# Cardiovascular disease prediction and screening using genomics

Jasmine Elina Gratton

University College London

Thesis submitted for the degree of Doctor of Philosophy

2022

I, Jasmine Gratton confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

**Abstract**

Polygenic scores, a measure of genome-wide allelic contribution for a trait, have gained attention in the medical research community in recent years and have led to polarised opinions in terms of their clinical importance. A growth in the number and size of genome-wide association studies, enabled by the assembly of large consortia of case-control and cohort studies, and the advent of national biobanks, has led to the discovery of millions of DNA sequence variants associated with thousands of continuous traits of biomedical relevance (e.g. blood pressure) and disease endpoints (e.g. coronary artery disease). This has contributed to the development of thousands of polygenic scores and a heightened interest in their use in disease prediction and screening.

This thesis evaluates the clinical utility of polygenic scores mainly in the context of cardiovascular disease prediction and screening. The poor performance of polygenic scores in disease prediction is first demonstrated by analysing the Polygenic Score Catalog that aggregates many published polygenic scores for various disease endpoints. The incremental predictive utility of polygenic scores to currently used cardiovascular risk prediction tools in the UK, based on non-genetic risk factors (e.g. QScores) is then evaluated for various cardiovascular disease endpoints using the appropriate metrics. The thesis also explores the potential application of polygenic scores for the discovery of individuals more likely to carry rare genetic variants, using the example of familial hypercholesterolaemia (FH), the most common monogenic disease, which is still currently highly underdiagnosed worldwide. This section begins by modelling a two-stage population screen for the systematic identification of FH cases in the general adult population, followed by an evaluation of the improvement in FH case detection by the inclusion of environmental predictors and a polygenic score for low-density lipoprotein cholesterol.

In conclusion, this thesis puts into perspective the incremental utility of polygenic scores in cardiovascular disease prediction, questioning the claims made on the performance of polygenic scores in prediction. The thesis also explores a potential new avenue of their utility as a tool for aiding with rare variant discovery.

**Impact statement**

This thesis investigates the clinical utility of polygenic scores (PGS) in cardiovascular disease (CVD) prediction and screening. Briefly, the work presented here provides additional evidence to the predictive utility of PGS when added to non-genetic CVD risk prediction models (the QScores developed in the UK). This thesis also proposes a novel two-stage population screening strategy for familial hypercholesterolaemia (FH) in adults, followed by the development of a prediction model for FH utilising PGS information. The impact of each results chapter is presented in more detail below.

Chapter 2 analyses data from the Polygenic Score Catalog, a publicly available website aggregating published PGS studies. The chapter examines various PGS performance metrics for a multitude of disease endpoints and converts them to a more clinically useful metric: the detection rate for a 5% false positive rate. This work provides an overview on the poor performance of PGS in detecting various diseases.

Chapter 4 investigates the predictive utility of PGS in non-genetic CVD clinical risk models used in the UK. It replicates results for QRISK3 and evaluates two other CVD subtype prediction models: QStroke and QDiabetes for the 10-year prediction of incident ischaemic stroke and incident type 2 diabetes. Additionally, the analyses were performed in the UK Biobank which matches the population in which/for whom the non-genetic risk prediction models (QScores) were derived in/for. This provides more robust comparisons than previous studies. The analyses indicate that in addition of being poor predictor of CVD and related endpoints, PGS do not add much predictive value to the clinically used non-genetic scores (QRISK3, QDiabetes and QStroke).

A two-stage adult population screening strategy for FH is tested in Chapter 5. This screening strategy is compared to a previous child-parent FH screening strategy that was rejected by the UK National Screening Committee. If employed, the current adult screening strategy proposed would provide a systematic approach for detecting novel FH cases in the general population (in line with the NHS Long Term Plan) and is anticipated to identify (and treat) individuals at risk of a premature coronary artery event and death.

In Chapter 6, a novel prediction model for FH is developed with the help of a penalised machine-learning algorithm (LASSO) using routinely available environmental variables and a PGS for low-density lipoprotein cholesterol. If used in the two-stage adult population screening strategy, this model is expected to increase the detection rate of FH positive individuals (for a fixed false positive rate), therefore ultimately reducing the burden of individuals sent for confirmatory genetic testing. This chapter also demonstrates that PGS can be clinically useful in helping with rare variant discovery.

**List of publications and presentations resulting from this PhD**

**Publications**:

- **Gratton J**, Finan C, Hingorani AD, Humphries SE, Futema M. LDL-C Concentrations and the 12-SNP LDL-C Score for Polygenic Hypercholesterolaemia in Self-Reported South Asian, Black and Caribbean Participants of the UK Biobank. *Front Genet.* 2022;13:683.
- **Gratton J**, Futema M, Humphries SE, Hingorani AD, Finan C, Schmidt AF. A machine learning model to aid detection of familial hypercholesterolaemia. medRxiv [Internet]. 2022 [cited 2022 Jul 12];2022.06.17.22276540. Available from: https://www.medrxiv.org/content/10.1101/2022.06.17.22276540v1
- Hingorani A, **Gratton J**, Finan C, Schmidt A, Patel R, Sofat R, Kuan V, Langenberg C, Hemingway H, Morris J, Wald N. Polygenic scores in disease prediction: evaluation using the relevant performance metrics. medRxiv [Internet]. 2022 [cited 2022 Jul 12];2022.02.18.22271049. Available from: https://www.medrxiv.org/content/10.1101/2022.02.18.22271049v1

**Publication in preparation**:

- **Jasmine Gratton**, Steve E. Humphries, A. Floriaan Schmidt, Riyaz Patel, Reecha Sofat, Chris Finan, Aroon D. Hingorani*, Marta Futema*. Modelling a two-stage adult population screen for autosomal dominant familial hypercholesterolaemia using UK Biobank.

**Oral presentations**:

- **European Atherosclerosis Society 90th Congress (2022)**: "Modelling a two-stage screen for autosomal dominant familial hypercholesterolaemia (FH) in UK Biobank"
- **The UK Pharmacogenetics & Stratified Medicine (UKPGx) Network Annual Open Meeting (2022)**: "Modelling a 2-stage screen for monogenic familial hypercholesterolaemia using UK Biobank"
- **HEART UK 35th Annual Medical & Scientific Conference (2022)**: Gratton J, Humphries SE, Schmidt AF, *et al.* Modelling a two-stage screen for autosomal dominant familial hypercholesterolaemia (FH) in UK Biobank. *Atherosclerosis Plus* 2022;**49**:S5. doi:10.1016/J.ATHPLU.2022.07.010

## UCL Research Paper Declaration Forms

## For Chapter 2:

| | | | |
|---|---|---|---|
| **1.** **For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:** | | | |
| **a)** **Where was the work published?** (e.g. journal name) | Click or tap here to enter text. | | |
| **b)** **Who published the work?** (e.g. Elsevier/Oxford University Press): | Click or tap here to enter text. | | |
| **c)** **When was the work published?** | Click or tap to enter a date. | | |
| **d)** **Was the work subject to academic peer review?** | Please select. | | |
| **e)** **Have you retained the copyright for the work?** | Please select. | | |
| [If no, please seek permission from the relevant publisher and check the box next to the below statement]:<br>☐ *I acknowledge permission of the publisher named under 1b to include in this thesis portions of the publication named as included in 1a.* | | | |
| **2.** **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:** | | | |
| **a)** **Has the manuscript been uploaded to a preprint server?** (e.g. medRxiv): | Yes | **If yes, which server?** medRxiv | |
| **b)** **Where is the work intended to be published?** (e.g. names of journals that you are planning to submit to) | Lancet | | |
| **c)** **List the manuscript's authors in the intended authorship order:** | Hingorani AD., Gratton J., Finan C., Schmidt AF., Patel R, Sofat R., Kuan V., Langenberg C., Hemingway H., Morris JK., Wald NJ. | | |
| **d)** **Stage of publication** | Not yet submitted | | |
| **3.** **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4**):** | | | |
| Click or tap here to enter text. | | | |
| **4.** **In which chapter(s) of your thesis can this material be found?** | | | |
| Chapter 2 | | | |
| **5.** **e-Signatures confirming that the information above is accurate** (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work): | | | |
| **Candidate:** | | **Date:** | 31/10/2022 |
| **Supervisor/ Senior Author** (where appropriate)**:** | | **Date:** | 01/11/2022 |

## For Chapter 5:

| | | |
|---|---|---|
| **1.** **For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:** | | |
| **a)** **Where was the work published?** (e.g. journal name) | Click or tap here to enter text. | |
| **b)** **Who published the work?** (e.g. Elsevier/Oxford University Press): | Click or tap here to enter text. | |
| **c)** **When was the work published?** | Click or tap to enter a date. | |
| **d)** **Was the work subject to academic peer review?** | Please select. | |
| **e)** **Have you retained the copyright for the work?** | Please select. | |
| [If no, please seek permission from the relevant publisher and check the box next to the below statement]:<br>☐ *I acknowledge permission of the publisher named under 1b to include in this thesis portions of the publication named as included in 1a.* | | |
| **2.** **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:** | | |
| **e)** **Has the manuscript been uploaded to a preprint server?** (e.g. medRxiv): | No | **If yes, which server?** Click or tap here to enter text. |
| **f)** **Where is the work intended to be published?** (e.g. names of journals that you are planning to submit to) | BMJ | |
| **g)** **List the manuscript's authors in the intended authorship order:** | Jasmine Gratton, Steve E. Humphries, A. Floriaan Schmidt, Riyaz Patel, Reecha Sofat, Chris Finan, Aroon D. Hingorani*, Marta Futema* | |
| **h)** **Stage of publication** | Not yet submitted | |
| **3.** **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4**):** | | |
| Click or tap here to enter text. | | |
| **4.** **In which chapter(s) of your thesis can this material be found?** | | |
| Chapter 5 | | |
| **5.** **e-Signatures confirming that the information above is accurate** (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work): | | |

| | | | |
|---|---|---|---|
| **Candidate:** | | **Date:** | 31/10/2022 |
| **Supervisor/ Senior Author** (where appropriate)**:** | | **Date:** | 01/11/2022 |

## For Chapter 6:

| | | |
|---|---|---|
| **1.** **For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:** | | |
| **a)** **Where was the work published?** (e.g. journal name) | Click or tap here to enter text. | |
| **b)** **Who published the work?** (e.g. Elsevier/Oxford University Press): | Click or tap here to enter text. | |
| **c)** **When was the work published?** | Click or tap to enter a date. | |
| **d)** **Was the work subject to academic peer review?** | Please select. | |
| **e)** **Have you retained the copyright for the work?** | Please select. | |
| [If no, please seek permission from the relevant publisher and check the box next to the below statement]:<br>☐ *I acknowledge permission of the publisher named under 1b to include in this thesis portions of the publication named as included in 1a.* | | |
| **2.** **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:** | | |
| **i)** **Has the manuscript been uploaded to a preprint server?** (e.g. medRxiv): | Yes | **If yes, which server?** medRxiv |
| **j)** **Where is the work intended to be published?** (e.g. names of journals that you are planning to submit to) | JACC Advances | |
| **k)** **List the manuscript's authors in the intended authorship order:** | Jasmine Gratton, Marta Futema, Steve E. Humphries, Aroon D. Hingorani, Chris Finan*, Amand F. Schmidt* | |
| **l)** **Stage of publication** | Undergoing revision after peer review | |
| **3.** **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4**):** | | |
| Click or tap here to enter text. | | |
| **4.** **In which chapter(s) of your thesis can this material be found?** | | |
| Chapter 6 | | |
| **5.** **e-Signatures confirming that the information above is accurate** (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work): | | |
| **Candidate:** | | **Date:** 31/10/2022 |
| **Supervisor/ Senior Author** (where appropriate)**:** | | **Date:** 01/11/2022 |

**Acknowledgements**

I would like to start by acknowledging the British Heart Foundation for funding my research.

Most importantly, this work would not have been possible without the expertise, guidance and help of my supervisors Prof Aroon Hingorani, Dr Chris Finan, Dr Floriaan Schmidt, Dr Marta Futema and Prof Reecha Sofat. I am forever grateful for the knowledge, skills and valuable time they have generously shared with me. I would also like to acknowledge Prof Steve Humphries for providing his FH expertise and for his collaboration on various projects, as well as Aroon's group members (and others) who have all contributed to an engaging and supportive work environment.

I would next like to thank my PhD friends, especially Roshni and Maria, for the good times and support that have made this journey even more enjoyable. I am very grateful to have made lifelong friendships.

And last but not least, I would like to thank my mum, my dad and my sister for providing moral and emotional support throughout the lows, and for celebrating the highs of these past years. Je vous aime fort.

**Table of contents**

**Abstract** ..... **3**
**Impact statement** ..... **4**
**List of publications and presentations resulting from this PhD** ..... **6**
**UCL Research Paper Declaration Forms** ..... **7**
**Acknowledgements** ..... **10**
**Table of contents** ..... **11**
**List of figures** ..... **16**
**List of tables** ..... **18**
**Abbreviations** ..... **20**
**1 Introduction** ..... **22**
1.1 Cardiovascular disease (CVD) and its risk factors ..... 22
1.1.1 Pathophysiology and risk factors of CVD ..... 22
1.1.2 Relationships between risk factors and disease endpoints ..... 23
1.1.3 Individual risk factors as predictive tests ..... 24
1.2 Evaluating clinical prediction models ..... 26
1.2.1 Model derivation ..... 26
1.2.2 Selecting model predictors ..... 27
1.2.3 Model calibration & recalibration ..... 28
1.2.4 Model discrimination: ROC curve & AUC ..... 29
1.2.5 Comparing models: net reclassification index (NRI) & decision curve analysis ..... 31
1.2.5.1 NRI ..... 31
1.2.5.2 Decision curve analysis ..... 32
1.2.6 Clinical utility ..... 33
1.3 Polygenic risk scores (PRS) in disease prediction ..... 34
1.3.1 Generating polygenic scores (PGS) ..... 35
1.3.1.1 Weighted and unweighted scores ..... 35
1.3.1.2 Score parameters: linkage disequilibrium (LD) & p-value thresholds ..... 36
1.3.1.3 Dataset considerations ..... 37
1.3.2 The prediction of complex heritable traits and incident diseases using PRS ..... 38
1.3.3 Comparing the predictive performance of clinical risk prediction models and PRS 39
1.3.4 Leveraging polygenic information for the discovery of rare genetic variants ..... 40
1.4 Familial hypercholesterolaemia (FH) ..... 41
1.4.1 Overview of FH ..... 41
1.4.2 Genetics of FH ..... 42
1.4.3 Clinical diagnosis of FH ..... 43
1.4.4 Cascade testing of index FH cases ..... 44

1.4.5 Population screening of FH ..... 45
1.5 Thesis overview ..... 46
1.6 References ..... 47
**2 Analysis of polygenic risk scores (PRS) in the Polygenic Score Catalog for disease screening, risk prediction, and population stratification ..... 59**
2.1 Abstract ..... 59
2.2 Introduction ..... 60
2.3 Methods ..... 61
2.3.1 Analysis overview ..... 61
2.3.2 Assumptions ..... 62
2.3.3 Deriving DR5 from HR or OR per SD ..... 62
2.3.4 Deriving DR5 from AUC and C-index ..... 65
2.3.5 Screening: calculating the likelihood ratio and odds of becoming affected given a positive test result (OAPR) ..... 68
2.3.6 Risk prediction: calculating the likelihood ratio and odds of becoming affected given a PRS result ..... 69
2.3.7 Risk stratification: calculating the likelihood ratio and odds of becoming affected for a particular PRS group ..... 70
2.4 Results ..... 71
2.4.1 Performance of PGS in the Polygenic Score Catalog ..... 71
2.4.2 Performance of PRS in disease screening ..... 73
2.4.3 Risk prediction: interpretation of PRS in an individual ..... 74
2.4.4 Performance of PRS in population stratification ..... 74
2.4.5 Using PRS in conjunction with conventional screening tests or risk factors ..... 75
2.5 Discussion ..... 78
2.6 References ..... 80
**3 The UK Biobank study ..... 83**
3.1 Overview ..... 83
3.2 Genotyping data ..... 83
3.3 Quality control (QC) ..... 84
3.3.1 Sample QC ..... 84
3.3.2 Genetic variant QC ..... 85
3.3.3 Relatedness ..... 85
3.3.4 Final QC step: merging the phenotype file ..... 86
3.4 References ..... 87
**4 Evaluating PRS in the prediction of cardiovascular disease (CVD), coronary heart disease (CHD), type 2 diabetes (T2D), and ischaemic stroke (IST) ..... 89**
4.1 Abstract ..... 89
4.2 Introduction ..... 90

4.3 Methods ..... 91
4.3.1 Implementing QRISK3, QStroke and QDiabetes in the UK Biobank ..... 91
4.3.2 The prediction models developed and tested ..... 92
4.3.3 Data QC of the UK Biobank ..... 93
4.3.4 Generating PRS for CVD, CHD, IST and T2D ..... 93
4.3.5 Assessment of the prediction models ..... 94
4.3.6 Software ..... 95
4.4 Results ..... 95
4.4.1 Characteristics of study participants ..... 95
4.4.2 PRS for CHD, IST and T2D ..... 97
4.4.3 Relationship of QScores and PRS to incident CVD, CHD, IST, T2D ..... 97
4.4.3.1 Score distributions ..... 97
4.4.3.2 Score odds ratio ..... 99
4.4.3.3 Correlation between PRS and QScores ..... 100
4.4.4 Calibration of the models tested ..... 101
4.4.5 Discrimination of the models tested ..... 103
4.4.6 Detection rate of cases by the models for a 5% false positive rate ..... 105
4.4.7 Comparison with similar studies ..... 106
4.5 Discussion ..... 108
4.5.1 Overview of the study ..... 108
4.5.2 An external validation of the QScores in the UK Biobank ..... 108
4.5.3 The performance of PRS in CVD, CHD, IST and T2D prediction ..... 109
4.5.4 The effects of age and sex in incident disease prediction ..... 109
4.5.5 The combined PRS and QScore models ..... 110
4.5.6 Study limitations ..... 111
4.5.7 Conclusion ..... 111
4.6 References ..... 112
4.7 Appendix ..... 116
**5 Modelling a two-stage adult population screen for autosomal dominant familial hypercholesterolaemia (FH) ..... 123**
5.1 Abstract ..... 123
5.2 Introduction ..... 124
5.3 Methods ..... 126
5.3.1 Participants ..... 126
5.3.2 LDL-C measurement ..... 126
5.3.3 Identification of carriers of FH-causing genetic variants ..... 126
5.3.4 Evaluation of two-stage adult FH screening performance ..... 128
5.3.5 Comparison of two-stage adult and child-parent screening ..... 128

5.3.6 Modelling cascade testing in families of index cases ..... 128
5.3.7 Achieving the NHS Long Term Plan target for FH case detection ..... 129
5.4 Results ..... 129
5.4.1 Demographic and other characteristics of study participants ..... 129
5.4.2 Participants with an FH-causing variant or variant of unknown significance ..... 130
5.4.3 Performance of two-stage adult screen for autosomal dominant FH ..... 132
5.4.4 Comparison of two-stage adult with child-parent screening ..... 136
5.4.5 Achieving the NHS target for FH case detection ..... 137
5.5 Discussion ..... 137
5.5.1 Overview of the study ..... 137
5.5.2 Limitations of the approach ..... 138
5.5.3 Potential setting of a two-stage adult screening programme ..... 139
5.5.4 DNA sequencing capacity in the NHS ..... 140
5.5.5 Cascade testing capacity ..... 140
5.5.6 Reaching the NHS Long Term Plan target ..... 140
5.5.7 Conclusion ..... 141
5.6 References ..... 141
5.7 Appendix ..... 147
**6 A machine learning model to aid detection of familial hypercholesterolaemia (FH) 156**
6.1 Abstract ..... 156
6.2 Introduction ..... 157
6.3 Methods ..... 158
6.3.1 Genomics data availability and FH case ascertainment ..... 158
6.3.2 LDL-C PGS generation ..... 158
6.3.3 Deriving a machine learning algorithm to prioritise participants with FH ..... 159
6.3.4 Evaluating the burden of genomic sequencing for FH ..... 160
6.3.5 Software ..... 161
6.4 Results ..... 161
6.4.1 Participant characteristics of our study cohort ..... 161
6.4.2 LDL-C PGS ..... 163
6.4.3 Multivariable machine learning model to prioritise FH variant carriers ..... 163
6.4.1 Evaluating the FH screening strategies through decision curve analysis ..... 165
6.4.2 Model FH classification ..... 166
6.4.3 Prioritising individuals for FH genomic testing in a two-stage population screening strategy168
6.5 Discussion ..... 171
6.6 References ..... 173

6.7 Appendix ..... 176
**7 General discussion ..... 182**
7.1 Overview of thesis ..... 182
7.1.1 PGS in CVD prediction ..... 182
7.1.2 PGS in rare variant discovery ..... 183
7.2 Wider perspective ..... 186
7.2.1 PGS in context ..... 186
7.2.2 Important considerations ..... 187
7.2.3 Future avenues ..... 188
7.3 Summary ..... 189
7.4 References ..... 189

**List of figures**

Figure 1.1 from Wald et al.[38] Distribution of risk factors in affected and unaffected individuals as predictors of disease. ....25
Figure 1.2 from E. Christensen[63] Relationship between risk factor distributions and the ROC curve. ....29
Figure 1.3 from Janssens et al.[64] Evaluating the discriminative ability of a model to distinguish between affected and unaffected individuals. ....30
Figure 1.4 from Vickers *et al.*[73] Decision curve analysis plot. ....32
Figure 1.5 Workflow for generating trait specific PGS and applying them to a target dataset. ....38
Figure 1.6 FH prevalence and the effects of statin treatment on survival. ....41
Figure 1.7 from Soutar & Naoumova.[109] The LDLR-mediated cellular uptake of LDL. ....42
Figure 2.1 Standardised normal distributions for affected and unaffected individuals and their relationship to the OR or HR per SD and the DR5. ....62
Figure 2.2 Standardised normal distributions for affected and unaffected individuals and their relationship to the AUC or C-index, the DR and the FPR. ....65
Figure 2.3 Standardised normal distributions for affected and unaffected individuals and their relationship to the likelihood ratio. ....69
Figure 2.4 Standardised normal distributions for affected and unaffected individuals and their relationship to the likelihood ratio for different quintiles of the distributions. ....71
Figure 2.5 Summary of data included in the Polygenic Score Catalog as of April 2022. ....72
Figure 2.6 Distribution of DR5 derived from HR per SD, OR per SD, AUC, and C-index values listed in the Polygenic Score Catalog. ....73
Figure 2.7 Flow diagram of a hypothetical cohort of 100,000 individuals modelled by Sun *et al.* for the detection of CAD and stroke cases using conventional risk factors (CRF) alone versus a model combining CRF and PRS.[16] ....76
Figure 3.1 Workflow of the QC steps for the UK Biobank genotype data. ....87
Figure 4.1 The CVD, CHD, IST and T2D PRS distributions for cases and controls. ....97
Figure 4.2 Logit transformed risk distributions for CVD, CHD, IST and T2D PRS, and QRISK3, QStroke and QDiabetes. ....98
Figure 4.3 Log mean incident disease (CVD, CHD, IST, T2D) per risk score decile for CVD, CHD, IST, T2D PRS and QRISK3, QStroke and QDiabetes. ....100
Figure 4.4 Correlation scatter plots of PRS and QScores for the prediction of incident outcomes. ....101

Figure 4.5 The calibration curves of the prediction models (CVD PRS, CHD PRS, IST PRS, QRISK3, QStroke, QDiabetes or combined models) tested for incident CVD, CHD, IST and T2D. ..... 102
Figure 4.6 The discrimination (C-statistic) of the prediction models (PRS, QScore and combined models) tested for incident CVD, CHD, IST and T2D. ..... 104
Figure 5.1 Relative frequency distributions of the adjusted LDL-C concentrations in monogenic FH variant carriers and non-carriers of the study cohort. ..... 132
Figure 5.2 The number of samples sequenced, and the number of FH cases detected using various LDL-C cut-off values in the two-stage adult screening population strategy for FH for a hypothetical population of 100,000 individuals. ..... 133
Figure 5.3 Illustration of the two-stage adult screen and subsequent cascade screening of first-degree relatives of index FH cases scaled to 100,000 individuals using an LDL-C cut-off value of 4.8 mmol/L in the first stage screen. ..... 135
Figure 6.1 Workflow of LDL-C PGS generation, FH case ascertainment and testing versus training data split of the UK Biobank's White British participants. ..... 159
Figure 6.2 Feature importance of the variables retained by LASSO regression predicting monogenic FH, and the density predicted probability distributions from this model for unaffected and affected FH individuals in White British participants of the UK Biobank. ..... 164
Figure 6.3 Discrimination and calibration of a multivariable algorithm including LDL-C PGS predicting FH carriership using independent testing data. ..... 165
Figure 6.4 Decision curve analysis of the FH prediction models ..... 166
Figure 6.5 Adult two-stage population screening strategy for monogenic FH. ..... 170

**Supplementary figures**

Supplementary Figure 4.1 The polygenic risk score (PRS) parameters tested (p-value and linkage disequilibrium (LD) cut-off values) and their C-statistic for incident coronary heart disease (CHD), incident ischaemic stroke (IST), and incident type 2 diabetes (T2D). ..... 122
Supplementary Figure 6.1 Correlation plot of the variables tested in the LASSO regression model for the prediction of monogenic FH. ..... 180
Supplementary Figure 6.2 LASSO regression model feature selection and importance for monogenic FH prediction. ..... 181

**List of tables**

Table 1.1 Two by two table illustrating the relationship between the DR, FPR, OAPR, PPV, NPV, odds of disease, absolute risk and LR. ....33
Table 1.2 The FH diagnostic variables included in the Simon Broome, Dutch Lipid Clinical Network, and the Make Early Diagnosis to Prevent Early Death criteria....44
Table 2.1 The DR5 values derived from the HR per SD, OR per SD, AUC, and C-index metrics reported in the Polygenic Score Catalog. ....72
Table 2.2 Effect of adding a PRS to non-genetic risk factors in prediction of CAD or CVD.....77
Table 4.1 The prediction models evaluated for incident CVD, CHD, IST and T2D. ....92
Table 4.2 The GWAS summary statistics used to generate the PRS. ....94
Table 4.3 Characteristics of 341,515 UK Biobank White British participants included in the analysis stratified by sex. ....96
Table 4.4 The mean predicted risk and SD of the distributions for cases and controls for CVD PRS, CHD PRS, IST PRS, T2D PRS, QRISK3 for incident CVD, QRISK3 for incident CHD, QStroke and QDiabetes....99
Table 4.5 Change in the C-statistic and detection rate for a 5% false positive rate between the QScore and the QScores x PRS models.... 105
Table 4.6 Comparison of PRS in previous key studies for CHD and CVD prediction.... 107
Table 5.1 Genetic coordinates of FH-causing genes.... 127
Table 5.2 Participant characteristic comparison between UK Biobank participants of the study cohort and the NHS Health Check 2017-2018.[43] .... 130
Table 5.3 Characteristics of the study participants.... 131
Table 5.4 Performance of a two-stage adult population screen for monogenic FH using different stage 1 LDL-C cut-offs.... 134
Table 5.5 Comparison of child-parent and adult two-stage screening strategies for FH.... 136
Table 6.1 UK Biobank participant characteristics stratified by carrying a FH-causing variant. . 162
Table 6.2 The classification accuracy of an algorithm for predicting monogenic FH using the multivariable model and LDL-C concentration accounting for statin use.... 167
Table 6.3 NRI table and estimates for a predicted probability threshold of 0.006 for FH comparing the multivariable model with LDL-C PGS to a simpler model of LDL-C concentration and statin use.... 168

**Supplementary tables**

Supplementary Table 4.1 The QRISK3, QDiabetes and QStroke variables and their corresponding UK Biobank data field numbers and ICD-10 diagnostic codes. ..... 116
Supplementary Table 4.2 The ICD-10 and OPCS-4 codes used to define incident cardiovascular disease, coronary heart disease, ischaemic stroke, and type 2 diabetes. ..... 117
Supplementary Table 4.3 Characteristics of 341,515 UK Biobank White British participants included in the analysis with missing data singly imputed stratified by sex. ..... 117
Supplementary Table 4.4 The p-values and linkage disequilibrium (LD) cut-off values that yielded the polygenic risk scores (PRS) with the highest C-statistic for each incident outcome studied. ..... 119
Supplementary Table 4.5 The C-statistic, calibration-in-the-large, calibration slope values and detection rate for a 5% false positive rate (DR5) of the PRS (CVD, CHD, IST, T2D), QScores (QRISK3, QStroke, QDiabetes) and combined models (age and sex; PRS, age and sex; PRS and QScore) for incident disease outcomes (CVD, CHD, IST and T2D). ..... 120
Supplementary Table 5.1 Autosomal dominant FH-causing mutation identified in the study cohort. ..... 147
Supplementary Table 5.2 List of variants of unknown significance (VUS) excluded from the analysis ..... 150
Supplementary Table 5.3 Study participants characteristics categorised by FH-causing gene .... 153
Supplementary Table 5.4 The counts obtained from the two-stage screen in the study cohort of 140,439 individuals for various LDL-C cut-off values ..... 155
Supplementary Table 6.1 UK Biobank participant characteristics post imputation of missing values stratified by FH carriership. ..... 176
Supplementary Table 6.2 The non-genetic variables and coefficients retained by LASSO regression for monogenic FH prediction. ..... 178
Supplementary Table 6.3 The variables and coefficients retained by LASSO regression for monogenic FH prediction. ..... 179

**Abbreviations**

| | |
|---|---|
| ACC | American College of Cardiology |
| AF | atrial fibrillation |
| AHA | American Heart Association |
| ALP | alkaline phosphatase |
| ALT | alanine aminotransferase |
| Apo-A1 | apolipoprotein A1 |
| Apo-B | apolipoprotein B |
| APOE | apolipoprotein E |
| AST | aspartate aminotransferase |
| AUC | area under the receiver operating characteristic curve |
| BMI | body mass index |
| BP | blood pressure |
| C-statistic (C-index) | concordance statistic |
| CAD | coronary artery disease |
| CDF | cumulative distribution function |
| CHD | coronary heart disease |
| CI | confidence interval |
| CIL | calibration-in-the-large |
| CRP | C-reactive protein |
| CS | calibration slope |
| CVD | cardiovascular disease |
| DBP | diastolic blood pressure |
| DLCN | Dutch Lipid Clinical Network |
| DR | detection rate |
| DR5 | detection rate for a 5% false positive rate |
| FH | familial hypercholesterolaemia |
| FPR | false positive rate |
| FRS | Framingham Risk Score |
| GLGC | Global Lipids Genetics Consortium |
| GWAS | genome-wide association study |
| HDL | high-density lipoprotein |
| HDL-C | high-density lipoprotein cholesterol |
| HES | hospital episode statistics |
| HF | heart failure |
| HR | hazard ratio |
| HST | haemorrhagic stroke |
| ICD-10 | International Classification of Disease 10 |
| INFO | imputation information |
| IQR | interquartile range |
| IST | ischaemic stroke |
| LASSO | least absolute shrinkage and selection operator |

| | |
|---|---|
| LD | linkage disequilibrium |
| LDL | low-density lipoprotein |
| LDL-C | low-density lipoprotein cholesterol |
| LDLR | low-density lipoprotein receptor |
| LDLRAP1 | low-density lipoprotein receptor adapter protein 1 |
| Lp(a) | lipoprotein A |
| LR | likelihood ratio |
| MAF | minor allele frequency |
| MEDPED | Make Early Diagnosis to Prevent Early Death |
| MI | myocardial infarction |
| MoM | multiple of the median |
| NHS | National Health Service |
| NICE | National Institute for Health and Care Excellence |
| NPV | negative predictive value |
| NRI | net reclassification index |
| NSAID | non-steroidal anti-inflammatory drug |
| OAPR | odds of being affected given a positive test result |
| OPCS-4 | Office of Population Censuses and Surveys 4 Classification of Interventions and Procedures |
| OR | odds ratio |
| PCSK9 | proprotein convertase subtilisin/kexin Type 9 |
| PGS | polygenic score |
| PPV | positive predictive value |
| PRS | polygenic risk score |
| QC | quality control |
| ROC | receiver-operating characteristic |
| SBP | systolic blood pressure |
| SD | standard deviation |
| SE | standard error |
| SNP | single nucleotide polymorphism |
| T2D | type 2 diabetes |
| UK | United Kingdom |
| US | United States |
| VEP | Variant Effect Predictor |
| VUS | variant of unknown significance |
| WES | whole-exome sequencing |

# 1 Introduction

## 1.1 Cardiovascular disease (CVD) and its risk factors

Cardiovascular disease (CVD) is a general term that encompasses many illnesses of the heart and vascular system. These include but are not limited to coronary heart disease (CHD), ischaemic stroke (IST), haemorrhagic stroke (HST), heart failure (HF) and atrial fibrillation (AF). Many of these illnesses have shared pathophysiology and a number of common risk factors, for example age, high blood pressure (BP), diabetes and elevated low-density lipoprotein cholesterol (LDL-C). CVD is the major cause of death globally with an estimated 17.9 million people losing their lives annually.[1] According to the British Heart Foundation, an estimated 168,000 deaths annually (24%) in the UK are caused by CVD, of which 48,000 are premature (defined as under the age of 75).[2]

### 1.1.1 Pathophysiology and risk factors of CVD

Many CVD events manifest following an obstruction of blood supply to the heart or brain. These obstructions can be caused by atherosclerotic plaques or, if these plaques rupture, by blood clots that can form on their surface.[3] An atherosclerotic plaque is a deposit of lipids, fibrous material and immune cells that accumulate over time in the inner layer of artery walls.[3] Elevated circulating LDL-C is a major cause of atherosclerosis, as evidenced by observational epidemiology, genetic studies and clinical trials of LDL-C lowering drugs.[4–6] As atherosclerotic plaques grow in size, they progressively narrow the arterial lumen and can lead to the reduction of blood flow to vital organs – a phenomenon known as ischaemia.[3] In the coronary arteries, that can lead to angina.[3] Atherosclerotic plaques can also rupture, releasing the fatty and fibrous elements into the circulation, and triggering local formation of a thrombus leading to an acute ischaemic event such as an IST or myocardial infarction (MI).[3]

AF is a common heart rhythm disturbance that can lead to blood stasis in the left atrium and the formation of left atrial thrombus.[7] Such thrombi can embolise to the brain causing a type of ischaemic stroke called a cardioembolic stroke, or to other organs causing systemic embolism.[7,8] Risk factors for AF include age, high BP, obesity, elevated alcohol and tobacco consumption, certain types of valvular heart disease, and HF of any cause.[9]

HF is characterised by abnormal pumping of the heart, leading to changes in blood ejection fraction or ventricular filling.[10] It is a progressive disease linked to many comorbidities and poor prognosis, with risk factors including ischaemic heart disease and hypertension.[10]

CVD has a long preclinical phase (decades) where many modifiable risk factors influence the progression and severity of disease.[3,11,12] Many risk factors are well studied, providing the opportunity for primary prevention and risk prediction.[11,12] Primary prevention has played a key role in trying to reduce the burden of CVD globally.[13] This relates to lifestyle modifications such as smoking cessation, alcohol intake reduction, regular exercise, and diet modifications (i.e. reducing saturated fat, salt and sugar intake, and increasing the consumption of fresh fruit, vegetables and high-fibre nutrients) or prescription medication (e.g. statins to reduce LDL-C levels, and antihypertensives to reduce elevated BP) to prevent the occurrence of a primary CVD event.[11,12] Over time, adverse lifestyle factors can accumulate and lead to hypertension, inflammation, type 2 diabetes and hypercholesterolemia, which are all major risk factors for CVD.[3,14] Secondary prevention of CVD aims to prevent a subsequent CVD event from occurring mostly via medical interventions such as prescription medication or surgery (e.g. coronary artery bypass).[3]

### 1.1.2 Relationships between risk factors and disease endpoints

Many CVD risk factors (e.g. BP, LDL-C) were first identified in prospective longitudinal cohort studies. The first major influential study of this kind was the Framingham Heart Study which began in 1947 in the United States (US).[15] The initial report described the sex, age, BP, cholesterol and body mass index (BMI) differences observed in patients with incident CHD.[16] Other major risk factors such as smoking and diabetes were later identified.[17,18] The inverse relationship between high-density lipoprotein cholesterol (HDL-C) levels and CHD incidence was also found.[19]. Since then, numerous global (e.g. MONICA, INTERHEART) and national (e.g. Whitehall I/II, the Reykjavik Heart Study, ARIC) cohort studies have further quantified the relationship between risk factors and CVD endpoints.[14,20–24]

The Emerging Risk Factors Collaboration and the Prospective Studies Collaboration have undertaken meta-analyses of the associations of risk factors and cardiovascular endpoints to provide detailed information on the strength and slope of associations.[25,26] These analyses have confirmed that the relationship between CVD and many of its risk factors is log-linear, meaning

that there is the same proportionate increase in CVD risk per unit change of the risk factor. For example, meta-analysing over 200 studies has shown that LDL-C has a dose-dependent and log-linear association with atherosclerotic CVD.[27] There is also a linear increase in the relative risk for CHD and the number of cigarettes smoked per day.[28] Similarly, there is a log-linear association between BP (systolic and diastolic) and the hazard for cardiovascular events (including cardiac death, MI, stroke and HF).[29,30] Although age is not a modifiable risk factor, it is the biggest predictor of CVD.[30]

There is also an additive relationship between CVD risk factors.[31] For example, in AF, there is a clear increase in risk with each additional risk factor (hypertension, low HDL-C, impaired fasting glucose, high waist circumference, elevated triglycerides).[32] Identifying a risk factor-disease association has potential clinical applications for treatment (for which the risk factor must be causal (e.g. BP, LDL-C)), or for disease prediction (for which the risk factor need not be causal).[33,34]

Since CVD risk factors are normally or log normally distributed in the population and exhibit a log-linear disease association, there is also no clear-cut threshold value above which disease risk is clearly increased.[35] These relationships help provide some basic information into how well a risk factor might perform as a predictive test, alone or in combination.

### 1.1.3 Individual risk factors as predictive tests

Most clinical CVD events occur among people with average risk factor levels because most risk factors are normally distributed and there are large numbers of individuals with near-average levels of risk factors at intermediate risk of disease.[36] This is sometimes referred to as Rose's Prevention Paradox: most cases of disease occur among the large number of individuals at intermediate risk rather than the small number of individuals at high risk.[37] This has important implications for CVD risk prediction.

The distributions of risk factors among people who do and do not develop CVD overlap substantially, so there is usually no clear cut-point for the risk factor that readily separates the two groups (**Figure 1.1** from Wald *et al.*[33]). This means that a CVD risk factor, even a causal one such as BP or LDL-C, poorly discriminates between affected and unaffected individuals and turns out to be a modest predictor of CVD (see section 1.2.4. of the Introduction for more information on model discrimination).[38] A risk factor has to be extremely strongly associated with a disease

for it be considered a useful screening test (this is the case for example for serum alpha fetoprotein in pregnancies affected by spina bifida) (**Figure 1.1**).[38] A causal risk factor is therefore not necessarily a good predictor of disease, but a risk factor also does not need to be causal for the disease for it to be a good predictor (e.g. alpha fetoprotein does not cause neural tube defects in spina bifida).

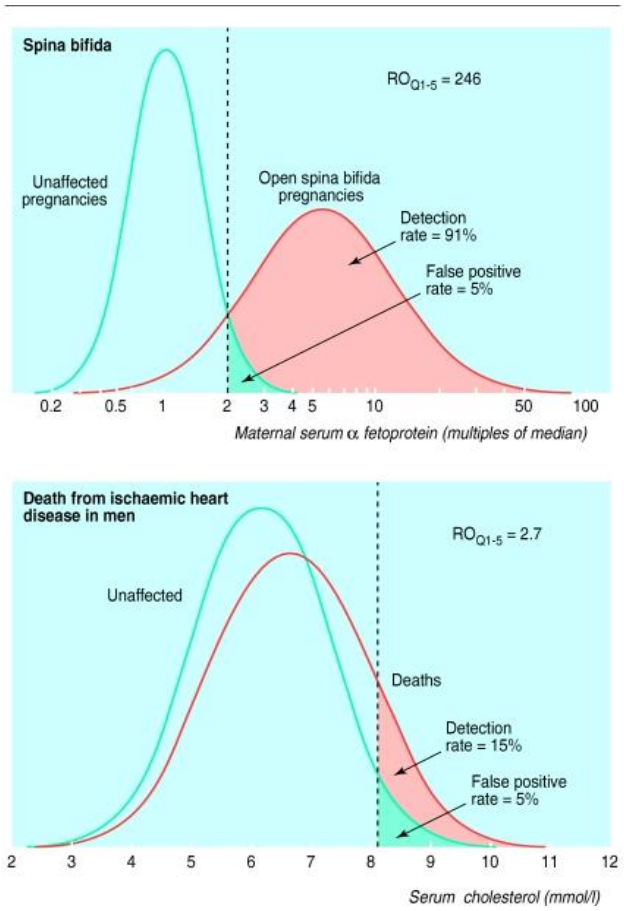



**Figure 1.1 from Wald et al.[38] Distribution of risk factors in affected and unaffected individuals as predictors of disease. A)** Distribution of alpha fetoprotein in affected and unaffected pregnancies for spina bifida. **B)** Distribution of serum cholesterol in men affected and unaffected by deadly ischaemic heart disease. The $RO_{Q1-5}$ (i.e. relative odds for people in the first versus fifth quintiles) is higher when the affected and unaffected distributions have a greater separation (see Introduction section 1.2.4 for a more detailed explanation).

The poor performance of a single CVD risk factor in disease prediction, coupled with the fact that most CVD events occur in the average-risk category, present many challenges for CVD prediction.[39,40] However, since there is an additive relationship between CVD risk factors and CVD risk, efforts have been made to develop CVD risk models that include multiple risk factors

in order to improve prediction and risk stratification of high-risk individuals.[41] These risk models have been developed using logistic regression analysis where the incident outcome (e.g. CVD incidence over a 10-year period) is regressed against independent variables (i.e. the predictors such as age, BP, LDL-C) to find the appropriate baseline risk value and prediction coefficients for each variable in the regression analysis. The development, evaluation and potential application of such models are described in more detail in the next section.

### 1.2 Evaluating clinical prediction models

The World Health Organisation estimates that 80% of CVD deaths are preventable, but the optimal strategy to prevent CVD events has been the subject of debate.[42] Rose's Prevention Paradox makes prediction challenging because most CVD events occur in individuals at intermediate risk, and there is no clear cut-point that delineates individuals who will and will not develop CVD. This has led the scientific community to question whether an individual-based (i.e. for high-risk individuals) or population-based strategy for CVD prevention would be more beneficial, especially considering the reduction in cost of some treatments (e.g. statins).[39,40]

In current clinical practice, individuals at high risk of CVD are identified by means of risk prediction models (despite their limitations). These models estimate a person's risk of developing CVD, after which a threshold value can be applied to guide early detection and prevention strategies. Examples of such models are the QRISK3 calculator in the United Kingdom (UK), the FINRISK-calculator in Finland, and the European SCORE risk charts.[43–45] CVD risk scores from the US include the Framingham Risk Score (FRS), the 2013 American College of Cardiology and American Heart Association (ACC/AHA) Pooled Cohort Equations, and the Reynolds Risk Score for women and men.[19,45–47] Clinical risk score models are available for many other cardiovascular-related traits. QResearch in the UK has developed QStroke for the 10-year prediction of IST, and QDiabetes for the 10-year prediction of type 2 diabetes (T2D).[48,49]

#### 1.2.1 Model derivation

In order for a clinical prediction model to be useful, it has to be accurate and easy to implement.[50] Typically, clinical predictors are readily or routinely collected variables such as age, sex, ethnicity, certain physiological measurements such as BP, height and weight, medical and family history.[51] Each variable carries a specific weight in the final risk score depending on how much it contributes

to the prediction of the disease endpoint. A baseline risk for time to follow up is also added in the risk calculations. Effective clinical prediction models should also be able to communicate risk in an interpretable manner.[52] For example, an absolute risk over a specified timeframe (as is used in the QRISK3 10-year CVD risk calculator) and other similar calculators.

All new clinical prediction models should be derived in a population that is representative of the population in which the prediction model is to be used, otherwise the newly derived model might not generalise very well.[53] Additionally, a simple but robust model is preferred over a complicated one as this could lead to overfitting of the model, which means that the model is too specific to the dataset where it is developed in and does not generalise well to others.[50,54] When developed, model predictors are carefully evaluated and dropped if they do not add much value to the overall predictive ability of the model.

### 1.2.2 Selecting model predictors

In disease prediction, predictors are often established risk factors for the disease (e.g. BP, LDL-C and BMI for CVD prediction).[41] However, more recently, novel biomarkers (e.g. C-reactive protein (CRP)) have been investigated as potential predictors of CVD.[55] The selection of model predictors is commonly done in a stepwise manner: this technique selects predictors by alternating between forward and backward selection.[54,56] Forward selection of predictors starts with an empty model. Predictors are added to the model one at a time and are retained if they add significant predictive value to the model.[56] Conversely, backward selection starts with a model that contains all the known predictors. These predictors are progressively removed from the model in a stepwise fashion and the predictive ability of the model is evaluated at each step. If the predictors do not add significant predictive value to the model, they are removed.[54,56] Implicit in the removal (or retention) of potential risk factors from a model is that certain risk factors tend to be correlated with one another and so having all may not add much to a risk model.[54] This is the case with CRP for example, where the addition of CRP to CVD prediction does not improve the discrimination of the prediction model by much, partly because CRP is already associated with the other CVD risk factors.[57,58] And while stepwise selection of model predictors is often performed, this can sometimes lead to overfitting of the resulting model.[54] Predictors can also be chosen by evaluating all the possible combinations of predictors and selecting the best ones.[56]

Some machine learning methods, such as least absolute shrinkage and selection operator (LASSO), automatically perform variable selection. LASSO does so by shrinking the predictor coefficients and excluding those that reach zero (as they would not provide an improvement in prediction).[53,59] This reduces the issue of model overfitting when performing variable selection.

Polygenic scores (PGS) differ in the way that their predictors (i.e. known genetic variants) are chosen: the variants are selected based on pre-specified p-value, linkage disequilibrium (LD) and minor allele frequency (MAF) cut-off values (see section 1.3. of the Introduction for more information on PGS).[60] While most non-genetic predictors such as hypertension and hypercholesterolemia for CVD are discovered through longitudinal population cohort studies, genetic risk factors are often identified from case-control genome-wide association studies (GWAS) as genetic variants remain unchanged throughout life.[60]

### 1.2.3 Model calibration & recalibration

Once the predictors have been selected, the performance of a model is evaluated based on its calibration and discrimination metrics. Calibration assesses the ability of a model to accurately predict the outcome of a group of interest. This can be done by plotting the mean observed risk against the mean predicted risk from the model using groups of individuals from low to high risk.[50,53] A well calibrated model will have a slope value close to one with a y-intercept value of zero. The calibration-in-the-large, which is a measure of the difference in the means of the observed and predicted risks, will be close to zero for well calibrated models.[50,53] The calibration estimate of a model in a training dataset will always be perfect because the model was made to fit (i.e. calibrated to) that specific dataset. Therefore, the success of calibration can only be defined in a dataset that is distinct from the training dataset and highlights the importance of having an independent testing dataset where the model can be impartially evaluated.[53]

And since model performance does not always translate well to new datasets, models can be recalibrated.[53] This also involves splitting the data into a training dataset and a testing dataset. In model recalibration, the training data's logistic regression model is defined as a linear predictor, and this linear predictor is used in a subsequent logistic regression model where the overall slope and intercept values are re-evaluated to adjust the model parameters for the new dataset.[61,62] The recalibrated model is then validated in the test data.

### 1.2.4 Model discrimination: ROC curve & AUC

The discrimination of a model is its ability to differentiate between people getting the outcome of interest (diseased/affected) or not (non-diseased/unaffected).[53] The model generates a continuous value, and a cut-point is chosen. For any cut-point, the detection rate (DR) gives the proportion of individuals with a predicted risk above the cut-point value who will develop the outcome of interest.[38] The DR is also known as the sensitivity, but "sensitivity" is often misinterpreted as the lower limit of detection for a biochemical test, hence why DR is preferred. The false positive rate (FPR) (1-specificity) is the proportion of individuals above the threshold value who will not develop the outcome of interest.[38] Assuming that the risk factor distributions for those who do or do not become affected overlap, the DR and FPR will be influenced by the cut-point chosen (**Figure 1.2** from E. Christensen[63]). The discriminative performance for different cut-points can be summarised using a receiver-operating characteristic (ROC) curve, which is a plot of the DR versus the FPR for different cut-points of predicted risk (**Figure 1.2**)).[63]

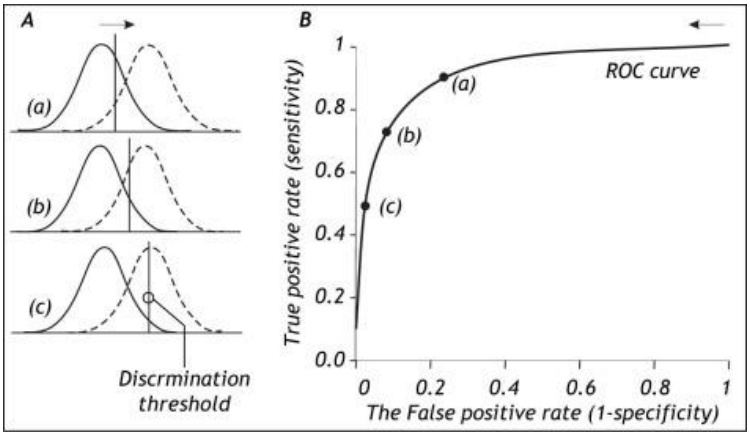



**Figure 1.2 from E. Christensen[63] Relationship between risk factor distributions and the ROC curve. (A)** Three identical unaffected and affected distributions with different discrimination thresholds. The discrimination threshold is the arbitrary cut-point value chosen for a predictive test. It is a compromise between the DR (sensitivity) of disease and the FPR (1-specificity). **(B)** The three cut-point values from the distributions in panel A are annotated on the corresponding ROC curve. The ROC curve is a visual summary of all the potential cut-point values that exist for the distributions. The higher the DR a cut-point allows, the higher the FPR. The shape of the ROC curve is the same for all the distributions in panel A as the cut-points do not influence it, but rather the difference in the means and the variance of the distributions influence the shape of the ROC curve. DR = detection rate; FPR = false positive rate; ROC = receiver-operating characteristic.

The ROC curve is related to the distributions of affected and unaffected individuals (**Figure 1.3** from Janssens *et al.*[64]). The area under the receiver-operating curve (AUC), provides information on how far apart these two normal distributions are from each other. The higher the mean predicted risk among affected versus unaffected, the better the separation of the distribution. A higher degree of separation between the risk distributions means that a model is better able to discriminate between affected and unaffected individuals, and is therefore a more useful prediction model (**Figure 1.3.a**).[64] The AUC represents the area between the ROC curve and the diagonal: the better the model, the further apart the two distributions are from each other, therefore the further apart the ROC curve is from the diagonal, and the closer the AUC is to 1 (**Figure 1.3.b**).[64] The AUC is often interpreted as the probability that an affected individual is correctly assigned as affected by the model compared to an unaffected individual.[53,65] For binary outcomes, the AUC and the concordance statistic (C-statistic or C-index) are equivalent and used synonymously.[53]

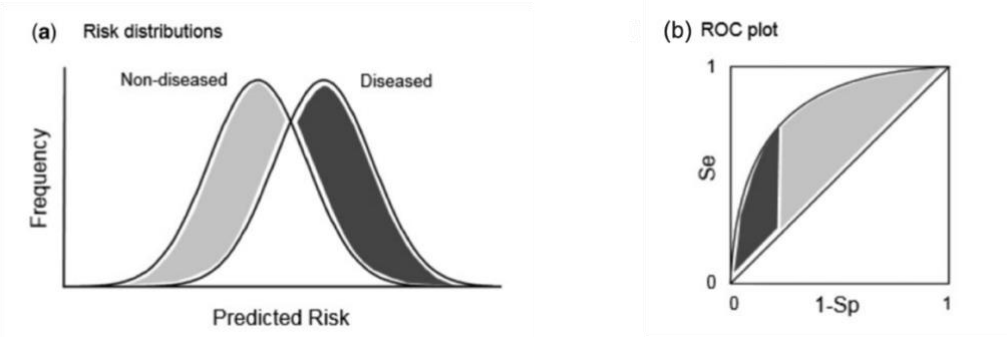



**Figure 1.3 from Janssens et al.[64] Evaluating the discriminative ability of a model to distinguish between affected and unaffected individuals. (a)** The risk distributions of affected and unaffected individuals given the predicted risk of a model. The overlap of both distributions is represented in white, and the non-overlapping regions are highlighted in two different shades of grey. The degree of separation between these distributions will have a direct effect on the size of the white and grey areas. **(b)** The corresponding ROC plot of the risk distributions in panel (a). The AUC is depicted by the grey areas. The degree of separation between the risk distributions influences the size of the grey areas, which determines the "height" of the ROC curve and the AUC. AUC = area under the receiver-operating curve; ROC = receiver-operating characteristic; Se = sensitivity; Sp = specificity.

The AUC is considered to be insensitive when it comes to adding a predictor to a model that already has good discrimination as the AUC will likely not visibly improve by much after a certain point.[66] This is because with every additional predictor that is added to a model, the separation of the means of the two distributions will increase, but the standard deviations (SD) will also increase, and hence the overlapping regions of the two distributions will not decrease as much as expected.[67] The effect of adding predictors to a model becomes increasingly modest with every new predictor, especially if they are not very strong. Model performance can still improve when adding predictors of small effect sizes, however a larger number of predictors will be needed to achieve the same discriminative ability as stronger predictors, and the improvement rate will also be much slower compared to predictors with large effect sizes.[58,68]

Another criticism of using the AUC as a measure of screening performance is that it covers all possible DR and FPR values for the test, including those that would never be considered useful in prediction.[69] Indeed, many studies with modest AUC values would never be considered as screening tests in clinical practice.[69] This is why it is important to specify the DR for different FPRs that would be considered acceptable.

### 1.2.5 Comparing models: net reclassification index (NRI) & decision curve analysis

The performance of a newly developed prediction model is often compared to that of other model(s) predicting the same outcome of interest. Usually, the models' discrimination metrics (AUC or C-index) are compared in a first instance, followed by a model classification analysis such as a net reclassification index (or net reclassification improvement) (NRI) analysis.

#### 1.2.5.1 NRI

A NRI analysis computes the probability that a model correctly and incorrectly reclassifies cases and controls compared to another model at a pre-specified threshold value.[70] This type of analysis is often done as it provides more tangible information regarding the impact of applying the prediction models in practice, which is not always concretely interpretable if only comparing the AUC or C-index of models.

However, the NRI has also been criticised as a metric. Studies have shown that the NRI tends to always be positive in large datasets even if a new variable added to a model is not predictive on its

own, likely due to overfitting.[71,72] The authors caution against using the NRI to evaluate the addition of a predictor to a pre-existing model. Another limitation of a NRI analysis is that it is performed at a specific cut-off value. This type of analysis does not provide an overview of the performance of the models at all possible threshold values, which might not always be required if the threshold value is set in practice.

#### 1.2.5.2 Decision curve analysis

A decision curve analysis overcomes this limitation of having to evaluate the performance of models based on a single pre-specific threshold value: a decision curve analysis computes the models' net benefit across all possible threshold values.[73] The net benefit of a model is calculated as the weighted difference between true and false positive cases (at all possible thresholds).[73] For example, a net benefit of 0.05 for a model at a specific probability threshold can be interpreted as overall 5 true positive cases detected (after subtracting all false positive cases from all true positive cases) for 100 individuals tested.[73] The decision curve analysis results are often plotted on a graph with the models' threshold values shown on the x-axis and the net benefit on the y-axis (**Figure 1.4** from Vikers *et al.*[73]). This provides a visual comparison of the changes in net benefit across all possible probability thresholds for various models.

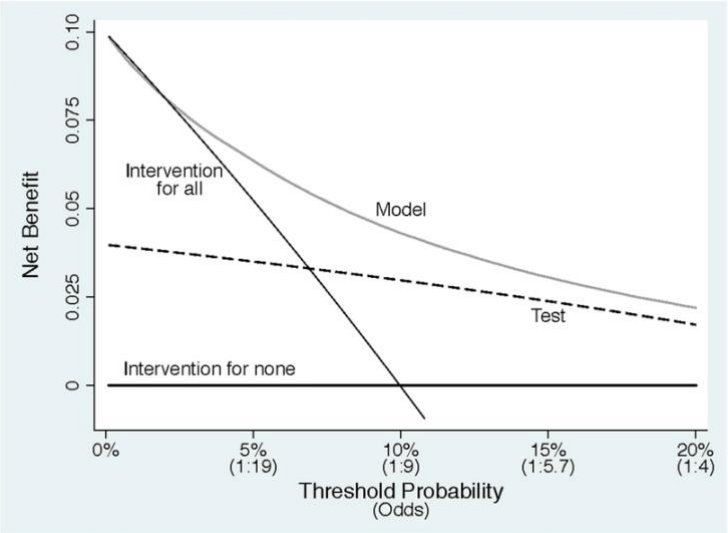



**Figure 1.4 from Vickers *et al.*[73] Decision curve analysis plot.** The threshold values are plotted on the x-axis against the net benefit of the models on the y-axis. This graph evaluates the net benefit of four possible scenarios for various probability thresholds: 1) intervention for none; 2) intervention for all; 3) model; 4) test.

