THE UNIVERSITY
of EDINBURGH

Biomarkers for cardiovascular risk prediction in people with type 2 diabetes

Anna Helen Price

Doctor of Philosophy

University of Edinburgh

Spring 2017

Contents

Acknowledgements ..... **i**

Declaration ..... **ii**

Publications and presentations relating to the work of this thesis ..... **iii**

Abstract ..... **iv**

1 Introduction: Type 2 diabetes mellitus, cardiovascular disease and biomarkers **1**

1.1 Diabetes Mellitus and Type 2 Diabetes 1

1.2 Cardiovascular disease 3

1.3 Conventional cardiovascular risk factors 4

1.4 Non-traditional biomarkers 6

1.4.1 Ankle Brachial Index 6

1.4.2 NT-proBNP 7

1.4.3 Troponin 7

1.4.4 GGT 8

1.4.5 Inflammatory biomarkers 8

1.5 Metabolomics 8

1.6 Cardiovascular risk scores 9

2 Statistical methods for risk prediction **12**

2.1 Characteristics of risk prediction models 12

2.2 Selecting biomarkers for risk prediction models 13

2.3 Omics data in risk prediction modelling 16

2.3.1 Univariate approach 17

2.3.2 Dimension reduction 18

2.3.3 Variable selection 19

2.4 Assessing a prediction model 20

2.4.1 Discrimination 21

2.4.2 Calibration 26

2.4.3 Global measures of fit 28

2.5 Validating a prediction model 28

3 Aims and objectives **30**

3.1 Aims 30

3.2 Thesis outline 30
4 Systematic Review: Cardiovascular risk scores for people with type 2 diabetes 32
4.1 Background 32
4.2 Aim 32
4.3 Methods 32
4.3.1 Inclusion and Exclusion Criteria 32
4.3.2 Search strategy 33
4.3.3 Selection of studies 35
4.3.4 Data extraction and management 35
4.4 Results 35
4.4.1 Study selection 35
4.4.2 Results tables 36
4.5 Discussion 61
4.6 Choosing a risk score 65
4.7 Conclusion 66
5 Data sources and methods 68
5.1 Edinburgh Type 2 Diabetes Study 68
5.1.1 Study population 69
5.1.2 Data collection 71
5.1.3 Variable measurement and definitions 73
5.1.3.1 Demographics 73
5.1.3.2 Diabetes and other medical history 74
5.1.3.3 Physical examination 74
5.1.3.4 Blood samples 75
5.1.3.5 Smoking 76
5.1.3.6 Prevalent cardiovascular events 77
5.1.3.7 Incident cardiovascular events 78
5.1.4 Ethical approval 81
5.1.5 Data management, cleaning and security 81
5.1.6 Missing data 81
5.1.7 Data analysis 82

5.1.7.1 Developing a basic model 82
5.1.7.2 General inflammation factor 86
5.2 UCLEB consortium cohorts 87
5.2.1 Contributing UCLEB studies 87
5.2.2 BRHS 88
5.2.3 BWHHS 89
5.2.4 SABRE 89
5.2.5 WHII 90
5.2.6 Data analysis 90
5.2.7 Variable definitions, data collation and missing data 92
6 Results I: Characteristics of ET2DS and descriptive statistics for cardiovascular events and biomarkers 95
6.1 Baseline demographic characteristics 95
6.2 Representativeness 97
6.3 Incident cardiovascular events 97
6.4 Descriptive statistics of biomarkers 97
6.5 Missing data 101
7 Results II: Improving cardiovascular risk prediction using individual and combined biomarkers in the ET2DS 103
7.1 Basic model 103
7.2 Associations between biomarkers at baseline 103
7.3 Relationships between biomarkers and cardiovascular risk 104
7.4 Adding individual biomarkers to the basic model 108
7.5 Adding combinations of biomarkers to the basic model 110
7.6 Conclusions 110
8 Results III: Associations between metabolomics data and cardiovascular disease in the UCLEB consortium cohorts 113
8.1 Missing data 113
8.2 Baseline characteristics of the UCLEB cohorts 116
8.3 CVD in the UCLEB cohorts 119
8.4 Descriptive statistics of metabolites 120

8.5 Associations between metabolites and cardiovascular disease in individual UCLEB studies 120
8.6 Associations between metabolites and CVD in the combined UCLEB data 125
8.6.1 Associations between metabolites and CVD adjusted for cohort 125
8.6.2 Associations between metabolites and CVD adjusted for cohort, age and sex 128
8.6.3 Associations between metabolites and CVD adjusted for cohort, age, sex and traditional cardiovascular risk factors 128
8.6.4 Associations between metabolites and CVD adjusted for cohort, age, sex, traditional cardiovascular risk factors, social status, BMI, eGFR and ethnicity 133
9 Discussion 136
9.1 Key findings 136
9.1.1 Improving cardiovascular risk prediction using individual and combined biomarkers 136
9.1.2 Associations between metabolomics data and CVD 136
9.2 Improving cardiovascular risk prediction using individual and combined biomarkers 137
9.2.1 Strengths of the ET2DS 137
9.2.1.1 Recruitment and representativeness 137
9.2.1.2 Completeness and accuracy of data collection 138
9.2.1.3 Cardiovascular follow up 138
9.2.1.4 Prospective design and sample size 138
9.2.2 Limitations of the ET2DS 139
9.2.2.1 Generalisability 139
9.2.2.2 Prevalent CVD and prescription of lipid-lowering medication 139
9.2.3 Strengths of the analysis plan 140
9.2.4 Limitations of the analysis plan 141
9.2.5 Comparisons of findings with previous studies 141
9.2.5.1 NT-proBNP 141
9.2.5.2 hs-cTnT 143

9.2.5.3 ABI 144
9.2.5.4 GGT 144
9.2.5.5 Inflammation factor 145
9.3 Associations between metabolomics data and CVD 146
9.3.1 Strengths of the UCLEB consortium studies 146
9.3.1.1 Sample size 146
9.3.1.2 Risk factors available 147
9.3.1.3 Analysis plan 147
9.3.2 Limitations of the UCLEB consortium data 147
9.3.2.1 Definition of outcome 147
9.3.2.2 Risk factor definitions and availability of variables 148
9.3.3 Comparisons of findings with previous studies 149
9.4 Risk prediction methods 150
9.4.1 Impact of the choice of statistical method 150
9.4.2 Model evaluation measures 151
9.4.3 Development of software for complex methods 152
9.5 Recommendations for future research 153
9.5.1 Improving cardiovascular risk prediction using individual and combined biomarkers 153
9.5.2 Associations between metabolomics data and CVD 154
10 Bibliography 156
11 Appendices A-1
Appendix A Publications and presentations A-1
Appendix B PAC application form B-1
Appendix C Descriptive statistics of metabolites C-1

List of Tables

Table 4-1 Cardiovascular risk models specifically developed in patients with type 2 diabetes 37
Table 4-2 Cardiovascular risk models developed in general populations with diabetes as a predictor 42
Table 4-3 Papers excluded at the full-text stage 56
Table 5-1 Summary of data collection in the Edinburgh Type 2 Diabetes Study 73
Table 5-2 PCA output for the ET2DS general inflammation factor g 87
Table 5-3 Variables available in the UCLEB cohorts 94
Table 6-1 Baseline characteristics of the ET2DS population 96
Table 6-2 Demographic and clinical characteristics of the ET2DS population and non-responders. Adapted from Marioni et al., 2010. 98
Table 6-3 Summary of first incident or recurrent cardiovascular events in the ET2DS 99
Table 6-4 Descriptive statistics of baseline biomarkers in ET2DS 99
Table 6-5 Missing data in the ET2DS at baseline 101
Table 7-1 Basic model coefficients and summary measures 105
Table 7-2 Correlation coefficients between biomarkers at baseline 106
Table 7-3 Associations between individual biomarkers and cardiovascular events and corresponding measures of model performance 109
Table 7-4 Top five models selected from the combined biomarkers and corresponding measures of model performance 112
Table 8-1 Available values for the risk factor variables in the UCLEB consortium cohorts 115
Table 8-2 Baseline characteristics of the UCLEB consortium cohorts 118
Table 8-3 Summary of cardiovascular events available in the UCLEB consortium cohorts 119
Table C-1 Medians, interquartile ranges (IQR) and missing values (NA; blank cell indicates no missing values) for 228 metabolites in the five individual UCLEB studies (full names of metabolites can be found in Table C-2) C-20
Table C-2 Data dictionary for full metabolite names C-29

List of Figures

Figure 2-1: Comparison of three ROC curves with varying discriminative abilities 22
Figure 2-2: Predictiveness curves for two competing models. Adapted from Pencina et al., 2010. 26
Figure 4-1 Flow chart of systematic review of studies presenting a cardiovascular risk score for use in people with type 2 diabetes 36
Figure 5-1 ET2DS recruitment flow diagram. Adapted from Price et al., 2008..... 70
Figure 6-1 Distributions of baseline biomarkers in ET2DS 100
Figure 7-1 Cardiovascular (CV) risk against categorised biomarkers in the ET2DS (Q: quintile)..... 107
Figure 8-1 Bar plots of p-values for univariate analysis of metabolites and cardiovascular disease in the individual UCLEB cohorts. 123
Figure 8-2 Odds ratios from univariate analysis of metabolites and cardiovascular disease in the individual UCLEB cohorts. 124
Figure 8-3 Bar plots of p-values for Model 1 analysis of metabolites and cardiovascular disease in the combined UCLEB cohorts. 126
Figure 8-4 Odds ratios from Model 1 analysis of metabolites and cardiovascular disease in the combined UCLEB cohorts..... 127
Figure 8-5 Bar plots of p-values for Model 2 analysis of metabolites and cardiovascular disease in the combined UCLEB cohorts. 129
Figure 8-6 Odds ratios from Model 2 analysis of metabolites and cardiovascular disease in the combined UCLEB cohorts..... 130
Figure 8-7 Bar plots of p-values for Model 3 analysis of metabolites and cardiovascular disease in the combined UCLEB cohorts. 131
Figure 8-8 Odds ratios from Model 3 analysis of metabolites and cardiovascular disease in the combined UCLEB cohorts..... 132
Figure 8-9 Bar plots of p-values for Model 4 analysis of metabolites and cardiovascular disease in the combined UCLEB cohorts. 134
Figure 8-10 Odds ratios from Model 4 analysis of metabolites and cardiovascular disease in the combined UCLEB cohorts..... 135

Figure C-1 Histograms of individual metabolites from the combined UCLEB data (full names of metabolites can be found in Table C-2)..... C-1

List of Abbreviations

ABI	Ankle brachial index
ABPI	Ankle brachial pressure index
AIC	Akaike information criterion
ApoA-1	Apolipoprotein A-1
ApoB	Apolipoprotein B
BIC	Bayes information criterion
BMI	Body mass index
BNP	Brain natriuretic peptide
BRHS	British Regional Heart Study
BWHHS	British Women's Heart and Health Study
CHD	Coronary heart disease
CI	Confidence interval
CKD	Chronic kidney disease
CRP	C-reactive protein
CVD	Cardiovascular disease
ECG	Electrocardiogram
eGFR	Estimated glomerular filtration rate
ET2DS	Edinburgh Type 2 Diabetes Study
FDR	False discovery rate
FWER	Family wide error rate
GGT	Gamma-glutamyl transpeptidase
GP	General practitioner
HbA_{1c}	Glycated haemoglobin
HDL	High-density lipoprotein
hs-cTnT	High-sensitivity cardiac troponin T
IDI	Integrated discrimination improvement
IDL	Intermediate-density lipoprotein
IHD	Ischaemic heart disease
IL-6	Interleukin-6
IP	Integrated specificity
IS	Integrated sensitivity
ISD	Information Services Division
LASSO	Least absolute selection and shrinkage operator
LDL	Low-density lipoprotein
LDR	Lothian Diabetes Register
MAR	Missing at random
MCAR	Missing completely at random
MI	Myocardial infarction
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NMR	Nuclear magnetic resonance
NPV	Negative predictive value
NRI	Net reclassification index
NR	Net reclassification
NT-proBNP	N-terminal pro-brain natriuretic peptide

OR	Odds ratio
PAC	Privacy Advisory Committee
PAD	Peripheral arterial disease
PCA	Principal components analysis
PLS	Partial least squares
PPV	Positive predictive value
ROC	Receiver operating characteristic
SABRE	Southall and Brent Revisited Study
sBP	Systolic blood pressure
SIGN	Scottish Intercollegiate Guidelines Network
SIMD	Scottish Index of Multiple Deprivation
SMR01	Scottish Morbidity Records
TIA	Transient ischaemic attack
TNF-α	Tumor necrosis factor alpha
UCLEB	UCL-LSHTM-Edinburgh-Bristol
VLDL	Very low-density lipoprotein
WHO	World Health Organisation
WHII	Whitehall-II

Acknowledgements

Firstly, I would like to take this opportunity to thank my supervisors, Professor Jackie Price and Professor Chris Weir, for their insightful feedback, constructive criticism and encouragement throughout the PhD. Their advice was invaluable and I could not have hoped for more supportive supervisors.

I would also like to acknowledge all of the staff and participants of the Edinburgh Type 2 Diabetes Study. Without them this thesis would not have been possible. Particular thanks go to all the (current and former) occupants of Room 666, who were an absolute pleasure to work alongside and in collaboration with.

I am so grateful to my PhD peers and academic colleagues who have been an invaluable source of help and support. Thanks to Dr Marco Colombo for his expert statistical, coding and computing advice; thanks to Dr Stela McLachlan for her help with data collation and manipulation; thanks to Marshall Dozier for her expert advice on systematic reviews; thanks to Hannah, Tracy and Sarah for offering listening ears and regular lunchbreak escapes; and thanks to all my PhD friends for providing distractions when they were needed most (and when they were not) and for keeping me sane for three years.

Finally, I wish to thank my parents, Norah Spears and David Price, for always encouraging me to do what I enjoy and what interests me in life; and thanks to Ross Mitchell for his (relative) patience during the PhD and his love and support for thirteen years.

Declaration

I declare that this thesis is of my own composition. The work presented here has not been submitted for any other degree or professional qualification.

The Edinburgh Type 2 Diabetes Study, which provided the data for a large proportion of the analyses presented here, had already completed the baseline, year 1 and year 4 phases by the time I joined the study, so the variables from these time points used for the purpose of analyses in this thesis were collected, cleaned and (in some cases) manipulated through the efforts of other members of the research team. I was responsible for the collection of eight-year follow-up data and the derivation of subsequent variables from this time point.

Signed:

Date:

Publications and presentations relating to the work of this thesis

A list of publications and presentations relating to the work carried out during this PhD project can be found below. Abstracts for presentations on work presented in this thesis can be found in Appendix A

Publications:

AH Price, CJ Weir, P Welsh, S McLachlan, MWJ Strachan, N Sattar, JF Price (2017) *Comparison of non-traditional biomarkers, and combinations of biomarkers, for vascular risk prediction in people with type 2 diabetes: The Edinburgh Type 2 Diabetes Study.* Submitted for publication – draft can be found in Appendix A

MF Suárez-Ortegón, S McLachlan, **AH Price**, M Fernández-Balsells, J Franch-Nadal, M Mata-Cases, J Barrot-de la Puente, X Mundet-Tudurí, D Mauricio, W Ricar, SH Wild, MWJ Strachan, JF Price, J-M Fernández-Real (2017) *Decreased iron stores are associated with cardiovascular disease in patients with type 2 diabetes: cross-sectional and longitudinal findings in two independent cohorts from Scotland and Catalonia*. Submitted for publication.

AH Price, P Welsh, CJ Weir, I Feinkohl, CM Robertson, JR Morling, S McLachlan, MWJ Strachan, N Sattar, JF Price (2014) *N-terminal pro-brain natriuretic peptide and risk of cardiovascular events in older patients with type 2 diabetes: the Edinburgh Type 2 Diabetes Study*. Diabetologia 57(12); 2505-12

R Bedenis, **AH Price**, CM Robertson, JR Morling, BM Frier, MWJ Strachan, JF Price (2014) *Association between severe hypoglycaemia, adverse macrovascular events, and inflammation in the Edinburgh Type 2 Diabetes Study.* Diabetes Care 37; 3301-3308

Conference Presentations:

European Association for the Study of Diabetes Annual Meeting 2016

Adding novel biomarkers to current cardiovascular risk scores for people with Type 2 diabetes: the Edinburgh Type 2 Diabetes Study (ET2DS). **AH Price**, CJ Weir, MWJ Strachan, N Sattar, S McLachlan, JF Price. Poster presentation – abstract can be found in Appendix A

Diabetes UK Professional Conference 2016

Improving cardiovascular (CV) risk scores with novel biomarkers in people with Type 2 diabetes: the Edinburgh Type 2 Diabetes Study (ET2DS). **AH Price**, CJ Weir, MWJ Strachan, N Sattar, S McLachlan, CM Robertson, JF Price. Diabetic Medicine 33 (Supp. 1) 2-208. Oral and poster presentation – abstracts can be found in Appendix A

Diabetes UK Professional Conference 2013

N-terminal pro-B-type natriuretic peptide (NT-proBNP) as an independent predictor of cardiovascular (CV) disease in people with Type 2 diabetes: the Edinburgh Type 2 Diabetes Study (ET2DS). **AH Price**, MWJ Strachan, CM Robertson, P Welsh, N Sattar, JF Price. Oral and poster presentation.

Abstract

Introduction: Type 2 diabetes continues to be one of the most common non-communicable diseases worldwide and complications due to type 2 diabetes, such as cardiovascular disease (CVD) can cause severe disability and even death. Despite advances in the development and validation of cardiovascular risk scores, those used in clinical practice perform inadequately for people with type 2 diabetes. Research has suggested that particular non-traditional biomarkers and novel omics data may provide additional value to risk scores over-and-above traditional predictors.

Aims: To determine whether a small panel of non-traditional biomarkers improve prediction models based on a current cardiovascular risk score (QRISK2), either individually or in combination, in people with type 2 diabetes. Furthermore, to investigate a set of 228 metabolites and their associations with CVD, independent of well-established cardiovascular risk factors, in order to identify potential new predictors of CVD for future research.

Methods: Analyses used the Edinburgh Type 2 Diabetes Study (ET2DS), a prospective cohort of 1066 men and women with type 2 diabetes aged 60-75 years at baseline. Participants were followed for eight years, during which time 205 had a cardiovascular event. Additionally, for omics analyses, four cohorts from the UCL-LSHTM-Edinburgh-Bristol (UCLEB) consortium were combined with the ET2DS. Across all studies, 1005 (44.73%) participants had CVD at baseline or experienced a cardiovascular event during follow-up.

Results: In the ET2DS, higher levels of high sensitivity cardiac troponin (hs-cTnT) and N-terminal pro-brain natriuretic peptide (NT-proBNP) and lower levels of ankle brachial pressure index (ABI) were associated with incident cardiovascular events, independent of QRISK2 and pre-existing cardiovascular disease (odds ratios per one SD increase in biomarker 1.35 (95% CI: 1.13, 1.61), 1.23 (1.02, 1.49) and 0.86 (0.73, 1.00) respectively). The addition of each biomarker to a model including just QRISK2 variables improved the c-statistic, with the biggest increase for hs-cTnT (from 0.722 (0.681, 0.763) to 0.732 (0.690, 0.774)). When multiple biomarkers were considered in combination, the greatest c-statistic was found for a model which included ABI, hs-cTnT and gamma-glutamyl transpeptidase (0.740 (0.699, 0.781)).

In the combined cohorts from the UCLEB consortium, a small number of high-density lipoprotein (HDL) particles were found to be significantly associated with CVD:

concentration of medium HDL particles, total lipids in medium HDL, phospholipids in medium HDL and phospholipids in small HDL. These associations persisted after adjustment for a range of traditional cardiovascular risk factors including age, sex, blood pressure, smoking and HDL to total cholesterol ratio.

Conclusions: In older people with type 2 diabetes, a range of non-traditional biomarkers increased predictive ability for cardiovascular events over-and-above the commonly used QRISK2 score, and a combination of biomarkers may provide the best improvement. Furthermore, a small number of novel omics biomarkers were identified which may further improve risk scores or provide better prediction than traditional lipid measurements such as HDL cholesterol.

1 Introduction: Type 2 diabetes mellitus, cardiovascular disease and biomarkers

1.1 Diabetes Mellitus and Type 2 Diabetes

Diabetes mellitus (commonly referred to as diabetes) is one of the most common non-communicable diseases in the world. Between 1980 and 2014 the number of people with diabetes increased from 108 million to 422 million (World Health Organization, 2016) and the mortality and morbidity directly linked to clinical complications such as cardiovascular disease and renal failure accounted for approximately five million deaths in 2015, 14.5% of global all-cause mortality in adults (International Diabetes Federation, 2015). Furthermore, the number of people with diabetes worldwide is predicted to rise to 642 million by 2040 (International Diabetes Federation, 2015). As well as resulting in loss of life, diabetes complications can cause severe disability and drastically reduce quality of life. It was estimated that in 2015 the cost of health spending on diabetes was at least USD 673 billion – approximately 11% of the total amount of money spent on health worldwide in the adult population (International Diabetes Federation, 2015).

The World Health Organization (WHO) defines diabetes as a "metabolic disorder of multiple aetiology characterised by chronic hyperglycaemia with disturbances of carbohydrate, fat and protein metabolism resulting from defects in insulin secretion, insulin action, or both" (World Health Organization, 2014). In other words, diabetes is a long-term metabolic disorder with multiple causes which is characterised by a persistent excess of glucose in the blood.

There are generally accepted to be two main types of diabetes: type 1 and type 2 diabetes. Type 1 diabetes, commonly referred to as "insulin-dependent diabetes", is caused by an autoimmune destruction of insulin-producing β-cells. It usually presents during early childhood or adolescence and requires patients to undergo regular, lifelong insulin injections. Type 2 diabetes can be caused by either a defect in insulin secretion or resistance to the action of insulin. It is by far the most common type of diabetes (representing 90% of diabetes cases worldwide) and usually develops in adulthood, although numbers are increasing in children and adolescents

(Pulgaron and Delamater, 2014). The development of type 2 diabetes is strongly related to obesity, lack of physical activity and bad diet (Kahn et al., 2006; Sigal et al., 2004). Type 2 diabetes can often be controlled using diet, exercise and weight loss alone, and in fact recent studies have suggested that type 2 diabetes could be reversible through the use of an extreme calorie-restrictive diet (Steven et al., 2016). However, in the majority of cases a combination of diet and exercise and either oral antidiabetic drugs (for example, sulphonylureas or biguanides) or the addition of insulin injections is eventually necessary.

The symptoms of type 2 diabetes are non-specific and may be minimal or not present for years, making the diagnosis of type 2 diabetes difficult. These symptoms can include increased urination, thirst or hunger, unexplained weight loss, numbness in the extremities, pain in the feet, blurred vision or recurring infections. In extreme cases loss of consciousness due to ketoacidosis can occur, though this is much more common in patients with type 1 diabetes. In order for clinicians to diagnose type 2 diabetes, an abnormal blood test can be used in conjunction with the presence of symptoms. If fasting plasma glucose concentration is greater than 7mmol/L (126mg/dL) then this is taken as confirmation of type 2 diabetes (World Health Organization, 2016). Furthermore, a test for glycated haemoglobin (HbA_{1C}) can be used to help diagnose type 2 diabetes. HbA_{1C} is an approximate measure of glucose control in the previous 2-3 months. The WHO recommend that an HbA_{1c} of 6.5% or greater should be used as a diagnosis for diabetes, although a lower HbA_{1c} does not exclude diabetes in the presence of a high glucose test (World Health Organization, 2011).

Treatment of type 2 diabetes aims to relieve or reverse symptoms and also to delay or prevent complications. Complications of diabetes occur due to damage to the small (microvascular) and large (macrovascular) blood vessels of the body caused by the chronic elevation of blood glucose: microvascular complications include retinopathy (damage to the eyes which can lead to blindness), nephropathy (damage to the kidneys which can lead to renal failure) and neuropathy (damage to the nerves which can lead to impotence and diabetic foot disorders) (Fowler, 2008). Macrovascular disease includes cardiovascular complications such as myocardial infarction (MI),

stroke, angina or transient ischaemic attack (TIA) (Fowler, 2008). Due to the risk of these problems, patients with type 2 diabetes require regular examination and screening (for example, eye exams, urine tests and foot care), as well as adequate education about potential complications in order to self-monitor for symptoms (NICE NG28, 2015).

The difficulty in diagnosing diabetes, the severe complications that require regular monitoring and the rapidly increasing number of diagnoses and diabetes-related deaths combine to pose one of the most challenging health problems of the 21st century.

1.2 Cardiovascular disease

Cardiovascular disease (CVD) is a general term which refers to a group of diseases of the heart and blood vessels (World Health Organization, 2015a). According to the WHO, CVD includes coronary heart disease (CHD), also referred to as coronary artery disease or ischaemic heart disease (IHD), cerebrovascular disease and peripheral arterial disease (PAD). CHD occurs when the blood flow to the heart muscle is blocked or reduced by a build-up of fatty deposits on the inner walls of the coronary arteries – a process referred to as atherosclerosis. As the arteries harden and swell, restricted blood flow to the heart can cause angina (chest pain). If eventually the artery is completely blocked then the heart is starved of oxygen which results in muscle cell death, or myocardial infarction (commonly known as a heart attack). Cerebrovascular disease encompasses both ischaemic stroke (cerebral cell death due to a lack of blood supply) and haemorrhagic stroke (bleeding from a cerebral blood vessel into the brain), as well as transient ischaemic attack (temporary reduction of blood flow to the brain without cell death, commonly called a “mini stroke”). Finally, PAD occurs when the blood vessels supplying the limbs, most commonly the legs, narrow due to atherosclerosis.

Despite considerable advances in the treatment of CVD, it remains the leading cause of death worldwide. Moreover, the rates of CVD and CVD deaths are expected to increase as the world population ages (Deaton et al., 2011). The incidence of and mortality due to CVD is particularly high in certain groups of the population,

including elderly people with type 2 diabetes (Halter et al., 2014). It is widely accepted that CVD is best prevented and treated in its early stages. Identifying patients who are most likely to benefit from an intervention and those for whom treatment is unnecessary can reduce patient morbidity, mortality and complications from unwanted side effects.

In the UK, CVD was the second most common cause of death in 2014, accounting for 27% of all deaths, according to the British Heart Foundation Cardiovascular Disease Statistics 2015 (Townsend N., 2015). The main causes of CVD deaths were CHD (45% of cardiovascular deaths) and stroke (25% of cardiovascular deaths). Of the four nations in the UK, Scotland had the highest CVD death rate for men and women, both combined and separately. Furthermore, CVD accounted for almost 1.7 million episodes in National Health Service (NHS) hospitals throughout the UK in 2015, a number which has been increasing in all UK nations over the last few years, and the incidence of CVD in the four nations is highest in Scotland among both men and women. Finally, recent studies have found that the CVD burden in the UK which can be attributed to diabetes is increasing (Kelly et al., 2009).

1.3 Conventional cardiovascular risk factors

Conventional cardiovascular risk factors can be categorised into two main groups: those which are non-modifiable (age, sex and ethnicity) and those which are modifiable (hypertension, dyslipidaemia, obesity, smoking, physical inactivity and, as previously discussed, diabetes).

It is well established that age is one of the biggest non-modifiable contributors to risk of CVD, with risk increasing as both sexes age (Castelli, 1984). Among adults, men are approximately twice as likely to develop or die from CVD as women (Lerner and Kannel, 1986). However, this difference between the sexes tends to reduce as women age and, in particular, reach the menopause (Kannel et al., 1976). Individuals from particular ethnic groups are known to be at an increased risk of CVD compared with those from other groups. For example, in the UK the prevalence of CHD is highest among Indian and Pakistani men (6% and 8% respectively). The incidence rate of MI is higher in South Asians than non-South Asians for both sexes and the stroke

incidence rate is higher in the Black ethnic group than in the White ethnic group for both sexes (Scarborough et al., 2010).

Modifiable risk factors can be categorised into two further general groups: high blood pressure and abnormal blood lipid measures, and lifestyle factors such as obesity, smoking and physical inactivity. High blood pressure, or hypertension, is one of the most powerful contributors to risk of CVD across all age groups and both sexes (Kannel, 1974), increasing risk by a factor of, on average, between 2- and 3-fold (Kannel, 1996). Abnormal blood lipid measures, also known as dyslipidaemia, occur when total cholesterol, low-density lipoprotein (LDL) cholesterol or triglyceride levels are raised, or alternatively when levels of high-density lipoprotein (HDL) cholesterol, a protective lipid, are low. The Prospective Studies Collaboration carried out in 2012 found a positive association between total cholesterol and CVD among adults, across all blood pressure levels, though no statistically significant association was found for stroke (Prospective Studies Collaboration, 2012).

The link between lifestyle factors and risk of CVD is well known. Obesity is most commonly measured using body mass index (BMI), the ratio of total body weight over height squared (kg/m^2), and individuals with a BMI greater than $30kg/m^2$ are considered to be obese. The global prevalence of obesity has increased dramatically over the last few decades, doubling between 1980 and 2008 (Bastien et al., 2014), and it has previously been shown that obesity is an independent risk factor for CVD (Poirier et al., 2006). The association between smoking and CVD is also well described and smoking is considered to be one of the most preventable causes of CVD (Lakier, 1992). A 50-year prospective cohort study carried out in the UK found that mortality rates tripled due to prolonged cigarette smoking, and specifically the mortality rate for CVD increased with smoking (Doll et al., 2004). Finally, it has been shown that increased physical activity has a protective effect for the development of CVD (Berlin and Colditz, 1990). A recent study found that physical inactivity causes 6% of the burden of disease worldwide that is due to CHD (Lee et al., 2012).

