age1, and 10 year risk = r. To do this we go back to start of step 3 and multiply the baseline cumulative hazard by HR, such that

$$risk_{age1,age1+10} = r$$

The HR which satisfies the above equation is what we solved for.

Step 5 – Calculating the conditional probabilities of an event in each year of follow up for a patient with a given 10 year risk

After step 4, when we have the HR which gives a 10 year risk of r, we retained the conditional probabilities associated with this HR. These are the $p_{c,i}, for\ i \in\ age1 \leq i \leq 90$, such that $risk_{age1,age1+10} = r$

These conditional probabilities of an event in each year, $p_{c,i}$, were used in the simulation. The patients 10 year risk will be equal to r, the required risk for the scenario of interest. If required, we could also calculate the risk between age1 and any age2, $risk_{age1,age2}$, using the method outlined in step 3.

Summary

It should be noted that the way in which we adjust the cumulative hazard function, in order to obtain 10 year risks of interest, is exactly how the lifetime risk model is adjusted in practice to identify patients of higher and lower risks. Proportional hazards are assumed in this scenario (how true this is in practice I very much doubt), and patients with higher and lower risks have their cumulative hazard function adjusted by a particular hazard ratio. While in practice that hazard ratio is calculated based on a patient's predictor variables, we just picked a hazard ratio in order to obtain the cumulative hazard function that is relevant to each scenario in the simulation.

A.6.2.3 Implementation of simulation

When the CVD event probabilities had been derived we allowed patients to move through the simulation in the following manner. Each scenario was based on age and 10-year CVD risk at the start of the simulation (risk assessment), gender, age at statin initiation and an assumed discontinuation and restarting rate.

First we calculated the CVD transition probabilities for a patient with a given age, gender and 10-year risks, using the methods from section A.6.2.2. Then we generated discontinuation and restarting times of statin treatment over the duration of follow up, based on the statin initiation date and the discontinuation and restarting rate. Next we adjusted the CVD transition probabilities each year if on statin treatment during that year, as outlined in the main manuscript. If a patient was on statin treatment for x% of a year, we reduced the treatment effect by this amount. For example if we assumed the relative risk when on statin treatment was 0.7, and a patient was on treatment for half a year, we only applied a risk reduction of 0.85.

We were then left with the final set of CVD transition probabilities each year, and probabilities of death each year (taken directly from ONS data). We simulated a CVD event time and death for a patient using these probabilities. We repeated this process 10,000 times for each scenario, and counted the number of CVD events that occurred prior to death over the course of the follow up. We compared the number of CVD events with the number occurring in a simulation where it was assumed each patient received no statin treatment for the duration of follow up, to get the number of events prevented per 100 people.

A.6.3 Calibration plots of Cox models from which transition probabilities were derived

The following section presents internal calibration plots for the Cox model used to derive the CVD transition probabilities, which is outlined in section A.6.2.2. It was important these models were well calibrated.

A.6.3.1 Methods

The process for developing the Cox model (age as time scale) was presented in section A.6.2.2. To produce the following plots, this model was developed on 80% of the primary prevention cohort and 20% retained for validation. 10-year risk scores were then generated for patients in the validation cohort. Patients were then allocated into 10 equal size groups based on their predicted risk (deciles). Within each group, the Kaplan Meier estimate of risk at 10 years follow up was calculated (observed risk) and the average predicted risk (expected risk). These were then plotted against each other to assess calibration.

A.6.3.2 Results

The calibration of the respective male and female models is shown in Figure A.6.1 and Figure A.6.2. While these calibration plots are optimistic (split sample internal calibration), they indicate both models were well calibrated.

Figure A.6.1: *Calibration of female Cox model used to derive CVD transition probabilities*

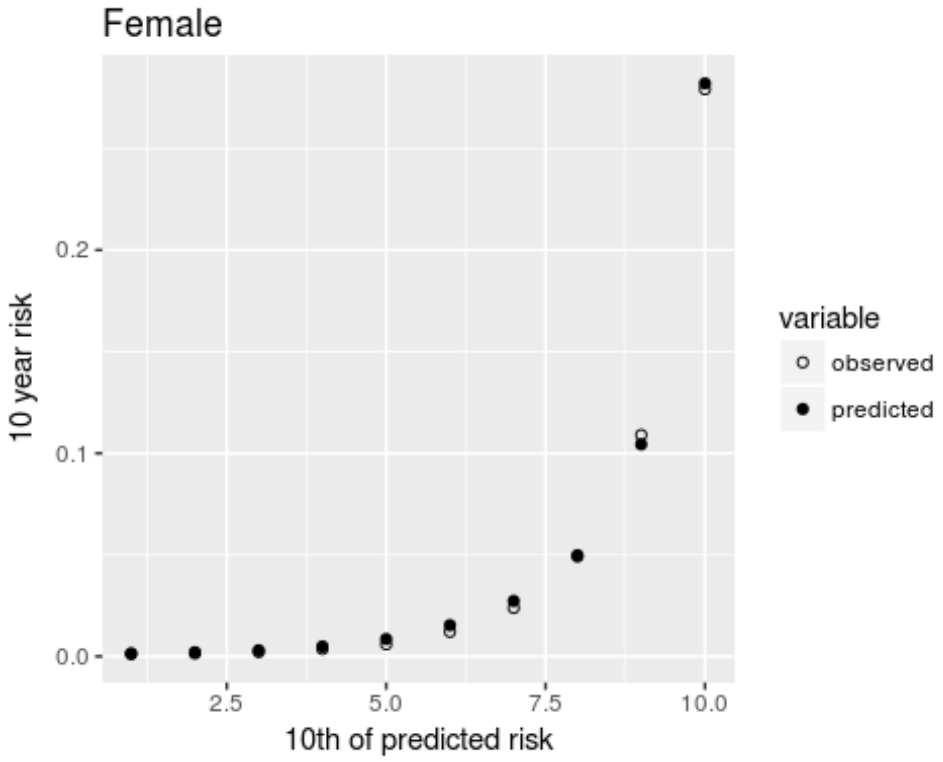
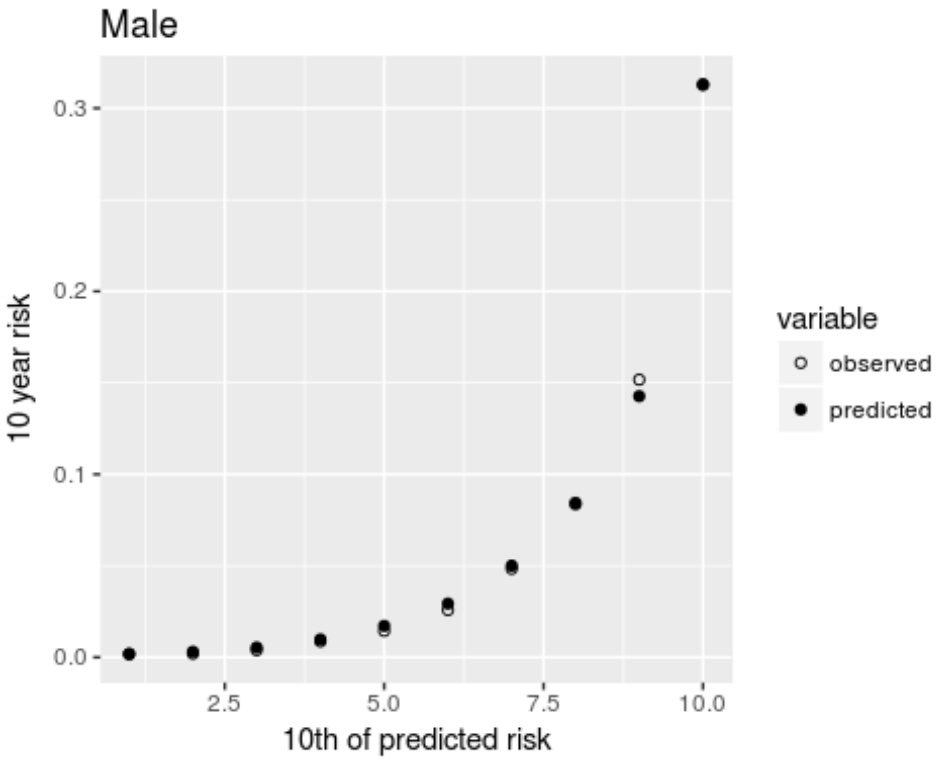


Figure A.6.2: *Calibration of female Cox model used to derive CVD transition probabilities*



A.6.4 Age stratification and extrapolation of discontinuation rates

A.6.4.1 Age stratification of discontinuation rates for first treatment period

Figure A.6.3 shows the discontinuation rates (Kaplan Meier plots) of patients during the first statin treatment period stratified by age (10 year age groups). We see there are fairly large differences in the discontinuation rates for different aged patients. We could not use these discontinuation rates in our simulation directly as we varied statin initiation age by one year intervals, and therefore need the discontinuation rate to be a function of age where age could take any integer.