### 1.2.6 Clinical utility

The metrics described previously provide a necessary measure of model performance, however, they do not provide tangible information as to the disease risk that they convey. The metrics more commonly used to assess the clinical usefulness of models are the DR (sensitivity), specificity (equal to 1 minus the FPR), and odds of being affected given a positive test result (OAPR).[74] These metrics require a pre-specified cut-off value for a model to classify individuals into "test positive" and "test negative" groups.

As mentioned in section 1.2.4 of the Introduction, the DR is a measure of the proportion of affected (diseased) individuals who test positive, and the FPR is a measure of the proportion of unaffected (non-diseased) individuals who incorrectly test positive (**Table 1.1**).[75] These metrics enable more comprehensible one-to-one comparisons of clinical models: the DR of tests can be directly compared if their FPR are fixed at a same level (e.g. comparing the DR of two tests for a 5% FPR). The likelihood ratio (LR) is calculated as the ratio between the DR and the FPR. This metric provides a relative measure of disease risk.[76]

| | **Disease +** | **Disease -** | *Total* | *Predictive values* |
|---|---|---|---|---|
| **Test +** | **80** | **45** | 125 | $OAPR = 80:45 \simeq 2:1$<br>$PPV = \frac{80}{80+45} \times 100 = 64\%$ |
| **Test -** | **20** | **855** | 975 | $NPV = \frac{855}{20+855} \times 100 = 97.7\%$ |
| *Total* | 100 | 900 | 1000 | $Absolute\ risk = \frac{100}{1000} \times 100 = 10\%$<br>$Odds\ before\ test = 1:9$ |
| *Sensitivity & specificity* | $DR = \frac{80 \times 100}{100} = 80\%$ | $FPR = \frac{45 \times 100}{900} = 5\%$ | | $LR = \frac{DR}{FPR} = \frac{80}{5} = 16$ |

**Table 1.1 Two by two table illustrating the relationship between the DR, FPR, OAPR, PPV, NPV, odds of disease, absolute risk and LR.** DR = detection rate; FPR = false positive rate; LR = likelihood ratio; NPV = negative predictive value; OAPR = odds of being affected given a positive test result; PPV = positive predictive value.

The OAPR is the positive predictive value (PPV) of a test expressed as an odd.[74] As illustrated in **Table 1.1**, it can be calculated as the ratio of affected to unaffected individuals with a positive test result (**Table 1.1**). The OAPR and PPV are examples of measures expressing disease risk in absolute terms. [74]

Indeed, disease risk can be expressed as a relative or an absolute measure. The relative risk of a disease refers to an estimate of the risk of one group in relation to another.[77] The absolute risk provides information on the individual or group risk of disease over a specified time (usually expressed as a probability or percentage).[77] Absolute risk is what clinical prediction models with multiple variables provide. When the DR and FPR are calculated, they can be converted to the absolute risk scale using transformation on average disease incidence rate (which is an absolute risk). This is done as follows:

- If the average absolute risk ($P$) over a specified period is 10%, the odds are equal to $1:9$ (where $odds = \frac{P}{1-P}$);
- The OAPR depends on the DR and FPR (or LR) of the test cut-off, such that $OAPR = DR \times 1 : FPR \times 9$ (or $OAPR = LR \times 1:9$) in this example.
    - For example, for a DR of 20% and a false positive of 5%, the $OAPR = 0.2 \times 1 : 0.05 \times 9 = 0.2 : 0.45 \simeq 1 : 2$
- The odds can be converted back to probability ($P$) using $P = \frac{odds}{odds+1}$, indicating an individual's probability of having the disease at that test cut-off.
    - In this example, $P = \frac{\frac{1}{2}}{\frac{1}{2}+1} = \frac{2}{6} = \frac{1}{3}$ or $\simeq 33\%$

While these measures are commonly used to assess the clinical utility of novel non-genetic prediction models or tests, as I show later in the thesis, PGS studies have yet to routinely incorporate these metrics when assessing their value in disease prediction.

### 1.3 Polygenic risk scores (PRS) in disease prediction

The terms "polygenic scores" (PGS), "polygenic risk scores" (PRS), "genetic scores" and "genetic risk scores" are used interchangeably in the literature. I will use the term PGS when referring to

them in a broader context of a genetic score for a trait (continuous or binary), and PRS when referring to them as risk scores for a binary disease endpoint.

### 1.3.1 Generating polygenic scores (PGS)

Over the past 15 years, GWAS have uncovered many single nucleotide polymorphisms (SNPs) associated with complex biological traits and common diseases. Each SNP has a small influence on its own, however, the idea behind PGS is that the aggregation (or burden) of these common variants in any individual's genome has a more substantial influence on an observable trait effect when pooled together.

#### 1.3.1.1 Weighted and unweighted scores

PGS are constructed from the sum of independent genetic variants associated with a trait (e.g. height, breast cancer, schizophrenia, educational attainment) in an individual's genome. The genetic variants used in the score are obtained from the summary statistics of a published GWAS for the trait. Two main types of PGS can be constructed: weighted and unweighted scores.[60]

$$PGSunweighted = \sum_{k=1}^{K} g_k \qquad PGSweighted = \sum_{k=1}^{K} \beta_k g_k \qquad \begin{matrix} g \in \{0, 1, 2\} \\ \beta \in \mathbb{R} \\ k \in \{1, \dots, K\} \end{matrix}$$

Unweighted scores require a simple addition of the total number of "risk" or trait alleles ($K$) obtained from the GWAS summary statistics; with 0 having no risk allele ($g = 0$), 1 having one risk allele ($g = 1$), and 2 having both risk alleles ($g = 2$). Weighted scores follow this same principle but have a specific genetic effect size estimate ($\beta_k$) associated with each variant ($g_k$), meaning that each genetic variant will carry a different weight in the overall score. The premise behind PGS is the higher the overall score for a trait, the higher the risk (or chances) of having or getting the trait in question.

The construction of these scores assumes that the effect of risk alleles is additive, meaning that the more risk alleles someone has for a disease, the higher their risk of having the disease.[78] This

assumption is based on the independent assortment of variants according to Mendel's law. And according to the central limit theorem, a variable that is the sum of many independent effects should be normally distributed, meaning that the distribution of risk variants for any given trait in a population is Gaussian in nature. This implies that a given unit increase in a PGS produces the same proportional increase in disease risk.[60]

#### 1.3.1.2 Score parameters: linkage disequilibrium (LD) & p-value thresholds

It is believed that the power of current GWAS is not large enough to detect all the genetic variants that would pass the genome-wide significance threshold due to sample size limitations.[79,80] For this reason, the variants included in the scores do not necessarily need to pass the genome-wide significance threshold (p-value of $5x10^{-8}$).[60]

The genetic architecture of the human genome is such that genetic variants in close physical proximity tend to be inherited together because they are less prone to being separated at meiosis by genetic recombination. Such variants are said to be in LD, and the extent of LD between pairs of variants is typically denoted by the squared coefficient of correlation ($r^2$). When calculating PGS, LD structure is accounted for to prevent score inflation.[60] Score inflation can arise from erroneously including multiple linked genetic variants due to LD in the score (the equivalent of counting a single variant multiple times). Clumping is a commonly used technique to account for LD structure. This method selects the most significant genetic variant per LD region defined.[60] Accounting for LD can also potentially lead to the removal genetic variants that are in LD with the most significant genetic variant of the LD region defined, but that are also independently associated with the studied trait.[68] This would lead to loss of information, and this is also a reason why multiple LD cut-offs are usually tested when generating PGS.[60]

Currently, there is no gold standard method when it comes to selecting which genetic variants should be included in a PGS. In practice, multiple scores are generated using different p-value, LD and MAF cut-off values to prevent underfitting of the prediction models.[60] Underfitting happens when the variables of a prediction model do not sufficiently predict the outcome of interest; usually because there are not enough useful predictors in the model. Testing different p-value, LD and MAF cut-off values ensures that the PGS generated include enough predictive genetic variants to prevent this issue of underfitting.

PGS can be constructed from a few tens of variants to millions of variants depending on the parameters chosen to generate the scores. The PGS that best predicts the outcome of interest is most commonly chosen based on the AUC (for binary endpoints) or $R^2$ (for continuous measures) of the score.

#### 1.3.1.3 Dataset considerations

PGS have emerged because of the increasing sample size of GWAS, leading to more discovered associations between genetic variants and traits, as well as bigger longitudinal cohort studies that have made the testing of these scores possible. Disease GWAS, which are used to discover genetic variants, are often case-control in design. These study designs cannot be used to evaluate the performance of predictors of future risk as they are cross-sectional.[81] To do so requires longitudinal cohort studies of initially healthy participants. There is a long history of such studies, and the UK Biobank is one of the most recent and largest.[82]

PGS are also often derived and then applied to an independent dataset (the target dataset) to avoid inflation of the observed effects.[83] PGS can be recalibrated in a subset of the independent data (the training data) and then applied to the rest of the independent data (the test data) to improve prediction (**Figure 1.5**). PGS can also be tested in an external dataset.

PGS tend to have poor predictive performance across diverse ancestries due to differences in allele frequencies and LD structure.[84,85] It is recommended to use GWAS summary statistics from the same ancestry background as the target dataset to prevent this issue.[60]

The unit of the final PGS depends on the unit of the genetic variants' weights in the GWAS they were derived from.[60] For example, a GWAS on height might express its weights in centimetres, which means that the final PGS are also expressed in centimetres. If the weights of the GWAS are mean centred and predict risk in terms of SD, the weights can then be multiplied by the SD of the GWAS trait, and the mean trait value can subsequently be added to the scores to convert them into the same unit as the trait in question. For case-control GWAS, the weights are usually expressed as log odds ratios (OR), and the PGS are derived relative to an individual who does not have any risk increasing alleles.[60] PGS therefore give a measure of relative risk rather than absolute risk.

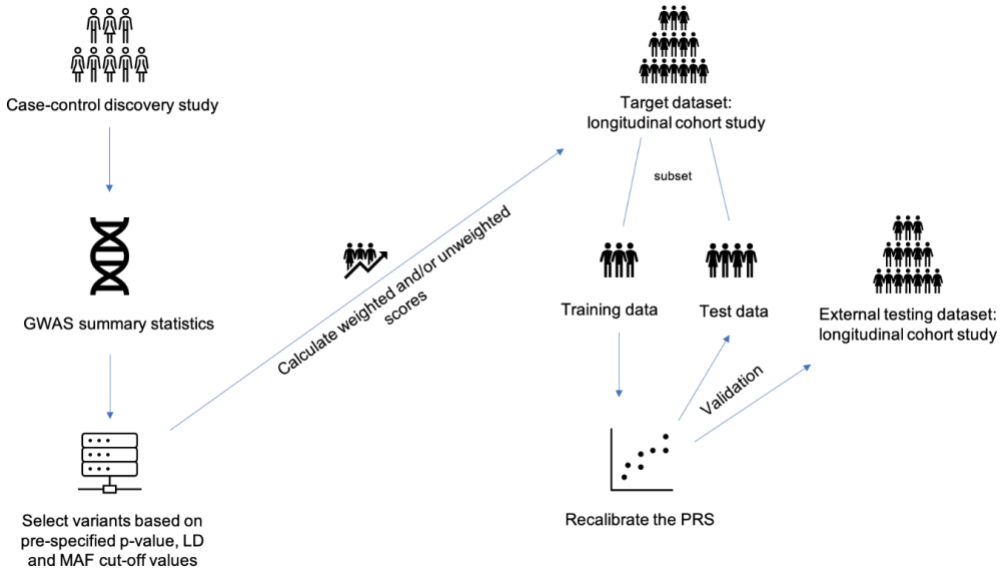



**Figure 1.5 Workflow for generating trait specific PGS and applying them to a target dataset.** Typically, the largest published GWAS summary statistic for the trait is chosen and multiple PGS are generated using different tuning parameters (p-value and LD cut-off thresholds). The PGS are tested in the target population (training dataset) and the PGS that best predicts the outcome of interest is chosen. The scores are then validated in a test dataset, which is either a subset of the target population (different from the training data), or an external dataset.

### 1.3.2 The prediction of complex heritable traits and incident diseases using PRS

One of the first PRS studies showed that the burden of common variants with small effect sizes play a role in the complex polygenic nature of schizophrenia and bipolar disorder.[86] Weighted PRS, or "allele scores", were generated using five different p-value thresholds that were not genome-wide significant (p-values of less than or equal to 0.1, 0.2, 0.3, 0.4 and 0.5). The genetic scores explained about 3% of the variance observed in schizophrenia; and cases had higher scores than controls. Nevertheless, the authors concluded that these scores did not have much clinical value in individual risk prediction. Since then, there have been numerous studies looking at the predictive performance of PRS in various other heritable complex diseases such as CVD and cancers, discussed in more detail later. A landmark paper showed that PRS can account for disease risk comparable to monogenic mutations for coronary artery disease (CAD), AF, T2D, inflammatory bowel disease and breast cancer.[87]

The main advantage of PRS is that they can be constructed early in a person's life as their genetic makeup remains unchanged throughout their lifetime, instead of restricting risk prediction to a 10-

year time frame later in life as is currently the case with most non-genetic risk models. This implies that people at high risk of a disease could potentially be targeted earlier than is currently done by some screening programs or intervention strategies. This approach would be extremely beneficial for people at risk of a premature CHD event for example, as they would greatly benefit from early lipid- or BP-lowering therapy.[81] Khera *et al.* developed a PGS for predicting and tracing BMI over a person's lifetime, suggesting a potential for early intervention and prevention strategies in individuals with a high PGS for BMI.[88] Another study showed that PRS could potentially inform the age of onset of a disease: PRS developed for five diseases (CHD, T2D, AF, breast cancer and prostate cancer) all showed a positive correlation between PRS and age of disease onset, although the mean age at diagnosis among those with the highest PRS was higher than the age at which national screening programmes are done (>67 for CHD, T2D, AF, breast and prostate cancer).[89] These findings point towards a potential use of PRS in targeted screening and prevention, and maybe also potentially delaying (or optimising) the screening age of certain conditions.

### 1.3.3 Comparing the predictive performance of clinical risk prediction models and PRS

The current literature on PRS shows that these scores exhibit an association with a variety of complex traits and disorders. However, associations with a significant p-value are not a guarantee that a risk score will be usefully predictive of disease.[38] If risk scores exhibit a relatively weak log-linear relationship with risk and are normally distributed, their distributions among those with and without disease will be highly overlapping.[33] No cut-point readily separates those with from those without disease. And while the OR between the top versus the bottom x% of the distribution might seem large (as is often reported in PRS publications), this way of presenting the data is often misleading as it does not necessarily translate to a high OR per SD.[90] The OR per SD has to be very high in order to provide useful discrimination.[33]

Furthermore, PRS and non-genetic clinical scores assess risk in different ways. PRS are a measure of relative risk over a lifetime for a particular trait or disease either based on the population studied or on a hypothetical individual who does not carry any risk increasing alleles.[60] They are not a measure of absolute disease risk unlike non-genetic clinical models that estimate the absolute risk of developing a disease over a defined period of time (e.g. QRISK3 predicts the 10-year risk of developing CVD), although this field of research is rapidly evolving.[43,91]

Then comes the issue of whether PRS add to existing non-genetic risk prediction tools. To assess this, their ability to predict future disease must be compared to the performance of current clinical risk prediction tools. So far, the general consensus seems to be that PRS do not necessarily perform better than traditional risk scores, but that prediction may be improved if they are combined. Inouye *et al.*'s metaGRS for CAD showed improved model discrimination based on the C-statistic when combined with six conventional risk factors for CAD.[92] Abraham *et al.* also developed a PRS for CHD that showed an increase of 0.016 in the C-statistic when combined with the FRS, and of 0.015 when combined with the ACC/AHA13 risk score.[93] A 53 genetic variant score for CVD improved the C-statistic of QRISK2 by 0.012.[94] And more recently, a study comparing QRISK3 to a PRS for CAD in the UK Biobank showed a significant although very modest added value of genetic information in risk prediction (an improvement of 0.02 in the C-statistic).[95] Similar results were obtained in FinnGen when comparing a PRS for CHD to the ASCVD risk calculator (C-statistic improvement of 0.003), a PRS for T2D to T2D risk factors (C-statistic improvement of 0.010), and a PRS for AF to CHARGE-AF (C-statistic improvement of 0.009).[89] Adding genetic information to clinical risk factors for IST also improved the ability to predict a future IST event (improvement of about 0.01 in the C-index); and the same was observed when adding a T2D PRS to the Framingham Offspring T2D risk score (improvement of 0.01 in the AUC).[96,97]

These incremental increases in the C-statistic are all very modest, and these studies have not evaluated the clinical utility of these new models using the appropriate clinically useful metrics of sensitivity, specificity and OAPR, as described in section 1.2.5 of the Introduction. Better research is needed to properly examine the utility of PRS in clinical risk prediction, screening, and risk stratification before their use in clinical practice can be recommended.

### 1.3.4 Leveraging polygenic information for the discovery of rare genetic variants

As seen previously, PGS are a measure of allelic burden for a trait or disease that are constructed from common variants. These scores have until now primarily been used to help predict the risk of common heritable diseases, but these scores could also be exploited to predict the value of a quantitative trait. This information might then hold value for identifying individuals harbouring rare genetic variants of large effect size for a trait.[98,99] The idea behind this being that individuals in whom the observed value of a continuous trait exceeds that predicted by their common allele burden (i.e. PGS) might be more likely to carry a rare variant of large effect size that would account

for this discrepancy. An example of this are individuals with LDL-C levels far exceeding those predicted by their LDL-C PGS who could be potential carriers of a monogenic variant for familial hypercholesterolaemia, which is explored in more detail in Chapter 6.[99]

## 1.4 Familial hypercholesterolaemia (FH)

### 1.4.1 Overview of FH

Familial hypercholesterolaemia (FH) is a monogenic disorder resulting from mutations in the *LDLR, APOB, PCSK9, APOE* and *LDLRAP1* genes, typically characterised by elevated LDL-C levels.[100,101] Individuals with FH have a 3 to 22 fold increased risk of CAD depending on age, and patients benefit from lipid-lowering therapies which reduce the risk of premature coronary events (**Figure 1.6.A** from Versmissen *et al.*[102]).[103,104]

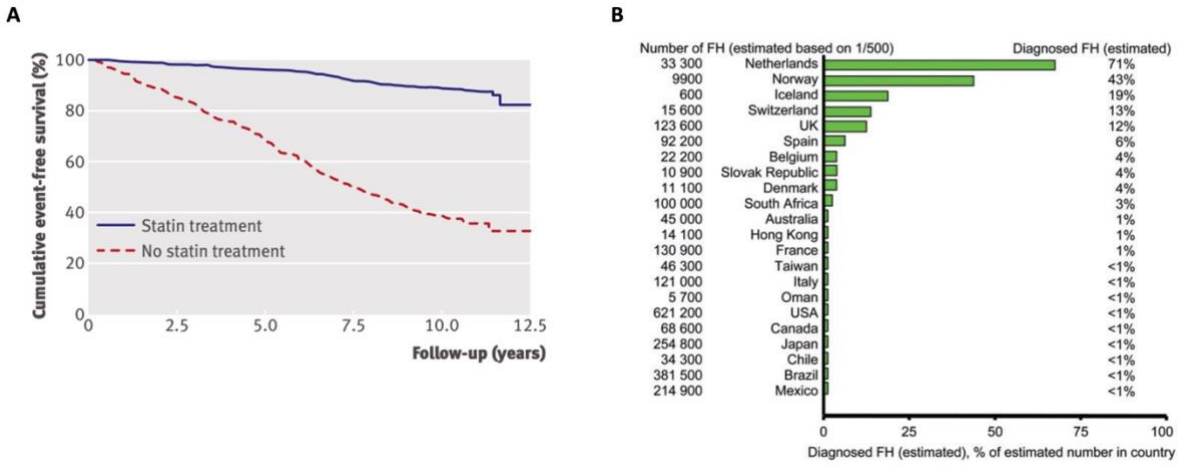



**Figure 1.6 FH prevalence and the effects of statin treatment on survival. (A)** from Versmissen *et al.*[102]: The cumulative CHD-free survival (%) for FH cases with and without statin treatment for a follow-up of 12.5 years. The blue curve (statin treatment) shows a higher cumulative event-free survival than the dotted red curve (no statin treatment). **(B)** from Nordestgaard *et al.*[107]: The estimated number of diagnosed FH cases (based on a prevalence of 1/500) per selected countries as of 2013. CHD = coronary heart disease; FH = familial hypercholesterolaemia.

The estimated population prevalence of FH is 1 in 250, making it the most common monogenic disorder worldwide.[105,106] Despite this, FH remains highly underdiagnosed in most countries (**Figure 1.6.B** from Nordestgaard *et al.*[107]). Increasing the identification of FH cases in the UK

has been listed as one of the objectives in the National Health Service's (NHS) Long Term Plan, but there is currently no proposal as to how this will be achieved.[108]

### 1.4.2 Genetics of FH

The pathogenic genetic variants responsible for FH are located in and near the *LDLR, APOB, PCSK9, APOE* and *LDLRAP1* genes, all of which encode proteins that play a role in lipid metabolism.[107] The low-density lipoprotein (LDL) receptor (LDLR) is a cell surface protein mainly expressed in the liver, responsible for the uptake of LDL-C from the blood, ultimately reducing circulating LDL-C concentration (**Figure 1.7** from Soutar & Naoumova[109]).[110]

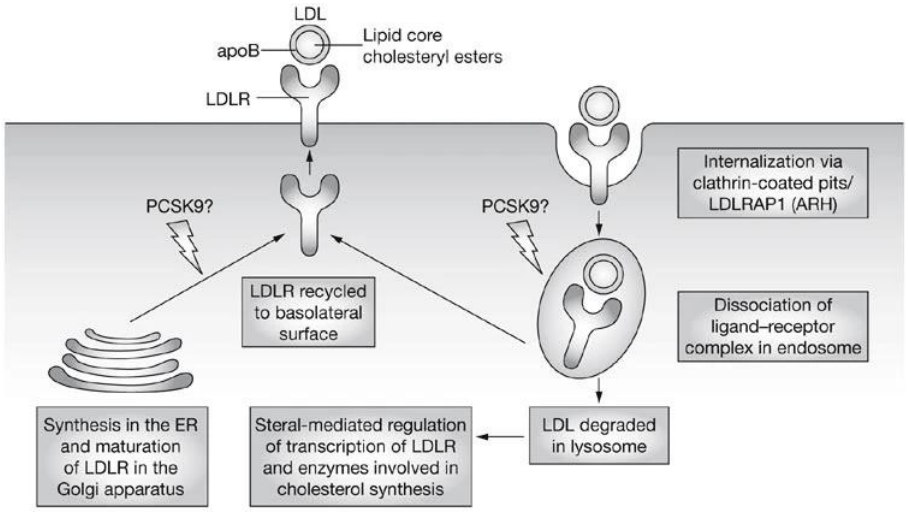



**Figure 1.7 from Soutar & Naoumova.[109] The LDLR-mediated cellular uptake of LDL.** The cytoplasmic internalisation of LDL is mediated through LDLR, a cell-surface protein, with the help of Apo-B and LDLRAP1. Once internalised, PCSK9 targets LDLR for lysosomal degradation.

Proprotein convertase subtilisin/kexin Type 9 (PCSK9) is a cytoplasmic protein that is secreted. It binds the LDLR on the cell surface and targets it for lysosomal degradation (**Figure 1.7**).[111] Inhibiting the function of PCSK9 prevents this process, leads to the retention of LDLRs on the cell surface, and consequently leads to a decrease in circulating LDL-C concentration. This effect represents the intended mechanism of action of PCSK9 inhibitor drugs for treating hypercholesterolaemia.[112]

Apolipoprotein B (Apo-B) and apolipoprotein E (APOE) are located on lipoprotein particles such as LDL (**Figure 1.7**).[113] FH-causing variants in *APOB* disrupt the binding affinity of LDL to LDLR, leading to hypercholesterolaemia.[114] For *APOE*, the gain-of-function FH-causing variant p.Leu167del increases the binding affinity of APOE to LDLR and thereby prevents the effective recycling of LDLR back to the cell surface, leading to hypercholesterolaemia.[115]

Homozygous or compound heterozygous mutations in *LDLRAP1,* a gene encoding the LDLR adapter protein 1 (LDLRAP1, also known as ARH) which mediates the internalisation of LDL-LDLR complexes, have also be found to cause a recessive form of FH (**Figure 1.7**).[116,117] Due to its recessive mode of inheritance, parents and children of affected individuals do not display the hypercholesterolaemia phenotype, but affected individuals have severe hypercholesterolaemia requiring treatment by LDL-apheresis.[118]

The spectrum of FH-causing variants makes it difficult to develop a simple genotype-based test for FH, unless the particular variant responsible for FH in a family is seen. For this reason, many clinical genetics services have developed exome sequencing protocols to confirm or refute a diagnosis of FH in suspected cases.[119]

### 1.4.3 Clinical diagnosis of FH

Several criteria have been developed to help diagnose FH in clinical practice. Examples include the Simon Broome Criteria in the UK, the Dutch Lipid Clinical Network (DLCN) criteria in the UK and the Netherlands, and the Make Early Diagnosis to Prevent Early Death (MEDPED) criteria in the US.[120] These tools work similarly to risk calculators: they provide a measure of certainty for having FH. The variables included in these scores vary but they all incorporate a measure of either LDL- or total cholesterol concentration (**Table 1.2**).[120] Genomic sequencing of suspected FH cases is necessary to confirm the diagnosis.

| | Simon Broome Criteria | Dutch Lipid Clinical Network (DLCN) | Make Early Diagnosis to Prevent Early Death (MEDPED) |
|---|---|---|---|
| Age | | | |
| Total cholesterol | >7.5 mmol/L | | |
| LDL-C | >4.9 mmol/L | >4.0 mmol/L | |
| Total cholesterol/LDL-C ratio | | | Age-dependent (from 5.7/4.0 mmol/L) |
| DNA-based evidence of a function *LDLR*, *PCSK9* and *APOB* mutation | | | |
| **Clinical history/examination** | | | |
| Tendon xanthomas | | | |
| Premature coronary heart disease | | | |
| Premature cerebral or peripheral vascular disease | | | |
| Arcus cornealis <45 years of age | | | |
| **Family history** | | | |
| Family history of premature CVD events | | | |
| Family history of extremely high cholesterol | | | |
| Family history of tendon xanthomas | | | |
| Family history of FH | | | |

**Table 1.2 The FH diagnostic variables included in the Simon Broome, Dutch Lipid Clinical Network, and the Make Early Diagnosis to Prevent Early Death criteria.** The shaded boxes in grey indicate the variables used in each score. CVD = cardiovascular disease; FH = familial hypercholesterolaemia; LDL-C = low-density lipoprotein cholesterol.

### 1.4.4 Cascade testing of index FH cases

Since FH is primarily an autosomal dominant disorder, first-degree relatives of patients have a 50% chance of being affected. Once a new index FH case has been identified and confirmed via genetic sequencing, the FH-causing genetic variant is cascade tested in close relatives (first, second, and third degree if possible) using simpler mutation detection approaches, as recommended by the National Institute for Health and Care Excellence (NICE).[121] This approach has been shown to be cost-effective in the UK, however, it relies on a means of identifying new index cases.[122] Currently, FH index case detection is opportunistic. For this reason, there is much interest in considering cost-effective approaches for FH population screening.

### 1.4.5 Population screening of FH

Currently, new FH cases are detected when individuals are found to have an elevated LDL or total cholesterol level, usually when blood tests are organised for another reason. Other patients are identified only after they suffer a CHD event at an early age. There is no systematic way of identifying new FH cases in the UK population, even though the NHS Long Term Plan published in 2019 stated that it aims to increase the diagnosis of FH cases from 7% to 25% in the next five years.[108]

Some studies have proposed population screening for FH, such as the child-parent screening strategy by Wald *et al.*[123] This involves measuring 15-month-old children's cholesterol levels via a blood spot while attending routine immunisation. Children with elevated cholesterol levels (e.g. >1.53 multiple of the median (MoM) for total cholesterol, or >1.84 MoM for LDL-C) are considered to be potential FH cases.[123] Their parents' cholesterol levels are subsequently measured, and the parent with the highest cholesterol concentration is considered to be the affected parent. Cascade screening is then initiated in close relatives (other children, siblings, and parents) of index cases for further FH case identification. This strategy was validated by Futema *et al.,* and similar ones are being trialled in Bavaria and Slovenia.[124–126] Despite its merits, the child-parent population screen for FH was rejected as an approach by the UK National Screening Committee because of concerns about the screened population being children who are not immediately eligible for lipid-lowering therapies.[127] Nevertheless, a pilot Child-Parent Screening Service is underway in 30,000 children aged 1 to 2 years old over a course of 24 months in select testing sites affiliated with the UK Academic Health and Science Network.[128,129]

A recent cost-effectiveness study from Australia also suggested that the genomic sequencing of all young adults for FH was cost-effective, assuming reasonably priced tests (<AU$250).[130] This type of cost-effectiveness study has not been done in the UK, and other more cost-effective screening strategies for FH in the general adult population have also yet to be explored. Considering the potential applications of PGS in rare variant discovery, PGS could play a role in population screening to identify FH cases in the general population.[98,99]

## 1.5 Thesis overview

The aim of this thesis was to evaluate the clinical utility of PGS in CVD prediction and screening. In the first part of the thesis (Chapter 2), I undertook an overview of the performance of the PGS reported in the Polygenic Score Catalog.

I then performed my own PGS analyses using data from the UK Biobank as it is the largest and most current longitudinal cohort study of the UK population to date, and it contains all the information needed to perform the various analyses detailed in this thesis. Chapter 3 describes the cohort study in more detail and the various quality control steps that were performed prior to utilising the dataset for the analyses.

Chapter 4 analyses the incremental predictive utility of PGS when added to the non-genetic clinical risk prediction calculators QRISK3, QStroke and QDiabetes for the 10-year prediction of incident CVD/CHD, IST and T2D respectively. It addresses some limitations of previous PGS studies and provides further evidence as to the clinical utility (or otherwise) of PGS in CVD prediction.

Chapter 5 introduces a new adult population screening strategy for FH and compares its performance to that of the child-parent screening strategy proposed by Wald *et al.*[123] This chapter addresses the current issues faced with the underdiagnosis of FH in the UK and provides a solution to address the target of identifying 25% of all FH cases as indicated in the NHS Long Term Plan.

Chapter 6 builds on Chapter 5 by developing a novel prediction model for detecting monogenic FH using machine learning (LASSO) to improve the efficiency of the two-stage adult population screen. The prediction model incorporates use of an LDL-C PGS together with other variables. This chapter provides further evidence as to the possible clinical applications of PGS in aiding monogenic FH detection and screening.

The thesis ends with a discussion in Chapter 7 on the various analyses undertaken throughout the thesis, and on the impact and implications that these have with regards to PGS in CVD prediction and screening.

## 1.6 References


1 World Health Organization. Cardiovascular diseases. 2020.https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1 (accessed 9 Nov 2020).

2 Health Intelligence Team B. BHF UK CVD Factsheet. 2022.

3 Peter Libby, Julie E. Buring, Lina Badimon, Göran K. Hansson, John Deanfield, Márcio Sommer Bittencourt LT& EFL. Atherosclerosis. *Nat Rev Dis Primers* 2019;**5**:57. doi:10.1038/s41572-019-0116-x

4 MacMahon S, Duffy S, Rodgers A, *et al.* Blood cholesterol and vascular mortality by age, sex, and blood pressure: a meta-analysis of individual data from 61 prospective studies with 55 000 vascular deaths. *The Lancet* 2007;**370**:1829–39. doi:10.1016/S0140-6736(07)61778-4

5 Ference BA, Yoo W, Alesh I, *et al.* Effect of Long-Term Exposure to Lower Low-Density Lipoprotein Cholesterol Beginning Early in Life on the Risk of Coronary Heart Disease: A Mendelian Randomization Analysis. *J Am Coll Cardiol* 2012;**60**:2631–9. doi:10.1016/J.JACC.2012.09.017

6 Baigent C, Blackwell L, Emberson J, *et al.* Efficacy and safety of more intensive lowering of LDL cholesterol: a meta-analysis of data from 170 000 participants in 26 randomised trials. *The Lancet* 2010;**376**:1670–81. doi:10.1016/S0140-6736(10)61350-5

7 Brundel BJJM, Ai X, Hills MT, *et al.* Atrial fibrillation. *Nature Reviews Disease Primers 2022 8:1* 2022;**8**:1–23. doi:10.1038/s41572-022-00347-9

8 Campbell BC V, De Silva DA, Macleod MR, *et al.* Ischaemic stroke. *Nat Rev Dis Primers* 2019;**5**:70. doi:10.1038/s41572-019-0118-8

9 Chung MK, Eckhardt LL, Chen LY, *et al.* Lifestyle and Risk Factor Modification for Reduction of Atrial Fibrillation: A Scientific Statement from the American Heart Association. *Circulation* 2020;**141**:E750–72. doi:10.1161/CIR.0000000000000748/FORMAT/EPUB

10 Ziaeian B, Fonarow GC. Epidemiology and aetiology of heart failure. *Nature Reviews Cardiology 2016 13:6* 2016;**13**:368–78. doi:10.1038/nrcardio.2016.25

11 Eckel RH, Jakicic JM, Ard JD, *et al.* 2013 AHA/ACC guideline on lifestyle management to reduce cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *J Am Coll Cardiol* 2014;**63**:2960–84. doi:10.1016/J.JACC.2013.11.003

12 Piepoli MF, Hoes AW, Agewall S, *et al.* 2016 European Guidelines on cardiovascular disease prevention in clinical practice: The Sixth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts)Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). *Eur Heart J* 2016;**37**:2315–81. doi:10.1093/EURHEARTJ/EHW106

13 Timmis A, Townsend N, Gale C, *et al.* European Society of Cardiology: Cardiovascular Disease Statistics 2017. *Eur Heart J* 2018;**39**:508–77. doi:10.1093/EURHEARTJ/EHX628

14 Wong ND. Epidemiological studies of CHD and the evolution of preventive cardiology. Nat Rev Cardiol. 2014;**11**:276–89. doi:10.1038/nrcardio.2014.26

15 Wong ND, Levy D. Legacy of the Framingham Heart Study: Rationale, Design, Initial Findings, and Implications. *Glob Heart* 2013;**8**:3. doi:10.1016/j.gheart.2012.12.001

16 DAWBER TR, MOORE FE, MANN G V. Coronary heart disease in the Framingham study. *Am J Public Health Nations Health* 1957;**47**:4–24. doi:10.2105/ajph.47.4_pt_2.4

17 Kannel WB, McGee DL. Diabetes and Cardiovascular Disease: The Framingham Study. *JAMA: The Journal of the American Medical Association* 1979;**241**:2035–8. doi:10.1001/jama.1979.03290450033020

18 DOYLE JT, DAWBER TR, KANNEL WB, *et al.* Cigarette smoking and coronary heart disease. Combined experience of the Albany and Framingham studies. *N Engl J Med* 1962;**266**:796–801. doi:10.1056/NEJM196204192661602

19 Wilson PWF, D'Agostino RB, Levy D, *et al.* Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;**97**:1837–47. doi:10.1161/01.CIR.97.18.1837

20 WHO MONICA Project Principal Investigators. The World Health Organization MONICA Project (monitoring trends and determinants in cardiovascular disease): a major international collaboration. WHO MONICA Project Principal Investigators. *J Clin Epidemiol* 1988;**41**:105–14. doi:10.1016/0895-4356(88)90084-4

21 Yusuf PS, Hawken S, Ôunpuu S, *et al.* Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): Case-control study. *Lancet* 2004;**364**:937–52. doi:10.1016/S0140-6736(04)17018-9

22 Marmot MG, Stansfeld S, Patel C, *et al.* Health inequalities among British civil servants: the Whitehall II study. *The Lancet* 1991;**337**:1387–93. doi:10.1016/0140-6736(91)93068-K

23 The Reykjavík Study | Hjarta. https://hjarta.is/en/research/reykjavikurrannsokn/ (accessed 20 Oct 2022).

24 Atherosclerosis Risk in Communities Study Description. https://sites.cscc.unc.edu/aric/description (accessed 20 Oct 2022).

25 Danesh J, Erqou S, Walker M, *et al.* The Emerging Risk Factors Collaboration: analysis of individual data on lipid, inflammatory and other markers in over 1.1 million participants in 104 prospective studies of cardiovascular diseases. *Eur J Epidemiol* 2007;**22**:839–69. doi:10.1007/S10654-007-9165-7

26 Iso H, Sato H, Chambless L, *et al.* Collaborative overview ('meta-analysis') of prospective observational studies of the associations of usual blood pressure and usual cholesterol levels with common causes of death: protocol for the second cycle of the Prospective Studies Collaboration. *J Cardiovasc Risk* 1999;**6**:315–20. doi:10.1177/204748739900600508

27 Ference BA, Ginsberg HN, Graham I, *et al.* Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement fromthe European Atherosclerosis Society Consensus Panel. *Eur Heart J* 2017;**38**:2459–72. doi:10.1093/eurheartj/ehx144

28 Pirie K, Peto R, Reeves GK, *et al.* The 21st century hazards of smoking and benefits of stopping: A prospective study of one million women in the UK. *The Lancet* 2013;**381**:133–41. doi:10.1016/S0140-6736(12)61720-6

29 Wan EYF, Yu EYT, Chin WY, *et al.* Association of Blood Pressure and Risk of Cardiovascular and Chronic Kidney Disease in Hong Kong Hypertensive Patients. *Hypertension* 2019;**74**:331–40. doi:10.1161/HYPERTENSIONAHA.119.13123

30 Oparil S, Acelajado MC, Bakris GL, *et al.* Hypertension. *Nat Rev Dis Primers* 2018;**4**:18014. doi:10.1038/nrdp.2018.14

31 KANNEL WB, DAWBER TR, KAGAN A, *et al.* Factors of risk in the development of coronary heart disease--six year follow-up experience. The Framingham Study. *Ann Intern Med* 1961;**55**:33–50. doi:10.7326/0003-4819-55-1-33

32 Lau DH, Nattel S, Kalman JM, *et al.* Modifiable Risk Factors and Atrial Fibrillation. *Circulation* 2017;**136**:583–96. doi:10.1161/CIRCULATIONAHA.116.023163

33 Wald NJ, Hackshaw AK, Frost CD. When can a risk factor be used as a worthwhile screening test? *BMJ : British Medical Journal* 1999;**319**:1562. doi:10.1136/BMJ.319.7224.1562

34 Burgess S, Foley CN, Zuber V. Inferring Causal Relationships Between Risk Factors and Outcomes from Genome-Wide Association Study Data. *Annu Rev Genomics Hum Genet* 2018;**19**:303. doi:10.1146/ANNUREV-GENOM-083117-021731

35 Law MR, Wald NJ. Risk factor thresholds: their existence under scrutiny. *BMJ : British Medical Journal* 2002;**324**:1570. doi:10.1136/BMJ.324.7353.1570

36 Rose G. Sick individuals and sick populations. *Int J Epidemiol* 2001;**30**:427–32. doi:10.1093/ije/30.3.427

37 Geoffrey R. Strategy of prevention: lessons from cardiovascular disease. *Br Med J (Clin Res Ed)* 1981;**282**:1847. doi:10.1136/BMJ.282.6279.1847

38 Wald NJ, Hackshaw AK, Frost CD. When can a risk factor be used as a worthwhile screening test? *BMJ* 1999;**319**:1562. doi:10.1136/bmj.319.7224.1562

39 Hingorani AD, Psaty BM. Primary prevention of cardiovascular disease: time to get more or less personal? *JAMA* 2009;**302**:2144–5. doi:10.1001/JAMA.2009.1698

40 Hingorani AD, Hemingway H. How should we balance individual and population benefits of statins for preventing cardiovascular disease? *BMJ* 2011;**342**:313–5. doi:10.1136/BMJ.C6244

41 Damen JAAG, Hooft L, Schuit E, *et al.* Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* 2016;**353**. doi:10.1136/BMJ.I2416

42 WHO/Europe | Cardiovascular diseases - Data and statistics. https://www.euro.who.int/en/health-topics/noncommunicable-diseases/cardiovascular-diseases/data-and-statistics (accessed 2 Jul 2020).

43 Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: Prospective cohort study. *BMJ (Online)* 2017;**357**. doi:10.1136/bmj.j2099

44 Vartiainen E, Laatikainen T, Peltonen M, *et al.* Predicting Coronary Heart Disease and Stroke: The FINRISK Calculator. Glob Heart. 2016;**11**:213–6. doi:10.1016/j.gheart.2016.04.007

45 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. doi:10.1016/S0195-668X(03)00114-3

46 Ridker PM, Paynter NP, Rifai N, *et al.* C-reactive protein and parental history improve global cardiovascular risk prediction: The Reynolds risk score for men. *Circulation* 2008;**118**:2243–51. doi:10.1161/CIRCULATIONAHA.108.814251

47 Ridker PM, Buring JE, Rifai N, *et al.* Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: The Reynolds Risk Score. *J Am Med Assoc* 2007;**297**:611–9. doi:10.1001/jama.297.6.611

48 Hippisley-Cox J, Coupland C, Brindle P. Derivation and validation of QStroke score for predicting risk of ischaemic stroke in primary care and comparison with other risk scores: A prospective open cohort study. *BMJ (Online)* 2013;**346**. doi:10.1136/bmj.f2573

49 Hippisley-Cox J, Coupland C. Development and validation of QDiabetes-2018 risk prediction algorithm to estimate future risk of type 2 diabetes: cohort study. *BMJ* 2017;**359**:j5019. doi:10.1136/bmj.j5019

50 Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 2014;**35**:1925–31. doi:10.1093/eurheartj/ehu207

51 Damen JAAG, Hooft L, Schuit E, *et al.* Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* 2016;**353**. doi:10.1136/BMJ.I2416

52 Naik G, Ahmed H, Edwards AGK. Communicating risk to patients and the public. *The British Journal of General Practice* 2012;**62**:213. doi:10.3399/BJGP12X636236

53 Steyerberg EW. *Clinical Prediction Models.* 1st ed. Springer-Verlag New York 2009. doi:10.1007/978-0-387-77244-8

54 Grant SW, Collins GS, Nashef SAM. Statistical Primer: developing and validating a risk prediction model. *European Journal of Cardio-Thoracic Surgery* 2018;**54**:203–8. doi:10.1093/EJCTS/EZY180

55 Wang J, Tan GJ, Han LN, *et al.* Novel biomarkers for cardiovascular risk prediction. *J Geriatr Cardiol* 2017;**14**:135. doi:10.11909/J.ISSN.1671-5411.2017.02.008

56 Chowdhury MZI, Turin TC. Variable selection strategies and its importance in clinical prediction modelling. *Fam Med Community Health* 2020;**8**:262. doi:10.1136/fmch-2019-000262

57 Hingorani AD, Sofat R, Morris RW, *et al.* Is it important to measure or reduce C-reactive protein in people at risk of cardiovascular disease? Eur Heart J. 2012;**33**:2258–64. doi:10.1093/eurheartj/ehs168

58 Shah T, Casas JP, Cooper JA, *et al.* Critical appraisal of CRP measurement for the prediction of coronary heart disease events: New data and systematic review of 31 prospective cohorts. Int J Epidemiol. 2009;**38**:217–31. doi:10.1093/ije/dyn217

59 Tibshirani R. THE LASSO METHOD FOR VARIABLE SELECTION IN THE COX MODEL. *Stat Med* 1997;**16**:385–95. doi:10.1002/(SICI)1097-0258(19970228)16:4

60 Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc. 2020;**15**:2759–72. doi:10.1038/s41596-020-0353-1

61 Steyerberg EW, Borsboom GJJM, van Houwelingen HC, *et al.* Validation and updating of predictive logistic regression models: A study on sample size and shrinkage. *Stat Med* 2004;**23**:2567–86. doi:10.1002/sim.1844

62 Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;**15**:361–87. doi:10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

63 Christensen E. Methodology of diagnostic tests in hepatology. Ann Hepatol. 2009;**8**:177–83. doi:10.1016/s1665-2681(19)31763-6

64 Cecile A, Janssens JW, Martens FK. Education Corner Reflection on modern methods: Revisiting the area under the ROC Curve Education Corner. *IEA International Epidemiological Association International Journal of Epidemiology*;**2020**:1–7. doi:10.1093/ije/dyz274

65 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;**143**:29–36. doi:10.1148/RADIOLOGY.143.1.7063747

66 Pepe MS, Janes HE. Gauging the Performance of SNPs, Biomarkers, and Clinical Factors for Predicting Risk of Breast Cancer. *J Natl Cancer Inst* 2008;**100**:978–9. doi:10.1093/jnci/djn215

67 Wald NJ, Morris JK, Rish S. The efficacy of combining several risk factors as a screening test. *J Med Screen* 2005;**12**:197–201. doi:10.1258/096914105775220642

68 Chatterjee N, Shi J, García-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat Rev Genet* 2016;**17**:392–406. doi:10.1038/nrg.2016.27

69 Wald NJ, Bestwick JP. Is the area under an ROC curve a valid measure of the performance of a screening or diagnostic test? *J Med Screen* 2014;**21**:51–6. doi:10.1177/0969141313517497

70 Leening MJG, Vedder MM, Witteman JCM, *et al.* Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide. *Ann Intern Med* 2014;**160**:122–31. doi:10.7326/M13-1522

71 Hilden J, Gerds TA. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. *Stat Med* 2014;**33**:3405–14. doi:10.1002/SIM.5804

72 Pepe MS, Fan J, Feng Z, *et al.* The Net Reclassification Index (NRI): A Misleading Measure of Prediction Improvement Even with Independent Test Data Sets. *Stat Biosci* 2015;**7**:282–95. doi:10.1007/S12561-014-9118-0/FULLTEXT.HTML

73 Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res* 2019;**3**:1–8. doi:10.1186/S41512-019-0064-7/FIGURES/3

74 Wald NJ, Morris JK. Assessing Risk Factors as Potential Screening Tests: A Simple Assessment Tool. *Arch Intern Med* 2011;**171**:286–91. doi:10.1001/ARCHINTERNMED.2010.378

75 Parikh R, Mathai A, Parikh S, *et al.* Understanding and using sensitivity, specificity and predictive values. *Indian J Ophthalmol* 2008;**56**:45. doi:10.4103/0301-4738.37595

76 McGee S. Simplifying Likelihood Ratios. *J Gen Intern Med* 2002;**17**:647. doi:10.1046/J.1525-1497.2002.10750.X

77 Noordzij M, van Diepen M, Caskey FC, *et al.* Relative risk versus absolute risk: one cannot be interpreted without the other. *Nephrology Dialysis Transplantation* 2017;**32**:ii13–8. doi:10.1093/NDT/GFW465

78 Visscher PM, Wray NR. Concepts and Misconceptions about the Polygenic Additive Model Applied to Disease. *Hum Hered* 2015;**80**:165–70. doi:10.1159/000446931

79 Dudbridge F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genet* 2013;**9**:e1003348. doi:10.1371/journal.pgen.1003348

80 Chatterjee N, Wheeler B, Sampson J, *et al.* Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat Genet* 2013;**45**:400–5. doi:10.1038/ng.2579

81 Holmes M V., Harrison S, Talmud PJ, *et al.* Utility of genetic determinants of lipids and cardiovascular events in assessing risk. *Nat Rev Cardiol* 2011;**8**:207–21. doi:10.1038/nrcardio.2011.6

82 Bycroft C, Freeman C, Petkova D, *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018;**562**:203–9. doi:10.1038/s41586-018-0579-z

83 Wray NR, Yang J, Hayes BJ, *et al.* Pitfalls of predicting complex traits from SNPs. *Nat Rev Genet* 2013;**14**:507–15. doi:10.1038/nrg3457

84 Kim MS, Patel KP, Teng AK, *et al.* Genetic disease risks can be misestimated across global populations. *Genome Biol* 2018;**19**:179. doi:10.1186/s13059-018-1561-7

85 Martin AR, Gignoux CR, Walters RK, *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am J Hum Genet* 2017;**100**:635–49. doi:10.1016/j.ajhg.2017.03.004

86 Purcell SM, Wray NR, Stone JL, *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 2009;**460**:748–52. doi:10.1038/nature08185

87 Khera A V., Chaffin M, Aragam KG, *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* 2018;**50**:1219–24. doi:10.1038/s41588-018-0183-z

88 Khera A V, Chaffin M, Wade KH, *et al.* Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood In Brief Article Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. *Ozlem Senol-Cosar* 2019;**177**:12. doi:10.1016/j.cell.2019.03.028

89 Mars N, Koskela JT, Ripatti P, *et al.* Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat Med* 2020;**26**:549–57. doi:10.1038/s41591-020-0800-0

90 Wald NJ, Old R. The illusion of polygenic disease risk prediction. *Genet Med* 2019;**21**:1705–7. doi:10.1038/S41436-018-0418-5

91 Pal Choudhury P, Maas P, Wilcox A, *et al.* iCARE: An R package to build, validate and apply absolute risk models. *PLoS One* 2020;**15**:e0228198. doi:10.1371/journal.pone.0228198

92 Inouye M, Abraham G, Nelson CP, *et al.* Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. *J Am Coll Cardiol* 2018;**72**:1883–93. doi:10.1016/j.jacc.2018.07.079

93 Abraham G, Havulinna AS, Bhalala OG, *et al.* Genomic prediction of coronary heart disease. *Eur Heart J* 2016;**37**:3267–78. doi:10.1093/eurheartj/ehw450

94 Morris RW, Cooper JA, Shah T, *et al.* Marginal role for 53 common genetic variants in cardiovascular disease prediction. *Heart* 2016;**102**:1640–7. doi:10.1136/heartjnl-2016-309298

95 Elliott J, Bodinier B, Bond TA, *et al.* Predictive Accuracy of a Polygenic Risk Score-Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease. *JAMA - Journal of the American Medical Association* 2020;**323**:636–45. doi:10.1001/jama.2019.22241

96 Abraham G, Malik R, Yonova-Doing E, *et al.* Genomic risk score offers predictive performance comparable to clinical risk factors for ischaemic stroke. *Nat Commun* 2019;**10**:1–10. doi:10.1038/s41467-019-13848-1

97 Talmud PJ, Cooper JA, Morris RW, *et al.* Sixty-five common genetic variants and prediction of type 2 diabetes. *Diabetes* 2015;**64**:1830–40. doi:10.2337/db14-1504

98 Zhou D, Yu D, Scharf JM, *et al.* Contextualizing genetic risk score for disease screening and rare variant discovery. *Nat Commun* 2021;**12**:1–14. doi:10.1038/s41467-021-24387-z

99 Lu T, Forgetta V, Richards JB, *et al.* Polygenic risk score as a possible tool for identifying familial monogenic causes of complex diseases. *Genetics in Medicine* 2022;**0**:1–11. doi:10.1016/J.GIM.2022.03.022

100 Austin MA, Hutter CM, Zimmern RL, *et al.* Genetic causes of monogenic heterozygous familial hypercholesterolemia: a HuGE prevalence review. *Am J Epidemiol* 2004;**160**:407–20. doi:10.1093/AJE/KWH236

101 Awan Z, Choi HY, Stitziel N, *et al.* APOE p.Leu167del mutation in familial hypercholesterolemia. *Atherosclerosis* 2013;**231**:218–22. doi:10.1016/J.ATHEROSCLEROSIS.2013.09.007

102 Versmissen J, Oosterveer DM, Yazdanpanah M, *et al.* Efficacy of statins in familial hypercholesterolaemia: a long term cohort study. *BMJ* 2008;**337**:223–6. doi:10.1136/BMJ.A2423

103 Tada H, Kawashiri MA, Nohara A, *et al.* Impact of clinical signs and genetic diagnosis of familial hypercholesterolaemia on the prevalence of coronary artery disease in patients with severe hypercholesterolaemia. *Eur Heart J* 2017;**38**:1573–9. doi:10.1093/EURHEARTJ/EHX004

104 Khera A V., Won HH, Peloso GM, *et al.* Diagnostic Yield and Clinical Utility of Sequencing Familial Hypercholesterolemia Genes in Patients With Severe Hypercholesterolemia. *J Am Coll Cardiol* 2016;**67**:2578–89. doi:10.1016/J.JACC.2016.03.520

105 Vallejo-Vaz AJ, Marco M De, Stevens CAT, *et al.* Overview of the current status of familial hypercholesterolaemia care in over 60 countries - The EAS Familial Hypercholesterolaemia Studies Collaboration (FHSC). *Atherosclerosis* 2018;**277**:234–55. doi:10.1016/J.ATHEROSCLEROSIS.2018.08.051

106 Akioyamen LE, Genest J, Shan SD, *et al.* Estimating the prevalence of heterozygous familial hypercholesterolaemia: a systematic review and meta-analysis. *BMJ Open* 2017;**7**. doi:10.1136/BMJOPEN-2017-016461

107 Nordestgaard BG, Chapman MJ, Humphries SE, *et al.* Familial hypercholesterolaemia is underdiagnosed and undertreated in the general population: Guidance for clinicians to

prevent coronary heart disease. *Eur Heart J* 2013;**34**:3478–90. doi:10.1093/eurheartj/eht273

108 NHS Long Term Plan » Cardiovascular disease. https://www.longtermplan.nhs.uk/online-version/chapter-3-further-progress-on-care-quality-and-outcomes/better-care-for-major-health-conditions/cardiovascular-disease/ (accessed 2 Mar 2022).

109 Soutar AK, Naoumova RP. Mechanisms of Disease: genetic causes of familial hypercholesterolemia. *Nature Clinical Practice Cardiovascular Medicine 2007 4:4* 2007;**4**:214–25. doi:10.1038/ncpcardio0836

110 Go GW, Mani A. Low-Density Lipoprotein Receptor (LDLR) Family Orchestrates Cholesterol Homeostasis. *Yale J Biol Med* 2012;**85**:19.

111 Horton JD, Cohen JC, Hobbs HH. PCSK9: a convertase that coordinates LDL catabolism. *J Lipid Res* 2009;**50**:S172–7. doi:10.1194/JLR.R800091-JLR200

112 Sabatine MS. PCSK9 inhibitors: clinical evidence and implementation. *Nature Reviews Cardiology 2018 16:3* 2018;**16**:155–65. doi:10.1038/s41569-018-0107-8

113 Dominiczak MH, Caslake MJ. Apolipoproteins: Metabolic role and clinical biochemistry applications. *Ann Clin Biochem* 2011;**48**:498–515. doi:10.1258/acb.2011.011111

114 Borén J, Lee I, Zhu W, *et al.* Identification of the low density lipoprotein receptor-binding site in apolipoprotein B100 and the modulation of its binding activity by the carboxyl terminus in familial defective apo-B100. *Journal of Clinical Investigation* 1998;**101**:1084. doi:10.1172/JCI1847

115 Rashidi OM, H.Nazar FA, Alama MN, *et al.* Interpreting the Mechanism of APOE (p.Leu167del) Mutation in the Incidence of Familial Hypercholesterolemia; An In-silico Approach. *Open Cardiovasc Med J* 2017;**11**:84. doi:10.2174/1874192401711010084

116 Zuliani G, Arca M, Signore A, *et al.* Characterization of a New Form of Inherited Hypercholesterolemia. *Arterioscler Thromb Vasc Biol* 1999;**19**:802–9. doi:10.1161/01.ATV.19.3.802

117 Michaely P, Li WP, Anderson RGW, *et al.* The modular adaptor protein ARH is required for low density lipoprotein (LDL) binding and internalization but not for LDL receptor clustering in coated pits. *J Biol Chem* 2004;**279**:34023–31. doi:10.1074/JBC.M405242200

118 Harada-Shiba M, Takagi A, Miyamoto Y, *et al.* Clinical Features and Genetic Analysis of Autosomal Recessive Hypercholesterolemia. *J Clin Endocrinol Metab* 2003;**88**:2541–7. doi:10.1210/JC.2002-021487

119 NHS England — London » Familial Hypercholesterolemia (FH). https://www.england.nhs.uk/london/london-clinical-networks/our-networks/cardiac/familial-hypercholesterolaemia/ (accessed 20 Oct 2022).

120 Hovingh GK, Davidson MH, Kastelein JJP, *et al.* Diagnosis and treatment of familial hypercholesterolaemia. *Eur Heart J* 2013;**34**:962–71. doi:10.1093/EURHEARTJ/EHT015

121 Recommendations | Familial hypercholesterolaemia: identification and management | Guidance | NICE. https://www.nice.org.uk/guidance/cg71/chapter/Recommendations#identifying-people-with-fh-using-cascade-testing (accessed 18 Apr 2022).

122 Kerr M, Pears R, Miedzybrodzka Z, *et al.* Cost effectiveness of cascade testing for familial hypercholesterolaemia, based on data from familial hypercholesterolaemia services in the UK. *Eur Heart J* 2017;**38**:1832–9. doi:10.1093/EURHEARTJ/EHX111

123 Wald DS, Bestwick JP, Morris JK, *et al.* Child–Parent Familial Hypercholesterolemia Screening in Primary Care. *New England Journal of Medicine* 2016;**375**:1628–37. doi:10.1056/NEJMOA1602777

124 Futema M, Cooper JA, Charakida M, *et al.* Screening for familial hypercholesterolaemia in childhood: Avon Longitudinal Study of Parents and Children (ALSPAC). *Atherosclerosis* 2017;**260**:47. doi:10.1016/J.ATHEROSCLEROSIS.2017.03.007

125 Groselj U, Kovac J, Sustar U, *et al.* Universal screening for familial hypercholesterolemia in children: The Slovenian model and literature review. *Atherosclerosis* 2018;**277**:383–91. doi:10.1016/j.atherosclerosis.2018.06.858

126 Sanin V, Schmieder R, Ates S, *et al.* Population-based screening in children for early diagnosis and treatment of familial hypercholesterolemia: design of the VRONI study. *Eur J Public Health* 2022;**32**:422–8. doi:10.1093/EURPUB/CKAC007

127 Child-family screening for familial hypercholesterolemia: ethical issues - GOV.UK. https://www.gov.uk/government/publications/child-family-screening-for-familial-hypercholesterolemia-ethical-issues/child-family-screening-for-familial-hypercholesterolemia-ethical-issues (accessed 20 Oct 2022).