1.4 Non-traditional biomarkers

Although the assessment of conventional cardiovascular risk factors remains vital for disease prediction and prevention, better risk stratification using additional biomarkers may allow for more targeted use of current prevention strategies and new treatments and the reduction of people on unnecessary treatment. Recent studies have suggested that there is the potential for a range of physical and circulating biomarkers (beyond the conventional risk factors already mentioned), to add value to vascular risk prediction (Gerszten and Wang, 2008). Such biomarkers, which will be discussed in the following sections, include the ankle brachial index (ABI), N-terminal pro-brain natriuretic peptide (NT-proBNP), high-sensitivity cardiac troponin T (hs-cTnT), gamma-glutamyl transpeptidase (GGT) and a group of inflammatory biomarkers – C-reactive protein (CRP), interleukin-6 (IL-6), tumor necrosis factor alpha (TNF-α) and fibrinogen. This panel of non-traditional biomarkers was selected based on the available vascular biomarkers measured in the Edinburgh Type 2 Diabetes Study (ET2DS), which was used for statistical analyses in this thesis. Whilst the ET2DS was originally designed to investigate risk factors for vascular cognitive impairment, it has been developed over 10 years to include a wide range of topical and relevant biomarkers relating to macrovascular risk prediction, through collaboration with clinical and biochemical experts in this field. Three available biomarkers were excluded from the panel in order to reduce the dimensionality of the data and allow the thorough investigation of a small number of predictors. Apolipoproteins A1 and B were included in the metabolomics data set which was used later in this thesis. Carotid intima-media thickness was only measured at year 1 of the ET2DS which would have resulted in a large number of missing data, impacting the investigation of all other biomarkers, and there was not compelling evidence in the literature of an association with cardiovascular events to support its inclusion in analyses.

1.4.1 Ankle Brachial Index

ABI (also referred to as ankle brachial pressure index, ABPI) is the ratio between the systolic blood pressure (sBP) in the ankle and that in the upper arm (the brachium), and is used clinically in the assessment of PAD of the lower limbs since low blood

pressure in the legs, in comparison to the arm, suggests narrowed arteries due to atherosclerosis. The Scottish Intercollegiate Guidelines Network (SIGN) clinical guidelines state that an ABI of less than 0.9 indicates PAD (SIGN, 2006). However, a very large ABI is also considered to be abnormal, indicating that the walls of the arteries have become hardened (a process referred to as calcification) and are incompressible. The exact upper cut-off point which marks an abnormal ABI measurement is debated, though a value of 1.4 is most commonly supported in the literature (Allison et al., 2008). As well as indicating PAD, it has been suggested that ABI is a marker of generalised CVD. A recent meta-analysis including nearly 50,000 people found that a low ABI was associated with total mortality, cardiovascular mortality and major coronary outcomes even after adjusting for the Framingham Risk Score (Fowkes et al., 2008).

1.4.2 NT-proBNP

Brain natriuretic peptide (BNP) is a 32 amino acid polypeptide which is released by the heart in response to increased stress on the heart wall. On secretion, BNP splits into the biologically active peptide and non-functional N-terminal fragment (NT-proBNP). Both plasma BNP and NT-proBNP levels are currently used in clinical practice to diagnose patients with heart failure, and to assess how severe the heart failure may be. Raised levels of BNP or NT-proBNP are typical in patients with acute heart failure and the National Institute for Health and Care Excellence (NICE) guidelines recommend that this measure is used in patients presenting with new suspected heart failure (NICE CG187, 2014). Several general population cohort studies have shown that NT-proBNP is also strongly associated with the risk of CVD and appears to have added prognostic value independent of conventional risk factors (Welsh et al., 2013; Linssen et al., 2010; Kistorp et al., 2005; Wang et al., 2004).

1.4.3 Troponin

Troponin is a complex of three proteins, troponin C, troponin I and troponin T, that are part of both skeletal and cardiac muscle. Cardiac troponin levels increase in response to clinical and subclinical myocardial ischaemia. In recent years, cardiac troponin T has been introduced to clinical practice in order to aid the diagnosis of MI in patients presenting with chest pain (NICE DG15, 2014). Specifically, high

sensitivity laboratory analysis now allows for precise measurements at very low concentrations of cardiac troponin T (hs-cTnT) (Hillis et al., 2014). In a number of general population studies, troponin has been shown to be strongly associated with the risk of CVD over and above the contribution of conventional risk factors (Saunders et al., 2011; Everett et al., 2011).

1.4.4 GGT

GGT is an enzyme most commonly found in the liver and is an important diagnostic biomarker in liver disease, since levels increase when the liver is damaged. In recent years two large general population cohort studies have found a significant association between GGT and CVD (Lee et al., 2006a and Jousilahti et al., 2000). Furthermore, a cohort study of 283,438 participants found an association between increased GGT and all types of cardiovascular mortality in both men and women (Kazemi-Shirazi et al., 2007).

1.4.5 Inflammatory biomarkers

A group of inflammatory biomarkers which can be measured to assess internal injury (CRP, IL-6, TNF-α and fibrinogen) are all proteins circulating in the blood whose levels increase in response to systemic inflammation. The potential of these non-traditional biomarkers as predictors of CVD risk has been investigated in recent studies, but their added value over and above conventional risk factors or even other non-traditional biomarkers remains unclear (Olsen et al., 2007; Wannamethee et al., 2011; Bettencourt et al., 2011; Berg and Scherer, 2005; The Emerging Risk Factors Collaboration, 2012).

1.5 Metabolomics

Omics data collectively describe the structure, function and dynamics of the body. They range from the genome, the most stable type of omics data, through the epigenome, transcriptome, proteome to the metabolome, the most unstable type of omics data (Chadeau-Hyam et al., 2013). Modern biological technologies are able to take advantage of well-phenotyped cohorts with stored biological samples and provide extensive omics data sets for novel analysis (Gehlenborg et al., 2010). Metabolomics are measurements of small molecules in biological samples, such as

blood, urine, saliva or tissue, and represent end products of cellular processes. Currently metabolomics data can be measured using two major technologies: nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (Larive et al., 2015; Keun and Athersuch, 2011; Di Girolamo et al., 2013). NMR is the only technique which does not require destruction of the biological sample, which can therefore be re-used for further analysis. Other key advantages of NMR are that it is highly reproducible and requires very simple sample preparation. However, it can be insensitive compared to the mass spectrometry technique which is highly sensitive. Mass spectrometry requires destruction of the sample, is less reproducible than NMR and produces complex data which can be difficult to analyse.

Recent studies have shown that a number of specific metabolites measured using omics technology are strongly associated with both all-cause mortality and specific cardiovascular outcomes independent of traditional risk factors (Soininen et al., 2015, Fischer et al., 2014). It has been suggested that the use of metabolomics data may help to add value to cardiovascular risk prediction and, in particular, lead to better prediction than the use of conventional lipid measurements such as total or HDL cholesterol (Würtz et al., 2012, Wurtz et al., 2015). Relevant investigations to date have been limited to a small number of metabolites and have not focused on the subgroup of the population with type 2 diabetes.

1.6 Cardiovascular risk scores

There are many different scoring systems that have been developed over the last few decades in order to estimate the risk of an individual developing CVD. Such scores rely heavily on conventional cardiovascular risk factors and are discussed in detail in this thesis in a systematic review of the literature in Chapter 4. However, only a few of these risk scores have been developed specifically for populations with type 2 diabetes, whereas most others have used general populations to create a risk score, which can then be applied to a subgroup of people with diabetes provided that the model development includes diabetes as a predictor. Furthermore, in general, current risk scores appear to perform inadequately for people with type 2 diabetes, either under- or over-estimating the risk of cardiovascular events (Simmons et al., 2009, van der Heijden et al., 2009, van Dieren et al., 2012).

The most well-known cardiovascular risk score is the Framingham Risk Score, which was developed using data from the Framingham Heart Study (Mahmood et al., 2014). The Framingham Heart Study is an on-going observational study, which began in 1948 with 5209 adult participants from the town of Framingham, USA, and is now on its third generation of subjects. The Framingham Heart Study has produced a wide range of interesting and useful results, many of which were previously unknown and are now well-accepted in public health. It also produced the Framingham Risk Score, first published in 1998 (Wilson et al., 1998), which is an algorithm that calculates the ten-year cardiovascular risk of an individual. The current version of the Framingham Risk Score was published in 2002 (Framingham Heart Study, 2002). The first version of the risk score included age, sex, LDL cholesterol, HDL cholesterol, blood pressure (and also whether the patient is treated or not for hypertension), diabetes and smoking. The updated version was modified to include dyslipidaemia, age, hypertension treatment, smoking and total cholesterol.

Other examples of cardiovascular risk scores developed in general populations are SCORE and DECODE. SCORE is the Systematic Coronary Risk Evaluation Project risk score, which was developed by The European Society of Cardiology (Conroy et al., 2003). It is a large dataset derived from 12 prospective European cohort studies (total number of participants is 205,178) and predicts ten-year risk of fatal atherosclerotic cardiovascular events. This risk estimation uses gender, age, smoking, sBP and total cholesterol as risk factors, but no diabetes variable. DECODE is the Diabetes Epidemiology: Collaborative Analysis of Diagnostic Criteria in Europe Study Group (Balkau et al., 2004). It consists of 14 European studies (a total of 25,413 subjects) and calculates both the five- and ten-year risk scores for cardiovascular mortality. The risk factors included in the models were age, cholesterol, smoking status, sBP, BMI and, for the first time in a risk score, fasting and 2-hour glucose measures.

A UK-specific risk score for the general population that has been developed is the QRISK2 score calculator (Hippisley-Cox et al., 2008). The QRISK2 calculator was developed by doctors and academics working in the UK NHS and is updated annually to keep it as accurate as possible. It is noted that although QRISK2 has been

developed for use in the UK, it can and is being used internationally. For non-UK use, since social status is based on UK postcodes, it is highlighted that users should be aware that CVD risk is likely to be under-estimated in patients from deprived areas and over-estimated for patients from affluent areas. QRISK2 calculates the 10-year risk of having MI or stroke and includes diabetes as one of 14 risk factors (including age, sex, ethnicity, social status and smoking status).

One example of a risk score which has been developed specifically in patients with type 2 diabetes is the UK Prospective Diabetes Study (UKPDS). UKPDS was a randomised trial involving 5102 participants with newly diagnosed type 2 diabetes which ran for twenty years between 1977 and 1997. In 1997 all surviving UKPDS patients were entered into a ten-year, post-trial monitoring programme. Data from the UKPDS was used to develop the UKPDS Risk Engine (Bannister et al., 2014), a type 2 diabetes specific risk score which can calculate the 10-year risk for non-fatal and fatal CHD, or non-fatal and fatal stroke. The models are based on current age, sex, ethnicity, smoking status, presence or absence of atrial fibrillation and levels of HbA1c, sBP, total cholesterol and HDL cholesterol.

This thesis explores the incorporation of biomarkers into cardiovascular risk scores aimed at improving the prediction of cardiovascular risk in people with type 2 diabetes. The specific aims and objectives are outlined in Chapter 3. In the following chapter (Chapter 2) statistical methods which are used to develop and evaluate risk scores are introduced and discussed.

2 Statistical methods for risk prediction

2.1 Characteristics of risk prediction models

Prediction models formally combine multiple predictors in order to calculate the risk of a chosen future outcome (this contrasts with a diagnostic model, which aims to identify an existing, but unknown, disease). Building prediction models requires a thorough and complex process. Moons et al., 2009, propose a three-stage procedure when developing new prediction models: development studies, validation studies and impact studies. Development studies allow for the initial building of a multivariate prediction model; validation studies are used to evaluate the chosen model's predictive performance; and then impact studies quantify whether the model is able to improve treatment in a practical clinical setting. Impact studies can also be used to explore the effect of using a prediction model on clinical management, patient outcome and cost effectiveness of treatment.

Prediction models use a combination of multiple variables to predict the risk of future disease outcomes for groups of patients or individuals. They are able to identify whether newly discovered risk factors can contribute to risk prediction. The results from such models inform patients about their future and can be used to guide treatment decisions for both the patient and their doctor. The use of prediction models also extends to the selection of patients for clinical trials and the comparison of the performance of different hospitals or health centres.

Since the clinical implications are significant, care must be taken when building a prediction model. This begins at the design stage of a study even before any statistical analysis takes place. Moons et al., 2009, outline the principles of a prediction study and state the following guidelines. The objective of a prediction study is to determine the risk of a future health outcome in a population using predictor variables. The ideal study design for a prediction study is a prospective cohort study. This is because prospective cohort studies are best suited to collecting complete and accurate data on the outcome and multiple predictors. Cohort studies also allow several different disease outcomes to be investigated. The outcome of interest should be specified in advance and be relevant to patients in a practical

setting. A range of predictors may be considered, but these must be well-defined, standardised, reproducible and measured using methods which are applicable in clinical practice. This allows the results of an eventual model to be utilised in a clinical setting and therefore directly benefit patients. The population under investigation should be clearly defined and the study sample is described as a group of people at risk from the specified outcome.

Once reliable and accurate data have been collected, a model can then be built using the outcome of interest and a selection of predictors. The predictor variables must be chosen from a potentially large list and Royston et al., 2009, give some proposals as to how this can best be done. They suggest that variables that have already been confirmed to be predictive should usually be included in a model. In general, predictors should only be included if measurements are high quality and can be compared across doctors and study centres. Royston et al., 2009, advise that continuous variables should not be dichotomised as otherwise valuable predictive information can be lost. There are various methods that can be used to select variables for a model, but Royston et al., 2009, warn that significance testing produces selection bias and over-optimistic results. Finally, it is recommended that once a final model has been chosen all the coefficients should be included in any reporting of results in order for the risk scores to be reproduced in other populations.

In general, a prediction model should be simple and easy to interpret, particularly for doctors if clinical use is the goal. It is rarely the case that one predictor gives an acceptable estimate of risk, therefore models will usually be multivariable. Prediction models require updating so that existing markers which inadequately predict risk can be removed or replaced. Furthermore, changes in treatment or clinical management over time may change health outcomes independently of the predictors in a prediction model.

2.2 Selecting biomarkers for risk prediction models

As previously discussed, in the study of CVD, modification of traditional vascular risk factors has a vital role to play in disease prevention. However, more recently, additional biomarkers have been explored as a means of supplementing information

obtained from the conventional risk factors (Gerszten and Wang, 2008). These new biomarkers require thorough investigation in order to assess and quantify any improvement in risk prediction that they may add to an existing risk prediction model. It should be noted that biomarkers can be categorised as predictive or causal, and that these categories can overlap to give a biomarker that is both predictive and causal. In this thesis I am interested in predictive biomarkers which may not necessarily contribute to the causal pathway. There are four key requirements that a new predictive biomarker must meet before it should be accepted in a model: statistical significance, large effect size, independent contribution over-and-above other measured risk factors and usefulness in a clinical setting (Pencina et al., 2008, Pencina et al., 2010). The most basic of these requirements is that there should be a statistically significant association between the predictor and the outcome of interest. This reduces the possibility that the association observed is due to chance, although it should be noted that this does not guarantee clinical relevance. A large effect size observed in a particular predictor usually indicates that there is some "gain" in the performance of the model. It is important to check that a new marker is associated with the outcome of interest independently of other existing risk factors, and this should be accounted for in the statistical analysis. An example of how this can be achieved is to use multivariate models that adjust for other known risk factors. Clinical significance, as well as statistical significance, should always be considered. In general we are most interested in whether a new biomarker, alone or in combination with established predictors, is able to more accurately categorise people into clinically meaningful high or low risk groups. It is also important to establish whether invasive or expensive markers have significant added value in comparison to cheaper or easily obtained predictors (Moons et al., 2009).

However, an important question to ask is how much improvement in a model is really possible. It should be expected that new markers are highly correlated with the original predictors already included in the model (Hand, 2006). If this is the case, the new marker's contribution to the model may be diminished and there will be a limited amount of gain in model performance. Furthermore, models are built using baseline variables which may change during the follow-up period, as may clinical practice. Both of these effects can limit the application of a developed model.

Pencina et al., 2010, conclude that due to these factors lack of perfection may be an inherent feature of risk prediction models, although this should not discourage future research into the subject.

The most commonly used and well-known variable selection methods are stepwise regression methods. Steyerberg outlines these techniques in his book "Clinical Prediction Models" (Steyerberg, 2009). In general terms, stepwise regression methods are automated processes of building a model. A final model is chosen by successively adding or removing candidate variables based on some pre-specified criterion such as an F- or t-test, the coefficient of determination (R^2), Akaike Information Criterion (AIC) (Akaike, 2011) or Bayes Information Criterion (BIC) (Neath and Cavanaugh, 2012). There are three main approaches to stepwise regression: forward selection, backwards selection (also known as elimination) and a combination of both forward and backward selection. Forward selection begins with no candidate variables in the model and adds the most significant candidate variable. This process is repeated until no significant variable remains outside the model. Once a variable enters the model it cannot be removed. Backward selection begins with a full model including all the candidate variables and removes the least significant candidate variable. Again, this process is repeated until no non-significant variables remain in the model. Forward and backward selection modifies the procedure of forward selection: after each iteration of the process all candidate variables currently included in the model are checked to see if their significance has dropped below the pre-specified level. If this is the case then the non-significant variable is removed from the model.

Several issues must be considered when choosing a stepwise regression approach. Forward selection can be an appropriate choice when carrying out an initial screening on a large number of variables in order to obtain a smaller panel of potential predictors. Furthermore, if multicollinearity, where two variables are highly correlated, is a concern then it is likely that forward selection will include neither variable in the model. However, this can result in important predictors being excluded from the model and therefore backward selection may be preferred. Backward selection may be preferred in the case of a smaller set of candidate

variables which have already been fine-tuned and which you wish to reduce further. Backward selection also has the advantage of starting with the full model which means that the effects of all candidate variables can be assessed simultaneously. One disadvantage of backwards selection is that variables which are not really necessary may end up in the final model.

An extension of stepwise regression approaches is the "all subsets regression" or "best subsets regression". In this approach all possible models derived from all possible combinations of candidate variables are assessed and the subset of predictors that does the best, according to a pre-specified criterion, is selected. All subsets regression has a key advantage over stepwise regression in that it can identify combinations of predictors not found by these forward or backward selection approaches. It also enables the identification of variables which consistently appear in most or all of the "best" models. However, Steyerberg warns that over fitting (where the model performs well on the original data, but performs poorly for future observations) can be a problem. Furthermore, different pre-specified criteria can result in different "best" models. Therefore it is vitally important that all subsets regression is not misused by claiming that it results in the one best model: rather, it provides a useful screening tool to reduce the number of possible regression models to a manageable amount which can be further explored, evaluated and refined in order to finally select one model. As discussed above, automatic methods are useful when the number of candidate variables is large, in which case it is not efficient to fit all possible models.

2.3 Omics data in risk prediction modelling

Specialised statistical methods must be used when dealing with omics data. This type of data creates the "n < p" situation where the number of predictors (p) is larger than the number of outcomes, or even the number of subjects (n). In this case the commonly used linear model no longer applies and, in general, standard statistical methods risk over fitting the data and/or observing false positives. For the situation of more predictors than outcomes, standard methods may lead to over fitted models that produce inaccurate results, so analysis must be carried out carefully (Pavlou et al., 2015).

Three approaches are presented by Chadeau-Hyam et al. in their paper "Deciphering the Complex: Methodological Overview of Statistical Models to Derive OMICS-Based Biomarkers" for analysing omics data appropriately: a univariate approach, dimension reduction and variable selection (Chadeau-Hyam et al., 2013). These three approaches are discussed in detail in the following sections.

2.3.1 Univariate approach

A univariate approach considers one predictor at a time, assuming the predictors to be independent. This approach is computationally efficient, accommodating extremely large numbers of predictors. It also allows for greater modelling flexibility than the other approaches, since the correlation structure between the predictors does not need to be modelled. It is straightforward to adjust models for potential confounders and models can be adapted to cope with all types of predictors and outcomes. However, since only the marginal effect of each predictor on the outcome is described, the models do not account for potential combined effects of predictors which may have an important role.

In a univariate approach the same model is fitted to each predictor, giving p models and p-values, where p is the total number of predictors. Each p-value is compared to an arbitrary threshold, α, which defines the risk of wrongly rejecting the null hypothesis (Type I Error). This leads to p conclusions regarding the significance of the associations. However, as the number of tests increases, the size of α also increases. Therefore the number of tests performed must be accounted for during analysis. Multiple testing corrections can be achieved either by adjusting the p-values or by altering α, and are usually carried out based on two main methods: using the Family Wise Error Rate (FWER) or the False Discovery Rate (FDR). The FWER is the probability of obtaining at least one false positive, and is a more severe control than the FDR which is the expected proportion of false positives among all significant associations. To use the FWER, a per-test significance level (a new α') is defined. An example of a multiple testing correction using the FWER is the commonly used Bonferroni correction which defines a new $\alpha' = \alpha/p$ (Bland and Altman, 1995). However, the Bonferroni correction can be too stringent (Simes, 1986) and so alternatively, provided tests can be assumed as independent, the Šidák

correction can be used which defines a new $\alpha' = 1-(1- \alpha)^{1/p}$ (Sidak, 1967). To use the FDR a per-test significance level, α', is defined ensuring that the FDR is bounded above by a desired value. An example of a multiple testing correction using the FDR is the Benjamini-Hochberg correction which is a step-up procedure comparing p-values sorted in ascending order to increasingly stricter cut-off values (Benjamini and Hochberg, 1995).

2.3.2 Dimension reduction

The aim of dimension reduction is to map the original predictors into a lower dimension using components which accurately reconstruct the structure of the original data. Dimension reduction approaches allow the visualisation of high dimensional data and are useful if there are irrelevant features of the data, such as noise features. They also permit the correlation structure of the data to be taken into account during the analysis, removing potential multi-collinearity. However, the key disadvantage of data reduction is that the interpretation of the new components can be difficult and unintuitive, especially given many original predictors. One option to improve interpretation is to ensure sparse results (that is, constraining the number of predictors included in a model), which will be discussed below.

There are two types of methods which can be used in dimension reduction: unsupervised and supervised methods. Unsupervised methods do not guarantee that the components are explanatory of the outcome and may be driven by noise in the data. However, a well-known example of an unsupervised method is Principal Components Analysis (PCA) (Everitt et al., 2013). The PCA procedure is to compute the correlation or covariance matrix and perform an eigen-analysis. This provides a set of new variables, or components, made up of uncorrelated linear combinations of the original variables. The resulting eigenvalues give the variance of each component and the eigenvectors give the loading of each component, which can be considered in order to establish appropriate interpretation for the independent components. By construction the components are independent and the total variance is preserved under the principal component transformation. Dimension reduction is usually achieved without substantial loss of information by working with the first *k* components. PCA is computationally efficient to perform, even for hundreds or

thousands of variables, and can cope with both continuous and discrete data. An extension of PCA is the sparse PCA which aims to remove irrelevant variables by computing sparse components. This can aid interpretation of components, as mentioned above.

An alternative supervised method for dimension reduction is Partial Least Squares (PLS) (Chadeau-Hyam et al., 2013) which chooses components in order to maximise the covariance between the predictors and the outcome. Therefore the final components will be those that are most correlated with the outcome. The method is an iterative procedure: in the first iteration, the linear combination of predictors which best describes the outcome is calculated, then subsequent iterations are performed to incorporate any further structure in the data. As with PCA, PLS can be extended to sparse PLS which selects only the relevant predictors by penalising the loading vector.

2.3.3 Variable selection

Variable selection aims to identify a sparse set of predictors that jointly predict the outcome. This means that variable selection approaches implicitly correct for multiple testing. Penalised regression is a general variable selection approach which estimates the model coefficients under certain constraints. It is computationally efficient, provides easily interpretable results and accommodates all outcome types. Two particular methods of penalised regression are ridge regression and Least Absolute Selection and Shrinkage Operator (LASSO) models (Chadeau-Hyam et al., 2013, Steyerberg, 2009). Ridge regression penalises the size of the regression coefficients, shrinking them towards zero. Specifically, given a response vector Y with *n* outcomes and a predictor matrix X with *p* predictors, the ridge regression coefficient β^{ridge} is defined as the value of β which minimises

$$\beta^{ridge} = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} {\beta_j}^2$$

where λ is a tuning parameter which controls the strength of the penalty term. Ridge regression has the advantage of being numerically stable when n < p and is beneficial

in the presence of multicollinearity, but it does not guarantee sparsity. LASSO models ensure sparsity of results through the following constraint which minimises

$$\beta^{LASSO} = \sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

where again λ is a tuning parameter. The nature of this constraint causes some coefficients to be shrunk all the way to zero and therefore allows LASSO to perform variable selection by removing unnecessary predictors.

Elastic Net is a flexible combination of ridge regression and LASSO where the constraint is defined as a weighted sum of both ridge regression and LASSO constraints. This method has the advantage that it is both numerically stable and sparse, but comes at the cost of an additional parameter to tune.

Both ridge regression and LASSO models require a tuning parameter, λ, to be selected which controls the strength of the penalty. The choice of λ can be determined using a cross-validation procedure where the optimal value of λ will be chosen to minimise the prediction mean square error. For LASSO models, as λ increases, the number of variables selected in the final model decreases. In the case of ridge regression, when $\lambda = 0$ the solution is similar to an ordinary least squares approach.

2.4 Assessing a prediction model

After biomarkers have been selected and a prediction model has been built, but before application in a clinical setting, it is vital to assess the performance of a prediction model in order to quantify its usefulness in statistical and clinical terms. There are numerous reasons why a model may not predict well: bad original study design, poor choice of modelling methods, over fitting, the absence of a key predictor or differences in either populations or methods of measurement. In prediction models the concepts of discrimination and calibration are of key interest as they both provide ways to assess the performance of a model, and these are discussed in detail in the following sections. It should be noted that there is always a trade-off between discrimination and calibration – a model cannot be perfect in both

(Cook, 2007). Therefore, both calibration and discrimination should be assessed when carrying out a model evaluation.

2.4.1 Discrimination

Discrimination describes the model's ability to distinguish between those patients who do or do not experience the health outcome of interest (Tripepi et al., 2013). Discrimination can be assessed in a number of different ways, ranging from the commonly used receiver operating characteristic (ROC) curves and c-statistics (Cook, 2008), to reclassification tables (Kerr et al., 2014), to more novel techniques such as net reclassification index (NRI) (Pencina et al., 2008), integrated discrimination improvement (IDI) (Pencina et al., 2008) or predictiveness curves (Pencina et al., 2010).

The most popular method of assessing discrimination is the ROC curve and the corresponding area under the curve (also referred to as the c-statistic). In order to calculate these measures we need to obtain the sensitivity and specificity of a model. Sensitivity is the probability of a "positive test" among those patients who do experience the outcome of interest. Specificity is the probability of a "negative test" among those patients who do not experience the outcome of interest. Here a "test" could be the result of a single binary prognostic test, but it could also be defined as the observation of a binary outcome, in which case we are interested in the classification of individuals into risk categories for the outcome. When comparing models, we favour those with higher sensitivity values, or a higher probability of classifying a patient who does have an event into a high risk category, and also those with higher specificity values, or a higher probability of classifying a patient who does not have an event into a low risk category. Specificity and sensitivity are unaffected by disease prevalence, though they can be effected by case mix, severity of disease, selection of control subjects, measurement technique, quality of the gold standard and risk factors (Cook, 2008, Cook, 2007). Additionally, specificity may be affected by the characteristics (for example age, gender, prevalence of risk factors) of people who do not experience the outcome of interest. The ROC curve is then a plot of sensitivity against 1-specificity (Figure 2-1) which provides a summary of

sensitivity and specificity, assessing how well a model separates individuals into two groups.

The c-statistic is the area under the ROC curve and is equal to the probability that a prediction model assigns a higher probability of an event, or higher risk score, to those who do actually experience an event or do actually belong in the high risk group. The c-statistic is based on the ranks of the predicted probabilities and compares these ranks in patients who do and do not experience the outcome of interest. The range of the c-statistic is between 0.5, which indicates no predictive ability in the model (marked by a green dotted line in Figure 2-1), and 1, which indicates that the model has perfect discrimination (marked by a blue solid line in Figure 2-1). Perfect discrimination only occurs if the scores for all those who do suffer from an event are higher than those who do not, with no overlap.

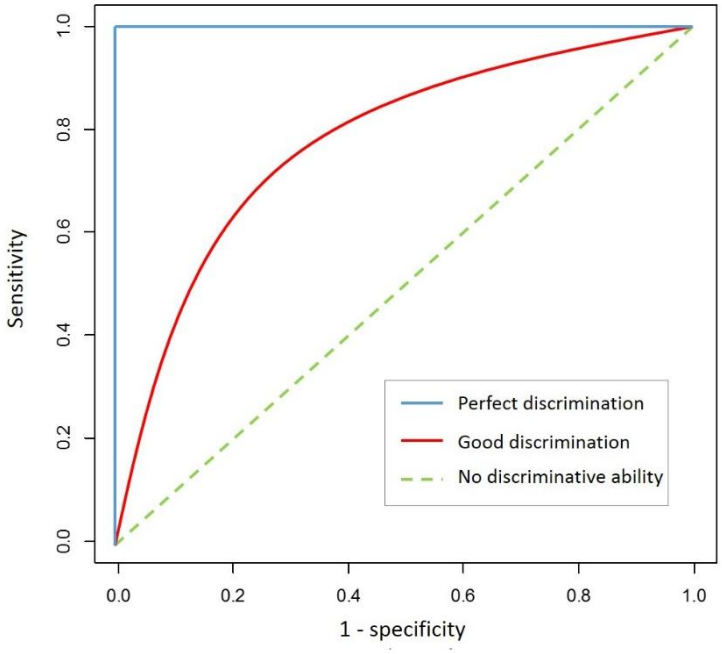


Figure 2-1: Comparison of three ROC curves with varying discriminative abilities

There is some discussion in the literature over the usefulness of the c-statistic in the statistical evaluation of prediction models. Cook, 2008, showed that the c-statistic can be insensitive when adding a new predictor to a model if the odds ratio (OR) for that marker is not extremely large (e.g. >16 per 2 standard deviations). Although a new predictor may have an independent and statistically significant contribution to a

prediction model, there can be very little improvement in the ROC curve. This is particularly noticeable when the basic model being used for comparison includes strong predictors and has a large c-statistic. The impact of the new predictor on the c-statistic will be lower since it is always hard to greatly improve a good model (Pencina et al., 2010). Finally, the c-statistic does not take the distribution of patients in terms of risk level into account, for example if there are a small number of patients who are at high risk and a large number of patients who are at very low risk (Cook, 2007). The conclusion of these papers is that researchers should not solely rely on the c-statistic when evaluating the discriminative ability of a prediction model, as novel biomarkers could still lead to a more accurate risk score despite little change in the c-statistic. However, it should still be valued as a summary of a model's discriminative ability, and is particularly useful when the goal is to estimate an optimal threshold for clinical use.