To do this we opted to fit a Cox proportional hazards model to the data with age as a predictor variable, and tested for fractional polynomials of age to allow a non-linear relationship between age and discontinuation rate. We felt this model was suitable as the survival curves presented in Figure A.6.3 are proportional up until the point where data starts to run low (i.e. not many age 70 – 80 year old patients have more than 5 years follow up on treatment), making the proportional hazards assumption viable. We felt it was appropriate to test the

proportional hazards assumption visually, as any formal test for the proportional hazards would find the assumption not to be true due to the large sample size (high power to detect any difference) and length of follow up is so long (no data is truly proportional over this length of time).

We found the optimal fractional polynomial of age to be: β_1*(age/100)^3 + β_2*((age/100)^3)*log(age/100). We then fitted the Cox model and discontinuation rates could be generated for any age. Figure A.6.4 shows the discontinuation rates for age = 45, 55, 65 and 75 from this model, for comparison to the age stratified discontinuation rates presented in Figure A.6.3. We felt the agreement between the two graphs was strong.

Figure A.6.3: *Kaplan Meier plots for patients during first treatment period, stratified by age*

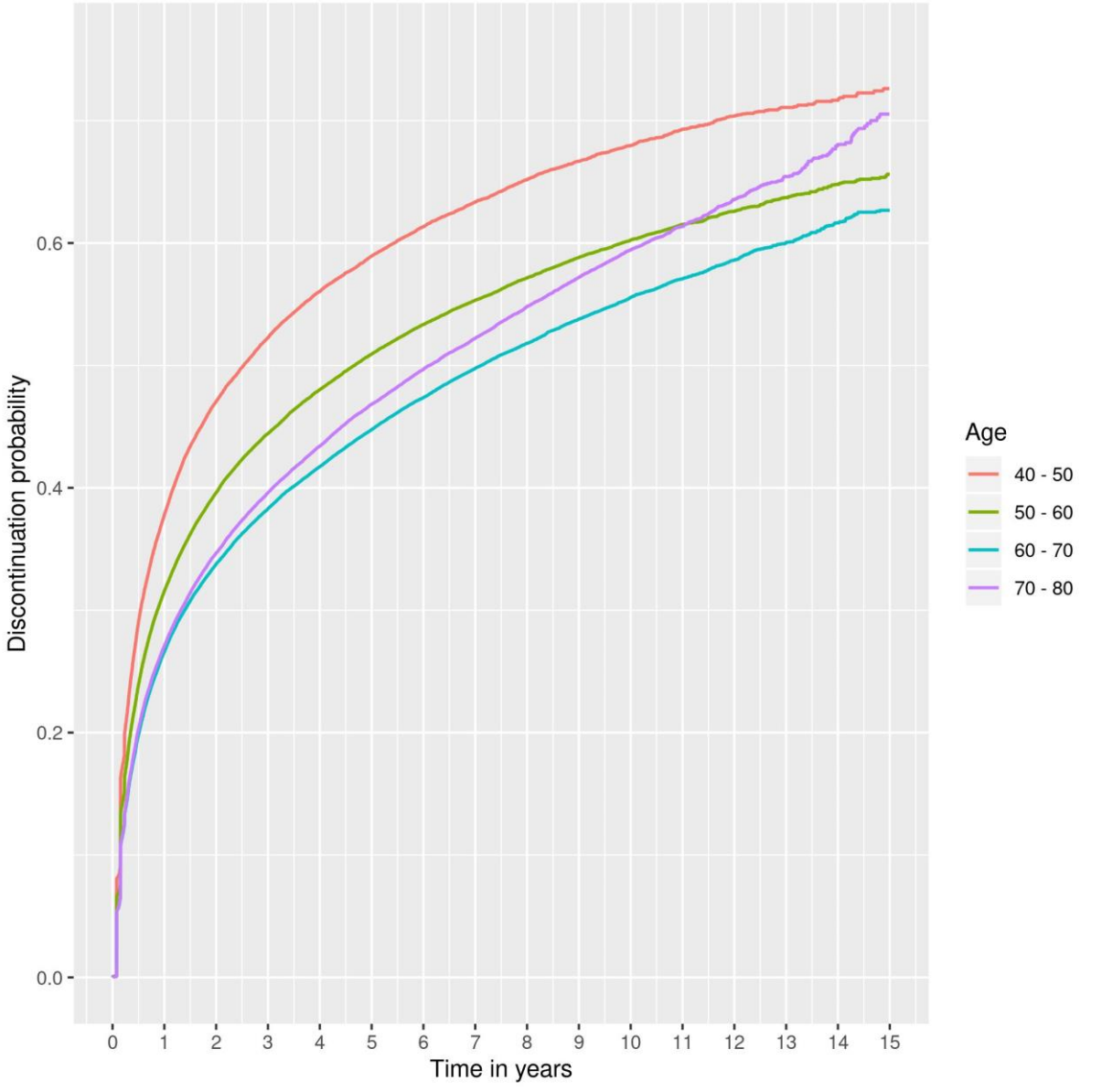
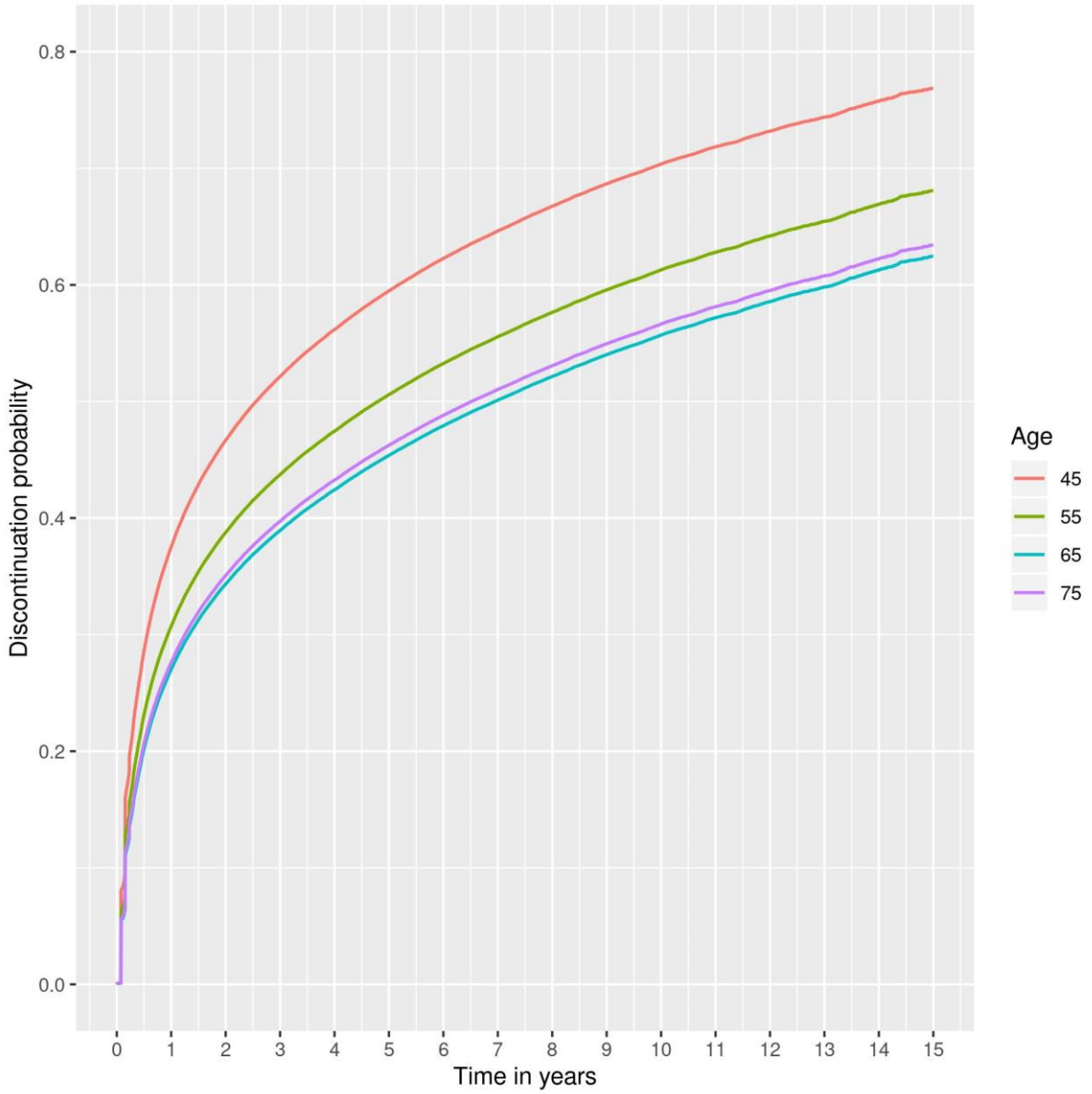