128 Wald DS, Neely D. The UK National Screening Committee's position on child–parent screening for familial hypercholesterolaemia. *J Med Screen* Published Online First: 22 July 2021. doi:10.1177/09691413211025426

129 CPSS – HEART UK. https://www.heartuk.org.uk/cholesterol/cpss-information-for-parents (accessed 5 May 2022).

130 Marquina C, Lacaze P, Tiller J, *et al.* Population genomic screening of young adults for familial hypercholesterolaemia: a cost-effectiveness analysis. *Eur Heart J* Published Online First: 11 November 2021. doi:10.1093/EURHEARTJ/EHAB770

# 2 Analysis of polygenic risk scores (PRS) in the Polygenic Score Catalog for disease screening, risk prediction, and population stratification

A preprint version of the following chapter is available on medRxiv and has been submitted for publication.[1]

## 2.1 Abstract

**Background:** There is interest in the potential use of polygenic risk scores (PRS) for disease prediction and screening but uncertainty on their performance. The aim of this chapter was to evaluate the PRS published in the Polygenic Score Catalog as predictive and screening tests using the relevant performance metrics.

**Methods:** I converted metrics curated in the Polygenic Score Catalog (odds ratios (OR), hazard ratios (HR), area under the receiver operating characteristic curve (AUC), C-index), into the sensitivity or detection rate (DR) for PRS cut-offs that define a 5% false positive rate (FPR). I evaluated the performance of PRS for disease in screening, risk prediction and population stratification by obtaining the odds of becoming affected calculated as the background odds of disease multiplied by the likelihood ratio. I also analysed the effect of adding a PRS to conventional risk factors in the prediction and primary prevention of coronary artery disease (CAD) and stroke.

**Results:** I identified 10,723 performance metrics for 2,194 polygenic scores for the prediction of 544 endpoints as of April 2022. At a 5% FPR, PRS detected between 8-19% (and therefore missed 81-92%) of affected individuals. For a CAD PRS with a DR for a 5% FPR (DR5) of 13% (DR5 = 13%) and a population 10-year risk of 8% (background odds of 1 to 12), the odds would be reduced to 1 to 20 with a PRS at the 25$^{th}$ centile and increased to 1 to 10 with a PRS at the 75$^{th}$ centile. Based on two data sources, adding a PRS to non-genetic risk prediction models for cardiovascular disease (CVD) and CAD, using a 10-year risk cut-off of 10% for initiation of statin treatment, produced numbers-needed-to-genotype to prevent one additional event of 5,882 and 8,879.

**Conclusion:** Analysis using the relevant metrics revealed the weak predictive performance of PRS for a wide range of disease endpoints (including various CVD endpoints), casting doubt on the future role of PRS in screening, population risk stratification, and individual risk prediction.

## 2.2 Introduction

Polygenic scores (PGS) represent the weighted sum of independent DNA sequence variants in a genome that increase risk of a particular trait. When the trait in question is a disease, PGS are more commonly referred to as polygenic risk scores (PRS). The weight assigned to each variant is based on the strength of its trait association in a genome wide association study (GWAS).[2] The increasing range and scale of GWAS over the last decade, now spanning over 900 diseases, has led to a proliferation in PGS and sparked widespread interest in their potential healthcare applications, capturing the attention of policy makers.[3] Individual consumers and healthcare providers can already access commercial genetic testing and software services based on PGS.[4–6] In a progression toward healthcare implementation, position papers have appeared on reporting standards and responsible clinical use from the Clinical Genome Resource (ClinGen) Complex Disease Working Group,[2] and the Polygenic Risk Score Task Force of the International Common Disease Alliance.[7] Yet, there is disagreement on the performance of PRS in disease screening, prediction, and risk stratification, and their eventual role in medicine and public health remains uncertain.[8–11]

Recently, Lambert and colleagues produced the Polygenic Score Catalog, a comprehensive, regularly updated, publicly accessible directory of studies on PGS for quantitative traits (e.g. blood pressure) and PRS for diseases (e.g. coronary artery disease (CAD)).[12] The catalogue lists the following performance metrics for PRS: the hazard (HR) and odds ratio (OR), both per one standard deviation (SD) increment in the score, and the area under the receiver operating characteristic curve (AUC), sometimes expressed as the C-index.

However, these widely reported metrics are not directly informative of performance in disease screening, individual risk prediction, or population risk stratification. The required metric is the odds of becoming affected given a result, which is the positive predictive value expressed as an odd. It is obtained by multiplying the background odds of disease in a population by the likelihood ratio. In each case (screening, prediction, and risk stratification), the likelihood ratio can be calculated from the PRS standard normal distributions. The aim of this chapter was to derive the required metrics from the reported metrics (HR per SD, OR per SD, AUC, and C-index) to properly evaluate the performance of PRS in their intended applications. The performance of PRS in established risk factor models in the prediction of CVD and CAD was also evaluated.

## 2.3 Methods

### 2.3.1 Analysis overview

The data analysed here was downloaded from the Polygenic Score Catalog in April 2022.[12] When referring to a particular PGS, I used the Polygenic Score Catalog reference number. Data analysis was done in R version 4.0.2, and figures were plotted using the R package *ggplot2* version 3.3.6.[13,14]

I excluded implausible values for the original metrics (OR or HR per SD <1; AUC and C-index either <0.5 or >1). I then used the overlap of the Gaussian distributions to calculate the detection rate (DR) for a pre-specified false positive rate (FPR). For simplicity and consistency, I set PRS cut-offs that define a 5% FPR and calculated the corresponding DR (the DR5) as explained below.[15]

The likelihood ratio in screening is defined as the ratio of the DR to the FPR; in risk prediction, as the ratio of the heights of the relative Gaussian distributions of PRS for affected and unaffected groups at a particular PRS value; and in risk stratification, as the relative areas under the distributions for affected and unaffected individuals in each PRS quantile (e.g., each fifth of the PRS distribution). PRS "centile" or "quantile" is in reference to the distribution in the unaffected group. In each of these three cases, multiplying the likelihood ratio by the background odds of disease gives the corresponding odds of becoming affected.

I re-analysed data from two original studies to quantify the extent to which the addition of information from PRS improves the prediction of CVD and CAD events.[16,17] I did so by computing the DR and FPR, with and without information from PRS, using risk cut-offs recommended in guidelines for the initiation of statin treatment. The DR was calculated as the ratio of true positive cases to the total number of positive cases (i.e. true positive cases plus false negative cases). The FPR was calculated as the ratio of false positive cases to the total number of negative cases (i.e. false positive plus true negative cases). I calculated the number of individuals who need to be genotyped to detect or prevent one additional CVD or CAD event.

### 2.3.2 Assumptions

The calculations assume that PRS exhibit a Gaussian distribution in a population: the proportional difference in disease risk is the same for any given difference in PRS value from any starting level (i.e. a log-linear relationship), and PRS distributions have the same SD (~1) in affected and unaffected individuals.[12]

### 2.3.3 Deriving DR5 from HR or OR per SD

From these assumptions, it is possible to mathematically derive the DR and FPR from the OR or HR per SD.[15,18] These derivations taken from Wald and colleagues' work are detailed in the methods section below.[18,19]

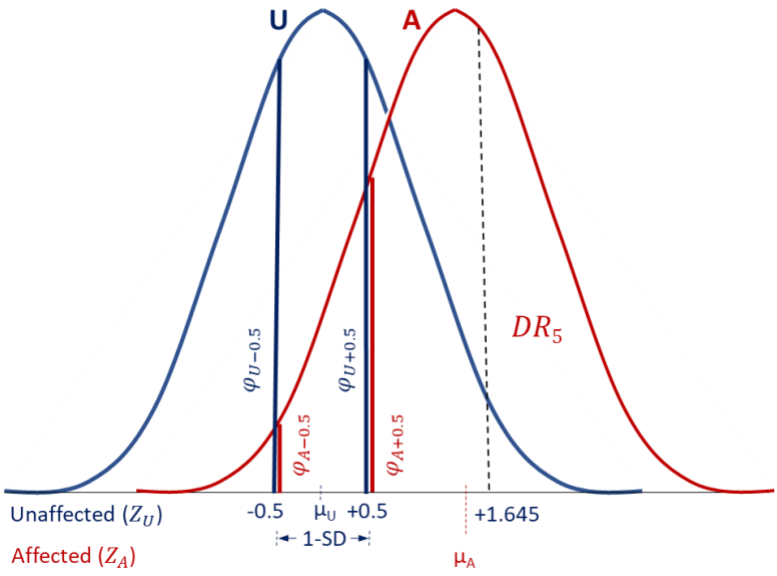



**Figure 2.1 Standardised normal distributions for affected and unaffected individuals and their relationship to the OR or HR per SD and the DR5.** The distribution for the affected (mean: $\mu_A$) is shown in red, and the distribution for the unaffected (mean: $\mu_U$) is depicted in blue. The standardised distributions have a SD of 1. The dotted vertical line refers to the 95$^{th}$ centile of the unaffected distribution, equivalent to the cut-off value for the 5% FPR. The DR5 refers to the DR for a 5% FPR, which is the area under the curve of the affected distribution to the right of the dotted line (i.e. the 5% FPR cut-off). The OR can be calculated by dividing the heights of the distributions at equivalent SDs (e.g. +0.5 and -0.5) from the mean of the unaffected distribution. DR = detection rate; DR5 = detection rate for a 5% false positive rate; FPR = false positive rate; HR = hazard ratio; OR = odds ratio; SD = standard deviation.

Consider two overlapping Gaussian PRS distributions with equal SD ($\sigma$) standardised using the Z-score: one for affected (A) and one for unaffected (U) with mean values $\mu_A$ and $\mu_U$ respectively (**Figure 2.1**). The OR per SD is equal to the ratios of the probability density function values ($\varphi$) for affected ($\varphi_A$) and unaffected ($\varphi_U$) individuals that are one SD apart, for example corresponding to values at +0.5 SD and -0.5 SD of the distribution of unaffected individuals. In **Figure 2.1**, the values are shown by the vertical red and blue lines respectively. This can be expressed as:

$$OR_{SD} = \frac{\left(\varphi_A / \varphi_U\right)_{+0.5}}{\left(\varphi_A / \varphi_U\right)_{-0.5}}$$

For a value $x$ on the x-axis, the probability density function ($\varphi$) of a Gaussian distribution is equal to:

$$\varphi(\mu, \sigma, x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Which means that for the distribution of the unaffected, where $\mu = 0$ and $\sigma = 1$:

$$\varphi_U(0,1,x) = \frac{1}{1\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x}{1}\right)^2}$$

$$\varphi_U(0,1,x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

At values of $x = 0.5$ and $x = -0.5$, we know that $\varphi_{U+0.5} = \varphi_{U-0.5}$, which means that:

$$OR_{SD} = \frac{(\varphi_{A+0.5})}{(\varphi_{A-0.5})}$$

For the distribution of the affected, $\mu = \mu_A$ and $\sigma = 1$; so at $x = 0.5$ and $x = -0.5$:

$$\varphi A(\mu_A, \sigma, -0.5) = \frac{1}{1\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{-\mu_A - 0.5}{1}\right)^2}$$

And,

$$\varphi A(\mu_A, \sigma, +0.5) = \frac{1}{1\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{0.5-\mu_A}{1}\right)^2}$$

Thus:

$$OR_{SD} = \frac{\frac{1}{1\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{0.5-\mu_A}{1}\right)^2}}{\frac{1}{1\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{-\mu_A-0.5}{1}\right)^2}}$$

$$OR_{SD} = \frac{e^{-\frac{1}{2}(0.5-\mu_A)^2}}{e^{-\frac{1}{2}(-\mu_A-0.5)^2}}$$

$$OR_{SD} = e^{\left[-\frac{1}{2}(0.5-\mu_A)^2+\frac{1}{2}(-\mu_A-0.5)^2\right]}$$

$$OR_{SD} = e^{-\frac{1}{2}\left[(0.5-\mu_A)^2-(-\mu_A-0.5)^2\right]}$$

$$OR_{SD} = e^{-\frac{1}{2}[-2\mu_A]}$$

$$OR_{SD} = e^{\mu_A}$$

$$\mu_A = \ln OR_{SD}$$

The DR at a cut point $x$ of a standard normal cumulative distribution can be calculated using the cumulative distribution function (CDF), which corresponds to the area under the normal distribution among affected at the right of $x$ (note that the entire area under the distribution is equal to 1):

$$DR = 1 - \emptyset(x - \mu_A)$$

And since the normal distribution is symmetrical:

$$DR = \emptyset(\mu_A - x)$$

The DR5 can be calculated using the 95$^{th}$ centile cut-off of the unaffected distribution (i.e. the threshold for a 5% FPR). This has a value of 1.645 on the unaffected distribution of $\mu_U = 0$ and $\sigma = 1$. Using the previous equations, we can obtain the DR5 and link it to the OR (or HR) per SD:

$$DR_5 = \emptyset(\mu_A - 1.645)$$

$$DR_5 = \emptyset(\ln OR_{SD} - 1.645)$$

### 2.3.4 Deriving DR5 from AUC and C-index

It is also possible to derive the DR5 from the AUC and C-index, as demonstrated by Wald and Bestwick.[19] The calculations below detail how this can be achieved.

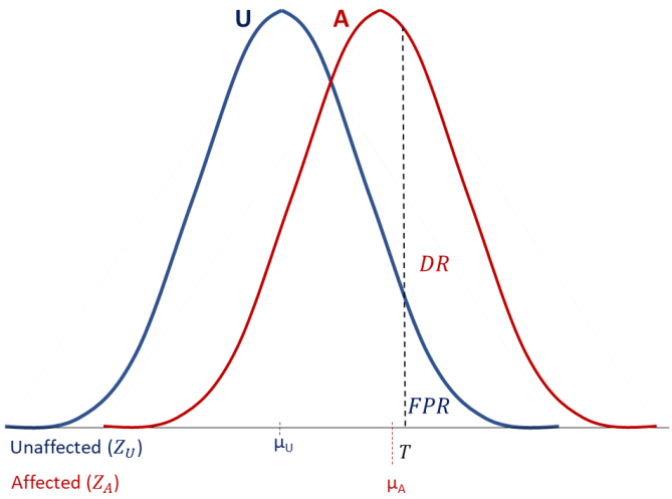



**Figure 2.2 Standardised normal distributions for affected and unaffected individuals and their relationship to the AUC or C-index, the DR and the FPR.** The affected distribution (mean: $\mu_A$) is depicted in red, and the unaffected distribution (mean: $\mu_U$) in blue. For a cut-off $T$, the DR and FPR can be calculated. AUC = area under the curve; DR = detection rate; FPR = false positive rate.

The DR for a cut-point value $T$ (**Figure 2.2**) is given by:

$$DR = 1 - \emptyset\left(\frac{T - \mu_A}{\sigma_A}\right)$$

And because of the symmetry of the distribution:

$$DR = \emptyset\left(\frac{\mu_A - T}{\sigma_A}\right)$$

$$\emptyset^{-1}(DR) = \frac{\mu_A - T}{\sigma_A}$$

Similarly, the FPR can be expressed as:

$$FPR = \emptyset\left(\frac{\mu_U - T}{\sigma_U}\right)$$

$$\emptyset^{-1}(FPR) = \frac{\mu_U - T}{\sigma_U}$$

Now writing both equations in terms of T:

$$T = \mu_A - \sigma_A\emptyset^{-1}(DR)$$

And

$$T = \mu_U - \sigma_U\emptyset^{-1}(FPR)$$

Combining the equations

$$\mu_A - \sigma_A\emptyset^{-1}(DR) = \mu_U - \sigma_U\emptyset^{-1}(FPR)$$

And making DR the subject,

$$\sigma_A \emptyset^{-1}(DR) = \mu_A - \mu_U + \sigma_U \emptyset^{-1}(FPR)$$

$$\emptyset^{-1}(DR) = \frac{\mu_A - \mu_U + \sigma_U \emptyset^{-1}(FPR)}{\sigma_A}$$

$$DR = \emptyset\left(\frac{\mu_A - \mu_U}{\sigma_A} + \frac{\sigma_U \emptyset^{-1}(FPR)}{\sigma_A}\right)$$

Considering the standard normal distribution and assuming equal SD in unaffected and affected groups ($\sigma_A = \sigma_U = 1$), with $\mu_U = 0$:

$$DR = \emptyset\left(\frac{\mu_A}{1} + \frac{1 \cdot \emptyset^{-1}(FPR)}{1}\right)$$

$$DR = \emptyset\big(\mu_A + \emptyset^{-1}(FPR)\big)$$

The AUC (or C-index) is the probability that an affected individual drawn at random ($A$) has a higher PRS than an unaffected individual drawn at random ($U$), i.e.:

$$P(A > U) = P(A - U > 0)$$

The AUC is therefore the CDF for the distribution of differences (the variances sum):

$$AUC = \emptyset\left(\frac{\mu_A - \mu_U}{\sqrt{\sigma_A^2 + \sigma_U^2}}\right)$$

$$\emptyset^{-1}(AUC) = \frac{\mu_A - \mu_U}{\sqrt{\sigma_A^2 + \sigma_U^2}}$$

Given that $\mu_U = 0$ and that the SD for the distributions for affected and unaffected individuals are equal to 1 ($\sigma_A = \sigma_U = 1$):

$$\emptyset^{-1}(AUC) = \frac{\mu_A}{\sqrt{2}}$$

$$\mu_A = \sqrt{2}\emptyset^{-1}(AUC)$$

From the previous DR equation,

$$DR = \emptyset\left(\mu_A + \emptyset^{-1}(FPR)\right)$$

To obtain the DR5 from the previous equations:

$$DR_5 = \emptyset\left(\mu_A + \emptyset^{-1}(0.05)\right)$$

$$DR_5 = \emptyset\left(\sqrt{2}\emptyset^{-1}(AUC) + \emptyset^{-1}(0.05)\right)$$

### 2.3.5 Screening: calculating the likelihood ratio and odds of becoming affected given a positive test result (OAPR)

In evaluating the performance of a PRS as a screening test, I calculated the likelihood ratio for a positive result (i.e. a PRS at or above a pre-specified cut-off) as the ratio $DR/FPR$ (**Figure 2.2**). The likelihood ratio for a test cut-off with a 5% FPR is given by $DR5/_{5\%}$ where DR5 is expressed as a percentage.

The odds of becoming affected is calculated by multiplying the background odds of disease by the likelihood ratio for a positive test result. For example, if the background odds of disease in the population is 1:9 and the DR5 is 15%, the likelihood ratio is $15/5 = 3$ and the $OAPR = (1 \times 3){:}\,9$ or $1{:}\,3$.

### 2.3.6 Risk prediction: calculating the likelihood ratio and odds of becoming affected given a PRS result

The likelihood ratio corresponds to how many times more likely a given PRS result is to arise from an affected individual than an unaffected individual. For evaluating PRS in risk prediction, the likelihood ratio (LR) can be calculated using the heights of the standard normal distribution curves for affected ($\varphi_A$) and unaffected ($\varphi_U$) individuals at a specific PRS value.

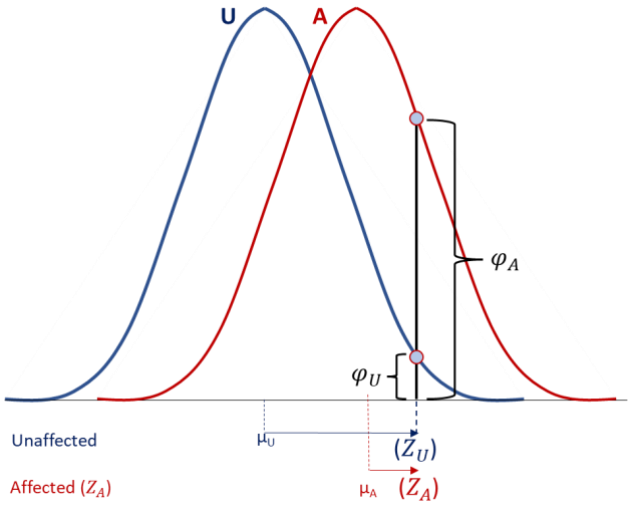



**Figure 2.3 Standardised normal distributions for affected and unaffected individuals and their relationship to the likelihood ratio.** The affected distribution (mean: $\mu_A$) is depicted in red, and the unaffected distribution (mean: $\mu_U$) in blue. $\varphi_A$ is the height of the distribution of affected individuals and $\varphi_U$ the height of the distribution of unaffected individuals at a same cut-point. The likelihood ratio is given by the ratio $\varphi_A/\varphi_U$.

In **Figure 2.3**, this is equal to:

$$LR = \varphi_A/\varphi_U$$

The heights of the distributions can be calculated using the equation for the Gaussian distribution:

$$LR = \frac{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(Z_A)^2}}{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(Z_U)^2}}$$

For a standard normal distribution of SD 1 ($\sigma = 1$):

$$LR = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(Z_A)^2}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(Z_U)^2}}$$

For example, for a PRS with a performance metric expressed as a HR per SD of 1.71 ($OR_{SD} = 1.71$): $\mu_A = \ln HR_{SD} = \ln 1.71 = 0.54$.

A PRS at the 75th centile of the distribution for unaffected individuals yields a Z-score of 0.67 for unaffected individuals ($Z_U = 0.67$). Using the formula above ($\varphi(0,1,x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$): $\varphi_U = 0.32$.

The Z-score for the affected distribution is equal to the difference between the Z-score of the unaffected distribution ($Z_U$) minus the mean of the affected distribution ($\mu_A$) (**Figure 2.3**): $Z_A = Z_U - \mu_A = 0.67 - 0.54 = 0.13$. Using the same formula above, this equals a height of $\varphi_A = 0.40$.

The likelihood ratio can then be calculated: $LR = {\varphi_A}/{\varphi_U} = {0.40}/{0.32} = 1.25$.

If the odds of disease in the population is 1:9, an individual whose PRS is at the 75th centile of the distribution among unaffected has an odd of becoming affected of $(1.25 \times 1): 9 \simeq 1: 7$.

### 2.3.7 Risk stratification: calculating the likelihood ratio and odds of becoming affected for a particular PRS group

In evaluating a PRS in risk stratification, the likelihood ratio was calculated as the ratio of areas under the distributions for affected and unaffected individuals in each PRS quantile (e.g. each fifth of the PRS distribution with respect to the unaffected) (**Figure 2.4**). The background odds of disease were then multiplied by the corresponding likelihood ratio to determine the odds of becoming affected for each quantile of the distribution. For example, for individuals in the fourth quintile, the odds of becoming affected are 1.47. If the background odds of disease in the population is 1:9, the odds of becoming affected for this group is $(1.47 \times 1): 9 \simeq 1: 6$.

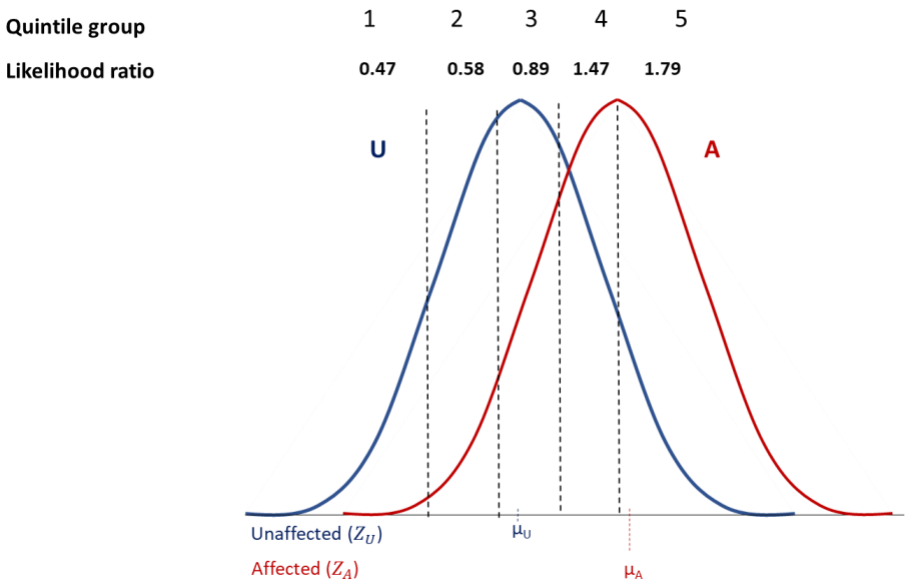



**Figure 2.4 Standardised normal distributions for affected and unaffected individuals and their relationship to the likelihood ratio for different quintiles of the distributions.** The affected distribution (A; mean $\boldsymbol{\mu_A}$) is depicted in red, and the unaffected distribution (U; mean $\boldsymbol{\mu_U}$) in blue. The unaffected distribution is split into quintiles. The likelihood ratio for each quintile is calculated as the ratio of the areas under the distributions for affected and unaffected individuals in each quintile.

## 2.4 Results

### 2.4.1 Performance of PGS in the Polygenic Score Catalog

By April 2022, the Polygenic Score Catalog had curated 10,723 performance metrics for 2,194 PGS, involving 544 diseases or traits, reported in 303 publications (**Figure 2.5**). Of the 10,723 metrics, 3,915 (37%) concerned disease endpoints, reported as OR per SD in 1,216, HR per SD in 378, AUC in 2,077 and C-index in 244 instances (**Figure 2.5**).

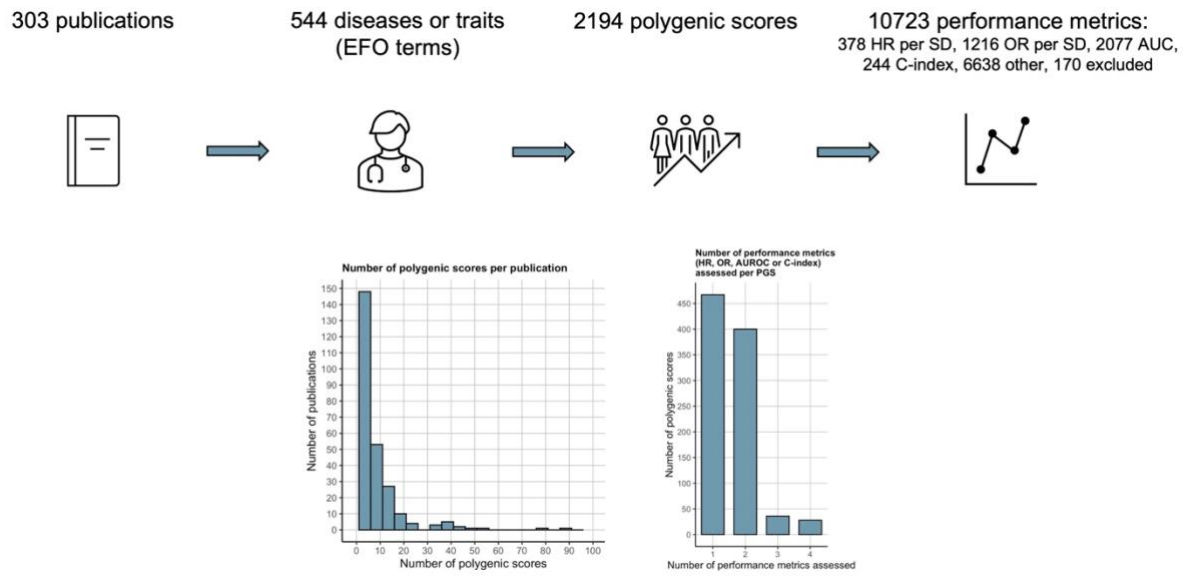



**Figure 2.5 Summary of data included in the Polygenic Score Catalog as of April 2022.** For ease of illustration, six outlier publications were removed from the first histogram. These included 4039, 3380, 302, 280, 221 and 170 PGS per publication.

The median DR5 values [interquartile range (IQR)] for PRS whose performance was reported using HR or OR per SD were 8.1% [7.0; 10.1] and 8.5% [6.3; 11.5] respectively, excluding 167 instances where the HR or OR per SD were recorded as <1 (**Table 2.1** and **Figure 2.6**).

| Metric | Count | Median DR5 | 25th centile | 75th centile | Maximum DR5 value | Minimum DR5 value |
|---|---|---|---|---|---|---|
| HR per SD | 378 | 8.1 | 7.0 | 10.1 | 51.3 | 5.0 |
| OR per SD | 1216 | 8.5 | 6.3 | 11.5 | 81.3 | 5.0 |
| AUC | 2077 | 13.5 | 9.9 | 22.1 | 96.9 | 5.1 |
| C-index | 244 | 19.1 | 12.8 | 25.3 | 58.6 | 6.3 |

**Table 2.1 The DR5 values derived from the HR per SD, OR per SD, AUC, and C-index metrics reported in the Polygenic Score Catalog.** AUC = area under the receiver operating characteristic curve; DR5 = detection rate for a 5% false positive rate; HR = hazard ratio; OR = odds ratio; SD = standard deviation.

For PRS performance reported using the AUC or C-index, the corresponding median DR5 values were 13.5% [9.9; 22.1] and 19.1% [12.8; 25.3] respectively, excluding two instances where the AUC was < 0.5, and one instance where the C-index was recorded as 632 (**Table 2.1** and **Figure 2.6**).

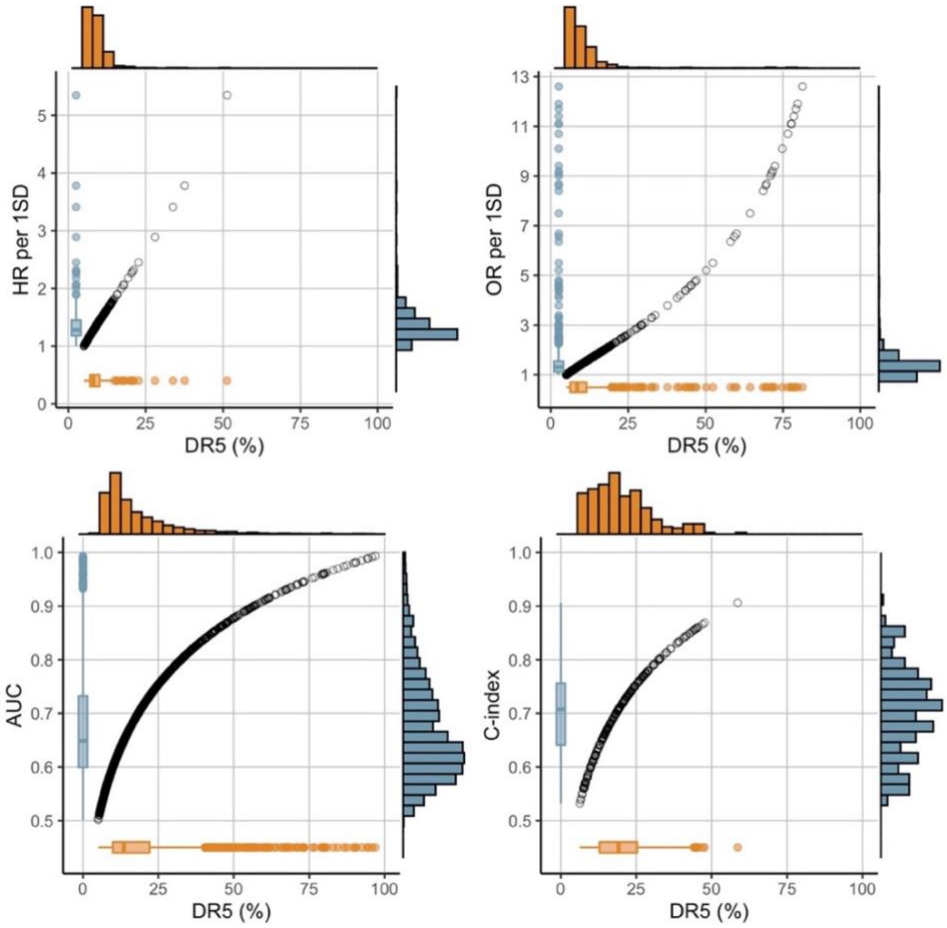



**Figure 2.6 Distribution of DR5 derived from HR per SD, OR per SD, AUC, and C-index values listed in the Polygenic Score Catalog.** AUC = area under the receiver operating characteristic curve; DR5 = detection rate for a 5% false positive rate; HR = hazard ratio; OR = odds ratio; SD = standard deviation.

### 2.4.2 Performance of PRS in disease screening

Based on this overview, PRS typically detect 8-19% (and therefore miss 81-92%) of affected individuals at a 5% FPR. For example, for a PRS for CAD (PGS000018), the DR5 is 12% (with 88% of cases missed). Applied in a middle-aged population with a 10-year CAD incidence of 10% (i.e. a background odds of 1:9), the OAPR is equal to $(0.12 \times 1):(0.05 \times 9) \simeq 1:4$. This can be interpreted as the false positives outnumber the true positives by around four to one. Changing the cut-off to reduce the FPR (e.g. to 1%) also reduces the DR (to 3% in this example, with 97% of cases missed). And retaining a 5% FPR but applying the test in a population with a 1% CAD incidence over the same period (i.e. a background odds of 1:99) (e.g. in younger individuals) yields an OAPR of 1:41 (i.e. the false positives outnumber the true positive by around 41 to one).

Achieving more effective discrimination requires much greater separation of the PRS distributions of affected and unaffected individuals than is observed in practice. For instance, achieving a DR5 of 85% requires an OR per SD of 15 (compared to the median observed value of 1.31) or an AUC of 0.96 (compared to the median observed value of 0.65) (**Figure 2.6**).[15,19] Only 11.4% of AUC values in the Polygenic Score Catalog exceeded 0.8 which equates to a DR5 of 32%, with most of these reflecting large effect variants at the HLA locus in a few autoimmune diseases.

### 2.4.3 Risk prediction: interpretation of PRS in an individual

The overlap in PRS distributions also enables calculation of the odds of becoming affected for an individual based on their PRS result (see Methods section). For example, a 8% 10-year risk of CAD for a 40-year old individual living in the UK is approximately equal to an odds of disease of $8:(100-8) \simeq 1:12$.[16] A CAD PRS (PGS000018) of HR per SD of 1.71 equals to a DR5 of 13%.[20] For an individual in the 75$^{th}$ PRS centile, the likelihood ratio is 1.25 (see Methods section 1.3.6) and the odds of CAD are increased from 1:12 to 1:10 (i.e. $(1.25 \times 1):12$). For an individual with a PRS at the 25$^{th}$ centile, the likelihood ratio is 0.6 and the 10-year odds of CAD is reduced from 1:12 to 1:20 (i.e. $(0.6 \times 1):12$). The change in odds is more substantial for individuals with a CAD PRS in either tail of the distribution: the odds are reduced from the background odds of 1:12 to 1:40 (i.e. $(0.30 \times 1):12$) at the 2.5$^{th}$ centile, and increased to 1:5 (i.e. $(2.48 \times 1):12$) at the 97.5$^{th}$ centile.

### 2.4.4 Performance of PRS in population stratification

Population stratification involves assorting individuals in a population into groups according to their disease risk. Using the previous CAD PRS example, we saw that the change in the odds is more substantial for individuals with PRS at either tail of the distribution: 1:40 at the 2.5$^{th}$ centile, and 1:5 at the 97.5$^{th}$ centile. However, the latter group only accounts for 7.8% of all CAD cases (the 97.5$^{th}$ centile of the unaffected distribution is equal to a Z-score of 1.42 in the affected distribution). In risk stratification, setting a more stringent PRS cut-off for designating individuals as high risk shortens the odds of disease in the high-risk group, but because group size diminishes, so does the proportion of cases that are available to detect.

### 2.4.5 Using PRS in conjunction with conventional screening tests or risk factors

It has been proposed that adding PRS to conventional risk factors for CAD and stroke (e.g. blood pressure and low-density lipoprotein cholesterol) could improve risk estimation to guide the prescription of statins for primary prevention.[16,17] **Figure 2.7** and **Table 2.2** shows a re-analysis of results from Sun *et al.* (from the publication's Supplementary Figure 13), based on a hypothetical cohort of 100,000 40-year-old individuals with a risk factor profile representative of the UK population and a background 10-year risk of CAD and stroke of 8%.[16] A conventional risk factor model incorporating age detects 60% of CAD and stroke cases at a 24% FPR (DR24 = 60%); and adding the PRS to the model detects 61% of CAD and stroke cases for a 23% FPR (DR23 = 61%) (**Figure 2.7**). Assuming a 10-year risk cut-off of 10% for prescribing statins, 100% adherence, and a statin risk-reduction for CAD and stroke of 20%, Sun *et al.* estimated that 974 events would be prevented using conventional risk factors and PRS together instead of 957 events using conventional risk factor prediction alone (**Figure 2.7**).[21] This gives a number-needed-to-genotype of 5,882 to prevent one additional event (**Table 2.2**).

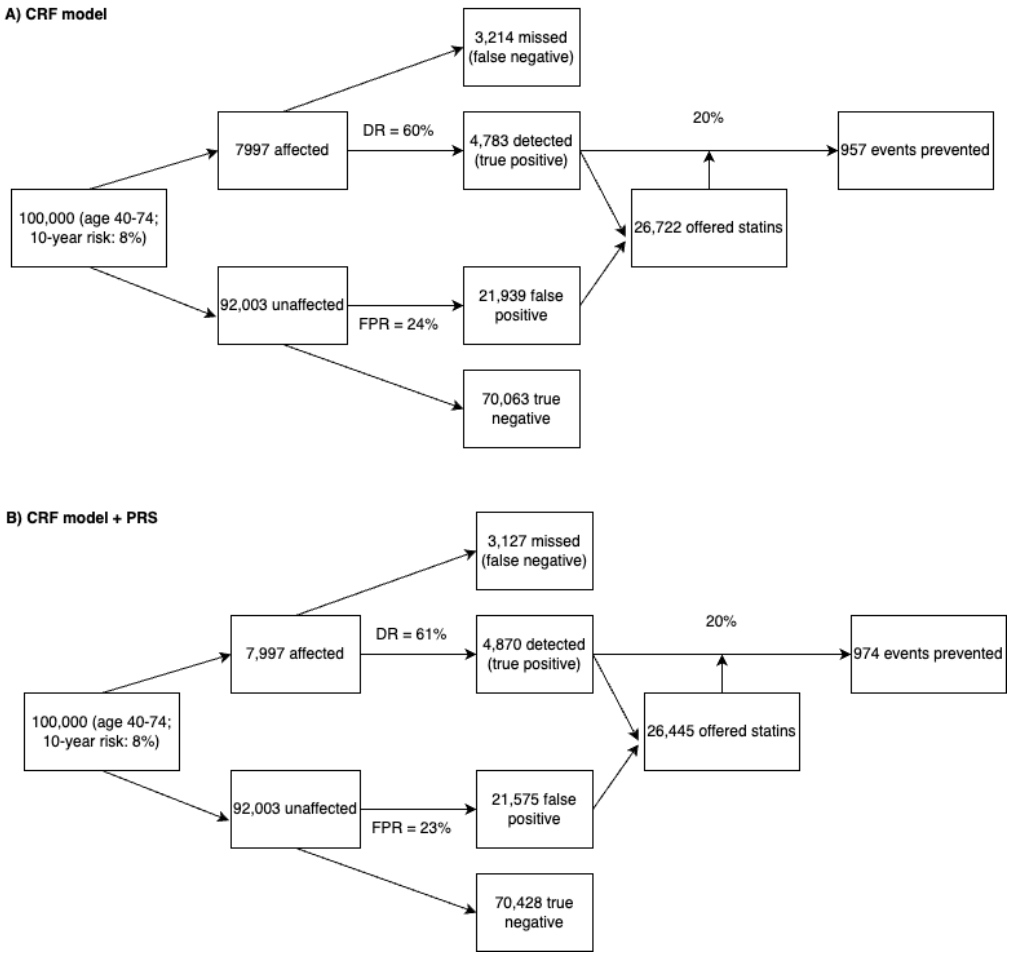



**Figure 2.7 Flow diagram of a hypothetical cohort of 100,000 individuals modelled by Sun *et al.* for the detection of CAD and stroke cases using conventional risk factors (CRF) alone versus a model combining CRF and PRS.[16]** 20% of detected CAD and stroke cases are expected to be prevented following statin treatment initiation. **A)** CRF model scenario. **B)** CRF + PRS model scenario. CAD = coronary artery disease; DR = detection rate; FPR = false positive rate; PRS = polygenic risk score.

Riveros-Mckay *et al.* also investigated the extent to which the addition of a PRS to conventional risk factors improves the identification of UK Biobank participants eligible to receive statins because their 10-year risk of CAD exceeds the cut-offs in UK or US primary prevention guidelines.[17] Re-analysis of their data (from the publication's Table 2 and Supplementary Table 4) reveals that the effect of adding information from a PRS is small. For example, using a 10-year risk cut-off of 10%, the QRISK3 model (based on conventional risk factors including age) detected 81% of cases at a 42% FPR (DR42 = 81%) overall for men and women (**Table 2.2**). Adding a PRS detected 84% of cases for a 41% FPR (DR41 = 84%). Using the 10% 10-year risk cut-off and assuming statins reduce CAD events by 20%, the number-needed-to-genotype to prevent one additional event based on this study is 8,879 (**Table 2.2**).

| Study | Risk tool | Screened | Genotyped | Risk cut-off | DR | FPR | # below cut-off | Events Below cut-off | # above cut-off | Events above cut-off | Additional events detected | Events avoided (statin) | Additional events avoided | # needed to genotype for detection of 1 additional case | # needed to genotype for prevention of 1 additional case |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sun *et al.* | CRF | 100000 | 0 | ≥10% | 60% | 24% | 73277 | 3214 | 26722 | 4783 | - | 957 | - | - | - |
| | CRF + PRS | 100000 | 100000 | ≥10% | 61% | 23% | 73554 | 3127 | 26445 | 4870 | 87 | 974 | 17 | 1149 | 5882 |
| Riveros-Mckay *et al.* | PCE | 186451 | 0 | >7.5% | 74% | 36% | 118082 | 1112 | 68369 | 3135 | - | 627 | - | - | - |
| | PCE + PRS | 186451 | 186451 | >7.5% | 80% | 36% | 117516 | 855 | 68935 | 3392 | 257 | 678 | 51 | 725 | 3656 |
| | QRISK3 | 186451 | 0 | >10% | 81% | 42% | 106697 | 797 | 79754 | 3450 | - | 690 | - | - | - |
| | QRISK3 + PRS | 186451 | 186451 | >10% | 84% | 41% | 108359 | 690 | 78092 | 3557 | 107 | 711 | 21 | 1743 | 8879 |

**Table 2.2 Effect of adding a PRS to non-genetic risk factors in prediction of CAD or CVD.** The values are based on a re-analysis of data reported by Sun *et al.* and Riveros-Mckay *et al.*[16,17] Both studies utilised data from UK Biobank. Sun *et al.* developed their own conventional risk factor score (CRF) and examined the effect of adding PRS for stroke and CAD on the prediction of CVD. They used a 10-year risk cut-off of 10% for offering statin treatment. Riveros-McKay *et al.* modelled screening performance in 18,6451 participants from UK Biobank based on either the Pooled Cohort Equation (PCE) developed for CVD prediction in the USA, using a 7.5% 10-year CVD risk cut-off, or using QRISK3 developed for CVD prediction in the UK, using a 10% CVD risk cut-off. The data on events reported by Riveros-Mckay *et al.* were for CAD alone rather than CVD (CAD and stroke). Calculations assume that all those exceeding the specified risk cut-off receive a statin and that statin treatment produces a 20% relative risk reduction ("events avoided" column). Number-needed-to-genotype refers to the number of individuals that need to be genotyped to detect or prevent one additional CVD event. CAD = coronary artery disease; CRF = conventional risk factor score; CVD = cardiovascular disease; DR = detection rate; FPR = false positive rate; PCE = Pooled Cohort Equation; PRS = polygenic risk score; # = number.

## 2.5 Discussion

This analysis converting 3,915 binary endpoint performance metrics to the relevant metrics indicates weak performance of PRS in disease screening, individual risk prediction, and population stratification, whether used alone or added to conventional risk factor models. Using PRS to identify the minority of individuals at very high risk necessitates genetic testing in all, generates many more false than true positives, and overlooks most cases which occur among those with average PRS.

These insights are not obvious from the widely reported (but less clinically informative) metrics curated by the Polygenic Score Catalog: the OR or HR per SD, AUC, and C-index.[12] However, by using these metrics to derive the odds of becoming affected for those with a "positive test" (in the case of screening), with a particular PGS value (in the case of risk prediction), or who occupy a particular PRS quantile (in the case of risk stratification), the limited performance of PRS in their intended applications become clearer. The conversion from less to more informative metrics involves first reconstituting the PRS distributions for affected and unaffected individuals. These distributions were found to overlap substantially for almost all conditions studied. Achieving more effective discrimination requires much greater separation of the distributions of affected and unaffected individuals than is observed in practice: it is this overlap of distributions that constrains the performance of PRS whether for screening, prediction, or risk stratification.

Studies have equated the predictive performance of PRS to that of non-genetic risk factors, such as blood pressure and low-density lipoprotein cholesterol for CAD.[22] However, although causally associated with CAD, they are also weak predictors of disease, making them bad comparators.[18] Furthermore, where a risk factor displays a Gaussian distribution in the population and a relatively weak log-linear association with disease risk, more cases occur among the average than among the few with extreme values.[23,24] For this reason, where there are safe, inexpensive preventative interventions (e.g. statins for the prevention of CAD and stroke), there is greater public health benefit in broadening rather than limiting their access.[25] Ascertaining a minority of individuals at very high risk (whether genetic or otherwise) may be justified if a preventative intervention is costly, resource limited, or potentially harmful. However, it entails testing in all and, aside from missing the many cases among those at average risk, generates many false positives. This could have substantial downstream resource implications for healthcare systems.

Despite this, it has been claimed that PRS generate substantial improvements in risk prediction, are likely to transform health care, and are ready to implement in practice.[26] Hopes may have been raised by several factors. PRS for disease prediction are appealing because of the low cost of genotyping and the invariant nature of genotype, which means that testing only needs to be done once in a lifetime to compute the risk for a wide range of diseases.[27] Many common disease-associated genetic variants are known, and together with the availability of large, longitudinal cohort studies with genetic and health outcome data, this has fuelled a growth in research on PGS. Pressure to demonstrate a tangible health impact of genetic research, coupled with increasing demand on healthcare systems has also forced consideration of new approaches to predict or detect disease. The appeal of PRS may also have been inflated by the depiction of their performance in research papers or company materials.[6,9,28] What is relevant in screening is the risk of an event in a group compared to that of the whole population, but published materials often illustrate comparisons between mutually exclusive groups such as those in opposite tails of a PGS distribution.[29]

The current analysis underlines the need for more responsible and pertinent presentation of the performance of PRS in disease prediction, screening, and population stratification. This could be achieved by 1) deriving the DR for a specified FPR; 2) primary studies always reporting the mean and SD of PRS among affected and unaffected individuals from which the overlap in distributions and the relevant performance metrics can then be derived; 3) authors reporting the performance of PGS with and without the inclusion of other variables (especially age and sex) which can markedly influence predictive performance so that users can judge the increment provided by the PGS itself; 4) commercial providers communicating individual test results to customers with greater clarity and relevance to screening performance (e.g. by reporting the odds of becoming affected, which requires additional information on population average risk at a particular age over a specified time). Finally, as others have already suggested, policy makers should consider tighter regulation of commercial PGS providers based on clinical (not just analytical) performance to protect already stretched out public health systems from being burdened by management of false positive results.

Although the current analysis unveils the limitations of PRS in disease screening, prediction, and risk stratification, they may find use in other applications. For example, they may explain the variable penetrance of rare mutations in monogenic diseases (e.g. familial hypercholesterolaemia)

and be employed to aid case detection.[30,31] There are also other predictive applications of genotyping (e.g. in pharmacogenetic testing) to optimise efficacy and safety of medicines.[32] The main healthcare benefit of common disease genomics may come from understanding the causes of disease and drug target discovery rather than disease prediction.[33]

The current analysis provides an overview of the PGS available on the Polygenic Score Catalog website.[12] The PRS analysed in this study were selected based on whether they had one of the following performance metrics associated with them: OR per SD, HR per SD, AUC, or C-index. These outcome measures were assumed to relate to binary disease outcomes (as opposed to non-disease traits). Furthermore, some PRS included in the current analysis might contain outliers that did not meet the assumptions in the Methods section for the derivation of the DR5, likelihood ratio, and OAPR. Finally, covariates were included in some of the reported PRS metrics but were not filtered out of the study, meaning that the performance of the PRS analysed here is likely to be an overestimation, highlighting the poor performance of PRS in disease screening, population risk stratification, and individual risk prediction.

## 2.6 References


1 Hingorani AD, Gratton J, Finan C, *et al.* Polygenic scores in disease prediction: evaluation using the relevant performance metrics. *medRxiv* 2022;:2022.02.18.22271049. doi:10.1101/2022.02.18.22271049

2 Wand H, Lambert SA, Tamburro C, *et al.* Improving reporting standards for polygenic scores in risk prediction studies. *Nature* 2021;**591**:211–9. doi:10.1038/s41586-021-03243-6

3 Genome UK: the future of healthcare - GOV.UK.

4 Horton R, Crawford G, Freeman L, *et al.* Direct-to-consumer genetic testing. *The BMJ* Published Online First: 2019. doi:10.1136/bmj.l5688

5 Genomics PLC - Genomics plc.

6 Allelica | Polygenic Risk Score.

7 Adeyemo A, Balaconis MK, Darnes DR, *et al.* Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. *Nature Medicine 2021 27:11* 2021;**27**:1876–84. doi:10.1038/s41591-021-01549-6

8 Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet* 2018;**19**:581–90. doi:10.1038/S41576-018-0018-X

9 Lambert SA, Abraham G, Inouye M. Towards clinical utility of polygenic risk scores. *Hum Mol Genet* 2019;**28**:R133–42. doi:10.1093/HMG/DDZ187

10 Wald NJ, Old R. The illusion of polygenic disease risk prediction. *Genet Med* 2019;**21**:1705–7. doi:10.1038/S41436-018-0418-5

11 Sud A, Turnbull C, Houlston R. Will polygenic risk scores for cancer ever be clinically useful? NPJ Precis Oncol. 2021. doi:10.1038/s41698-021-00176-1

12 Lambert SA, Gil L, Jupp S, *et al.* The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. Nat Genet. 2021. doi:10.1038/s41588-021-00783-5

13 R: The R Project for Statistical Computing. https://www.r-project.org/ (accessed 11 Feb 2022).

14 Create Elegant Data Visualisations Using the Grammar of Graphics • ggplot2. https://ggplot2.tidyverse.org/ (accessed 14 Oct 2022).

15 Wald NJ, Morris JK. Assessing Risk Factors as Potential Screening Tests: A Simple Assessment Tool. *Arch Intern Med* 2011;**171**:286–91. doi:10.1001/ARCHINTERNMED.2010.378

16 Sun L, Pennells L, Kaptoge S, *et al.* Polygenic risk scores in cardiovascular risk prediction: A cohort study and modelling analyses. *PLoS Med* 2021;**18**:e1003498. doi:10.1371/JOURNAL.PMED.1003498

17 Riveros-Mckay F, Weale ME, Moore R, *et al.* Integrated Polygenic Tool Substantially Enhances Coronary Artery Disease Prediction. *Circ Genom Precis Med* 2021;**14**:E003304. doi:10.1161/CIRCGEN.120.003304

18 Wald NJ, Hackshaw AK, Frost CD. When can a risk factor be used as a worthwhile screening test? *BMJ : British Medical Journal* 1999;**319**:1562. doi:10.1136/BMJ.319.7224.1562

19 Wald NJ, Bestwick JP. Is the area under an ROC curve a valid measure of the performance of a screening or diagnostic test? *J Med Screen* Published Online First: 2014. doi:10.1177/0969141313517497

20 Inouye M, Abraham G, Nelson CP, *et al.* Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. *J Am Coll Cardiol* 2018;**72**:1883–93. doi:10.1016/J.JACC.2018.07.079

21 Overview | Cardiovascular disease: risk assessment and reduction, including lipid modification | Guidance | NICE.

22 Law MR, Wald NJ. Risk factor thresholds: Their existence under scrutiny. Br Med J. 2002. doi:10.1136/bmj.324.7353.1570

23 Pain O, Gillett AC, Austin JC, *et al.* A Tool for Translating Polygenic Scores onto the Absolute Scale Using Summary Statistics. *medRxiv* 2021;:2021.04.16.21255481. doi:10.1101/2021.04.16.21255481

24 Rose G. Sick individuals and sick populations. *Int J Epidemiol* Published Online First: 1985. doi:10.1093/ije/14.1.32

25 Hingorani AD, Hemingway H. How should we balance individual and population benefits of statins for preventing cardiovascular disease? *BMJ* Published Online First: 2011. doi:10.1136/bmj.c6244

26 Knowles JW, Ashley EA. Cardiovascular disease: The rise of the genetic risk score. *PLoS Med* Published Online First: 2018. doi:10.1371/journal.pmed.1002546

27 Holmes M V., Harrison S, Talmud PJ, *et al.* Utility of genetic determinants of lipids and cardiovascular events in assessing risk. *Nat Rev Cardiol* 2011;**8**:207–21. doi:10.1038/NRCARDIO.2011.6

28 Thompson DJ, Wells D, Selzam S, *et al.* UK Biobank release and systematic evaluation of optimised polygenic risk scores for 53 diseases and quantitative traits. *medRxiv* 2022;:2022.06.16.22276246. doi:10.1101/2022.06.16.22276246

29 Polygenic Risk Scores - Genomics plc.

30 Zhou D, Yu D, Scharf JM, *et al.* Contextualizing genetic risk score for disease screening and rare variant discovery. *Nat Commun* 2021;**12**:1–14. doi:10.1038/s41467-021-24387-z

31 Lu T, Forgetta V, Richards JB, *et al.* Polygenic risk score as a possible tool for identifying familial monogenic causes of complex diseases. *Genetics in Medicine* 2022;**0**:1–11. doi:10.1016/J.GIM.2022.03.022

32 Johnson D, Wilke MAP, Lyle SM, *et al.* A Systematic Review and Analysis of the Use of Polygenic Scores in Pharmacogenomics. *Clin Pharmacol Ther* 2022;**111**:919–30. doi:10.1002/CPT.2520

33 Schmidt AF, Hingorani AD, Finan C. Human Genomics and Drug Development. *Cold Spring Harb Perspect Med* Published Online First: 2021. doi:10.1101/cshperspect.a039230

# 3 The UK Biobank study

## 3.1 Overview

The UK Biobank is a large longitudinal cohort study of approximately 500,000 individuals living in the UK.[1] The participants have been thoroughly phenotyped through self-reported questionnaires, physical measurements (including physiological measures, biomarker measures and imaging data), and medical record linkage (hospital episode statistics (HES) and primary care data). This information is organised into "data fields", which are coded columns containing the relevant measures. Genetic data is also available for most participants through genotyping arrays and whole exome sequencing. Whole genome sequencing of the UK Biobank participants is underway and is expected to be released for all participants in early 2023.

Recruitment of participants started in 2006 up until 2010 in 22 assessment centres across England, Wales and Scotland. Participants' age ranged from 37 to 69 at the time of recruitment. These individuals come from various socio-demographic backgrounds and ethnicities, although the large majority of people (~94%) self-identify as White British and are known to be healthier than the average British population.[2]

## 3.2 Genotyping data

The UK Biobank whole-genome genotyping was done using the UK BiLEVE Axiom Array and the UK Biobank Axiom Array. The arrays are 95% similar, with the first array genotyping 807,411 markers and the second array genotyping 820,967 single nucleotide polymorphisms (SNPs) and short insertions and deletions (indel) markers.[1] 49,979 participants were genotyped using the UK BiLEVE Axiom Array, and the rest of the participants (438,427) were genotyped using the UK Biobank Axiom Array. The genotyping data was phased using SHAPEIT3 and imputed with IMPUTE4 (a recoded version of IMPUTE2) using the Haplotype Reference Consortium and the merged UK10K haplotype and 1000 Genomes phase 3 reference panels.[1,3,4] A total of 93,095,623 autosomal SNPs, indels and large structural variants were imputed. The data was aligned to the positive strand of the reference in chromosome build 37.

The Wellcome Trust Centre for Human Genetics performed an initial marker-based and sample-based quality control (QC) of the data.[1] Briefly, poor quality markers were set to missing, and

unreliable markers across multiple genotype batches were altogether removed from the dataset. Lists of sample IDs (referred to as "eid" in the UK Biobank) with sex mismatches and outliers for missingness and heterozygosity were generated and assigned to specific UK Biobank data fields but were not removed from the study. The only participants removed from the initial QC were sample duplicates and those who wished to withdraw from the study.

The imputed genotype data (Version 3) is available to download in BGEN v1.2 format (.bgen, .sample, .bgi files).[5] The additional QC steps that we performed on these data are explained below.

### 3.3 Quality control (QC)

The UK Biobank is a large multi-ancestry dataset that underwent very minimal QC of its genotyping data prior to release. A thorough QC of the genetic data was therefore required before generating the PGS. Errors arising from poor genotyping, insufficient imputation quality, or even sample mishandling from human error can all lead to significant bias in the analyses if they are not accounted for. The QC steps that were performed by the UK Biobank and those undertaken by our group are detailed in this section. The UK Biobank project ID that was approved for this work is 40721.

#### 3.3.1 Sample QC

As mentioned previously, the UK Biobank provides a list of recommended sample exclusions based on their initial QC of the data. These include outliers for heterozygosity and missingness (UK Biobank data field 22027: 968 individuals) and putative chromosome aneuploidy (UK Biobank data field 22019: 652 individuals).