A further issue that can arise with regards to the c-statistic is in the case of time to event data and subsequent survival analysis. Ignoring the information captured by the time to event can result in a biased c-statistic. Therefore, an alternative statistic was proposed by Harrell et al., 1996, called the concordance index. The concordance index is defined as the proportion of usable pairs of patients, one with and one without the outcome of interest, in which the patient who did experience the event has a higher predicted probability. An unusable pair is regarded as a case where the patient with the shorter follow-up time did not experience the event and hence the true order of the pair remains unknown (Gerds et al., 2013). Similar to the c-statistic described above, the concordance index ranges from 0.5 (no discrimination) to 1 (perfect discrimination). Harrell et al., 1996, warn that, as with the c-statistic, the concordance index is not sensitive to minor differences between the discriminative abilities of two models.

Two alternative measures of discrimination were proposed by Pencina et al., 2008: the NRI and the IDI. The NRI evaluates the proportions of people moving up and down risk categories and calculates these proportions by examining people who do and do not experience the outcome of interest separately. The NRI can be used to measure the change in predictive performance when a new biomarker is added to a

basic model. It is only affected by people who change risk categories, since the model would give the same prediction for those who remain in the same category. A movement upwards in risk category for someone who has the outcome of interest is deemed as improving risk classification, because we want to increase the predicted probability of an event for someone who does indeed experience that event. For people who do not experience an event, downwards movement results in better risk classification. Pencina et al., 2008, suggest using the NRI as a simple, but meaningful check of classification accuracy. The NRI is calculated using the following formula:

$$\text{NRI} = [\text{Proportion(up|event)} - \text{Proportion(down|event)}]$$

$$- [\text{Proportion(up|no event)} - \text{Proportion(down|no event)}]$$

Pencina et al., 2008, highlight a key disadvantage of the NRI: that the method is dependent on the choice of risk categories. It is useful if clinically relevant risk categories already exist, but otherwise an alternative method is required. This alternative measure is the IDI, which is an extension of the NRI when no risk categories are established. The integrated sensitivity (IS) is defined as the integral of sensitivity between 0 and 1, and can be considered as the average sensitivity. The integrated specificity (IP) is the integral of '1-specificity' between 0 and 1, and can be considered as the average '1-specificity'. The IDI is then the difference in differences between the integrated sensitivity and 1-specificity for models with and without the new marker and is calculated as follows:

$$\text{IDI} = (\text{IS}_{\text{new}} - \text{IS}_{\text{old}}) - (\text{IP}_{\text{new}} - \text{IP}_{\text{old}})$$

where "new" is the model including the new marker of interest and "old" is the basic model without the marker. The IDI is equal to the difference between improvement (or decline) in average sensitivity and improvement (or decline) in average 1-specificity. Pencina et al., 2008, conclude that they still suggest that the c-statistic should be the first check of discrimination, but that, due to some of the aforementioned issues with that measure, NRI and IDI should also be taken into consideration.

A recent paper by Kerr et al., 2014, also cautions the use of NRI. They warn that care must be taken when interpreting the NRI, because although it combines four proportions, the NRI is not a proportion itself. They also highlight the fact that irrelevant information is included in the NRI, since the NRI does not account for the size of changes in predicted risk – a small and clinically irrelevant change in risk score will still contribute to the NRI. In the case of more than two risk categories, a move from the lowest category to the middle category is treated as the same as a move from the lowest category to the highest category. Furthermore, there is a danger that an uninformative new biomarker could appear to have considerable predictive value, even if the NRI is calculated using a large, independent validation dataset. This is not a problem for the c-statistic. The authors suggest that the presentation of reclassification tables may give a more informative summary of the classification with a new biomarker. One approach to the NRI could be to use it as a descriptive tool after model checking to demonstrate what would happen to risk scores in a clinical setting if the new model was used, but not to rely on it as a formal model comparison tool.

Other measures of discrimination are the positive and negative predictive values. The positive predictive value (PPV) is the probability of the outcome of interest given a "positive test" result. The negative predictive value (NPV) is the probability of no event given a "negative test" result. The PPV and NPV can be useful measures to complement sensitivity and specificity, although they are dependent on the prevalence of disease (Cook, 2008).

Even more novel ways to explore discrimination than the NRI have been proposed, such as using predictiveness curves, which are a general method of assessing the usefulness of a prediction model and the classification performance of that model. The curves are constructed by ordering all the predicted probabilities from lowest to highest and plotting them against the observed risk percentiles (Figure 2-2) (Pepe et al., 2008). The ideal shape of a predictiveness curve is one that stays close to the horizontal axis and then increases rapidly (Pencina et al., 2010).

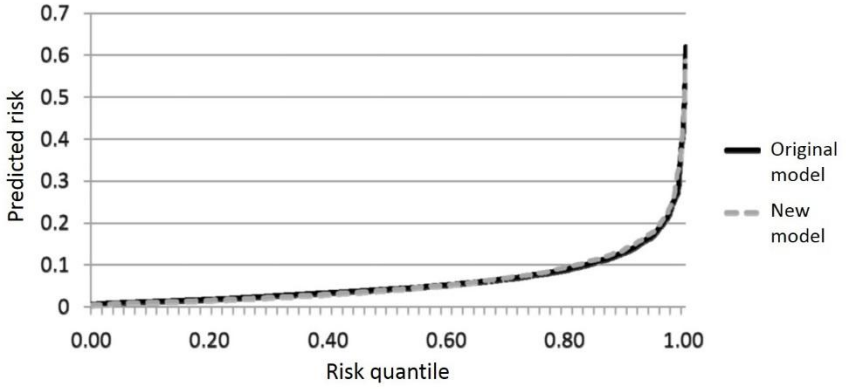


Figure 2-2: Predictiveness curves for two competing models. Adapted from Pencina et al., 2010.

In summary, there are numerous ways of assessing the discriminative ability of a prediction model. Although there is no agreed procedure for this evaluation, the literature gives the impression that a suitable approach would be to first calculate the ROC curve and corresponding c-statistic and also to present a reclassification table. The importance of this table should not be ignored, since a key concern for clinicians is whether or not a new biomarker accurately classifies patients into higher or lower clinically meaningful risk categories. It may be useful to report the positive and negative predictive values, as well as information on the context of the situation e.g. prevalence rates, alongside the sensitivity and specificity. An additional measure such as NRI could be used to supplement this information, in order to show how the risk score would change in a real life setting, but this should not be used alone and the weaknesses of this measure should be noted.

2.4.2 Calibration

As well as discrimination, calibration should be assessed when evaluating a new prediction model. Calibration is a model's ability to correctly estimate the risk of a future event and is a measure of how well the predicted probabilities agree with the observed risk that later develops (Tripepi et al., 2010). Calibration directly compares observed and predicted event rates for groups of patients. If a model is well calibrated then the event rates predicted by the model should closely correspond with those that are observed in practice. As with discrimination, there are various ways to evaluate calibration. The Hosmer-Lemeshow test (Lemeshow and Hosmer, 1982) can be used, or alternatively less formal methods such as plotting observed events against predicted events for different ranges of predicted risk.

The Hosmer-Lemeshow test is the most widely used measure of calibration and assesses the "goodness-of-fit" of a model, by comparing the observed number of events with the number predicted by the model. The null hypothesis assumes a well calibrated model and states that the predicted and observed probabilities of the event do not differ. The test is performed by forming subgroups of the data and, within each subgroup, computing the estimated and observed probabilities of an event for each subject. The test statistic is calculated using the following formula:

$$X^2 = \sum_{i=1}^{g} \frac{(observed_i - expected_i)^2}{expected_i}$$

and follows a χ^2 distribution with g-2 degrees of freedom, where g is the number of subgroups (Lemeshow and Hosmer, 1982). However, there are weaknesses with the Hosmer-Lemeshow test which researchers should be aware of when using it to assess model calibration. The Hosmer-Lemeshow test is sensitive to the choice of subgroups (Hosmer et al., 1997) and can be seen as a fairly crude measure since we cannot know the underlying risk for each patient, but only observe whether they experience an event or not (Cook, 2008). Kramer and Zimmerman, 2007, advise that caution should be taken when interpreting the results of the Hosmer-Lemeshow test. Although they do suggest that an evaluation of model calibration should include the test, a significant result does not always mean that the model is not useful. They suggest that additional information could be presented, such as the overall number of patients and the observed and predicted probabilities within each decile. A further issue with the Hosmer-Lemeshow test is that it can be over-sensitive to large samples because it has high power to detect small differences in risk which may not be clinically relevant (McGeechan et al., 2008). McGeechan et al., 2008, propose that the result of the Hosmer-Lemeshow test can be accompanied by a bar chart showing the average observed and expected risks for deciles of risk, allowing inspection of whether the differences between the two measures are large enough to be clinically significant.

2.4.3 Global measures of fit

Global measures of fit combining both calibration and discrimination are also available. The AIC and BIC tell us the likelihood that the fitted model would produce the data that is observed in practise. These measures both impose penalties for increasing numbers of covariates in order to discourage over fitting, but they do not give any indication of how clinically useful the addition of a biomarker is and so should not be used on their own (McGeechan et al., 2008). The Brier score measures the accuracy between the predicted and observed events (Steyerberg, 2009). It is a quadratic scoring rule and is computed by squaring the differences between the observed and expected outcome probabilities. The range of the Brier score is between 0 (perfectly fitted model) and 0.25 (uninformative model) and it is possible to adjust the score in order to apply it to time dependent data that include censored results (Steyerberg et al., 2010). Finally, we can also consider using a pseudo R^2 measure of goodness of fit (Cameron and Windmeijer, 1997), a measure of model performance in terms of the proportion of variability in the data that is explained by the model.

2.5 Validating a prediction model

In order for a model to be useful for clinical practice, we need to know that it can be generalised to other groups of patients and that it will perform well for them. This process will require some degree of clinical judgement, but can also be assessed by carrying out a validation study on some form of new data. Altman et al., 2009, explore three methods of using new data to validate a model: internal validation, temporal validation and external validation. In internal validation the original dataset is split randomly into two groups (often in a 2:1 ratio). The first group is referred to as the "training set" and is used to build the prediction model. The second group, the "test set", is then used to assess the accuracy of the predictions from that model. Although internal validation can be helpful, it can produce optimistic results and, since it is carried out using one dataset, the model produced cannot be generalised to other populations. Temporal validation can be seen as a compromise between internal and external validation. It evaluates the prediction model on new patients sampled from the same centre. Although this new data is independent of the original

set and model, the patients are likely to have certain characteristics in common. External validation uses new data from a different population and is able to properly assess the generalizability of the model. Steyerberg et al., 2013, state that a newly developed model must be validated using external validation before being implemented in clinical practice in order to ensure that it is reliable.

3 Aims and objectives

3.1 Aims

The overall aim of this thesis was to explore the incorporation of multiple biomarkers into existing cardiovascular risk scores in people with type 2 diabetes, in order to develop models which better predict the risk of major cardiovascular events in this high risk group of individuals.

Specific objectives were to:

1. Determine whether a panel of pre-selected biomarkers, either individually or in combination, add value to a cardiovascular risk score currently used in people with type 2 diabetes. In order to achieve this, record linkage to identify incident cardiovascular events in an established prospective cohort study, the ET2DS, was undertaken, followed by detailed data analysis.
2. Determine associations between CVD and a large number of cardiometabolic metabolites measured using omics technology, in order to identify potential new biomarkers which may improve risk prediction models in people with type 2 diabetes in the future. In order to achieve this, data from multiple cohort studies in the University College London-London School of Hygiene and Tropical Medicine-Edinburgh-Bristol (UCLEB) consortium were harmonised and analysed statistically.

3.2 Thesis outline

Background information on the clinical topics of type 2 diabetes and CVD, together with information on statistical modelling approaches suitable for risk prediction research has already been provided (Chapters 1 and 2). The next chapter (Chapter 4) is a systematic review of cardiovascular risk scores that can be used in people with type 2 diabetes. The results from this review informed my choice of risk score as the basic predictive model for subsequent analyses.

The methods chapter (Chapter 5) describes the design and data collection used in the individual epidemiological studies included in this thesis (the ET2DS and the

suitable cohorts from the UCLEB consortium), together with the statistical methods for the two key aims. Chapter 6 provides relevant descriptive results from the ET2DS and Chapters 7 and 8 present results of statistical modelling to meet the two key aims respectively.

The thesis concludes with a discussion (Chapter 9) summarising the key findings of the research, discussing the strengths and limitations of the studies and the analyses carried out and comparing the findings of this thesis with previous studies. Finally, recommendations for future research on this topic are outlined.

4 Systematic Review: Cardiovascular risk scores for people with type 2 diabetes

4.1 Background

Although there are numerous risk scores available for clinical use, it is not clear which score should be used to guide the treatment of patients with type 2 diabetes. A systematic review by van Dieren et al., 2012, provided an overview of all CVD models which could be applied to people with type 2 diabetes. It found a total of 45 models, 12 of which were specifically developed in diabetic cohorts and 33 of which were developed in a general population but included diabetes as a risk factor in the score. However, this systematic search was carried out in 2011 and since then new models have been developed both in diabetes and in general populations which can be used for patients with type 2 diabetes. The van Dieren et al., 2012, paper used a search strategy which was deemed to capture a suitably wide range of papers and the criteria used to select and reject papers was appropriate for the purpose of this review.

4.2 Aim

The aim of this systematic review was to update the original van Dieren et al., 2012, review in order to identify and summarise all the available cardiovascular risk scores that can be applied to people with type 2 diabetes. A final score was then chosen from this list, using a set of pre-specified criteria, for use as a basic model to which non-traditional biomarkers were added to assess their added predictive ability.

4.3 Methods

4.3.1 Inclusion and Exclusion Criteria

The following inclusion criteria were based on the original van Dieren et al., 2012, search strategy. A study was included when:

(1) The prediction model was either developed in people with type 2 diabetes or included diabetes as a predictor.

(2) The outcome of the prediction model was a 'hard' cardiovascular end-point. Accepted outcomes were stroke, heart failure, MI or a general CVD category which included some or all of these events, such as CHD.
(3) It presented a prediction model which was not developed exclusively in patients with previous CVD or other vascular condition such as hypertension. In other words, the study had not set out to recruit *only* subjects with previous CVD.
(4) It presented a new mathematical model for the risk score, rather than focusing on the added predictive value of new risk factors to an existing prediction model.

The two main exclusion criteria, which were built into the search strategy, were non-human studies and non-English studies.

4.3.2 Search strategy

The search strategy followed by van Dieren et al., 2012, was described in their paper as follows:

((Validat$ OR Predict$.ti. OR Rule$) OR (Predict$ AND (Outcome$ OR Risk$ OR Model$)) OR (Decision$ AND (Model$ OR Clinical$ OR Logistic Models/)) OR (Prognostic AND (History OR Variable$ OR Criteria OR Scor$OR Characteristic$ OR Finding$ OR Factor$ OR Model$)) OR ("risk score"[All fields] OR "prediction model"[All fields] OR "prediction rule"[All fields] OR "risk assessment" [All fields] OR "algorithm"[All fields])) AND (cardiovascular OR coronary OR cerebrovascular OR heart OR stroke) AND (diabetes OR "diabetes mellitus" OR "type 2 diabetes") NOT (Animals[MeSH] NOT Humans[MeSH].

The first challenges in replicating the search were that some details had not been specified (for example, which fields certain terms had been searched in) and also some of the search syntax was not consistent with the current Ovid Medline code. A test search was carried out, following the original search strategy as strictly as possible. The search was restricted to results between 2 April 2011 and 1 May 2015, and it was assumed that the field used was "All fields" unless otherwise specified. The start date of the search was chosen to directly follow on from the van Dieren

search dates, which were from beginning of database to 1 April 2011. This initial test search returned a considerably larger number of potentially eligible papers (8091) than expected for the time frame, given the result of the earlier search by van Dieren (6803 papers identified from beginning of database to 1 April 2011). Therefore, it was concluded that the search strategy needed to be modified to make it more specific.

The first step taken was to change all the fields to “mp” (the multipurpose field which searches the Title, Original Title, Abstract, Subject Heading, Name of Substance, and Registry Word fields). A further restriction was placed on the “cardiovascular” search terms, requiring them to appear within 5 words of the “prediction” search terms. Finally, the arrangement of the “prediction” search terms was modified in order that the search focused more strictly on papers which were relevant to prediction modelling. Creating a search strategy for articles about prediction is a particularly challenging task as there is a range of interchangeable words and phrases used across the literature.

After the search strategy had been modified it was tested against the original search by van Dieren. This was done by changing the dates to match the original search (beginning of database to 1 April 2011) and searching for key papers identified by van Dieren et al., 2012. The results showed that all of these papers in the original review would still have been captured using the modified search strategy. In addition this search resulted in approximately the same number of papers as identified by van Dieren et al., 2012. This led to the conclusion that the final search strategy was a reasonable replication of the original.

The final search terms used in the MEDLINE search were as follows:

(("risk score" or "prediction model" or "prediction rule" or "risk assessment" or ((Predict$ or prognos$) and (Outcome$ or Risk$ or Model$ or Rule$ or "algorithm")) or (Decision$ and (Model$ or Clinical$ or Logistic Models)) or (Prognostic and (History or Variable$ or Criteria or Scor$ or Characteristic$ or Finding$ or Factor$ or Model$))) adj5 (cardiovascular or coronary or cerebrovascular or heart or stroke)).mp. AND (diabetes or "diabetes mellitus" or

"type 2 diabetes").mp. AND (201104* or 201105* or 201106* or 201107* or 201108* or 201109* or 20111* or 2012* or 2013* or 2014* or 2015*).ed. NOT Animals/ not Humans/ LIMIT to (dutch or english)

The systematic review update search was carried out on 13/05/2015 in Ovid MEDLINE.

4.3.3 Selection of studies

I screened all titles and abstracts, and then the full text of the articles. (Although this process would ideally have been carried out by two reviewers working independently, as per Cochrane guidelines (Higgins and Green, 2011), this was not practical for my PhD project).

4.3.4 Data extraction and management

Data were extracted from the final articles and put into results tables (Table *4-1* and Table *4-2*), containing the following information: title of paper and reference, population under study, recruitment method, number of events and total size of sample, type of statistical model used, predicted years modelled, number and full list of predictors included in model, model evaluation (discrimination and calibration), model validation and style of model presentation. These tables were split depending on whether the score was developed in a general population or specifically in patients with type 2 diabetes.

4.4 Results

4.4.1 Study selection

The updated systematic review search returned 3069 results, which were exported to Endnote. Figure 4-1 describes the systematic selection process. The total number of duplicated papers identified using Endnote was 370. 185 papers were therefore removed due to duplication. This left the number of articles to be screened at 2832. These 2832 papers were screened by title and abstract and 2787 records were removed as they were irrelevant to the systematic review. This resulted in 45 records remaining for a full-text assessment based on the eligibility criteria discussed above.

After examination of the full texts, 12 studies remained. Two of the final studies were developed specifically for a diabetic population (see Table *4-1* for a summary of these models) and 10 included diabetes as a factor in the model (see Table *4-2*). Nine of the final models were developed in European, American or combined populations, two models were developed in Asian populations and one model was developed in an Australian population. The majority of the scores were developed for a general cardiovascular outcome such as CVD or CHD, although three studies restricted the outcome to stroke and one to heart failure. The development study samples ranged from 607 to 3.6 million participants. 33 articles were excluded (see Table *4-3* for full details of exclusions) as they violated at least one of the four inclusion criteria.

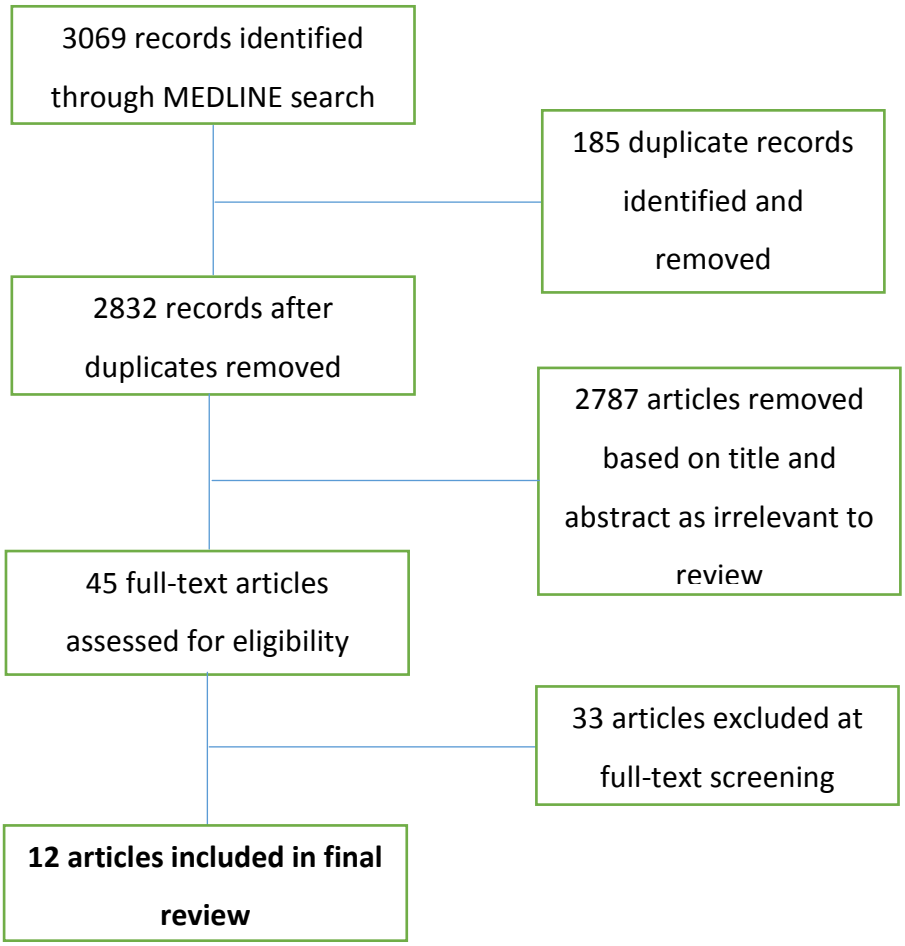


Figure 4-1 Flow chart of systematic review of studies presenting a cardiovascular risk score for use in people with type 2 diabetes

4.4.2 Results tables

Table 4-1 Cardiovascular risk models specifically developed in patients with type 2 diabetes

Title	Reference	Population	Recruitment method	N events/ N total	Type of model	Out-come	Predicted years	N predictors	Predictors	Discrimi-nation (c-statistic)	Calibrati-on (p-value H-L test)	Valida-tion	Present-ation of risk model
New studies identified in this update													
Risk score model for the assessment of coronary artery disease in asymptomatic patients with type 2 diabetes	Park et al., 2015	Type 2 diabetes, Korea	Patients at the Asan Medical Center with type 2 diabetes, who had undergone CCTA evaluation, prospectively enrolled	83/ 607	Logistic	CAD	5	8	Age, sex, duration of diabetes, hypertension, smoking status, family history of premature CAD, history of stroke, chol:HDL ratio and neuropathy	0.75	0.99	Not carried out	Original model and scoring chart
Prediction and classification of cardiovascular disease risk in older adults with diabetes	Mukamal et al., 2013	65+, diabetes, USA	Recruited from Medicare-eligibility lists in Pittsburgh, Sacramento, Hagerstown and Forsyth County	265/ 782	Cox	CVD	10	11	Age, sBP, smoking status, total cholesterol, HDL cholesterol, creatinine, oral hypoglycaemic agent or insulin use, CRP, ABI, ECG LV hypertrophy and internal CIMT	0.68	0.65	External	Original model
Studies identified in van Dieren et al., 2012, review													
Contemporary model for cardiovascular risk prediction in people with type	Kengne et al., 2011a	55+, Type 2 diabetes, 20 countries	Patients with established type 2 diabetes from 20 countries in Asia, Australia, Europe and	473/ 7168	Cox	CVD	4	10	Age at diagnosis, sex, duration of diabetes, pulse pressure, hypertension, AF,	0.70	0.76	Internal and external	Original model and scoring chart

2 diabetes (ADVANCE study)			North America were recruited from health centres						non-HDL cholesterol, HbA1c, ACR and retinopathy				
An Australian cardiovascular risk equation for type 2 diabetes: the Fremantle Diabetes Study	Davis et al., 2010	Type 2 diabetes, Australia	Longitudinal observational study of patients within a postcode-defined region of Australia. A variety of recruiting strategies were used.	185/ 1240	Logistic	CVD	5	7	Age, sex, ethnicity, prior CVD, HDL-cholesterol, HbA1c and ACR	0.80	0.74	External	Original model
Derivation and validation of a new cardiovascular risk score for people with type 2 diabetes: the New Zealand diabetes cohort study	Elley et al., 2010	Type 2 diabetes, New Zealand	Patients from the "Get Checked" program, a national primary-care annual review program using routinely collected data	6479/ 36,127	Cox	CVD and CHD	5	10	Age at diagnosis, sex, duration of diabetes, ethnicity, sBP, smoking status, chol:HDL ratio, HbA1c, ACR and medication status	CVD: 0.68 CHD: 0.71	Good	External	Original model
Risk prediction of cardiovascular disease in type 2 diabetes: a risk equation from the Swedish National Diabetes Register	Cederholm et al., 2008	Type 2 diabetes, Sweden	Observational study of patients from the Swedish National Diabetes Register	1482/ 11,646	Cox	CVD	5	9	Age at diagnosis, sex, duration of diabetes, sBP, BMI, smoking status, HbA1c and antihypertensive and lipid-reducing drugs	0.70	0.08	NR	Original model

Development and validation of a total coronary heart disease risk score in type 2 diabetes mellitus	Yang et al., 2008b	Type 2 diabetes, China	Patients from the Hong Kong Diabetes Registry	351/ 7067	Cox	CVD	5	7	Age, sex, duration of diabetes, smoking status, non–HDL cholesterol , ACR and eGFR	0.70	Good, p>0.05	Internal	Original model
Development and validation of a risk score for hospitalization for heart failure in patients with Type 2 diabetes mellitus	Yang et al., 2008a	Type 2 diabetes, China	Patients from the Hong Kong Diabetes Registry	274/ 7067	Cox	Heart failure	5	6	Age, BMI, coronary heart disease during follow-up, HbA1c, ACR and blood haemoglobin at baseline	0.85	Good, p>0.10	Internal	Original model
Development and validation of stroke risk equation for Hong Kong Chinese patients with type 2 diabetes: the Hong Kong Diabetes Registry	Yang et al., 2007	Type 2 diabetes, China	Patients from the Hong Kong Diabetes Registry	332/ 7209	Cox	Stroke	5	4	Age, history of CHD, HbA1c and ACR	0.75	Good, p>0.05	Internal	Original model
Derivation and validation of a prediction score for major coronary heart disease events in a UK type 2 diabetic population	Donnan et al., 2006	Type 2 diabetes, Scotland	All subjects with type 2 diabetes and registered with a GP in Tayside, Scotland (diabetes ascertained using the Diabetes Audit and Research in Tayside database	243/ 4569	Weibull	CHD	10	9	Age at diagnosis, sex, duration of diabetes, sBP, height, smoking status, total cholesterol, HbA1c and treated hypertension	0.71	0.54	External	Original model

(DARTS study)

Prediction of coronary heart disease in middle-aged adults with diabetes (ARIC study)	Folsom et al., 2003	Type 2 diabetes, USA	Prospective cohort, sampled from four US communities	128/ 1273	Cox	CHD	10	19	Age, ethnicity, sBP, BMI, waist-to-hip ratio, smoking status, total and HDL cholesterol, lipoprotein(a), antihypertensives, albumin, creatinine, WBC, fibrinogen, factor VIII, sport activity, residual FEV1, Keys score and pack-years smoking	0.77 (women), 0.74 (men)	NR	Not carried out	Basic and full models
UKPDS 60: risk of stroke in type 2 diabetes estimated by the UK Prospective Diabetes Study risk engine	Kothari et al., 2002	Newly diagnoised type 2 diabetes, UK	Patients referred by GPs in the catchment areas of 23 UK hospitals	188/ 4549	MLE, Newton Raphson method	Stroke	10	7	Age, sex, duration of diabetes, sBP, smoking status, AF and chol:HDL ratio	NR	NR	Not carried out	Original model, risk software
The UKPDS risk engine: a model for the risk of coronary heart disease in Type II diabetes	Stevens et al., 2001	Newly diagnosed type 2 diabetes, UK	Patients referred by GPs in the catchment areas of 23 UK hospitals	NR /4540	MLE, Newton Raphson method	CHD	10	8	Age, sex, duration of diabetes, ethnicity, smoking status, sBP, HbA1c and total chol:HDL ratio	NR	NR	Not carried out	Original model, risk software