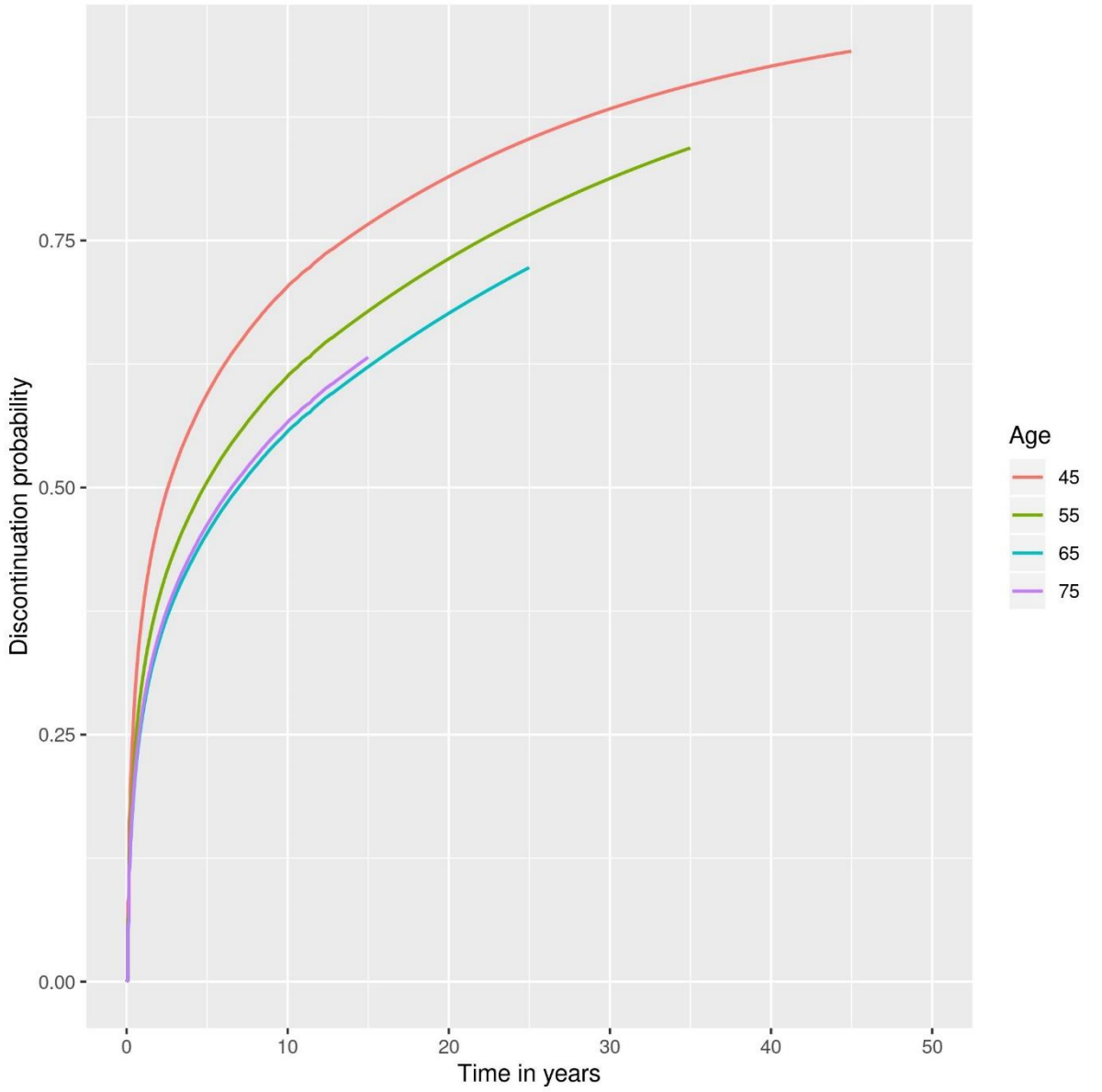
Figure A.6.4: *Age stratified discontinuation rates derived from the Cox proportional hazards model*



A.6.4.2 Extrapolation of discontinuation and restarting rates

The discontinuation/restarting rates were extrapolated using a constant rate from year 13 (first discontinuation), year 10 (second treatment period, and first and second restarting periods) and year 8 (third treatment period) onwards. The discontinuation probability used in each day beyond the cut-off point was calculated as the mean daily discontinuation probability of the final year prior to the cut-off point. To follow are Figures of the extrapolation of the discontinuation and restarting rates. For the first treatment period (Figure A.6.5), we plot the extrapolation of the discontinuation rate for a range of ages on the same graph, as this discontinuation rate is age stratified. The extrapolated second and third discontinuation rates are then plotted together (Figure A.6.6), and the extrapolated first and second restarting rate (Figure A.6.7).

Figure A.6.5: *Extrapolation of the discontinuation rates derived from CPRD for the first treatment period, stratified by age, extrapolation made from year 13 onwards.*



**Note we only need to extrapolate as far as 90 years old, hence the different extrapolation lengths*

Figure A.6.6: *Extrapolation of the discontinuation rates derived from CPRD for the second and third treatment periods, extrapolation made from year 10 and year 8 onwards.*