Sex mismatches between self-reported sex (UK Biobank data field 31) and genetic sex (UK Biobank data field 22001) were identified and participants were excluded (378 individuals). 14,248 participants with missing genetic sex information (i.e. not genotyped) were also excluded. These participants were removed from the dataset using QCTOOL v2.[6]

For the work in this thesis, I only QCed the data for participants of European ancestry from UK Biobank data field 22006 (409,616 individuals). This UK Biobank data field was derived using self-

reported "White British" ethnicity (UK Biobank data field 21000) and principal components analysis of their genotypes. The participant eids from data field 22006 were extracted and added as inclusion samples when running QCTOOL v2.[6]

### 3.3.2 Genetic variant QC

The QC for variant data was done in two stages. The sample exclusions described in the previous section and the first part of the variant QC were performed in a single step, followed by a second variant QC and sample relatedness exclusion step (see **Figure 3.1** for the full workflow of the QC procedure).

The first variant QC stage consisted of excluding all variants with an imputation information (INFO) score <0.3 using bgenix.[7] The INFO score gives a measure of the certainty of the imputation: a value of 1 means complete certainty or directly genotyped, while a value of 0 means complete uncertainty about the imputation of the genetic variant.[8]

Newly recalculated minor allele frequencies (MAF) were also generated for the subset of participants of European ancestry. MAF is the frequency of the rarest allele for a given variant in a cohort and will therefore fluctuate based on the ancestry of the cohort. These MAF values were subsequently used in the second stage of the variant QC.

For the second variant QC stage, I implemented four different genetic variant QC parameters as was done by the Bristol MRC-IEU: genetic variants were removed if they had a MAF <0.1%, MAF <0.5% and INFO <0.9, MAF <1% and INFO <0.8, MAF <3% and INFO <0.6.[9] It is often the case that low MAFs are a result of poor imputation quality (low INFO score). In order to retain a maximum number of genetic variants with low MAFs, we ensured that their INFO score was high: the higher the imputation quality (and certainty), the lower the MAF I allowed. The final number of variants retained after the genetic QC steps was 15,125,437.

### 3.3.3 Relatedness

The inclusion of related individuals in a study cohort can skew the results of some analyses by erroneously enriching some SNPs. An in-house script was written to maximise the number of individuals to keep while removing up to and including third degree relatives (equivalent to a

kinship coefficient greater or equal to 0.0442). The UK Biobank has pre-calculated kinship coefficients for pairs of participants using the KING toolset[10]. The information is provided in the format of a table with three columns: the first two columns represent pairs of participants' unique identifying numbers ("eid") with their kinship coefficient in the third column. All pairs of individuals with a kinship coefficient below 0.0442 were excluded from the table and kept in our analysis as they were more distantly related than third degree relatives.

The "remove_relatedness" script works by first identifying the individual with the most relatives in the UK Biobank. This is done by counting the number of times an eid appears in the table. The eid that appears the most times is the individual who has the largest number of first to third degree relatives in the UK Biobank. That individual and their pairs are then removed from the table, the eid counts are regenerated, and so on. This loop ends when there are only unique eids left in the table, at which point one eid per pairs of individuals left (i.e. either from the first or the second column of the table) is arbitrarily remove. In the end, I obtained a list of 74,704 unique eid to be removed from the QCed dataset.

### 3.3.4 Final QC step: merging the phenotype file

As a final step, the QCed data above was merged with an updated UK Biobank phenotype file to drop any participants who had withdrawn from the study. The phenotype file is continuously updated by the UK Biobank to ensure such removal of participants throughout the course of the study. At the time of this project, the final count of participants of European ancestry obtained after the QC steps was 341,515.

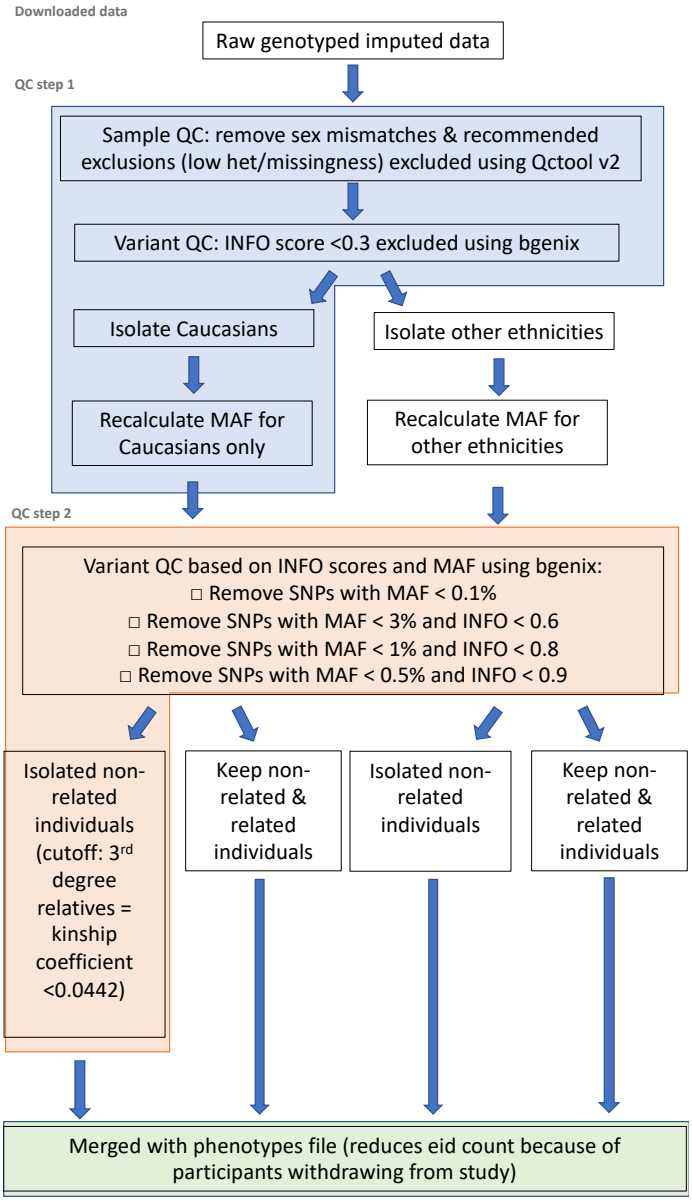



**Figure 3.1 Workflow of the QC steps for the UK Biobank genotype data.** QC step 1 is highlighted in purple and QC step 2 in orange. The highlighted section in green represents the final merge between the genetic data and the sample list. The sample list is continuously updated with participant withdrawals. The non-highlighted QC steps have not been completed. Caucasians refer to individuals of European (White British) ancestry.

## 3.4 References

1 Bycroft C, Freeman C, Petkova D, *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018;**562**:203–9. doi:10.1038/s41586-018-0579-z

2 Fry A, Littlejohns TJ, Sudlow C, *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* 2017;**186**. doi:10.1093/aje/kwx246

3 O'Connell J, Sharp K, Shrine N, *et al.* Haplotype estimation for biobank-scale data sets. *Nature Genetics* 2016;**48**:817–20. doi:10.1038/ng.3583

4 Howie B, Fuchsberger C, Stephens M, *et al.* Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics* 2012;**44**:955–9. doi:10.1038/ng.2354

5 Band G, Marchini J. BGEN : a binary file format for imputed genotype and haplotype data. *bioRxiv* 2018;:1–6. doi:10.1101/308296

6 QCTOOL v2. https://www.well.ox.ac.uk/~gav/qctool/index.html#documentation (accessed 20 Mar 2020).

7 gavinband / bgen / wiki / bgenix — Bitbucket. https://bitbucket.org/gavinband/bgen/wiki/bgenix (accessed 20 Mar 2020).

8 Genotype imputation and genetic association studies of UK Biobank.

9 Mitchell RE, Hemani G, Dudding T, *et al.* UK Biobank Genetic Data: MRC-IEU Quality Control, version 2, 18/01/2019.

10 Manichaikul A, Mychaleckyj JC, Rich SS, *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics (Oxford, England)* 2010;**26**:2867–73. doi:10.1093/bioinformatics/btq559

# 4 Evaluating PRS in the prediction of cardiovascular disease (CVD), coronary heart disease (CHD), type 2 diabetes (T2D), and ischaemic stroke (IST)

## 4.1 Abstract

**Background:** Current clinical risk prediction tools do not incorporate polygenic information, but studies have shown that the addition of polygenic risk scores (PRS) to non-genetic risk prediction tools might improve disease prediction for various disease endpoints. The utility of PRS has not been thoroughly evaluated for the clinically utilised QRISK3, QDiabetes and QStroke prediction tools in the UK population which are used respectively in the 10-year prediction of incident cardiovascular disease (CVD)/coronary heart disease (CHD), type 2 diabetes (T2D) and ischaemic stroke (IST). In this chapter, I evaluated the performance of each of these risk models with and without the incorporation of PRS.

**Methods:** Utilising data from 341,515 White British UK Biobank participants, I externally validated the QRISK3, QStroke and QDiabetes prediction models and expanded these to include PRS for CVD/CHD, IST and T2D, respectively. Disease endpoints were defined using available hospital episode statistics data (ICD-10 and OPCS-4 codes). Weighted PRS for CHD, T2D and IST were generated using external genetic variants and weights from the genome-wide association studies Cardiogram, Diagram, and Megastroke respectively. Model performance was evaluated based on discrimination (C-statistic) and calibration (calibration-in-the-large and calibration slope). The detection rate for a 5% false positive rate (DR5) was calculated to inform the clinical utility of the models. Five different models were evaluated for the prediction of each disease endpoint: 1) PRS, 2) QScore, 3) PRS, age and sex, 4) PRS and QScore, 5) age and sex.

**Results:** There were 14,010 incident CHD events, 23,389 incident CVD events, 2,909 incident IST events and 12,599 incident T2D events observed in the ten years post-enrolment of the study cohort. The mean predicted risk of the PRS were slightly lower for unaffected individuals than the affected individuals for incident CVD (0.068 (SD: 0.019) vs 0.074 (SD: 0.022)), CHD (0.041 (SD: 0.009) vs 0.043 (SD: 0.009)), IST (0.008 (SD: 0.001) vs 0.009 (SD: 0.001)) and T2D (0.036 (SD: 0.021) vs 0.049 (SD: 0.027)). Very little to no correlation was observed between the PRS and their respective QScore (highest correlation: $r = 0.11$ (p-value $<2.2 \times 10^{-16}$) for T2D PRS and QDiabetes). All models calibrated well, except for QStroke in men with or without IST PRS (calibration slope without IST PRS = 2.24 (95% CI: 2.06; 2.41); calibration slope with IST PRS = 20.84 (95% CI:

18.60; 23.08)). The PRS for CVD, CHD and IST poorly discriminated between cases and controls (C-statistic range from 0.55 (95% CI: 0.53; 0.58) to 0.59 (95% CI: 0.58; 0.59)). The addition of PRS to the QScores (QRISK3, QDiabetes, QStroke) did not significantly improve the C-statistic of the QScore models, with the highest increase in the C-statistic (of 0.015) observed for CVD PRS and QRISK3 in men. This translates to an improvement of 1.5% in the DR5. Overall, men benefited more from the addition of PRS to QRISK3 and QDiabetes, while the opposite was found for women.

**Conclusion:** The overlap in PRS distributions between affected and unaffected individuals, coupled with the low C-statistics and DR5 of the PRS models for all outcomes studied indicated that the PRS did not predict the 10-year risk of incident CVD, CHD, IST and T2D very well. I also did not observe a substantial improvement in model discrimination and calibration when adding the PRS to QRISK3, QStroke and QDiabetes. PRS have limited clinical utility in the 10-year prediction of CVD/CHD, IST and T2D.

## 4.2 Introduction

Cardiovascular disease (CVD) is the major cause of death globally. It has a long preclinical phase where many modifiable risk factors influence disease progression and severity over decades.[1] These risk factors are well studied and provide the opportunity for primary prevention, secondary prevention and risk prediction.

Identifying individuals at increased risk has been beneficial in reducing the burden of CVD globally. Many countries have since developed and calibrated unique CVD risk prediction models for their own populations, such as QRISK3 in the United Kingdom (UK), the Framingham Risk Score and the 2013 ACC/AHA Pooled Cohort Equations in the United States, the European SCORE risk charts in Europe, and the FINRISK-calculator in Finland.[2–7] These predictors typically include readily ascertained or routinely collected variables (e.g. age, sex, ethnicity, physiological measurements, family and medical history) that have been specifically calibrated for their populations.

In recent years there has been an interest in incorporating polygenic information into clinical practice for disease prediction.[8–10] Many studies have analysed the predictive utility of polygenic scores in combination with non-genetic CVD risk factors.[11–14] The current consensus is that

polygenic scores provide modest additional discriminative ability, with the non-genetic factors explaining most of the CVD incidence variability.[14–17] Some have argued that the modest additional improvement in discrimination that these genetic scores provide might not warrant integration into clinical care.[15,17–19] While these polygenic scores have been analysed alongside non-genetic risk factors and risk prediction models, these studies have not always applied the non-genetic risk prediction tools in the same population they were developed in and for (e.g. QRISK3 in the UK for the 10-year prediction of CVD when using the UK Biobank dataset), which could lead to less accurate and more variable model performance estimates, and potentially influence (limit or inflate) the added benefit of polygenic score information.[20,21]

In early 2022, the UK National Health Service (NHS) and Genomics plc announced a joint trial combining QRISK3 with polygenic information to identify individuals at high risk of CVD.[22] While this trial will deliver valuable data, I here provide preliminary evidence on the utility of adding polygenic information to the freely available UK-based QRISK3, QStroke and QDiabetes risk prediction models in the UK Biobank for the 10-year prediction of incident CVD/coronary heart disease (CHD), ischaemic stroke (IST), and type 2 diabetes (T2D), respectively.

## 4.3 Methods

### 4.3.1 Implementing QRISK3, QStroke and QDiabetes in the UK Biobank

The UK Biobank is a large ongoing longitudinal cohort study of approximately 500,000 individuals living in the UK, with recruitment of participants in 2006-2010.[23] Participants have been thoroughly phenotyped through self-reported questionnaires, physical measurements (including physiological measures, biomarker measures and imaging data), and medical record linkage (HES and primary care data).

QResearch in the UK have developed publicly available clinical (non-genetic) risk prediction tools for the prediction of various disease endpoints: CVD for QRISK3, IST for QStroke, and T2D for QDiabetes. The implementation of the QRISK3-2018, QStroke-2012 and QDiabetes-2018 algorithms in the UK Biobank in this study used variables referenced from baseline, defined as the date of first attendance of the participants at the UK Biobank assessment centre (**Supplementary Table 4.1**).[2,24,25] Incident events were capped at 10 years post study enrolment. Variables with missing data were imputed using the R package MICE.[26] The QDiabetes variable "learning

difficulties" was not included in the 10-year calculation of incident T2D because the data was not available (**Supplementary Table 4.1**).

### 4.3.2 The prediction models developed and tested

The four incident disease outcomes evaluated in this study were coronary heart disease (CHD), T2D, IST, and CVD (defined as a composite of CHD, all stroke, heart failure and atrial fibrillation). HES data, including the International Classification of Diseases (ICD-10) and the Office of Population Censuses and Surveys Classification of Interventions and Procedures (OPCS-4) codes, were used to define the four disease endpoints in the UK Biobank (**Supplementary Table 4.2**).

Five prediction models were tested for each of these incident disease outcomes (**Table 4.1**). These included a baseline model of age and sex evaluated for all outcomes (model 5). Disease-specific polygenic risk scores (PRS) were generated for each endpoint studied (CVD, CHD, IST, T2D) (model 1), and the QScores (QRISK3-2018, QStroke-2012 and QDiabetes-2018) (model 2) were evaluated in an outcome-specific manner (e.g. QDiabetes for the prediction of incident T2D) (**Table 4.1**). A combined model of PRS with age and sex was tested for each endpoint (model 3), and a model combining the QScore and PRS information (in an outcome-specific manner) was developed to evaluate the incremental predictive utility of adding PRS information to the QScores studied (model 4) (**Table 4.1**). The combined models 3, 4, and 5 in **Table 4.1** used interaction terms in the logistic regression analyses (denoted by the "x" sign), allowing for non-linear interactions between model predictors. The QScores and PRS were logit transformed prior to generating the combined models, and the output of the combined models was transformed back to the risk probability scale.

| Incident CVD | Incident CHD | Incident IST | Incident T2D |
|---|---|---|---|
| 1. PRS for CVD | 1. PRS for CHD | 1. PRS for IST | 1. PRS for T2D |
| 2. QRISK3 | 2. QRISK3 | 2. QStroke | 2. QDiabetes |
| 3. CVD PRS x age x sex | 3. CHD PRS x age x sex | 3. IST PRS x age x sex | 3. T2D PRS x age x sex |
| 4. CVD PRS x QRISK3 | 4. CHD PRS x QRISK3 | 4. IST PRS x QStroke | 4. T2D PRS x QDiabetes |
| 5. Age x sex | 5. Age x sex | 5. Age x sex | 5. Age x sex |

**Table 4.1 The prediction models evaluated for incident CVD, CHD, IST and T2D.** " x " denotes the interaction term used in the models. CHD = coronary heart disease; CVD = cardiovascular disease; IST = ischaemic stroke; PRS = polygenic risk score; T2D = type 2 diabetes.

### 4.3.3 Data QC of the UK Biobank

The UK Biobank genotyping data underwent minimal quality control (QC) prior to its release. The QC of the dataset is explained in more detail in **Chapter 3**. Briefly, I performed additional QC steps based on Bristol's MRC-IEU protocol and excluded the UK Biobank's recommended exclusions of outliers for heterozygosity and missingness (data field 22027), and putative chromosome aneuploidy (data field 22019).[27] I removed individuals with sex mismatches between self-reported sex (data field 31) and genetic sex (data field 22001), and with missing genetic sex information. I selected individuals of White British ancestry (data field 22006) and removed up to 3$^{rd}$ degree relatives. All genetic variants with an imputation information (INFO) score of <0.3 were excluded. Genetic variants were also removed if they had a minor allele frequency (MAF) <0.1%, MAF <0.5% and INFO <0.9, MAF <1% and INFO <0.8, MAF <3% and INFO <0.6.[27] The total number of participants remaining in the study cohort following these QC steps was 341,515.

### 4.3.4 Generating PRS for CVD, CHD, IST and T2D

Weighted PRS for CHD, T2D and IST were generated for the entire QCed UK Biobank dataset using external genome-wide association study (GWAS) summary statistics genetic variants and weights from Cardiogram, Diagram, and Megastroke, respectively (**Table 4.2**).[28–30] The CVD PRS was calculated by combining a PRS for CHD, PRS for all stroke, PRS for heart failure, and PRS for atrial fibrillation in a logistic regression analysis with incident CVD as the outcome and the PRS as predictors (allowing for interactions between PRS predictors in the regression analysis) (**Table 4.2**).[28,30–32] The GWAS summary statistics were restricted to participants of European ancestries, matching the ancestry of the study cohort. For each PRS, a combination of p-value ($5x10^{-8}$, $5x10^{-7}$, $5x10^{-6}$, $5x10^{-5}$, $5x10^{-4}$) and linkage disequilibrium (LD) (0.8, 0.6, 0.4, 0.2, 0.01) cut-off values were tested. For each disease endpoint, the PRS with the highest area under the curve (AUC), equivalent to the C-statistic for binary outcomes, was retained (**Supplementary Figure 4.1**).

| Trait | Consortium | Source | Article link | Data download link | Access to data download | Studied SNPs in discovery GWAS | Cases | Controls | Additional information |
|---|---|---|---|---|---|---|---|---|---|
| Coronary artery disease (CHD) | Cardiogram | Nikpay *et al.* 2015[28] | https://www.nature.com/articles/ng.3396 | http://www.cardiogramplusc4d.org/data-downloads/ | 13th January 2020 | 9,455,778 | 60,801 | 123,504 | CAD additive; populations: European, South Asian, East Asian |
| Type 2 diabetes (T2D) | Diagram | Scott *et al.* 2017[29] | https://diabetes.diabetesjournals.org/content/66/11/2888 | https://diagram-consortium.org/downloads.html | 4th February 2020 | 10,221,232 | 26,676 | 132,532 | Not adjusted for BMI; European ancestry |
| Any ischaemic stroke (IST) | Megastroke | Malik *et al.* 2018[30] | https://www.nature.com/articles/s41588-018-0058-3 | http://www.megastroke.org/download.html | 25th November 2019 | 1,216,870 | 34,217 | 406,111 | European ancestry |
| Atrial fibrillation (AF) | HUNT, deCODE, MGI, DiscovEHR, UK Biobank, AFGen | Nielsen *et al.* 2018[32] | https://www.nature.com/articles/s41588-018-0171-3 | http://csg.sph.umich.edu/willer/public/afib2018/ | 29th January 2021 | 34,740,186 | 60,620 | 970,216 | European ancestry |
| Heart failure (HF) | HERMES | Shah *et al.* 2020[31] | https://www.nature.com/articles/s41467-019-13690-5 | https://cvd.hugeamp.org/dinspector.html?dataset=GWAS_HERMES_eu | 29th January 2021 | 3,468,278 | 47,309 | 930,014 | European ancestry |
| Any stroke | Megastroke | Malik *et al.* 2018[30] | https://www.nature.com/articles/s41588-018-0058-3 | https://www.megastroke.org/download.html | 25th November 2019 | 8,254,765 | 67,162 | 454,450 | European ancestry |

**Table 4.2 The GWAS summary statistics used to generate the PRS.** Cardiogram for CHD PRS, Diagram for T2D PRS, and Megastroke for IST PRS. The PRS for AF, HF and any stroke were combined with the CHD PRS to form the CVD PRS. AF = atrial fibrillation; CAD = coronary artery disease; CHD = coronary heart disease; GWAS = genome-wide association study; HF = heart failure; IST = ischaemic stroke; PRS = polygenic risk score; SNP = single nucleotide polymorphism; T2D = type 2 diabetes.

### 4.3.5 Assessment of the prediction models

The QCed UK Biobank cohort of 341,515 White British participants was split into approximately 50% training (n = 171,338) and 50% testing (n = 170,180) data. The models in the test data were recalibrated using data from the training set: the logistic regression models were first fitted onto the training data and the calibration slope and intercept values were re-adjusted in the test data. The scores were then transformed back to the risk probability scale. The calibration (calibration-in-the-large and calibration slope) and discrimination (C-statistic/AUC) were evaluated for each model in **Table 4.1** of the test data. The increase in the odds of disease per one standard deviation (SD) of the PRS and QScores were obtained after calculating the Z-scores and fitting a logistic regression with the respective incident diseases as outcomes and the Z-scores as predictors. The

Z-scores were calculated by subtracting the mean of the scores from the score for each individual and dividing the whole by the SD of the score distributions. The correlation between PRS and their respective QScore was computed using the non-parametric Spearman correlation (*rho*) from the R package stats version 4.0.2.[33] The detection rate (or sensitivity) for a 5% false positive rate (DR5) of the models were calculated based on the C-statistic of the models. These measures are commonly used in clinical settings for evaluating clinical models and tests. The calculations are described in more detail in **Chapter 2** and in the manuscript by Hingorani *et al.*[18]

### 4.3.6 Software

All data analysis was performed in R version 4.0.2.[33] The study's participant characteristics table was produced using the R package tableone version 0.12.0, and the p-values of group differences between sexes in **Table 4.3** were calculated using the Kruskal-Wallis Rank sum nonparametric test for continuous variables and the Man-Whitney U test for binary variables.[34] The plots were generated with ggplot2 version 3.3.5.[35]

## 4.4 Results

### 4.4.1 Characteristics of study participants

In total, 341,515 White British participants from the UK Biobank were included in the analysis whose summary characteristics are shown in **Table 4.3** and **Supplementary Table 4.3**. There were 14,010 incident CHD cases, 23,389 incident CVD cases, 2,909 incident IST cases and 12,599 incident T2D cases observed in the cohort for a follow up time of 10 years post study enrolment. Significant group difference between sexes were observed for all variables with the exception of prevalent severe mental illness (p-value = 0.997) and congestive cardiac failure (p-value = 0.076) (**Table 4.3**).

| | **Female** | **Male** | **P-value of group differences** | **Missing (%)** |
|---|---|---|---|---|
| n (%) | 183651 (53.8%) | 157864 (46.2%) | | |
| Age (median [IQR]) | 58.0 [50.0, 63.0] | 59.0 [51.0, 64.0] | <0.001 | 0.0 |
| BMI, kg/m2 (median [IQR]) | 26.1 [23.4, 29.6] | 27.3 [25.0, 30.0] | <0.001 | 0.3 |
| Cholesterol ratio (median [IQR]) | 3.7 [3.1, 4.4] | 4.3 [3.6, 5.2] | <0.001 | 12.8 |
| Systolic blood pressure, mmHg (median [IQR]) | 133.5 [121.5, 147.5] | 140.0 [129.0, 152.0] | <0.001 | 0.2 |
| Smoking status (%) | | | <0.001 | 3.6 |

| | | | | |
|---|---|---|---|---|
| Non-smoker | 109168 (60.9) | 77149 (51.4) | | |
| Former smoker | 58387 (32.6) | 61682 (41.1) | | |
| Light smoker (<10 cigarettes/day) | 2802 (1.6) | 1678 (1.1) | | |
| Moderate smoker (10-19 cigarettes/day) | 5376 (3.0) | 4397 (2.9) | | |
| Heavy Smoker (>20 cigarettes/day) | 3589 (2.0) | 5148 (3.4) | | |
| Townsend deprivation index (median [IQR]) | -2.8 [-3.8, 0.0] | -2.7 [-3.8, 0.1] | 0.004 | 0.1 |
| Family history of CHD (%) | 84064 (50.0) | 64025 (46.4) | <0.001 | 10.4 |
| Family history of type 2 diabetes (%) | 38740 (23.6) | 30308 (22.7) | <0.001 | 12.7 |
| **Prescription history** | | | | |
| Statins (%) | 17638 (9.6) | 28171 (17.8) | <0.001 | 0.0 |
| Atypical antipsychotics (%) | 453 (0.2) | 474 (0.3) | 0.003 | 0.0 |
| Corticosteroids (%) | 2885 (1.6) | 2631 (1.7) | 0.028 | 0.0 |
| Erectile dysfunction (%) | 0 (0.0) | 1897 (1.2) | <0.001 | 0.0 |
| Treated hypertension (%) | 8833 (4.8) | 11070 (7.0) | <0.001 | 0.0 |
| NSAIDs (%) | 46701 (25.7) | 45247 (29.0) | <0.001 | 1.1 |
| Anticoagulants (%) | 1099 (0.6) | 2653 (1.7) | <0.001 | 0.0 |
| **Prevalent medical conditions** | | | | |
| Atrial fibrillation (%) | 1310 (0.7) | 3388 (2.1) | <0.001 | 0.0 |
| Congestive cardiac failure (%) | 5 (0.0) | 12 (0.0) | 0.076 | 0.0 |
| Coronary heart disease (%) | 2640 (1.4) | 8743 (5.5) | <0.001 | 0.0 |
| Chronic kidney disease (stage 4 or 5) (%) | 172 (0.1) | 312 (0.2) | <0.001 | 0.0 |
| Cardiovascular disease (%) | 4124 (2.2) | 11737 (7.4) | <0.001 | 0.0 |
| Gestational diabetes (%) | 220 (0.1) | 0 (0.0) | <0.001 | 0.0 |
| Ischaemic stroke (%) | 373 (0.2) | 874 (0.6) | <0.001 | 0.0 |
| Manic depression/schizophrenia (%) | 397 (0.2) | 449 (0.3) | <0.001 | 0.0 |
| Migraine (%) | 2171 (1.2) | 652 (0.4) | <0.001 | 0.0 |
| Polycystic ovary syndrome (%) | 142 (0.1) | 0 (0.0) | <0.001 | 0.0 |
| Rheumatoid arthritis (%) | 1054 (0.6) | 457 (0.3) | <0.001 | 0.0 |
| Systemic Lupus Erythematosus (%) | 156 (0.1) | 31 (0.0) | <0.001 | 0.0 |
| Severe mental illness (%) | 677 (0.4) | 583 (0.4) | 0.997 | 0.0 |
| Valvular heart disease (%) | 636 (0.3) | 1042 (0.7) | <0.001 | 0.0 |
| Type 2 diabetes (%) | 3100 (1.7) | 5622 (3.6) | <0.001 | 0.0 |
| Type 1 diabetes (%) | 214 (0.1) | 271 (0.2) | <0.001 | 0.0 |
| Renal disease (%) | 226 (0.1) | 388 (0.2) | <0.001 | 0.0 |
| **Incident disease** | | | | |
| CHD (%) | 4609 (2.5) | 9401 (6.0) | <0.001 | 0.0 |
| CVD (%) | 8626 (4.7) | 14763 (9.4) | <0.001 | 0.0 |
| Ischaemic stroke (%) | 1109 (0.6) | 1800 (1.1) | <0.001 | 0.0 |
| Type 2 diabetes (%) | 5051 (2.8) | 7548 (4.8) | <0.001 | 0.0 |

**Table 4.3 Characteristics of 341,515 UK Biobank White British participants included in the analysis stratified by sex.** The p-values of group differences between sexes were obtained using the Kruskal-Wallis Rank sum nonparametric test for continuous variables, and the Man-Whitney U test for binary variables. The percentage of missing data for each variable are shown in the last column. BMI = body mass index; CHD = coronary heart disease; CVD = cardiovascular disease; IQR = interquartile range; NSAID = non-steroidal anti-inflammatory drug.

### 4.4.2 PRS for CHD, IST and T2D

The PRS studied were selected based on the highest C-statistic calculated for each incident disease endpoint after testing various p-value and LD cut-off values. The optimal p-value threshold was $5x10^{-4}$ in most cases, while the optimal LD cut-off values were 0.2 and 0.6 (**Supplementary Figure 4.1** and **Supplementary Table 4.4**). The C-statistic of the PRS ranged from 0.547 to 0.658, with the lowest one obtained for the IST PRS in the prediction of incident CVD, and the highest one obtained for the T2D PRS for the prediction of incident T2D (**Supplementary Figure 4.1** and **Supplementary Table 4.4**).

### 4.4.3 Relationship of QScores and PRS to incident CVD, CHD, IST, T2D

#### 4.4.3.1 Score distributions

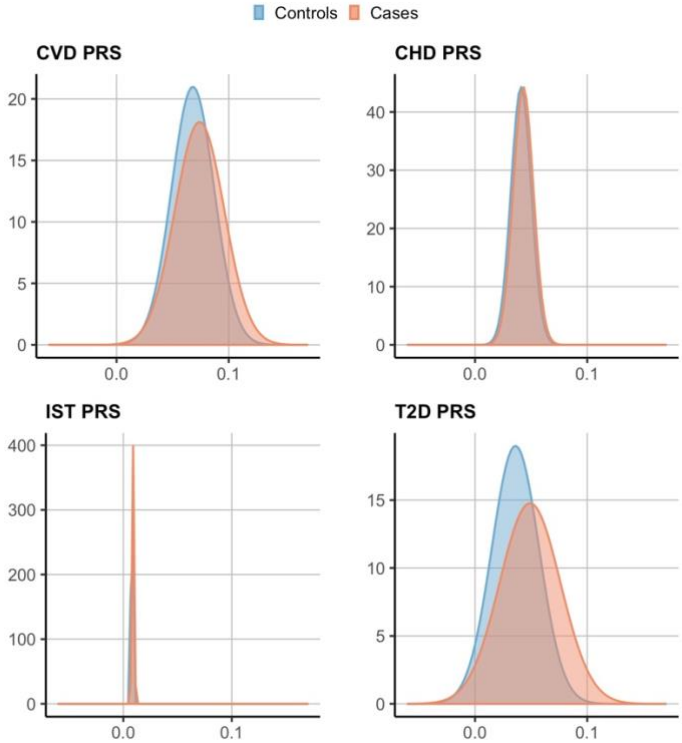



**Figure 4.1 The CVD, CHD, IST and T2D PRS distributions for cases and controls.** The standard normal distributions (based on the mean and SD) of the PRS are plotted on the same scale on the x-axis against the density of the distributions on the y-axis. Cases are in red and controls in blue. CHD = coronary heart disease; CVD = cardiovascular disease; IST = ischaemic stroke; PRS = polygenic risk score; SD = standard deviation; T2D = type 2 diabetes.

The distributions of the QScores (QRISK3, QStroke and QDiabetes) and the PRS for CVD, CHD, IST and T2D were overlapping for cases and controls (**Figures 4.1** and **4.2**). The separation of the mean of the distributions for cases and controls was bigger for all QScores than their respective PRS, indicating better discrimination (**Figure 4.2** and **Table 4.4**). This separation of the means for cases and controls was outcome-specific: the biggest separations were observed for incident T2D (QDiabetes and T2D PRS), and the smallest for incident IST (QStroke and IST PRS) (**Figures 4.1, 4.2** and **Table 4.4**).

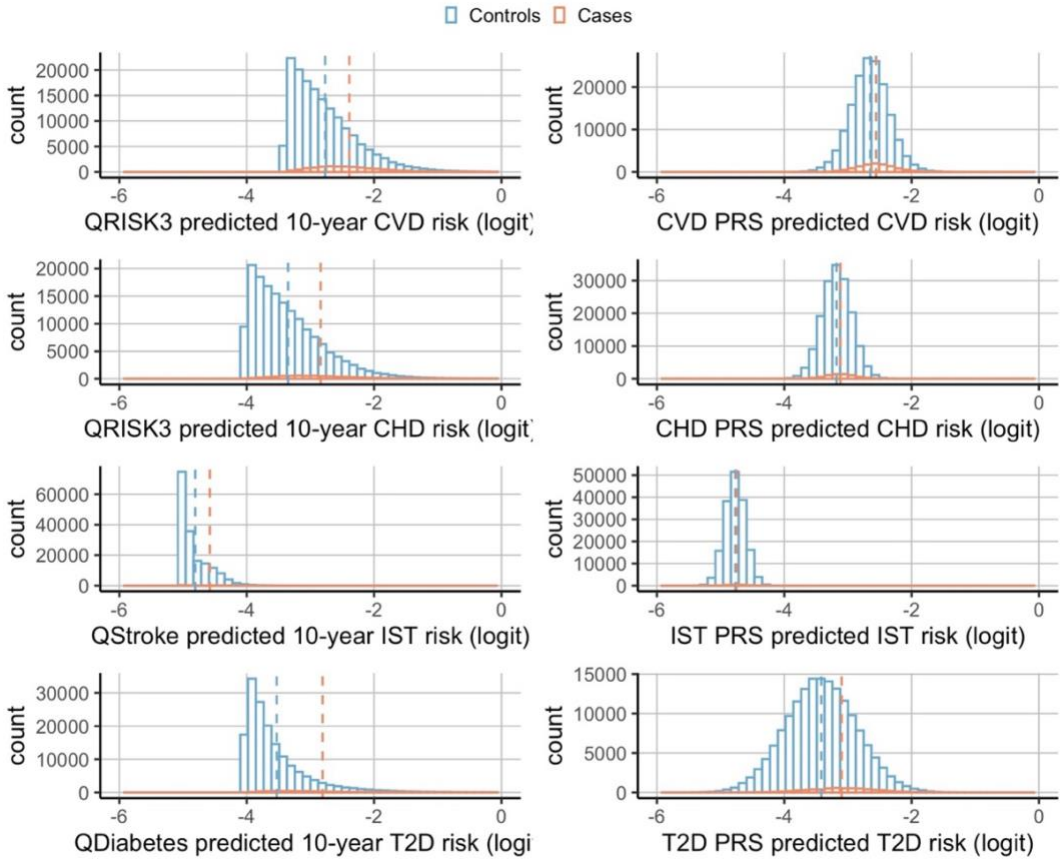



**Figure 4.2 Logit transformed risk distributions for CVD, CHD, IST and T2D PRS, and QRISK3, QStroke and QDiabetes.** The plots were generated using the test data. The logit predicted risks are plotted on the same x-axis scale. The mean of the distributions is depicted by the dotted lines. CHD = coronary heart disease; CVD = cardiovascular disease; IST = ischaemic stroke; PRS = polygenic risk score; T2D = type 2 diabetes.

| | Controls: mean (SD) | Cases: mean (SD) | Mean difference |
|---|---|---|---|
| **CVD PRS** | 0.068 (0.019) | 0.074 (0.022) | 0.006 |
| **CHD PRS** | 0.041 (0.009) | 0.043 (0.009) | 0.002 |
| **IST PRS** | 0.008 (0.001) | 0.009 (0.001) | 0.001 |
| **T2D PRS** | 0.036 (0.021) | 0.049 (0.027) | 0.013 |
| **QRISK3 for incident CVD** | 0.067 (0.047) | 0.095 (0.062) | 0.028 |
| **QRISK3 for incident CHD** | 0.040 (0.036) | 0.068 (0.063) | 0.028 |
| **QStroke** | 0.008 (0.011) | 0.012 (0.037) | 0.004 |
| **QDiabetes** | 0.035 (0.046) | 0.080 (0.098) | 0.045 |

**Table 4.4 The mean predicted risk and SD of the distributions for cases and controls for CVD PRS, CHD PRS, IST PRS, T2D PRS, QRISK3 for incident CVD, QRISK3 for incident CHD, QStroke and QDiabetes.** CHD = coronary heart disease; CVD = cardiovascular disease; IST = ischaemic stroke; PRS = polygenic risk score; SD = standard deviation; T2D = type 2 diabetes.

#### 4.4.3.2 Score odds ratio

Mean disease incidence increased per risk score decile for QRISK3, QDiabetes, CVD PRS, CHD PRS, and T2D PRS, while this increase was less important for QStroke and IST PRS (**Figure 4.3**). This means that there was a log-linear increase in risk, and the relationship between incident disease risk and score increase can be described in a single value: the odds ratio (OR) per SD. The OR were higher for QRISK3 than for their respective PRS (CVD and CHD) (e.g. 1.4 (standard error (SE): 1.01) for QRISK3 versus 1.32 (SE: 1.01) for CVD PRS), while the IST PRS and T2D PRS showed a higher OR of incident disease per one SD than their respective QScores (e.g. 1.18 (SE: 1.03) for the IST PRS versus 1.06 (SE: 1.01) for QStroke) (**Figure 4.3**).

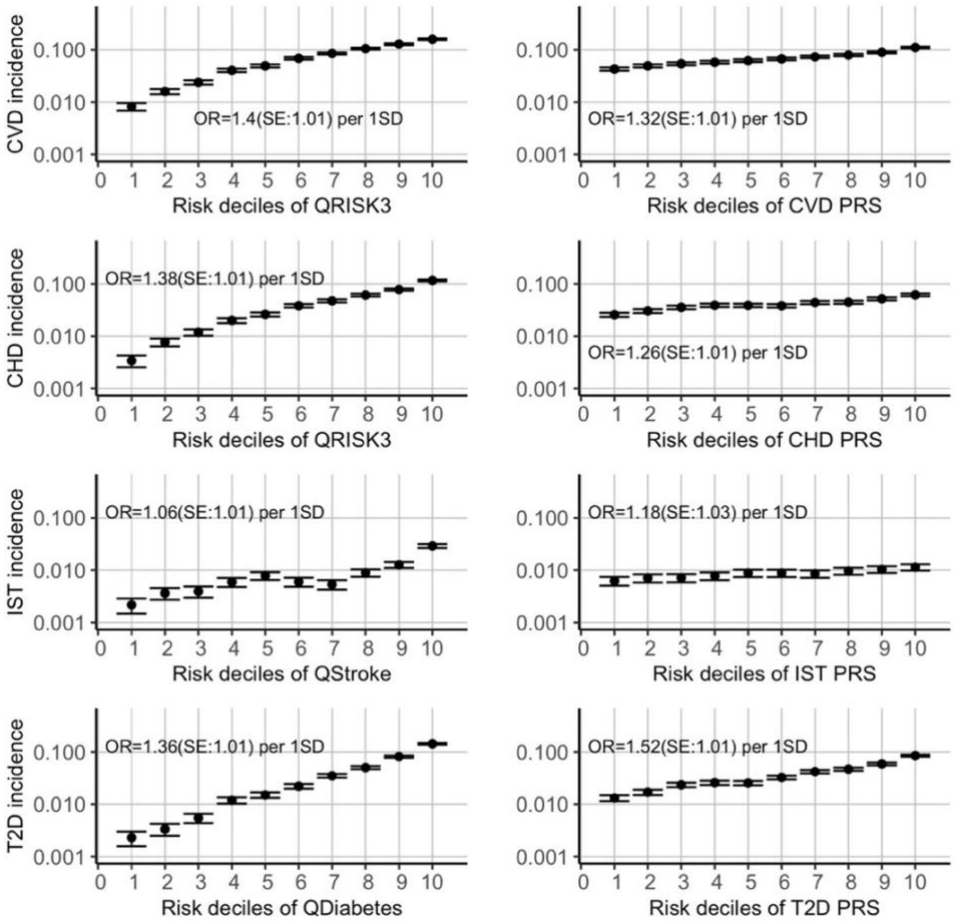



**Figure 4.3 Log mean incident disease (CVD, CHD, IST, T2D) per risk score decile for CVD, CHD, IST, T2D PRS and QRISK3, QStroke and QDiabetes.** The y-axis represents the log mean incident disease per risk score decile (on the x-axis). The OR increase per one SD of the scores are indicated on the plots by the diamonds with the SE depicted by the horizontal lines. The QScores (QRISK3, QStroke, QDiabetes) are shown on the left side of the figure, and the PRS on the right side of the figure. CHD = coronary heart disease; CVD = cardiovascular disease; IST = ischaemic stroke; OR = odds ratio; PRS = polygenic risk score; SE = standard error; SD = standard deviation; T2D = type 2 diabetes.

#### 4.4.3.3 Correlation between PRS and QScores

There was very little correlation observed between the PRS and their respective QScore. The highest spearman correlation coefficient *rho* (*r*) obtained was for T2D PRS and QDiabetes ($r$ = 0.11; p-value $<2.2 \times 10^{-16}$) and the lowest one observed was for IST PRS and QStroke ($r$ = 0.017; p-value $<1.5 \times 10^{-12}$) (**Figure 4.4**). This suggests that PRS could potentially provide additional information for disease prediction not currently captured by the QScores.

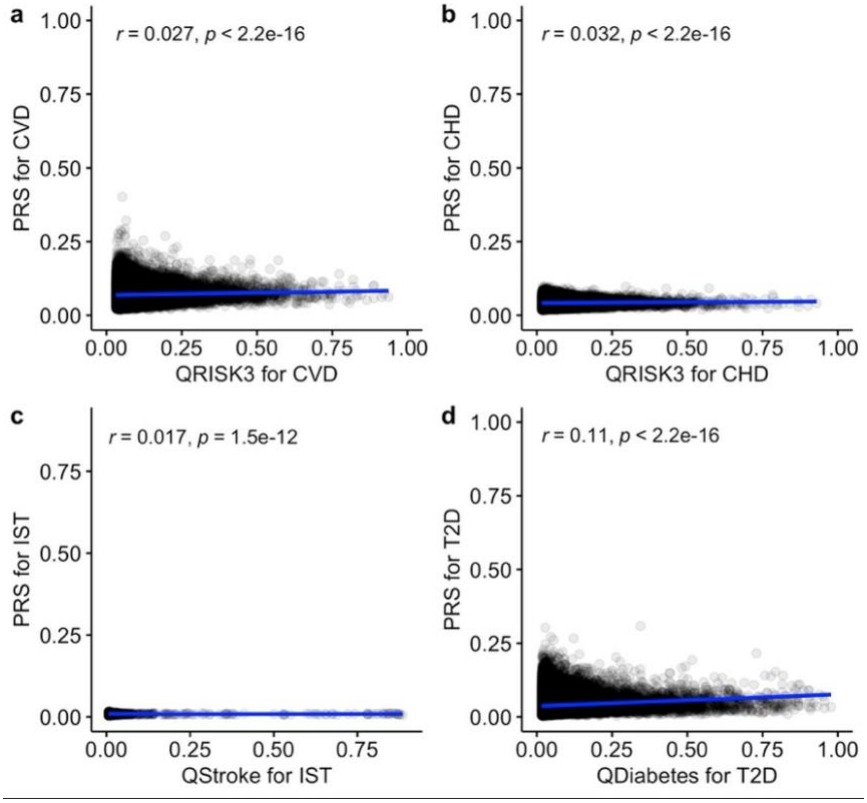



**Figure 4.4 Correlation scatter plots of PRS and QScores for the prediction of incident outcomes. a)** between CVD PRS and QRISK3 for the prediction of incident CVD; **b)** between CHD PRS and QRISK3 for the prediction of incident CHD; **c)** between IST PRS and QStroke for the prediction of incident IST; **d)** between T2D PRS and QDiabetes for the prediction of incident T2D. *r* refers to the spearman correlation coefficient *rho*, and *p* is the p-value of the strength of association of the correlations. CHD = coronary heart disease; CVD = cardiovascular disease; IST = ischaemic stroke; PRS = polygenic risk score; T2D = type 2 diabetes.

### 4.4.4 Calibration of the models tested

The calibration of the models was assessed by plotting the mean observed and mean predicted risks for each model, depicting agreement. For all outcomes studied, men had a higher observed risk of disease than that predicted by their PRS (**Figure 4.5**). For both men and women combined (overall), the calibration-in-the-large of the PRS models (**Table 4.1** model 1) showed good agreement between observed and predicted risk (with a difference in means close to 0) (**Supplementary Table 4.5**).

The age and sex combined model (**Table 4.1** model 5) also calibrated well for all disease outcomes (**Figure 4.5**). The inclusion of PRS to the age and sex models (**Table 4.1** model 3) did not affect model calibration (**Figure 4.5** and **Supplementary Table 4.5**).

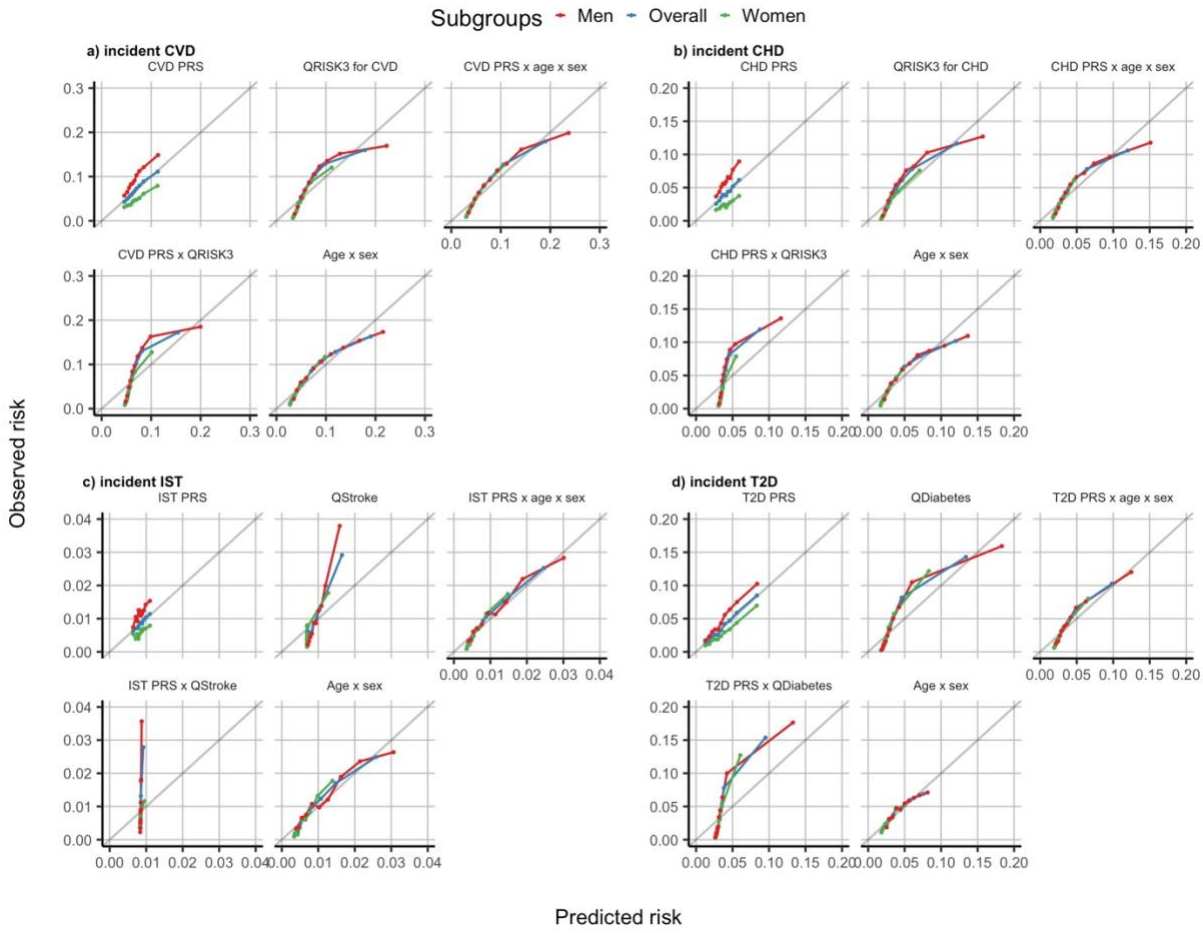



**Figure 4.5 The calibration curves of the prediction models (CVD PRS, CHD PRS, IST PRS, QRISK3, QStroke, QDiabetes or combined models) tested for incident CVD, CHD, IST and T2D.** The mean observed risk (y-axis) for each decile is plotted against the mean predicted risk (x-axis) for each decile in the test data. The deciles are shown as dots on the graph. A subgroup analysis by sex is depicted in red for men, green for women, and both (overall) in blue. The "x" sign refers to the interaction term used in the regression analyses. **(a)** Models for CVD prediction include CVD PRS, QRISK3, CVD PRS x age x sex, CVD PRS x QRISK3, age x sex. **(b)** Models for CHD prediction include CHD PRS, QRISK3, CHD PRS x age x sex, CHD PRS x QRISK3, age x sex. **(c)** Models for IST prediction include IST PRS, QStroke, IST PRS x age x sex, IST PRS x QStroke, age x sex. **(d)** Models for T2D prediction include T2D PRS, QDiabetes, T2D PRS x age x sex, T2D PRS x QDiabetes, age x sex. CHD = coronary heart disease; CVD = cardiovascular disease; IST = ischaemic stroke; PRS = polygenic risk score; T2D = type 2 diabetes.

The calibration of the QScores (**Table 4.1** model 2) was dependent on the score: QStroke showed the poorest calibration with individuals in the highest risk deciles having their risk of incident IST underestimated (especially in men: calibration slope = 2.24 (95% CI: 2.06; 2.41)), while QRISK3 and QDiabetes slightly underestimated the risk of incident CVD, CHD and T2D in the middle

risk deciles (**Figure 4.5**). The calibration of the QScores combined with the PRS (**Table 4.1** model 4) followed the calibration patterns of the QScores (**Figure 4.5**).

### 4.4.5 Discrimination of the models tested

A model's ability to differentiate accurately between cases and controls was evaluated by comparing the C-statistic of the models where 0.5 indicates no discrimination and 1 indicates perfect discrimination. The PRS (**Table 4.1** model 1) had the lowest C-statistic for incident CVD, CHD and IST (C-statistic range from 0.548 (95% CI: 0.530; 0.567) for IST PRS in men to 0.588 (95% CI: 0.581; 0.595) for CVD PRS in men) (**Figure 4.6** and **Supplementary Table 4.5**). The C-statistic of the T2D PRS outperformed that of the age and sex model (**Table 4.1** model 5) (C-statistic of 0.659 (95% CI: 0.652; 0.666) for T2D PRS versus C-statistic of 0.636 (95% CI: 0.629; 0.642) for the age and sex model) (**Figure 4.6** and **Supplementary Table 4.5**).

The C-statistics of the PRS models improved when combining them with the age and sex variables (**Table 4.1** model 3) for all disease endpoints (C-statistic range from 0.668 (95% CI: 0.661; 0.675) for CHD PRS with age and sex in men to 0.724 (95% CI: 0.704; 0.745) for IST PRS with age and sex in women) (**Figure 4.6** and **Supplementary Table 4.5**).

In terms of the QScores' (**Table 4.1** model 2) discriminative ability in the UK Biobank, QDiabetes had the highest C-statistic (0.802 (95% CI: 0.797; 0.807)) and QStroke the lowest (0.698 (95% CI: 0.684; 0.712)) (**Figure 4.6** and **Supplementary Table 4.5**).

The addition of polygenic information to the QScores (**Table 4.1** model 4) did not greatly improve the C-statistic of the QScore models for all incident disease outcomes studied (**Figure 4.6** and **Table 4.5**). The highest increase in the C-statistic (of 0.015) was observed when combining the CVD PRS with QRISK3 in men for the 10-year prediction of incident CVD (**Table 4.5**). This is equivalent to an increase of 1.5% in the detection rate for a 5% false positive rate (see methods section). In some instances, the addition of a PRS to the QScore worsened the discrimination of the QScore, as was the case with the IST PRS and QStroke in men, women and overall (**Table 4.5**).

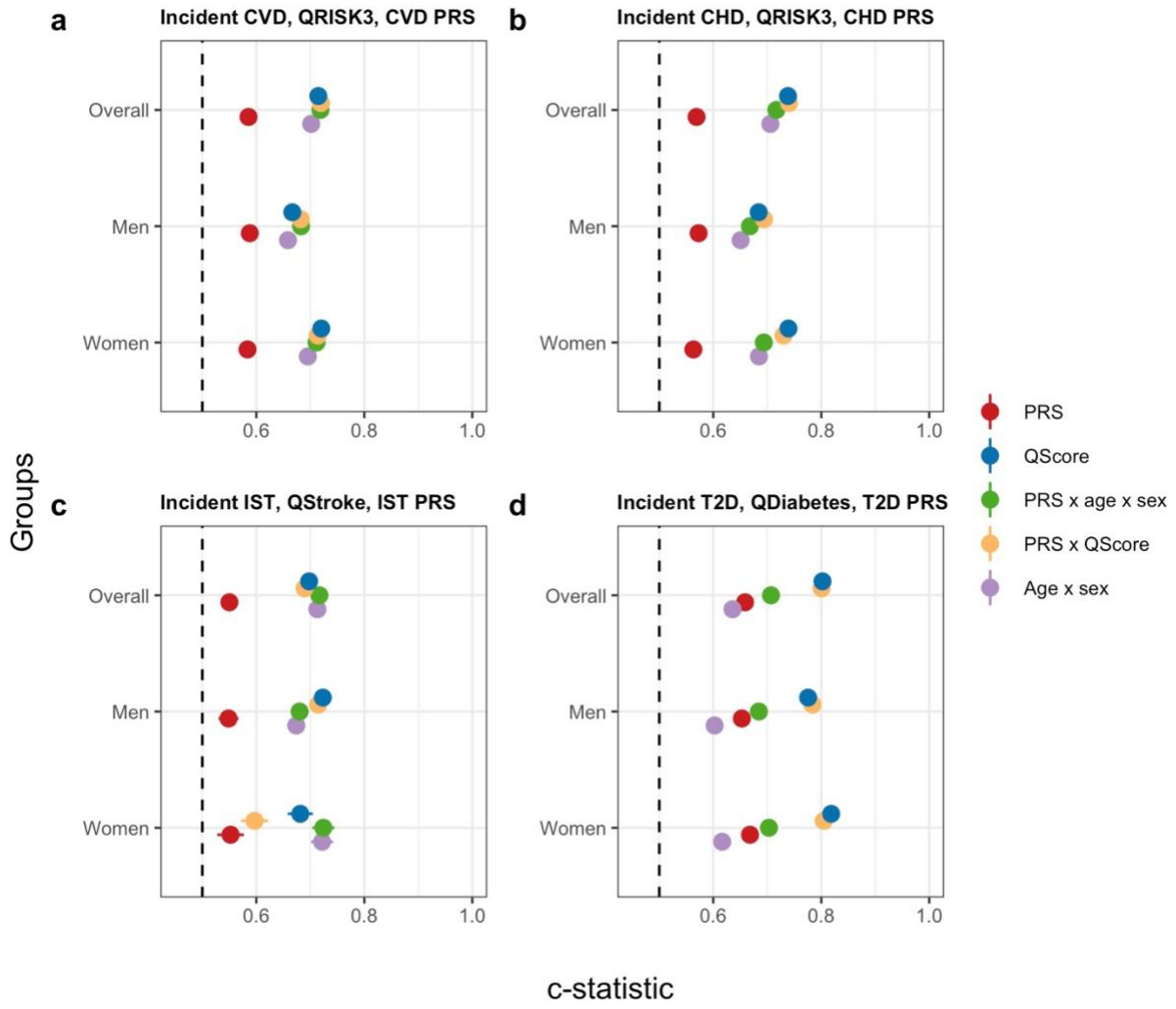



**Figure 4.6 The discrimination (C-statistic) of the prediction models (PRS, QScore and combined models) tested for incident CVD, CHD, IST and T2D.** The discriminative ability of the models in the test data is quantified by the C-statistic on the x-axis. The 95% confidence intervals (CI) are depicted by lines protruding from the dots as shown in the figure legend (narrow 95% CI are not visible as they are incorporated into the datapoint). Model discrimination is shown by subgroup of sex (y-axis). The "x" sign refers to the interaction term used in the regression analyses. **(a)** Models for CVD prediction include CVD PRS, QRISK3, CVD PRS x age x sex, CVD PRS x QRISK3, age x sex. **(b)** Models for CHD prediction include CHD PRS, QRISK3, CHD PRS x age x sex, CHD PRS x QRISK3, age x sex. **(c)** Models for IST prediction include IST PRS, QStroke, IST PRS x age x sex, IST PRS x QStroke, age x sex. **(d)** Models for T2D prediction include T2D PRS, QDiabetes, T2D PRS x age x sex, T2D PRS x QDiabetes, age x sex. CHD = coronary heart disease; CVD = cardiovascular disease; IST = ischaemic stroke; PRS = polygenic risk score; T2D = type 2 diabetes.