Developing risk stratification charts for diabetic and nondiabetic subjects	Yudkin and Chaturvedi, 1999	Type 2 diabetes, USA	Respondents from a random sample of the adult population of Framingham, Massachusetts	NR /2138	CHD	10	5	Age, sBP, smoking status, chol:HDL ratio and microalbuminuria	NR	NR	Not carried out	Scoring chart

ACR: albumin creatinine ratio
AF: atrial fibrillation
CAD: coronary artery disease
CIMT: carotid intima-media thickness
ECG: electrocardiogram
eGFR: estimated glomerular filtration rate
FEV: forced expiratory volume
H-L: Hosmer-Lemeshow test
LV: left ventricular
NR: not reported
WBC: white blood cell

Table 4-2 Cardiovascular risk models developed in general populations with diabetes as a predictor

Title	Reference	Population	Recruitment method	N events/ N total	Type of model	Out-come	Predicted years	N predictors	Predictors	Discrimi-nation (c-statistic)	Calibrati-on (p-value H-L test)	Valida-tion	Present-ation of risk model
New studies identified in this update													
Heart failure risk prediction in the Multi-Ethnic Study of Atherosclerosis (MESA)	Chahal et al., 2015	USA GP	Recruited from 6 towns in the USA, aiming to capture equal number of males and females and according to specified race/ethnicity proportions	176/ 6809	Cox	HF	5	8	Age, sex, sBP, heart rate, BMI, smoking status, NT-proBNP and diabetes	0.87	0.86	Internal	Original models, scoring chart
Improving long-term prediction of first cardiovascular event: the contribution of family history of coronary heart disease and social status	Veronesi et al., 2014	Italian GP	Data from three independent population-based surveys in Northern Italy	356/ 3956	Cox	CHD	20	10	Age, sBP, smoking status, family history of CHD, total cholesterol, HDL cholesterol, hypertensive medication, education and diabetes	Men: 0.77 Women: 0.84	Good	Not carried out	Reclassification table
Development of a point-based prediction model for the incidence of total stroke: Japan public	Yatsuya et al., 2013	Japanese GP	Study cohorts established in 9 public health-center areas. Study population defined as all residents with	790/ 15,672	Cox	Stroke	NR	7	Age, sex, BMI, blood pressure, smoking status, antihypertensive medication and	0.73	Good	External	

health center study			Japanese nationality aged 40 to 69 years						diabetes				
ASCORE: an up-to-date cardiovascular risk score for hypertensive patients reflecting contemporary clinical practice developed using the (ASCOT-BPLA) trial data	Prieto-Merino et al., 2013	European GP	Patients recruited from family general practices in the UK, Ireland and the Nordic countries	1240/ 15,955	Cox	CVD	5	10	Age, sex, sBP, smoking status , total cholesterol, HDL cholesterol, fasting glucose, creatinine, previous blood pressure treatment and diabetes	0.66	0.22	Internal	Original models, scoring chart
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	Hippisley-Cox et al., 2013	England and Wales GP	Open cohort of patients aged 25-84 years at the study entry date, drawn from patients registered with eligible general practices	77,578/ 3.5 million	Cox	Stroke or TIA	10	18	Age, sex, ethnicity, hypertension, sBP, BMI, smoking status, AF, congestive cardiac failure, CHD, valvular heart disease, family history of coronary disease, chol:HDL ratio, deprivation score, rheumatoid arthritis, chronic kidney disease and type 1 diabetes, type 2 diabetes	Men: 0.87 Women: 0.88	NR	External	Original model

Layperson-oriented vs. clinical-based models for prediction of incidence of ischemic stroke: National FINRISK Study	Qiao et al., 2012	Finnish GP	Random population representative samples	840/ 30,361	Cox	Stroke	10	10	Age, sBP, BMI, smoking status, hypertension medication, happy marriage, capability to walk 500 m, regular exercise, vegetable/fruit intake and diabetes	Men: 0.82 Women: 0.82	NR	Not carried out	Original model
A score for the prediction of cardiovascular events in the hypertensive aged (ANBP2 study)	Nelson et al., 2012	Australian GP	Prospective randomized trial conducted in Australian general practices in hypertensive patients	1431/ 6083	Cox	CVD	NR	7	Age, sex, family history of heart disease or stroke, anticoagulant, antihypertensive medication, physical activity and diabetes medication	0.65	NR	Internal	Original model
Prediction model to estimate presence of coronary artery disease: retrospective pooled analysis of existing cohorts	Genders et al., 2012	USA and European GP	Patients from 18 hospitals enrolled in single centre studies. 14 datasets consisted of consecutive patients enrolled in a prospective study for other research objectives. 4 datasets consisted of patients retrospectively identified as eligible via electronic radiology reporting systems.	1634/ 5677	Logistic	CAD	NR	8	Age, sex, hypertension, smoking status, chest pain, dyslipidaemia, coronary calcium score and diabetes	0.88	NR	External	Original models

Coronary risk assessment among intermediate risk patients using a clinical and biomarker based algorithm developed and validated in two population cohorts (CHDRA model)	Cross et al., 2012	USA GP	Participants came from the Marshfield Clinic Personalized Medicine Research Project, a population-based sample repository in Wisconsin	385/ 10,623	Cox	CHD	5	11	Age, sex, family history of MI, combined with serum levels of seven biomarkers (CTACK, Eotaxin, Fas Ligand, HGF, IL-16, MCP-3, and sFas) and diabetes	0.65	NR	External	Original model
One risk assessment tool for cardiovascular disease, type 2 diabetes, and chronic kidney disease	Alssema et al., 2012	Netherlands GP	Randomly selected inhabitants of Rotterdam, Hoorn and Groningen	1075/ 6780	Logistic	CVD	7	7	Age, BMI, waist circumference, smoking status, antihypertensive medication, family history of MI or stroke and family history of diabetes	Men: Women: 0.82	NR	Internal	Original model, scoring chart
Studies identified in van Dieren et al., 2012, review													
Constructing the prediction model for the risk of stroke in a Chinese population: report from a cohort study in Taiwan	Chien et al., 2010	Chinese GP	Study participants were residents of the Chin-Shan area in Taiwan and recruited using contact with local population authorities to identify appropriate households	240/ 3602	Cox	Stroke	10	7	Age, sex, sBP, dBP, family history of stroke, AF and diabetes	0.77	NR	Internal	Original model, scoring chart, nomogram

Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database	Hippisley-Cox et al., 2010	British GP	A prospective cohort study recruiting primary care patients from the QResearch database	121,623 / 1,267,159	Cox	CVD	Lifetime	12	Ethnicity, hypertension, sBP, BMI, smoking status, chol:HDL ratio, AF, family history of CHD in first degree relative aged under 60 years, deprivation score, rheumatoid arthritis, chronic kidney disease and type 2 diabetes	Women: 0.84 Men: 0.83	Good	Internal	Original model
Estimating modifiable coronary heart disease risk in multiple regions of the world: the INTERHEART Modifiable Risk Score (IHMRS)	McGorrian et al., 2011	GP from 52 countries	Cases of first MI admitted to coronary care or equivalent units	12,438/ 27,043	Logistic	MI	NR	6	Age, sex, hypertension, smoking status, second-hand smoke exposure, apolipoprotein B:A1 ratio and diabetes	0.71	0.0004	Internal	Original model
Development and validation of a cardiovascular risk prediction model for Japanese: the Hisayama study	Arima et al., 2009	Japanese GP	A long-term prospective cohort study recruiting residents of Hisayama Town	216/ 2742	Cox	CVD	14	7	Age, sex, sBP, smoking status, HDL cholesterol, LDL cholesterol and diabetes	0.81	0.60	Internal	Original model, scoring chart
Risk charts illustrating the 10-year risk of stroke among residents of	Ishikawa et al., 2009	Japanese GP	Residents of 12 rural districts in Japan recruited by local government offices, who issued invitations	255/ 12,276	Cox	Stroke	10	5	Sex, age, sBP, smoking status and diabetes	NR	NR	Not carried out	Scoring chart

Japanese rural communities: the JMS Cohort Study			to all people who were eligible for the mass screening for CVD program in Japan										
Risk charts illustrating the 10-year risk of myocardial infarction among residents of Japanese rural communities: the JMS Cohort Study	Matsumoto et al., 2009	Japanese GP	Residents of 12 rural districts in Japan recruited by local government offices, who issued invitations to all people who were eligible for the mass screening for CVD program in Japan	92/ 12,323	Cox	MI	10	6	Age, sex, sBP, smoking status, total cholesterol and diabetes	NR	NR	Not carried out	Scoring chart
Predicting the 30-year risk of cardiovascular disease: the Framingham heart study	Pencina et al., 2009	USA GP	Respondents from a random sample of the adult population of Framingham, Massachusetts	671/ 4506	Cox	CVD	30	8	Sex, age, sBP, smoking status, total cholesterol, HDL cholesterol, antihypertensive treatment and diabetes	0.80	0.89	Internal	Original model
General cardiovascular risk profile for use in primary care: The Framingham heart study	D'Agostino et al., 2008	USA GP	Respondents from a random sample of the adult population of Framingham, Massachusetts	641/ 8491	Cox	CVD	10	7	Age, sBP, smoking status, total cholesterol, HDL cholesterol and diabetes	Men: 0.76 Women: 0.79	Men: 0.14 Women:0.56	Not carried out	Original models, scoring chart
Predicting cardiovascular risk in England and Wales: prospective	Hippisley-Cox et al., 2008	British GP	A prospective cohort study recruiting primary care patients from the QResearch	140,115 / 1,53558	Cox	CVD	10	14	Age, sex, ethnicity, hypertension, sBP, BMI, smoking status, AF, family history CHD in first	Women: 0.82 Men: 0.79	Good	Internal	Original model

derivation and validation of QRISK2			database	3					degree relative under 60 years, chol:HDL ratio, deprivation score, kidney disease, rheumatoid arthritis and diabetes				
Assessing risk of myocardial infarction and stroke: new data from the Prospective Cardiovascular Munster (PROCAM) study	Assmann et al., 2007	German GP	Employees of 52 companies and local government authorities in Germany were recruited	596/ 35,100	CHD: Weibull Stroke: Cox	CHD and stroke	10	CHD: 6 Stroke: 5	CHD: sBP, smoking status, LDL cholesterol, HDL cholesterol, triglycerides and diabetes Stroke: age, sex, sBP, smoking status and diabetes	CHD: 0.82 Stroke: 0.78	NR	Internal	Original models, scoring chart
Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score	Ridker et al., 2007	USA GP	Derived from the Women's Health Study, a nationwide cohort of US women	504/ 24,558	Cox	CVD	10	9	Age, sBP, smoking status, parental history of MI, Lipoprotein A, Apolipoprotein B, Apolipoprotein A1, HbA1c, CRP and diabetes	0.81	0.38	Internal	Original models
Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish	Woodward et al., 2007	Scottish GP	Includes overlapping studies: the Scottish Heart Health Study, which recruited random samples of residents across 25 districts of Scotland,	422/ 13,297	Cox	CVD	10	8	Age, sBP, smoking status, family history, total cholesterol, HDL cholesterol, SIMD	Men: 0.73 Women: 0.77	NR	Not carried out	Original model

Heart Health Extended Cohort (SHHEC)			and the Scottish MONICA Project, which recruited residents in Edinburgh and north Glasgow.						and diabetes			
Coronary risk prediction for those with and without diabetes (Asia Pacific Cohort Studies Collaboration)	, 2006	Asian GP	Study participants came from the Asia Pacific Cohort Studies Collaboration, which is an overview of cohort studies carried out in Asia	2265/ 364,566	Cox	CHD mortality	8	6	Age, sBP, smoking status, total cholesterol and diabetes	Men: 0.82 Women: 0.88	p<0.001	Not carried out	Original model
Prediction of coronary heart disease in a population with high prevalence of diabetes and albuminuria: the Strong Heart Study	Lee et al., 2006b	American Indian GP	American Indian men and women were recruited form 13 Indian tribes/communities in Arizona, North and South Dakota and Oklahoma	724/ 4372	Cox	CHD	10	9	Age, sex, hypertension, smoking status, total cholesterol, HDL cholesterol, LDL cholesterol, albuminuria and diabetes	Men: 0.73 Women: 0.71	Men: 0.45 Women: 0.51	Internal	Original model
A coronary heart disease risk score based on patient-reported information (HEART)	Mainous et al., 2007	USA GP	Data came from the Atherosclerosis Risk In Communities study public use database, a large scale prospective cohort	1108/ 14,343	Cox	CHD	10	Men: 7 Women: 6	Men: Age, hypertension, smoking status, hypercholesterolemia, family history, physical activity and diabetes Women: age, hypertension, BMI, smoking status, hypercholesterolem	Men: 0.65 Women: 0.79	NR	Internal	Original model, scoring chart

									ia and diabetes			
Estimation of 10-year risk of fatal and nonfatal ischemic cardiovascular diseases in Chinese adults	Wu et al., 2006	Chinese GP	Participants were randomly selected in clusters e.g. villages, household or companies from 4 districts	742/ 9903	Cox	Ischaemic CVD	10	6	Age, sBP, BMI, smoking status, total cholesterol, and diabetes	Men: 0.80 Women: 0.79	Men: 0.73 Women: 0.27	External	Original models, scoring chart
Prediction of coronary events in a low incidence population. Assessing accuracy of the CUORE Cohort Study prediction equation.	Ferrario et al., 2005	Italian male GP	Participants came from 11 population-based cohorts in different regions of Italy, where random samples of residents were taken	312/ 6865	Cox	CHD	10	8	Age, sBP, smoking status, family history of CHD, total cholesterol, HDL cholesterol, hypertension treatment and diabetes	0.74	p>0.05	Not carried out	Original model
Riskard 2005. New tools for prediction of cardiovascular disease risk derived from Italian population studies	Menotti et al., 2005	Italian GP	Data came from 9 population-based studies provided by the members of the Research Group	1382/ 17,153	Weibull	CVD	5, 10, 15	9	Age, sex, BMI, blood pressure, heart rate, smoking status, HDL cholesterol, non-HDL cholesterol and diabetes	NR	NR	Not carried out	Original model, risk chart, risk software
Prediction of the risk of cardiovascular mortality using a score that includes glucose	Balkau et al., 2004	European GP	Data came from population-based cohorts in Europe	791/ 25,413	Cox	CVD death	5, 10	6	Age, sBP, smoking status, cholesterol, fasting and 2-h glucose (including cases of known diabetes), fasting	NR	NR	Not carried out	Original model

as a risk factor (DECODE study)									glucose alone (including cases of known diabetes)				
Predictive value for the Chinese population of the Framingham CHD risk assessment tool compared with the Chinese Multi-Provincial Cohort Study (CMCS)	Liu et al., 2004	Chinese GP	Participants came from 16 centers in 11 provinces of China, and a multi-stage sampling method was used (centers were non-randomly selected and then stratified random sampling was performed in each center)	816/ 30,121	Cox	CHD and mortality	10	6	Age, blood pressure, smoking status, total cholesterol, HDL cholesterol and diabetes	Men: 0.74 Women: 0.76	Men: 0.13 Women: 0.08	Not carried out	Original model
Heart to Heart: a computerized decision aid for assessment of coronary heart disease risk and the impact of risk-reduction interventions for primary prevention	Pignone et al., 2004	NR	Respondents from a random sample of the adult population of Framingham, Massachusetts	NR	NR	CHD	5, 10	8	Age, sex, sBP, smoking status, total cholesterol, HDL cholesterol, LV hypertrophy and diabetes	NR	NR	Not carried out	Risk software
Development and validation of a model to estimate stroke incidence in a population	Schau et al., 2003	NR	Model was developed based on over 100 relevant articles and information from appropriate organisations e.g. the American Heart	NR	NR	Stroke	NR	8	Age, sex, ethnicity, hypertension, smoking status, AF, IHD and diabetes	NR	NR	External	Risk software

			Association										
Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Munster (PROCAM) study	Assmann et al., 2002	German male GP	Employees of 52 companies and local government authorities in Germany were recruited	325/ 5345	Cox	CHD	10	8	Age, sBP, smoking status, family history of premature MI, HDL cholesterol, LDL cholesterol, triglycerides and diabetes	0.83	p>0.3	Internal	Original model, scoring chart
A stroke prediction score in the elderly: validation and Web-based application (CHS)	Lumley et al., 2002	US elderly GP	Participants came from random samples of Health Care Financing Administration Medicare eligibility lists	399/ 5888	Cox	Stroke	5	9	Age, sBP, ECG diagnosis of AF, ECG diagnosis of LVH, confirmed history of CVD, creatinine, time to walk 15 ft. and diabetes	Men: 0.65 Women: 0.77	NR	Internal	Original model, scoring chart, risk software
The risk functions incorporated in Riskard 2002: a software for the prediction of cardiovascular risk in the general population based on Italian data	Menotti et al., 2002	Italian GP	Data came from 9 population-based studies provided by the members of the Research Group	544/ 9771	Weibull	CHD, CVA and CVD	5	9	Age, sex, blood pressure, heart rate, BMI, smoking status, HDL cholesterol, non-HDL cholesterol and diabetes	CHD: 0.76 CVA: 0.86	NR	Not carried out	Original model, risk software

Prediction of stroke in the general population in Europe (EUROSTROKE): Is there a role for fibrinogen and electrocardiography?	Moons et al., 2002	European GP	Data came from 10 ongoing European population-based prospective follow-up studies	219/ 698	Logistic	Stroke	7	6	Age, hypertension, dBP, smoking status, stroke history and diabetes	0.69	p>0.5	Internal	Original model
A new method for CHD prediction and prevention based on regional risk scores and randomized clinical trials; PRECARD and the Copenhagen Risk Score	Thomsen et al., 2001	European GP	Data from two randomly-sampled Danish population studies were pooled	509/ 24,508	Cox	MI	5, 10, 20	9	Age, sex, sBP, BMI, smoking status, previous heart disease, familial predisposition, HDL cholesterol and diabetes	NR	NR	Not carried out	Original model, risk software
Multivariate risk estimation for coronary heart disease: the Busselton Health Study	Knuiman et al., 1998	Australian GP	Data came from the 1978 Busselton Health Survey participants	519/ 2258	Cox	Mortality or CHD	10	9	Age, blood pressure, smoking status, LVH and previous history of CHD, anti-hypertensive medication, total cholesterol, HDL cholesterol and diabetes	NR	NR	Not carried out	Original model

Prediction of coronary heart disease using risk factor categories	Wilson et al., 1998	US GP	Respondents from a random sample of the adult population of Framingham, Massachusetts	610/ 5345	Cox	CHD	10	7	Age, blood pressure, smoking status, total cholesterol, HDL cholesterol, LDL cholesterol and diabetes	Men: 0.74 Women: 0.77	NR	Not carried out	Original model, score chart
Prevention of coronary heart disease in clinical practice. Recommendations of the Second Joint Task Force of European and other Societies on Coronary Prevention	Wood et al., 1998	NR	Not relevant	NR	NR	CHD	10	7	Age, sex, sBP, smoking status, CHD, cholesterol and diabetes	NR	NR	Not carried out	Risk chart
A risk scoring system for prediction of coronary heart disease based on multivariate analysis: development and validation	Zodpey et al., 1994	Indian GP	Data came from a pair-matched case-control study at the Govt Medical College, Nagpur, India	154/ 308	Logistic	CHD	NR	5	Hypertension, total cholesterol, socioeconomic status, physical inactivity and diabetes	0.80	NR	External	Scoring chart
Cardiovascular disease risk profiles (Framingham)	Anderson et al., 1991a	USA GP	Respondents from a random sample of the adult population of Framingham, Massachusetts	NR/ 5573	Weibull	CHD, stroke, CVD, CVD mortality	Variable	6	Blood pressure, smoking status, total cholesterol, HDL cholesterol and diabetes	NR	NR	Not carried out	Original model

An updated coronary risk profile. A statement for health professionals (Framingham)	Anderson et al., 1991b	USA GP	Respondents from a random sample of the adult population of Framingham, Massachusetts	626/ 5573	Weibull	CHD	5, 10	8	Age, sex, sBP, smoking status, LVH, total cholesterol, HDL cholesterol and diabetes	NR	NR	Not carried out	Original model, scoring chart

AF: atrial fibrillation
CAD: coronary artery disease
CVA: cerebrovascular accident
ECG: electrocardiogram
GP: general population
HF: heart failure
H-L: Hosmer-Lemeshow test
LDL: low-density lipoprotein
LV: left ventricular
LVH: left ventricular hypertrophy
NR: not reported

Table 4-3 Papers excluded at the full-text stage

Title of paper	Reference	Inclusion criteria not met	Details of why inclusion criteria was not met
Risk scoring system to predict 3-year survival in patients treated for asympotomatic carotid stenosis	Alcocer et al., 2013	2, 3	**(2)** Outcome is not cardiovascular specific (3-year all-cause mortality) **(3)** The model was developed exclusively in patients with previous CVD (patients treated for asymptomatic carotid stenosis)
Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: the CHARGE-AF consortium	Alonso et al., 2013	2	**(2)** Outcome is a surrogate end-point (atrial fibrillation)
One-month to 10-year survival in the Copenhagen stroke study: interactions between stroke severity and other prognostic indicators	Andersen and Olsen, 2011	2, 3	**(2)** Outcome is mortality **(3)** The model was developed exclusively in patients with previous CVD (stroke)
Heart failure prognostic model	Axente et al., 2011	3	**(3)** The model was developed exclusively in patients with previous CVD (heart failure diagnosis)
A novel risk classification paradigm for patients with impaired glucose tolerance and high cardiovascular risk	Bethel et al., 2013	2	**(2)** Outcome is incident diabetes onset, not CVD
Survival of patients undergoing rescue percutaneous coronary intervention: development	Burjonroppa	3	**(3)** The model was developed exclusively in patients with

and validation of a predictive tool	et al., 2011		previous CVD (continuing or recurrent MI)
Validation of continuous clinical indices of cardiometabolic risk in a cohort of Australian adults	Carroll et al., 2014	1	**(1)** The model was not developed in people with diabetes, and also did not include diabetes as a predictor
New prognostic score for stable coronary disease evaluation	Coutinho Storti et al., 2011	3	**(3)** The model was developed exclusively in patients with previous CVD (patients with stable multi-vessel coronary artery disease and preserved ventricular function)
Prospective development and validation of a model to predict heart failure hospitalisation	Cubbon et al., 2014	3	**(3)** The model was developed exclusively in patients with previous CVD (patients with stable chronic heart failure)
Semi-parametric risk prediction models for recurrent cardiovascular events in the LIPID study	Cui et al., 2010	3	**(3)** The model was developed exclusively in patients with previous CVD (history of MI or hospitalization for unstable angina 3-36 months previously)
CHADS2 score predicts functional outcome of stroke in patients with a history of coronary artery disease	Hoshino et al., 2013	4	**(4)** The paper does not present a new model - it aims to evaluate the CHADS2 scoring system
Risk score for predicting recurrence in patients with ischemic stroke: the Fukuoka stroke risk score for Japanese.	Kamouchi et al., 2012	3	**(3)** The model was developed exclusively in patients with previous CVD (stroke)
The ADVANCE cardiovascular risk model and current strategies for cardiovascular disease risk evaluation in people with diabetes	Kengne, 2013	4	**(4)** The paper does not present a new model - it aims to evaluate the ADVANCE risk model
Prediction of stroke or TIA in patients without atrial fibrillation using CHADS2 and CHA2DS2-VASc	Mitchell et al.,	4	**(4)** The paper does not present a new model - it aims to

scores	2014		evaluate the CHADS2 and CHA2DS2-VAs scores
Validation of the atherosclerotic cardiovascular disease Pooled Cohort risk equations	Muntner et al., 2014	4	**(4)** The paper does not present a new model - it aims to evaluate the Pooled Cohort risk equations
Prediction of major vascular events after stroke: the stroke prevention by aggressive reduction in cholesterol levels trial	Ovbiagele et al., 2014	3	**(3)** The model was developed exclusively in people with previous CVD (stroke or TIA)
Cardiovascular risk prediction in diabetic men and women using hemoglobin A1c vs diabetes as a high-risk equivalent	Paynter et al., 2011	4	**(4)** The paper does not present a new model - it aims to improve current risk scores by adding HbA1c to the models
Risk for cardiovascular events in an Italian population of patients with type 2 diabetes	Pellegrini et al., 2011	4	**(4)** The paper does not present a new model - it aims to compare four current risk scores (Framingham, UKPDS, Riskard and Progetto Cuore)
A clinical risk score for heart failure in patients with type 2 diabetes and macrovascular disease: an analysis of the PROactive study	Pfister et al., 2013	3	**(3)** The model was developed exclusively in people with previous CVD (history of MI or stroke, percutaneous coronary intervention or coronary artery bypass surgery, acute coronary syndrome, or objective evidence of coronary artery disease or obstructive arterial disease in the leg)
Metabolic syndrome model definitions predicting type 2 diabetes and cardiovascular disease	Povel et al., 2013	1	**(1)** Outcome is developing type 2 diabetes
Prognostic models for stable coronary artery disease based on electronic health record cohort of 102 023 patients	Rapsomaniki et al., 2014	3	**(3)** The model was developed exclusively in patients with previous CVD (stable angina, patients with history of MI, coronary artery bypass graft , or percutaneous coronary intervention prior to the start of the study period and patients with a diagnosis of ACS within the study period (unstable angina

			or acute MI))
New Zealand Diabetes Cohort Study cardiovascular risk score for people with Type 2 diabetes: validation in the PREDICT cohort	Robinson et al., 2012	4	**(4)** The paper does not present a new model - it aims to validate the New Zealand adaptation of the Framingham score
A new risk scheme to predict ischemic stroke and other thromboembolism in atrial fibrillation: the ATRIA study stroke risk score	Singer et al., 2013	3	**(3)** The model was developed exclusively in patients with previous CVD (atrial fibrillation)
Validation of the ABCD3-I score to predict stroke risk after transient ischemic attack	Song et al., 2013	4	**(4)** The paper does not present a new model - it aims to validate the ABCD(3)-I score
A new scoring system for evaluating the risk of heart failure events in Japanese patients with atrial fibrillation	Suzuki et al., 2012	3	**(3)** The model was developed exclusively in patients with previous CVD (atrial fibrillation)
Does aortic stiffness improve the prediction of coronary heart disease in elderly? The Rotterdam Study	Verwoert et al., 2012	4	**(4)** The paper does not present a new model - it aims to improve the Framingham model
An international model to predict recurrent cardiovascular disease	Wilson et al., 2012	3	**(3)** The model was developed exclusively in people with previous CVD
A new model for 5-year risk of cardiovascular disease in Type 1 diabetes; from the Swedish National Diabetes Register	Cederholm et al., 2011	1	**(1)** The model was developed in patients with type 1 diabetes

An evidence-based score to detect prevalent peripheral artery disease (PAD)	Duval et al., 2012	2	**(2)** Outcome is not a hard CV end-point (development of PAD)
Personalized prediction of lifetime benefits with statin therapy for asymptomatic individuals: a modeling study	Ferket et al., 2012	1	**(1)** Only some models include diabetes as a predictor (CHD, 6-months CHD event mortality and other CVD mortality); Also, the study gives statins as an intervention and therefore is not an appropriate type of study for this review.
Risk equations to predict life expectancy of people with Type 2 diabetes mellitus following major complications: a study from Western Australia.	Hayes et al., 2011	2	**(2)** Outcome is not CV-specific (death)
Cardiovascular risk prediction models for people with severe mental illness: results from the prediction and management of cardiovascular risk in people with severe mental illnesses (PRIMROSE) research program	Osborn et al., 2015	1	**(1)** The development population has an unrelated disease (severe mental illness) so is not relevant to this review
Contemporary model for cardiovascular risk prediction in people with type 2 diabetes	Kengne et al., 2011b	NA	The paper was already included in the original van Dieren search

ACS: acute coronary syndrome; CHD: coronary heart disease; CV: cardiovascular; CVD: cardiovascular disease; MI: myocardial infarction; NA: not applicable; PAD: peripheral arterial disease; TIA: transient ischemic attack

4.5 Discussion

van Dieren et al., 2012, identified a total of 45 models, 12 of which were specifically developed in diabetic cohorts and 33 of which were developed in a general population but included diabetes as a risk factor in the score. They found that although there were a number of cardiovascular risk scores available for use in people with type 2 diabetes, less than a third of these had been externally validated. Only a few models demonstrated excellent discriminative ability (c-statistic greater than 0.80) and most models developed in general populations showed poor calibration.

Only two new risk scores developed specifically in patients with type 2 diabetes were found in this updated review in addition to those already identified by van Dieren et al., 2012. The first of these was the score developed by Mukamal et al., 2013, in 782 patients with type 2 diabetes living in the USA. Participants were aged 65 years or older and were recruited from Medicare-eligibility lists in four states in the USA. The score predicts the 10 year risk of a cardiovascular event and has been externally validated. The second new score developed specifically in patients with type 2 diabetes was developed by Park et al., 2015, in a sample of 607 patients living in Seoul, South Korea. The score predicts the 5 year risk of a coronary artery disease event and shows good discrimination and calibration, though validation has not yet been carried out. It should be noted that both these studies captured relatively small numbers of events (83 and 265 respectively).

Ten new risk scores developed in general populations, but including diabetes as a predictor in the model, were found in the updated review. A variety of cardiovascular endpoints were used: six scores had a general outcome such as CVD or CHD, three used stroke or TIA and one used heart failure. The length of follow-up of the studies ranged from five to 20 years and the sample sizes ranged from 3956 to 3.6 million. Eight of the models were developed in European, American or combined populations, one model was developed in a Japanese population and one model was developed in an Australian population. When reported, all models showed moderate to good calibration and discrimination. Validation was carried out for eight of the scores.