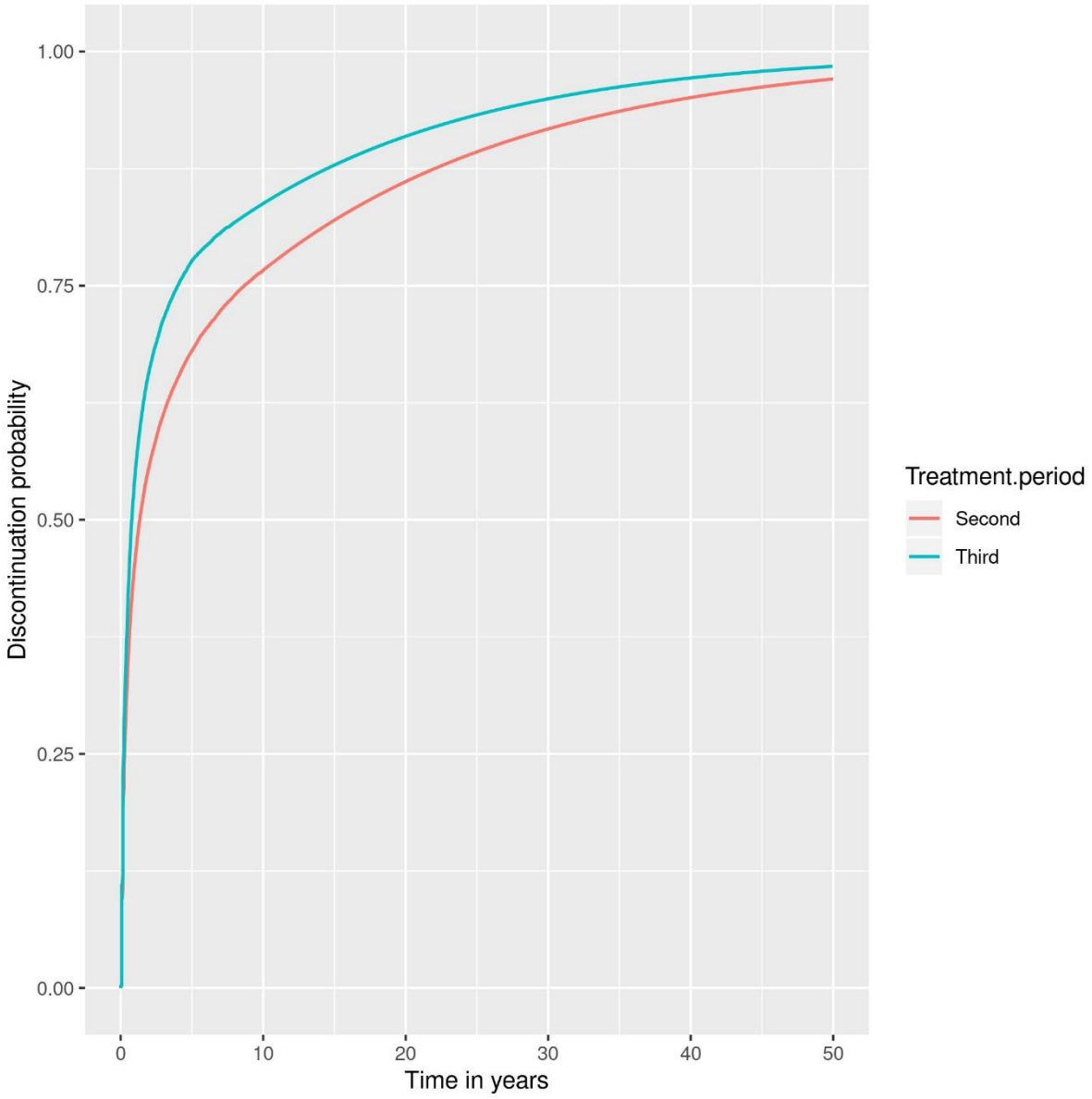
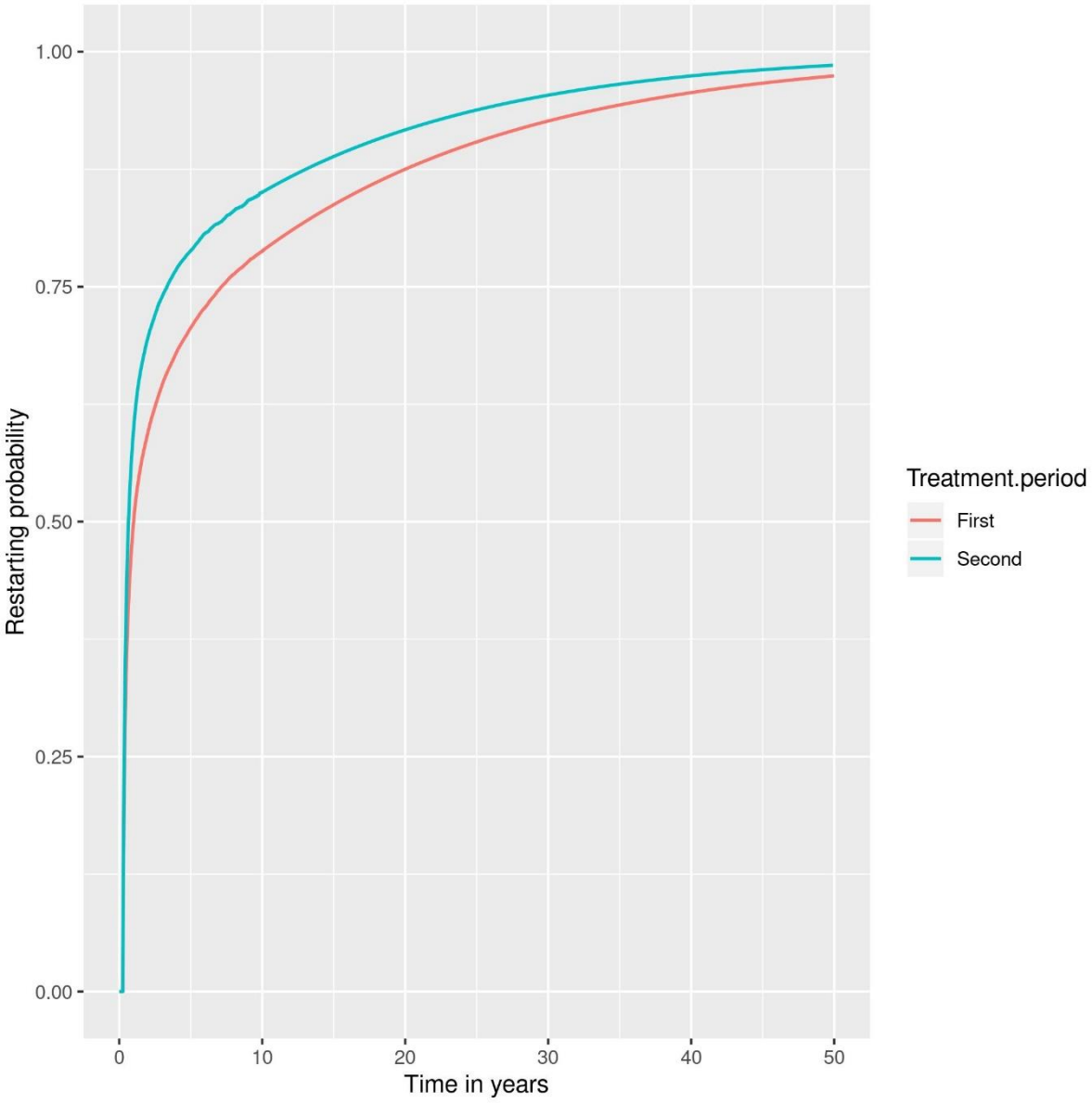


Figure A.6.7: *Extrapolation of the restarting rates derived from CPRD for the first and second treatment period. Extrapolation made from 10 years onwards.*



A.6.5 Discontinuation and restarting rates in the cohort of statin users where treatment periods with only one prescription are removed

The Kaplan Meier plots in Figure A.6.8 and Figure A.6.9 show the discontinuation and restarting rates for each treatment period, for the cohort of patients used in the sensitivity analysis, where treatment periods of length 1 are removed. These plots are the equivalent of Figure 6.2, but for this cohort.

Figure A.6.8: *Kaplan Meier plots of the time until discontinuing statins for the first, second, third and fourth discontinuation, single prescription treatment periods removed*

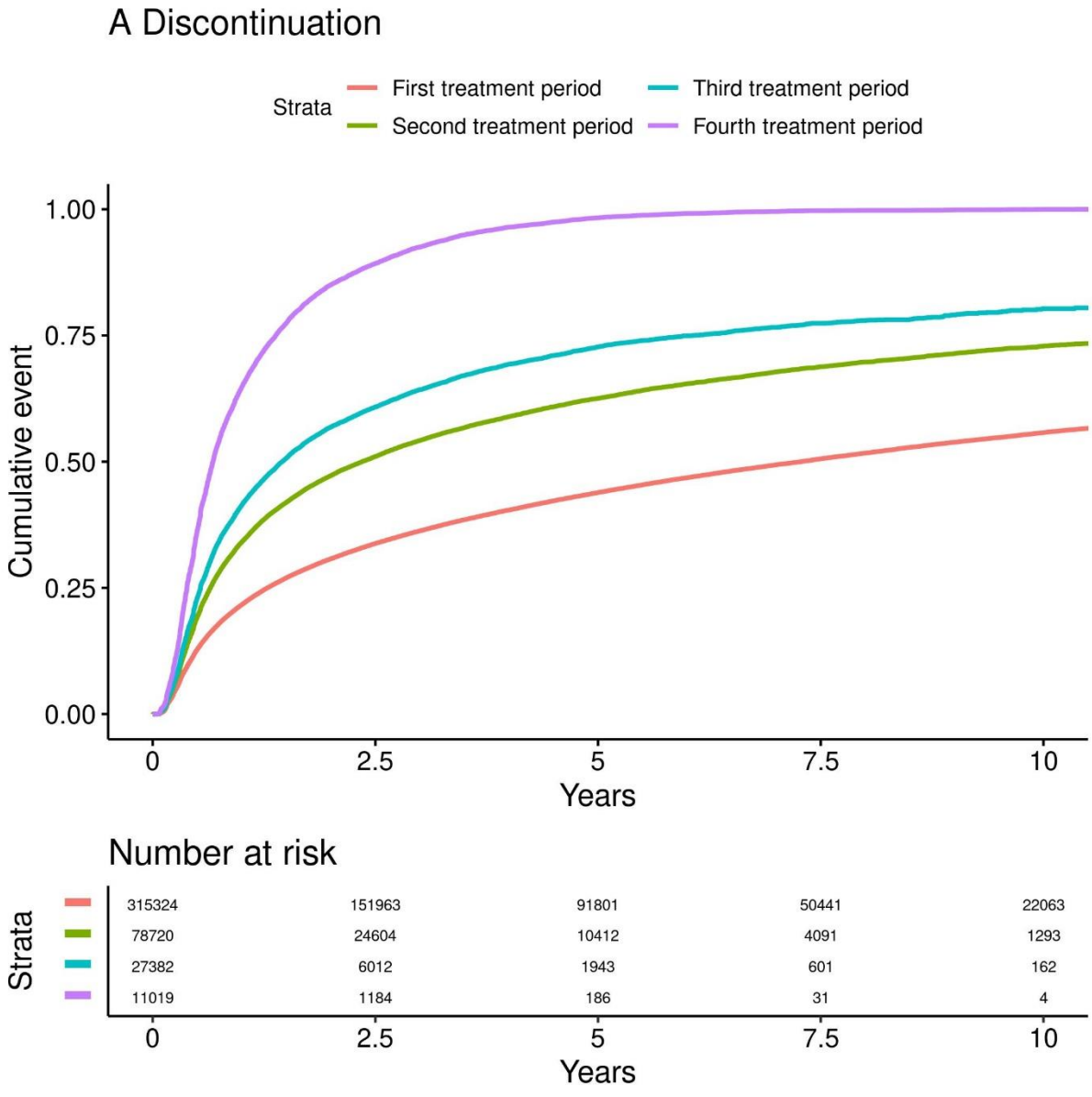
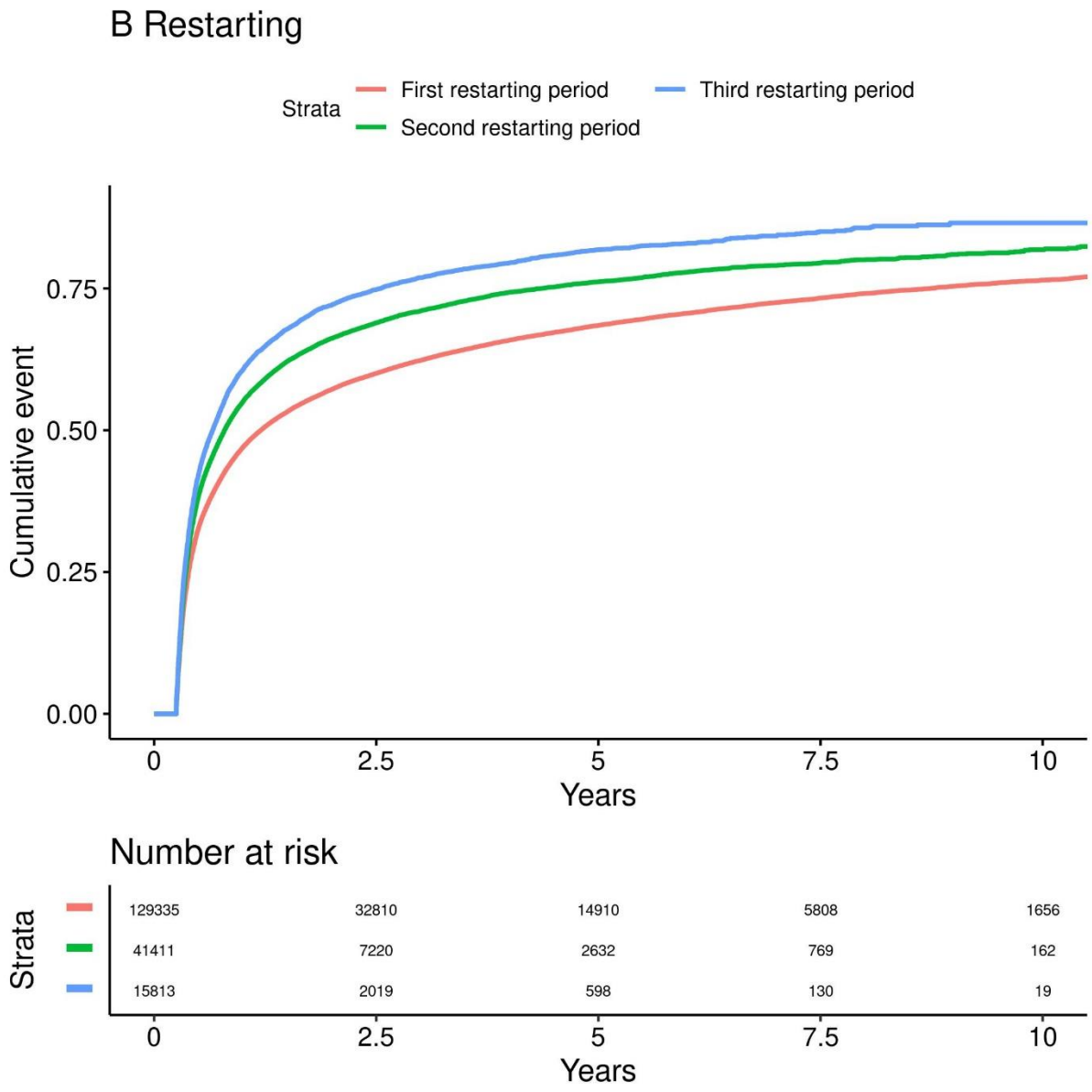