Model discrimination was also found to be sex-specific in some cases. For example, the C-statistic for the age and sex model (**Table 4.1** model 5), for the PRS combined with age and sex (**Table 4.1** model 3) for all outcomes studied, for QRISK3 and for QDiabetes were all higher in women than men (**Figure 4.6** and **Supplementary Table 4.5**). Meanwhile, the CHD PRS had a slightly

higher C-statistic in men (0.599 (95% CI: 0.581; 0.595)) than in women (0.563 (95% CI: 0.551; 0.575)), as did QStroke (C-statistic of 0.723 (95% CI: 0.707; 0.739) in men and of 0.681 (95% CI: 0.658; 0.705) in women) (**Figure 4.6** and **Supplementary Table 4.5**). Overall, men benefited more (in terms of improvement in the C-statistic) from the addition of PRS to QScores (with the exception of QStroke), while the opposite was observed for women for all outcomes studied (**Table 4.5**).

| Models: QScore vs QScore x PRS | Outcome | Subgroup | Change in C-statistic | Change in DR5 |
|---|---|---|---|---|
| QRISK3 vs CVD PRS x QRISK3 | Incident CVD | Overall | 0.005 | 0.006 |
| QRISK3 vs CVD PRS x QRISK3 | Incident CVD | Female | -0.006 | -0.007 |
| QRISK3 vs CVD PRS x QRISK3 | Incident CVD | Male | 0.015 | 0.015 |
| QRISK3 vs CHD PRS x QRISK3 | Incident CHD | Overall | 0.002 | 0.002 |
| QRISK3 vs CHD PRS x QRISK3 | Incident CHD | Female | -0.009 | -0.012 |
| QRISK3 vs CHD PRS x QRISK3 | Incident CHD | Male | 0.01 | 0.01 |
| QStroke vs IST PRS x QStroke | Incident IST | Overall | -0.008 | -0.008 |
| QStroke vs IST PRS x QStroke | Incident IST | Female | -0.084 | -0.067 |
| QStroke vs IST PRS x QStroke | Incident IST | Male | -0.009 | -0.011 |
| QDiabetes vs T2D PRS x QDiabetes | Incident T2D | Overall | -0.002 | -0.003 |
| QDiabetes vs T2D PRS x QDiabetes | Incident T2D | Female | -0.013 | -0.025 |
| QDiabetes vs T2D PRS x QDiabetes | Incident T2D | Male | 0.008 | 0.013 |

**Table 4.5 Change in the C-statistic and detection rate for a 5% false positive rate between the QScore and the QScores x PRS models.** DR5 = detection rate for a 5% false positive rate; CHD = coronary heart disease; CVD = cardiovascular disease; IST = ischaemic stroke; PRS = polygenic risk score; T2D = type 2 diabetes.

### 4.4.6 Detection rate of cases by the models for a 5% false positive rate

The detection rate for a 5% false positive rate (DR5) is a well-established performance metric that is commonly used to evaluate new clinical models in screening and prediction. The DR5 of the models tested in this study are shown in **Supplementary Table 4.5.** Since the AUC and the DR5 are linked, the higher the C-statistic, the higher the DR5. For example, for the QScore that showed the best improvement in the C-statistic when adding a PRS (i .e. the QRISK3 x CVD PRS model

in men): the DR5 for QRISK3 was equal to 15%, while the DR5 for CVD PRS was equal to 9.2%, and the DR5 of the combined QRISK3 x CVD PRS model was equal to 16.5% (**Supplementary Table 4.5**). This means that for a 5% false positive rate, adding a PRS improved the detection rate of the non-genetic clinical model by 1.5% at best out of all the QScores and sex-subgroups tested (**Table 4.5**).

### 4.4.7 Comparison with similar studies

Comparisons of the study results with previous key publications evaluating CHD and/or CVD PRS in disease prediction are shown in **Table 4.6**. The C-statistic (or AUC) of the PRS ranged from 0.56 (95% CI: 0.56; 0.57) to 0.81 (95% CI: 0.81; 0.81) (**Table 4.6**). This large range can be explained by the various covariates that were included (or not) in the PRS models when evaluated for CVD or CHD prediction. A PRS with over 3.5 million variants (and without other covariates in the model) (C-statistic = 0.662 (95% CI: 0.658; 0.665)) did not greatly outperform a PRS with 88 times less variants ($n$ = 40,079) (C-statistic = 0.61 (95% CI: 0.60; 0.62)) (**Table 4.6**).[15,36] The incremental increase in the C-statistic when adding PRS to a non-genetic model in these studies ranged from 0.002 to 0.03 (**Table 4.6**).

| Publication | # individuals | Outcome definitions | # cases | PRS construction method | # variants in PRS | Performance metric (95% confidence interval) | Covariates in model | Non-genetic model evaluated | Change in C-statistic with PRS added to non-genetic model | Detection rate for a 5% false positive rate |
|---|---|---|---|---|---|---|---|---|---|---|
| **Khera *et al.* (Nat Gen 2018)**[37] | Validation dataset: n = 120,280; Testing dataset: n = 288,978 | CAD: self-reported, ICD-9, ICD-10, OPCS-4 codes | Prevalent CAD in testing dataset: 8,676 | Weighted PRS (LDPred) | 6,630,150 | AUC = 0.81 (0.81; 0.81) | Age, sex, genotyping array, first 4 principal components of ancestry | N/A | N/A | 34% |
| **Inouye et al. (J Am Coll Cardiol 2018)**[14] | 482,629 | CAD: self-reported, ICD-9, ICD-10, OPCS-4 codes | 22,242 total: 12,513 incident cases + 9,729 prevalent cases | Weighted PRS from 3 published genetic risk scores | 1.7 million | C-index (prevalent cases) = 0.623 (0.615; 0.630) | Genotyping array, 10 first genetic principal components of ancestry | 6 risk factors: smoking, diabetes, family history of heart disease, BMI, hypertension, high cholesterol | 0.073 | 12% |
| **Elliott *et al.* (JAMA 2020)**[15] | Derivation dataset: n = 15,947; Testing dataset: n = 352,660 | CAD: mortality data, ICD-9, ICD-10, OPCS-4 codes; CVD = CAD + angina + stroke | Incident CAD: 6,272; incident CVD: 13,753 | Weighted PRS (lassosum) | CAD PRS: 40,079; CVD PRS: 297,862 | C-statistic for CAD = 0.61 (0.60; 0.62); C-statistic for CVD = 0.56 (0.56; 0.57) | N/A | Pooled cohort equation | 0.02 | 11% for CAD PRS; 8% for CVD PRS |
| **Sun *et al.* (PLOS Med 2021)**[8] | 306,654 | CVD = CHD + IST | 5,680 incident CVD = 3,333 CHD + 2,347 IST | Weighted PRS from 3 published genetic risk scores | CAD PRS: 1.7 million; IST PRS: 3.2 million | HR CHD PRS = 1.31 (1.27; 1.34); HR IST PRS = 1.18 (1.15; 1.21); C-statistic N/A | Age, smoking status, history of diabetes, systolic BP, total cholesterol, HDL-C | 7 risk factors: age, sex, smoking status, history of diabetes, systolic BP, total cholesterol, high LDL-C | CVD PRS: 0.012; CAD PRS: 0.022; IST PRS: 0.003 | 8% for CHD PRS; 7% for IST PRS |
| **Riveros-Mckay *et al.* (Circ Genom Precis Med 2021)**[36] | Training dataset: n = 60,000; Testing dataset: n = 212,563 | CAD: self-reported, ICD-9, ICD-10, OPCS-4 codes | 4,247 incident CAD | Weighted PRS (LDPred) | >3.5 million | C-statistic = 0.662 (0.658; 0.665) | N/A | Pooled cohort equation | 0.03 | 15% |
| **Current analysis** | Training dataset: n = 171,335; Testing dataset: n = 170,180 | CVD = CHD + all stroke + HF + AF: ICD-10, OPCS-4 codes | Incident CVD: 11,673; incident CHD: n = 6,973 | Weighted PRS | 2,642 | C-statistic for CVD PRS = 0.586 (0.580; 0.591); C-statistic for CHD PRS = 0.569 (0.562; 0.576) | N/A | QRISK3 | CVD PRS: 0.005; CHD PRS: 0.002 | 9% for CVD PRS; 8% for CHD PRS |

**Table 4.6 Comparison of PRS in previous key studies for CHD and CVD prediction.** The incremental increase in the C-statistic of the models when comparing the non-genetic model to the combined model (non-genetic + PRS) is shown. AF = atrial fibrillation; AUC = area under the receiver-operating characteristic curve; BMI = body mass index; BP = blood pressure; CAD = coronary artery disease; CHD = coronary heart disease; CVD = cardiovascular disease; HES = hospital episode statistic; HDL-C = high-density lipoprotein cholesterol; HF = heart failure; HR = hazard ratio; ICD = International Classification of Disease; IST = ischaemic stroke; LDL-C = low-density lipoprotein cholesterol; MI = myocardial infarction; OPCS = Office of Population Censuses and Surveys 4 Classification of Interventions and Procedures; PRS = polygenic risk score.

## 4.5 Discussion

### 4.5.1 Overview of the study

This study evaluated whether the addition of PRS to the clinically used QScores (QRISK3, QStroke and QDiabetes) in the UK population improved model performance for the 10-year prediction of incident CVD, CHD, IST and T2D.

Since the QScores evaluated here were derived specifically for the UK population, the dataset used for this study was also from the UK: the UK Biobank. The UK Biobank is a large ongoing longitudinal cohort study of approximately 500,000 participants that provides genetic data and information on non-genetic variables required for the calculation of the non-genetic and genetic risk scores.[23]

QRISK3 calculates a person's 10-year risk of CVD. There is currently no GWAS on the CVD definition that matches that of QRISK3's; the closest GWAS being for CHD. While I also generated a PRS for CVD by combining PRS for CHD, all stroke, heart failure, and atrial fibrillation (more detail in the methods section), matching the outcome CVD definition of QRISK3; evaluating both incident CVD and CHD ensured a more accurate like-for-like comparison of the genetic and non-genetic risk prediction models assessed in this study.

### 4.5.2 An external validation of the QScores in the UK Biobank

This study provided an external validation of QRISK3, QStroke and QDiabetes. QRISK3 and QDiabetes had good calibration and discrimination, despite obtaining a lower AUC/C-statistic than the original publications.[2,24,25] Overall, QDiabetes had the best discrimination (C-statistic = 0.802 (95% CI: 0.797; 0.807)) and calibration (calibration-in-the-large = 0.003 (95% CI: -0.023; 0.029); calibration slope = 0.975 (95% CI: 0.950; 1.001)) metrics out of all the QScores studied. QStroke did not calibrate well for men in the UK Biobank and discriminated relatively poorly in women (C-statistic = 0.681 (95% CI: 0.658; 0.705)). This could be because the freely available C code for QStroke dates to 2012, whereas the current interactive calculator was updated in 2018. The weights of the QStroke variables in this study are likely to be outdated. There were also overall less incident IST cases in the UK Biobank than the other disease outcomes studied, which influences the confidence of the predictions (**Supplementary Table 4.2**).

### 4.5.3 The performance of PRS in CVD, CHD, IST and T2D prediction

While a higher PRS was associated with a higher rate of disease incidence for all outcomes studied, the CVD, CHD and IST PRS showed the worst calibration and discrimination (C-statistic <0.588 (95% CI: 0.581; 0.595)) metrics out of all the models tested. The T2D PRS outperformed the age and sex model. Likely explanations include variable disease aetiology (e.g. earlier age of onset of T2D), the quality and size of published GWAS that are available for generating the outcome-specific PRS (e.g. 10 million single nucleotide polymorphism hits for the T2D Diagram GWAS versus 1 million hits for IST Megastroke GWAS), and the fact that CVD-related endpoints (especially CVD, CHD and IST) are highly age and sex-dependent.[29,30,38]

### 4.5.4 The effects of age and sex in incident disease prediction

The addition of the age and sex variables to the PRS model improved the C-statistic for all the incident outcomes studied. This is important to highlight as some publications do not clearly report all the variables included in their prediction models when evaluating their PRS in disease prediction.[18,37] The high AUC (or C-statistic) of these PRS models is likely to be driven by other covariates in the model more than by the genetic score itself (**Table 4.6**). This was clearly shown by Lello *et al.* where the inclusion of age and sex improved the C-statistic of 16 PRS for complex traits.[39]

The magnitude of this effect was also seen to be outcome specific: the addition of T2D PRS to the age and sex model for the prediction of incident T2D in this study lead to the highest increase in the C-statistic (of 0.03) compared to the age and sex model alone out of all the outcomes studied. This is likely because the T2D PRS had relatively good discrimination on its own (see previous section) and therefore its combination with uncorrelated variables (age and sex) resulted in this higher increase observed. A study looked at the relative effect of adding a T2D genetic risk score to modifiable and non-modifiable risk factors and observed that this effect was greater in younger and leaner participants.[40] While the median age of UK Biobank participants was 58 to 59 at baseline (**Table 4.3**), they are known to be healthier than the general population, which might explain why this larger increase in the C-statistic was observed for T2D.[41] The influence of genetics in the development of T2D is more important in individuals with lower adverse factors for T2D (such as lower BMI and age).

For CVD prediction, the CHD and CVD PRS alone performed poorly in this study. When combining the PRS with age and sex, the C-statistic obtained were similar to that of QRISK3 which also includes the age and sex variables. The C-statistic of the age and sex model only was also very similar (**Supplementary Table 4.5**). Wald *et al.* have highlighted the importance that age plays in CVD prediction, proposing that age alone is an appropriate and more cost-effective screening strategy for CVD than current clinical risk models such as the Framingham risk score.[38] Furthermore, a study by Khan *et al.* analysed the effects of adding PRS to a non-genetic risk score for CVD stratified by age.[42] They observed that the C-statistic of the combined model only marginally improved in young adults after adding the PRS, while it did not improve in midlife adults. This shows that the relevance of PRS in risk prediction differs with age, and that PRS provide minimal added benefits to non-genetic risk model for the prediction of CVD in later life when it is usually employed.

### 4.5.5 The combined PRS and QScore models

The minimal correlation between the QScores and their respective PRS seemed to suggest that PRS could potentially add novel predictive information to these scores (**Figure 4.4**). To evaluate the incremental predictive utility of polygenic information in the QScores, I generated combined models in an outcome-specific way: CVD PRS and QRISK3 for the incidence of CVD, CHD PRS and QRISK3 for the incidence of CHD, IST PRS and QStroke for the incidence of IST, and T2D PRS and QDiabetes for the incidence of T2D. All combined models had a similar calibration and discrimination as their respective QScores with overlapping confidence intervals (**Figures 4.5**, **4.6** and **Supplementary Table 4.5**). This suggests that the PRS generated in this study provided minimal added predictive value to QRISK3, QStroke and QDiabetes for the prediction of incident CVD/CHD, IST and T2D, respectively. This might be because the non-genetic variables included in QRISK3, QStroke and QDiabetes might already sufficiently mediate the PRS associations with the outcomes studied (e.g. the inclusion of family history of premature CVD as a predictor in QRISK3 and QStroke). The modest improvement in the C-statistic when comparing QRISK3 to a combined model of QRISK3 and PRS (in this study an increase of 0.015 in men, and of 0.006 for men and women combined) was also previously reported by Elliott *et al.* (an increase of 0.02 in the C-statistic) and Riveros-Mckay *et al.* (an increase of 0.03 in the C-statistic).[15,36] This minimal improvement in the C-statistic can be more readily interpreted by converting it to a

detection rate for a 5% false positive rate: a 0.015 improvement in the C-statistic is equivalent to an improvement of 1.5% in the detection rate for a 5% false positive rate (see results section).

### 4.5.6 Study limitations

A limitation of the study is that the UK Biobank cohort is known to be healthier than the general UK population, which might impact the prediction metrics obtained here.[41] However, considering this, PRS might play a more important role in incident disease prediction in this study than they would in a general population where adverse environmental variables have a bigger influence on disease risk. Regardless, more evidence is needed to properly assess the utility of including polygenic information to QRISK3, such as the combined trial by the NHS and Genomics plc.[22]

Another study consideration is that I limited the analyses to individuals of White British ancestry. The reason for this was the poor availability of GWAS in non-European ancestries. Subgroup analyses in different ethnic groups might produce variable results; however, due to the poor transferability of PRS in different ancestries, I do not expect to see an improvement in the performance of these combined genetic and non-genetic risk prediction models.[43]

The GWAS summary statistics used to generate the atrial fibrillation PRS and the heart failure PRS in this study contained data from the UK Biobank. These PRS were combined with the CHD PRS and the all stroke PRS to form the CVD PRS analysed here. The reason for using these data was that there were no other GWAS datasets available for atrial fibrillation and heart failure that did not contain UK Biobank information. This could potentially lead to an overinflation in the CVD PRS predictions observed in this study and could mean that these effects would be attenuated if implemented in an external and independent dataset, emphasising the limited utility of PRS in CVD prediction.[44]

### 4.5.7 Conclusion

The results of this study echo what previous studies have found: PRS can add some discriminative value to non-genetic CVD scores (in this case QRISK3), but this improvement is minimal, meaning that the clinical utility of PRS is still unconvincing at this point. The inclusion of PRS to QStroke and QDiabetes also did not improve the discrimination of the models for men and

women combined; highlighting that the utility of PRS information in disease prediction is outcome- and model-specific, likely driven by the underlying aetiology of the disease in question**.** The benefits of PRS in disease prediction are also likely to be more useful in younger individuals where external environmental risk factors have had less of an effect on disease progression and risk. However, considering the large impact that non-genetic risk factors (including age and sex) have on CVD risk, it is unlikely that PRS would be of much use for predicting CVD events in later life. Further critical evaluation of the inclusion of polygenic information to non-genetic risk prediction models is required prior to clinical implementation. Cost-effectiveness studies will also be needed to appropriately guide discussions on possible implementation in clinical settings. Other uses of polygenic scores in the clinic besides risk prediction are also worth exploring, such as its potential in rare variant discovery.

### 4.6 References


1 DAWBER TR, MOORE FE, MANN G V. Coronary heart disease in the Framingham study. *Am J Public Health Nations Health* 1957;**47**:4–24. doi:10.2105/ajph.47.4_pt_2.4

2 Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* 2017;**357**. doi:10.1136/BMJ.J2099

3 Wilson PWF, D'Agostino RB, Levy D, *et al.* Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;**97**:1837–47. doi:10.1161/01.CIR.97.18.1837

4 Goff DC, Lloyd-Jones DM, Bennett G, *et al.* 2013 ACC/AHA guideline on the assessment of cardiovascular risk: A report of the American college of cardiology/American heart association task force on practice guidelines. Circulation. 2014;**129**:49–73. doi:10.1161/01.cir.0000437741.48606.98

5 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. doi:10.1016/S0195-668X(03)00114-3

6 Vartiainen E, Laatikainen T, Peltonen M, *et al.* Predicting Coronary Heart Disease and Stroke: The FINRISK Calculator. Glob Heart. 2016;**11**:213–6. doi:10.1016/j.gheart.2016.04.007

7 Damen JAAG, Hooft L, Schuit E, *et al.* Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* 2016;**353**. doi:10.1136/BMJ.I2416

8 Sun L, Pennells L, Kaptoge S, *et al.* Polygenic risk scores in cardiovascular risk prediction: A cohort study and modelling analyses. *PLoS Med* 2021;**18**:e1003498. doi:10.1371/JOURNAL.PMED.1003498

9 Polygenic risk: What's the score? https://www.nature.com/articles/d42473-019-00270-w (accessed 1 Apr 2022).

10 Lambert SA, Abraham G, Inouye M. Towards clinical utility of polygenic risk scores. *Hum Mol Genet* 2019;**28**:R133–42. doi:10.1093/HMG/DDZ187

11 Abraham G, Rutten-Jacobs L, Inouye M. Risk Prediction Using Polygenic Risk Scores for Prevention of Stroke and Other Cardiovascular Diseases. *Stroke* 2021;:2983–91. doi:10.1161/STROKEAHA.120.032619

12 Läll K, Mägi R, Morris A, *et al.* Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores. *Genetics in Medicine 2017 19:3* 2016;**19**:322–9. doi:10.1038/gim.2016.103

13 Weale ME, Riveros-Mckay F, Selzam S, *et al.* Validation of an Integrated Risk Tool, Including Polygenic Risk Score, for Atherosclerotic Cardiovascular Disease in Multiple Ethnicities and Ancestries. *Am J Cardiol* 2021;**148**:157–64. doi:10.1016/J.AMJCARD.2021.02.032

14 Inouye M, Abraham G, Nelson CP, *et al.* Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. *J Am Coll Cardiol* 2018;**72**:1883. doi:10.1016/J.JACC.2018.07.079

15 Elliott J, Bodinier B, Bond TA, *et al.* Predictive Accuracy of a Polygenic Risk Score–Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease. *JAMA* 2020;**323**:636–45. doi:10.1001/JAMA.2019.22241

16 Mars N, Koskela JT, Ripatti P, *et al.* Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat Med* 2020;**26**:549–57. doi:10.1038/s41591-020-0800-0

17 Mosley JD, Gupta DK, Tan J, *et al.* Predictive Accuracy of a Polygenic Risk Score Compared With a Clinical Risk Score for Incident Coronary Heart Disease. *JAMA* 2020;**323**:627–35. doi:10.1001/JAMA.2019.21782

18 Hingorani AD, Gratton J, Finan C, *et al.* Polygenic scores in disease prediction: evaluation using the relevant performance metrics. *medRxiv* 2022;:2022.02.18.22271049. doi:10.1101/2022.02.18.22271049

19 Lyssenko V, Jonsson A, Almgren P, *et al.* Clinical Risk Factors, DNA Variants, and the Development of Type 2 Diabetes. *New England Journal of Medicine* 2008;**359**:2220–32. doi:10.1056/nejmoa0801869

20 Farzadfar F. Cardiovascular disease risk prediction models: challenges and perspectives. *Lancet Glob Health* 2019;**7**:e1288–9. doi:10.1016/S2214-109X(19)30365-1

21 Brautbar A, Pompeii LA, Dehghan A, *et al.* A genetic risk score based on direct associations with coronary heart disease improves coronary heart disease risk prediction in the Atherosclerosis Risk in Communities (ARIC), but not in the Rotterdam and Framingham Offspring, Studies. *Atherosclerosis* 2012;**223**:421–6. doi:10.1016/J.ATHEROSCLEROSIS.2012.05.035

22 NHS launches new polygenic scores trial for heart disease - Genomics Education Programme. https://www.genomicseducation.hee.nhs.uk/blog/nhs-launches-new-polygenic-scores-trial-for-heart-disease/ (accessed 5 Apr 2022).

23 Bycroft C, Freeman C, Petkova D, *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018;**562**:203–9. doi:10.1038/S41586-018-0579-Z

24 Hippisley-Cox J, Coupland C. Development and validation of QDiabetes-2018 risk prediction algorithm to estimate future risk of type 2 diabetes: cohort study. *BMJ* 2017;**359**:j5019. doi:10.1136/BMJ.J5019

25 Hippisley-Cox J, Coupland C, Brindle P. Derivation and validation of QStroke score for predicting risk of ischaemic stroke in primary care and comparison with other risk scores: a prospective open cohort study. *BMJ* 2013;**346**. doi:10.1136/BMJ.F2573

26 van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *J Stat Softw* 2011;**45**:1–67. doi:10.18637/jss.v045.i03

27 Mitchell RE, Hemani G, Dudding T, *et al.* UK Biobank Genetic Data: MRC-IEU Quality Control, version 2, 18/01/2019.

28 Nikpay M, Goel A, Won H-H, *et al.* A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nat Genet* 2015;**47**:1121–30. doi:10.1038/ng.3396

29 Scott RA, Scott LJ, Mägi R, *et al.* An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes* 2017;**66**:2888–902. doi:10.2337/db16-1253

30 Malik R, Chauhan G, Traylor M, *et al.* Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat Genet* 2018;**50**:524–37. doi:10.1038/s41588-018-0058-3

31 Shah S, Henry A, Roselli C, *et al.* Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure. *Nature Communications 2020 11:1* 2020;**11**:1–12. doi:10.1038/s41467-019-13690-5

32 Nielsen JB, Thorolfsdottir RB, Fritsche LG, *et al.* Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nature Genetics 2018 50:9* 2018;**50**:1234–9. doi:10.1038/s41588-018-0171-3

33 R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. doi:ISBN 3-900051-07-0

34 Panos A, Mavridis D. TableOne: an online web application and R package for summarising and visualising data. *Evid Based Ment Health* 2020;**23**:127–30. doi:10.1136/EBMENTAL-2020-300162

35 Wickham H. ggpolt2 Elegant Graphics for Data Analysis. *Use R! series* 2016;:211.

36 Riveros-Mckay F, Weale ME, Moore R, *et al.* Integrated Polygenic Tool Substantially Enhances Coronary Artery Disease Prediction. *Circ Genom Precis Med* 2021;**14**:192–200. doi:10.1161/CIRCGEN.120.003304

37 Khera A V., Chaffin M, Aragam KG, *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics 2018 50:9* 2018;**50**:1219–24. doi:10.1038/s41588-018-0183-z

38 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. doi:10.1371/journal.pone.0018742

39 Lello L, Raben TG, Yong SY, *et al.* Genomic Prediction of 16 Complex Disease Risks Including Heart Attack, Diabetes, Breast and Prostate Cancer. *Scientific Reports 2019 9:1* 2019;**9**:1–16. doi:10.1038/s41598-019-51258-x

40 Langenberg C, Sharp SJ, Franks PW, *et al.* Gene-Lifestyle Interaction and Type 2 Diabetes: The EPIC InterAct Case-Cohort Study. *PLoS Med* 2014;**11**:e1001647. doi:10.1371/journal.pmed.1001647

41 Fry A, Littlejohns TJ, Sudlow C, *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* 2017;**186**. doi:10.1093/aje/kwx246

42 Khan SS, Page C, Wojdyla DM, *et al.* Predictive Utility of a Validated Polygenic Risk Score for Long-Term Risk of Coronary Heart Disease in Young and Middle-Aged Adults. *Circulation* 2022;:101161CIRCULATIONAHA121058426. doi:10.1161/CIRCULATIONAHA.121.058426

43 Martin AR, Kanai M, Kamatani Y, *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* 2019;**51**:584–91. doi:10.1038/s41588-019-0379-x

44 Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols 2020 15:9* 2020;**15**:2759–72. doi:10.1038/s41596-020-0353-1

## 4.7 Appendix

**Supplementary Table 4.1 The QRISK3, QDiabetes and QStroke variables and their corresponding UK Biobank data field numbers and ICD-10 diagnostic codes.** CHD = coronary heart disease; ICD = International Classification of Disease.

| | QRISK3 variables (model B) | QStroke variables | QDiabetes variables | UK Biobank data field number | UK Biobank data field 41270: ICD-10 diagnosis codes |
|---|---|---|---|---|---|
| sex | | | | 31 | |
| age | | | | 21003 | |
| ethnicity | | | | 22006 | |
| Townsend deprivation index | | | | 189 | |
| body mass index | | | | 21001 | |
| treated hypertension | | | | 20002 | I10 |
| smoking | | | | 20116, 3456 | |
| type 1 diabetes | | | | | E10 |
| type 2 diabetes | | | | | E11 |
| systolic blood pressure | | | | 93, 4080 | |
| total cholesterol:high density lipoprotein cholesterol ratio | | | | 30690, 30760 | |
| family history of CHD in a 1st degree relative <60 yo | | | | 20107, 20110, 20111 | |
| rheumatoid arthritis | | | | | M06.9, M05.0, M05.1, J99.0, M06.1, M06.4 |
| atrial fibrillation | | | | | I48 |
| major chronic renal disease | | | | | N04, N03.2, N11.1, Z99.2, Z94.0 |
| chronic kidney disease (stages 3, 4 or 5) | | | | | N18.0, N18.3, N18.4, N18.5, N18.9 |
| migraine | | | | | G43, G44.0 |
| corticosteroids | | | | | |
| systemic lupus erythematosus | | | | | M32.9, M32.1, I39 |
| atypical antipsychotics | | | | | |
| severe mental illness (schizophrenia, bipolar disorder, moderate/severe depression) | | | | | F23, F28, F29, F20, F31 |
| erectile dysfunction diagnosis | | | | | N52.9 |
| chronic kidney disease (stages 4 or 5) | | | | | N18.0, N18.4, N18.5, N18.9 |

| | | | | | |
|---|---|---|---|---|---|
| congestive cardiac failure | | | | | I50.0 |
| coronary heart disease | | | | | G45, I20, I21, I22, I23, I24, I25 |
| valvular heart disease | | | | | I34, I35, I05, I06, I08 |
| cardiovascular disease | | | | | G45, I20, I21, I22, I23, I24, I25, I63, I64 |
| gestational diabetes | | | | | O244, O249 |
| polycystic ovary syndrome | | | | | E282 |
| statins | | | | | |
| family history of diabetes | | | | 20107 | |
| manic depression or schizophrenia | | | | | F29, F20, F31 |
| learning difficulties | | | | N/A | |

**Supplementary Table 4.2 The ICD-10 and OPCS-4 codes used to define incident cardiovascular disease, coronary heart disease, ischaemic stroke, and type 2 diabetes.** The number of incident cases relative to the study baseline (i.e. date of first attendance at the UK Biobank assessment centre) as of January 2021. ICD = International Classification of Diseases; OPCS = Office of Population Censuses and Surveys Classification of Interventions and Procedures.

| | ICD-10 codes | OPCS-4 codes | Number of incident cases in entire dataset | Number of incident cases in the test data |
|---|---|---|---|---|
| **Cardiovascular disease** | I21, I22, I23, I24, I25.0, I25.1, I25.2, I25.3, I25.5, I25.6, I25.8, I25.9, I48, I50, I11.0, I13.0, I13.2, I32.2, I61, I63 | K40, K41, K42, K43, K44.1, K44.8, K44.9, K45.1, K45.2, K45.3, K45.4, K45.5, K45.8, K45.9, K46.1, K46.2, K46.3, K46.4, K46.8, K46.9, K47.1, K49, K50, K75, K52.1, K57.1, K57.5, K62.1, K62.2, K62.3, K62.4, K62.5, X50.1, X50.2 | 23,389 | 11,673 |
| **Coronary heart disease** | I21, I22, I23, I24, I25.0, I25.1, I25.2, I25.3, I25.5, I25.6, I25.8, I25.9 | K40, K41, K42, K43, K44.1, K44.8, K44.9, K45.1, K45.2, K45.3, K45.4, K45.5, K45.8, K45.9, K46.1, K46.2, K46.3, K46.4, K46.8, K46.9, K47.1, K49, K50, K75 | 14,010 | 6,973 |
| **Ischaemic stroke** | I63 | | 2,909 | 1,456 |
| **Type 2 diabetes** | E11 | | 12,599 | 6,293 |

**Supplementary Table 4.3 Characteristics of 341,515 UK Biobank White British participants included in the analysis with missing data singly imputed stratified by sex.** The p-values of group differences between sexes were obtained using the Kruskal-Wallis Rank sum nonparametric test for continuous variables, and the Man-Whitney U test for binary variables. The percentage of missing data for each variable are shown in the last column. BMI = body mass index; CHD = coronary heart disease; CVD = cardiovascular disease; IQR = interquartile range; NSAID = non-steroidal anti-inflammatory drug.

| | Female | Male | p-value for group difference |
|---|---|---|---|
| n (%) | 183651 (53.8%) | 157864 (46.2%) | |
| Age (median [IQR]) | 58.00 [50.00, 63.00] | 59.00 [51.00, 64.00] | <0.001 |

| | | | |
|---|---|---|---|
| BMI, kg/m2 (median [IQR]) | 26.06 [23.43, 29.62] | 27.30 [24.99, 30.05] | <0.001 |
| Cholesterol ratio (median [IQR]) | 3.68 [3.11, 4.42] | 4.34 [3.61, 5.17] | <0.001 |
| Systolic blood pressure, mmHg (median [IQR]) | 133.50 [121.50, 147.50] | 140.00 [129.00, 152.00] | <0.001 |
| Smoking status (%) | | | <0.001 |
| Non-smoker | 111512 (60.7) | 81469 (51.6) | |
| Former smoker | 59985 (32.7) | 64533 (40.9) | |
| Light smoker (<10 cigarettes/day) | 2879 (1.6) | 1807 (1.1) | |
| Moderate smoker (10-19 cigarettes/day) | 5557 (3.0) | 4651 (2.9) | |
| Heavy Smoker (>20 cigarettes/day) | 3718 (2.0) | 5404 (3.4) | |
| Townsend deprivation index (median [IQR]) | -2.37 [-3.75, -0.01] | -2.36 [-3.76, 0.12] | 0.003 |
| Family history of CHD (%) | 91654 (49.9) | 73856 (46.8) | <0.001 |
| Family history of diabetes (%) | 43140 (23.5) | 36094 (22.9) | <0.001 |
| **Prescription history** | | | |
| Statins (%) | 17638 (9.6) | 28171 (17.8) | <0.001 |
| Atypical antipsychotics (%) | 453 (0.2) | 474 (0.3) | 0.003 |
| Corticosteroids (%) | 2885 (1.6) | 2631 (1.7) | 0.028 |
| Erectile dysfunction (%) | 16 (0.0) | 1897 (1.2) | <0.001 |
| Treated hypertension (%) | 8833 (4.8) | 11070 (7.0) | <0.001 |
| NSAIDs (%) | 47160 (25.7) | 45788 (29.0) | <0.001 |
| Anticoagulants (%) | 1099 (0.6) | 2653 (1.7) | <0.001 |
| **Prevalent medical conditions** | | | |
| Atrial fibrillation (%) | 1310 (0.7) | 3388 (2.1) | <0.001 |
| Congestive cardiac failure (%) | 5 (0.0) | 12 (0.0) | 0.076 |
| Coronary heart disease (%) | 2640 (1.4) | 8743 (5.5) | <0.001 |
| Chronic kidney disease (stage 4 or 5) (%) | 172 (0.1) | 312 (0.2) | <0.001 |
| Cardiovascular disease (%) | 4124 (2.2) | 11737 (7.4) | <0.001 |
| Gestational diabetes (%) | 220 (0.1) | 0 (0.0) | <0.001 |
| Ischaemic stroke (%) | 373 (0.2) | 874 (0.6) | <0.001 |
| Manic depression/schizophrenia (%) | 397 (0.2) | 449 (0.3) | <0.001 |
| Migraine (%) | 2171 (1.2) | 652 (0.4) | <0.001 |
| Polycystic ovary syndrome (%) | 142 (0.1) | 0 (0.0) | <0.001 |
| Rheumatoid arthritis (%) | 1054 (0.6) | 457 (0.3) | <0.001 |
| Systemic Lupus Erythematosus (%) | 156 (0.1) | 31 (0.0) | <0.001 |
| Severe mental illness (%) | 677 (0.4) | 583 (0.4) | 0.997 |
| Valvular heart disease (%) | 636 (0.3) | 1042 (0.7) | <0.001 |
| Type 2 diabetes (%) | 3100 (1.7) | 5622 (3.6) | <0.001 |
| Type 1 diabetes (%) | 214 (0.1) | 271 (0.2) | <0.001 |
| Renal disease (%) | 226 (0.1) | 388 (0.2) | <0.001 |
| **Incident disease** | | | |
| CHD (%) | 4609 (2.5) | 9401 (6.0) | <0.001 |
| CVD (%) | 8626 (4.7) | 14763 (9.4) | <0.001 |
| Ischaemic stroke (%) | 1109 (0.6) | 1800 (1.1) | <0.001 |
| Type 2 diabetes (%) | 5051 (2.8) | 7548 (4.8) | <0.001 |

**Supplementary Table 4.4 The p-values and linkage disequilibrium (LD) cut-off values that yielded the polygenic risk scores (PRS) with the highest C-statistic for each incident outcome studied.** CHD = coronary heart disease; CVD = cardiovascular disease.

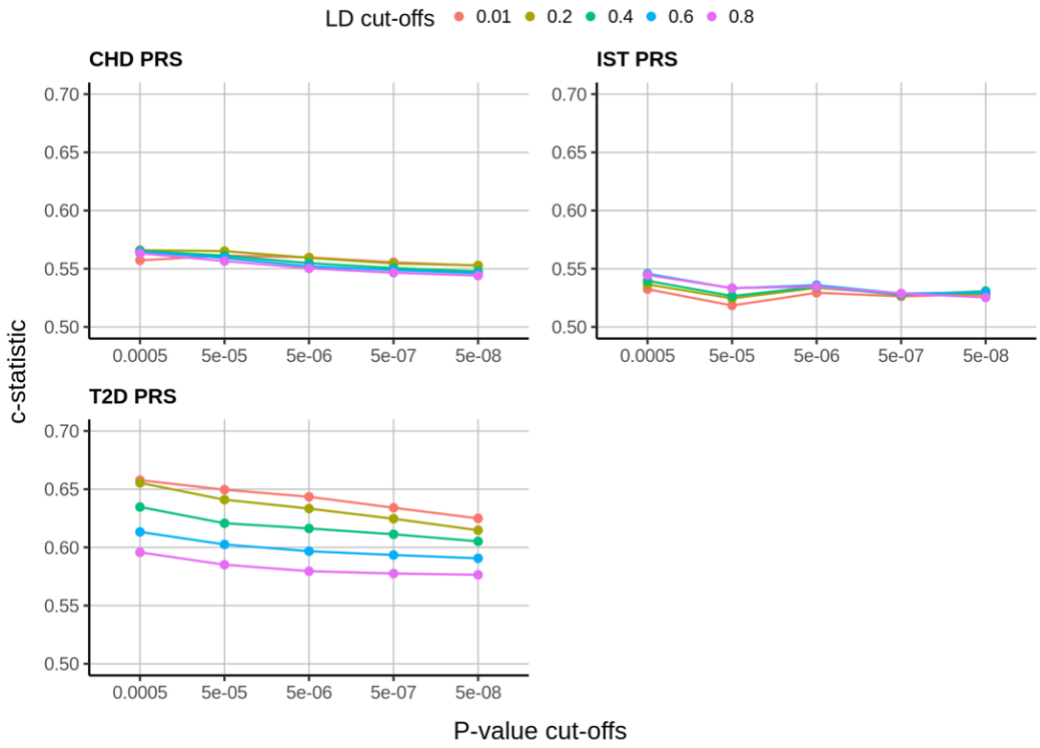
| PRS | Incident endpoint predicted | P-value cut-off | LD cut-off | C-statistic |
|---|---|---|---|---|
| CHD | CHD | $5x10^{-4}$ | 0.2 | 0.566 |
| Ischaemic stroke | Ischaemic stroke | $5x10^{-4}$ | 0.6 | 0.547 |
| Type 2 diabetes | Type 2 diabetes | $5x10^{-4}$ | 0.01 | 0.658 |
| CHD | CVD | $5x10^{-5}$ | 0.2 | 0.535 |
| Heart failure | CVD | $5x10^{-4}$ | 0.6 | 0.537 |
| All stroke | CVD | $5x10^{-8}$ | 0.6 | 0.522 |
| Atrial fibrillation | CVD | $5x10^{-4}$ | 0.2 | 0.571 |

**Supplementary Table 4.5 The C-statistic, calibration-in-the-large, calibration slope values and detection rate for a 5% false positive rate (DR5) of the PRS (CVD, CHD, IST, T2D), QScores (QRISK3, QStroke, QDiabetes) and combined models (age and sex; PRS, age and sex; PRS and QScore) for incident disease outcomes (CVD, CHD, IST and T2D).** The values obtained are from the independent test dataset. CI = confidence interval; CHD = coronary heart disease; CVD = cardiovascular disease; IST = ischaemic stroke; PRS = polygenic risk score; T2D = type 2 diabetes.

| Models | Outcome | Subgroup | C-statistic | C-statistic lower 95% CI | C-statistic upper 95% CI | Calibration-in-the-large | Calibration-in-the-large lower 95% CI | Calibration-in-the-large upper 95% CI | Calibration slope | Calibration slope lower 95% CI | Calibration slope upper 95% CI | DR5 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CVD PRS | Incident CVD | Overall | 0.586 | 0.58 | 0.591 | 0.001 | -0.018 | 0.02 | 0.997 | 0.937 | 1.057 | 0.091 |
| CVD PRS | Incident CVD | Female | 0.584 | 0.575 | 0.593 | -0.409 | -0.439 | -0.378 | 0.998 | 0.901 | 1.095 | 0.089 |
| CVD PRS | Incident CVD | Male | 0.588 | 0.581 | 0.595 | 0.345 | 0.321 | 0.369 | 1.007 | 0.93 | 1.084 | 0.092 |
| CHD PRS | Incident CHD | Overall | 0.569 | 0.562 | 0.576 | -0.002 | -0.026 | 0.022 | 1.083 | 0.977 | 1.189 | 0.081 |
| CHD PRS | Incident CHD | Female | 0.563 | 0.551 | 0.575 | -0.528 | -0.569 | -0.486 | 1.013 | 0.829 | 1.197 | 0.078 |
| CHD PRS | Incident CHD | Male | 0.573 | 0.564 | 0.581 | 0.398 | 0.369 | 0.428 | 1.133 | 1.003 | 1.262 | 0.083 |
| IST PRS | Incident IST | Overall | 0.55 | 0.535 | 0.565 | 0.009 | -0.043 | 0.06 | 1.101 | 0.77 | 1.433 | 0.071 |
| IST PRS | Incident IST | Female | 0.552 | 0.528 | 0.577 | -0.369 | -0.454 | -0.284 | 1.035 | 0.49 | 1.579 | 0.072 |
| IST PRS | Incident IST | Male | 0.548 | 0.53 | 0.567 | 0.321 | 0.256 | 0.386 | 1.134 | 0.716 | 1.552 | 0.07 |
| T2D PRS | Incident T2D | Overall | 0.659 | 0.652 | 0.666 | 0.006 | -0.02 | 0.031 | 1.031 | 0.986 | 1.077 | 0.143 |
| T2D PRS | Incident T2D | Female | 0.668 | 0.657 | 0.679 | -0.296 | -0.336 | -0.257 | 1.109 | 1.037 | 1.181 | 0.151 |
| T2D PRS | Incident T2D | Male | 0.653 | 0.644 | 0.662 | 0.276 | 0.243 | 0.309 | 0.98 | 0.921 | 1.039 | 0.138 |
| QRISK3 | Incident CVD | Overall | 0.715 | 0.711 | 0.719 | -0.002 | -0.021 | 0.017 | 1.008 | 0.98 | 1.037 | 0.2 |
| QRISK3 | Incident CVD | Female | 0.72 | 0.713 | 0.728 | -0.139 | -0.17 | -0.108 | 1.353 | 1.292 | 1.413 | 0.206 |
| QRISK3 | Incident CVD | Male | 0.667 | 0.661 | 0.673 | 0.093 | 0.069 | 0.118 | 0.774 | 0.738 | 0.81 | 0.15 |
| QRISK3 | Incident CHD | Overall | 0.738 | 0.733 | 0.744 | -0.009 | -0.033 | 0.016 | 1.032 | 1.001 | 1.063 | 0.229 |
| QRISK3 | Incident CHD | Female | 0.739 | 0.73 | 0.749 | -0.222 | -0.264 | -0.18 | 1.356 | 1.289 | 1.424 | 0.23 |
| QRISK3 | Incident CHD | Male | 0.684 | 0.677 | 0.691 | 0.119 | 0.089 | 0.149 | 0.802 | 0.764 | 0.841 | 0.167 |
| QStroke | Incident IST | Overall | 0.698 | 0.684 | 0.712 | 0 | -0.052 | 0.052 | 1.019 | 0.93 | 1.108 | 0.181 |
| QStroke | Incident IST | Female | 0.681 | 0.658 | 0.705 | -0.257 | -0.343 | -0.17 | 0.619 | 0.505 | 0.732 | 0.164 |
| QStroke | Incident IST | Male | 0.723 | 0.707 | 0.739 | 0.18 | 0.115 | 0.245 | 2.235 | 2.063 | 2.407 | 0.21 |
| QDiabetes | Incident T2D | Overall | 0.802 | 0.797 | 0.807 | 0.003 | -0.023 | 0.029 | 0.975 | 0.95 | 1.001 | 0.328 |
| QDiabetes | Incident T2D | Female | 0.818 | 0.811 | 0.826 | -0.047 | -0.087 | -0.007 | 1.292 | 1.243 | 1.341 | 0.359 |
| QDiabetes | Incident T2D | Male | 0.776 | 0.769 | 0.783 | 0.042 | 0.007 | 0.076 | 0.83 | 0.799 | 0.861 | 0.284 |
| Age x sex | Incident CVD | Overall | 0.702 | 0.697 | 0.706 | -0.002 | -0.021 | 0.017 | 1.037 | 1.008 | 1.065 | 0.185 |
| Age x sex | Incident CVD | Female | 0.695 | 0.688 | 0.703 | -0.022 | -0.053 | 0.009 | 1.595 | 1.522 | 1.667 | 0.178 |
| Age x sex | Incident CVD | Male | 0.659 | 0.653 | 0.665 | 0.011 | -0.014 | 0.035 | 0.868 | 0.829 | 0.907 | 0.143 |
| Age x sex | Incident CHD | Overall | 0.706 | 0.7 | 0.711 | -0.008 | -0.032 | 0.016 | 1.048 | 1.013 | 1.084 | 0.19 |
| Age x sex | Incident CHD | Female | 0.685 | 0.674 | 0.695 | -0.076 | -0.118 | -0.034 | 1.887 | 1.763 | 2.011 | 0.168 |
| Age x sex | Incident CHD | Male | 0.651 | 0.643 | 0.658 | 0.028 | -0.002 | 0.058 | 0.857 | 0.807 | 0.906 | 0.137 |
| Age x sex | Incident IST | Overall | 0.713 | 0.701 | 0.725 | 0.003 | -0.049 | 0.055 | 1.109 | 1.032 | 1.185 | 0.198 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Age x sex | Incident IST | Female | 0.722 | 0.701 | 0.742 | -0.041 | -0.126 | 0.044 | 1.702 | 1.517 | 1.886 | 0.208 |
| Age x sex | Incident IST | Male | 0.674 | 0.657 | 0.691 | 0.03 | -0.035 | 0.095 | 0.931 | 0.83 | 1.032 | 0.157 |
| Age x sex | Incident T2D | Overall | 0.636 | 0.629 | 0.642 | 0.002 | -0.023 | 0.027 | 1.056 | 0.999 | 1.113 | 0.124 |
| Age x sex | Incident T2D | Female | 0.616 | 0.606 | 0.627 | 0.001 | -0.039 | 0.04 | 1.493 | 1.345 | 1.64 | 0.11 |
| Age x sex | Incident T2D | Male | 0.602 | 0.594 | 0.611 | 0.003 | -0.03 | 0.036 | 0.952 | 0.864 | 1.041 | 0.1 |
| CVD PRS x age x sex | Incident CVD | Overall | 0.719 | 0.714 | 0.723 | -0.005 | -0.024 | 0.014 | 1.021 | 0.995 | 1.048 | 0.205 |
| CVD PRS x age x sex | Incident CVD | Female | 0.712 | 0.704 | 0.719 | -0.077 | -0.108 | -0.046 | 1.336 | 1.278 | 1.394 | 0.197 |
| CVD PRS x age x sex | Incident CVD | Male | 0.683 | 0.677 | 0.689 | 0.043 | 0.018 | 0.068 | 0.855 | 0.821 | 0.889 | 0.166 |
| CHD PRS x age x sex | Incident CHD | Overall | 0.717 | 0.711 | 0.723 | -0.01 | -0.035 | 0.014 | 1.031 | 0.998 | 1.064 | 0.202 |
| CHD PRS x age x sex | Incident CHD | Female | 0.694 | 0.683 | 0.704 | -0.093 | -0.135 | -0.051 | 2.005 | 1.882 | 2.127 | 0.177 |
| CHD PRS x age x sex | Incident CHD | Male | 0.668 | 0.661 | 0.675 | 0.034 | 0.005 | 0.064 | 0.841 | 0.797 | 0.884 | 0.151 |
| IST PRS x age x sex | Incident IST | Overall | 0.717 | 0.704 | 0.729 | 0.002 | -0.049 | 0.054 | 1.101 | 1.028 | 1.173 | 0.202 |
| IST PRS x age x sex | Incident IST | Female | 0.724 | 0.704 | 0.745 | -0.063 | -0.148 | 0.022 | 1.384 | 1.242 | 1.526 | 0.211 |
| IST PRS x age x sex | Incident IST | Male | 0.68 | 0.663 | 0.697 | 0.043 | -0.022 | 0.109 | 0.953 | 0.855 | 1.05 | 0.163 |
| T2D PRS x age x sex | Incident T2D | Overall | 0.707 | 0.701 | 0.713 | 0.005 | -0.02 | 0.031 | 1.062 | 1.025 | 1.098 | 0.191 |
| T2D PRS x age x sex | Incident T2D | Female | 0.703 | 0.693 | 0.713 | -0.077 | -0.116 | -0.037 | 1.306 | 1.235 | 1.376 | 0.186 |
| T2D PRS x age x sex | Incident T2D | Male | 0.684 | 0.676 | 0.693 | 0.066 | 0.033 | 0.1 | 0.916 | 0.87 | 0.962 | 0.167 |
| CVD PRS x QRISK3 | Incident CVD | Overall | 0.72 | 0.715 | 0.724 | -0.002 | -0.021 | 0.017 | 0.994 | 0.96 | 1.027 | 0.206 |
| CVD PRS x QRISK3 | Incident CVD | Female | 0.714 | 0.706 | 0.721 | -0.256 | -0.287 | -0.226 | 1.414 | 1.333 | 1.496 | 0.199 |
| CVD PRS x QRISK3 | Incident CVD | Male | 0.682 | 0.676 | 0.688 | 0.192 | 0.167 | 0.216 | 0.732 | 0.694 | 0.771 | 0.165 |
| CHD PRS x QRISK3 | Incident CHD | Overall | 0.74 | 0.735 | 0.746 | -0.007 | -0.032 | 0.017 | 1.048 | 1.005 | 1.092 | 0.231 |
| CHD PRS x QRISK3 | Incident CHD | Female | 0.73 | 0.72 | 0.74 | -0.4 | -0.442 | -0.358 | 1.591 | 1.47 | 1.712 | 0.218 |
| CHD PRS x QRISK3 | Incident CHD | Male | 0.694 | 0.687 | 0.701 | 0.265 | 0.235 | 0.295 | 0.763 | 0.716 | 0.811 | 0.177 |
| IST PRS x QStroke | Incident IST | Overall | 0.69 | 0.675 | 0.704 | 0.007 | -0.044 | 0.059 | 1.052 | 0.839 | 1.265 | 0.173 |
| IST PRS x QStroke | Incident IST | Female | 0.597 | 0.571 | 0.622 | -0.371 | -0.456 | -0.286 | 0.96 | 0.7 | 1.219 | 0.097 |
| IST PRS x QStroke | Incident IST | Male | 0.714 | 0.698 | 0.731 | 0.32 | 0.255 | 0.385 | 20.838 | 18.598 | 23.079 | 0.199 |
| T2D PRS x QDiabetes | Incident T2D | Overall | 0.8 | 0.795 | 0.806 | 0.003 | -0.023 | 0.029 | 0.939 | 0.905 | 0.972 | 0.325 |
| T2D PRS x QDiabetes | Incident T2D | Female | 0.805 | 0.796 | 0.813 | -0.176 | -0.215 | -0.136 | 1.537 | 1.454 | 1.62 | 0.334 |
| T2D PRS x QDiabetes | Incident T2D | Male | 0.784 | 0.777 | 0.791 | 0.156 | 0.122 | 0.19 | 0.758 | 0.721 | 0.794 | 0.297 |

**Supplementary Figure 4.1 The polygenic risk score (PRS) parameters tested (p-value and linkage disequilibrium (LD) cut-off values) and their C-statistic for incident coronary heart disease (CHD), incident ischaemic stroke (IST), and incident type 2 diabetes (T2D).**



# 5 Modelling a two-stage adult population screen for autosomal dominant familial hypercholesterolaemia (FH)

A manuscript for the following chapter is underway.

## 5.1 Abstract

**Background**: Autosomal dominant familial hypercholesterolaemia (FH) is highly underdiagnosed worldwide and increased detection of FH cases has been identified as one of the goals of the NHS' Long Term Plan. There is currently no population screening strategy in place for FH in the UK. Instead, new cases are identified opportunistically. Increasing case detection could require a population screening strategy with the potential to reduce and prevent premature coronary heart disease and death. The aim of the work in this chapter was to evaluate the performance of a two-stage adult population screen for FH and compare it to child-parent screening.

**Methods**: I modelled use of different low-density lipoprotein cholesterol (LDL-C) cut-offs (stage 1) to select individuals for DNA sequencing to identify FH-causing variants in *LDLR, APOB, PCSK9* and *APOE* (stage 2) in 140,439 unrelated participants of European ancestry from the UK Biobank with information on circulating LDL-C concentration and exome sequencing data. For different LDL-C cut-offs, I estimated the stage 1 detection and false positive rate, the proportion of individuals requiring DNA sequencing (stage 1 screen positive rate), and the number of FH cases identified by population screening. I also modelled the number of additional FH cases that might be detected by cascade testing of first-degree relatives of index cases and compared this approach with child-parent screening for FH.

**Results**: I identified 488 individuals with an FH-causing variant and 139,951 without (prevalence: 1 in 288). An LDL-C cut-off of >4.8 mmol/L had a stage 1 detection rate (DR; sensitivity) of 40% (95% CI: 36-44%) for a false positive rate of 10% (95% CI: 10-11%). Using this LDL-C cut-off to screen 100,000 individuals (among whom there would be an estimated 347 FH cases) would generate 10,398 stage 1 screen positives for sequencing, detect 138 FH cases, miss 209, with a further 207 cases potentially being detected through two-generation cascade testing of first-degree relatives of index cases. This is about a third as many FH cases as childhood screening with three generation cascade testing, for twice the sequencing burden. Detecting 25% of all FH cases (~49,000 additional cases; the target set in the NHS Long Term Plan) with the two-stage adult

screening strategy would require screening around 14 million adults and sequencing of 1.5 million of them, or screening 4.6 million children and sequencing ¼ million of them with cascade testing.

**Conclusion**: Two-stage adult population screening for FH could help achieve the FH case detection target in the NHS Long Term Plan, but less efficiently than childhood screening and with a greater sequencing requirement.

## 5.2 Introduction

Autosomal dominant familial hypercholesterolaemia (FH) is the commonest monogenic disorder with a prevalence of about 1 in 250.[1–3] People with FH have a heterozygous DNA variant in either the *LDLR*, *APOB, PCSK9* or *APOE* genes,[4,5] leading to at least a 4.8-fold age-adjusted risk of coronary artery disease compared to the general population.[6] If detected, affected individuals can benefit from drugs that lower low-density lipoprotein cholesterol (LDL-C) to reduce the risk of a coronary event.[7,8]

First-degree relatives of index FH cases have a one in two chance of carrying the same causative genetic variant, enabling further case detection through each index case by family-based cascade testing.[9,10] However, cascade testing is limited by the stream of index cases identified, and this is currently opportunistic rather than systematic. FH cases are currently detected when they present with coronary disease at a young age or are found by chance to have an extreme elevation of LDL-C concentration when assessed as part of a healthcare contact for another reason. As a consequence, only 19,000 (7%) of an estimated 270,000 FH cases in the UK are known.[11,12]

Systematic identification of individuals with FH in the population would address underdiagnosis and the missed opportunity for coronary disease prevention. The NHS Long Term plan has set a target of detecting at least 25% of FH cases (~67,500) over the next 5 years,[13] but does not elaborate on how this will be achieved. Measurement of circulating LDL-C concentration in adults performs poorly as a screening test in distinguishing people with FH from those with other causes of a high LDL-C (e.g., due to diet, lifestyle, or carriage of a high burden of common genetic variants that affect LDL-C concentration).[14,15]

Although individuals with elevated LDL-C regardless of cause can benefit from LDL-C lowering, cascade testing is only relevant in the families of those with autosomal dominant FH.[16]

Moreover, adults with autosomal dominant FH may have a higher risk of coronary disease than people who have a similar LDL-C but without a causative genetic variant.[17] LDL-C concentration in childhood identifies individuals with FH more accurately than measurement in adults and underpins the concept of child-parent screening. In this approach, affected parents, older siblings and grandparents are identified through the screening of children by the age of two by measurement of LDL-C, followed by genetic testing of stored samples with an LDL-C beyond a pre-specified cut-off.[18,19] Despite demonstration of the feasibility and efficiency of this approach,[18] it did not receive the endorsement of the UK National Screening Committee when last reviewed.[20]

Different genes and DNA variants may cause FH in different families. Thus, sequencing of the four relevant FH-causing genes is needed to identify the causative variant in an index case after which simpler, cheaper mutation detection methods can be employed for cascade testing of family members. Although DNA sequencing is more accurate than biochemical screening, and could be used at any age, it is currently too expensive to be considered as the primary screening method for FH.

An alternative approach that minimises sequencing burden while avoiding concern about FH screening in childhood is a two-stage screening design in adults. In this approach, LDL-C (an inexpensive test with a high false positive rate in adults) is measured at stage 1, followed by targeted sequencing of FH genes (a more expensive test with a low false positive rate) at stage 2 for individuals whose LDL-C concentration exceeds a specified cut-off. However, the effectiveness of this approach, evaluated on the basis of the number of index FH cases detected and missed, the additional cases detected by cascade testing, and the sequencing burden, has not been evaluated nor compared with child-parent screening.