A notable feature of the results from all studies, including those found in the van Dieren et al., 2012, review, is that, in most cases, the predictors used in each score remain very similar across the models. Age, sex, blood pressure, hypertension, cholesterol, smoking status and family history of CVD are common across the vast majority of models, both for those developed in general and diabetes populations. In the models developed specifically for patients with type 2 diabetes the following additional variables are also included in most scores: duration of diabetes, HbA1c and a measure of kidney disease such as microalbuminuria or albumin-creatinine ratio. One score found in the updated review, the National FINRISK Study (Qiao et al., 2012), took a different approach to building a risk prediction model for the general population. The aim was to develop a simple model which a patient could use themselves, without the assistance of a clinician. Therefore, although some of the conventional risk factors are used in the model such as age, blood pressure, hypertension and smoking status, there are a number of uncommon predictors included: happy marriage, ability to walk 500m, regular physical activity and fruit and vegetable intake.

Another notable feature of the risk scores is the lack of non-traditional biomarkers included in most of the models: in total, seven of a potential 57 scores (12%) used non-traditional biomarkers in the set of predictors. The first risk score developed in a general population to include such a predictor was the Reynolds Risk Score (Ridker et al., 2007), which was developed from the Women's Health Study, a cohort of women living in the USA, and included apolipoprotein B, apolipoprotein A1 and CRP. The INTERHEART Modifiable Risk Score (McGorrian et al., 2011), a general population study including participants from 52 countries, included apolipoprotein B and A1 as a ratio. Prior to these studies, the ARIC score (Folsom et al., 2003), which used a sample of patients with type 2 diabetes living in the USA, included fibrinogen in the prediction model as a marker of inflammation. The CHD Risk Assessment score (Cross et al., 2012) developed in a general North American population, combined traditional cardiovascular risk factors with a larger panel of seven protein biomarkers: cutaneous T cell-attracting chemokine, Eotaxin, Interleukin 16, monocyte chemotactic protein-3, hepatocyte growth factor, Fas Ligand and sFas. Most recently, Mukamal et al., 2013, included CRP in their model for elderly

patients with type 2 diabetes living in the USA and the MESA score (Chahal et al., 2012) included NT-proBNP in their study of a general USA population.

Finally, there is a clear trend in the improvement of the quality of evaluation and validation of the risk prediction models. The earliest papers identified in the original van Dieren et al., 2012, review did not carry out or report any type of model evaluation measures, such as the c-statistic for model discrimination or the Hosmer-Lemeshow test for model calibration. Furthermore no validation, either internal or external, was investigated as part of the model development. In contrast, the majority of papers published since around 2002 include measures to evaluate at least one of discrimination and calibration, and also include some form of validation. There are exceptions throughout the articles, but overall the reporting of model evaluation and validation has greatly improved over time. The most recent models, developed in the last five years, tend to include a variety of measures which summarise the calibration and discrimination. All 12 papers from the current review update reported a measure of model discrimination (c-statistic) greater than 0.6 which indicates moderate discriminative ability. Six papers (50%) reported a value greater than 0.8, which shows excellent discriminative ability. Nine papers (75%) carried out some form of model validation: five of these (42%) used external validation which, as discussed at the end of Chapter 2, is considered essential for a reliable risk prediction model (Steyerberg et al., 2013).

In terms of papers which were excluded at the full-text stage, most of these exclusions were made based on a violation of either inclusion criterion number three (presenting a model which was not developed exclusively in patients with previous CVD) or number four (presenting a new mathematical model for the risk score). In the case of papers which developed a model exclusively in patients with previous CVD, the definition of CVD varied widely. Definitions included heart failure, MI and angina, stroke and TIA, atrial fibrillation or a general CVD category. Many papers were excluded as the aim of the study was to evaluate or improve an existing cardiovascular risk score (such as the ADVANCE, Framingham or CHADS2 scores), rather than develop a new mathematical model. Four papers were excluded because the outcome of the risk score was not a hard cardiovascular end-point, which was the

requirement of inclusion criterion number two. Alonso et al., 2013, and Duval et al., 2012, used surrogate end-points (atrial fibrillation and development of PAD respectively) as the outcome, Hayes et al., 2011, used all-cause mortality as the outcome and Bethel et al., 2013, studied incident diabetes onset as the outcome, not CVD. In addition, two papers (Alcocer et al., 2013 and Andersen and Olsen, 2011) violated both inclusion criteria numbers two and three: Alcocer et al., 2013, studied all-cause mortality as the outcome and developed the model in patients with previous CVD; Andersen and Olsen, 2011, also studied mortality as the outcome and developed the model in patients who had previously suffered from stroke. Five papers were excluded because they violated the first inclusion criterion (model development was in people with type 2 diabetes or included diabetes as a predictor). Furthermore, Cederholm et al., 2011, developed a cardiovascular risk score for patients with type 1 diabetes using the Swedish National Diabetes Register and Osborn et al., 2015, studied a development population which had an unrelated disease (severe mental illness) and therefore both papers were not relevant to the specific aim of this review. Finally, the model developed by Kengne et al., 2011a, was already included in the original van Dieren et al., 2012, search and was therefore removed from the final list.

The difficulty of updating the original van Dieren et al., 2012, search strategy has been discussed above. The insufficient information reported regarding the search terms meant that the updated search strategy could not replicate exactly what had been done previously. However, all efforts were made to replicate the search as closely as possible. Furthermore, there is an intrinsic difficulty in carrying out this type of search due to the large number of interchangeable terms which are used by authors. For example, when referring to a risk score an author may use the term "score", but they may also use any number of "model", "rule", "algorithm" or "assessment". An author may also use only one of "predict", "prognosis" or "risk". Therefore a search with this particular aim needs to be wide enough to capture all relevant articles, but not too general so that the number of papers to be screened is impractical. Considering all of these factors, the final search strategy updated the van Dieren et al., 2012, search as accurately as possible.

4.6 Choosing a risk score

As stated in Section 4.2, one of the aims of this review was ultimately to select a risk score to be used as the basis of a basic model to which non-traditional biomarkers could be added. A set of characteristics which the "perfect" risk score would have was devised. These were based on the recommendations for the development of prediction models outlined by Moons et al., 2009, in their four-part publication and also on the desire for the results of the research to be clinically useful. The risk score should have been:

- Developed using sound methodology
- Recently evaluated, showing good calibration and discrimination
- Externally validated
- Developed for a general CVD outcome, rather than restricted to one specific type of cardiovascular event such as stroke or MI
- Recommended in the clinical guidelines

Based on these characteristics, three scores from the full list of 57 were considered for use as the basic model for subsequent analyses: the ADVANCE, UKPDS and QRISK2 scores. The ADVANCE score was developed in 2011 in 11,140 participants with established type 2 diabetes from 20 countries. The model shows good discrimination and calibration and has been both internally and externally validated. The score predicts the 4-year risk of CVD and includes 10 risk factors (age at diagnosis of diabetes, sex, duration of diabetes, pulse pressure, hypertension, atrial fibrillation, non-HDL cholesterol, HbA_{1c}, albumin-creatinine ratio and retinopathy). The UKPDS score was developed in 2001 in 4549 newly diagnosed patients with type 2 diabetes in the UK. Although calibration, discrimination and external validation were not carried out during the original development, multiple studies have since shown that the UKPDS has moderate to poor discrimination and calibration in people with type 2 diabetes (van Dieren et al., 2011; Simmons et al., 2009; Guzder et al., 2005). There are two versions of the UKPDS score: the 10-year risk of CHD and the 10-year risk of stroke. To obtain overall CVD risk the two scores could be added together, though this is not recommended (UKPDS, 2011).

The QRISK2 score is updated annually and was developed in 3.6 million participants in England and Wales (48,889 participants had type 2 diabetes at baseline). The model shows good discrimination and calibration for both general and diabetic populations, and has been both internally and externally validated (Hippisley-Cox et al., 2014). The score predicts the 10-year risk of CVD and, importantly, is recommended by the NICE clinical guidelines for use in patients with or without type 2 diabetes (NICE CG181, 2016). Although ADVANCE and UKPDS met many of the characteristics outlined above, it was for this final key reason that the QRISK2 score was selected for use in subsequent analysis.

4.7 Conclusion

This systematic review presents an overview of all cardiovascular risk scores up to May 2015 which are suitable for use in patients with type 2 diabetes: either developed in a type 2 diabetes population or including diabetes as a predictor in the model. A systematic review carried out in 2011 (van Dieren et al., 2012) found 12 models which were specifically designed for people with type 2 diabetes and 33 models developed in general populations but which included diabetes as a predictor. This systematic review aimed to update the original review in order to provide a current summary of all available risk scores. Two new models were found which were developed exclusively in people with type 2 diabetes and ten new models were found which were developed in a general population but were suitable for use in people with type 2 diabetes. Almost all papers in the update include a variety of measures which summarise the calibration and discrimination of the model, and report on either internal or external validation. These results are in marked contrast to the earliest papers found in the van Dieren et al., 2012, review which did not report any measures of either calibration or discrimination, or carry out any model validation, either internal or external. These two findings mark the extreme ends of a trend of vast improvement in model evaluation and validation over time.

Inspecting the predictors in the risk scores also highlights the lack of non-traditional biomarkers which are included in model development. Traditional cardiovascular risk factors such as age, sex, blood pressure and smoking are favoured in the majority of risk scores. However, more recently a select few biomarkers such as NT-

proBNP, CRP and IL-6 have been assessed and included in risk scores. This suggests a gap in the research for further investigation of additional biomarkers, including those which are now available through omics data such as metabolomics. New studies should consider the added benefit of biomarkers which have previously not been considered for inclusion. For any findings to be reliable, it is also essential that these studies follow the trend discussed above and report appropriate measures of model performance and validation.

Finally, using a set of pre-specified characteristics for the ideal cardiovascular risk score, the QRISK2 score was selected for use throughout this thesis as the basic model to which additional biomarkers will be added. Further discussion on the implementation of QRISK2 as the basic model for subsequent analyses in this thesis can be found in the next chapter, Chapter 5.

5 Data sources and methods

This chapter details the methodology of the studies used as data sources in this thesis. The design of the Edinburgh Type 2 Diabetes Study (ET2DS) is described, as well as the data relevant to this thesis which were collected at the baseline, year 1 and year 4 phases of the study, prior to the research reported in this thesis. Data collection and derivation of variables for the eight-year follow-up phase of the study and methods of retrieving missing data are also described, both of which I carried out. The methods of the ET2DS have been described in accordance with the REporting of studies Conducted using Observational Routinely-collected Data statement for reporting observational studies (RECORD, 2015). Additionally, the statistical analysis plan is presented for the use of the ET2DS data to investigate the added predictive value of a panel of non-traditional biomarkers to a current cardiovascular risk score (Chapters 6 and 7).

The designs of the cohorts from the UCLEB consortium which were used in this thesis are also outlined, and the relevant data are described as well as the statistical analysis plan. The data from these studies (including the ET2DS) were used for the analysis of potential associations between 228 metabolites and CVD (Chapter 8).

5.1 Edinburgh Type 2 Diabetes Study

The ET2DS is a population-based prospective cohort study of 1066 men and women aged between 60 and 75 years at baseline (2006-2007) with established type 2 diabetes, living in the Lothian region in Scotland. The original aim of the study was to investigate the role of potential risk factors in complications of diabetes, such as micro and macrovascular disease, cognitive impairment and non-alcoholic fatty liver disease. Although it is possible to use a retrospective cohort design to study risk prediction, prospective cohort studies are preferable and allow for the well-defined selection of participants, clear and consistent definitions of predictors and outcomes and the addition or exclusion of particular variables during follow-up (Steyerberg, 2009). The paper published by Price et al., 2008, provides a detailed description of the full study protocol for the ET2DS and has been used, along with other published

material and my own experience in collecting follow-up data, as the basis for the following sections.

5.1.1 Study population

Potentially eligible participants for the ET2DS were selected from the Lothian Diabetes Register (LDR), an automated database established in 2001 which contains clinical information on almost all patients with known type 2 diabetes living in Lothian, central Scotland. Although the register defines diabetes according to WHO criteria, the study required re-validation of type 2 diabetes in order to ensure that the diagnosis was robust. In order to obtain further confirmation of type 2 diabetes, several different diagnoses were accepted:

- individual was taking oral anti-diabetic medication and/or insulin
- if the individual was managing their diabetes via dietary modification alone then they had an HbA_{1c} measure of > 6.5% at the baseline clinic.

Individuals who were controlling diabetes through diet and who had an HbA_{1c} measurement below 6.5% had their clinical records reviewed by a consultant diabetologist to confirm a diagnosis. Furthermore, individuals who started insulin treatment within one year of diagnosis, reported evidence of pancreatic disease at the clinic or those who were treated with insulin and aged under 35 years at diagnosis, were carefully reviewed. If it was not possible to confirm a clinical diagnosis of type 2 diabetes by review of hospital and/or GP records then such individuals were excluded. Additionally, non-English speakers, individuals with poor eyesight (corrected visual acuity worse than 6/36 for distance vision or unable to read large print text), individuals unable or unwilling to provide consent and those who were physically unable to complete all assessment elements were excluded. The requirements of English and adequate eyesight were necessary in order to allow participants to complete the cognitive tasks.

The study aimed to recruit 1000 subjects, which would allow 90% power at the two-sided 5% significance level to detect a Pearson correlation coefficient of ≥ 0.10 between a continuous outcome measure and predictor variable. This sample size was also estimated to allow for detection of any risk factor that contributed 1% or more to

the variance in the outcome for observed associations, both at baseline and during follow-up.

On the 1st August 2006, people aged between 60 and 74 years were identified on the LDR, and grouped by sex and 5-year age bands. Between 20th June 2006 and 1st June 2007, 5454 potential participants were contacted by mail through the custodians of the LDR. Of the 3286 individuals who responded, 1252 people expressed an interest in participating in the ET2DS, and 1077 of these attended the baseline research clinic. Of the 1077 individuals attending the baseline clinics, four people were subsequently excluded as they were too physically or emotionally distressed and it was not appropriate to continue with the physical or cognitive examinations. Seven people did not meet the strict criteria for type 2 diabetes and were also excluded. In practice no one had to be excluded from the study for the other pre-specified exclusion criteria. This resulted in 1066 participants who were willing and eligible to take part in the ET2DS. An overview of this process is summarised in Figure 5-1.

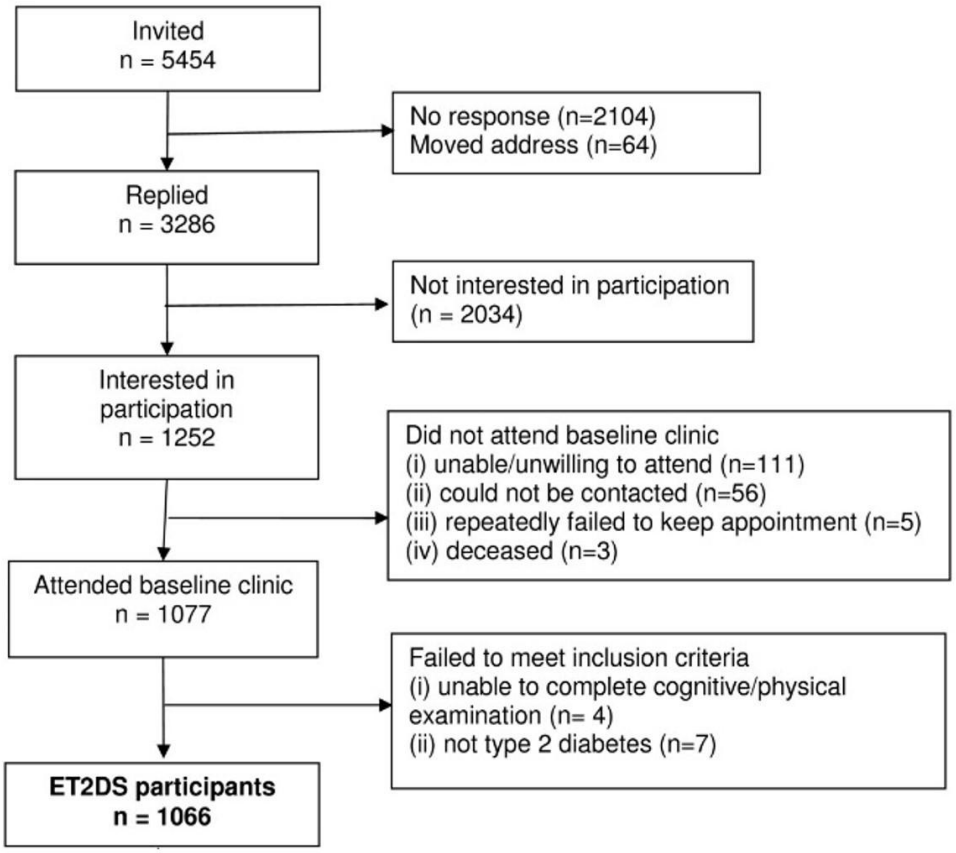


Figure 5-1 ET2DS recruitment flow diagram. Adapted from Price et al., 2008

In order to assess the impact of non-responders, demographic characteristics and clinical features of study participants and non-respondents were compared. This analysis showed that the ET2DS population is largely representative of all patients aged between 60 and 75 years with type 2 diabetes living in Lothian (Marioni et al., 2010).

5.1.2 Data collection

To date the ET2DS has had four phases of data collection: a baseline clinic (2006-2007), a liver sub-study clinic after one year, and 4 and 8 year follow-up phases. Table *5-1* provides a summary of the data collection at each phase of the study.

Baseline visit

Baseline research clinics were carried out at the Wellcome Trust Clinical Research Facility, Western General Hospital, Edinburgh, UK. In order to ensure that all individuals who had agreed to participate were assessed, taxis to and from the clinic were arranged if required, travel expenses were reimbursed and flexibility in the date and time of appointments or rescheduling was offered. Participants were required to undergo an overnight fast before attending the clinic, undergo venepuncture to provide a blood sample, provide a urine sample and undergo a 12-lead electrocardiogram (ECG). Physical examinations were performed by six specially trained research nurses and measurement technicians, who were following pre-specified standard operating procedures and data collection forms, ensuring consistency both between and within the nurses' assessment. Participants also returned completed questionnaires which included information on demographic characteristics, diabetes history and treatment, CVD, medications and smoking habits.

At baseline, data linkage to the Scottish Morbidity Records (SMR01) scheme (acute hospital discharge records) was carried out via the Information Services Division (ISD) of NHS Scotland to provide information for all participants on all medical and surgical discharges from wards in Scottish hospitals between 1981 and September 2007. These data were used to confirm self-reported history of CVD and other medical conditions such as atrial fibrillation and rheumatoid arthritis. Additionally,

selected historical data held on the LDR were retrieved, such as HbA_{1c}, blood pressure, serum cholesterol levels and estimated glomerular filtration rate (eGFR).

1 year clinic

One year after recruitment to the study, all surviving participants were invited to return for further examination at a 1 year clinic. The primary purpose of this clinic was to assess liver function, however it also provided the opportunity for further vascular assessment and venous blood sampling. A total of 940 participants attended the 1 year clinic (88.18% of the baseline study participants); 17 of the original ET2DS participants had died, 8 had indicated that they did not wish to be re-contacted when attending the baseline clinic or were considered unsuitable for further contact by the study team, 61 were unable or unwilling to attend and 40 either could not be contacted or did not attend their clinic appointment. For the purposes of subsequent analyses in this thesis, a very small number of specific biological measurements taken at year 1 (total cholesterol for one participant, HDL cholesterol for three participants and GGT for three participants) have been considered as baseline data since it is assumed that changes in these variables will not be clinically significant over the course of one year (Morling et al., 2015, Morling et al., 2014a, Morling et al., 2014b).

4 year follow-up

In 2010, appointments for a 4 year clinic visit were arranged, with the primary aim of assessing the development of CVD during the follow-up period. 974 participants were invited to the 4 year clinic (11 participants had withdrawn from the study after the baseline clinic and 81 had died). Of those invited, 830 participants attended the 4 year clinic (77.86% of the baseline study population). 15 participants were not contactable, 100 participants declined to attend and 30 participants withdrew from the study.

Physical examinations were carried out by four researchers and included fasting blood sampling, ultrasound scanning and a 12-lead ECG. Self-completed questionnaires were also returned by participants at the clinics. Data linkage via the

ISD was carried out to provide information on all hospital discharges and death certificates during the follow-up period, and access to clinical case notes was obtained if required.

8 year follow-up

In March 2015, repeat ISD record linkage was carried out and I was responsible for obtaining these data and subsequently identifying and confirming new cardiovascular events that had occurred since the 4 year follow-up phase. I submitted a successful application for Privacy Advisory Committee (PAC) approval to ISD, shown in Appendix B Information was obtained on all hospital discharges and death certificates for participants since the 4 year follow-up using probabilistic record linkage based on the participants' full name, address, date of birth and gender.

Table 5-1 Summary of data collection in the Edinburgh Type 2 Diabetes Study

Baseline	*Year 1 clinic*	*Year 4 follow-up*	*Year 8 follow-up*	*Data collection type*
✓	✓	✓		General questionnaire
✓		✓		Cardiovascular questionnaire
✓	✓	✓		Physical examination
✓	✓	✓		Fasting venous blood sample
✓		✓		ECG
✓		✓	✓	Data linkage

5.1.3 Variable measurement and definitions

This section describes the measurement procedures and the definitions used for those ET2DS variables used in this thesis; the numerous other data collected as part of the ET2DS are outlined in Price et al., 2008.

5.1.3.1 Demographics

At baseline, subjects were required to complete a questionnaire prior to attending the clinic. This questionnaire included questions on date of birth (for age), sex and

ethnicity. Deprivation was assessed using the Scottish Index of Multiple Deprivation (SIMD) 2006 which was calculated from the participants' home postcodes at baseline and for this study was grouped by quintile (The Scottish Government, 2012). The SIMD is a composite index which combines 38 indicators across seven domains, covering income; employment; health; education, skills and training; housing; geographic access; and crime.

5.1.3.2 Diabetes and other medical history

The self-report questionnaire at baseline included a question about the date of diagnosis of diabetes, which was used to calculate the duration of diabetes.

The definition of diabetes treatment type used a combination of answers from the baseline self-report questionnaire and medication lists which were brought to the clinic. Treatment type was defined as: (i) diet controlled only, (ii) oral anti-diabetic medication only (including metformin, sulphonylureas and thiazolidinediones), and (iii) insulin use (possibly with oral anti-diabetic medication additionally).

The self-report questionnaire at baseline also asked participants about their medical history and current medications. This information was used in combination with ISD record linkage and/or physical examinations to confirm previous diagnosis of atrial fibrillation, rheumatoid arthritis or hypertension. Atrial fibrillation was recorded if a subject self-reported use of digoxin, had the relevant hospital discharge code (ICD-10 code I48) or if atrial fibrillation was present on the ECG at baseline. Rheumatoid arthritis was recorded from a combination of self-report and linkage to ISD medical and surgical discharge records (ICD-10 code M06). Hypertension was defined as self-report of anti-hypertensive medication.

5.1.3.3 Physical examination

In order to obtain BMI, height (in metres) was measured without shoes using a wall-mounted vertical ruler and weight (in kilograms) was measured without outdoor clothing or shoes using electronic scales. BMI was then calculated as weight/height^2 (kg/m^2).

sBP was measured in the right arm to the nearest 2 mmHg, with the participant in the supine position. In order to obtain the ABI, the sphygmomanometer cuff was placed around the arm and inflated to 30mmHg above the estimated sBP. The pressure was reduced at a rate of 2-3mmHg per second and the sBP was recorded when the first clear sound was detected. This process was repeated in both arms and both ankles (dorsalis pedis and posterior tibial arteries) and subsequently ABI was calculated as the lowest ankle pressure divided by the highest arm (brachial) pressure.

5.1.3.4 Blood samples

At baseline, plasma from fasting venous blood samples was frozen for storage. Subsequently total cholesterol, HDL cholesterol, HbA_{1c} and eGFR were measured, all determined using a Vitros Fusion chemistry system (Ortho Clinical Diagnostics, Bucks, UK) at the Western General Hospital, Edinburgh, UK.

Chronic kidney disease (CKD) was defined as an eGFR of less than 60ml/min on two of three consecutive measurements in the 12 to 24 months prior to baseline.

Fasting venous blood samples were also assessed for measurement of potential cardiovascular biomarkers: NT-proBNP, hs-cTnT, GGT, CRP, fibrinogen, IL-6 and TNF-α. Plasma NT-proBNP and hs-cTNT were measured using the Elecsys 2010 electrochemiluminescence method (Roche Diagnostics, Burgess Hill, UK) and calibrated using the manufacturer's reagents. The manufacturer's controls were used with limits of acceptability defined by the manufacturer. GGT was analysed using a Vitros Fusion chemistry system (Ortho Clinical Diagnostics, High Wycombe, UK) at the Western General Hospital, Edinburgh, UK. Assays for plasma TNF-α, IL-6, CRP and fibrinogen were carried out in the University Department of Medicine, Glasgow Royal Infirmary. TNF-α and IL-6 antigen levels were determined using high-sensitivity ELISA kits (R&D Systems, Oxon, UK). CRP was assayed using a high-sensitivity immunonephelometric assay. Fibrinogen assays were performed using stored plasma anticoagulated with trisodium citrate and the automated Clauss assay (MDA-180 coagulometer, Organon Teknika).

Finally, fasting venous blood samples from baseline were used to obtain metabolomics data using an automated high-throughput serum NMR platform. The

process of metabolomics measurement using this platform is described in detail by Soininen et al., 2015. The platform provides information on 228 metabolic measures which can be summarised using the following 11 molecular groups:

- Very low-density lipoprotein (VLDL)
- Intermediate-density lipoprotein (IDL)
- Low-density lipoprotein (LDL)
- High-density lipoprotein (HDL)
- Lipoprotein particle sizes (for each subclass of lipid above)
- Apolipoproteins, namely apolipoprotein A-1 (ApoA-1) and apolipoprotein B (ApoB)
- Fatty acids, including omega fatty acids, saturated fatty acids and total fatty acids
- Glycolysis related metabolites produced during the process of extracting energy from glucose, including glucose, glycerol and lactate
- Amino acids, including glutamine and glycine
- Ketone bodies produced by the liver from fatty acids, including acetate and acetoacetate
- Fluid balance molecules, namely creatinine and albumin
- Inflammation, measured by glycoprotein levels, mainly a1-acid glycoprotein

5.1.3.5 Smoking

A questionnaire on smoking was self-completed at baseline. Participants were asked whether they currently smoked, and if so how many cigarettes, cigars and/or ounces of tobacco they typically smoked per week. If participants did not currently smoke, they were asked about their smoking history. Participants who reported having quit smoking in the previous six months were considered to be current smokers for the analyses in this thesis (Marioni et al., 2010). Smoking of cigars and pipes was relatively uncommon, and these quantities were converted to equivalent numbers of cigarettes based on estimated tobacco content (one cigar equivalent to four cigarettes and one ounce of tobacco equivalent to 50 cigarettes (Feinkohl et al., 2015)). Smoking was then defined categorically in line with the QRISK2 definition of smoking (Hippisley-Cox et al., 2010) as follows: (1) non-smoker, (2) ex-smoker, (3)

light current smoker (<10 cigarettes or equivalent/day), (4) moderate current smoker (10-19 cigarettes or equivalent/day), (5) heavy current smoker (20+ cigarettes or equivalent/day).

5.1.3.6 Prevalent cardiovascular events

Prevalent cardiovascular events were identified using multiple sources in order to ensure that all possible information was included. At baseline, participants self-reported medical diagnoses and/or treatment (either surgical or medication) for angina, MI, stroke or PAD, and completed a WHO chest pain questionnaire. A resting 12-lead ECG was carried out and coded using the Minnesota coding system. Events identified using these methods were further confirmed with data linkage from ISD. Any events that were not self-reported by subjects but were identified on ISD linkage data were recorded. If necessary, notes from general practitioners (GP) and hospitals were obtained in order to confirm events.

The following pre-determined criteria were used to define prevalent cardiovascular events:

MI: (1) primary or secondary diagnosis ICD-10 code (World Health Organization, 2015b) for MI (I21-I23, I252) on discharge record, and either self-report of a doctor diagnosis of MI, positive confirmation of MI on the WHO chest pain questionnaire, report of MI on GP notes or ECG evidence of MI; or (2) clinical criteria for MI met following inspection of hospital and/or GP notes.

Angina: (1) primary or secondary diagnosis ICD-10 code for angina (I20-I25) on discharge record; or (2) at least 2 of (a) self- report of a doctor diagnosis of angina or of taking angina medication, (b) ECG confirmation of angina, and (c) positive confirmation of angina on the WHO chest pain questionnaire; or (3) clinical diagnosis of angina on inspection of hospital notes.

Stroke: (1) primary diagnosis ICD-10 code for stroke (I61, I63-I66, I679, I694) on discharge record; or (2) clinical criteria for stroke met on inspection of hospital notes in subjects with either self-report of stroke or a non-primary ICD-10 hospital discharge/death code for stroke.

Transient ischemic attack (TIA): (1) primary or secondary diagnosis ICD-10 code for TIA (G45, G659) on discharge record; or (2) clinical criteria for TIA met on inspection of hospital notes in subjects with either self-report of stroke (such as "mini stroke" or "slight stroke) or with non-primary ICD-10 discharge code for stroke or TIA.

Coronary intervention: Office of Population Censuses and Surveys -4 code (Office of Population Censuses and Surveys, 1993) for coronary intervention (K40-K44, K49) on discharge record.

5.1.3.7 Incident cardiovascular events

Four year follow-up

After four years, incident or recurrent cardiovascular events were identified using repeat self-reported questionnaires completed at the four year clinic, GP questionnaires, ECG results and record linkage from ISD (Morling et al., 2015). If required, hospital notes were searched for further information and if there were doubts regarding whether criteria had been met, particular cases were discussed by a panel of researchers and a consensus decision was made.