Figure A.6.9: *Kaplan Meier plots of the time until restarting statins for the first, second and third time, single prescription treatment periods removed*



A.6.6 Results from the female analysis and the sensitivity analyses

The results from the primary analysis for the female cohort are presented here, in Figure A.6.10 and Figure A.6.11. The results from all the sensitivity analyses are available on the GitHub page,[103] as they are numerous and provide little extra context.

The primary analysis for the female cohort brings similar conclusions to the male cohort. The peak of the trajectories was driven by age, rather than risk score, although the peak is shifted by about five years to just after 60. This reflects female's lower risk of death. There are slightly larger gains to be made by delaying statin initiation until risks higher than 10% than there was for the men.

The sensitivity analyses echo the findings from the primary analysis. Reducing the relative rate (increasing the treatment effect) caused a higher number of events prevented, and therefore greater gains to be made by delaying statin initiation. However the shape of the trajectories remained the same and the maxima was around the same age. When the cohort of statin users excludes treatment periods with only one prescription, the events prevented increased slightly but the change is not large. Once again, the maxima of the trajectories were at a similar point.

Supplementary illustrative example 1: If we prescribe statins to a cohort of 50-year old women with a 10% 10-year CVD risk, we prevent 4.53 events per 100 individuals over the course of 40 years. If we took this same cohort of women, but instead waited 10 years before initiating statins, at which point there 10-year risk of CVD would be around 20%, then we would prevent 5.49 events per 100 individuals over the 40 years period of follow up.

Supplementary illustrative example 2: Consider prescribing statins to a cohort of 50-year old women with a 4% 10-year CVD risk (median for that group). Per 100 individuals, 3.64 events are prevented if discontinuation rates remain as normal, 5.79 events if discontinuation is reduced by a sixth, 7.24 events if discontinuation is reduced by a third, 8.25 events if discontinuation is halved, and 9.47 events if there is no discontinuation. The equivalent number of events prevented for a cohort with 10% 10-year CVD risk are 4.54, 7.21, 9.05, 10.27 and 11.83.

Figure A.6.10: *Number of cardiovascular events prevented over the duration of follow up with different time delays in starting statins, stratified by baseline age and 10 year CVD risk using the discontinuation rates as observed in the statin cohort (female cohort)*