Participants in UK Biobank, a national, population-based cohort study, have already had LDL-C measurement and exome sequencing, which offers an opportunity to model the performance of two-stage adult population screening for FH. The age range of UK Biobank participants at recruitment also overlaps with that of individuals who, until the COVID-19 pandemic, were invited to NHS Health Checks in England.[21] NHS Health Checks evaluate a range of risk factors and blood is routinely drawn for the measurement of circulating lipid concentration. Since genomic sequencing could subsequently be undertaken from a stored blood sample drawn at the time of

assessment in those with an LDL-C above a pre-specified cut-off, the NHS Health Check could serve as a setting for an adult FH screening programme.

Here, I model the performance of two-stage adult population screening to identify index FH cases and compare it with the previously reported performance of child-parent screening for FH, when both are followed by cascade testing of relatives of index cases.[10,18] I also estimate the time needed to reach the NHS Long Term Plan Goal.

## 5.3 Methods

### 5.3.1 Participants

UK Biobank recruited ~500,000 participants between 40 and 75 years of age between 2006 and 2010.[22] Participants completed questionnaires, undertook a variety of physical assessments, and provided biological samples for genotyping, sequencing, and other measurements.[23,24]

### 5.3.2 LDL-C measurement

In a total of 486,459 UK Biobank participants, serum was obtained from a blood draw in the non-fasting state at the time of recruitment and stored at -80ºC and liquid nitrogen for later analysis. LDL-C was measured directly by enzymatic protective selection analysis with a Beckman Coulter AU5800 and the values were recorded in mmol/L.[25] Where an LDL-C measurement was missing for an included participant, I imputed it using single imputation with the R package MICE version 3.10.0.[26] Where an included participant was already recorded as receiving a statin, I adjusted their LDL-C concentration using the correction coefficient 1.43.[27]

### 5.3.3 Identification of carriers of FH-causing genetic variants

A blood sample was drawn for DNA analysis in 454,787 participants of UK Biobank and stored at -80ºC.[28] Exome capture was done using the IDT xGen Exome Research Panel v1.0 which included additional probes, and exome-sequencing was performed on the Illumina NovaSeq 6000 platform.[29]

I identified 140,439 European ancestry participants from data-field 22006 of the UK Biobank with exome sequence data and valid (directly typed or imputed) measurements of LDL-C at the time of analysis and included them in the study. Each participant was assigned to one of three groups: (1) individuals with an established FH causing variant in *LDLR*, *PCSK9*, *APOB* or *APOE* genes based on annotation in ClinVar and manual curation (**Supplementary Table 5.1**); (2) individuals with variants of unknown significance (VUS) in these four genes (**Supplementary Table 5.2**); and (3) individuals with no FH-causing variant or VUS. I classified individuals from the first group as "affected" and those from the other two groups as "unaffected".

The *LDLR*, *PCSK9*, *APOB* or *APOE* gene regions were extracted from the UK Biobank exome data (**Table 5.1**). Multiallelic sites were converted to biallelic sites using BCFtools version 1.11.,[30] and genetic variants were annotated using Ensembl's Variant Effect Predictor (VEP) release 103.1.[31] The *LDLR* variants in the canonical transcript ENST00000558518 were filtered for a minor allele frequency (MAF) of 0.0006, which is equal to the frequency of the most common FH variant (p.Arg3527Gln in *APOB*) according to the gnomAD database version 2.1.1.[32] Further variant filtering steps included a minimum read depth of 10 and genotype quality of 20. The SAMtools plugin split-vep was used to keep variants that had a predicted consequence of missense or worse, and the resulting variants with a SIFT annotation of "tolerated" or a PolyPhen annotation of "benign" were excluded.[30,33,34] These filtering steps were followed by manual curation of the variants by two expert reviewers who respected the Association for Clinical Genomics Science (ACGS) guidelines and the evidence accrued from the LOVD database for *LDLR*.[35,36] For the *APOE* gene, the heterozygous p.Leu167del in-frame deletion was considered to be FH-causing, and the pathogenic variants in *PCSK9* and *APOB* were filtered based on a pre-defined list of curated variants with functional assay backing.[5,37]

| Gene name | Chromosome number | Start coordinate | End coordinate |
|---|---|---|---|
| *LDLR* | 19 | 11,089,262 | 11,133,820 |
| *APOB* | 2 | 21,001,429 | 21,044,073 |
| *APOE* | 19 | 44,905,791 | 44,909,393 |
| *PCSK9* | 1 | 55,039,347 | 55,064,852 |

**Table 5.1 Genetic coordinates of FH-causing genes.** Genetic coordinates are mapped to GRCh38.

### 5.3.4 Evaluation of two-stage adult FH screening performance

I counted the number of individuals with an FH-causing variant above and below different LDL-C cut-off values and used this to estimate the stage 1 detection rate (the proportion of eligible participants with an FH-causing variant whose LDL-C value exceeded the cut-off), the stage 1 false positive rate (the proportion of eligible participants with no FH-causing variant whose LDL-C value exceeded the cut-off), the odds of being affected give a positive result (the ratio of true to false positives), and the stage 1 screen positive rate (the proportion of individuals whose LDL-C exceeded the cut-off regardless of FH-causing variant status). I assumed that all individuals with an LDL-C value above the cut-off would undergo targeted sequencing (stage 2). I assumed that targeted sequencing has a 100% detection rate for individuals with FH-causing variants, and that individuals with a VUS identified on sequencing would not be taken forward into the cascade testing phase.

### 5.3.5 Comparison of two-stage adult and child-parent screening

I used data from the current analysis and from previous reports to compare the performance of two-stage adult and child-parent screening in terms of samples requiring sequencing, index FH cases detected, and additional cases detected through cascade testing.[9,10,18,38] For the comparison, I assumed that samples from children or adults with an LDL-C concentration beyond a pre-specified cut-off would undergo targeted sequencing of FH-causing variants.

### 5.3.6 Modelling cascade testing in families of index cases

Using the methods and assumptions described by Morris *et al.* and Wald *et al.*,[9,18] I estimated the number of additional FH cases identified by cascade testing in the families of each index case identified by the two population screening strategies. I assumed cascade testing of first-degree relatives only: since each index case in the two-stage adult screen would be between 40 and 75 years of age, I assumed their parents would either have died or be too old for screening. I also assumed that families comprise of two children on average (as estimated by Office of National Statistics) such that each index case would have an average of one sibling and two offspring.[39]For the child-parent screen, I assumed that this strategy would enable screening of three generations (the sibs, parents, and grandparents of the index child), as opposed to two generations (the sibs

and children of the index adult) in the two-stage adult screen. Calculations were based on a best-case scenario for cascade testing where all affected first-degree relatives would be detected.

### 5.3.7 Achieving the NHS Long Term Plan target for FH case detection

To estimate the amount of time it would take for the child-parent and the adult two-stage screen to reach the NHS Long Term Plan goal of 25% FH case detection, I assumed that current FH case detection in the UK is at 7%, and that an additional 49,000 FH cases would need to be diagnosed to reach the 25% target set out by the NHS.[11–13] For the child-parent screen, I assumed that 95% of children would be immunised and eligible for FH screening.[40,41] For the two-stage adult screen, I assumed that around one million adults attend their NHS Health Check.[42] I then calculated the number of FH cases that would be detected per year if implementing the child-parent or the adult-two stage screen followed by cascade testing, based on the previous modelling. I extrapolated these figures to estimate the number of years it would take to detect the additional 49,000 FH case target set out by the NHS. These calculations do not account for related individuals being picked up from the population screen who were previously detected via cascade testing (or *vice versa*).

## 5.4 Results

### 5.4.1 Demographic and other characteristics of study participants

I identified 140,439 White British participants from UK Biobank with a valid (directly measured or imputed) LDL-C value and exome sequence data available at the time of analysis. I compared the UK Biobank participant characteristics with individuals who participated in the NHS Health Check in 2017-2018, where FH screening could be rolled out. The UK Biobank participants in the current analysis were slightly older than individuals evaluated in the NHS Health Check but had a similar gender distribution (**Table 5.2**). About 15.8% of those undergoing NHS Health Checks self-report as non-White, whereas the dataset I analysed from UK Biobank was limited to those of who self-reported as being White (**Table 5.2**).[43]

| | NHS Health Check (2017-2018) | UK Biobank study cohort |
|---|---|---|
| Count of participants | 1,108,841 | 140,439 |
| Sex (male) | 509,752 (46.0%) | 63878 (45.5%) |
| Age | | |
| 39 | 0 (0%) | 2 (0.001%) |
| 40 to 44 | 240,438 (21.7%) | 13,338 (9.5%) |
| 45 to 49 | 205,722 (18.6%) | 17,577 (12.5%) |
| 50 to 54 | 209,088 (18.9%) | 21,066 (15.0%) |
| 55 to 59 | 180,624 (16.3%) | 25,385 (18.1%) |
| 60 to 64 | 147,444 (13.3%) | 35,473 (25.3%) |
| 65 to 69 | 125,525 (11.3%) | 26,938 (19.2%) |
| 70 to 74 | 0 (0%) | 660 (0.5%) |
| Self-reported ethnicity | | |
| Any other ethnic group | 17,531 (1.6%) | 0 (0%) |
| Asian or Asian British | 98,692 (8.9%) | 0 (0%) |
| Black or African or Caribbean or Black British | 45,674 (4.1%) | 0 (0%) |
| Mixed or multiple ethnic groups | 13,498 (1.2%) | 0 (0%) |
| White | 864,173 (77.9%) | 140,439 (100%) |
| Ethnicity not stated or recorded | 69,273 (6.2%) | 0 (0%) |

**Table 5.2 Participant characteristic comparison between UK Biobank participants of the study cohort and the NHS Health Check 2017-2018.[43]**

### 5.4.2 Participants with an FH-causing variant or variant of unknown significance

Of the 140,439 participants studied (median age 58 years, 45% male), I identified 488 with an FH-causing variant interpreted as "pathogenic" or "likely pathogenic" (**Table 5.3** and **Supplementary Table 5.1**) as per ACGS guidelines,[44] giving a prevalence of 1 in 288. A further 660 (1 in 213) individuals were found to carry a VUS (**Supplementary Table 5.2**). In those with an FH-causing variant, the variant was in *LDLR* in 374 (1 in 376), in *APOB* in 101 (1 in 1,390), and in *APOE* in 13 participants (1 in 10,803) (**Table 5.3** and **Supplementary Table 5.1**). None of those analysed carried an FH-causing variant in *PCSK9*.

| | No FH-causing variant | FH causing variant | P-value of group differences |
|---|---|---|---|
| n | 139291 | 488 | |
| *LDLR* variant (%) | 0 (0.0) | 374 (76.6) | |
| *APOB* variant (%) | 0 (0.0) | 101 (20.7) | |
| *APOE* variant (%) | 0 (0.0) | 13 (2.7) | |
| Age (median [IQR]) | 58 [51, 63] | 58 [51, 63] | 0.803 |
| Sex (male) (%) | 63382 (45.5) | 207 (42.4) | 0.187 |
| BMI, kg/m2 (median [IQR]) | 26.7 [24.1, 29.8] | 27.1 [23.9, 29.8] | 0.689 |
| Townsend deprivation index (median [IQR]) | -2.4 [-3.8, 0.0] | -2.2 [-3.7, 0.2] | 0.346 |
| Smoking status (%) | | | 0.827 |
| Non-smoker | 79618 (57.2) | 281 (57.6) | |
| Former smoker | 51177 (36.7) | 173 (35.5) | |
| Light smoker (<10 cigarettes/day) | 2021 (1.5) | 7 (1.4) | |
| Moderate smoker (10-19 cigarettes/day) | 3497 (2.5) | 13 (2.7) | |
| Heavy Smoker (>20 cigarettes/day) | 2978 (2.1) | 14 (2.9) | |
| Statin use (%) | 18139 (13.0) | 165 (33.8) | <0.001 |
| Family history of CHD (%) | 67013 (48.1) | 306 (62.7) | <0.001 |
| **Blood biomarkers** | | | |
| LDL-C (unadjusted), mmol/L (median [IQR]) | 3.5 [3.0, 4.1] | 3.9 [3.2, 4.8] | <0.001 |
| LDL-C (adjusted for statin users), mmol/L (median [IQR]) | 3.7 [3.1, 4.2] | 4.4 [3.7, 5.4] | <0.001 |
| Total cholesterol, mmol/L (median [IQR]) | 5.7 [4.9, 6.4] | 6.1 [5.2, 7.3] | <0.001 |
| Triglycerides, mmol/L (median [IQR]) | 1.5 [1.1, 2.2] | 1.3 [0.9, 1.9] | <0.001 |
| HDL-C, mmol/L (median [IQR]) | 1.4 [1.2, 1.7] | 1.4 [1.2, 1.6] | 0.086 |
| **Disease incidence & prevalence** | | | |
| CHD prevalence (%) | 3890 (2.8) | 40 (8.2) | <0.001 |
| CHD incidence (%) | 5370 (3.9) | 32 (6.6) | 0.003 |
| CVD prevalence (%) | 5686 (4.1) | 45 (9.2) | <0.001 |
| CVD incidence (%) | 9038 (6.5) | 46 (9.4) | 0.011 |
| Type 2 diabetes prevalence (%) | 3593 (2.6) | 11 (2.3) | 0.757 |
| Type 2 diabetes incidence (%) | 4948 (3.6) | 19 (3.9) | 0.776 |

**Table 5.3 Characteristics of the study participants.** P-values were obtained following Kruskal-Wallis Rank sum nonparametric testing. BMI = body mass index; CHD = coronary heart disease; CVD = cardiovascular disease (defined as CHD, ischaemic and haemorrhagic stroke, heart failure, and atrial fibrillation); FH = familial hypercholesterolaemia; HDL-C = high-density lipoprotein cholesterol; IQR = interquartile range; LDL-C = low-density lipoprotein cholesterol.

LDL-C concentration was higher in those with an FH-causing variant (median 4.43 mmol/L, IQR [3.67; 5.43]) than those without (median 3.67 mmol/L, IQR [3.14; 4.24]) (**Figure 5.1** and **Table 5.3**). LDL-C concentration was on average higher among those with an FH-causing variant in the *APOB* gene than among those with mutations in the other FH-causing genes (**Supplementary Table 5.3**).

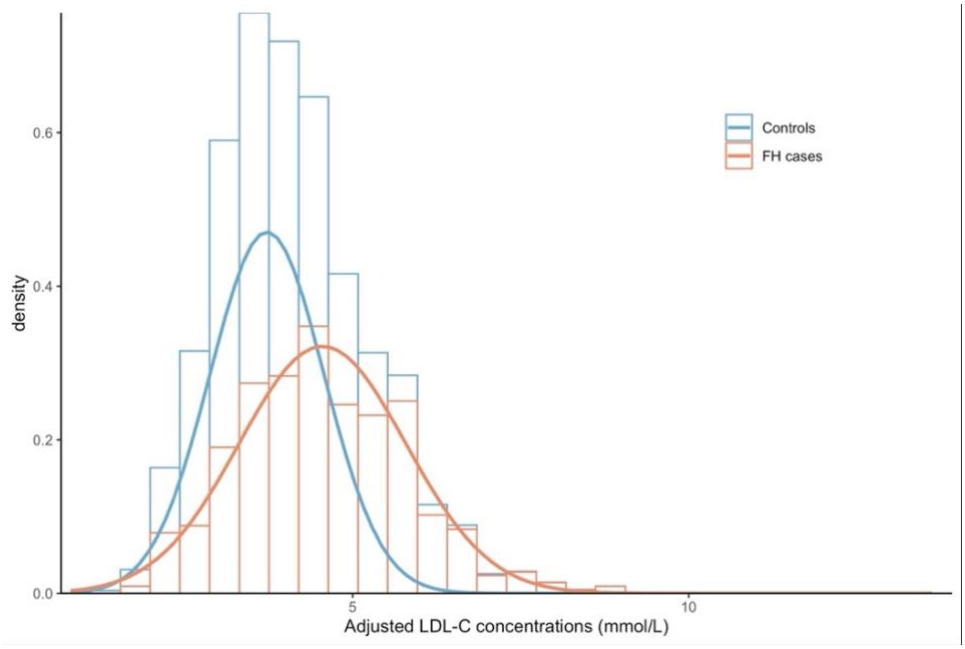



**Figure 5.1 Relative frequency distributions of the adjusted LDL-C concentrations in monogenic FH variant carriers and non-carriers of the study cohort.** Unaffected individuals are shown in blue and affected individuals in red.

There was no significant difference between those with and without FH-causing variants in age, sex, body mass index (BMI), Townsend deprivation index, smoking status, high-density lipoprotein cholesterol (HDL-C) and lipoprotein(a) concentration (**Table 5.3**). Of those with FH-causing variants, 34% were prescribed statins, compared to 13% of those without (p-value <0.001) (**Table 5.3**). There was a higher proportion of people with a family history of coronary heart disease (CHD) in those with FH-causing variants than those without (63% versus 48%; p-value <0.001), as well as more than double the prevalence of both CHD (8% versus 3%; p-value <0.001) and cardiovascular disease (CVD) (a composite of CHD, ischaemic and haemorrhagic stroke, heart failure, and atrial fibrillation) (9% versus 4%; p-value <0.001) (**Table 5.3**). The 10-year CHD (7% vs 4%) and CVD incidence (9% vs 7%) were also significantly higher in the FH-causing variant positive group (p-value <0.05) (**Table 5.3**).

### 5.4.3 Performance of two-stage adult screen for autosomal dominant FH

For different LDL-C cut-offs between 3 and 8.5 mmol/L, I estimated the detection and false positive rate of stage 1 screening, the proportion of samples eligible for sequencing (stage 1 screen

positive rate), and the number and proportion of FH cases identified by the two-stage population screen (**Supplementary Table 5.4**). The lower the LDL-C cut-off, the higher the detection rate but also the false-positive rate and therefore the number of stage 1 screen positive samples that would be submitted for sequencing (**Figure 5.2** and **Table 5.4**). For example, using an LDL-C cut-off of 5.0 mmol/L (1.36 multiples of the median (MoM) LDL-C value), gave a detection rate of 35% (95% CI: 31-39%) for a 7% (95% CI: 7%-7.3%) false-positive rate at stage 1, with a screen positive rate of 7% (**Table 5.4**). An LDL-C cut-off of 4.0 mmol/L (1.09 MoM) gave a stage 1 detection rate of 65% (95% CI: 60%-69%) for a 35% (95% CI: 35-35%) false-positive rate, with a stage 1 screen positive rate of 35%.

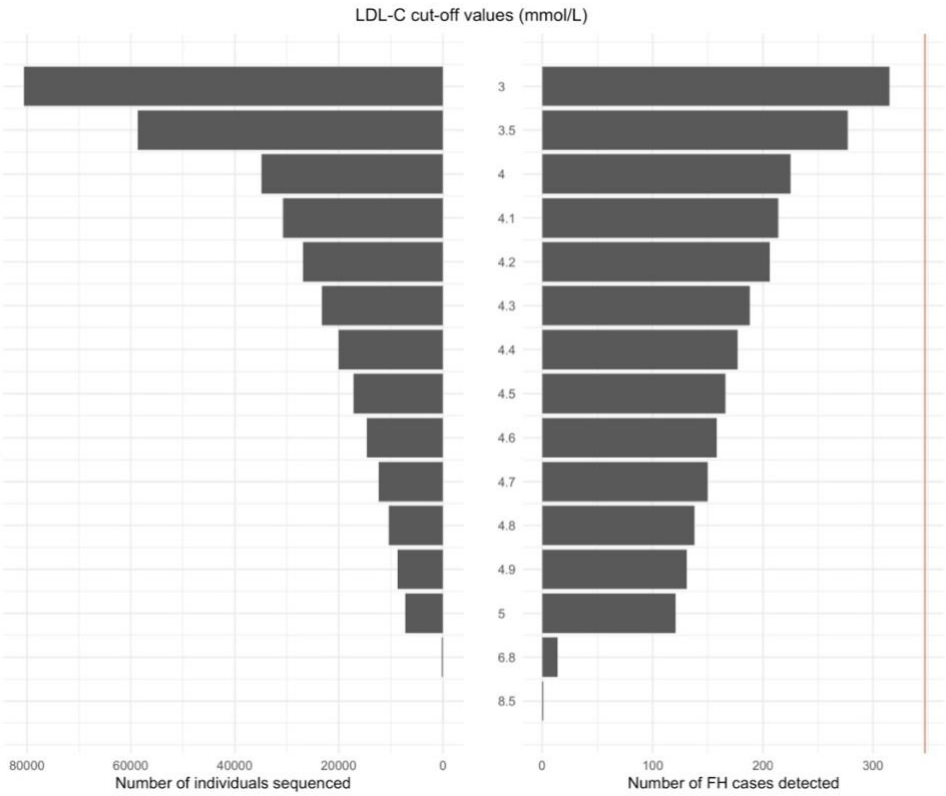



**Figure 5.2 The number of samples sequenced, and the number of FH cases detected using various LDL-C cut-off values in the two-stage adult screening population strategy for FH for a hypothetical population of 100,000 individuals.** The orange vertical line represents the total number of 347 FH cases in the hypothetical sample population of 100,000 individuals (for an FH prevalence of 1:288).

| LDL-C cut-off (mmol/L) | MoM | Stage 1 (LDL-C) | | | | | | Stage 2 (Sequencing) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Detection rate (%) | False positive rate (%) | OAPR | Cases detected | Cases missed | False positives | Number sequenced | Cases confirmed | VUS |
| **3** | 0.82 | 91 (88-93) | 81 (80-81) | 1:255 | 315 | 32 | 80258 | 80573 | 315 | 417 |
| **3.5** | 0.95 | 80 (76-83) | 59 (58-59) | 1:211 | 277 | 70 | 58346 | 58623 | 277 | 331 |
| **4** | 1.09 | 65 (60-69) | 35 (35-35) | 1:154 | 225 | 122 | 34678 | 34903 | 225 | 218 |
| **4.1** | 1.12 | 62 (57-66) | 31 (30-31) | 1:143 | 214 | 133 | 30556 | 30770 | 214 | 196 |
| **4.2** | 1.14 | 59 (55-64) | 27 (27-27) | 1:129 | 206 | 141 | 26628 | 26834 | 206 | 177 |
| **4.3** | 1.17 | 54 (50-59) | 23 (23-23) | 1:123 | 188 | 159 | 23075 | 23263 | 188 | 159 |
| **4.4** | 1.2 | 51 (46-55) | 20 (20-20) | 1:112 | 177 | 170 | 19835 | 20012 | 177 | 139 |
| **4.5** | 1.22 | 48 (43-52) | 17 (17-17) | 1:102 | 166 | 181 | 16997 | 17163 | 166 | 124 |
| **4.6** | 1.25 | 46 (41-50) | 15 (14-15) | 1:92 | 158 | 189 | 14470 | 14628 | 158 | 108 |
| **4.7** | 1.28 | 43 (39-48) | 12 (12-12) | 1:81 | 150 | 197 | 12195 | 12345 | 150 | 92 |
| **4.8** | 1.31 | 40 (36-44) | 10 (10-11) | 1:74 | 138 | 209 | 10260 | 10398 | 138 | 79 |
| **4.9** | 1.33 | 38 (34-42) | 9 (9-9) | 1:65 | 131 | 216 | 8572 | 8703 | 131 | 70 |
| **5** | 1.36 | 35 (31-39) | 7 (7-7) | 1:59 | 121 | 226 | 7116 | 7237 | 121 | 59 |
| **6.8** | 1.85 | 4 (3-6) | 0.2 (0.2-0.2) | 1:15 | 14 | 333 | 211 | 225 | 14 | 1 |
| **8.5** | 2.31 | 0.4 (0.1-2) | 0 (0-0) | 1:15 | 1 | 346 | 15 | 16 | 1 | 0 |

**Table 5.4 Performance of a two-stage adult population screen for monogenic FH using different stage 1 LDL-C cut-offs.** Reported counts are based on a screened population of 100,000 adults with 347 monogenic FH cases and 470 individuals with a VUS. FH = familial hypercholesterolaemia; MoM = multiple of the median; OAPR = odds of being affected given a positive test result; VUS = variant of uncertain significance.

**Figure 5.3** illustrates the performance of the two-stage screen scaled to a cohort of 100,000 people, using a stage 1 LDL-C cut-off of 4.8 mmol/L (1.32 MoM). The 10,398 stage 1 screen positive individuals include 138 (40%) of the 347 FH cases, all of whom would be identified by sequencing at stage 2. Individuals who are screen positive from stage 1 would also include 10,181 individuals with no FH-causing variant or VUS, as well as 79 individuals with no FH-causing variant but with a VUS; the two groups together giving a stage 1 false positive rate of 10% (95% CI: 10%-11%). All these individuals would be correctly classified by DNA sequencing at stage 2, giving a stage 2 false positive rate of 0%, with a VUS rate of 0.8%. **Table 5.4** documents the corresponding values for a range of LDL-C cut-offs. Where cascade screening is seeded by the two-stage adult screen with a stage 1 LDL-C cut-off of 4.8 mmol/L (1.31 MoM), based on the previous assumptions about family size, the 138 index cases identified for every 100,000 screened would lead to an additional 207 FH cases being identified through cascade testing, giving 345 cases in all (**Figure 5.3**). Overall, one FH case would be identified for every 290 individuals recruited to the two-stage population screen using the LDL-C cut-off of 4.8 mmol/L.

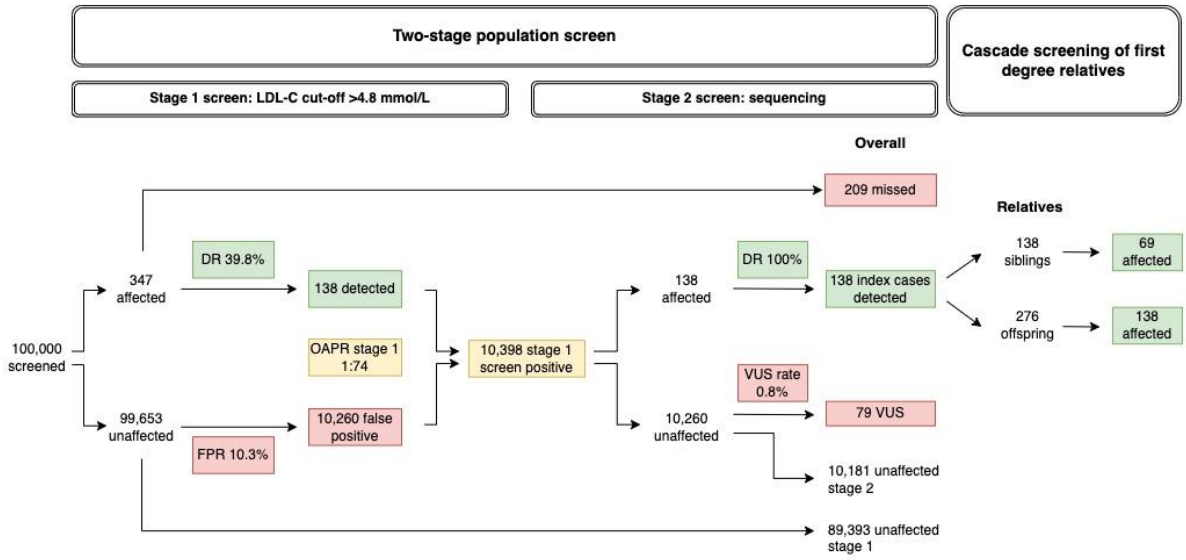



**Figure 5.3 Illustration of the two-stage adult screen and subsequent cascade screening of first-degree relatives of index FH cases scaled to 100,000 individuals using an LDL-C cut-off value of 4.8 mmol/L in the first stage screen.** DR = detection rate (sensitivity); FPR = false positive rate (1-specificity); OAPR = odds of being affected given a positive result; VUS = variant of uncertain significance.

### 5.4.4 Comparison of two-stage adult with child-parent screening

| | Child-parent | Adult two-stage |
|---|---|---|
| **Population screen** | | |
| Target population | Children | Adults |
| Number screened | 100,000 | 100,000 |
| Estimated FH population prevalence | 1:288 | 1:288 |
| Estimated FH cases | 347 | 347 |
| Test | Total cholesterol | LDL-C |
| Test cut-off | ≥1.35 MoM | 1.31 MoM (>4.8 mmol/L) |
| Index FH cases missed | 44 | 209 |
| Index FH cases detected | 303 | 138 |
| False positives | 4,681 | 10,260 |
| Number eligible for sequencing | 4,984 | 10,398 |
| Index FH cases confirmed on sequencing | 303 | 138 |
| **Cascade testing** | | |
| Generations screened | 3 | 2 |
| 1st degree relatives of index cases detected[1] | 455 | 207 |
| 2nd degree relatives of index cases detected[1] | 303 | 0 |
| **Combined** | | |
| Number of FH cases detected | 1,061 | 345 |
| **Screening efficiency** | | |
| Number needed to screen to identify one FH case (population screen) | 330 | 725 |
| Number needed to sequence to identify one FH case (population screen) | 16 | 75 |
| Number needed to screen to identify one FH case (combined) | 94 | 290 |
| Number needed to sequence to identify one FH case (combined) | 5 | 30 |

**Table 5.5 Comparison of child-parent and adult two-stage screening strategies for FH.** The FH prevalence of 1:288 of the UK Biobank was applied to the child-parent screen and counts were adjusted accordingly.[10,18] [1]Estimates are based on figures from the Office of National Statistics where the average UK family comprises two children.[39]

Two-stage adult screening of 100,000 individuals based on an LDL-C cut-off of >4.8 mmol/L (1.31 MoM) would identify just under half as many FH cases (138 versus 303) for twice the burden of sequencing (10,398 versus 4,984 samples) when compared to child-parent screening of 100,000 participants at age two using a total cholesterol cut-off of ≥1.35 MoM (**Table 5.5**).[10,18] Child-parent screening identifies 1 index case per 330 individuals screened and 16 sequenced, whereas the corresponding values for two-stage adult screening are 725 and 75 (**Table 5.5**). Because child-parent screening can seed cascade testing of three generations rather than two, it identifies about three times as many cases as two-stage adult screening (1,061 versus 345 cases for 100,000 individuals screened) (**Table 5.5**). Child-parent screening identifies 1 index case per 330 individuals screened and 16 sequenced, whereas the corresponding values for two-stage adult screening are

725 and 75. Overall (combining screening and cascade testing of index cases), child-parent screening identifies 1 index case per 94 individuals screened and 5 sequenced, whereas the corresponding values for two-stage adult screening are 290 and 30 (**Table 5.5**).

### 5.4.5 Achieving the NHS target for FH case detection

In 2019 the NHS Long Term Plan set to increase FH detection from ~7% to 25% (~49,000 additional cases) in the next five years but has not outlined how this will be achieved.[13] Here I compare how long it would take to reach this target if the child-parent screen or the two-stage adult screen were to be implemented as population screening strategies in the UK.

In 2020 there were 681,560 live births in England and Wales, and immunisation uptake was ~95%.[40,41] This equates to around 650,000 children eligible for FH screening each year using the child-parent screening approach. Assuming that 1,061 FH cases are detected for 100,000 individuals screened with the child-parent screening strategy when followed by cascade testing (**Table 5.5**), 4.6 million children would have to be screened and ¼ million sequenced to identify 49,000 additional FH cases, which would take approximately 7 years to reach the NHS Long Term Plan goal's FH target identification of 25%.

In 2017/2018, around one million adults attended their NHS Health Check.[42] This means that 14 million adults would have to be screened and 1.5 million sequenced to identify 25% of all UK FH cases (**Table 5.5**), which would take approximately 14 years to reach the NHS' Long Term Plan goal with cascade testing.

## 5.5 Discussion

### 5.5.1 Overview of the study

I have modelled the performance of adult two-stage screen for FH in individuals between the age of 40 and 75 years (median age 58 years) using data available from the UK Biobank. Combining an inexpensive test with a high false positive rate at stage 1 (LDL-C concentration), with an expensive test with low false positive rate at stage 2 (DNA sequencing) was estimated to detect 40% (95% CI: 36-44%) of cases for a false positive rate of 10% (95% CI: 10-11%), using an LDL-C cut-off of 4.8 mmol/L (1.31 MoM). This cut-off would result in 10,398 samples being sequenced

for every 100,000 people screened. Lowering the LDL-C cut-off would increase the number of people with FH who are detected but increase the sequencing burden and *vice versa*.

The two-stage adult screen modelled here is less efficient than the child-parent screening approach proposed and evaluated previously.[18,19] Two-stage adult screening with two-generation cascade testing identifies about a third as many FH cases as childhood screening with three generation cascade testing, for twice the sequencing burden. Detecting 25% of all FH cases (the target set in the NHS Long Term Plan) requires screening around 14 million adults and sequencing of 1.5 million of them, or 4.6 million children and sequencing ¼ million of them if followed by cascade testing. Nevertheless, child-parent screening was not endorsed by the National Screening Committee when last reviewed on the grounds that it does not immediately benefit the children who are screened at around one year of age, because they do not become eligible to receive cholesterol lowering treatment until the age of ten.[20] The developers of the approach have countered this and other concerns,[45,46] but at present, child-parent screening is not in general use in the UK.

### 5.5.2 Limitations of the approach

Several assumptions were made to carry out this proof-of-principle study. Firstly, calculations were based on the average UK family size, and on the assumption that cascade testing identifies all affected close relatives, which is likely to be an overestimation.[47,48] In the calculations, I also did not account for the fact that the pool of unrelated index cases diminishes as more FH cases are identified through screening and cascade testing. However, the same assumptions were applied to both the child-parent and the adult two-stage screening strategies, which is unlikely to significantly affect the comparisons made between both strategies.

Secondly, I assumed that parents of affected adults were either deceased or too old to benefit from treatment in the cascade testing phase of the adult two-stage screen, which might not always be the case and is likely to be an underestimation of the performance of this screening strategy compared to the child-parent screen. Nevertheless, without taking cascade testing into account, the child-parent screen still outperformed the adult two-stage screen for FH in the study (**Table 5.5**).

Thirdly, this cohort was not fully representative of the UK population who attended the NHS Health Checks (**Table 5.2**). I used data available in UK Biobank because participants (aged between 40 and 75 at the time of recruitment) have had LDL-C measurements and exome sequencing data available for analysis. However, the median age of participants (58 years) is older than might be considered optimal for adult FH screening. Screening at a younger age (e.g., 40 to 50 years) would have the advantage of a potentially higher stage 1 detection rate, because LDL-C concentration better separates FH cases from those with an elevated LDL-C for other reasons at younger than older ages, and because screening parents of index cases as well as siblings and children may be possible (three rather than two generation screening). This study was a proof-of-principle, and a pilot of this study in the general UK population would provide invaluable information as to the feasibility and outcome of this two-stage screen. Cost-effectiveness studies would also provide additional information regarding the benefits of this screening strategy in the prevention of premature CHD and death.

Finally, the potential success of two-stage adult screening for FH in practice also requires consideration of several practical issues including (1) the potential setting of a screening programme; (2) the capacity of the NHS Genomic Medicine Service to undertake sequencing on the necessary scale; and (3) the capacity to undertake cascade testing of first-degree relatives.

### 5.5.3 Potential setting of a two-stage adult screening programme

The NHS Long-Term Plan has a stated aim of increasing the proportion of detected FH cases from 7% to 25% but does not elaborate on how this is to be achieved.[13] Cascade testing has been endorsed by the National Institute of Health and Care Excellence (NICE),[16] but the efficiency of cascade testing is dependent on the detection of index cases. The approach I have modelled involves a two-stage population screen in adults in which the high false positive rate of a stage 1 LDL-C measurement is mitigated by the low false positive rate of a stage 2 DNA sequencing test. Although less efficient than child-parent screening, it avoids concerns about screening for FH in childhood.

Until the COVID-19 pandemic, an NHS Health Check was operating in England since 2009, and was offered to men and women aged 40 to 74 without previously diagnosed hypertension, diabetes mellitus, FH, CHD, heart failure, atrial fibrillation, stroke or transient ischaemic attack, peripheral arterial disease, or chronic kidney disease. Those already on statins or known to have a 10-year

CVD risk of ≥20% were also excluded. Of those invited, about 50% attended (about 1 million people per annum);[43] thus, if re-introduced, the NHS Health Check could, in principle, provide the setting for a two-stage population screen for FH modelled here.

### 5.5.4 DNA sequencing capacity in the NHS

The National Genomic Medicine Service was established to enable the NHS to harness the power of genomic technology and science to improve the health of the population and deliver on the commitments of the NHS Long Term Plan.[49] One of its stated aims is the "early detection and treatment of high-risk conditions including expanding genomic testing for familial hypercholesterolaemia".[49] To service a two-stage population screen of the type I have modelled, with around 1 million people per annum undergoing stage 1 of the screen via the NHS Health Check, the National Genomic Medicine Service would need develop capacity to offer targeted sequencing of FH-causing genes in around 400,000 people per annum (assuming a detection rate of 40%; **Table 5.5**).

### 5.5.5 Cascade testing capacity

In its guidelines, NICE has already drawn up recommendations for cascade testing in families where an FH-causing variant has been detected in an index case.[16] Under the current guidance, NICE suggests case finding should be based on identification of individuals whose total cholesterol concentration >7.5 mmol/L below and >9.0 mmol/L above 30 years of age, which is very high and therefore likely to miss many FH-variant carriers. However, it makes no recommendation on the systematic measurement of cholesterol concentration but implies that potential cases are identified through surveys of existing health records, which is not an approach that has been evaluated directly. Thus, although the mechanisms and financial support for cascade testing are already in place, the means to identify affected individuals through systematic population screening is not.

### 5.5.6 Reaching the NHS Long Term Plan target

The NHS Long-Term Plan's aim of increasing the proportion of detected FH cases from 7% to 25% requires the further identification of approximately 49,000 FH cases in the UK population. Using the child-parent screen followed by cascade testing and the assumptions listed in this study,

it would take approximately 7 years to reach this target (confirming Wald and Bestwick's estimate in a previous report),[51] and 14 years to reach it if employing the adult two-stage screen combined with cascade testing. These two estimations are well over the 5-year target set by the NHS Long Term Plan, which has not specified how it will achieve it.[13] However, these two screening strategies are not mutually exclusive and could both be rolled out if approved by the UK National Screening Committee. If sequencing capacities are not able to cope with increased demand, it is likely that the NHS might have to reconsider their Long Term Plan goal of identifying 25% of UK FH cases in 5 years, or detail how this might be achieved.

### 5.5.7 Conclusion

In summary, I have used data from the UK Biobank to model the performance of two-stage adult population screen to identify index FH cases and to estimate the performance of cascade screening in affected families. I compared its performance to child-parent screening, and although I found it to be less efficient, two-stage adult screening could be used to meet the target of detecting 25% of FH cases stated in the NHS Long Term Plan in 14 years. The approach could be evaluated prospectively, and if feasible and cost-effective, then the foundations for a national programme may already be in place through the NHS Health Check, the NHS Genomic Medicine Service, and the NICE endorsed frameworks for cascade testing.

## 5.6 References


1 Vallejo-Vaz AJ, Marco M De, Stevens CAT, *et al.* Overview of the current status of familial hypercholesterolaemia care in over 60 countries - The EAS Familial Hypercholesterolaemia Studies Collaboration (FHSC). *Atherosclerosis* 2018;**277**:234–55. doi:10.1016/J.ATHEROSCLEROSIS.2018.08.051

2 Nordestgaard BG, Chapman MJ, Humphries SE, *et al.* Familial hypercholesterolaemia is underdiagnosed and undertreated in the general population: guidance for clinicians to prevent coronary heart disease Consensus Statement of the European Atherosclerosis Society. *Eur Heart J* 2013;**34**:3478–90. doi:10.1093/EURHEARTJ/EHT273

3 Akioyamen LE, Genest J, Shan SD, *et al.* Estimating the prevalence of heterozygous familial hypercholesterolaemia: a systematic review and meta-analysis. *BMJ Open* 2017;**7**. doi:10.1136/BMJOPEN-2017-016461

4 Austin MA, Hutter CM, Zimmern RL, *et al.* Genetic causes of monogenic heterozygous familial hypercholesterolemia: a HuGE prevalence review. *Am J Epidemiol* 2004;**160**:407–20. doi:10.1093/AJE/KWH236

5 Awan Z, Choi HY, Stitziel N, *et al.* APOE p.Leu167del mutation in familial hypercholesterolemia. *Atherosclerosis* 2013;**231**:218–22. doi:10.1016/J.ATHEROSCLEROSIS.2013.09.007

6 Benn M, Watts GF, Tybjaerg-Hansen A, *et al.* Familial hypercholesterolemia in the danish general population: prevalence, coronary artery disease, and cholesterol-lowering medication. *J Clin Endocrinol Metab* 2012;**97**:3956–64. doi:10.1210/JC.2012-1563

7 Tada H, Kawashiri MA, Nohara A, *et al.* Impact of clinical signs and genetic diagnosis of familial hypercholesterolaemia on the prevalence of coronary artery disease in patients with severe hypercholesterolaemia. *Eur Heart J* 2017;**38**:1573–9. doi:10.1093/EURHEARTJ/EHX004

8 Khera A V., Won HH, Peloso GM, *et al.* Diagnostic Yield and Clinical Utility of Sequencing Familial Hypercholesterolemia Genes in Patients With Severe Hypercholesterolemia. *J Am Coll Cardiol* 2016;**67**:2578–89. doi:10.1016/J.JACC.2016.03.520

9 Morris JK, Wald DS, Wald NJ. The evaluation of cascade testing for familial hypercholesterolemia. *Am J Med Genet A* 2012;**158A**:78–84. doi:10.1002/AJMG.A.34368

10 Wald DS, Wald NJ. Integration of child–parent screening and cascade testing for familial hypercholesterolaemia. *J Med Screen* Published Online First: 14 October 2018. doi:10.1177/0969141318796856

11 Population estimates for the UK, England and Wales, Scotland and Northern Ireland - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2020 (accessed 28 Feb 2022).

12 Green P, Saunders T. Detecting familial hypercholesterolaemia (FH) in general practice: An audit within the Medway Clinical Commissioning Group. *Atherosclerosis* 2014;**236**:e307. doi:10.1016/J.ATHEROSCLEROSIS.2014.08.012

13 NHS Long Term Plan » Cardiovascular disease. https://www.longtermplan.nhs.uk/online-version/chapter-3-further-progress-on-care-quality-and-outcomes/better-care-for-major-health-conditions/cardiovascular-disease/ (accessed 2 Mar 2022).

14 Carson JAS, Lichtenstein AH, Anderson CAM, *et al.* Dietary cholesterol and cardiovascular risk: A science advisory from the American heart association. *Circulation* 2020;**141**:E39–53. doi:10.1161/CIR.0000000000000743

15 Talmud PJ, Shah S, Whittall R, *et al.* Use of low-density lipoprotein cholesterol gene score to distinguish patients with polygenic and monogenic familial hypercholesterolaemia: A case-control study. *The Lancet* 2013;**381**:1293–301. doi:10.1016/S0140-6736(12)62127-8/ATTACHMENT/66BC0F03-0A61-42C7-ADC9-213680782539/MMC1.PDF

16 Overview | Familial hypercholesterolaemia: identification and management | Guidance | NICE. https://www.nice.org.uk/guidance/cg71 (accessed 28 Feb 2022).

17 Abul-Husn NS, Manickam K, Jones LK, *et al.* Genetic identification of familial hypercholesterolemia within a single U.S. health care system. *Science* 2016;**354**. doi:10.1126/SCIENCE.AAF7000

18 Wald DS, Bestwick JP, Morris JK, *et al.* Child–Parent Familial Hypercholesterolemia Screening in Primary Care. *New England Journal of Medicine* 2016;**375**:1628–37. doi:10.1056/NEJMOA1602777

19 Futema M, Cooper JA, Charakida M, *et al.* Screening for familial hypercholesterolaemia in childhood: Avon Longitudinal Study of Parents and Children (ALSPAC). *Atherosclerosis* 2017;**260**:47. doi:10.1016/J.ATHEROSCLEROSIS.2017.03.007

20 Familial Hypercholesterolaemia (child) - UK National Screening Committee (UK NSC) - GOV.UK. https://view-health-screening-recommendations.service.gov.uk/familial-hypercholesterolaemia-child/ (accessed 28 Feb 2022).

21 Duddy C, Wong G, Gadsby EW, *et al.* NHS Health Check programme: a protocol for a realist review. *BMJ Open* 2021;**11**:e048937. doi:10.1136/BMJOPEN-2021-048937

22 Sudlow C, Gallacher J, Allen N, *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015;**12**. doi:10.1371/JOURNAL.PMED.1001779

23 Kerr M, Pears R, Miedzybrodzka Z, *et al.* Cost effectiveness of cascade testing for familial hypercholesterolaemia, based on data from familial hypercholesterolaemia services in the UK. *Eur Heart J* 2017;**38**:1832–9. doi:10.1093/EURHEARTJ/EHX111

24 Bycroft C, Freeman C, Petkova D, *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018;**562**:203–9. doi:10.1038/S41586-018-0579-Z

25 Fry D, Almond R, Moffat S, *et al.* UK Biobank Biomarker Project Companion Document to Accompany Serum Biomarker Data. 2019.

26 van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *J Stat Softw* 2011;**45**:1–67. doi:10.18637/jss.v045.i03

27 Trinder M, Francis GA, Brunham LR. Association of Monogenic vs Polygenic Hypercholesterolemia with Risk of Atherosclerotic Cardiovascular Disease. *JAMA Cardiol* 2020;**5**:390–9. doi:10.1001/jamacardio.2019.5954

28 Backman JD, Li AH, Marcketta A, *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* 2021;**599**:628–34. doi:10.1038/S41586-021-04103-Z

29 Van Hout C V., Tachmazidou I, Backman JD, *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* 2020;**586**:749. doi:10.1038/S41586-020-2853-0

30 Danecek P, Bonfield JK, Liddle J, *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* 2021;**10**. doi:10.1093/GIGASCIENCE/GIAB008

31 McLaren W, Gil L, Hunt SE, *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* 2016;**17**:1–14. doi:10.1186/S13059-016-0974-4

32 Karczewski KJ, Francioli LC, Tiao G, *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;**581**:434–43. doi:10.1038/s41586-020-2308-7

33 Adzhubei IA, Schmidt S, Peshkin L, *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* 2010;**7**:248–9. doi:10.1038/nmeth0410-248

34 Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* Published Online First: 2009. doi:10.1038/nprot.2009.86

35 Ellard S, Baple EL, Callaway A, *et al.* ACGS Best Practice Guidelines for Variant Classification in Rare Disease 2020 Recommendations ratified by ACGS Quality Subcommittee on 4 th. *bioRxiv* Published Online First: 2020. doi:10.1101/531210

36 Leigh S, Futema M, Whittall R, *et al.* The UCL low-density lipoprotein receptor gene variant database: pathogenicity update. *J Med Genet* 2017;**54**:217–23. doi:10.1136/JMEDGENET-2016-104054

37 Lázaro Conxi, Lerner-Ellis Jordan, Spurdle Amanda. *Clinical DNA variant interpretation theory and practice.* Academic Press 2021.

38 Wald DS, Bestwick JP, Wald NJ. Child-parent screening for familial hypercholesterolaemia: screening strategy based on a meta-analysis. *BMJ* 2007;**335**:599. doi:10.1136/BMJ.39300.616076.55

39 Families and households in the UK - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/bulletins/familiesandhouseholds/2020 (accessed 1 Mar 2022).

40 Births in England and Wales - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/birthsummarytablesenglandandwales/2020 (accessed 12 Jul 2022).

41 Childhood Vaccination Coverage Statistics - England - 2019-20 - NHS Digital. https://digital.nhs.uk/data-and-information/publications/statistical/nhs-immunisation-statistics/england---2019-20 (accessed 12 Jul 2022).

42 NHS Health Check Programme, Patients Recorded as Attending and Not Attending, 2012-13 to 2017-18 - NHS Digital. https://digital.nhs.uk/data-and-information/publications/statistical/nhs-health-check-programme/2012-13-to-2017-18 (accessed 12 Jul 2022).

43 NHS Health Check programme, Patients Recorded as Attending and Not Attending, 2012-13 to 2017-18 - NHS Digital. https://digital.nhs.uk/data-and-information/publications/statistical/nhs-health-check-programme/2012-13-to-2017-18 (accessed 3 Mar 2022).

44 Richards S, Aziz N, Bale S, *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;**17**:405–24. doi:10.1038/GIM.2015.30

45 Wald DS, Martin AC. Decision to reject screening for familial hypercholesterolaemia is flawed. *Arch Dis Child* 2021;**106**:525–6. doi:10.1136/ARCHDISCHILD-2020-319168

46 Wald DS, Neely D. The UK National Screening Committee's position on child–parent screening for familial hypercholesterolaemia. *J Med Screen* 2021;**28**:217. doi:10.1177/09691413211025426

47 Bhatnagar D, Morgan J, Siddiq S, *et al.* Outcome of case finding among relatives of patients with known heterozygous familial hypercholesterolaemia. *BMJ* 2000;**321**:1497–500. doi:10.1136/BMJ.321.7275.1497

48 Hadfield SG, Horara S, Starr BJ, *et al.* Family tracing to identify patients with familial hypercholesterolaemia: the second audit of the Department of Health Familial Hypercholesterolaemia Cascade Testing Project. *Ann Clin Biochem* 2009;**46**:24–32. doi:10.1258/ACB.2008.008094

49 NHS England » NHS Genomic Medicine Service. https://www.england.nhs.uk/genomics/nhs-genomic-med-service/ (accessed 2 Mar 2022).

50 NHS England » National Genomic Test Directory. https://www.england.nhs.uk/publication/national-genomic-test-directories/ (accessed 2 Mar 2022).