The following pre-defined criteria were used to define incident or recurrent cardiovascular events between baseline and year four:

Fatal MI: (1) primary or secondary ICD-10 code for MI on death certificate; or (2) clinical criteria for non-fatal MI within 4 weeks of unexplained/sudden death.

MI: (1) ICD-10 code for new MI on discharge record after baseline; or (2) at least one of self-report of doctor diagnosis of MI after baseline, new confirmation of MI on the WHO chest pain questionnaire, ECG evidence for MI that was not present at baseline or GP report of MI; or (3) clinical criteria for MI met following inspection of hospital and/or GP notes for subjects with one or more individual indicators of a possible MI but not meeting the full criteria.

Angina: no indication of angina at baseline plus either (1) ICD-10 code for angina as primary diagnosis code on discharge record after baseline; or (2) at least two of self-

report of doctor diagnosis or taking medication for angina after baseline, new ECG evidence for angina or new confirmation of angina on the WHO chest pain questionnaire; or (3) clinical criteria for angina met following inspection of hospital and/or GP notes for subjects with one or more individual indicators of a possible angina but not meeting the full criteria.

Fatal stroke: (1) ICD-10 code for stroke on death certificate; or (2) clinical criteria for non-fatal stroke within 6 weeks of unexplained or sudden death.

Non-fatal stroke: (1) ICD-10 code for stroke as primary diagnosis on discharge record after baseline; or (2) self-report of stroke confirmed by inspection of clinical notes; or (3) non-primary ICD-10 codes for stroke confirmed by inspection of clinical notes.

TIA: (1) ICD-10 code for TIA as primary diagnosis on discharge record after baseline; or (2) self-report of stroke confirmed as TIA on inspection of clinical notes; or (3) non-primary ICD-10 code for stroke or TIA confirmed as TIA on inspection of clinical notes.

Coronary intervention: Office for Population Censuses and Surveys -4 code for coronary intervention on discharge record.

Fatal other IHD: subject did not meet any of the criteria for fatal MI and had an ICD-10 code for IHD (I209, I249, I258, I259) as primary cause of death.

Eight year follow-up

Eight years after baseline, a further ISD data linkage was carried out. I identified possible cardiovascular events, operations and procedures since the four year follow-up based on ICD-10 codes and developed criteria for new incident or recurrent events. If there was doubt as to whether criteria had been met, hospital notes were obtained and individual cases were discussed by researchers and clinicians to reach a consensus.

The following pre-specified criteria were used to define incident or recurrent cardiovascular events between year four and year eight follow-up phases (note that

these definitions differ from the year four cardiovascular event definitions, since data sources were limited to ISD linkage and clinical notes at the year eight follow-up):

Fatal MI: (1) ICD-10 code for MI (I21-I23) as primary cause of death; or (2) non-primary ICD-10 code for MI on death record confirmed with inspection of clinical notes.

MI: (1) ICD-10 code for MI (I21-I23) as primary diagnosis on discharge record; or (2) ICD-10 code for MI or old MI (I252) as non-primary diagnosis confirmed with inspection of clinical notes.

Angina: no indication of angina at baseline or year four follow-up plus ICD-10 code for angina (I20) as diagnosis code on discharge record.

Fatal stroke: (1) ICD-10 code for stroke (I61, I63-I64) as primary cause of death; or (2) non-primary ICD-10 code for stroke on death record confirmed with inspection of clinical notes.

Non-fatal stroke: (1) ICD-10 code for stroke (I61, I63-I64) as primary diagnosis on discharge record; or (2) ICD-10 code for stroke as non-primary diagnosis confirmed with inspection of clinical notes.

TIA: (1) ICD-10 code for TIA (G45) as primary diagnosis on discharge record; or (2) non-primary ICD-10 code for TIA confirmed on inspection of clinical notes.

Coronary intervention: Office for Population Censuses and Surveys -4 code for coronary intervention (K40-44, K49) on discharge record.

Fatal other IHD: subject did not meet any of the criteria for fatal MI and had an ICD-10 code for IHD (I209, I249, I258, I259) as primary cause of death.

After combining year four and year eight follow-up data, incident cardiovascular events were defined as the first fatal or non-fatal MI, diagnosis of angina, fatal or non-fatal stroke, TIA, coronary intervention or fatal other IHD experienced by a participant since baseline.

5.1.4 Ethical approval

All participants gave informed consent prior to data collection. Ethical approval for the study was granted by the Lothian Medical Research Ethics Committee. Furthermore, full ethical permission was granted for ISD data linkage performed at baseline and all follow-up waves.

5.1.5 Data management, cleaning and security

Data from the baseline, year 1 and year 4 questionnaires and data collection forms were coded and manually entered into a master database using Microsoft Access 2003/2010, Microsoft Corporation, Washington, USA. Laboratory data were included in the same file, obtained either from paper records or electronic files provided by the participating laboratories. At baseline all of the data from paper records were double entered into the database, and any discrepancies were resolved by referring back to the original paper documents. At year 4 a random 10% sample of the data was double entered and the error rates were found to be low for all variables so it was felt that there was no necessity to double enter the remaining records. Data from the year 4 and year 8 ISD record linkages were coded from the original electronic files and entered into a master database using SPSS v19.0 (SPSS Inc., Illinois, USA).

After data entry, descriptive analyses were carried out on measurement data and inspected for outliers and missing values. Specious results (those which were considered inaccurate based on medical knowledge or laboratory detection limits) were scrutinized by referring back to the original paper records and correcting mistakes in data entry if required.

The master databases are held and backed up on a secure university server, requiring electronic permission to access the storage drive via a unique username and password. Paper records are stored in secured filing cabinets in a locked office with only authorized access permitted.

5.1.6 Missing data

Despite previous data cleaning, the baseline variables used for this thesis still included missing data and, prior to carrying out formal analysis, I undertook a

missing data retrieval process in order to ensure a minimum amount of missing data in the final dataset. Original paper records were checked for missing data on medications, sBP, BMI and laboratory measurements such as cholesterol. Previously missing values were changed if this information was available.

Retrieving missing BMI values

One subject had a missing BMI value due to being heavier than the scale maximum and so this value was changed to a BMI of 48.2 based on the known scale maximum of 160kg and a year four home weight measurement of 151kg.

Retrieving missing CKD values

Five subjects had a missing CKD value because no routine data records prior to baseline were available. In these cases, a decision on whether the subject had CKD or not was taken based on an eGFR value less than 60ml/min from the baseline clinic visit alone. One subject had a missing CKD value due to a glitch in the hospital record linkage process, and this was rectified by obtaining the correct records prior to baseline and completing the missing data.

Retrieving missing laboratory measurements

A number of subjects were missing one or more laboratory measurement, such as serum total or HDL cholesterol or one of the pre-selected biomarkers. If missing data persisted after checking the original baseline clinic records, the value was taken as the closest routinely collected measurement within the previous or subsequent 6 months from the baseline clinic, or, if no such measurement was available, from the year 1 clinic measurements if available.

5.1.7 Data analysis

5.1.7.1 Developing a basic model

As discussed in the systematic review in Chapter 4, the QRISK2 score was chosen as the basis for an initial model in the ET2DS to which additional biomarkers would subsequently be added. This section describes the process of building a basic model based on this score.

Model updating procedure

There are various options for updating a given statistical model such as QRISK2, from applying the original prediction model to re-estimating all parameters and extending the model to include additional parameters (Steyerberg, 2009). Since the model coefficients for the QRISK2 score are not made publicly available it was not possible to carry out some of the intermediary steps which may initially be desirable in this context, such as applying the original model or updating the intercept term. Therefore, the decision was taken to recalibrate the QRISK2 model by building a new model for the ET2DS data which followed the QRISK2 variables as closely as possible.

QRISK2 variables not in the ET2DS

The QRISK2 model includes the following 14 risk factors: age, sex, ethnicity, social status, smoking status, diabetes status, family history of CVD, CKD, atrial fibrillation, hypertension, rheumatoid arthritis, ratio of total to HDL cholesterol, BMI and sBP. Family history of CVD is not available for the ET2DS as this data was not collected and all participants are confirmed to have type 2 diabetes, so these two variables were dropped from the model.

Information on ethnicity was collected as part of the ET2DS, but due to low numbers of non-white participants (n=17), the model was restricted to Caucasian participants only (n=1049). It was considered that the ET2DS did not provide enough variability in the ethnicity of the participants to justify including this categorical variable in a statistical model.

Variable definitions in QRISK2 vs ET2DS

The QRISK2 score uses the Townsend index (Townsend et al., 1988) as the measure of social status. The Townsend index is an area-based score of social deprivation incorporating four factors (unemployment, non-car ownership, non-house ownership and household crowding). However, the Townsend index is only calculated for

England and Wales and therefore is not available for the ET2DS participants who all live in the Lothian region of Scotland. The measure of social status which is available for the ET2DS is the SIMD. As previously described, the SIMD is also an area-based score of social deprivation, but it incorporates a different set of factors (current income, employment, health, education, skills and training, housing, geographic access and crime) which make the basis for calculations of Townsend and SIMD completely different and hence the final scores incomparable. In order to investigate the issue of social status in the ET2DS, a model was created with no measure of social deprivation included and then the SIMD was added in for comparison. The model performance improved with the addition of SIMD (c-statistic and pseudo R^2 increased from 0.683 to 0.706 and 7.59% to 10.3% respectively), so the SIMD was included as a measure of social status in the basic model for the ET2DS.

The definition of CKD used in QRISK2 was given as a clinical diagnosis of CKD based on clinical codes (Hippisley-Cox et al., 2010). However, the list of clinical codes is not made publicly available. Using a similar doctor-diagnosis definition based on hospital and surgical discharge records, a variable for CKD was created for the ET2DS. This variable identified 1.7% of the cohort as suffering from CKD. Although this was similar to the result found in the subgroup with type 2 diabetes in the QRISK2 development cohort (Hippisley-Cox et al., 2014), it is much lower than the anticipated rate of CKD among an elderly population with type 2 diabetes, which would be expected to be approximately a third (Retnakaran et al., 2006, Koro et al., 2009). A second variable for CKD was already available in the ET2DS, based on eGFR and discussed in section 5.1.3.4. This definition is equivalent to Stage 3-5 CKD (Levey and Coresh, 2012). This new variable identified 24.4% of the cohort as suffering from CKD, which is in line with what would be expected for this population. A comparison was made between models including the QRISK2 definition and the new ET2DS definition, and the latter was shown to improve model performance (c-statistic and pseudo R^2 increased from 0.687 to 0.699 and 7.90% to 9.11% respectively). Therefore, as this was considered to be a more accurate definition of CKD, and one that would be used by a GP in clinical practice, it was used to build the basic model for the ET2DS.

Accounting for prevalent CVD and lipid lowering medication

In addition to the variables discussed above, the basic model was extended to include two additional covariates: prevalence of CVD and lipid lowering medication. QRISK2 excluded participants who had previous CVD or were taking statins, but as a large proportion of the ET2DS cohort had prevalent CVD (*n*=367, 35%) or took lipid lowering medication at baseline (*n*=912, 86%), representing the situation in the target population of elderly people with type 2 diabetes, this exclusion method was not considered feasible while retaining adequate statistical power for analyses.

Modelling method

Although QRISK2 used Cox hazard regression modelling to build their risk score, the decision was taken to use logistic regression to build the ET2DS basic model. This was for a number of reasons:

- The additional underlying assumption of proportional hazards was not required for logistic regression. This assumption was violated for some of the key covariates.
- Although the scope of the analysis plan changed during the course of my PhD due to time constraints, complex variable selection methods for analysing multiple biomarkers (such as LASSO regression) using Cox regression had not yet been fully developed, and I wanted to keep the model choice consistent throughout my PhD project.
- Previously published cardiovascular risk scores have used differing modelling methods, including logistic regression (Park et al., 2015; Davis et al., 2010; Genders et al., 2012, Alssema et al., 2012; McGorrian et al., 2011).

In addition, I carried out a sensitivity analysis on the initial models produced, comparing the results of logistic and Cox regression, and found that the results were very similar. Therefore, binary logistic regression models were used to build the basic model and then evaluate the relationships between each biomarker and cardiovascular events. It should be noted that participants who died from non-CVD causes (*n*=121) were not excluded from statistical analyses and are regarded as

having had ‘no event’ in order to retain adequate statistical power. Such participants would be regarded as censored observations for the purposes of Cox regression analysis.

The added predictive value of including each biomarker in the model, over and above the QRISK2 variables, was assessed. The c-statistic was calculated for all models to provide a measure of model discrimination. Additionally, the net reclassification (NR) was calculated separately for participants who did experience a cardiovascular event and those who did not. The NR compares two models (here, the basic model and a new model incorporating one or more non-traditional biomarkers) and gives the increase or decrease in the proportion of subjects correctly classified by the new model, according to pre-specified cardiovascular risk categories (0-10%, 10-20% and >20%). Calibration was assessed using the Hosmer-Lemeshow test in which the null hypothesis assumes a well-calibrated model, so a p-value > 0.05 indicates good calibration. In this thesis the test is reported for 10 subgroups, though a number of group sizes for this test were run in order to check for a consistent overall result. Three global measures of model fit were calculated, the AIC, deviance and a pseudo R^2 value for logistic regression.

All subsets regression was used to compare all possible combinations of biomarkers and obtain the best five models, according to a pre-specified statistical criterion (the AIC). In addition, a model was fitted which included the QRISK2 risk factors and the full panel of biomarkers.

A p-value of <0.05 was taken to be statistically significant.

5.1.7.2 General inflammation factor

The four inflammation biomarkers selected for analysis in the ET2DS (TNF-α, IL-6, CRP and fibrinogen) are highly correlated (see Table *7-2* in Chapter 7). For this reason, recent studies have suggested that they can be combined into one general factor which describes the overall inflammatory burden and reduces the dimensionality of the data from four to one (Alman et al., 2013; Kumar et al., 2016; Hsu et al., 2009; Bedenis et al., 2014). The four inflammatory biomarkers were combined into one general inflammation factor, *g*, using an unrotated PCA. All four

biomarkers loaded quite strongly onto the first principal component (Table *5-2*), which explained 49% of the total variability, and this was used to calculate *g*.

Table 5-2 PCA output for the ET2DS general inflammation factor g

Inflammation biomarker	*Factor loading*	*Principal component*	*Variance explained (%)*
TNF-α	0.44	1	49
IL-6	0.75	2	25
CRP	0.80	3	15
Fibrinogen	0.76	4	11

n=1021

Log transformed values were used for TNF-α, IL-6 and CRP

5.2 UCLEB consortium cohorts

The UCLEB consortium was established in order to facilitate in-depth exploration of genetic associations using the Metabochip array (Shah et al., 2013b). The consortium involves 12 UK-based, well-established prospective observational studies and consists of over 30,000 participants in total. Metabolomics data were available in seven of the participating studies: the British Regional Heart Study (BRHS), the British Women's Heart and Health Study (BWHHS), the Caerphilly Prospective Study (CaPS), the ET2DS, the Medical Research Council National Survey of Health and Development (MRC NSHD), the Southall and Brent Revisited Study (SABRE) and the Whitehall-II Study (WHII). The following sections describe the process of data collation, the available variables and the design and key features of the studies which were analysed for this thesis. The design of the ET2DS has been described in detail previously in this chapter.

5.2.1 Contributing UCLEB studies

Six of the seven UCLEB studies with metabolomics data measured the panel of 228 metabolites detailed in section 5.1.3.4 of this chapter. However, CaPS measured a different set of metabolites which included additional metabolites not found in the panel of 228 and excluded some metabolites which were in this panel. Additionally, CaPS did not have data for most of the outcome variables (prevalent MI,

revascularisation, angina or stroke, or incident angina) or some key cardiovascular risk factors (ethnicity, anti-hypertensive medication and lipid lowering medication). Finally, CaPS contributed only 34 participants to the collated data, so any adjustment for cohort would be likely to produce a poor estimate. For all of these reasons, CaPS was excluded prior to analysis.

MRC NSHD was also excluded prior to analysis based on the fact that metabolomics data were collected between 2006 and 2010, seven to eleven years after the collection of data on prevalent diabetes, risk factors and CVD which was carried out in 1999. Therefore, most importantly, at the time of metabolomics measurement it was not known whether participants had the outcome of CVD. Furthermore, MRC NSHD only contributed 25 participants to the combined data.

Therefore, the final studies chosen to be analysed in this thesis were BRHS, BWHHS, ET2DS, SABRE and WHII.

5.2.2 BRHS

The BRHS is a prospective study of 7735 middle-aged men recruited from general health practices in 24 towns around the UK (Shaper et al., 1981). Participants, aged 40-59, were recruited between 1978 and 1980 and continue to be followed-up over 30 years later. At baseline and at intervals thereafter, a wide range of risk factors were measured such as HDL and total cholesterol, blood pressure and BMI. Demographic data such as age, social status and ethnicity were also collected, as well as information on smoking habits and medication. Between 1998 and 2000 the blood samples subsequently used for NMR metabolomics measurements were collected (Shah et al., 2013b) and risk factors were re-measured, including key components of QRISK2, as well as the presence or absence of diabetes and prevalent CVD. Follow-up for cardiovascular events includes data on incident fatal and non-fatal MI (Wannamethee et al., 2016). A non-fatal MI was diagnosed according to WHO criteria and evidence of non-fatal MI was obtained through informal reports from GPs and inspection of hospital notes and clinical correspondence.

5.2.3 BWHHS

The BWHHS is a prospective cohort study of 4286 women aged between 60-79 years at baseline, who were recruited from general practices in 23 towns in England, Scotland and Wales between 1999 and 2001 (Lawlor et al., 2003). The study began the fourth wave of follow-up in 2010. Physical examinations were carried out at baseline clinics in order to assess prevalent diabetes and also to measure a range of physical risk factors including BMI, blood pressure, cholesterol and creatinine. Blood samples taken at baseline were subsequently used for NMR metabolomics measurement (Shah et al., 2013b). Self-completed questionnaires were completed in order to provide information on participant demographic data such as age, sex, social status and ethnicity, as well as smoking habits and medications. Prevalent CVD was determined using a combination of self-reported diagnosis and GP records with confirmation from a doctor in the practice for major events such as MI and stroke. Incident cardiovascular events were determined using a combination of review of medical records, self-report questionnaires and linkage to the NHS Central Register for death records (Shah et al., 2013a).

5.2.4 SABRE

SABRE is a prospective cohort study of 4858 participants aged 40-69 years at baseline, who were recruited from GP practices and workplaces in west and north-west London between 1988 and 1991 (Tillin et al., 2012). Participants have been followed for a total of 25 years. At baseline, participants attended a clinic where physical examinations were carried out in order to assess prevalent diabetes, blood pressure, BMI, total and HDL cholesterol and creatinine. Blood samples taken at the clinics were subsequently used for NMR metabolomics measurement. Prior to examination, participants completed self-report questionnaires to provide demographic data such as age, sex, social status and ethnicity, as well as information on smoking habits and medications. Prevalent cardiovascular events were identified using either ECG abnormalities at baseline or an indication of pre-baseline CVD on primary care records at follow-up. Incident cardiovascular events were determined using a combination of ICD-10 codes on hospital discharge and death records and self-report questionnaires, either completed by the participant themselves or by their GP if the participant had moved outside the local area (Tillin et al., 2012). The ICD-

10 codes used to define cardiovascular events have been described in detail by Tillin et al., 2014.

5.2.5 WHII

WHII is a prospective cohort study of 10,308 participants aged 35-55 at baseline, who were recruited from the British Civil Service in London between 1985 and 1988 (Ferrie et al., 2002). Participants have been followed-up every five years, and this is expected to continue until 2030. At baseline clinics, physical examinations were undertaken in order to collect data on blood pressure, BMI and HDL and total cholesterol levels. Questionnaires were used to collect demographic information such as age, sex, social status and ethnicity, as well as information on smoking habits and medication. Between 1997 and 1999, during phase 5 of the study, the blood samples used subsequently for NMR metabolomics measurement were collected, and repeat measures of cardiovascular risk factors were made as well. The presence or absence of diabetes and prevalent CVD was also reassessed. Incident cardiovascular events were determined using a combination of record linkage to routinely collected data, self-report questionnaires, physical examinations and clinical notes. The paper by Hinnouho et al., 2015 describes in detail the methods for determining incident cardiovascular events: non-fatal MI was assessed using self-report questionnaires, ECGs, cardiac biomarkers and clinical notes; diagnosis of new angina since baseline was assessed based on self-reports of symptoms, confirmed by inspection of clinical notes or ECG abnormalities; and stroke was confirmed using ICD-10 codes on hospital discharge records and included TIAs.

5.2.6 Data analysis

The definitions of "incident" events across the studies and across event types was not consistent. In the ET2DS and for angina in all studies, incident events were defined as the first event during follow-up and therefore were either incident or recurrent events. In all other cases, incident events were defined as the first ever event and therefore included prevalent events. Since the raw data was not available for all studies in order to re-code the events to be consistent with the ET2DS definitions (including participants with recurrent events), there were two options for the cardiovascular outcome used for subsequent analyses: either to exclude all

participants with prevalent cardiovascular events or to combine prevalent and incident events into a general CVD outcome. However, excluding all participants with prevalent cardiovascular events would not have been possible while retaining adequate statistical power for the subsequent analysis of a large metabolomics data set. Although it was not my initial intention, and I would have preferred to use only incident or recurrent events, a general CVD outcome was considered to be the best use of the data given its limitations. Therefore, the outcome of interest for this analysis was CVD diagnosed at baseline or developing during the study follow-up period. This definition involved a participant experiencing one or more of the following: prevalent MI, prevalent revascularisation, prevalent angina, prevalent stroke, incident angina, incident fatal or non-fatal MI, incident revascularisation, incident fatal or non-fatal stroke or any incident fatal or non-fatal CHD.

Binary logistic regression models were used to evaluate the relationships between each metabolite and CVD and results were summarised using ORs and Manhattan-style plots (Miquel, 2016) showing corresponding p-values. The level of statistical significance was adjusted for multiple comparisons using three different methods: a FDR of 1%, a FDR of 5% and a Bonferroni correction.

Analyses explored four different models for the combined UCLEB data: a model adjusted only for cohort; a model adjusted for cohort, age and sex; a model adjusted additionally for the most widely accepted traditional cardiovascular risk factors (cohort, age, sex, HDL to total cholesterol ratio, sBP, smoking status, anti-hypertensive medication and lipid lowering medication); and a model adjusted for a small panel of additional key risk factors included in the majority of previously developed cardiovascular risk scores in either diabetic or general population studies (social status, BMI, eGFR and ethnicity), as discussed in the systematic review in Chapter 4. This final model did not match the basic ET2DS model discussed in this chapter in section 5.1.7.1, since the same variables were not available in all the UCLEB cohorts. A number of intermediary models could have been carried out using the variables listed above, but to avoid over-analysing the data statistical analyses were restricted to these four models.

5.2.7 Variable definitions, data collation and missing data

Based on the above four models, the following variables were desirable from each of the chosen UCLEB studies (in addition to robust data on prevalent diabetes): age, sex, HDL and total cholesterol, sBP, smoking status, anti-hypertensive medication, lipid lowering medication, social status, BMI, eGFR and ethnicity. In addition the outcome variables of prevalent MI, revascularisation, angina, stroke and incident angina, MI, revascularisation, stroke and CHD were required. These data were extracted, where available, from each study for those participants with prevalent type 2 diabetes at baseline by the ET2DS Data Manager. The UCLEB consortium definition for prevalent type 2 diabetes is any one of: self-report of type 2 diabetes at baseline; type 2 diabetes found on medical history review; participant taking glucose lowering medication at baseline; or participant fasting glucose level greater than 7mmol/L at baseline. Relevant data on all subjects from the ET2DS were extracted since every participant in this study has type 2 diabetes at baseline. The extracted data for each study was saved on the secure UCLEB server at University College London, in a folder only accessible by those working on this project. All analyses on the UCLEB cohorts were carried out on this external server.

Table *5-3* shows the data available for each UCLEB cohort and details the definitions for these variables. The majority of these definitions are set by UCLEB in order to achieve consistency across the studies. However, the definition of the social status variable was not harmonised across all studies used in this thesis. The UCLEB social status variable, based on occupation, is defined by six categories: unskilled; semi-skilled; manual skilled; non-manual skilled; managerial and lower professional; and professional. BRHS, BWHHS and SABRE use this definition, with an extra seventh category for armed forces in BRHS. Social status data from the ET2DS is coded according to the National Statistics Socio-economic classification (Office for National Statistics, 2010) which defines five groups as follows: semi-routine and routine occupations; lower supervisory and technical occupations; small employers and own account workers; intermediate occupations; managerial, administrative and professional occupations. Finally, WHII, a cohort recruited from civil servants in London, defines social status according to three categories: clerical and support; administrative; and professional and executive. In order to be able to adjust the final

model for social status, I decided to collapse the original UCLEB variable down to three groups: unskilled; skilled (including semi-, manual and non-manual skilled); and professional (including managerial, lower professional and professional). In the ET2DS, participants in semi-routine and routine occupations were classified as unskilled, participants in lower supervisory and technical operations were classified as skilled and all other participants were classified as professional. In WHII, participants in the clerical, support and administrative categories were classified as skilled and the remaining participants as professional. Finally, the four individuals in BRHS who were in the armed forces were coded as missing, since not enough information was available (i.e. ranking within the military) in order to classify these participants according to the new system.

Information on eGFR was not available for the BRHS and WHII cohorts, and a number of CVD outcomes were not available for BRHS. Since eGFR is only included as a risk factor in the final multivariable model, the decision was taken to drop the BRHS and WHII cohorts from analysis at this stage. Despite the missing CVD outcome types in the BRHS cohort, the available cardiovascular events were considered to be a subgroup of the overall cardiovascular condition. Therefore, rather than exclude this cohort from analysis, the outcome for BRHS was defined as a participant experiencing one or more of the following: prevalent MI, prevalent stroke, incident fatal or non-fatal MI or incident revascularisation.

Missing data retrieval for the ET2DS has been discussed in detail in section 5.1.6 of this chapter. Although the other UCLEB cohorts were subject to missing values, missing data retrieval was not possible for most of the other UCLEB cohorts since there was no direct access to the original data sources. However, after contacting the researchers at WHII, information on the data collection of medication variables was obtained and a large proportion of missing values for these variables (anti-hypertensive and lipid lowering medication) were changed to ‘no’.

Table 5-3 Variables available in the UCLEB cohorts

Variable	*UCLEB variable definition (units)*	***BRHS*** *n*=141	***BWHHS*** *n*=352	***ET2DS*** *n*=1058	***SABRE*** *n*=448	***WHII*** *n*=248
Age	Age (years)	✓	✓	✓	✓	✓
Sex	Gender (male/female)	✓	✓	✓	✓	✓
Total cholesterol	Total cholesterol (mmol/l)	✓	✓	✓	✓	✓
HDL cholesterol	High-density lipoprotein cholesterol (mmol/l)	✓	✓	✓	✓	✓
sBP	Systolic blood pressure (mmHg)	✓	✓	✓	✓	✓
Smoking status	Participant ever smoked (yes/no)	✓	✓	✓	✓	✓
Anti-hypertensive medication	Participant taking anti-hypertensive medication (yes/no)	✓	✓	✓	✓	✓
Lipid lowering medication	Participant taking lipid lowering medication (yes/no)	✓	✓	✓	✓	✓
BMI	Body mass index (kg/m^2)	✓	✓	✓	✓	✓
eGFR	Estimated glomerular filtration rate (ml/min)	NA	✓	✓	✓	NA
Ethnicity	1: White; 2: Asian; 3: Afro-Caribbean; 4: Chinese/Oriental; 5: Other	✓	✓	✓	✓	✓
Social status	1: Unskilled; 2: Skilled; 3: Professional	✓	✓	✓	✓	✓
Prevalent MI	MI prior to baseline	✓	✓	✓	✓	✓
Prevalent revascularisation	Revascularisation prior to baseline	NA	✓	✓	✓	✓
Prevalent angina	Diagnosis of angina prior to baseline	NA	✓	✓	✓	✓
Prevalent stroke	Stroke prior to baseline	✓	✓	✓	✓	✓
Incident angina	Diagnosis of angina during follow-up period	NA	✓	✓	✓	✓
Incident fatal or non-fatal MI or revascularisation	Fatal or non-fatal MI or revascularisation during follow-up period	✓	✓	✓	✓	✓
Incident fatal or non-fatal stroke	Fatal or non-fatal stroke during follow-up period	NA	✓	✓	✓	✓
Incident fatal or non-fatal CHD	Fatal or non-fatal coronary heart disease event during follow-up period	NA	✓	✓	✓	✓

n: number of participants with prevalent type 2 diabetes at baseline

BMI: body mass index; BRHS: British Regional Heart Study; BWHHS: British Women's Heart and Health Study; CHD: coronary heart disease; eGFR: estimated glomerular filtration rate; ET2DS: Edinburgh Type 2 Diabetes Study; HDL: high-density lipoprotein; MI: myocardial infarction; NA: not available; SABRE: Southall and Brent Revisited Study; sBP: systolic blood pressure; WHII: Whitehall-II Study

6 Results I: Characteristics of ET2DS and descriptive statistics for cardiovascular events and biomarkers

This chapter describes the baseline characteristics and representativeness of the ET2DS study population. Incident or recurrent cardiovascular events which occurred during follow-up are also summarised and descriptive statistics are presented for the panel of non-traditional biomarkers discussed in section 1.4 of Chapter 1, which will be included in subsequent analysis. Finally, missing data is discussed and summarised.