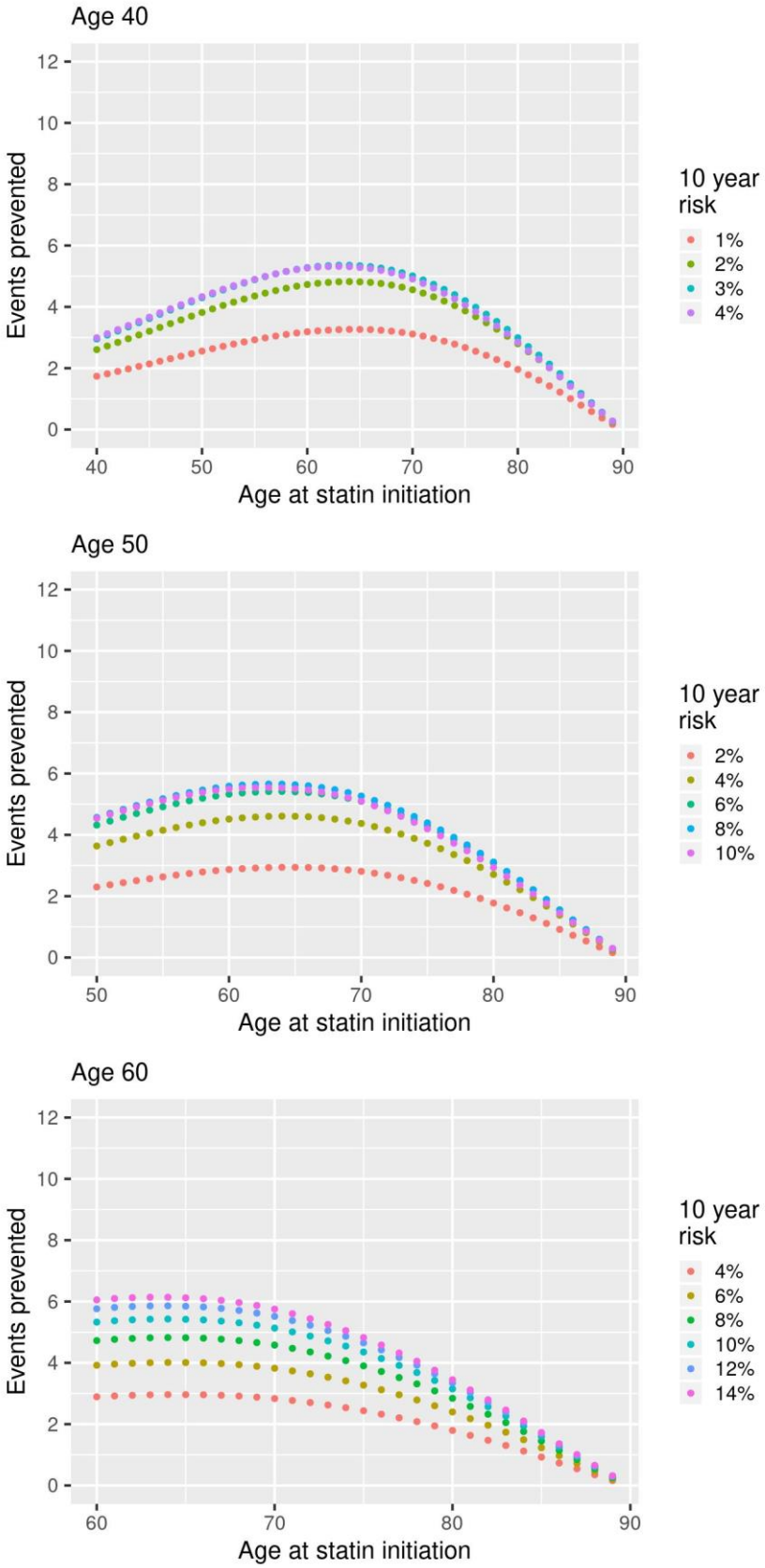
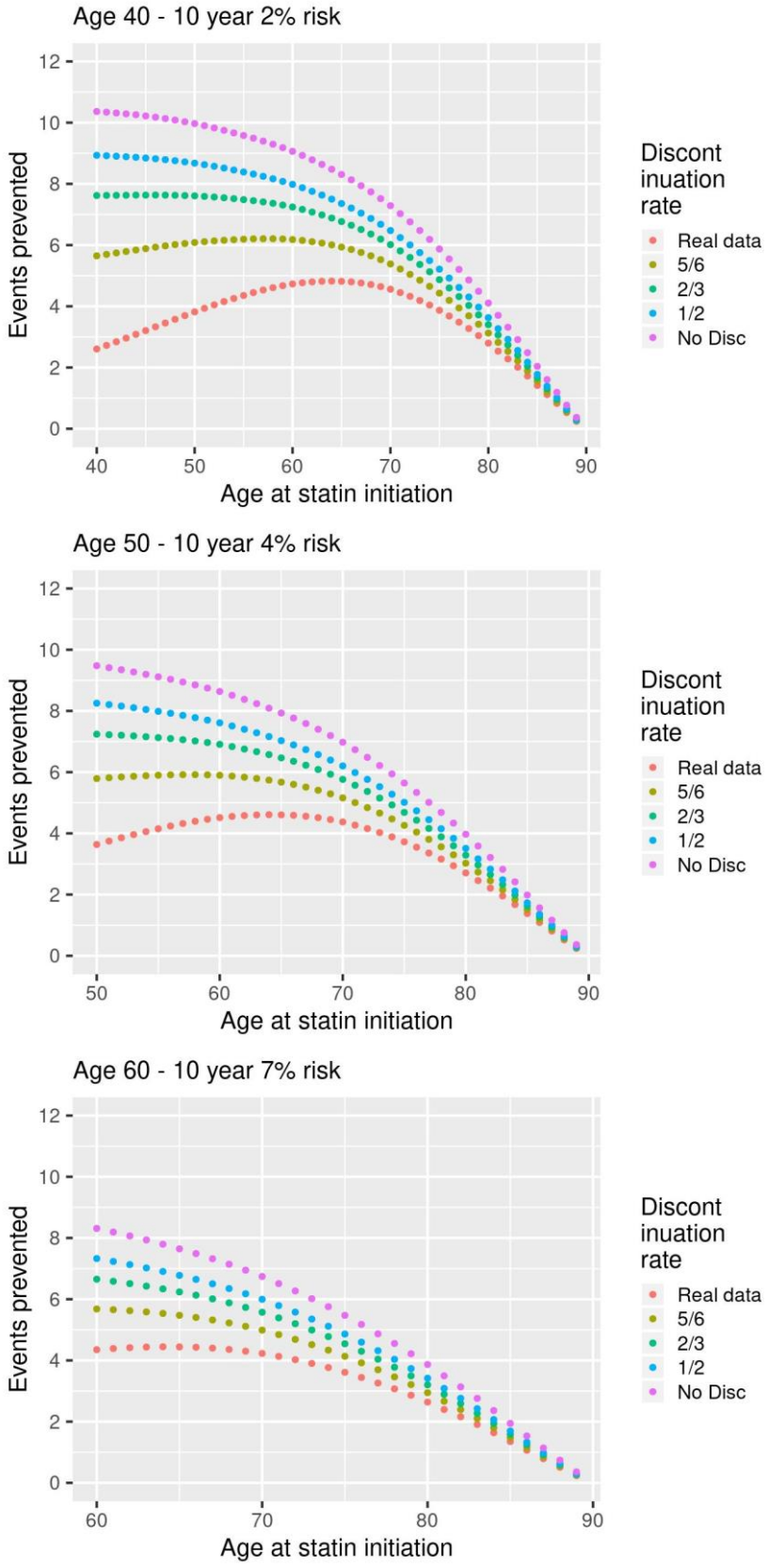


Figure A.6.11: *Number of cardiovascular events prevented over the duration of follow up with different time delays in starting statins, stratified by baseline age and discontinuation rate (female cohort)*



A.7 Chapter 7 appendices

A.7.1 Density and convergence plots of imputed variables

A.7.1.1 Methods

Amount of missing data

The levels of missing data were as follows: cholesterol/HDL ratio [17.56% for females and 16.91% for males], SBP [1.60% and 2.26%], SBP variability [6.26% and 9.77%], Smoking [9.71% and 8.50%] and BMI [18.44% and 20.65%]. Missing data in Ethnicity was combined with white to create a 'white or not stated' category, as is the case in QRISK3.

Methods for running the imputation

Multiple imputation by chained equations was used to impute missing data for body mass index (BMI), systolic blood pressure (SBP) and SBP variability, cholesterol, HDL and smoking status. The program used to impute the data was the R package 'mice'.[130] There were 20 imputation procedures carried out, and 30 iterations for each one. Variables included in the imputation model were all predictor variables required to produce a risk score using QRISK3 (including interaction terms). All continuous variables were imputed using predictive mean matching, and polytomous regression for categorical variables.[130] Interactions terms were imputed empirically from the two component variables (not stochastically), and interactions terms were not used to impute their component variables.

Methods for assessment performance of imputation process

For continuous data the density plots shown assess whether there were any systematic differences in for the non-missing data and the imputed data. This also enabled us to check that the distribution of imputed values was reasonable (i.e. no extreme values, or a distribution shape which clearly indicates an issue with the imputation procedure). In the plots, each red line is a density plot of the imputed data in one of the imputed datasets, and the blue line is the density plot of the non-missing data.

The convergence plots assess whether the Markov chain in the imputation process had reached a steady state by the final iteration. The x-axis is the iteration number, y-axis the

mean or standard deviation of the imputed values, and each coloured line a different imputation process. For categorical variables, the distribution of the variable from each imputation stream are presented, as well as the distribution of non-missing values.

A.7.1.2 Results for imputation of statin initiation cohort

All convergence plots reached a steady state very quickly, far before the 30th iteration. All density plots had reasonable distributions with no extreme values. All plots presented below.

BMI

Figure A.7.1: *BMI convergence plot for imputation of statin initiation cohort*

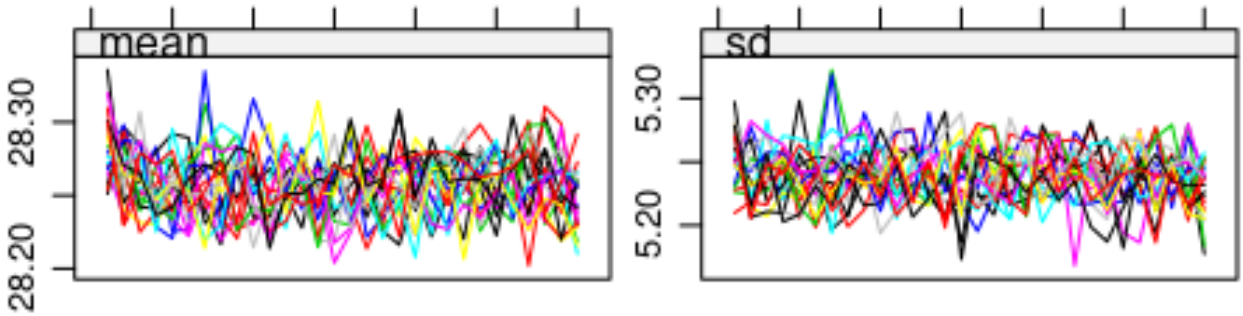
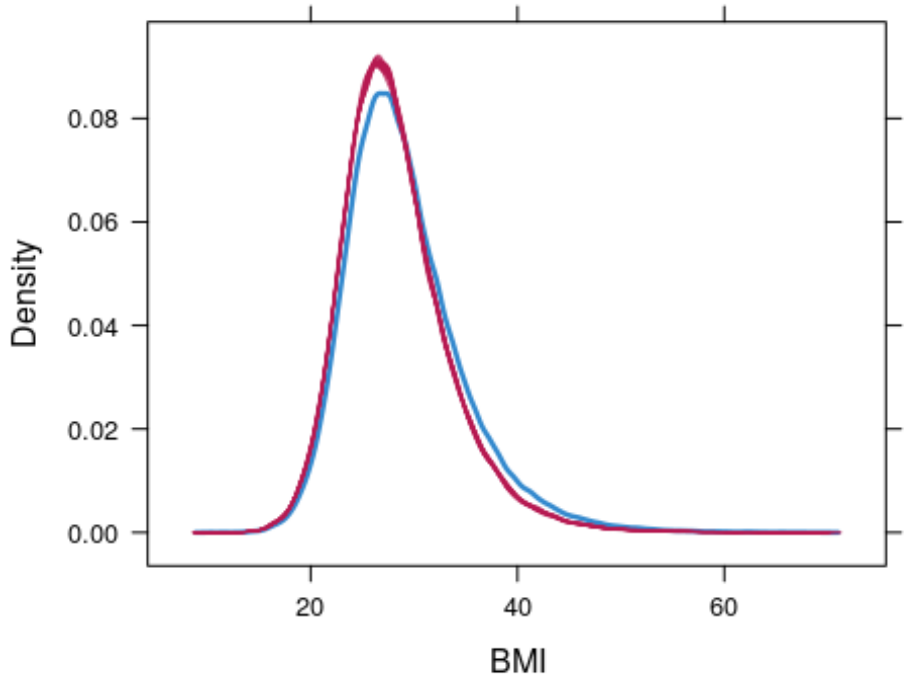