51 Wald DS, Bestwick JP. Reaching detection targets in familial hypercholesterolaemia: Comparison of identification strategies. *Atherosclerosis* 2020;**293**:57–61. doi:10.1016/j.atherosclerosis.2019.11.028

## 5.7 Appendix

**Supplementary Table 5.1 Autosomal dominant FH-causing mutation identified in the study cohort.** Genetic coordinates are mapped to GRCh38.

| Gene | Chromosome | Position | Reference allele | Alternate allele | Nucleotide change | Protein | Number of carriers | UKB frequency (1/n) |
|---|---|---|---|---|---|---|---|---|
| *APOB* | 2 | 21006289 | G | A | c.10579C>T | p.Arg3527Trp | 2 | 70,220 |
| | | 21006288 | C | T | c.10580G>A | p.Arg3527Gln | 99 | 1,419 |
| *APOE* | 19 | 44908791 | GCTC | G | c.499_501del | p.Leu167del | 13 | 10,803 |
| *LDLR* | 19 | 11100236 | C | G | c.81C>G | p.Cys27Trp | 1 | 140,439 |
| | | 11100291 | T | G | c.136T>G | p.Cys46Gly | 1 | 140,439 |
| | | 11100294 | G | A | c.139G>A | p.Asp47Asn | 5 | 28,088 |
| | | 11102705 | C | T | c.232C>T | p.Arg78Cys | 13 | 10,803 |
| | | 11102714 | C | T | c.241C>T | p.Arg81Cys | 2 | 70,220 |
| | | 11102732 | T | G | c.259T>G | p.Trp87Gly | 6 | 23,407 |
| | | 11102741 | G | A | c.268G>A | p.Asp90Asn | 5 | 28,088 |
| | | 11102765 | G | A | c.292G>A | p.Gly98Ser | 10 | 14,044 |
| | | 11102774 | G | A | c.301G>A | p.Glu101Lys | 12 | 11,703 |
| | | 11102787 | G | A | c.313+1G>A | . | 5 | 28,088 |
| | | 11102787 | G | C | c.313+1G>C | . | 1 | 140,439 |
| | | 11102787 | G | GT | c.313+2dup | . | 2 | 70,220 |
| | | 11105249 | C | T | c.343C>T | p.Arg115Cys | 2 | 70,220 |
| | | 11105268 | G | T | c.362G>T | p.Cys121Phe | 2 | 70,220 |
| | | 11105324 | G | A | c.418G>A | p.Glu140Lys | 1 | 140,439 |
| | | 11105339 | GTGCTCACCTGTGGTCCCGCCAGC | G | c.435_457del | p.Leu146ProfsTer26 | 1 | 140,439 |
| | | 11105407 | C | A | c.501C>A | p.Cys167Ter | 2 | 70,220 |
| | | 11105408 | G | A | c.502G>A | p.Asp168Asn | 14 | 10,031 |
| | | 11105415 | AC | A | c.513del | p.Asp172ThrfsTer34 | 1 | 140,439 |
| | | 11105448 | C | G | c.542C>G | p.Pro181Arg | 2 | 70,220 |
| | | 11105549 | C | T | c.643C>T | p.Arg215Cys | 4 | 35,110 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 11105567 | G | A | c.661G>A | p.Asp221Asn | 2 | 70,220 |
| | | 11105568 | A | G | c.662A>G | p.Asp221Gly | 5 | 28,088 |
| | | 11105585 | GAC | G | c.680_681del | p.Asp227GlyfsTer12 | 4 | 35,110 |
| | | 11105585 | GAC | GAG | c.681delinsG | p.Asp227Glu | 2 | 70,220 |
| | | 11105588 | G | T | c.682G>T | p.Glu228Ter | 2 | 70,220 |
| | | 11105589 | AG | A | c.685del | p.Glu229LysfsTer36 | 1 | 140,439 |
| | | 11106579 | C | T | c.709C>T | p.Arg237Cys | 1 | 140,439 |
| | | 11106588 | G | A | c.718G>A | p.Glu240Lys | 20 | 7,022 |
| | | 11106592 | T | C | c.722T>C | p.Phe241Ser | 1 | 140,439 |
| | | 11106631 | A | C | c.761A>C | p.Gln254Pro | 1 | 140,439 |
| | | 11107432 | C | A | c.858C>A | p.Ser286Arg | 1 | 140,439 |
| | | 11107433 | G | A | c.859G>A | p.Gly287Ser | 4 | 35,110 |
| | | 11107436 | G | A | c.862G>A | p.Glu288Lys | 1 | 140,439 |
| | | 11107461 | G | A | c.887G>A | p.Cys296Tyr | 1 | 140,439 |
| | | 11107481 | C | T | c.907C>T | p.Arg303Trp | 2 | 70,220 |
| | | 11107486 | C | G | c.912C>G | p.Asp304Glu | 4 | 35,110 |
| | | 11107512 | G | A | c.938G>A | p.Cys313Tyr | 2 | 70,220 |
| | | 11110660 | G | A | c.949G>A | p.Glu317Lys | 35 | 4,013 |
| | | 11110678 | G | A | c.967G>A | p.Gly323Ser | 1 | 140,439 |
| | | 11110714 | G | A | c.1003G>A | p.Gly335Ser | 3 | 46,813 |
| | | 11110738 | G | A | c.1027G>A | p.Gly343Ser | 8 | 17,555 |
| | | 11110759 | C | T | c.1048C>T | p.Arg350Ter | 4 | 35,110 |
| | | 11110760 | G | C | c.1049G>C | p.Arg350Pro | 4 | 35,110 |
| | | 11111571 | G | A | c.1118G>A | p.Gly373Asp | 1 | 140,439 |
| | | 11111619 | C | T | c.1166C>T | p.Thr389Met | 8 | 17,555 |
| | | 11113286 | G | A | c.1195G>A | p.Ala399Thr | 1 | 140,439 |
| | | 11113287 | C | A | c.1196C>A | p.Ala399Asp | 1 | 140,439 |
| | | 11113292 | CTCTTC | CTCT | c.1205_1206del | p.Phe403HisfsTer37 | 1 | 140,439 |
| | | 11113307 | C | T | c.1216C>T | p.Arg406Trp | 5 | 28,088 |
| | | 11113308 | G | A | c.1217G>A | p.Arg406Gln | 4 | 35,110 |
| | | 11113313 | G | A | c.1222G>A | p.Glu408Lys | 1 | 140,439 |
| | | 11113322 | A | G | c.1231A>G | p.Lys411Glu | 1 | 140,439 |
| | | 11113329 | C | T | c.1238C>T | p.Thr413Met | 14 | 10,031 |
| | | 11113337 | C | T | c.1246C>T | p.Arg416Trp | 2 | 70,220 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 11113419 | G | C | c.1328G>C | p.Trp443Ser | 1 | 140,439 |
| | | 11113426 | C | G | c.1335C>G | p.Asp445Glu | 5 | 28,088 |
| | | 11113554 | CA | C | c.1379del | p.His460ProfsTer47 | 1 | 140,439 |
| | | 11113590 | G | T | c.1414G>T | p.Asp472Tyr | 6 | 23,407 |
| | | 11113608 | G | A | c.1432G>A | p.Gly478Arg | 2 | 70,220 |
| | | 11113612 | T | C | c.1436T>C | p.Leu479Pro | 2 | 70,220 |
| | | 11113620 | G | A | c.1444G>A | p.Asp482Asn | 29 | 4,843 |
| | | 11113650 | G | A | c.1474G>A | p.Asp492Asn | 1 | 140,439 |
| | | 11113678 | C | T | c.1502C>T | p.Ala501Val | 5 | 28,088 |
| | | 11113705 | C | T | c.1529C>T | p.Thr510Met | 3 | 46,813 |
| | | 11113743 | G | A | c.1567G>A | p.Val523Met | 1 | 140,439 |
| | | 11116095 | T | G | c.1588T>G | p.Phe530Val | 10 | 14,044 |
| | | 11116125 | G | A | c.1618G>A | p.Ala540Thr | 2 | 70,220 |
| | | 11116141 | G | A | c.1634G>A | p.Gly545Glu | 1 | 140,439 |
| | | 11116198 | A | G | c.1691A>G | p.Asn564Ser | 2 | 70,220 |
| | | 11116873 | C | T | c.1720C>T | p.Arg574Cys | 2 | 70,220 |
| | | 11116898 | T | C | c.1745T>C | p.Leu582Pro | 1 | 140,439 |
| | | 11116918 | G | A | c.1765G>A | p.Asp589Asn | 1 | 140,439 |
| | | 11116928 | G | A | c.1775G>A | p.Gly592Glu | 1 | 140,439 |
| | | 11116936 | C | T | c.1783C>T | p.Arg595Trp | 6 | 23,407 |
| | | 11116937 | G | A | c.1784G>A | p.Arg595Gln | 2 | 70,220 |
| | | 11116976 | C | G | c.1823C>G | p.Pro608Arg | 1 | 140,439 |
| | | 11120091 | G | A | c.1846-1G>A | . | 1 | 140,439 |
| | | 11120106 | G | T | c.1860G>T | p.Trp620Cys | 1 | 140,439 |
| | | 11120110 | GAT | G | c.1867_1868del | p.Ile623HisfsTer21 | 1 | 140,439 |
| | | 11120143 | C | T | c.1897C>T | p.Arg633Cys | 9 | 15,604 |
| | | 11120144 | G | A | c.1898G>A | p.Arg633His | 1 | 140,439 |
| | | 11120152 | G | A | c.1906G>A | p.Gly636Ser | 3 | 46,813 |
| | | 11120212 | C | A | c.1966C>A | p.His656Asn | 8 | 17,555 |
| | | 11120370 | G | A | c.1988G>A | p.Gly663Glu | 1 | 140,439 |
| | | 11120408 | G | A | c.2026G>A | p.Gly676Ser | 5 | 28,088 |
| | | 11120436 | C | T | c.2054C>T | p.Pro685Leu | 12 | 11,703 |
| | | 11120441 | A | T | c.2059A>T | p.Ile687Phe | 5 | 28,088 |
| | | 11120442 | T | TC | c.2061dup | p.Asn688GlnfsTer29 | 1 | 140,439 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 11123200 | G | T | c.2167G>T | p.Glu723Ter | 1 | 140,439 |
| | | 11128027 | C | CA | c.2332dup | p.Arg778LysfsTer4 | 1 | 140,439 |

**Supplementary Table 5.2 List of variants of unknown significance (VUS) excluded from the analysis.** Genetic coordinates are mapped to GRCh38. Count refers to the number of participants having the VUS.

| Gene | Chromosome number | Position | Reference allele | Alternate allele | HGVSc | HGVSp | Count |
|---|---|---|---|---|---|---|---|
| *APOB* | 2 | 21001939 | ACTG | A | ENST00000233242:c.13480_13482delCAG | ENSP00000233242.1:p.Gln4494del | 132 |
| | | 21006196 | C | T | ENST00000233242:c.10672C>T | ENSP00000233242.1:p.Arg3558Cys | 299 |
| | | 21006239 | C | G | ENST00000233242:c.10629C>G | ENSP00000233242.1:p.Asn3543Lys | 3 |
| | | 21006349 | C | T | ENST00000233242:c.10519C>T | ENSP00000233242.1:p.Arg3507Trp | 1 |
| | | 21015387 | G | C | ENST00000233242:c.3491G>C | ENSP00000233242.1:p.Arg1164Thr | 1 |
| *PCSK9* | 1 | 55044021 | A | G | ENST00000302118:c.386A>G | ENSP00000303208.5:p.Asp129Gly | 2 |
| | | 55052698 | G | A | ENST00000302118:c.706G>A | ENSP00000303208.5:p.Gly236Ser | 4 |
| | | 55058543 | C | G | ENST00000302118:c.1399C>G | ENSP00000303208.5:p.Pro467Ala | 3 |
| *LDLR* | 19 | 11100261 | G | C | ENST00000558518.6:c.106G>C | ENSP00000454071.1:p.Asp36His | 1 |
| | | 11100322 | C | T | ENST00000558518.6:c.167C>T | ENSP00000454071.1:p.Ser56Phe | 1 |
| | | 11100328 | A | T | ENST00000558518.6:c.173A>T | ENSP00000454071.1:p.Glu58Val | 2 |
| | | 11100340 | C | T | ENST00000558518.6:c.185C>T | ENSP00000454071.1:p.Thr62Met | 10 |
| | | 11102720 | A | T | ENST00000558518.6:c.247A>T | ENSP00000454071.1:p.Ile83Phe | 1 |
| | | 11105262 | G | C | ENST00000558518.6:c.356G>C | ENSP00000454071.1:p.Gly119Ala | 1 |
| | | 11105337 | C | T | ENST00000558518.6:c.431C>T | ENSP00000454071.1:p.Pro144Leu | 1 |
| | | 11105379 | C | T | ENST00000558518.6:c.473C>T | ENSP00000454071.1:p.Ser158Phe | 1 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | 11105414 | G | A | ENST00000558518.6:c.508G>A | ENSP00000454071.1:p.Asp170Asn | 22 |
| | | 11105415 | AC | GC | ENST00000558518.6:c.509delinsG | ENSP00000454071.1:p.Asp170Gly | 1 |
| | | 11106580 | G | A | ENST00000558518.6:c.710G>A | ENSP00000454071.1:p.Arg237His | 10 |
| | | 11106593 | C | A | ENST00000558518.6:c.723C>A | ENSP00000454071.1:p.Phe241Leu | 3 |
| | | 11106601 | C | G | ENST00000558518.6:c.731C>G | ENSP00000454071.1:p.Ser244Cys | 1 |
| | | 11106639 | C | T | ENST00000558518.6:c.769C>T | ENSP00000454071.1:p.Arg257Trp | 2 |
| | | 11107472 | A | G | ENST00000558518.6:c.898A>G | ENSP00000454071.1:p.Arg300Gly | 1 |
| | | 11111538 | A | C | ENST00000558518.6:c.1085A>C | ENSP00000454071.1:p.Asp362Ala | 60 |
| | | 11111558 | G | A | ENST00000558518.6:c.1105G>A | ENSP00000454071.1:p.Val369Met | 3 |
| | | 11111609 | G | T | ENST00000558518.6:c.1156G>T | ENSP00000454071.1:p.Asp386Tyr | 5 |
| | | 11113278 | G | T | ENST00000558518.6:c.1187G>T | ENSP00000454071.1:p.Gly396Val | 1 |
| | | 11113287 | C | T | ENST00000558518.6:c.1196C>T | ENSP00000454071.1:p.Ala399Val | 1 |
| | | 11113292 | CTCTTC | CTCTTG | ENST00000558518.6:c.1206delinsG | ENSP00000454071.1:p.Phe402Leu | 1 |
| | | 11113362 | C | T | ENST00000558518.6:c.1271C>T | ENSP00000454071.1:p.Pro424Leu | 3 |
| | | 11113374 | A | C | ENST00000558518.6:c.1283A>C | ENSP00000454071.1:p.Asn428Thr | 1 |
| | | 11113409 | A | G | ENST00000558518.6:c.1318A>G | ENSP00000454071.1:p.Arg440Gly | 4 |
| | | 11113561 | TCTCTTCCTA | TCTCTTACTA | ENST00000558518.6:c.1391delinsA | ENSP00000454071.1:p.Ser464Tyr | 2 |
| | | 11113625 | G | T | ENST00000558518.6:c.1449G>T | ENSP00000454071.1:p.Trp483Cys | 1 |
| | | 11113751 | T | G | ENST00000558518.6:c.1575T>G | ENSP00000454071.1:p.Asp525Glu | 6 |
| | | 11113762 | G | T | ENST00000558518.6:c.1586G>T | ENSP00000454071.1:p.Gly529Val | 1 |
| | | 11116101 | T | C | ENST00000558518.6:c.1594T>C | ENSP00000454071.1:p.Tyr532His | 1 |
| | | 11116132 | T | A | ENST00000558518.6:c.1625T>A | ENSP00000454071.1:p.Ile542Asn | 1 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | 11116205 | C | G | ENST00000558518.6:c.1698C>G | ENSP00000454071.1:p.Ile566Met | 1 |
| | | 11116885 | G | A | ENST00000558518.6:c.1732G>A | ENSP00000454071.1:p.Val578Ile | 2 |
| | | 11116914 | C | G | ENST00000558518.6:c.1761C>G | ENSP00000454071.1:p.Ser587Arg | 4 |
| | | 11116949 | T | C | ENST00000558518.6:c.1796T>C | ENSP00000454071.1:p.Leu599Ser | 4 |
| | | 11116970 | C | A | ENST00000558518.6:c.1817C>A | ENSP00000454071.1:p.Ala606Asp | 14 |
| | | 11120454 | C | T | ENST00000558518.6:c.2072C>T | ENSP00000454071.1:p.Ser691Leu | 4 |
| | | 11120484 | G | T | ENST00000558518.6:c.2102G>T | ENSP00000454071.1:p.Gly701Val | 1 |
| | | 11120507 | A | G | ENST00000558518.6:c.2125A>G | ENSP00000454071.1:p.Arg709Gly | 1 |
| | | 11123315 | C | T | ENST00000558518.6:c.2282C>T | ENSP00000454071.1:p.Thr761Met | 11 |
| | | 11128062 | C | A | ENST00000558518.6:c.2366C>A | ENSP00000454071.1:p.Ala789Asp | 1 |
| | | 11129553 | G | C | ENST00000558518.6:c.2430G>C | ENSP00000454071.1:p.Trp810Cys | 1 |
| | | 11129573 | A | T | ENST00000558518.6:c.2450A>T | ENSP00000454071.1:p.Asn817Ile | 1 |
| | | 11129582 | G | A | ENST00000558518.6:c.2459G>A | ENSP00000454071.1:p.Ser820Asn | 1 |
| | | 11129633 | A | G | ENST00000558518.6:c.2510A>G | ENSP00000454071.1:p.His837Arg | 18 |
| | | 11129653 | G | A | ENST00000558518.6:c.2530G>A | ENSP00000454071.1:p.Gly844Ser | 1 |
| | | 11131299 | G | C | ENST00000558518.6:c.2566G>C | ENSP00000454071.1:p.Glu856Gln | 1 |

**Supplementary Table 5.3 Study participants characteristics categorised by FH-causing gene.** P-value of group differences between FH-causing genes are shown and obtained from the Kruskal-Wallis Rank sum test. Missing (%) refers to the proportion of missing data in each field. BMI = body mass index; CHD = coronary heart disease; CVD = cardiovascular disease (defined as CHD, ischaemic and haemorrhagic stroke, heart failure, and atrial fibrillation); HDL-C = high-density lipoprotein cholesterol; IQR = interquartile range; LDL-C = low-density lipoprotein cholesterol.

| | ***LDLR*** | ***APOB*** | ***APOE*** | **P-value Kruskal-Wallis Rank sum test** | **Missing (%)** |
|---|---|---|---|---|---|
| **n** | 374 | 101 | 13 | | |
| **Age (median [IQR])** | 58.00 [51.00, 63.00] | 57.00 [51.00, 62.00] | 63.00 [53.00, 66.00] | 0.378 | 0.0 |
| **Sex (male) (%)** | 156 (41.7) | 45 (44.6) | 6 (46.2) | 0.844 | 0.0 |
| **Townsend deprivation index (median [IQR])** | -2.22 [-3.64, 0.16] | -2.20 [-3.98, 0.13] | -1.37 [-3.47, -0.12] | 0.848 | 0.4 |
| **Smoking status (%)** | | | | 0.825 | 0.0 |
| **Non-smoker** | 217 (58.0) | 59 (58.4) | 5 (38.5) | | |
| **Former smoker** | 131 (35.0) | 35 (34.7) | 7 (53.8) | | |
| **Light smoker (<10 cigarettes/day)** | 5 (1.3) | 2 (2.0) | 0 (0.0) | | |
| **Moderate smoker (10-19 cigarettes/day)** | 10 (2.7) | 2 (2.0) | 1 (7.7) | | |
| **Heavy Smoker (>20 cigarettes/day)** | 11 (2.9) | 3 (3.0) | 0 (0.0) | | |
| **BMI, kg/m2 (median [IQR])** | 26.92 [23.84, 29.61] | 27.78 [24.02, 30.06] | 25.73 [24.81, 27.26] | 0.245 | 0.2 |
| **Family history of CHD (%)** | 237 (63.4) | 61 (60.4) | 8 (61.5) | 0.857 | 0.0 |
| **Statin use (%)** | 128 (34.2) | 33 (32.7) | 4 (30.8) | 0.932 | 0.0 |
| **Biomarkers** | | | | | |
| **LDL-C (unadjusted), mmol/L (median [IQR])** | 3.74 [3.05, 4.71] | 4.35 [3.81, 5.32] | 3.55 [3.01, 4.25] | <0.001 | 0.0 |
| **LDL-C (adjusted for statin users), mmol/L (median [IQR])** | 4.28 [3.56, 5.23] | 5.01 [4.28, 5.76] | 3.68 [3.55, 4.98] | <0.001 | 0.0 |
| **HDL-C, mmol/L (median [IQR])** | 1.38 [1.18, 1.64] | 1.34 [1.15, 1.58] | 1.73 [1.50, 1.86] | 0.037 | 14.5 |
| **Total cholesterol, mmol/L (median [IQR])** | 5.93 [5.06, 7.02] | 6.56 [5.70, 8.03] | 5.58 [5.21, 6.58] | 0.001 | 6.1 |
| **Triglycerides, mmol/L (median [IQR])** | 1.28 [0.92, 1.91] | 1.26 [1.00, 1.93] | 0.76 [0.69, 1.13] | 0.019 | 6.1 |

| | | | | | |
|---|---|---|---|---|---|
| **Lipoprotein(a), nmol/L (median [IQR])** | 29.00 [11.52, 60.52] | 26.35 [8.52, 50.35] | 8.50 [4.77, 57.71] | 0.245 | 26.8 |
| **Apolipoprotein A, g/L (median [IQR])** | 1.46 [1.33, 1.64] | 1.44 [1.26, 1.60] | 1.67 [1.56, 1.75] | 0.029 | 15.0 |
| **Apolipoprotein B, g/L (median [IQR])** | 1.11 [0.96, 1.34] | 1.31 [1.12, 1.46] | 0.99 [0.81, 1.16] | <0.001 | 7.2 |
| **C-reactive protein, mg/L (median [IQR])** | 1.19 [0.58, 2.20] | 1.25 [0.69, 2.64] | 1.45 [0.76, 1.93] | 0.402 | 6.4 |
| **Disease prevalence & incidence** | | | | | |
| **CHD prevalence (%)** | 30 (8.0) | 10 (9.9) | 0 (0.0) | 0.457 | 0.0 |
| **CHD incidence (%)** | 26 (7.0) | 6 (5.9) | 0 (0.0) | 0.586 | 0.0 |
| **CVD prevalence (%)** | 34 (9.1) | 11 (10.9) | 0 (0.0) | 0.435 | 0.0 |
| **CVD incidence (%)** | 38 (10.2) | 8 (7.9) | 0 (0.0) | 0.395 | 0.0 |
| **Type 2 diabetes prevalence (%)** | 10 (2.7) | 1 (1.0) | 0 (0.0) | 0.514 | 0.0 |
| **Type 2 diabetes incidence (%)** | 17 (4.5) | 2 (2.0) | 0 (0.0) | 0.379 | 0.0 |

**Supplementary Table 5.4 The counts obtained from the two-stage screen in the study cohort of 140,439 individuals for various LDL-C cut-off values.** LDL-C = low-density lipoprotein cholesterol; OAPR = odds of being affected given a positive test result; VUS = variant of unknown significance.

| Cut-off | Detection rate (sensitivity) | False positive rate (1-specificity) | Positive predictive value (PPV) | Negative predictive value (NPV) | OAPR | Cases missed | True positive cases | False positive cases | Number sent for sequencing | Number of VUS above threshold |
|---|---|---|---|---|---|---|---|---|---|---|
| **3** | 90.6 (87.7-92.9) | 80.5 (80.3-80.7) | 0.4 (0.4-0.4) | 99.8 (99.8-99.9) | 1:255 | 46 | 442 | 112713 | 113155 | 586 |
| **3.5** | 79.7 (75.9-83) | 58.5 (58.3-58.8) | 0.5 (0.4-0.5) | 99.8 (99.8-99.9) | 1:211 | 99 | 389 | 81940 | 82329 | 465 |
| **4** | 64.8 (60.4-68.9) | 34.8 (34.6-35) | 0.6 (0.6-0.7) | 99.8 (99.8-99.8) | 1:154 | 172 | 316 | 48702 | 49018 | 306 |
| **4.1** | 61.5 (57.1-65.7) | 30.7 (30.4-30.9) | 0.7 (0.6-0.8) | 99.8 (99.8-99.8) | 1:143 | 188 | 300 | 42913 | 43213 | 275 |
| **4.2** | 59.2 (54.8-63.5) | 26.7 (26.5-27) | 0.8 (0.7-0.9) | 99.8 (99.8-99.8) | 1:129 | 199 | 289 | 37396 | 37685 | 248 |
| **4.3** | 54.1 (49.7-58.5) | 23.2 (22.9-23.4) | 0.8 (0.7-0.9) | 99.8 (99.8-99.8) | 1:123 | 224 | 264 | 32406 | 32670 | 223 |
| **4.4** | 50.8 (46.4-55.2) | 19.9 (19.7-20.1) | 0.9 (0.8-1) | 99.8 (99.8-99.8) | 1:112 | 240 | 248 | 27856 | 28104 | 195 |
| **4.5** | 47.7 (43.3-52.2) | 17.1 (16.9-17.3) | 1 (0.9-1.1) | 99.8 (99.8-99.8) | 1:102 | 255 | 233 | 23871 | 24104 | 174 |
| **4.6** | 45.5 (41.1-49.9) | 14.5 (14.3-14.7) | 1.1 (0.9-1.2) | 99.8 (99.7-99.8) | 1:92 | 266 | 222 | 20322 | 20544 | 151 |
| **4.7** | 43.2 (38.9-47.7) | 12.2 (12.1-12.4) | 1.2 (1.1-1.4) | 99.8 (99.7-99.8) | 1:81 | 277 | 211 | 17127 | 17338 | 129 |
| **4.8** | 39.8 (35.5-44.2) | 10.3 (10.1-10.5) | 1.3 (1.2-1.5) | 99.8 (99.7-99.8) | 1:74 | 294 | 194 | 14409 | 14603 | 111 |
| **4.9** | 37.7 (33.5-42.1) | 8.6 (8.5-8.7) | 1.5 (1.3-1.7) | 99.8 (99.7-99.8) | 1:65 | 304 | 184 | 12038 | 12222 | 99 |
| **5** | 34.8 (30.7-39.2) | 7.1 (7-7.3) | 1.7 (1.4-1.9) | 99.8 (99.7-99.8) | 1:59 | 318 | 170 | 9994 | 10164 | 83 |
| **6.8** | 3.9 (2.5-6) | 0.2 (0.2-0.2) | 6 (3.9-9.2) | 99.7 (99.6-99.7) | 1:16 | 469 | 19 | 297 | 316 | 2 |
| **8.5** | 0.4 (0.1-1.5) | 0 (0-0) | 8.7 (2.4-26.8) | 99.7 (99.6-99.7) | 1:10 | 486 | 2 | 21 | 23 | 0 |

# 6 A machine learning model to aid detection of familial hypercholesterolaemia (FH)

A preprint version of the following chapter can be found on medRxiv and has been submitted for publication.[1]

## 6.1 Abstract

**Background:** People with monogenic familial hypercholesterolaemia (FH) are at an increased risk of premature coronary heart disease and death. With a prevalence of 1:250, FH is relatively common; but there is no population screening strategy in place for FH so most carriers are only identified late in life, delaying timely and cost-effective interventions. The previous chapter modelled a two-stage adult screen with low-density lipoprotein cholesterol (LDL-C) concentration in stage 1, followed by confirmatory FH variant sequencing in stage 2. The aim of this chapter was to derive an algorithm to improve the identification of people with suspected monogenic FH in stage 1 for subsequent confirmatory genomic testing and cascade screening.

**Methods:** A penalised (LASSO) logistic regression model was used to identify predictors that most accurately identified people with a higher probability of monogenic FH in 139,779 unrelated participants of the UK Biobank. Candidate predictors included information on medical and family history, anthropometric measures, blood biomarkers, and a LDL-C polygenic score (PGS). Model derivation and evaluation was performed using a random split of 80% training and 20% testing data.

**Results:** 488 FH variant carriers were identified using whole exome-sequencing of the *LDLR, APOB, APOE* and *PCSK9* genes. A 14-variable algorithm for monogenic FH was derived, where the top five variables included triglyceride, LDL-C, and apolipoprotein A1 concentrations, self-reported statin use, and an LDL-C PGS. Model evaluation in the test data resulted in an area under the curve (AUC) of 0.77 (95% CI: 0.71; 0.83), and appropriate calibration (calibration-in-the-large: -0.07 (95% CI: -0.28; 0.13); calibration slope: 1.02 (95% CI: 0.85; 1.19)). Excluding the PGS as a candidate feature resulted in a 9-variable model with a comparable AUC: 0.76 (95% CI: 0.71; 0.82). Both multivariable models (with or without the PGS) outperformed screening-prioritisation based on LDL-C adjusted for statin use.

**Conclusion:** The detection of individuals with monogenic FH can be improved with the inclusion of additional non-genetic variables as well as a PGS for LDL-C. This reduces the burden of genetic sequencing required in an adult two-stage population screening strategy for FH.

## 6.2 Introduction

Familial hypercholesterolaemia (FH) is an autosomal dominant disorder caused by variants in the *LDLR, APOB, PCSK9,* or *APOE* genes. It is characterised by elevated low-density lipoprotein (LDL-C) concentration and premature coronary heart disease (CHD).[2] FH-causing variants are found in about 1 in 250 individuals (95% CI: 1:345; 1:192),[3] however the condition remains highly underdiagnosed worldwide with only an estimated 1% to 10% of cases diagnosed.[4,5] Affected individuals are at increased risk of premature CHD, where early initiation of lipid-lowering treatment is paramount for risk management.[4] There is currently no systematic way of identifying new index FH cases in the general population, although cascade testing in families of affected individuals has been shown to be highly cost-effective in many countries.[6–9] Currently, patient diagnosis often happens after the development of CHD symptoms or by opportunistic measurement of lipid profile and at the discretion of clinicians. Diagnosis is made using tools such as the Dutch Lipid Clinical Network (DLCN) and the Simon Broome criteria, which have not been designed to be used as population screening tools.[2]

In 2016, Wald *et al.* suggested screening children aged 15 months of age by measurement of total or LDL-C to systematically identify index monogenic FH cases in the general population as a prelude to testing parents and other family members.[10] Futema *et al.* showed that measurement of LDL-C alone at age 9 may be insufficiently accurate in reliably distinguishing FH-variant carriers from those with an elevated cholesterol as a consequence diet and lifestyle factors, or carriage of a high burden of common cholesterol-raising alleles, and suggested adding a confirmatory targeted-sequencing step to reduce the number of false positive cases detected.[11]

The increased availability of routine health checks in adults either through work-place schemes or local healthcare providers offers an opportunity to systematically identify adult carriers of FH-causing variants.[12] Positioning adult FH screening within routine health checks, which typically record a substantial number of other clinical measurements, offers the opportunity to consider additional predictors for FH. This may be important because, while the effect of FH on CHD risk is mediated through elevated circulating LDL-C concentration, it is well-known that LDL-C

concentration associates with other variables such as blood and liver biomarkers, diet, and also with common, genetic variants.[13] Combining multiple environmental factors and a polygenic score for LDL-C raising genetic variants may improve the detection of people with monogenic FH for prioritisation for confirmatory genetic testing.[14,15] This is because individuals with monogenic FH are likely to have a measured LDL-C concentration that is higher than can be accounted for by these other variables.

In the current chapter I utilise the UK Biobank data to evaluate the detection rate and testing burden of four prioritisation strategies to identify people with suspected FH-causing variants for confirmatory genetic testing: 1) no prioritisation (i.e., referring all participants for sequencing), 2) a plasma LDL-C-based prioritisation model adjusting for statin treatment, 3) a multivariable machine learning prioritisation model with non-genetic variables, 4) a multivariable machine learning prioritisation model which includes a polygenic score (PGS) for LDL-C.

## 6.3 Methods

### 6.3.1 Genomics data availability and FH case ascertainment

I identified 472,147 UK Biobank participants of White British ancestry (data-field 21000) as part of the approved project ID 40721. After performing genomic quality control steps (**Chapter 3**), 341,515 individuals remained, including 140,439 with whole-exome sequencing (WES) data necessary to identify those who carry an FH-causing variant. Causal FH variants were searched for in the WES data encompassing the *LDLR*, *APOB*, *PCSK9* and *APOE* genes (**Chapter 5** Methods). A total of 488 pathogenic and likely pathogenic FH variants were identified (**Chapter 5 Supplementary Table 5.1**). Additionally, 660 participants were found to carry FH variants of uncertain significance (VUS) (**Chapter 5 Supplementary Table 5.2**). These were excluded from the analysis because more evidence is required to interpret the effect of those VUS.

### 6.3.2 LDL-C PGS generation

I next generated a PGS for LDL-C concentration using an independent data subset of 173,672 White British participants without lipid-lowering medication or WES data (**Figure 6.1**). An initial list of 10,137 genetic variants with a p-value threshold of $<5x10^{-4}$ was obtained from the Global Lipids Genetics Consortium (GLGC) genome-wide association study (GWAS) summary statistics

for LDL-C.[16] To reduce the number of potentially redundant variants and optimise LDL-C prediction, I next applied a least absolute shrinkage and selection operator (LASSO) regression algorithm using the biglasso package in R.[17] The degree of penalisation was determined through 15-fold cross-validation, maximising the explained variance (R-squared), which resulted in a 1,466 genetic variant LDL-C PGS.

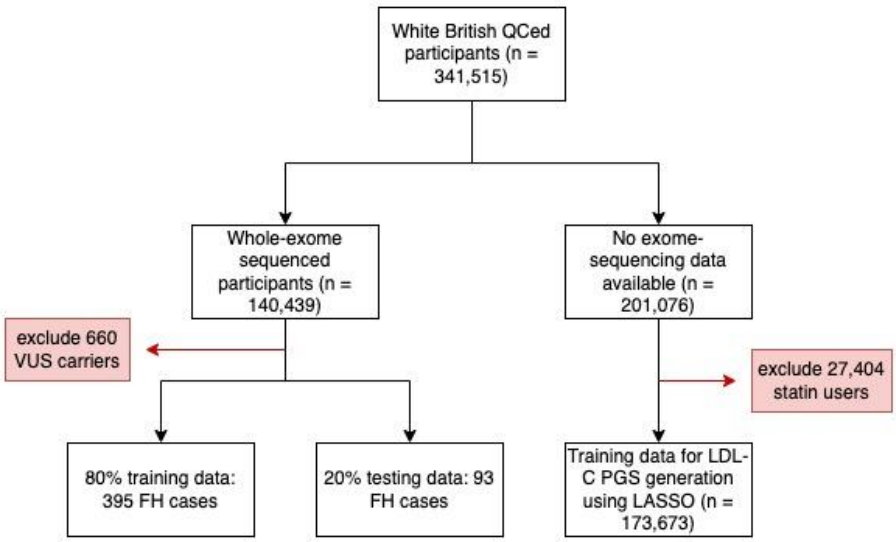



**Figure 6.1 Workflow of LDL-C PGS generation, FH case ascertainment and testing versus training data split of the UK Biobank's White British participants.** The data was split according to the availability of whole-exome sequencing data. FH = familial hypercholesterolaemia; LASSO = least absolute shrinkage and selection operator; PGS = polygenic score; QC = quality control; VUS = variants of uncertain significance.

### 6.3.3 Deriving a machine learning algorithm to prioritise participants with FH

I extracted data on a total of 24 candidate FH predictors, specifically: LDL-C, high-density lipoprotein cholesterol (HDL-C), total cholesterol, triglycerides, lipoprotein A (Lp(a)), apolipoprotein A1 (Apo-A1), apolipoprotein B (Apo-B), C-reactive protein (CRP), aspartate aminotransferase (AST), alanine aminotransferase (ALT), alkaline phosphatase (ALP), sex, body mass index (BMI), age, self-reported statin use, alcohol use, systolic blood pressure (SBP), diastolic blood pressure (DBP), Townsend deprivation index, smoking status, family history of CHD, type 2 diabetes diagnosis, hypertension, and LDL-C PGS. This was expanded by including 10 product terms between: age and LDL-C, age and LDL-C PGS, LDL-C PGS and LDL-C, age$^2$, LDL-C$^2$, statin use and LDL-C, family history of CHD and sex, family history of CHD and statin use, family

history of CHD and alcohol use, family history of CHD and hypertension. The limited missing data (**Supplementary Table 6.1**) were singly imputed using the R package MICE.[18]

Model derivation was performed using the WES data, applying a 80% training data split of 111,824 subjects, retaining 20% testing data (containing 93 carriers out of 27,955 subjects) to unbiasedly evaluate model performance (**Figure 6.1**). To prevent potential model instability, highly correlated variables (i.e. multicollinear) were removed. These included Apo-B and total cholesterol (**Supplementary Figure 6.1**). Variables were standardised to mean zero and standard deviation (SD) one. Finally, I applied a binomial regression model with LASSO penalisation to derive a discrimination-optimised FH prediction model. Specifically, optimal penalisation was determined through 15-fold cross-validation maximising the C-statistic (i.e., the area under the receiver operating characteristic (AUC-ROC) curve).[17] A first multivariable model was derived with non-genetic variables only (i.e. without LDL-C PGS), and a second model was generated with the inclusion of LDL-C PGS.

Model performance was evaluated using the 20% testing data based on its discriminative ability (C-statistic), appropriate calibration of predicted and observed probability of having an FH variant (using calibration plots, calibration-in-the-large, and calibration slope), and classification metrics (sensitivity, specificity (or its compliment the false positive rate), positive predictive value, and the negative predicted value).

### 6.3.4 Evaluating the burden of genomic sequencing for FH

While genetic sequencing is the gold standard for FH diagnosis, it may often be prohibitively expensive to offer it to an entire population as a screening strategy. I therefore explored whether prioritising people with suspected FH can reduce the screening burden with an acceptable number of false-negative results. I evaluated the following prioritisation strategies: 1) no prioritisation (i.e. referring all participants for sequencing), 2) prioritisation based on LDL-C concentration (adjusting for statin use), 3) a multivariable model built from clinical biomarkers and environmental predictors only, 4) a multivariable model built from genetic, clinical biomarkers and environmental predictors.

These prioritisation strategies were evaluated on the number of subjects that would need to be sequenced, the proportion of FH carriers who would be missed, and the ratio of FH carriers correctly prioritised by the number of non-carriers unnecessarily offered sequencing. A decision

curve analysis was performed comparing a model's net-benefit across various probability thresholds for confirmatory FH screening. Here, "net benefit" is calculated as the weighted difference between true and false positives at a specific threshold.[19] Additionally, prioritisation based on LDL-C concentrations (adjusted for statin use) was compared to prioritisation using the multivariable model with the PGS with the help of a net reclassification index (NRI) analysis. The choice of 0.006 as a probability threshold was used as an example. This choice of threshold was based on the results of the decision curve analysis (i.e. located within the plausible range of probability thresholds).

### 6.3.5 Software

The data analysis and figures were done in R version 4.0.2.[20] The R package tableone version 0.12.0. was used to make **Table 6.1**.[21] The receiver operating characteristic (ROC) curves were plotted with the R package pROC version 1.16.2.[22] The NRI analysis was done using the R package nricens version 1.6,[23] the decision curve analysis was performed using the R package dcurves version 0.3.0.[24]

## 6.4 Results

### 6.4.1 Participant characteristics of our study cohort

Using the UK Biobank WES data, I identified 488 pathogenic or likely pathogenic FH variant carriers (list of variants shown in **Chapter 5 Supplementary Table 5.1**) and 139,291 non-carriers. FH variant carriers had a significantly higher frequency of a family history of CHD (62.7% versus 48.1% in controls), higher prevalence (8.2% versus 2.8% in controls) and incidence (6.6% versus 3.9% in controls) of CHD (**Table 6.1**).

| | FH-variant negative | FH-variant positive | p-value | Missing (%) |
|---|---|---|---|---|
| n | 139291 | 488 | | |
| Sex (male) (%) | 63382 (45.5) | 207 (42.4) | 0.187 | 0.0 |
| Age (median [IQR]) | 58.0 [51.0, 63.0] | 58.0 [51.0, 63.0] | 0.803 | 0.0 |
| Townsend deprivation index (median [IQR]) | -2.4 [-3.8, 0.0] | -2.2 [-3.7, 0.2] | 0.346 | 0.1 |
| BMI, kg/m2 (median [IQR]) | 26.7 [24.1, 29.8] | 27.1 [23.9, 29.8] | 0.689 | 0.3 |
| Smoking status (%) | | | 0.685 | 3.7 |
| Non-smoker | 76862 (57.3) | 262 (56.2) | | |

| | | | | |
|---|---|---|---|---|
| Former smoker | 49302 (36.7) | 171 (36.7) | | |
| Light smoker (<10 cigarettes/day) | 1952 (1.5) | 6 (1.3) | | |
| Moderate smoker (10-19 cigarettes/day) | 3296 (2.5) | 13 (2.8) | | |
| Heavy Smoker (>20 cigarettes/day) | 2796 (2.1) | 14 (3.0) | | |
| Alcohol use (%) | | | 0.492 | 0.0 |
| Prefer not to answer | 88 (0.1) | 1 (0.2) | | |
| 1/day | 29719 (21.3) | 93 (19.1) | | |
| 3-4 times/week | 34015 (24.4) | 135 (27.7) | | |
| 1-2 times/week | 36823 (26.4) | 130 (26.6) | | |
| 1-3 times/month | 15498 (11.1) | 54 (11.1) | | |
| Special occasions | 14383 (10.3) | 45 (9.2) | | |
| Never | 8765 (6.3) | 30 (6.1) | | |
| Family history of CHD (%) | 67013 (48.1) | 306 (62.7) | <0.001 | 0.0 |
| Systolic blood pressure, mmHg (median [IQR]) | 136.5 [125.0, 149.5] | 135.0 [124.5, 148.5] | 0.119 | 0.2 |
| Diastolic blood pressure, mmHg (median [IQR]) | 82.0 [75.0, 89.0] | 81.0 [74.0, 87.0] | 0.024 | 0.2 |
| Statin use (%) | 18139 (13.0) | 165 (33.8) | <0.001 | 0.0 |
| Hypertension (median [IQR]) | 7946 (5.7) | 35 (7.2) | 0.195 | 0.0 |
| LDL-C PGS (median [IQR]) | 3.7 [3.5, 3.9] | 3.7 [3.5, 3.9] | 0.652 | 0.0 |
| **Biomarkers** | | | | |
| LDL-C (unadjusted for statin use), mmol/L (median [IQR]) | 3.5 [3.0, 4.1] | 3.9 [3.2, 4.9] | <0.001 | 5.0 |
| HDL-C, mmol/L (median [IQR]) | 1.4 [1.2, 1.7] | 1.4 [1.2, 1.6] | 0.086 | 12.5 |
| Total cholesterol, mmol/L (median [IQR]) | 5.7 [4.9, 6.4] | 6.1 [5.2, 7.3] | <0.001 | 4.8 |
| Lipoprotein(a), nmol/L (median [IQR]) | 20.0 [9.3, 59.8] | 27.6 [10.3, 59.2] | 0.083 | 24.3 |
| Apolipoprotein A1, g/L (median [IQR]) | 1.5 [1.4, 1.7] | 1.5 [1.3, 1.6] | <0.001 | 13.0 |
| Apolipoprotein B, g/L (median [IQR]) | 1.0 [0.9, 1.2] | 1.2 [1.0, 1.4] | <0.001 | 5.3 |
| Triglycerides, mmol/L (median [IQR]) | 1.5 [1.1, 2.2] | 1.3 [0.9, 1.9] | <0.001 | 4.9 |
| C-reactive protein, mg/L (median [IQR]) | 1.3 [0.7, 2.7] | 1.2 [0.6, 2.3] | 0.065 | 5.1 |
| Aspartate aminotransferase, um (median [IQR]) | 24.4 [21.0, 28.8] | 25.1 [21.0, 29.6] | 0.111 | 5.2 |
| Alanine aminotransferase, um (median [IQR]) | 20.1 [15.4, 27.3] | 20.2 [15.6, 27.2] | 0.848 | 4.9 |
| Alkaline phosphatase, um (median [IQR]) | 80.1 [67.1, 95.4] | 80.6 [66.8, 96.1] | 0.506 | 4.8 |
| **Disease prevalence & incidence** | | | | |
| CHD prevalence (%) | 3890 (2.8) | 40 (8.2) | <0.001 | 0.0 |
| CHD incidence (%) | 5370 (3.9) | 32 (6.6) | 0.003 | 0.0 |
| CVD prevalence (%) | 5686 (4.1) | 45 (9.2) | <0.001 | 0.0 |
| CVD incidence (%) | 9038 (6.5) | 46 (9.4) | 0.011 | 0.0 |
| Type 2 diabetes prevalence (%) | 3593 (2.6) | 11 (2.3) | 0.757 | 0.0 |
| Type 2 diabetes incidence (%) | 4948 (3.6) | 19 (3.9) | 0.776 | 0.0 |

**Table 6.1 UK Biobank participant characteristics stratified by carrying a FH-causing variant.** The p-values shown in the table are from the Kruskal-Wallis Rank Sum test for continuous variables, and from the Man-Whitney U test for binary variables. BMI = body mass index; CHD = coronary heart disease; CVD = cardiovascular disease; FH = familial hypercholesterolaemia; HDL-C = high-density lipoprotein cholesterol; IQR = interquartile range; LDL-C = low-density lipoprotein cholesterol; PGS = polygenic score.

### 6.4.2 LDL-C PGS

The p-value cut-off of $5x10^{-4}$ for the variants of GLGC's LDL-C GWAS summary statistics included 10,137 variants. After running the LASSO regression on these variants using the LASSO training dataset described in the methods section, 1,466 genetic variants were retained by the model. The LDL-C PGS r-squared was 0.14 (95% CI: 0.13-0.15) in the independent test data.

### 6.4.3 Multivariable machine learning model to prioritise FH variant carriers

Nine non-genetic variables were retained by the LASSO regression model which did not include the LDL-C PGS (**Supplementary Table 6.2**). These predictors were age, statin use, SBP, DBP, Apo-A1, and triglyceride concentrations, family history of CHD, and two interaction terms: LDL-$C^2$, and statin use and LDL-C. The test data AUC for this model was 0.76 (95% CI: 0.71; 0.82).

14 out of the 32 variables were retained by the LASSO regression model which included a LDL-C PGS for the prediction of FH (**Figure 6.2.a**, **Supplementary Figure 6.2**, and **Supplementary Table 6.3**): triglyceride, Apo-A1, ALT and CRP concentrations, statin use, LDL-C PGS, family history of CHD, DBP, BMI, and prevalent T2D. Additionally, the following product terms were selected: LDL-$C^2$, statin use and LDL-C, age and LDL-C PGS. The test data AUC for this model was comparable but superior to the previous model: 0.77 (95% CI: 0.71; 0.83), with a training data AUC of 0.78 (95% CI: 0.75; 0.81). Calibration statistics (calibration-in-the-large: -0.073 (95% CI: -0.28; 0.13) and calibration slope: 1.02 (95% CI: 0.85; 1.19)) indicated the predicted probability agreed well with the observed probability (**Figure 6.3.a**). The median predicted probability of having monogenic FH by this multivariable model was around three folds higher in FH carriers (0.64%, interquartile range (IQR): 0.31; 1.62) compared to non-carriers (0.23%, IQR: 0.14; 0.38), with partial overlap between FH carriers and non-carriers (**Figure 6.2.b**).

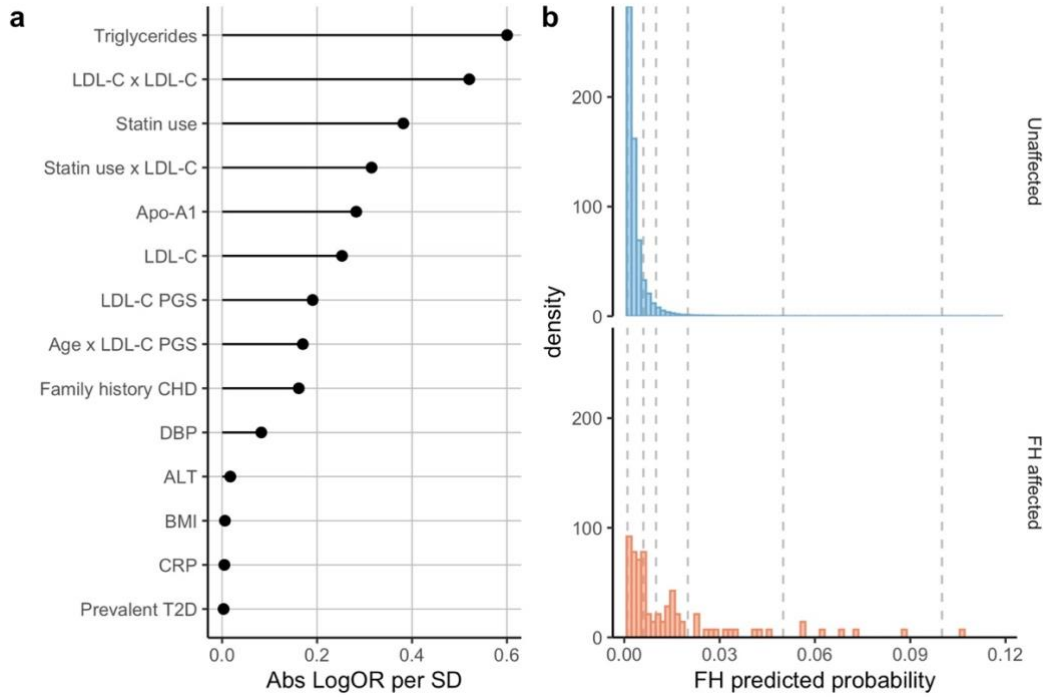



**Figure 6.2 Feature importance of the variables retained by LASSO regression predicting monogenic FH, and the density predicted probability distributions from this model for unaffected and affected FH individuals in White British participants of the UK Biobank.** **a)** The 14 predictors retained by LASSO regression ordered by absolute log odds ratio (OR) per standard deviation (SD). The "x" sign is used to indicate an interaction term. Abs = absolute; ALT = alanine aminotransferase; Apo-A1 = apolipoprotein A1; BMI = body mass index; CHD = coronary heart disease; CRP = C-reactive protein; DBP = diastolic blood pressure; LDL-C = low-density lipoprotein cholesterol; PGS = polygenic score; T2D = type 2 diabetes. **b)** The density predicted probability distributions for affected (in orange) and unaffected (in blue) familial hypercholesterolaemia (FH) participants in our test cohort as predicted by the multivariable model. 14 unaffected individuals had a monogenic FH predicted probability above 0.12 and are not shown on the plot for legibility purposes. The vertical dotted lines represent the various FH predicted probability thresholds evaluated in **Table 6.2**.

Both multivariable machine learning models outperformed the model which only considered LDL-C (AUC: 0.62, 95% CI: 0.56; 0.68), as well as the model which corrected for statin use (AUC: 0.71, 95% CI: 0.65; 0.77) (**Figure 6.3.b**).

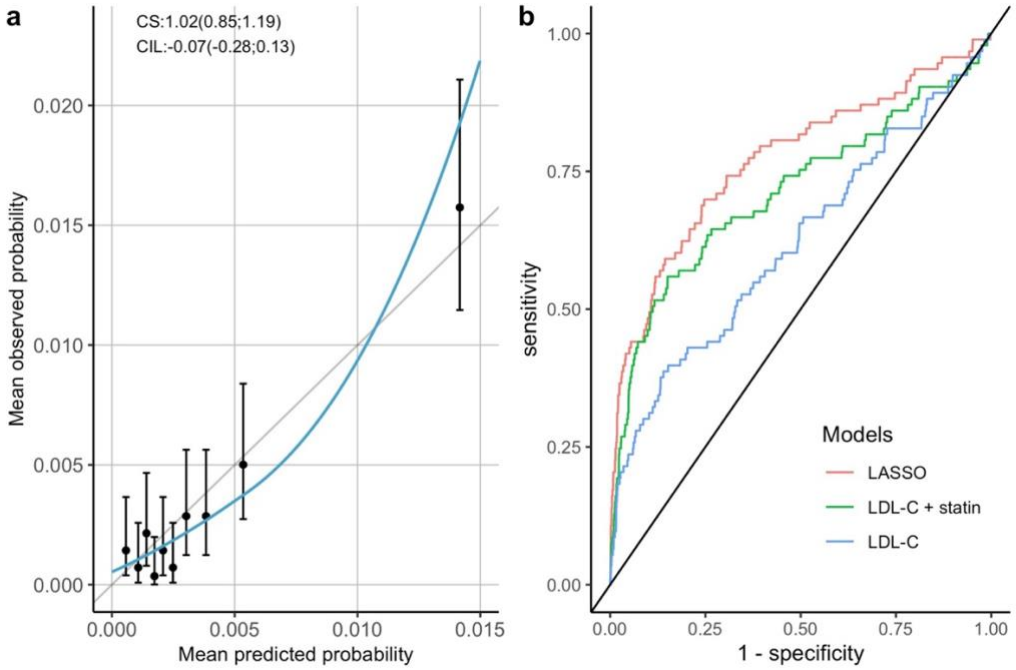



**Figure 6.3 Discrimination and calibration of a multivariable algorithm including LDL-C PGS predicting FH carriership using independent testing data. a)** The calibration plot for the multivariable model where the mean predicted and mean observed probability for each decile of the test data are depicted by the datapoints with their 95% confidence intervals (CI). Perfect calibration is indicated by the vertical black line. The calibration-in-the-large (CIL) and the calibration slope (CS) values are indicated on the plot with their 95% CI in brackets. The loess line was fitted with FH-causing variant status as the outcome and mean predicted probability as the predictor. **b)** The receiver operating characteristic (ROC) curves for the multivariable model with LDL-C PGS (in red), LDL-C concentration and statin model (in green), and LDL-C concentration only model (in blue). The area under the curve (AUC) for each of these models are equal to 0.77 (95% CI: 0.71; 0.83), 0.71 (95% CI: 0.65; 0.77) and 0.62 (95% CI: 0.56; 0.68) respectively. LDL-C = low-density lipoprotein cholesterol; PGS = polygenic score.

### 6.4.1 Evaluating the FH screening strategies through decision curve analysis

We next determined at which probability threshold the net benefit of the various models was larger than the "sequence all" strategy (**Figure 6.4**). The net benefit of the "sequence all" strategy was lower than that of the other models tested at a threshold of 0.0013 (0.13%). This implies that model-based prioritisation for confirmatory FH sequencing is more beneficial if one decided to screen 1/0.0013 = 769 or more people to detected one FH case. Irrespective of the probability threshold, the multivariable machine learning models had a larger net benefit than the LDL-C adjusted for statin use model. At a threshold of 0.0050 (0.5%), the multivariable model with the LDL-C PGS had the largest net benefit out of all the models tested (**Figure 6.4**).

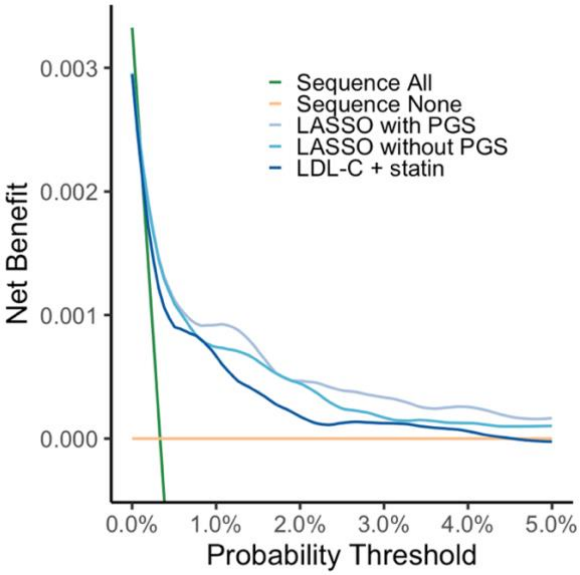



**Figure 6.4 Decision curve analysis of the FH prediction models.** The highest curve indicates the highest net benefit which considers the benefits and harms of a model. "Sequence all" refers to screening and sequencing the entire population, while "sequence none" refers to no FH screening. The "LDL-C + statin" model is a model based on LDL-C concentration adjusted for statin use. The LASSO models are the multivariable machine learning models that either included or excluded the LDL-C PGS.

### 6.4.2 Model FH classification

Next, I evaluated the performance of FH classification of the multivariable model with the largest net benefit (i.e. the model including the LDL-C PGS) (**Figure 6.4**) using six probability thresholds of having an FH variant (from 0.001 to 0.10) in the test dataset. The sensitivity increased from 1.1% (95% CI: 0.2; 5.8) for a predicted probability of 0.10, to 94.6% (95% CI: 88.0; 97.7) for a predicted probability of 0.001; with the false positive rate similarly increasing from 0.1% (95% CI: 0.0; 0.1) to 87.0% (95% CI: 86.6; 87.4) (**Table 6.2**). I further compared the performance of these thresholds to a simpler model of LDL-C concentration adjusted for statin, which underperformed (**Table 6.2**).

| **Predicted probability cut-off** | % sensitivity (95%CI) | % false positive rate (95%CI) | % positive predictive value (95%CI) | % negative predictive value (95%CI) | FH-causing variants below threshold | FH-causing variants above threshold | Controls above threshold |
|---|---|---|---|---|---|---|---|
| **Multivariable model** | | | | | | | |
| **0.1** | 1.1 (0.2;5.8) | 0.1 (0.0;0.1) | 5.6 (1.0;25.8) | 99.7 (99.6;99.7) | 92 | 1 | 17 |
| **0.05** | 7.5 (3.7;14.7) | 0.2 (0.1;0.2) | 13.0 (6.4;24.4) | 99.7 (99.6;99.8) | 86 | 7 | 47 |
| **0.02** | 20.4 (13.5;29.7) | 1.1 (0.9;1.2) | 6.0 (3.9;9.2) | 99.7 (99.7;99.8) | 74 | 19 | 296 |
| **0.01** | 41.9 (32.4;52.1) | 4.5 (4.2;4.7) | 3.0 (2.2;4.1) | 99.8 (99.7;99.8) | 54 | 39 | 1244 |
| **0.006** | 54.8 (44.7;64.6) | 11.9 (11.5;12.3) | 1.5 (1.2;2.0) | 99.8 (99.8;99.9) | 42 | 51 | 3311 |
| **0.001** | 94.6 (88.0;97.7) | 87.0 (86.6;87.4) | 0.4 (0.3;0.4) | 99.9 (99.7;99.9) | 5 | 88 | 24240 |
| **Model: LDL-C concentration + statin use** | | | | | | | |
| **0.1** | 0.0 (0.0;4.0) | 0.0 (0.0;0.1) | 0.0 (0.0;35.4) | 99.7 (99.6;99.7) | 93 | 0 | 7 |
| **0.05** | 1.1 (0.2;5.8) | 0.1 (0.1;0.2) | 3.2 (0.6;16.2) | 99.7 (99.6;99.7) | 92 | 1 | 30 |
| **0.02** | 12.9 (7.5;21.2) | 1.1 (1.0;1.2) | 3.8 (2.2;6.5) | 99.7 (99.6;99.8) | 81 | 12 | 304 |
| **0.01** | 38.7 (29.4;48.9) | 5.6 (5.4;5.9) | 2.2 (1.6;3.1) | 99.8 (99.7;99.8) | 57 | 36 | 1574 |
| **0.006** | 52.7 (42.6;62.5) | 14.6 (14.2;15.0) | 1.2 (0.9;1.6) | 99.8 (99.8;99.9) | 44 | 49 | 4067 |
| **0.001** | 90.3 (82.6;94.8) | 84.0 (83.5;84.4) | 0.4 (0.3;0.4) | 99.8 (99.6;99.9) | 9 | 84 | 23393 |

**Table 6.2 The classification accuracy of an algorithm for predicting monogenic FH using the multivariable model and LDL-C concentration accounting for statin use.** There are 93 FH-causing variant positive participants in the test data comprising of a total of 27,955 participants. CI = confidence interval; FH = familial hypercholesterolaemia; LDL-C = low-density lipoprotein cholesterol.

The net reclassification index (NRI) comparing the LDL-C and statin use model to the multivariable model, indicated that the improved performance of the latter was due to it assigning a higher predicted probability to FH variant carriers. At a predicted probability threshold of 0.006, the probability for FH carriers being reclassified as having an FH variant was equal to 0.097 (95% CI: 0.038; 0.159), as opposed to the probability of 0.075 (95% CI: 0.026; 0.130) of being down-classified as not having an FH variant (**Table 6.3**).

| LDL-C + statin use model | Multivariable model<br>**< 0.006 predicted probability threshold** | **>= 0.006 predicted probability threshold** | **Total** |
|---|---|---|---|
| **< 0.006 predicted probability threshold** | 22,846 | 993 | 23,839 |
| **>= 0.006 predicted probability threshold** | 1,747 | 2,369 | 4,116 |
| **Total** | 24,593 | 3,362 | 27,955 |
| **NRI estimates** | | | |
| **NRI:** | 0.049 (-0.037; 0.131) | | |
| **Event NRI:** | 0.022 (-0.063; 0.104) | | |
| **Non-event NRI:** | 0.027 (0.023; 0.031) | | |
| **Pr(Up \| Case)** | 0.097 (0.038; 0.159) | | |
| **Pr(Down \| Case)** | 0.075 (0.026; 0.130) | | |
| **Pr(Down \| Ctrl)** | 0.062 (0.060; 0.065) | | |
| **Pr(Up \| Ctrl)** | 0.035 (0.033; 0.037) | | |

**Table 6.3 NRI table and estimates for a predicted probability threshold of 0.006 for FH comparing the multivariable model with LDL-C PGS to a simpler model of LDL-C concentration and statin use.** The predicted probability threshold of 0.006 was chosen to illustrate the NRI analysis between the multivariable model and the LDL-C with statin use model. The test dataset of 27,955 participants was used, which included 93 FH variant carriers. NRI estimates were obtained via percentile bootstrap method. Case = positive for an FH-causing variant; Ctrl = control (negative for an FH-causing variant); Down = reclassified to the lower category; FH = familial hypercholesterolaemia; LDL-C = low-density lipoprotein cholesterol; NRI = net reclassification index; Pr = probability; Up = reclassified to the higher category.

### 6.4.3 Prioritising individuals for FH genomic testing in a two-stage population screening strategy

Finally, I evaluated the performance of a two-stage population screen for identifying new index FH cases, where the second stage consisted of targeted sequencing of FH variants (**Figure 6.5**). The multivariable (with LDL-C PGS) and LDL-C with statin use models were compared using a common threshold of 0.006, which falls within the plausible range found using the decision curve analysis (**Figure 6.4**). On average, seven additional FH carriers would be detected for 100,000 individuals screened when using the multivariable model compared to the LDL-C and statin use model. Per 100,000 individuals screened, the multivariable model with LDL-C PGS would refer 12,033 individuals (12%) for genetic sequencing, compared to 14,730 (15%) with the LDL-C and statin use model, resulting in a 18% reduction in genetic testing at this specific threshold.

Furthermore, if I assume that FH has a population prevalence of 1 in 286 (equal to our cohort's prevalence) and that one FH case has on average 1.5 first-degree relatives ((2 children + 1 sibling) / 2) who are also affected by FH (discovered through cascade testing),[25] then overall one FH case would be identified for every ~219 people screened when using the multivariable model with LDL-C PGS, compared to one FH case for every ~228 individuals screened with the LDL-C and statin use model.