6.1 Baseline demographic characteristics

Baseline characteristics of the ET2DS cohort are presented in Table *6-1*. The baseline cohort comprised of 1066 individuals, with an average age of 67.9 years. 547 (51.31%) participants were male. 127 (11.91%) were in the first quintile of the SIMD (most deprived), 208 (19.51%) were in the second quintile, 188 (17.64%) were in the third quintile, 194 (18.20%) were in the fourth quintile and 349 (32.74%) were in the top quintile (least deprived). At baseline, 912 (85.55%) participants were taking lipid lowering medication and 872 (81.80%) were taking anti-hypertensive medication. Baseline prevalences of MI, angina, stroke, TIA and coronary intervention were 150 (14.07%), 298 (27.95%), 62 (5.82%), 31 (2.91%) and 110 (10.32%) respectively.

Table 6-1 Baseline characteristics of the ET2DS population

Variable	
Age (years)	67.91 ± 4.20
Sex (male)	547 (51.31)
Lipid-lowering medication	912 (85.55)
Anti-hypertensive medication	872 (81.80)
Smoking status	
Non-smoker	419 (39.31)
Ex-smoker	499 (46.81)
Current smoker – light (<10 cigarettes/day)	31 (2.91)
Current smoker – moderate (10-19 cigarettes/day)	48 (4.50)
Current smoker – heavy (20+ cigarettes/day)	69 (6.47)
Atrial fibrillation	69 (6.47)
Chronic kidney disease	260 (24.39)
Rheumatoid arthritis	39 (3.66)
Scottish Index of Multiple Deprivation	
Quintile 1	127 (11.91)
Quintile 2	208 (19.51)
Quintile 3	188 (17.64)
Quintile 4	194 (18.20)
Quintile 5	349 (32.74)
Body mass index (kg/m^2)	31.43 ± 5.69
Systolic blood pressure (mmHg)	133.29 ± 16.44
Total cholesterol (mmol/L)	4.31 ± 0.90
High-density lipoprotein cholesterol (mmol/L)	1.29 ± 0.36
Cardiovascular disease at baseline[a]	
Myocardial infarction	150 (14.07)
Angina	298 (27.95)
Stroke	62 (5.82)
Transient ischemic attack	31 (2.91)
Coronary intervention	110 (10.32)

Data are presented as means ± SD or *n* (%)

[a]Note that there is overlap among these subgroups

Maximum *n* = 1066

6.2 Representativeness

At baseline researchers in the ET2DS assessed the representativeness of the recruited ET2DS population (Marioni et al., 2010). Table *6-2* compares the key socio-demographic and clinical characteristics of the people recruited into the study against the people invited from the LDR, but who did not participate in the final study ("non-responders"). Marioni et al., 2010, found no statistically significant differences between the ET2DS population and non-responders for age, duration of diabetes, HbA_{1c}, diabetes treatment by insulin or social status (assessed using SIMD). Although they did find a statistically significant difference between the ET2DS population and non-responders for sex, sBP and total cholesterol, the actual sizes of these differences were small and the authors question the clinical significance of these differences. Furthermore, the authors investigated the similarities between non-responders and ET2DS participants in these characteristics by sex and 5-year age bands and found similar conclusions.

6.3 Incident cardiovascular events

Outcome events were defined as the first fatal or non-fatal cardiovascular event a participant experienced during the eight year follow-up period. Therefore events could be true incident events, if the participant had not experienced this type of outcome prior to baseline, or could be a recurrent event, if the participant had experienced this type of outcome prior to baseline. During the eight year follow-up period a total of 208 outcome events were recorded (19.51% of the study population). The breakdown according to type of cardiovascular event is shown in Table *6-3*.

6.4 Descriptive statistics of biomarkers

Table *6-4* presents descriptive statistics for baseline biomarkers. The distributions of NT-proBNP, hs-cTNT, GGT, TNF-α, CRP and IL-6 were skewed (shown in Figure 6-1) and therefore a log transformation (using the natural logarithm) has been used in all subsequent analyses. ABI has a reverse J-shaped relationship with cardiovascular risk and values greater than 1.4 measure medial arterial calcinosis rather than

atherosclerosis, so in line with previous studies participants with an ABI > 1.4 (*n*=17) were omitted from subsequent analysis (McDermott et al., 2005).

Table 6-2 Demographic and clinical characteristics of the ET2DS population and non-responders. Adapted from Marioni et al., 2010.

	ET2DS population	Non-responders
n	1066	4386
Age (years)	67.9 ± 4.2	67.9 ± 4.4
Sex (male)	547 (51.3)	1839 (41.9)**
Duration of diabetes		
Up to 5 years	516 (48.4)	2315 (48.7)
≥ 5 years	550 (51.6)	2251 (51.3)
HbA_{1c}	7.4 (1.1)	7.4 (1.4)
Insulin treatment	185 (17.4)	704 (16.1)
sBP (mmHg)	133.3 ± 16.4	137.2 ± 18.2**
Total cholesterol (mmol/l)	4.3 (0.9)	4.2 (0.9)*
Scottish Index of Multiple Deprivation		
Quintile 1	127 (11.9)	736 (16.8)
Quintile 2	208 (19.5)	1134 (25.9)
Quintile 3	188 (17.6)	820 (18.7)
Quintile 4	194 (18.2)	782 (17.8)
Quintile 5	349 (32.7)	897 (20.5)

Data are presented as means ± SD or *n* (%)

* p-value < 0.01; ** p-value < 0.001 for χ^2 test for independence or *t* test for differences between groups

HbA_{1c}: glycated haemoglobin

sBP: systolic blood pressure

Table 6-3 Summary of first incident or recurrent cardiovascular events in the ET2DS

Cardiovascular Event Type	*n* (%)
Myocardial infarction (fatal/non-fatal)	61 (5.72)
Angina	38 (3.56)
Stroke (fatal/non-fatal)	53 (4.97)
Transient ischemic attack (fatal/non-fatal)	12 (1.13)
Coronary intervention	26 (2.44)
Fatal ischemic heart disease	18 (1.69)

Table 6-4 Descriptive statistics of baseline biomarkers in ET2DS

Biomarker	
Ankle brachial index	1.0 ± 0.2
N-terminal pro-brain natriuretic peptide (pg/ml)	75.0 (37.0, 169.3)
High-sensitivity cardiac troponin T (ng/l)	9.6 (6.9, 13.8)
Gamma-glutamyl transpeptidase (U/L)	18.0 (11.0, 32.0)
C-reactive protein (mg/l)	1.9 (0.9, 4.4)
Interleukin-6 (pg/ml)	2.9 (2.0, 4.5)
Tumor necrosis factor alpha (pg/ml)	1.1 (0.7, 1.6)
Fibrinogen (g/L)	3.6 ± 0.7

Data are presented as means ± SD or median (lower IQR, upper IQR)

Maximum *n* = 1066

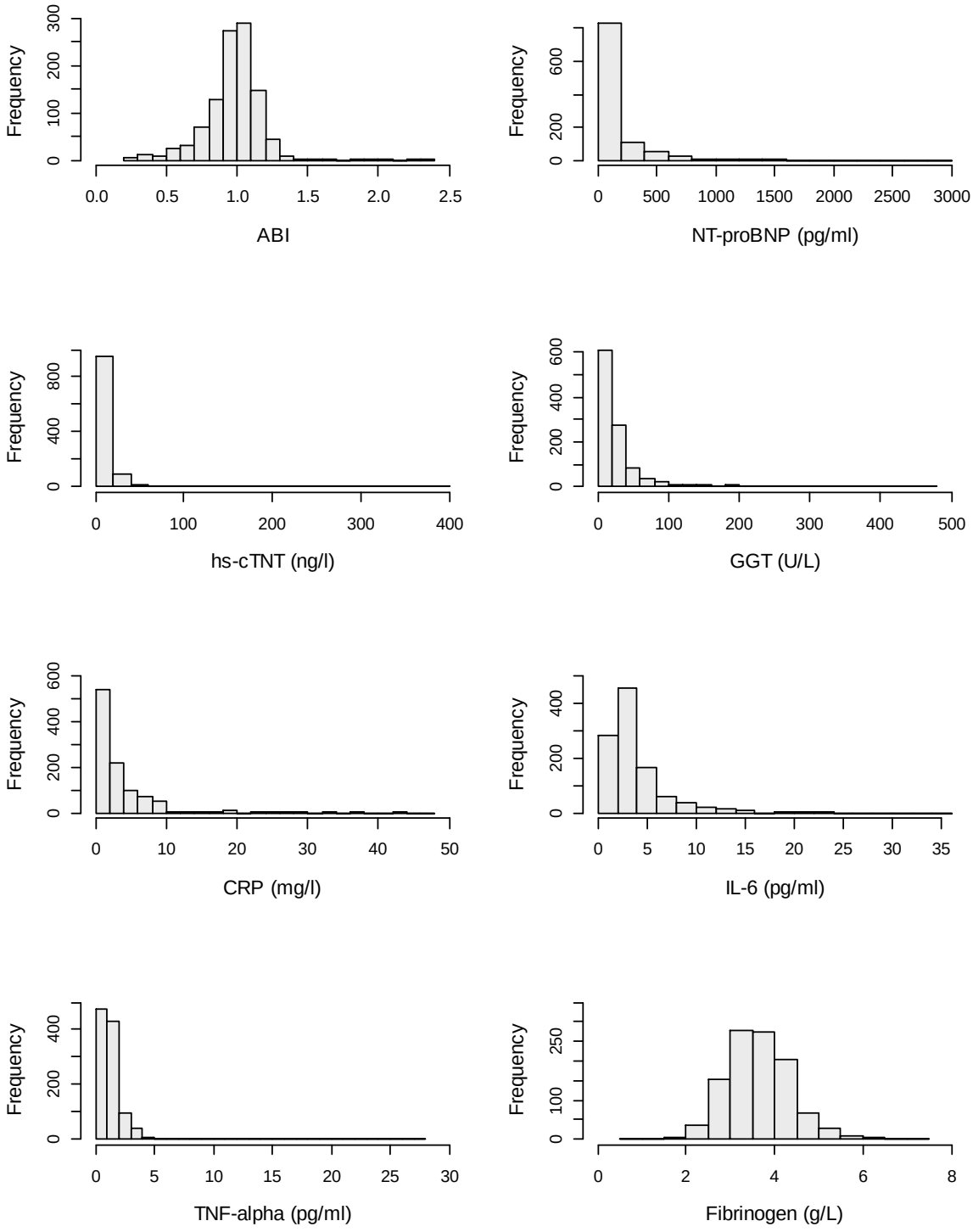


Figure 6-1 Distributions of baseline biomarkers in ET2DS

6.5 Missing data

Missing data on baseline risk factors and biomarkers are summarised in Table *6-5*, including the amount of missing data for each variable before and after final retrieval of data missing from the ET2DS database by myself for the purposes of this analysis, and reasons for remaining missing data where available. Data was complete for age, sex, social status, smoking status, atrial fibrillation, arthritis and prevalent events. The remaining variables had very little missing data (4.13% of cases missing due to one or more missing value after retrieving missing data).

Table 6-5 Missing data in the ET2DS at baseline

Variable	*n* missing before retrieving missing data	*n* missing after retrieving missing data	Reason for remaining missing data (if available)
Chronic kidney disease	6	0	
Anti-hypertensive medication	7	0	
Lipid lowering medication	2	0	
Body mass index	2	1	Participant was unable to stand at baseline clinic
Systolic blood pressure	2	0	
Total cholesterol	8	0	
HDL cholesterol	8	1	
Ankle brachial index	7	7	1 amputation below the knee and remaining participants found it too painful to have blood pressure taken
N-terminal pro-brain natriuretic peptide	14	14	
High-sensitivity cardiac troponin T	12	12	
Gamma-glutamyl transpeptidase	11	1	
C-reactive protein	24	24	
Interleukin-6	2	2	
Tumor necrosis factor alpha	3	3	
Fibrinogen	3	3	

It is not possible to test whether the pattern of missing data is missing at random (MAR) (Steyerberg, 2009), but in order to determine whether the remaining missing data was missing completely at random (MCAR), associations between key predictors (age and sex) and missingness were assessed. Additionally, the association between cardiovascular outcomes and missingness was investigated. No statistically significant associations were found, so the assumption of MCAR required for complete case analysis was not violated (Baraldi and Enders, 2010). Furthermore, since the loss of cases due to missing data was less than 5%, the risks associated with complete case analysis of biased estimates or reduced statistical power were considered to be negligible (Graham, 2009).

7 Results II: Improving cardiovascular risk prediction using individual and combined biomarkers in the ET2DS

This chapter presents analyses of the panel of non-traditional biomarkers added, individually and in combination, to the basic risk prediction model based on QRISK2. The basic model is summarised and the pairwise correlations among biomarkers are presented. The results of adding the non-traditional biomarkers to the basic model one at a time are then shown, followed by the results of an all subsets regression which considered the biomarkers in combination.

7.1 Basic model

The basic model used for subsequent analyses was based on QRISK2, adapted for use in the ET2DS cohort, as outlined previously. In order to allow consistent comparisons between subsequent models, only subjects with complete data for all basic model variables and biomarkers were used (n=989). Table *7-1* summarises the exponentiated coefficients (the ORs) for each variable in the basic model. A set of model summary measures is also presented. The c-statistic for the basic model was 0.722 (95% CI: 0.681, 0.763) and the model showed good calibration (Hosmer-Lemeshow test non-significant). The pseudo R^2 measure gives the variability explained by the basic model as 17.08%.

7.2 Associations between biomarkers at baseline

At baseline, moderate to strong relationships were observed among most of the biomarkers, as shown in Table *7-2*. In particular, the group of inflammatory biomarkers (TNF-α, IL-6, CRP and fibrinogen) were positively correlated with each other, and all of these associations were found to be strongly statistically significant (p<0.001). By design, the general inflammation factor *g* was strongly correlated with all inflammatory biomarkers. Moderate correlations were found between NT-proBNP and both ABI and hs-cTnT (r=-0.21 and 0.38 respectively; both p<0.001). ABI and hs-cTnT were weakly negatively associated (r=-0.10, p<0.01). GGT correlated weakly with three of the inflammatory biomarkers (TNF-α, IL-6 and CRP; r = 0.08, 0.16 and 0.24 respectively).

7.3 Relationships between biomarkers and cardiovascular risk

In order to explore the shape of the relationships between the biomarkers and cardiovascular risk, each biomarker was categorised into five evenly sized groups. Figure 7-1 shows the log odds of an incident or recurrent cardiovascular event in each group, separately for each biomarker. A decreasing trend between ABI and cardiovascular risk was observed, while increasing trends were observed between each of the other biomarkers and cardiovascular risk. These increasing trends appear to be strongest for NT-proBNP and hs-cTnT, while the remaining biomarkers show relatively weaker relationships with cardiovascular risk. All the associations were considered to be roughly linear in shape.

Table 7-1 Basic model coefficients and summary measures

Variable	ORs (95% CI)
Age (per year)	1.06 (1.02, 1.11)
Sex [Female]	0.80 (0.55, 1.15)
Smoking [Ex-smoker][a]	1.07 (0.73, 1.57)
Smoking [Current – light] [a]	1.15 (0.36, 3.10)
Smoking [Current – moderate] [a]	1.18 (0.44, 2.83)
Smoking [Current – heavy] [a]	1.04 (0.49, 2.08)
Atrial fibrillation [Yes]	1.87 (1.04, 3.32)
Chronic kidney disease [Yes]	1.85 (1.26, 2.69)
Rheumatoid arthritis [Yes]	1.24 (0.51, 2.74)
Hypertension [Yes]	1.00 (0.62, 1.65)
Body mass index (per kg/m^2)	1.01 (0.98, 1.05)
Systolic blood pressure (per mmHg)	1.01 (1.00, 1.02)
Total:high-density lipoprotein cholesterol	1.36 (1.16, 1.59)
Scottish Index of Multiple Deprivation [Quintile 1 reference category]	
Quintile 2	0.46 (0.26, 0.82)
Quintile 3	0.80 (0.45, 1.41)
Quintile 4	0.62 (0.35, 1.12)
Quintile 5	0.51 (0.30, 0.89)
Prevalent cardiovascular disease [Yes]	2.07 (1.45, 2.96)
Lipid lowering medication [Yes]	1.44 (0.84, 2.59)
Model summary measures:	
Deviance	859.38
Akaike's information criterion	899.38
c-statistic (95% CI)	0.722 (0.681, 0.763)
Hosmer-Lemeshow p-value	0.97
R^2 (%)	17.08

n = 989

[a] compared to reference category of non-smokers

Table 7-2 Correlation coefficients between biomarkers at baseline

	ABI	NT-proBNP	hs-cTnT	GGT	TNF-α	IL-6	CRP	Fibrinogen	*g*
ABI	1	-0.21***	-0.10**	-0.05	-0.07*	-0.12***	-0.13***	-0.18***	-0.18***
NT-proBNP		1	0.38***	-0.03	0.16***	0.18***	0.11***	0.21***	0.23***
hs-cTnT			1	0.03	0.19***	0.17***	-0.03	0.05	0.12***
GGT				1	0.08*	0.16***	0.24***	-0.06	0.16***
TNF-α					1	0.31***	0.12***	0.12***	0.43***
IL-6						1	0.42***	0.34***	0.75***
CRP							1	0.54***	0.80***
Fibrinogen								1	0.76***
g									1

Max *n* = 1032; missing data ranges from 8 to 39 data points

* Pearson correlation test p-value < 0.05

** Pearson correlation test p-value < 0.01

*** Pearson correlation test p-value <0.001

ABI: ankle brachial index; CRP: C-reactive protein; *g*: inflammation factor; GGT: gamma-glutamyl transpeptidase; hs-cTnT: high-sensitivity cardiac troponin T; IL-6: interleukin-6; NT-proBNP: N-terminal pro-brain natriuretic peptide; TNF-α: tumor necrosis factor alpha

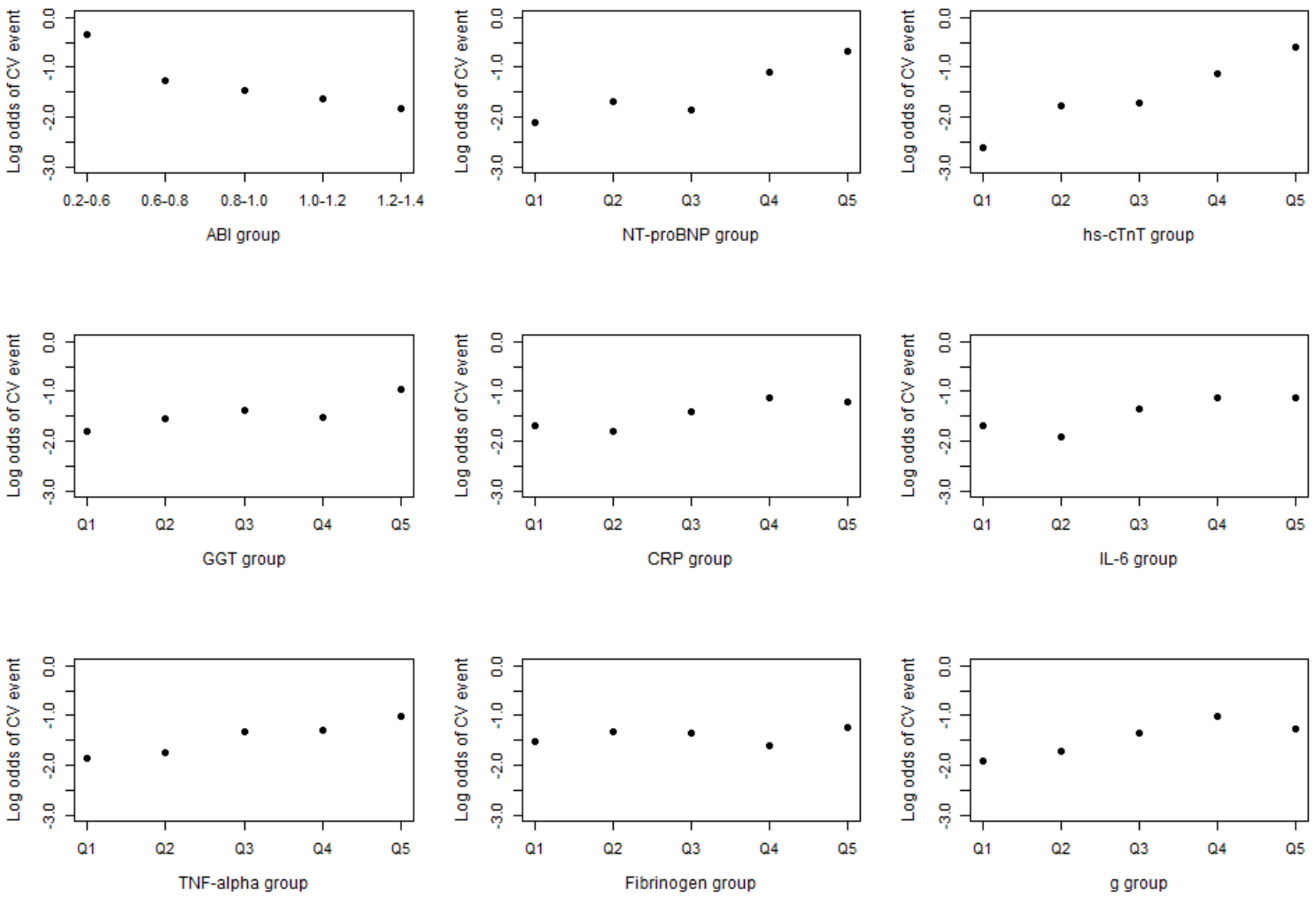


Figure 7-1 Cardiovascular (CV) risk against categorised biomarkers in the ET2DS (Q: quintile)

7.4 Adding individual biomarkers to the basic model

Five logistic regression models were fitted, adding each individual biomarker to the basic model, and the results are summarised in Table *7-3*. Increased levels of individual circulating biomarkers and the inflammatory factor were associated with an increased incidence of cardiovascular events over-and-above the basic model. However, only the associations for NT-proBNP and hs-cTnT were statistically significant. The strongest association was observed for hs-cTnT (OR for a 1 SD increase in hs-cTnT 1.35; 95% CI: 1.13, 1.61). A higher ABI was associated with a lower incidence of events, although the confidence interval just included 1 (OR for a 1 SD increase in ABI 0.86, 95% CI: 0.73, 1.00).

The basic model had a c-statistic of 0.722 (0.681, 0.762) and was well-calibrated. The addition of each individual biomarker increased the c-statistic, with the greatest increase seen for hs-cTnT (c-statistic increased by 0.01 to 0.732, 95% CI: 0.690, 0.774). The addition of individual biomarkers also improved the risk classification for participants who did not experience a cardiovascular event, although this generally resulted in poorer risk classification for participants who did experience a cardiovascular event. The addition of hs-cTnT resulted in poorer risk classification by 1.6% for participants who experienced a cardiovascular event, but improved risk classification by 2.2% for those who did not. All the models were shown to be well-calibrated (Hosmer-Lemeshow $p > 0.05$).

The AIC decreased for each model which included an additional biomarker compared to the basic model, which indicates that these models have an improved fit despite the additional covariates. Similarly, the deviance decreased after the addition of each individual biomarker confirming this improved statistical fit. Finally, the variation in the data explained by the model increased marginally for each additional biomarker, with the biggest increase in the pseudo R^2 found for hs-cTnT.

Table 7-3 Associations between individual biomarkers and cardiovascular events and corresponding measures of model performance

Model	Predictors in the model, additional to conventional risk factors[a]	OR for a one SD increase in biomarker (95% CI)	c-statistic (95% CI)	NR – Event[b] (%)	NR – No event[b] (%)	Hosmer-Lemeshow p value	Deviance	AIC	R^2 (%)
Basic model	-	-	0.722 (0.681, 0.763)	-	-	0.97	859.38	899.38	17.08
+ ABI	ABI	0.86 (0.73, 1.00)	0.725 (0.684, 0.766)	-2.2	2.0	0.83	855.63	861.63	17.44
+ NT-proBNP	NT-proBNP	1.23 (1.02, 1.49)	0.726 (0.685, 0.767)	-2.2	1.5	0.81	854.56	860.56	17.55
+ hs-cTnT	hs-cTnT	1.35 (1.13, 1.61)	0.732 (0.690, 0.774)	-1.6	2.2	0.09	848.60	854.60	18.12
+ GGT	GGT	1.16 (0.98, 1.37)	0.726 (0.685, 0.766)	-2.7	1.1	0.40	856.42	862.42	17.37
+ *g*	*g*	1.07 (0.90, 1.27)	0.724 (0.683, 0.765)	0.5	1.2	0.90	858.81	864.81	17.13

n = 989

[a] Conventional risk factors based on QRISK2: age, sex, smoking, atrial fibrillation, chronic kidney disease, arthritis, hypertension, body mass index, systolic blood pressure, total:high-density lipoprotein cholesterol, social status, baseline cardiovascular disease status (myocardial infraction, angina, transient ischemic attack and stroke) and lipid lowering medication

[b] n = 186 for event, n = 803 for no event

ABI: ankle brachial index; AIC: Akaike's information criterion; *g*: inflammation factor; GGT: gamma-glutamyl transpeptidase; hs-cTnT: high-sensitivity cardiac troponin T; NT-proBNP: N-terminal pro-brain natriuretic peptide; OR: odds ratio; SD: standard deviation

7.5 Adding combinations of biomarkers to the basic model

An all subsets regression was carried out in order to identify the top five models according to a pre-specified statistical criterion (AIC), after adjusting for the QRISK2 risk factors, from all possible combinations of biomarkers. These top five models are shown in Table *7-4*. All five models selected hs-cTnT and none included the general inflammation factor *g*. The best model found using this method added ABI, hs-cTnT and GGT to the basic model based on QRISK2. This model was well-calibrated and had a c-statistic of 0.740 (CI: 0.699, 0.781), an increase of 0.018 compared with the basic model. The addition of the three biomarkers resulted in slightly poorer risk classification by 1.1% for participants who experienced a cardiovascular event, but improved risk classification by 4.4% for those who did not. The second best model was well-calibrated and showed the same increase in the c-statistic as the top model, but the net reclassification was poorer both for participants who experienced a cardiovascular event (-2.7%) and for those who did not (3.4%) compared with the top model. For comparison, the full model including all biomarkers is also shown in Table *7-4*. The c-statistic showed the same increase as the top model, suggesting an upper limit to model performance. The addition of all biomarkers resulted in poorer risk classification by 1.6% for participants who experienced a cardiovascular event, but improved risk classification by 5.2% for those who did not.

All five top models, and the full model, showed a decrease in both AIC and deviance, indicating improved statistical fit over the basic model based on QRISK2. The variation in the data explained by the model increased for all five top models, and for the full model containing all five biomarkers.

7.6 Conclusions

The addition of each individual biomarker from a panel of pre-selected non-traditional vascular biomarkers improved the predictive ability of a basic model based on QRISK2, with the greatest improvement found for hs-cTnT. When considered in combination the greatest improvement in risk prediction was found for a model including ABI, hs-cTnT and GGT in addition to the QRISK2 predictors.

However, in all the models fitted, there remains a large amount of unexplained variability according to the pseudo R^2 measurement (highest R^2 found was 18.87%, for both the second best combined model and the full model including all biomarkers). This indicates that there is still considerable room for improvement in risk prediction and motivated the novel metabolomics analyses presented in the following chapter.

Table 7-4 Top five models selected from the combined biomarkers and corresponding measures of model performance

Model	Predictors in the model, additional to conventional risk factors[a]	c-statistic (95% CI)	NR – Event[b] (%)	NR – No event[b] (%)	Hosmer-Lemeshow p value	Deviance	AIC	R^2 (%)
Basic	-	0.722 (0.681, 0.763)	-	-	0.97	859.38	899.38	17.08
Top five models chosen using all-subsets regression selection:								
1	ABI + hs-cTnT + GGT	0.740 (0.699, 0.781)	-1.1	4.4	0.15	842.46	852.46	18.71
2	ABI + hs-cTnT + GGT + NT-proBNP	0.740 (0.699, 0.780)	-2.7	3.5	0.34	840.87	852.87	18.87
3	hs-cTnT + GGT + NT-proBNP	0.738 (0.697, 0.779)	-1.6	5.1	0.47	843.25	853.25	18.64
4	ABI + hs-cTnT	0.735 (0.694, 0.776)	-3.2	5.4	0.35	845.40	853.40	18.43
5	hs-cTnT + GGT	0.738 (0.697, 0.778)	-1.1	3.9	0.21	845.49	853.49	18.42
Full model:								
	ABI + hs-cTnT + GGT + NT-proBNP + *g*	0.740 (0.699, 0.781)	-1.6	5.2	0.39	840.87	854.87	18.87

n = 989

[a] Conventional risk factors based on QRISK2: age, sex, smoking, atrial fibrillation, chronic kidney disease, arthritis, hypertension, body mass index, systolic blood pressure, total:high-density lipoprotein cholesterol, social status, baseline cardiovascular disease status (myocardial infraction, angina, transient ischemic attack and stroke) and lipid lowering medication

[b] n = 186 for event, n = 803 for no event

ABI: ankle brachial index; AIC: Akaike's information criterion; *g*: inflammation factor; GGT: gamma-glutamyl transpeptidase; hs-cTnT: high-sensitivity cardiac troponin T; NT-proBNP: N-terminal pro-brain natriuretic peptide; OR: odds ratio; SD: standard deviation

8 Results III: Associations between metabolomics data and cardiovascular disease in the UCLEB consortium cohorts

This chapter describes the missing data and baseline characteristics of the five UCLEB cohorts used in this thesis: BRHS, BWHHS, ET2DS, SABRE and WHII. Prevalent and incident cardiovascular events are summarised and descriptive analyses are presented for the 228 metabolites measured. Associations between the metabolites and CVD are presented initially for each cohort individually, unadjusted for any risk factors. Finally, associations between the metabolites and CVD are presented for the data from the five cohorts combined, using four different models to adjust for risk factors.