Figure A.7.2: *BMI density plot for imputation of statin initiation cohort*



SBP

Figure A.7.3: *SBP convergence plot for imputation of statin initiation cohort*

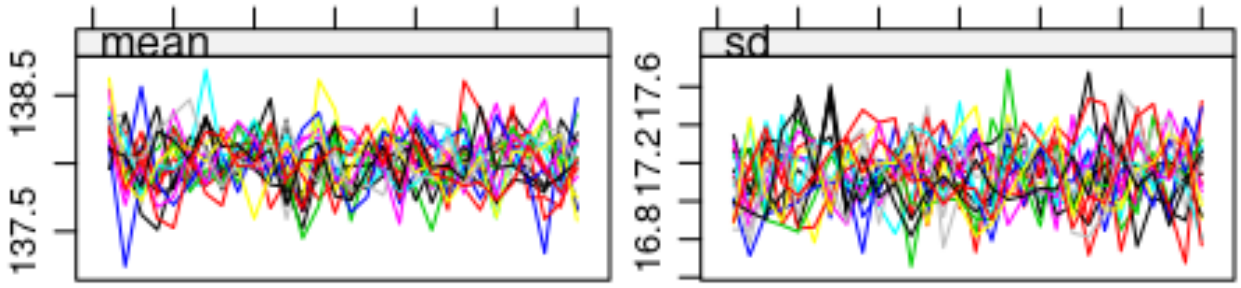
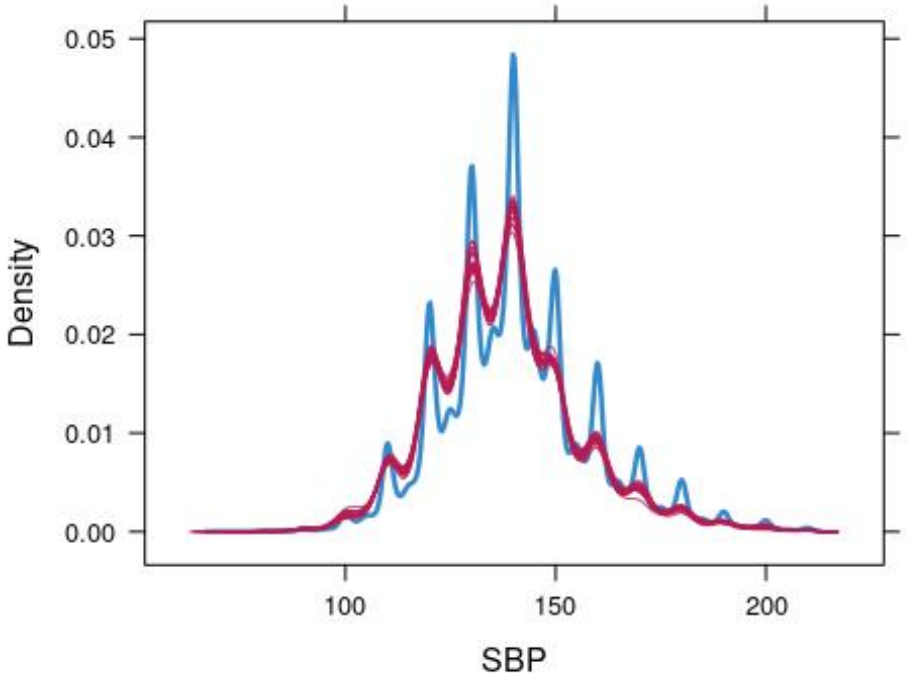


Figure A.7.4: *SBP density plot for imputation of statin initiation cohort*



SBP variability

Figure A.7.5: *SBP variability convergence plot for imputation of statin initiation cohort*

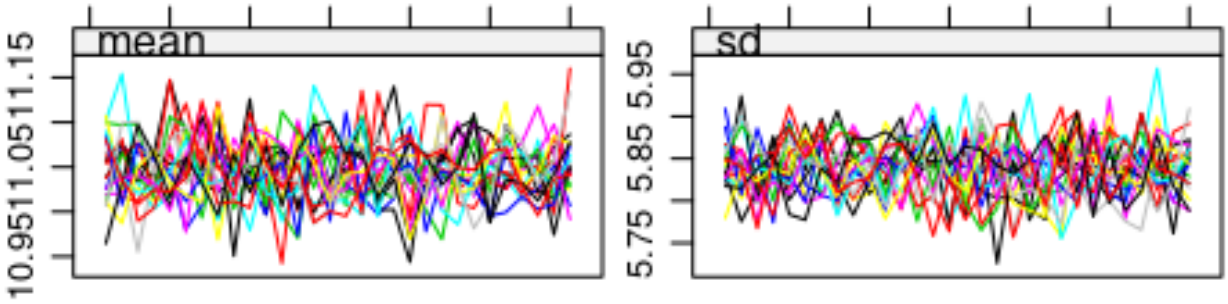
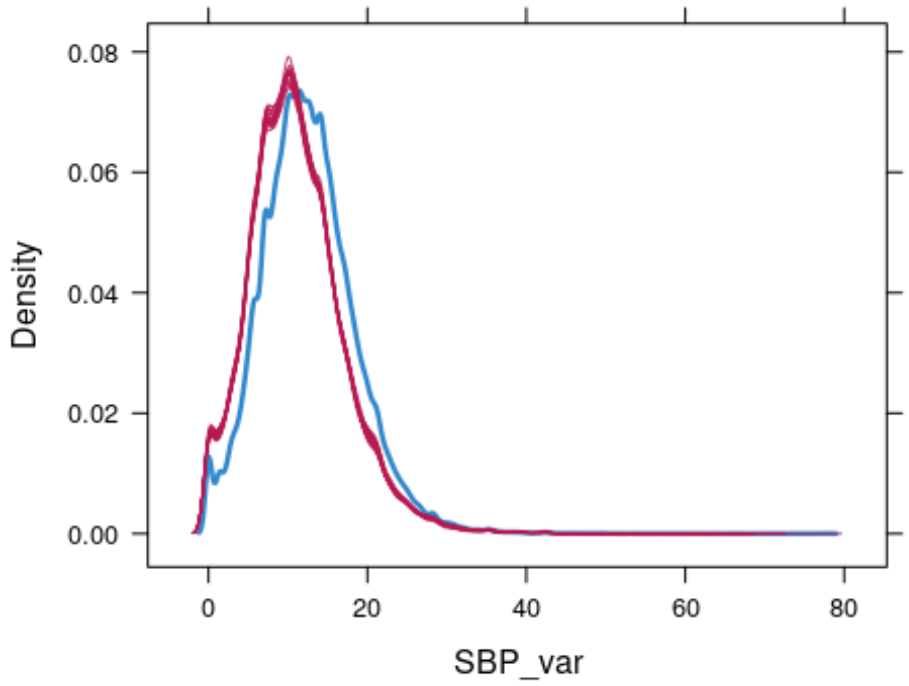


Figure A.7.6: *SBP variability density plot for imputation of statin initiation cohort*



Cholesterol

Figure A.7.7: *Cholesterol convergence plot for imputation of statin initiation cohort*