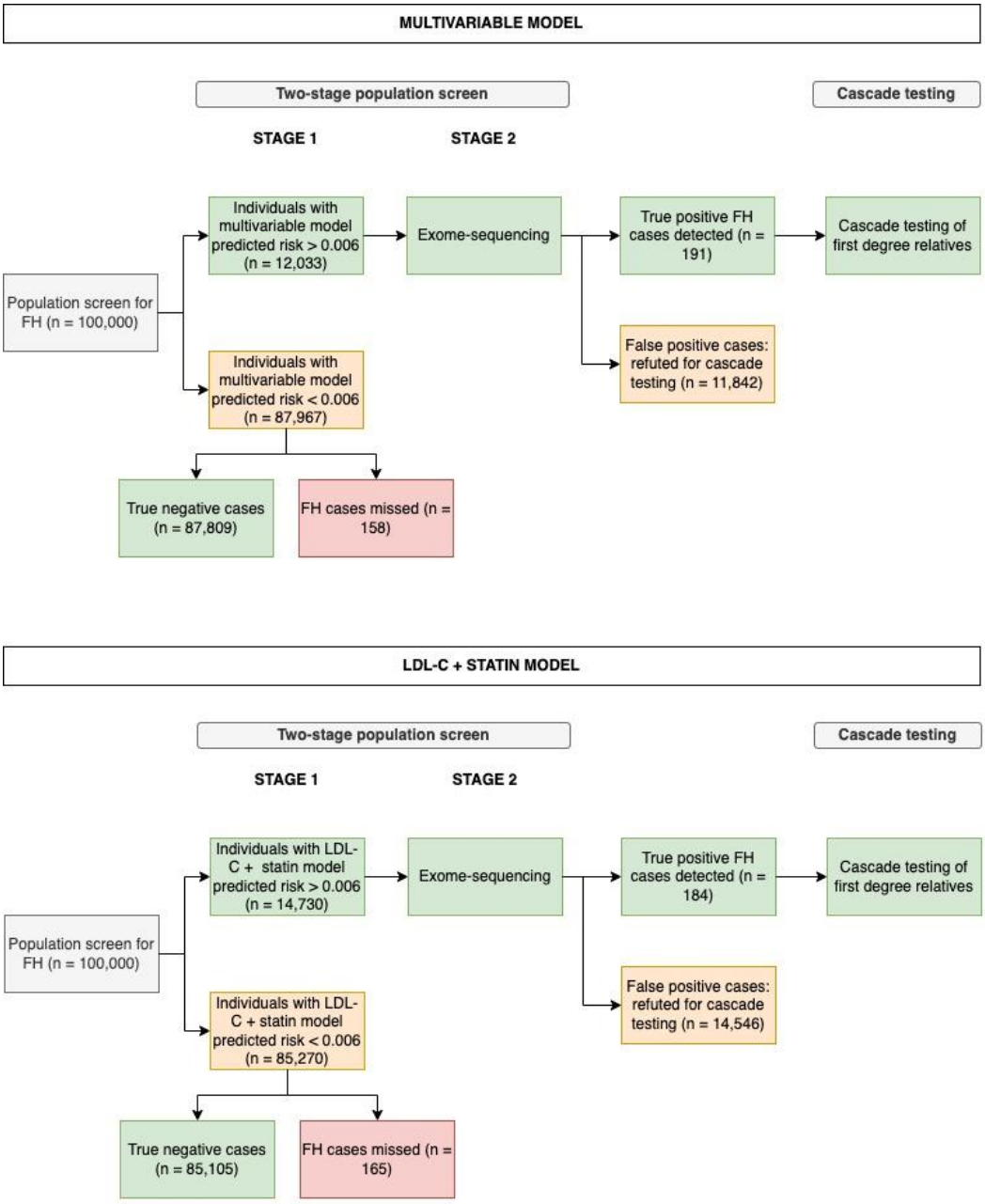



**Figure 6.5 Adult two-stage population screening strategy for monogenic FH.** Stage 1 screen identifies individuals with a predicted probability by the LASSO model (with LDL-C PGS) above a pre-specified threshold value, followed by a second stage of exome-sequencing. FH cases detected following this two-stage screen are brought forward for cascade testing of first-degree relatives. The sensitivity and false positive rate of the first stage depends on the threshold value chosen for the model. I assume perfect discrimination in the second stage of exome-sequencing (sensitivity of 100% and false positive rate of 0%). Cascade testing is expected to yield 1.5 additional FH cases detected for every FH case identified through the two-stage population screening strategy, as described in the results section. FH = familial hypercholesterolaemia; LASSO = least absolute shrinkage and selection operator; LDL-C = low-density lipoprotein cholesterol; PGS = polygenic score.

## 6.5 Discussion

In the current chapter I derived a multivariable machine learning model to identify people with suspected FH for confirmatory DNA sequencing in the context of population screening. Using LASSO regression, I derived a 14-feature model consisting of LDL-C, Apo-A1, triglyceride, ALT, and CRP concentrations, self-reported statin use, family history of CHD, DBP, BMI, type 2 diabetes diagnosis, three product terms, and an LDL-C PGS. The multivariable algorithm was able to discriminate between FH and non-FH carriers with an AUC of 0.77 (95% CI: 0.71; 0.83), with good calibration, outperforming a simpler model consisting of LDL-C and an indicator for statin prescription, and a multivariable model without LDL-C PGS as a predictor.

Above a classification threshold of 0.0013 (0.13%), the multivariable algorithm that contained the LDL-C PGS showed the highest net benefit out of all the models tested (**Figure 6.4**), and was able to decrease the number of subjects referred to genetic sequencing (e.g. from 100,000 individuals without any prioritisation, to 14,730 with prioritisation using the LDL-C and statin use model, and to 12,033 with prioritisation using the multivariable model for a predicted probability threshold of carrying a variant for monogenic FH of 0.006; equivalent to approximately a 18% decrease in individuals needed to be sequenced between the last two models (**Figure 6.5**)). These differences become more significant if extrapolating the values to a population-wide scale comprising of millions of participants screened. Our results provide support for opportunistic screening and seeding of cascade testing for FH, which could be integrated within existing health checks offered to employers or local healthcare providers (e.g. NHS Health Checks).[12]

Previously, Banda *et al.* used a machine learning method to detect monogenic FH cases from electronic health records.[26] While their model showed an impressive AUC of 0.94, one of their most important features was referral to a cardiology clinic, which is in very close proximity to confirmatory FH testing, limiting the model's utility as a prospective tool for FH diagnosis. Besseling *et al.* developed a multivariable model to identify FH carriers validated in study participants consisting of FH cases and their relatives, again limiting applicability to the general population.[27] Our model instead considers FH prioritisation in a non-GP-referred population and is more generalisable as a systematic population screening tool.

Our multivariable model included three terms for LDL-C (LDL-C itself, LDL-C squared, and an interaction with statin prescription), which combined makes it the most important predictor.

Additionally, our model also identified novel predictors for FH such as triglyceride and Apo-A1 concentrations, with triglycerides having the largest absolute OR per SD (0.60). In this study, I find that FH carriers had significantly lower triglyceride concentrations than non-carriers (**Table 6.1**), which resulted in a negative association, indicating that triglyceride concentrations can be useful in discriminating between individuals who have hypercholesterolaemia due to lifestyle factors or other causes (e.g. combined hyperlipidaemia) as opposed to an FH-causing variant. I also found that higher Apo-A1 concentrations, a protein found on high-density lipoprotein (HDL) particles, was associated with a decreased probability of FH.

The variables included in our multivariable algorithm should not be interpreted as causal risk factors for monogenic FH; they simply help to distinguish non-monogenetic sources of variation in LDL-C concentrations from monogenic causes (as was discussed in more detail previously with triglyceride concentrations). This also provides the rational for including an LDL-C PGS in the model: a large discrepancy between predicted LDL-C concentrations (by the LDL-C PGS) and observed LDL-C concentrations might be indicative of FH carriership,[14,15] demonstrated here by a negative coefficient for LDL-C PGS in the model (**Supplementary Table 6.3**). I note that a previous LDL-C PGS by Wu *et al.* had a substantially larger R-squared (0.21 (95% CI: 0.20-0.22)) than reported here (0.14 (95% CI: 0.13-0.15)).[32] Unlike Wu *et al.* who identified genetic variants from an internal UK Biobank LDL-C GWAS overlapping with the PGS training data; I identified variants based on an independent dataset from GLGC,[16] guarding against overfitting through 'data-leakage' between the training and testing datasets and providing a more robust estimate of explained variance. Currently, PGS information is not routinely used or collected in clinical practice, which is why I also derived a multivariable model without LDL-C PGS, which did not meaningfully differ (**Supplementary Table 6.2**). Previous studies have suggested that PGS could be used to identify individuals with a rare variant for certain diseases, such as FH.[14,15] Our study confirms the utility of the PGS for FH prioritisation; however, given its correlation with environmental variables (e.g. lipid levels), this genetic information can be readily replaced with information from non-genetic data.

A study limitation to consider is the exclusion of individuals with VUS from our study cohort. There is conflicting evidence as to the causal effects of these VUS in FH. I anticipate that some are likely to be FH-causing while others are not, but more research is needed. As more VUS are classified as either FH-causing or not, the model can be readily updated to reflect our growing understanding of FH. Additionally, it is impossible to know whether some study participants have

been genetically tested for carrying an FH variant, and whether they might have modified their behaviour (e.g. diet) following their diagnosis. This could potentially impact the accuracy of the multivariable model developed here; however, considering that only approximately 7% of FH cases have been diagnosed in the UK,[33] this low number of diagnoses is unlikely to have a significant effect on the model and results presented here.

I have tested our multivariable model in a dataset which was independent from the training data, with no significant difference between training and testing AUC (difference of 0.01), suggesting limited model overfitting to the current sample. Nevertheless, considering the health discrepancies observed between the UK Biobank and the general UK population,[34] I suggest that this model is locally validated and updated before applying it to distinct settings. Model validation should especially be conducted when considering populations of non-European ancestry. Irrespective of the important considerations regarding model transferability, prior to integrating the model in clinical care, an informed decision should be made on the optimal predicted probability threshold for monogenic FH classification. I wish to highlight that the choice of 0.006 as a threshold in **Table 6.3** and **Figure 6.5** is simply an illustration, and depending on the available healthcare resources, a different threshold might be preferred (**Figure 6.4**).

In conclusion, I derived a multivariable classification model for detecting monogenic FH variant carriers that outperformed a model based on LDL-C concentration (adjusted for statin use) for FH screening, and that offers an opportunity to prioritise suspected FH carriers for genetic sequencing.

## 6.6 References


1 Gratton J, Futema M, Humphries SE, *et al.* A machine learning model to aid detection of familial hypercholesterolaemia. *medRxiv* 2022;:2022.06.17.22276540. doi:10.1101/2022.06.17.22276540

2 McGowan MP, Hosseini Dehkordi SH, Moriarty PM, *et al.* Diagnosis and treatment of heterozygous familial hypercholesterolemia. *J Am Heart Assoc* 2019;**8**. doi:10.1161/JAHA.119.013225

3 Akioyamen LE, Genest J, Shan SD, *et al.* Estimating the prevalence of heterozygous familial hypercholesterolaemia: a systematic review and meta-analysis. *BMJ Open* 2017;**7**. doi:10.1136/BMJOPEN-2017-016461

4 Vallejo-Vaz AJ, Stevens CAT, Lyons ARM, *et al.* Global perspective of familial hypercholesterolaemia: a cross-sectional study from the EAS Familial Hypercholesterolaemia Studies Collaboration (FHSC). *The Lancet* 2021;**398**:1713–25. doi:10.1016/S0140-6736(21)01122-3/ATTACHMENT/9B02EB98-E580-4B03-84A0-777C13E004A9/MMC1.PDF

5 Tromp TR, Hartgers ML, Hovingh GK, *et al.* Worldwide experience of homozygous familial hypercholesterolaemia: retrospective cohort study. *The Lancet* 2022;**399**:719–28. doi:10.1016/S0140-6736(21)02001-8/ATTACHMENT/EDA26D05-0227-4BA4-9063-70E486D14264/MMC1.PDF

6 Kerr M, Pears R, Miedzybrodzka Z, *et al.* Cost effectiveness of cascade testing for familial hypercholesterolaemia, based on data from familial hypercholesterolaemia services in the UK. *Eur Heart J* 2017;**38**:1832–9. doi:10.1093/EURHEARTJ/EHX111

7 Marquina C, Lacaze P, Tiller J, *et al.* Population genomic screening of young adults for familial hypercholesterolaemia: a cost-effectiveness analysis. *Eur Heart J* Published Online First: 11 November 2021. doi:10.1093/EURHEARTJ/EHAB770

8 Jackson CL, Huschka T, Borah B, *et al.* Cost-effectiveness of cascade genetic testing for familial hypercholesterolemia in the United States: A simulation analysis. *Am J Prev Cardiol* 2021;**8**:100245. doi:10.1016/J.AJPC.2021.100245

9 Lázaro P, Pérez de Isla L, Watts GF, *et al.* Cost-effectiveness of a cascade screening program for the early detection of familial hypercholesterolemia. *J Clin Lipidol* 2017;**11**:260–71. doi:10.1016/J.JACL.2017.01.002

10 Wald DS, Bestwick JP, Morris JK, *et al.* Child–Parent Familial Hypercholesterolemia Screening in Primary Care. *New England Journal of Medicine* 2016;**375**:1628–37. doi:10.1056/NEJMOA1602777

11 Futema M, Cooper JA, Charakida M, *et al.* Screening for familial hypercholesterolaemia in childhood: Avon Longitudinal Study of Parents and Children (ALSPAC). *Atherosclerosis* 2017;**260**:47. doi:10.1016/J.ATHEROSCLEROSIS.2017.03.007

12 Duddy C, Wong G, Gadsby EW, *et al.* NHS Health Check programme: a protocol for a realist review. *BMJ Open* 2021;**11**:e048937. doi:10.1136/BMJOPEN-2021-048937

13 Talmud PJ, Drenos F, Shah S, *et al.* Gene-centric Association Signals for Lipids and Apolipoproteins Identified via the HumanCVD BeadChip. *Am J Hum Genet* Published Online First: 2009. doi:10.1016/j.ajhg.2009.10.014

14 Zhou D, Yu D, Scharf JM, *et al.* Contextualizing genetic risk score for disease screening and rare variant discovery. *Nat Commun* 2021;**12**:1–14. doi:10.1038/s41467-021-24387-z

15 Lu T, Forgetta V, Richards JB, *et al.* Polygenic risk score as a possible tool for identifying familial monogenic causes of complex diseases. *Genetics in Medicine* 2022;**0**:1–11. doi:10.1016/J.GIM.2022.03.022

16 Willer CJ, Schmidt EM, Sengupta S, *et al.* Discovery and refinement of loci associated with lipid levels. *Nat Genet* 2013;**45**:1274–83. doi:10.1038/ng.2797

17 Zeng Y, Breheny P. The biglasso Package: A Memory- and Computation-Efficient Solver for Lasso Model Fitting with Big Data in R. *R Journal* 2017;**12**:1–14. doi:10.32614/rj-2021-001

18 van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *J Stat Softw* 2011;**45**:1–67. doi:10.18637/jss.v045.i03

19 Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res* 2019;**3**:1–8. doi:10.1186/S41512-019-0064-7/FIGURES/3

20 R: The R Project for Statistical Computing. https://www.r-project.org/ (accessed 11 Feb 2022).

21 Panos A, Mavridis D. TableOne: an online web application and R package for summarising and visualising data. *Evid Based Ment Health* 2020;**23**:127–30. doi:10.1136/EBMENTAL-2020-300162

22 Robin X, Turck N, Hainard A, *et al.* pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011;**12**:1–8. doi:10.1186/1471-2105-12-77/TABLES/3

23 Inoue E. nricens: NRI for Risk Prediction Models with Time to Event and Binary Response Data. 2018.

24 Sjoberg DD. dcurves: Decision Curve Analysis for Model Evaluation. https://github.com/ddsjoberg/dcurves (accessed 5 Sep 2022).

25 Families and households in the UK - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/bulletins/familiesandhouseholds/2020 (accessed 1 Mar 2022).

26 Banda JM, Sarraju A, Abbasi F, *et al.* Finding missed cases of familial hypercholesterolemia in health systems using machine learning. *NPJ Digit Med* Published Online First: 2019. doi:10.1038/s41746-019-0101-5

27 Besseling J, Reitsma JB, Gaudet D, *et al.* Selection of individuals for genetic testing for familial hypercholesterolaemia: development and external validation of a prediction model

for the presence of a mutation causing familial hypercholesterolaemia. *Eur Heart J* 2017;**38**:565–73. doi:10.1093/EURHEARTJ/EHW135

28 Laufs U, Parhofer KG, Ginsberg HN, *et al.* Clinical review on triglycerides. *Eur Heart J* 2020;**41**:99–109c. doi:10.1093/EURHEARTJ/EHZ785

29 Futema M, Ramaswami U, Tichy L, *et al.* Comparison of the mutation spectrum and association with pre and post treatment lipid measures of children with heterozygous familial hypercholesterolaemia (FH) from eight European countries. *Atherosclerosis* 2021;**319**:108–17. doi:10.1016/J.ATHEROSCLEROSIS.2021.01.008

30 gavinband / bgen / wiki / bgenix — Bitbucket. https://bitbucket.org/gavinband/bgen/wiki/bgenix (accessed 20 Mar 2020).

31 Haralambos K, Whatley SD, Edwards R, *et al.* Clinical experience of scoring criteria for Familial Hypercholesterolaemia (FH) genetic testing in Wales. *Atherosclerosis* 2015;**240**:190–6. doi:10.1016/J.ATHEROSCLEROSIS.2015.03.003

32 Wu H, Forgetta V, Zhou S, *et al.* Polygenic Risk Score for Low-Density Lipoprotein Cholesterol Is Associated with Risk of Ischemic Heart Disease and Enriches for Individuals with Familial Hypercholesterolemia. *Circ Genom Precis Med* 2021;**14**:3106. doi:10.1161/CIRCGEN.120.003106

33 NHS England and NHS Improvement London » Familial Hypercholesterolemia (FH). https://www.england.nhs.uk/london/london-clinical-networks/our-networks/cardiac/familial-hypercholesterolaemia/ (accessed 9 May 2022).

34 Fry A, Littlejohns TJ, Sudlow C, *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* 2017;**186**. doi:10.1093/aje/kwx246

## 6.7 Appendix

**Supplementary Table 6.1 UK Biobank participant characteristics post imputation of missing values stratified by FH carriership.** The p-values shown in the table are from the Kruskal-Wallis Rank Sum test for continuous variables, and from the Man-Whitney U test for binary variables. BMI = body mass index; CHD = coronary heart disease; CVD = cardiovascular disease; FH = familial hypercholesterolaemia; HDL-C = high-density lipoprotein cholesterol; IQR = interquartile range; LDL-C = low-density lipoprotein cholesterol; PGS = polygenic score.

| | Mutation negative | Mutation positive | p-value of differences |
|---|---|---|---|
| n | 139291 | 488 | |
| Sex (male) (%) | 63382 (45.5) | 207 (42.4) | 0.187 |
| Age (median [IQR]) | 58.0 [51.0, 63.0] | 58.0 [51.0, 63.0] | 0.803 |

| | | | |
|---|---|---|---|
| Townsend deprivation index (median [IQR]) | -2.4 [-3.8, 0.0] | -2.2 [-3.7, 0.1] | 0.367 |
| BMI, kg/m2 (median [IQR]) | 26.7 [24.1, 29.8] | 27.1 [23.9, 29.8] | 0.647 |
| Smoking status (%) | | | 0.827 |
| Non-smoker | 79618 (57.2) | 281 (57.6) | |
| Former smoker | 51177 (36.7) | 173 (35.5) | |
| Light smoker (<10 cigarettes/day) | 2021 (1.5) | 7 (1.4) | |
| Moderate smoker (10-19 cigarettes/day) | 3497 (2.5) | 13 (2.7) | |
| Heavy Smoker (>20 cigarettes/day) | 2978 (2.1) | 14 (2.9) | |
| Alcohol consumption (%) | | | 0.492 |
| Prefer not to answer | 88 (0.1) | 1 (0.2) | |
| 1/day | 29719 (21.3) | 93 (19.1) | |
| 3-4 times/week | 34015 (24.4) | 135 (27.7) | |
| 1-2 times/week | 36823 (26.4) | 130 (26.6) | |
| 1-3 times/month | 15498 (11.1) | 54 (11.1) | |
| Special occasions | 14383 (10.3) | 45 (9.2) | |
| Never | 8765 (6.3) | 30 (6.1) | |
| Family history of CHD (%) | 67013 (48.1) | 306 (62.7) | <0.001 |
| Systolic blood pressure, mmHg (median [IQR]) | 136.5 [125.0, 149.5] | 135.0 [124.5, 148.5] | 0.109 |
| Diastolic blood pressure, mmHg (median [IQR]) | 82.0 [75.0, 89.0] | 81.0 [74.0, 87.0] | 0.024 |
| Hypertension (%) | 7946 (5.7) | 35 (7.2) | 0.195 |
| Statin use (%) | 18139 (13.0) | 165 (33.8) | <0.001 |
| LDL-C PGS, mmol/L (median [IQR]) | 3.7 [3.5, 3.9] | 3.7 [3.5, 3.9] | 0.652 |
| **Blood biomarkers** | | | |
| LDL-C (unadjusted for statin use), mmol/L (median [IQR]) | 3.5 [3.0, 4.1] | 3.9 [3.2, 4.8] | <0.001 |
| LDL-C (adjusted for statin use), mmol/L (median [IQR]) | 3.7 [3.1, 4.2] | 4.4 [3.7, 5.4] | <0.001 |
| HDL-C, mmol/L (median [IQR]) | 1.4 [1.2, 1.7] | 1.4 [1.2, 1.7] | 0.199 |
| Total cholesterol, mmol/L (median [IQR]) | 5.7 [4.9, 6.4] | 6.0 [5.1, 7.2] | <0.001 |
| Lipoprotein(a), nmol/L (median [IQR]) | 17.9 [9.8, 55.3] | 21.3 [12.3, 53.1] | 0.223 |
| Apolipoprotein A1, g/L (median [IQR]) | 1.5 [1.4, 1.7] | 1.5 [1.3, 1.7] | <0.001 |
| Apolipoprotein B, g/L (median [IQR]) | 1.0 [0.9, 1.2] | 1.1 [1.0, 1.4] | <0.001 |
| Triglycerides, mmol/L (median [IQR]) | 1.5 [1.1, 2.2] | 1.3 [0.9, 1.9] | <0.001 |
| C-reactive protein, mg/L (median [IQR]) | 1.3 [0.7, 2.7] | 1.2 [0.6, 2.4] | 0.045 |
| Aspartate aminotransferase, um (median [IQR]) | 24.4 [21.0, 28.8] | 25.2 [21.0, 29.5] | 0.089 |
| Alanine aminotransferase, um (median [IQR]) | 20.1 [15.4, 27.3] | 20.2 [15.6, 27.4] | 0.830 |
| Alkaline phosphatase, um (median [IQR]) | 80.1 [67.1, 95.5] | 80.6 [66.5, 95.8] | 0.571 |
| **Disease prevalence & incidence** | | | |
| CHD prevalence (%) | 3890 (2.8) | 40 (8.2) | <0.001 |
| CHD incidence (%) | 5370 (3.9) | 32 (6.6) | 0.003 |
| CVD prevalence (%) | 5686 (4.1) | 45 (9.2) | <0.001 |
| CVD incidence (%) | 9038 (6.5) | 46 (9.4) | 0.011 |
| Type 2 diabetes prevalence (%) | 3593 (2.6) | 11 (2.3) | 0.757 |
| Type 2 diabetes incidence (%) | 4948 (3.6) | 19 (3.9) | 0.776 |

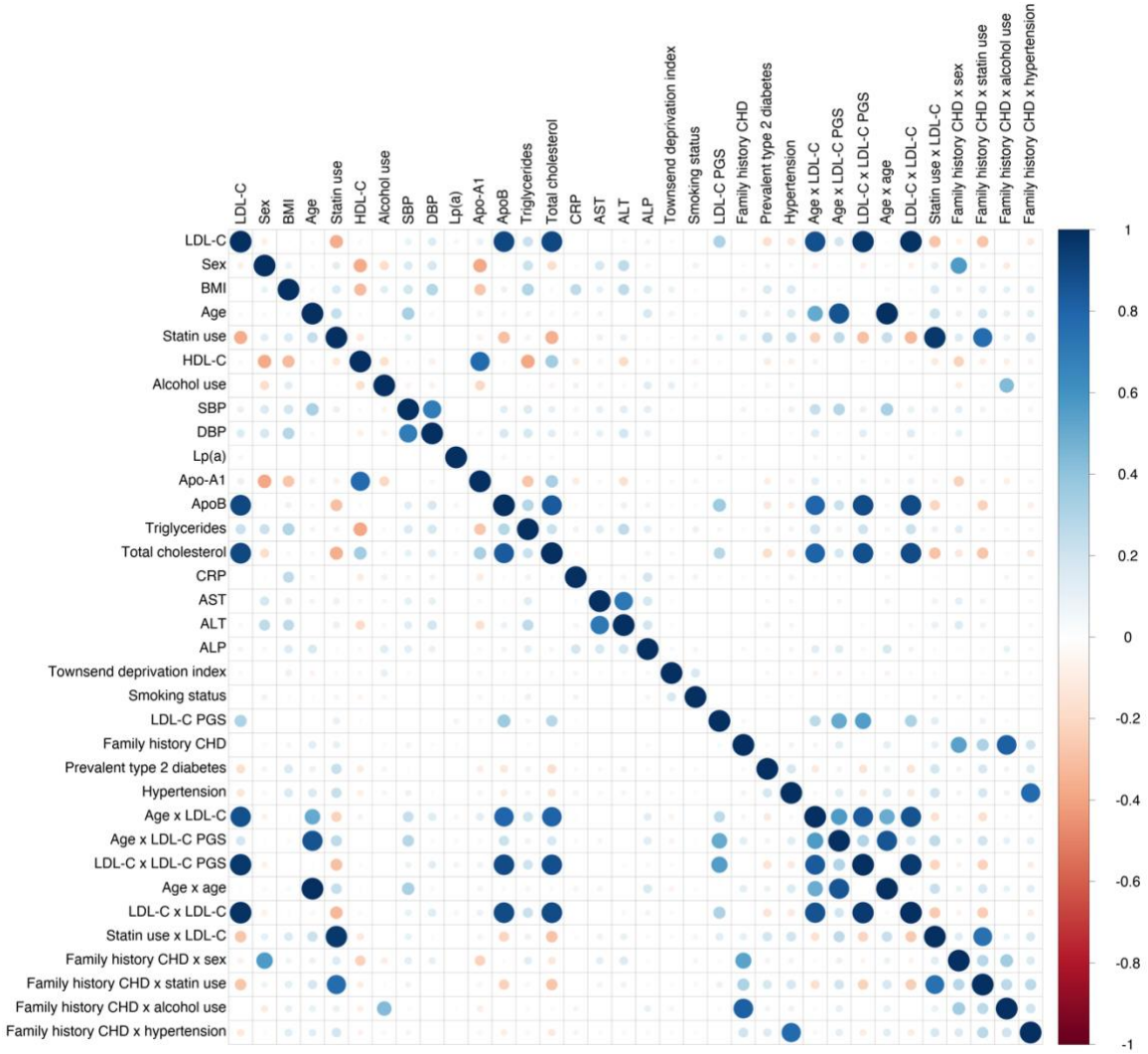
**Supplementary Table 6.2 The non-genetic variables and coefficients retained by LASSO regression for monogenic FH prediction.** The C-statistic of the independent test dataset was equal to 0.76 (95% CI: 0.71; 0.82). The variables were standardised prior to running the LASSO regression: the mean and SD are given in the table. Apo-A1 = apolipoprotein A1; CHD = coronary heart disease; CI = confidence interval; FH = familial hypercholesterolaemia; LASSO = least absolute shrinkage and selection operator; LDL-C = low-density lipoprotein cholesterol; PGS = polygenic score; SD = standard deviation.

| | Coefficients | Mean | SD |
|---|---|---|---|
| **(Intercept)** | -6.014496 | | |
| **Age** | -0.071679 | 56.86088 | 7.971519 |
| **Statin use** | 0.180699 | 0.1314255 | 0.3378712 |
| **Systolic blood pressure** | -0.005447 | 138.0973 | 18.44399 |
| **Diastolic blood pressure** | -0.060420 | 82.26036 | 10.08779 |
| **Apo-A1** | -0.258628 | 1.552397 | 0.2721509 |
| **Triglycerides** | -0.561805 | 1.738264 | 1.009821 |
| **Family history of CHD** | 0.146255 | 0.4811304 | 0.4996527 |
| **LDL-C x LDL-C** | 0.626778 | 13.53496 | 6.539555 |
| **Statin use x LDL-C** | 0.406210 | 0.3680052 | 0.9784316 |

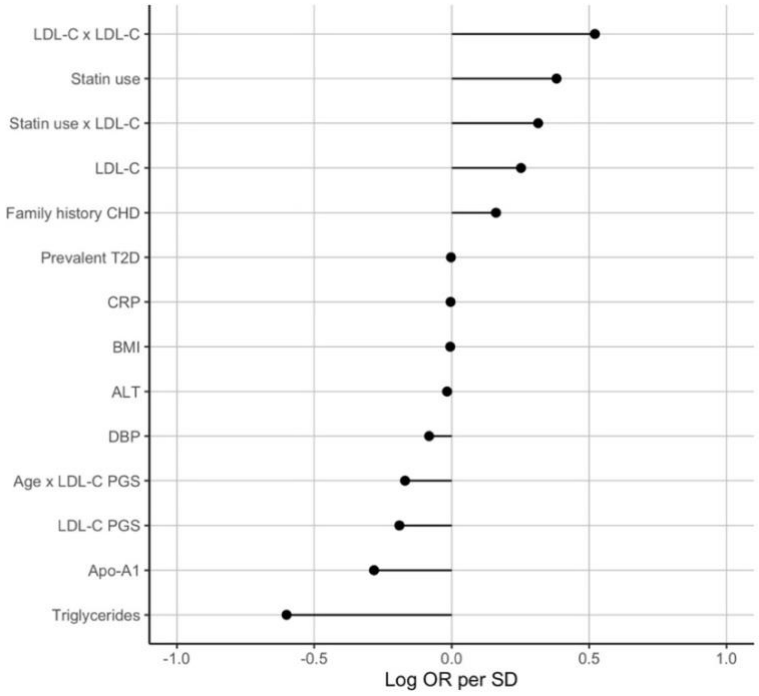
**Supplementary Table 6.3 The variables and coefficients retained by LASSO regression for monogenic FH prediction.** The variables were standardised prior to running the LASSO regression: the mean and SD are given in the table. ALT = Alanine aminotransferase; Apo-A1 = apolipoprotein A1; BMI = body mass index; CHD = coronary heart disease; CRP = C-reactive protein; FH = familial hypercholesterolaemia; HDL-C = high-density lipoprotein cholesterol; LASSO = least absolute shrinkage and selection operator; LDL-C = low-density lipoprotein cholesterol; PGS = polygenic score; SD = standard deviation.

| | Coefficients | Mean | SD |
|---|---|---|---|
| **(Intercept)** | -6.061379 | | |
| **LDL-C** | 0.252485 | 3.575719 | 0.8655747 |
| **BMI** | -0.005819 | 27.30579 | 4.653879 |
| **Statin use** | 0.381582 | 0.1314255 | 0.3378712 |
| **Diastolic blood pressure** | -0.082472 | 82.26036 | 10.08779 |
| **Apo-A1** | -0.282415 | 1.552397 | 0.2721509 |
| **Triglycerides** | -0.600269 | 1.738264 | 1.009821 |
| **CRP** | -0.004564 | 2.521555 | 4.347809 |
| **ALT** | -0.017213 | 23.38488 | 13.94569 |
| **LDL-C PGS** | -0.190587 | 3.706099 | 0.3052653 |
| **Family history of CHD** | 0.161401 | 0.4811304 | 0.4996527 |
| **Prevalent type 2 diabetes** | -0.002958 | 0.02464675 | 0.1550489 |
| **Age x LDL-C PGS** | -0.169897 | 210.7329 | 34.35274 |
| **LDL-C x LDL-C** | 0.520575 | 13.53496 | 6.539555 |
| **Statin use x LDL-C** | 0.314738 | 0.3680052 | 0.9784316 |

**Supplementary Figure 6.1 Correlation plot of the variables tested in the LASSO regression model for the prediction of monogenic FH.** The data shown here is from the training dataset as this was used to evaluated highly correlated variables prior to running LASSO regression. ALP = alkaline phosphatase; ALT = alanine transaminase; Apo-A1 = apolipoprotein A1; ApoB = apolipoprotein B; AST = aspartate aminotransferase; BMI = body mass index; CHD = coronary heart disease; CRP = C-reactive protein; DBP = diastolic blood pressure; FH = familial hypercholesterolaemia; HDL-C = high-density lipoprotein cholesterol; LASSO = least absolute shrinkage and selection operator; LDL-C = low-density lipoprotein cholesterol; Lp(a) = lipoprotein A; PGS = polygenic score; SBP = systolic blood pressure; T2D = type 2 diabetes.



**Supplementary Figure 6.2 LASSO regression model feature selection and importance for monogenic FH prediction.** Feature importance is ordered by value of log odds ratio (OR) per standard deviation (SD). ALT = alanine transaminase; Apo-A1 = apolipoprotein A; BMI = body mass index; CHD = coronary heart disease; CRP = C-reactive protein; DBP = diastolic blood pressure; FH = familial hypercholesterolaemia; LASSO = least absolute shrinkage and selection operator; LDL-C = low-density lipoprotein cholesterol; PGS = polygenic score; T2D = type 2 diabetes.



# 7 General discussion

## 7.1 Overview of thesis

### 7.1.1 PGS in CVD prediction

This thesis explored the utility of polygenic scores (PGS) in cardiovascular disease (CVD) prediction and screening. The first project (**Chapter 2**) started with an analysis of 2,194 previously published PGS for 544 disease endpoints aggregated in the freely available Polygenic Score Catalog.[1] In this chapter, I converted the hazard ratios (HR) and odds ratio (OR) per one standard deviation (SD) of the PGS, and the area under the receiver operating characteristic curve (AUC) (or C-index/C-statistic) of these scores into a detection rate for a 5% false positive rate (DR5). The detection rate (or sensitivity) is a measure that is commonly used to evaluate the clinical utility of novel models but has yet to be widely incorporated into PGS studies.[2] This measure also provides a more tangible understanding of incremental changes in the HR, OR and AUC/C-index of models in terms of clinical impact.[3,4] The results of the study indicated that the overall strength of association of PGS with various disease endpoints is weak, with the median DR5 [interquartile range (IQR) %] values for the HR and OR per one SD equalling 8% [7; 10] and 9% [6; 12], and the median DR5 for the AUC and C-index reaching 14% [10; 22] and 19% [13; 25] respectively. This means that on average, PGS in the Polygenic Score Catalog missed 81%-92% of affected individuals at a 5% false positive rate.

The chapter further put into perspective the performance of polygenic risk scores (PRS) in individual risk prediction, population stratification, and disease screening by respectively evaluating the odds of being affected given a PRS result, the odds of being affected given the occupancy of a particular PRS quintile, and the odds of being affected given a positive test result (OAPR). These are clinically useful metrics to evaluate tests as they provide an absolute risk that considers the background odds of disease in a population over a specified timeframe, as opposed to PRS which measure relative risk in a population. The normal distributions of the PRS can be exploited to derive the likelihood ratio, which is then used to calculate the odds of being affected. Taking example PRS from the Polygenic Score Catalog, this chapter illustrated the interpretation of a PRS in individual CVD risk prediction, where an average 40-year old individual with a coronary artery disease (CAD) PRS (DR5 = 13%) at the 75% centile of the distribution had their 10-year background odds of CAD increased from 1:12 to 1:10, and at the 25% centile of the PRS distribution had their odds reduced to 1:20. For PRS in population stratification, individuals in the

97.5$^{th}$ centile of a CAD PRS had their odds increased from 1:12 to 1:5, but this tail of the distribution only accounts for 7.5% of all CAD cases, which is important to consider if the aim of risk stratification is to identify many additional at-risk individuals. And in disease screening, the false positives outnumbered the true positives by four to one for a CAD PRS (DR5 = 12%) in a middle-aged population with a background 10-year odds of disease of 1:9, and 41 to one for a population were the background 10-year risk of disease was 1% (e.g. in younger individuals). This shows that the background odds of disease heavily influence the number of false positive cases detected in screening, which is not immediately visible if only relying on a PRS. The conversion of the relative risk that PRS confer to an absolute risk scale (i.e. the odds of being affected) not only provides a more concrete understanding of personal risk on an individual level, but also gives insight into the performance and limitations of PRS in population stratification and disease screening.

The following project (**Chapter 4**) investigated whether PGS might improve the predictive ability of non-genetic clinical risk prediction models for a range of cardiovascular and related outcomes (the QScores) when combined. More specifically, whether they improved the 10-year risk estimation of incident CVD/coronary heart disease (CHD), type 2 diabetes and ischaemic stroke in QRISK3, QDiabetes and QStroke respectively. These non-genetic scores were developed in and for the UK population, and it was therefore appropriate to use the UK Biobank as a test dataset for this project. The quality control and data cleaning steps applied to the UK Biobank data are detailed in **Chapter 3**. The models generated were compared with one another based on their odds of incident disease per one SD in the scores, their discrimination (C-statistic), their calibration (calibration-in-the-large and calibration slope), and their detection rate for a 5% false positive rate. The results showed that the effects of adding PGS to these non-genetic QScores were outcome- and sex-specific and provided at best minimal improvements in the C-statistic (highest improvement obtained: 0.015, equivalent to an increase of 1.5% in the DR5). These results were similar to the ones obtained by previous studies, however, as is often the case with the current PGS literature, the interpretation of the results vary widely.[5,6]

### 7.1.2 PGS in rare variant discovery

The next section of the thesis investigated a novel potential clinical application of PGS: whether PGS information could be used to aid with rare variant discovery for disease screening.[7] Individuals with a PGS for a trait that does not match their observed trait value might be more

likely to harbour a rare monogenic variant of large effect size for the trait in question; as the variant of large effect size would not be captured by the PGS and could explain the discrepancy between observed and expected values.[8]

To study this, I focused on familial hypercholesterolaemia (FH) which is the most common monogenic disorder. FH is caused by deleterious genetic variants in the *LDLR*, *APOB*, *PCSK9* or *APOE* genes, which lead to elevated circulating levels of low-density lipoprotein cholesterol (LDL-C) and increase the risk of premature CHD and death. FH is highly underdiagnosed (only around 7% of UK cases have been detected) as there is currently no population screening strategy in place for it, which is a missed primary prevention opportunity.[9] The NHS Long Term Plan set a goal to increase the diagnosis of FH cases from 7% to 25% in five years but has not specified how this will be achieved.[10]

The first part (**Chapter 5**) of the project consisted of developing a two-stage population screening strategy for FH in adults, which was then subsequently improved by developing a novel prediction model using a machine learning algorithm that included a PGS for LDL-C (**Chapter 6**). This novel population screening approach for the systematic identification of FH patients is in line with the NHS Long Term Plan to increase the diagnosis of FH cases from 7% to 25% in the UK.[10] The two-stage adult screen relies on a first stage where LDL-C concentrations are measured in individuals aged 40 and above attending their NHS Health Check. Measurement of LDL-C is low cost, but insufficiently accurate on its own to properly discriminate FH cases in adults. However, it could be used as part of a two-stage screen in which individuals with LDL-C concentrations exceeding a certain pre-specified threshold value are referred to genomic sequencing in the second stage (a more expensive but highly accurate stage) to either confirm or refute the presence of a monogenic variant for FH.[11] Cascade testing of close relatives of newly identified index cases could then be initiated. This strategy was compared to a similar population screening approach in which index cases are ascertained in childhood: the child-parent screen proposed by Wald *et al.*. This approach was considered previously by the UK National Screening Committee but rejected as an option on the grounds that it does not immediately benefit the children who are screened (at around one to two years of age), as they only become eligible to receive cholesterol-lowering treatment from the age of ten.[12–15] The comparison of both approaches showed that it would take twice a long (approximately 14 years versus 7 years) to reach the NHS' Long Term Plan goal of identifying 25% of UK FH cases if implementing the two-stage adult screen instead of the child-parent screen. It is worth noting that both methods are not mutually exclusive and that

implementing both screening strategies (and/or other strategies) would shorten the amount of time needed to reach the 25% target detection rate set out by the NHS, depending on available resources.

The second part of the project (**Chapter 6**) investigated whether the two-stage adult screening strategy could be improved by reducing the burden of false positive cases sent for sequencing in the second stage. For this, I developed a novel FH prediction model using a machine learning algorithm (LASSO) and various clinical variables that are readily obtained. It also included a PGS for LDL-C. This novel prediction model improved FH case detection and reduced the number of false positive cases sent for genomic sequencing; however, the feasibility of clinical implementation was not evaluated in this instance.

This project also provided evidence that PGS might have additional clinical benefits other than disease prediction, for which I showed in **Chapter 2** that their performance is limited. One application is in aiding the identification of individuals with a monogenic form of a disease.[7,8] Indeed, the negative coefficient of the LDL-C PGS in the model indicated that individuals with hypercholesterolaemia due to a high LDL-C PGS were given a lower probability of having hypercholesterolaemia from carriage of a monogenic variant for FH. However, when I developed a separate model for FH case detection where the PGS for LDL-C was not included, I observed that model performance was almost identical to the one that included the LDL-C PGS. This was likely because the non-genetic variables present in the model (e.g. LDL-C and triglyceride concentrations) acted as proxies for the LDL-C PGS. It is worth noting that the cohort participants were middle aged; and given that the weight of PGS information in prediction seems to be more important in younger individuals, it is possible that the model with the LDL-C PGS for FH detection would perform significantly better than the model without it in younger individuals (such as in children).[16,17] This remains to be tested on a larger scale. For now, the results of this chapter show that polygenic information can be applied in principle as one means to improve the detection of monogenic FH cases, but that this information can also be replaced by measured risk factor variables (in adults).

## 7.2 Wider perspective

### 7.2.1 PGS in context

Research into PGS has substantially increased in the past 10 years. This is partly due to the increasing size and number of published genome-wide association studies (GWAS) providing freely available summary statistics for various traits and diseases. The development of large longitudinal cohort studies (such as the UK Biobank) has enabled the development and testing of these scores. PGS have been shown to be significantly associated with many heritable traits, and the concept of utilising this information to improve healthcare is understandably attractive. The nature of PGS means that they only need to be measured once in a person's lifetime, can be applied to a variety of traits and diseases, can be measured with high technical accuracy, and are cheap to obtain with genotyping costs rapidly declining. It is for these reasons that many private and public entities are funding research into PGS and advocating their use in healthcare.[18–22]

Although PGS are undisputedly associated with many disease traits, the strength of this association is not sufficient to mean that they are clinically useful in disease prediction, screening, and risk stratification. Other scientists have demonstrated that causal risk factors for disease are not necessarily good predictors of disease; and that for a risk factor to be a good predictor, it has to be very highly associated with the disease in question.[23,24] These important points seem to have been overlooked in the field of PGS research.

This can be partly attributed to the way in which the performance of PRS is depicted in scientific publications. Many papers compare the very top of a PGS distribution to other sections of the distribution (such as the bottom end), instead of comparing the top end of the distribution to the rest of the distribution.[21,25,26] These analyses lead to inflated risk ratios that are misleading.

Furthermore, most studies base their conclusions of clinical utility on the OR or HR per SD and the AUC/C-index of PGS models, which are measures that provide information on association but are not directly informative of clinical utility.[3,4] As seen previously, the clinical utility of models is better understood in terms of the detection rate for a pre-specified false positive rate and the OAPR for disease screening, the odds of being affected for a particular test result for individual risk prediction, and the odds of being affected as a result of occupancy of a particular PRS group or quintile for risk stratification, which PGS studies have failed to report.[2,27]

PGS are normally distributed in a population and the overlap in the distributions of affected and unaffected individuals is substantial. The degree of overlap between these distributions indicates the level of discrimination that PGS provide (see **Chapter 1**). However, these distributions (or the mean and SD of the distributions) are not often reported by PGS studies.[25,26,28,29]

Another point to consider is that most PGS studies include other variables in their models, particularly age, but do not often clearly state this.[28] This thesis (**Chapter 4**) and many scientific publications have shown how the inclusion of other variables in PGS models (such as age and sex) have a major influence on disease prediction, especially CVD.[6,16,30] There needs to be clearer transparency and objectiveness in PGS reporting, and when evaluating their clinical utility.

Many studies have also looked at whether the inclusion of PGS in non-genetic risk prediction models for disease (such as CVD) improve prediction and risk stratification of individuals. There are multiple points to consider here. While the inclusion of PGS have shown to increase the C-statistic of most non-genetic risk prediction models, this increase has not been substantial (**Chapter 4**).[6,16,26,31] Net reclassification index (NRI) tables in these papers show that the inclusion of PGS in non-genetic risk prediction models improves the reclassification of individuals into higher or lower risk categories. While this might be true, these numbers tend to be low, as most cases occur in the middle of the risk distribution rather than the upper tail.[27] Therefore, for diseases where there are cheap, non-invasive, and safe preventative options (such as statins for CVD), there might be more benefit in lowering the age of treatment commencement than to include a PGS in the risk prediction models.[32] It is also worth noting that NRI tables have been criticised as a measure of clinical utility.[33] A more intuitive measure is the DR for a given FPR (e.g. DR5): **Chapter 4** showed that the inclusion of a PGS in non-genetic CVD models improved the DR5 by at best 1.5%.[2,27]

### 7.2.2 Important considerations

In this polarised field of research, some scientists believe that the magnitude change in the C-statistic that PGS provide to non-genetic CVD risk prediction tools is clinically useful in improving disease prediction, disease screening, and patient risk stratification, while others interpret these results with more caution.[5,34–36] Further evidence from trials, such as the one announced by the NHS and Genomics plc in the UK in early 2022, will provide invaluable information on the

feasibility of PGS implementation in clinical care and on their utility in patient risk stratification and disease prediction.[37] Proper cost-effectiveness studies will also be needed.

As it currently stands, there are important points to consider if moving forward with the clinical implementation of PGS in disease prediction and screening. The first point being that the improvement in discrimination and in the detection rate for a false positive rate that these scores provide is still minimal. With the increasing size and diversity of genome-wide association studies from which these PGS are derived, and with the development of statistical and computational methods that improve the predictive power of these scores, it is conceivable to imagine that PGS might one day provide benefits to patients. However, progress and research in this field is still much needed prior to clinical implementation.

Another point to consider are the valid concerns behind the poor transferability of PGS to other ancestry groups.[38] Many efforts are underway to address these issues, but currently the implementation of such scores could cause more harm than good by widening health disparities.[38]

And finally, the issue of communicating the meaning of PGS (and risk) to the general public (and to scientists) is still a major one. The most damaging being the false interpretation or over-interpretation of PGS "determinism" in terms of the magnitude risk they confer to various traits and diseases. An example of this are the various start-ups that are now incorporating PGS screening into embryo selection, which will not only likely disappoint parents-to-be, but also presents significant ethical concerns that borders on eugenics.[21]

### 7.2.3 Future avenues

While PGS seem to have limited added benefit in terms of CVD prediction and screening in the general population, they could still be useful in other contexts. The combination of multiple PGS for improving risk prediction has yet to be fully explored for example. Their utility in risk prediction or stratification may also depend on the outcome studied, or on patient subgroups rather than the general population (as was seen in the improved prediction of incident CVD in type 2 diabetes patients).[39]

It has also been proposed that PGS could be useful in aiding with rare variant discovery, which was explored in **Chapter 6**.[7,8] Indeed, there is a need to identify individuals with rare variants for Mendelian disorders (such as for FH), especially where interventions are available. There are two main reasons for this: treating the index case, and cascade testing of close relatives. In the case of FH, multiple treatment options are available such as oral (statins, ezetimibe) and injectable (PCSK9 inhibitors, monoclonal antibodies) therapies, and the emergence of base editing therapies (e.g. CRISPR for *PCSK9*).[40,41] Cascade testing of index cases has also been shown to be highly cost-effective in many countries.[42–45] The role that PGS play in modulating monogenic variant penetrance is also being studied.[21]

## 7.3 Summary

Research into PGS and how they might improve clinical care has grown significantly in the past few years. The current consensus in the scientific community seems to be that they are useful and will play a part in future clinical care, but this viewpoint is still contested by many. This thesis puts the utility of PGS in CVD prediction, screening and risk stratification into perspective and provides evidence of them being poor predictors of incident CVD in the general population, but they may still find other purposes. Further critical research is needed in the field, and careful practical and ethical considerations will be essential prior to clinical implementation. More research into other applications of PGS, such as in rare variant discovery and how they might benefit patient care, is also worth pursuing and is expected to grow in this fast-evolving field of research.

## 7.4 References


1 Lambert SA, Gil L, Jupp S, *et al.* The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. Nat Genet. 2021. doi:10.1038/s41588-021-00783-5

2 Wald NJ, Old R. The illusion of polygenic disease risk prediction. *Genet Med* 2019;**21**:1705–7. doi:10.1038/S41436-018-0418-5

3 Wald NJ, Bestwick JP. Is the area under an ROC curve a valid measure of the performance of a screening or diagnostic test? *J Med Screen* Published Online First: 2014. doi:10.1177/0969141313517497

4 Wald NJ, Morris JK. Assessing Risk Factors as Potential Screening Tests: A Simple Assessment Tool. *Arch Intern Med* 2011;**171**:286–91. doi:10.1001/ARCHINTERNMED.2010.378

5 Riveros-Mckay F, Weale ME, Moore R, *et al.* Integrated Polygenic Tool Substantially Enhances Coronary Artery Disease Prediction. *Circ Genom Precis Med* 2021;**14**:192–200. doi:10.1161/CIRCGEN.120.003304

6 Elliott J, Bodinier B, Bond TA, *et al.* Predictive Accuracy of a Polygenic Risk Score–Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease. *JAMA* 2020;**323**:636–45. doi:10.1001/JAMA.2019.22241

7 Lu T, Forgetta V, Richards JB, *et al.* Polygenic risk score as a possible tool for identifying familial monogenic causes of complex diseases. *Genetics in Medicine* 2022;**0**:1–11. doi:10.1016/J.GIM.2022.03.022

8 Zhou D, Yu D, Scharf JM, *et al.* Contextualizing genetic risk score for disease screening and rare variant discovery. *Nat Commun* 2021;**12**:1–14. doi:10.1038/s41467-021-24387-z

9 NHS England — London » Familial Hypercholesterolemia (FH). https://www.england.nhs.uk/london/london-clinical-networks/our-networks/cardiac/familial-hypercholesterolaemia/ (accessed 22 Oct 2022).

10 NHS Long Term Plan » Cardiovascular disease. https://www.longtermplan.nhs.uk/online-version/chapter-3-further-progress-on-care-quality-and-outcomes/better-care-for-major-health-conditions/cardiovascular-disease/ (accessed 2 Mar 2022).

11 NHS Health Check programme, Patients Recorded as Attending and Not Attending, 2012-13 to 2017-18 - NHS Digital. https://digital.nhs.uk/data-and-information/publications/statistical/nhs-health-check-programme/2012-13-to-2017-18 (accessed 3 Mar 2022).

12 Wald DS, Bestwick JP, Morris JK, *et al.* Child–Parent Familial Hypercholesterolemia Screening in Primary Care. *New England Journal of Medicine* 2016;**375**:1628–37. doi:10.1056/NEJMOA1602777

13 Wald DS, Bestwick JP, Wald NJ. Child-parent screening for familial hypercholesterolaemia: screening strategy based on a meta-analysis. *BMJ* 2007;**335**:599. doi:10.1136/BMJ.39300.616076.55

14 Wald DS, Martin AC. Decision to reject screening for familial hypercholesterolaemia is flawed. *Arch Dis Child* 2021;**106**:525–6. doi:10.1136/ARCHDISCHILD-2020-319168

15 Wald DS, Neely D. The UK National Screening Committee's position on child–parent screening for familial hypercholesterolaemia. *J Med Screen* 2021;**28**:217. doi:10.1177/09691413211025426

16 Khan SS, Page C, Wojdyla DM, *et al.* Predictive Utility of a Validated Polygenic Risk Score for Long-Term Risk of Coronary Heart Disease in Young and Middle-Aged Adults. *Circulation* 2022;:101161CIRCULATIONAHA121058426. doi:10.1161/CIRCULATIONAHA.121.058426

17 Futema M, Shah S, Cooper JA, *et al.* Refinement of Variant Selection for the LDL Cholesterol Genetic Risk Score in the Diagnosis of the Polygenic Form of Clinical Familial Hypercholesterolemia and Replication in Samples from 6 Countries. *Clin Chem* 2015;**61**:231–8. doi:10.1373/CLINCHEM.2014.231365

18 Allelica | Polygenic Risk Score.

19 Genome UK: 2021 to 2022 implementation plan - GOV.UK. https://www.gov.uk/government/publications/genome-uk-2021-to-2022-implementation-plan/genome-uk-2021-to-2022-implementation-plan (accessed 1 Sep 2022).

20 Polygenic score pilot for heart disease begins - Genomics Education Programme. https://www.genomicseducation.hee.nhs.uk/blog/polygenic-score-pilot-for-heart-disease-begins/ (accessed 1 Sep 2022).

21 Turley P, Meyer MN, Wang N, *et al.* Problems with Using Polygenic Scores to Select Embryos. *New England Journal of Medicine* 2021;**385**:78–86. doi:10.1056/NEJMSR2105065/SUPPL_FILE/NEJMSR2105065_DISCLOSURES.PDF

22 Lambert SA, Abraham G, Inouye M. Towards clinical utility of polygenic risk scores. *Hum Mol Genet* 2019;**28**:R133–42. doi:10.1093/HMG/DDZ187

23 Wald NJ, Hackshaw AK, Frost CD. When can a risk factor be used as a worthwhile screening test? *BMJ : British Medical Journal* 1999;**319**:1562. doi:10.1136/BMJ.319.7224.1562

24 Schooling CM, Jones HE. Clarifying questions about 'risk factors': Predictors versus explanation. *Emerg Themes Epidemiol* 2018;**15**:1–6. doi:10.1186/S12982-018-0080-Z/TABLES/1

25 Mars N, Koskela JT, Ripatti P, *et al.* Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nature Medicine 2020 26:4* 2020;**26**:549–57. doi:10.1038/s41591-020-0800-0

26 Riveros-Mckay F, Weale ME, Moore R, *et al.* Integrated Polygenic Tool Substantially Enhances Coronary Artery Disease Prediction. *Circ Genom Precis Med* 2021;**14**:E003304. doi:10.1161/CIRCGEN.120.003304

27 Groenendyk JW, Greenland P, Khan SS. Incremental Value of Polygenic Risk Scores in Primary Prevention of Coronary Heart Disease: A Review. *JAMA Intern Med* Published Online First: 22 August 2022. doi:10.1001/JAMAINTERNMED.2022.3171

28 Khera A v., Chaffin M, Aragam KG, *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics 2018 50:9* 2018;**50**:1219–24. doi:10.1038/s41588-018-0183-z

29 Mosley JD, Gupta DK, Tan J, *et al.* Predictive Accuracy of a Polygenic Risk Score Compared With a Clinical Risk Score for Incident Coronary Heart Disease. *JAMA* 2020;**323**:627–35. doi:10.1001/JAMA.2019.21782

30 Lello L, Raben TG, Yong SY, *et al.* Genomic Prediction of 16 Complex Disease Risks Including Heart Attack, Diabetes, Breast and Prostate Cancer. *Scientific Reports 2019 9:1* 2019;**9**:1–16. doi:10.1038/s41598-019-51258-x

31 Sun L, Pennells L, Kaptoge S, *et al.* Polygenic risk scores in cardiovascular risk prediction: A cohort study and modelling analyses. *PLoS Med* 2021;**18**:e1003498. doi:10.1371/JOURNAL.PMED.1003498

32 Hingorani AD, Hemingway H. How should we balance individual and population benefits of statins for preventing cardiovascular disease? *BMJ* Published Online First: 2011. doi:10.1136/bmj.c6244

33 Pepe MS, Fan J, Feng Z, *et al.* The Net Reclassification Index (NRI): A Misleading Measure of Prediction Improvement Even with Independent Test Data Sets. *Stat Biosci* 2015;**7**:282–95. doi:10.1007/S12561-014-9118-0/FULLTEXT.HTML

34 Inouye M, Abraham G, Nelson CP, *et al.* Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. *J Am Coll Cardiol* 2018;**72**:1883. doi:10.1016/J.JACC.2018.07.079

35 Mosley JD, Gupta DK, Tan J, *et al.* Predictive Accuracy of a Polygenic Risk Score Compared With a Clinical Risk Score for Incident Coronary Heart Disease. *JAMA* 2020;**323**:627–35. doi:10.1001/JAMA.2019.21782

36 Lyssenko V, Jonsson A, Almgren P, *et al.* Clinical Risk Factors, DNA Variants, and the Development of Type 2 Diabetes. *New England Journal of Medicine* 2008;**359**:2220–32. doi:10.1056/nejmoa0801869

37 NHS launches new polygenic scores trial for heart disease - Genomics Education Programme. https://www.genomicseducation.hee.nhs.uk/blog/nhs-launches-new-polygenic-scores-trial-for-heart-disease/ (accessed 5 Apr 2022).

38 Martin AR, Kanai M, Kamatani Y, *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* 2019;**51**:584–91. doi:10.1038/s41588-019-0379-x

39 Dziopa K, Chaturvedi N, Gratton J, *et al.* Combining stacked polygenic scores with clinical risk factors improves cardiovascular risk prediction in people with type 2 diabetes. *medRxiv* 2022;:2022.09.01.22279477. doi:10.1101/2022.09.01.22279477

40 McGowan MP, Hosseini Dehkordi SH, Moriarty PM, *et al.* Diagnosis and treatment of heterozygous familial hypercholesterolemia. *J Am Heart Assoc* 2019;**8**. doi:10.1161/JAHA.119.013225

41 Ledford H. CRISPR 'cousin' put to the test in landmark heart-disease trial. *Nature* 2022;**607**:647–647. doi:10.1038/D41586-022-01951-1

42 Kerr M, Pears R, Miedzybrodzka Z, *et al.* Cost effectiveness of cascade testing for familial hypercholesterolaemia, based on data from familial hypercholesterolaemia services in the UK. *Eur Heart J* 2017;**38**:1832–9. doi:10.1093/EURHEARTJ/EHX111

43 Marquina C, Lacaze P, Tiller J, *et al.* Population genomic screening of young adults for familial hypercholesterolaemia: a cost-effectiveness analysis. *Eur Heart J* Published Online First: 11 November 2021. doi:10.1093/EURHEARTJ/EHAB770

44 Jackson CL, Huschka T, Borah B, *et al.* Cost-effectiveness of cascade genetic testing for familial hypercholesterolemia in the United States: A simulation analysis. *Am J Prev Cardiol* 2021;**8**:100245. doi:10.1016/J.AJPC.2021.100245

45 Lázaro P, Pérez de Isla L, Watts GF, *et al.* Cost-effectiveness of a cascade screening program for the early detection of familial hypercholesterolemia. *J Clin Lipidol* 2017;**11**:260–71. doi:10.1016/J.JACL.2017.01.002