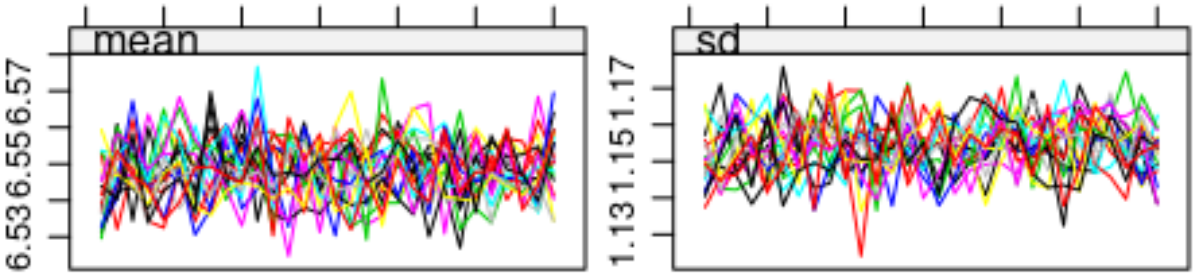
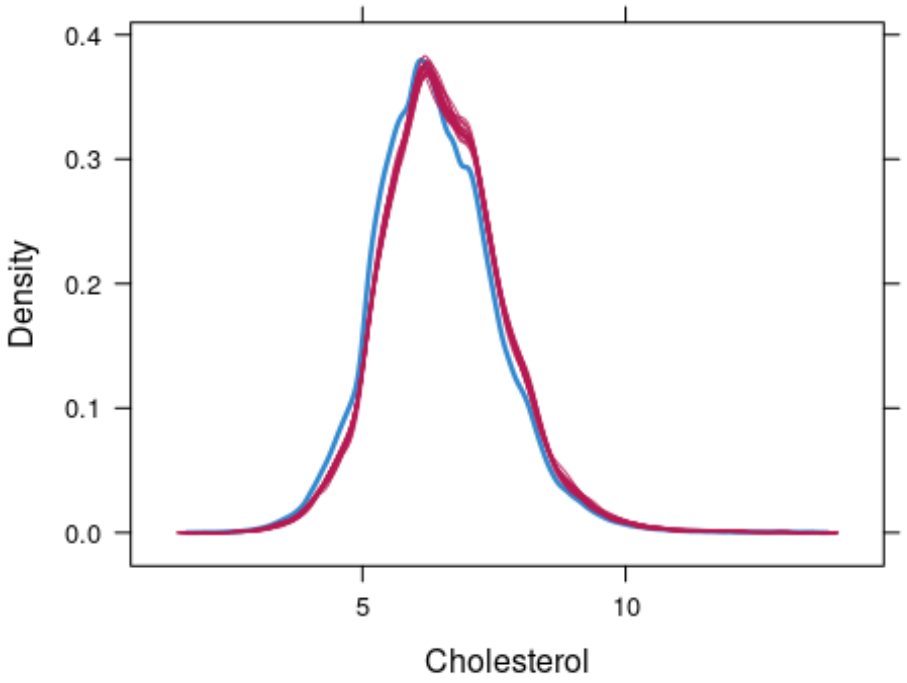


Figure A.7.8: *Cholesterol density plot for imputation of statin initiation cohort*



HDL

Figure A.7.9: *HDL convergence plot for imputation of statin initiation cohort*

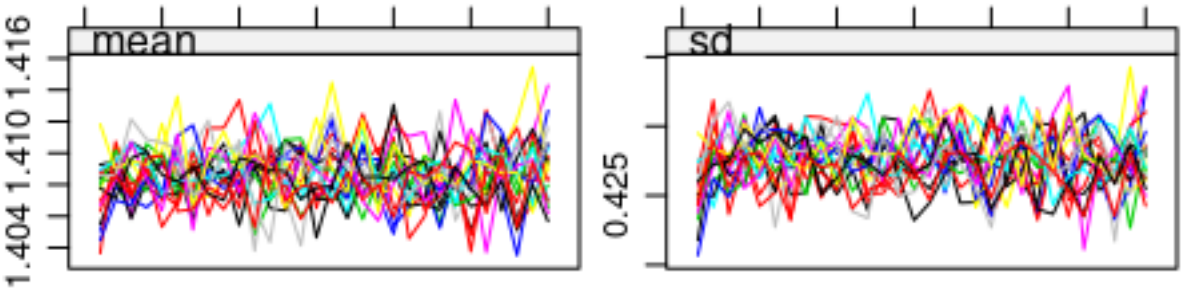
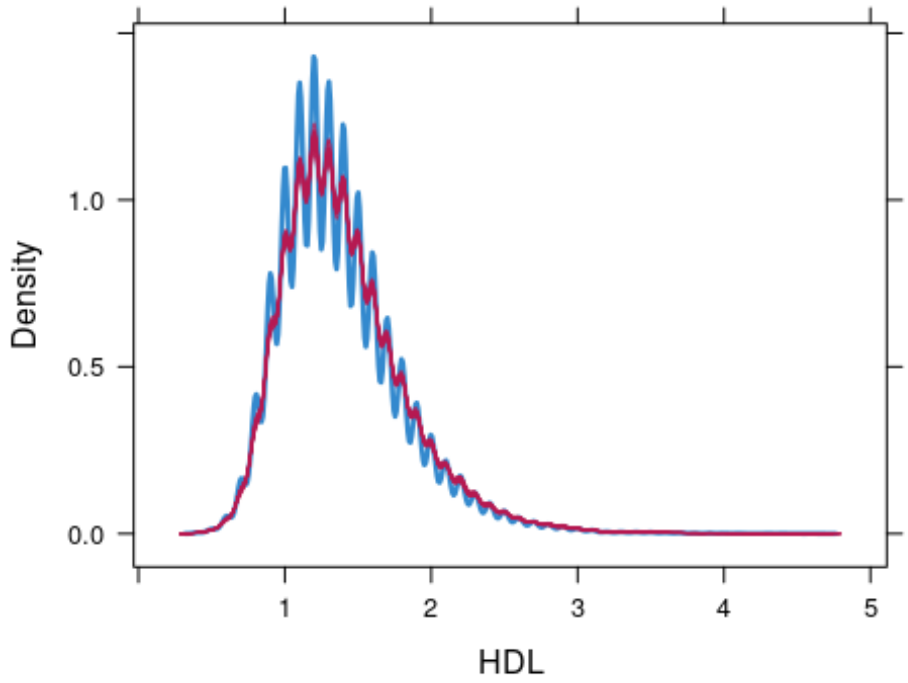


Figure A.7.10: *HDL density plot for imputation of statin initiation cohort*



Smoking status

Table A.7.1: *Distribution of real data and imputed data (%) for smoking status in imputation of statin initiation cohort*

	Smoking status (%)				
Imputation	Never	Ex	Light	Moderate	Heavy
Real data	38.92	35.85	9.75	8.06	7.43
1	37.58	41.82	7.19	7.26	6.16
2	37.92	40.82	7.10	7.58	6.58
3	37.00	41.55	7.39	7.95	6.11
4	36.93	41.71	7.64	7.37	6.35
5	37.12	41.63	7.04	7.57	6.64
6	37.30	41.29	7.14	7.78	6.49
7	37.04	41.47	7.50	7.92	6.07
8	37.36	41.72	7.26	7.45	6.20
9	37.01	40.88	7.43	8.02	6.66
10	37.46	41.52	7.29	7.26	6.47
11	36.75	42.54	7.00	7.72	5.99
12	36.36	42.05	7.69	7.84	6.06
13	37.30	41.40	7.20	7.74	6.36
14	37.42	41.06	7.51	7.57	6.44
15	37.48	41.46	7.26	7.79	6.01
16	37.79	41.10	7.15	7.11	6.84
17	37.19	41.50	7.16	7.68	6.47
18	37.20	41.92	7.25	7.45	6.19
19	37.33	41.59	7.45	7.57	6.07
20	37.13	41.57	7.34	7.58	6.38

A.7.2 Supplementary tables and figures

Table A.7.2: *Baseline demographics of statin cohort*

	Female	Male
N	166,209	185,344
Continuous variables		
Age	63.5 (11.05)	60.08 (11.08)
Systolic blood pressure	140.33 (18.35)	140.61 (17.2)
Systolic blood pressure variability	13.07 (5.8)	12.12 (5.89)
Body mass index	29.26 (6.35)	28.96 (5.05)
Cholesterol/HDL ratio	4.64 (1.42)	5.21 (1.53)
Categorical variables		
Atrial fibrillation	2.85%	3.62%
Atypical antipsychotic medication	0.86%	0.76%
Corticosteroid use	2.00%	1.23%
Chronic kidney disease stage 3/4/5	13.61%	7.17%
Diabetes (type 1)	1.32%	1.70%
Diabetes (type 2)	21.19%	22.25%
Ethnicity: Bangladesh	0.13%	0.16%
Black African	0.40%	0.40%
Black Caribbean	0.44%	0.34%
Chinese	0.13%	0.11%
Indian	0.90%	1.06%
Other	0.79%	0.83%
Other Asian	0.59%	0.64%
Pakistani	0.32%	0.39%
White	96.30%	96.08%

Family history of CVD	29.49%	23.22%
HIV	0.04%	0.13%
Treated Hypertension	48.74%	43.70%
Migraine	10.64%	4.41%
Rheumatoid Arthritis	2.10%	0.88%
Smoking: Never	46.47%	32.24%
Ex	30.60%	40.49%
Current	22.93%	27.27%
Systemic lupus erythematosus	0.23%	0.03%
Severe mental illness	15.87%	8.56%