8.1 Missing data

Analyses of the metabolomics data explored four different models for the combined UCLEB data: a model adjusted only for cohort (model 1); a model adjusted for cohort, age and sex (model 2); a model adjusted additionally for the most widely accepted traditional cardiovascular risk factors (model 3); and a model adjusted additionally for a small panel of additional key risk factors included in the majority of previously developed cardiovascular risk scores (model 4). Table *8-1* summarises the amount of missing data for each risk factor in the five UCLEB cohorts required for these four models. Data was complete for age, sex, anti-hypertensive medication and lipid lowering medication in all five studies. This meant that risk factor data was complete for the model adjusted for cohort only (model 1) and for the model additionally adjusted for age and sex (model 2). Most of the remaining variables had very little missing data, though most notably the BWHHS was missing 25% of the values for social status. In addition, WHII was missing about 13% of the values for HDL cholesterol and 12% of the values for BMI. Overall, the proportions of cases missing in the combined data due to one or more missing value for models 3 (adjusted additionally for HDL to total cholesterol ratio, sBP, smoking status and lipid lowering and anti-hypertensive medication) and 4 (adjusted additionally for social status, BMI, eGFR and ethnicity) were 4.98% and 24.48% respectively.

The amount of missing data for models 1, 2 and 3 was below the 5% level considered acceptable for complete case analysis. Although the amount of missing data for model 4 was higher than would be desirable for a complete case analysis, alternative missing data methods such as multiple imputation were not considered feasible for this setting. In the case of combining multiple cohorts, a multilevel multiple imputation method would be required in order to allow for between-study heterogeneity and avoid biased results (Jolani et al., 2015). This could be implemented using novel techniques such as joint modelling or fully conditional specification models and would require study to be regarded as a fixed effect (Audigier et al., 2017). However, such methods are still under development and implementation has many limitations: such imputation models can have convergence problems leading to biased results; the method can yield biased estimates of measures of uncertainty; substantial computational power is required to create a large number of imputed datasets; and finally, such methods were still under development for statistical software packages at the time of analysis (Koopman et al., 2008; Jolani et al., 2015). Furthermore, associations between missingness and both key predictors (age and sex) and cardiovascular disease were investigated and no statistically significant associations were found. This indicated that the assumption of MCAR, required for a complete case analysis, was not violated. For all these reasons, a complete case analysis was carried out for each model.

The amount of missing data in the metabolites are summarised in Table *C-1* in Appendix C As discussed above, multilevel multiple imputation was not considered possible in this context, therefore each subsequent model is calculated for the maximum number of metabolites available.

Table 8-1 Available values for the risk factor variables in the UCLEB consortium cohorts

	BRHS	BWHHS	ET2DS	SABRE	WHII
n	*141*	*352*	*1058*	*448*	*248*
n - Complete data for model 3[a] (%)	*135 (95.74%)*	*343 (97.44%)*	*1049 (99.15%)*	*414 (92.41%)*	*194 (78.23%)*
n - Complete data for model 4[b] (%)	-	*255 (72.44%)*	*1039 (98.20%)*	*403 (89.96%)*	-
Age	100.00%	100.00%	100.00%	100.00%	100.00%
Sex	100.00%	100.00%	100.00%	100.00%	100.00%
High-density lipoprotein cholesterol	97.16%	97.73%	99.34%	92.41%	86.69%
Total cholesterol	99.29%	98.30%	99.34%	99.78%	100.00%
Systolic blood pressure	98.58%	99.72%	99.81%	100.00%	99.60%
Smoking status	100.00%	100.00%	100.00%	100.00%	91.13%
Anti-hypertensive medication	100.00%	100.00%	100.00%	100.00%	100.00%
Lipid lowering medication	100.00%	100.00%	100.00%	100.00%	100.00%
Body mass index	97.87%	99.15%	99.91%	99.55%	87.50%
Estimated glomerular filtration rate	NA	99.72%	99.34%	99.78%	NA
Ethnicity	95.74%	99.72%	100.00%	100.00%	100.00%
Social status	97.16%	75.00%	99.05%	97.77%	100.00%

[a]Model 3 adjustment covariates: age, sex, high-density lipoprotein:total cholesterol, systolic blood pressure, smoking status, anti-hypertensive medication and lipid lowering medication

[b]Model 4 adjustment covariates: age, sex, high-density lipoprotein:total cholesterol, systolic blood pressure, smoking status, anti-hypertensive medication, lipid lowering medication, body mass index, estimated glomerular filtration rate, ethnicity and social status

BRHS: British Regional Heart Study; BWHHS: British Women's Heart and Health Study; ET2DS: Edinburgh Type 2 Diabetes Study; NA: not available; SABRE: Southall and Brent Revisited Study; WHII: Whitehall-II Study

8.2 Baseline characteristics of the UCLEB cohorts

Baseline characteristics, including traditional cardiovascular risk factors, of the five UCLEB cohorts used in this thesis are presented in Table *8-2*. The baseline cohorts of BRHS, BWHHS, ET2DS, SABRE and WHII comprised 141, 352, 1058, 448 and 248 individuals with prevalent type 2 diabetes at baseline respectively.

At baseline in the BRHS, the average age was 69.53 years and all participants in this cohort were male. 76 (53.90%) participants were taking anti-hypertensive medication and 21 (14.89%) were taking lipid lowering medication. The vast majority of the cohort are white (*n*=132, 97.78%) and the social status of the participants was split as follows: 13 (9.49%) unskilled, 108 (78.83%) skilled and 16 (11.68%) professional.

In the BWHHS, the average age at baseline was 69.47 years and all participants in this cohort were female. 185 (52.56%) participants were taking anti-hypertensive medication and 58 (16.48%) were taking lipid lowering medication. The vast majority of the cohort are white (*n*=348, 99.15%) and the social status of participants was split as follows: 4 (1.52%) unskilled, 193 (73.11%) skilled and 67 (25.38%) professional.

The average age at baseline in the ET2DS was 67.89 and 513 (48.49%) participants were female. 867 (81.95%) participants were taking anti-hypertensive medication and 905 (85.54%) were taking lipid lowering medication. Most of the ET2DS cohort were white (*n*=1041, 98.39%) and the social status of participants was split as follows: 177 (16.89%) unskilled, 166 (15.84%) skilled and 705 (67.27%) professional.

In SABRE, the average age at baseline was 54.24 years and only 50 (11.16%) of the participants were female. 102 (22.77%) participants were taking anti-hypertensive medication and only 3 (0.67%) were recorded as taking lipid lowering medication. The ethnicity of participants was broken down as follows: 87 (19.42%) white, 324 (72.32%) Asian and 37 (8.26%) Afro-Caribbean. The social status was split as follows: 49 (11.19%) unskilled, 332 (75.80%) skilled and 57 (13.01%) professional.

Finally, the average age at baseline in WHII was 57.79 years and 77 (31.05%) participants were female. 80 (32.36%) participants were taking anti-hypertensive medication and 21 (8.47%) participants were taking lipid lowering medication. The ethnicity of participants was broken down as follows: 193 (77.82%) white and 55 (22.18%) Asian. All participants in WHII were categorised as skilled in terms of social status which may reflect a low proportion of participants considered "professional" in the overall WHII study or a lower prevalence of type 2 diabetes among the "professional" group.

When combined, the total number of participants available for analysis was 2247. 992 (44.15%) participants were female and the average age at baseline was 64.41 years. 1310 (58.30%) participants were taking anti-hypertensive medication and 1008 (44.86%) participants were taking lipid lowering medication. In the three cohorts used for the final model including BMI, social status, eGFR and ethnicity (BWHHS, ET2DS and SABRE), 1476 (79.44%) participants were white, 337 (18.14%) participants were Asian, 41 (2.21%) participants were Afro-Caribbean and 3 participants were classified as "other" ethnicity. Finally, the social status of these three cohorts was split as follows: 230 (12.38%) unskilled, 691 (37.19%) skilled and 829 (44.62%) professional.

Table 8-2 Baseline characteristics of the UCLEB consortium cohorts

Variable	***BRHS*** *(max n=141)*	***BWHHS*** *(max n=352)*	***ET2DS*** *(max n=1058)*	***SABRE*** *(max n=448)*	***WHII*** *(max n=248)*	***Combined cohorts*** *(max n=2247)*
Age (years)	69.53 ± 5.21	69.47 ± 5.68	67.89 ± 4.20	54.24 ± 6.82	57.79 ± 6.09	64.41 ± 8.07
Sex [female]	All male	All female	513 (48.49)	50 (11.16)	77 (31.05)	992 (44.15)
Total cholesterol (mmol/L)	5.60 ± 1.09	6.41 ± 1.38	4.31 ± 0.90	6.01 ± 1.10	5.97 ± 1.13	5.24 ± 1.39
HDL cholesterol (mmol/L)	1.14 ± 0.27	1.48 ± 0.42	1.29 ± 0.36	1.18 ± 0.34	1.37 ± 0.38	1.30 ± 0.38
sBP (mmHg)	151.80 ± 25.64	154.46 ± 26.06	133.20 ± 16.39	132.19 ± 19.26	128.74 ± 17.94	136.99 ± 21.51
Smoking status [ever]	15 (10.64)	168 (47.73)	647 (61.15)	161 (35.94)	32 (14.16)	1023 (45.98)
Anti-hypertensive medication [yes]	76 (53.90)	185 (52.56)	867 (81.95)	102 (22.77)	80 (32.26)	1310 (58.30)
Lipid lowering medication [yes]	21 (14.89)	58 (16.48)	905 (85.54)	3 (0.67)	21 (8.47)	1008 (44.86)
BMI (kg/m^2)	28.16 ± 3.59	29.88 ± 5.79	31.42 ± 5.67	27.20 ± 4.07	27.01 ± 4.50	29.69 ± 5.51
eGFR (mL/min)	NA	66.56 ± 11.80	78.32 ± 23.13	97.41 ± 35.82	NA	89.73 ± 35.49
Ethnicity						
White	132 (97.78)	348 (99.15)	1041 (98.39)	87 (19.42)	193 (77.82)	1801 (80.40)
Asian	0	0	13 (1.24)	324 (72.32)	55 (22.18)	392 (17.50)
Afro-Caribbean	2 (1.48)	2 (0.57)	2 (0.19)	37 (8.26)	0	43 (1.92)
Other	1 (0.74)	1 (0.28)	2 (0.19)	0	0	4 (0.18)
Social status						
Unskilled	13 (9.49)	4 (1.52)	177 (16.89)	49 (11.19)	0	243 (11.38)
Skilled	108 (78.83)	193 (73.11)	166 (15.84)	332 (75.80)	248 (100)	1047 (49.04)
Professional	16 (11.68)	67 (25.38)	705 (67.27)	57 (13.01)	0	845 (39.58)

Data are presented as means ± SD or *n* (%)

BMI: body mass index; eGFR: estimated glomerular filtration rate; HDL: high-density lipoprotein; NA: not available; sBP: systolic blood pressure

8.3 CVD in the UCLEB cohorts

Table *8-3* shows the breakdown of both prevalent and incident cardiovascular events available for each of the five UCLEB cohorts. Unfortunately, the definitions of "incident" events across the studies and across event types was not consistent. Additionally, the exclusion of all participants with previous CVD in order to obtain incident events only would not have been possible while retaining adequate statistical power for subsequent analyses (see Chapter 5 for a detailed discussion of these issues). Therefore, the outcome of CVD was defined as either a participant having prevalent CVD at baseline or experiencing a fatal or non-fatal cardiovascular event during the study follow-up period. In the BRHS, BWHHS, ET2DS, SABRE and WHII cohorts there were 52 (36.88%), 174 (49.43%), 451 (42.63%), 253 (56.47%) and 74 (29.84%) participants with the outcome of CVD respectively, giving a total of 1005 (44.73%) participants with CVD across all studies.

Table 8-3 Summary of cardiovascular events available in the UCLEB consortium cohorts

Cardiovascular event	***BRHS*** *(n=141)*	***BWHHS*** *(n=352)*	***ET2DS*** *(n=1058)*	***SABRE*** *(n=448)*	***WHII*** *(n=248)*
Prevalent MI	20 (14.2)	27 (7.7)	147 (13.9)	66 (14.7)	16 (6.5)
Prevalent revascularisation	NA	10 (2.8)	107 (10.1)	1 (0.3)	5 (2.2)
Prevalent angina	NA	76 (21.6)	295 (27.9)	25 (5.6)	35 (15.2)
Prevalent stroke	15 (10.6)	36 (10.2)	62 (5.9)	21 (4.7)	0
Incident angina	NA	70 (19.9)[b]	44 (4.2)[b]	64 (17.7)[b]	27 (11.0)[b]
Incident fatal or non-fatal MI or revascularisation	33 (23.4)[a]	79 (22.4)[a]	97 (9.2)[b]	212 (51.1)[a]	26 (12.3)[a]
Incident fatal or non-fatal stroke	NA	74 (21.0)[a]	59 (5.6)[b]	106 (25.6)[a]	13 (5.3)[a]
Incident fatal or non-fatal CHD or stroke	NA	127 (36.1)[a]	149 (14.1)[b]	251 (60.0)[a]	37 (17.5)[a]

Data are presented as *n* (%)

[a] Incident events defined as first event ever

[b] Incident events defined as first event during follow-up

BRHS: British Regional Heart Study; BWHHS: British Women's Heart and Health Study; CHD: coronary heart disease; ET2DS: Edinburgh Type 2 Diabetes Study; MI: myocardial infarction; NA: not available; SABRE: Southall and Brent Revisited Study; WHII: Whitehall-II Study

8.4 Descriptive statistics of metabolites

Summaries of the distributions of each metabolite for the combined UCLEB data can be found in Figure C-1 in Appendix C in the form of histograms. Distributions of the metabolites in each individual study were investigated and found to be consistent with the combined results.

Descriptive statistics, including medians, interquartile ranges (IQR) and number of missing values, were compared across individual studies (results presented in Table C-1 in Appendix C). Medians and IQRs were found to be reasonably consistent across the studies. Most metabolites in the individual studies had little missing data (5% or less missing data per variable) and missing data ranged from zero to 347. The highest amount of missing data was found in the ET2DS (*n*=347, 32.80%) for five of the derived metabolites (phospholipids to total lipids ratio in very large HDL; total cholesterol to total lipids ratio in very large HDL; cholesterol esters to total lipids ratio in very large HDL; free cholesterol to total lipids ratio in very large HDL and triglycerides to total lipids ratio in very large HDL). These measures are all computed as a ratio between two different metabolites. Therefore, if the level of one of these metabolites is zero then the derived variable cannot be calculated and the value is defined as missing. As discussed previously, despite missing values in the metabolomics data, multilevel multiple imputation was not considered possible in this context and each subsequent model is calculated for the maximum number of metabolites available in the combined data.

8.5 Associations between metabolites and cardiovascular disease in individual UCLEB studies

In order to obtain an initial impression of associations in the data, unadjusted univariate models were fitted between each metabolite and the outcome of CVD for each UCLEB cohort separately. The results of these models are summarised using p-values and ORs in Figure 8-1 and Figure 8-2 respectively. The results have been colour coded to differentiate the 11 groups of metabolites discussed in section 5.1.3.4 of Chapter 5. As discussed in Chapter 5, a range of adjusted thresholds for statistical

significance are presented in Figure 8-1: a 5% FDR, a 1% FDR and a Bonferroni correction.

Significant associations between any of the metabolites and CVD, according to all three selected significance thresholds, were found only for the ET2DS. This result was unsurprising given that the ET2DS was the largest cohort (*n*=1058) and had the highest number of CVD outcomes (*n*=451, 42.63%). The metabolites which showed the most convincing statistically significant associations (that is, retaining statistical significance even above the strictest adjustment of the Bonferroni correction) in the ET2DS were: one of the VLDL particles, phospholipids to total lipids ratio in medium VLDL particles (displayed in turquoise); two of the IDL particle, total cholesterol to total lipids ratio and triglycerides to total lipids ratio in IDL (displayed in light yellow); a number of the HDL particles (displayed in red) and lipoprotein particle sizes (displayed in blue); two of the apolipoproteins, apolipoprotein A1 and ratio of apolipoprotein B to apolipoprotein A1 (displayed in orange); and creatinine (displayed in light green). Although statistically significant associations were only found in ET2DS, the overall pattern of the p-values was similar across all five studies. In particular, the group of HDL particles showed a strong association with CVD, as well as creatinine and the inflammation metabolite glycoprotein (the rightmost metabolite, displayed in yellow). In the WHII cohort, strong associations were also observed between CVD and a number of the VLDL particles and fatty acids, although this relationship was not replicated in the other studies. Overall, the pattern of associations was fairly consistent across all five studies.

The pattern of the effect sizes and directions of the associations was assessed using ORs in order to determine whether these were also consistent across the studies and therefore indicate that it would be reasonable to combine the data in the next stage of analyses. Figure 8-2 shows that results were reasonably similar across all five studies. It was observed in all the cohorts that the first half of the VLDL particle group had almost identical ORs for each measure. Initially, I thought that this could have been caused by the way that cases and controls had been selected in some studies, but after further investigation this was discovered not to be the case since

none of the studies used this design. The most likely explanation is that these metabolites essentially measure the same particle in different ways.

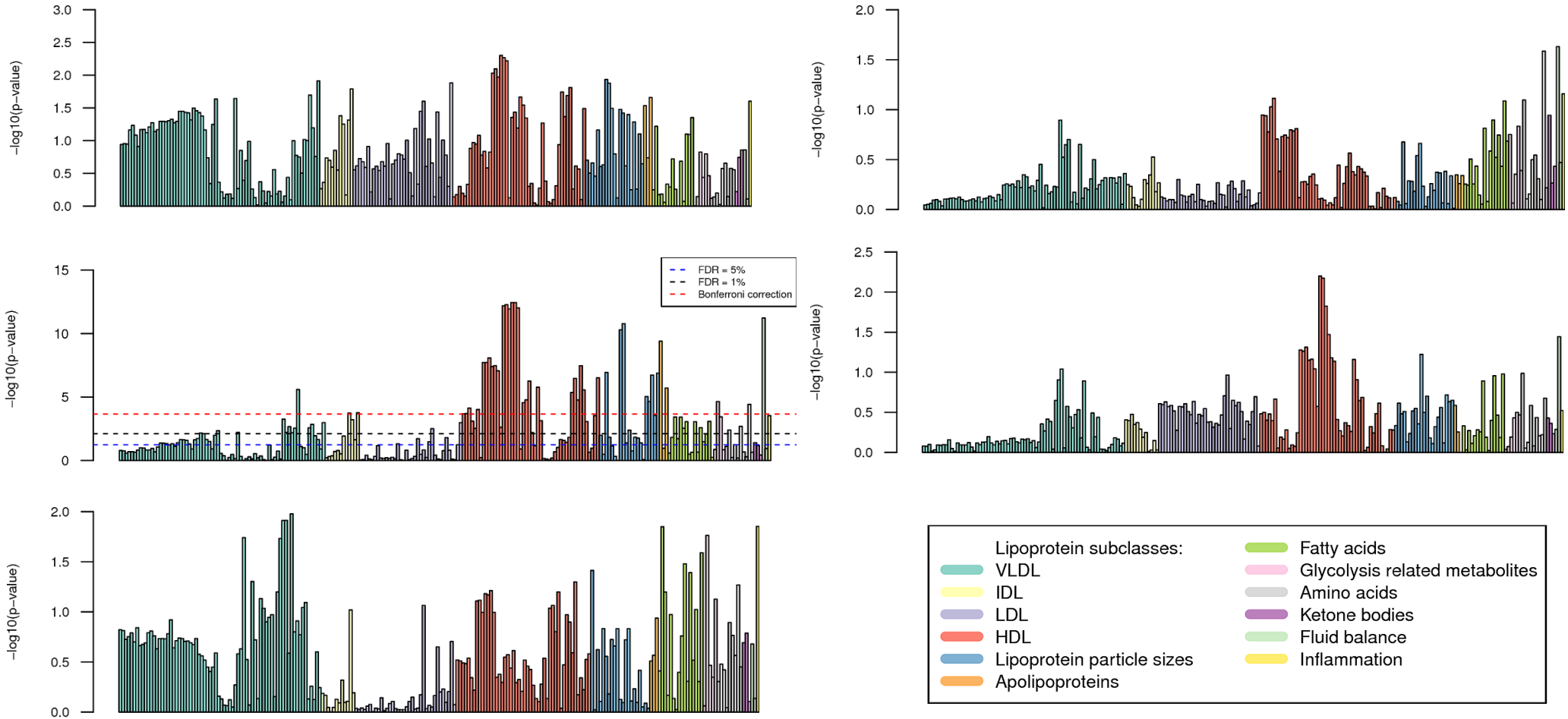


Figure 8-1 Bar plots of p-values for univariate analysis of metabolites and cardiovascular disease in the individual UCLEB cohorts.

Top left: BRHS; top right: BWHHS; middle left: ET2DS; middle right: SABRE; bottom left: WHII; bottom right: metabolite key.

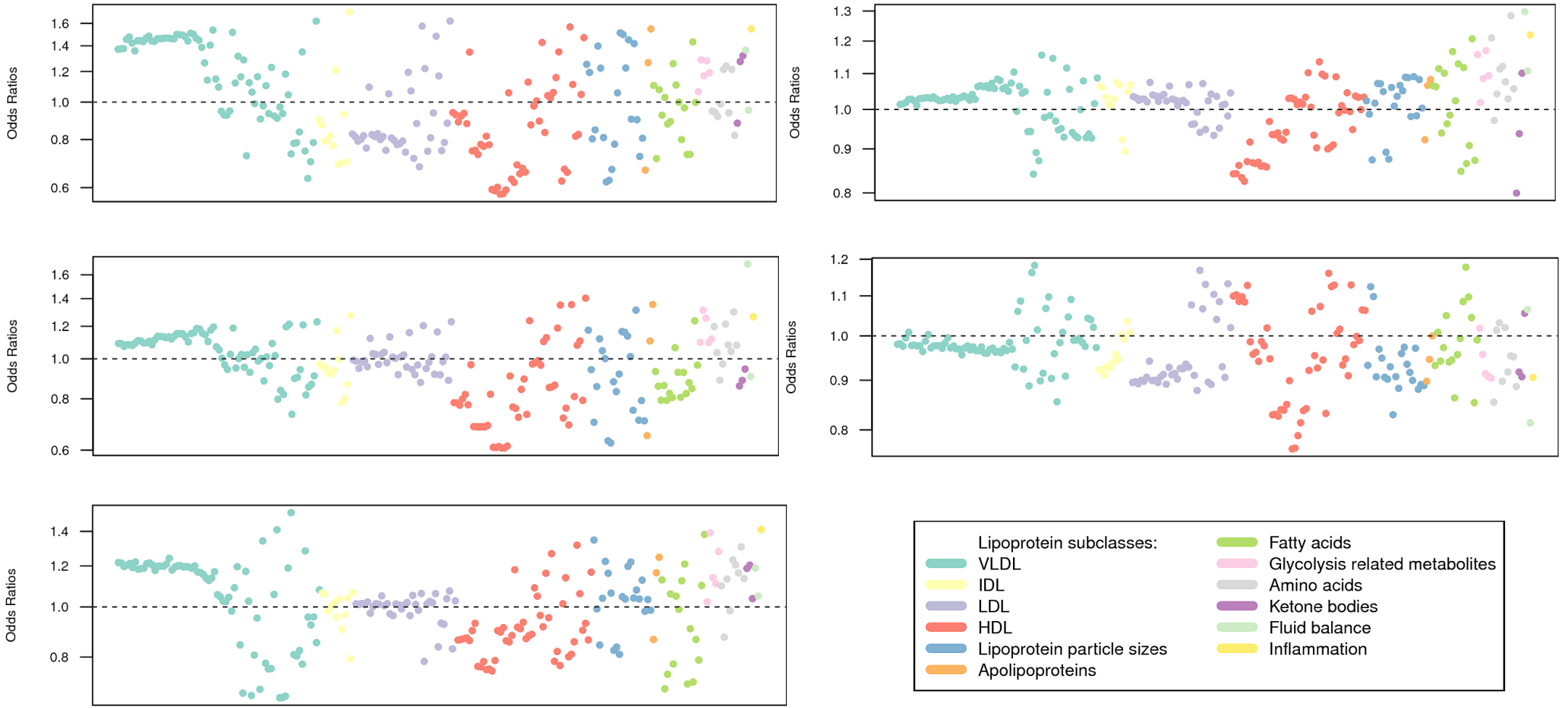


Figure 8-2 Odds ratios from univariate analysis of metabolites and cardiovascular disease in the individual UCLEB cohorts.

Top left: BRHS; top right: BWHHS; middle left: ET2DS; middle right: SABRE; bottom left: WHII; bottom right: metabolite key.

8.6 Associations between metabolites and CVD in the combined UCLEB data

The final stage of metabolomics analysis involved combining the data from all five cohort studies and investigating the associations between the metabolites and CVD using the following four models:

Model 1: adjusted only for cohort

Model 2: adjusted for cohort, age and sex

Model 3: adjusted for cohort, age, sex, HDL to total cholesterol ratio, sBP, smoking status, anti-hypertensive medication and lipid lowering medication

Model 4: adjusted for cohort, age, sex, HDL to total cholesterol ratio, sBP, smoking status, anti-hypertensive medication, lipid lowering medication, social status, BMI, eGFR and ethnicity.

The maximum *n* for the subsequent models is 2247 and the number of participants with CVD in the combined cohorts is 1005 (44.73%). Note that at the final stage of analysis (model 4), two of the cohort studies, BRHS and WHII, are dropped since eGFR was not available.

8.6.1 Associations between metabolites and CVD adjusted for cohort

The first models fitted for the combined data were adjusted only for cohort. The results from these models are summarised in Figure 8-3 and Figure 8-4, using p-values and ORs respectively. Many of the metabolites remained statistically significantly associated with CVD even using the most stringent multiple testing adjustment, the Bonferroni correction. These metabolites included many of the HDL particles, a number of the lipoprotein particle sizes, two of the apolipoprotein measures and creatinine. This pattern was similar to those observed in the analyses of the individual studies (Figure 8-1). Similarly, the pattern of the ORs was consistent with the previous results. The largest effect sizes were observed for a number of the HDL particles and creatinine.

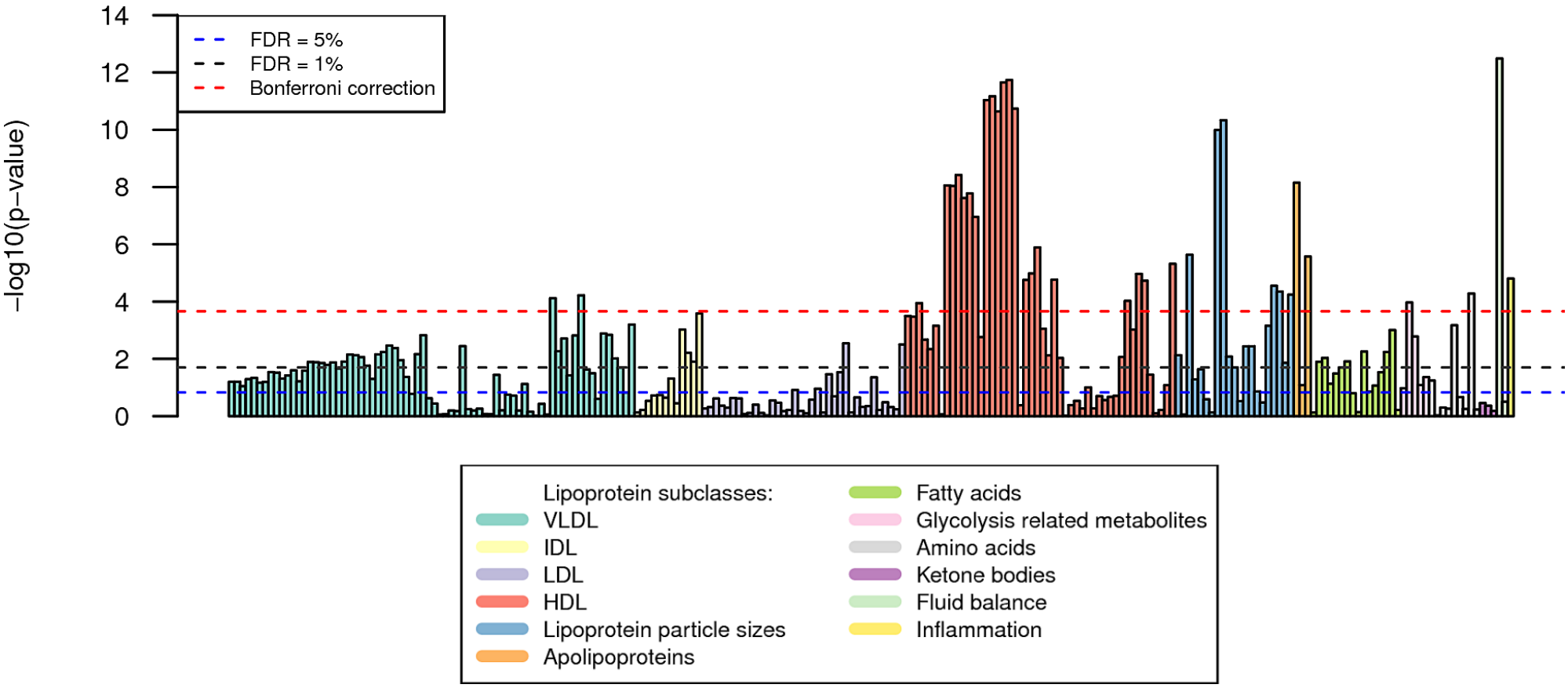


Figure 8-3 Bar plots of p-values for Model 1 analysis of metabolites and cardiovascular disease in the combined UCLEB cohorts.

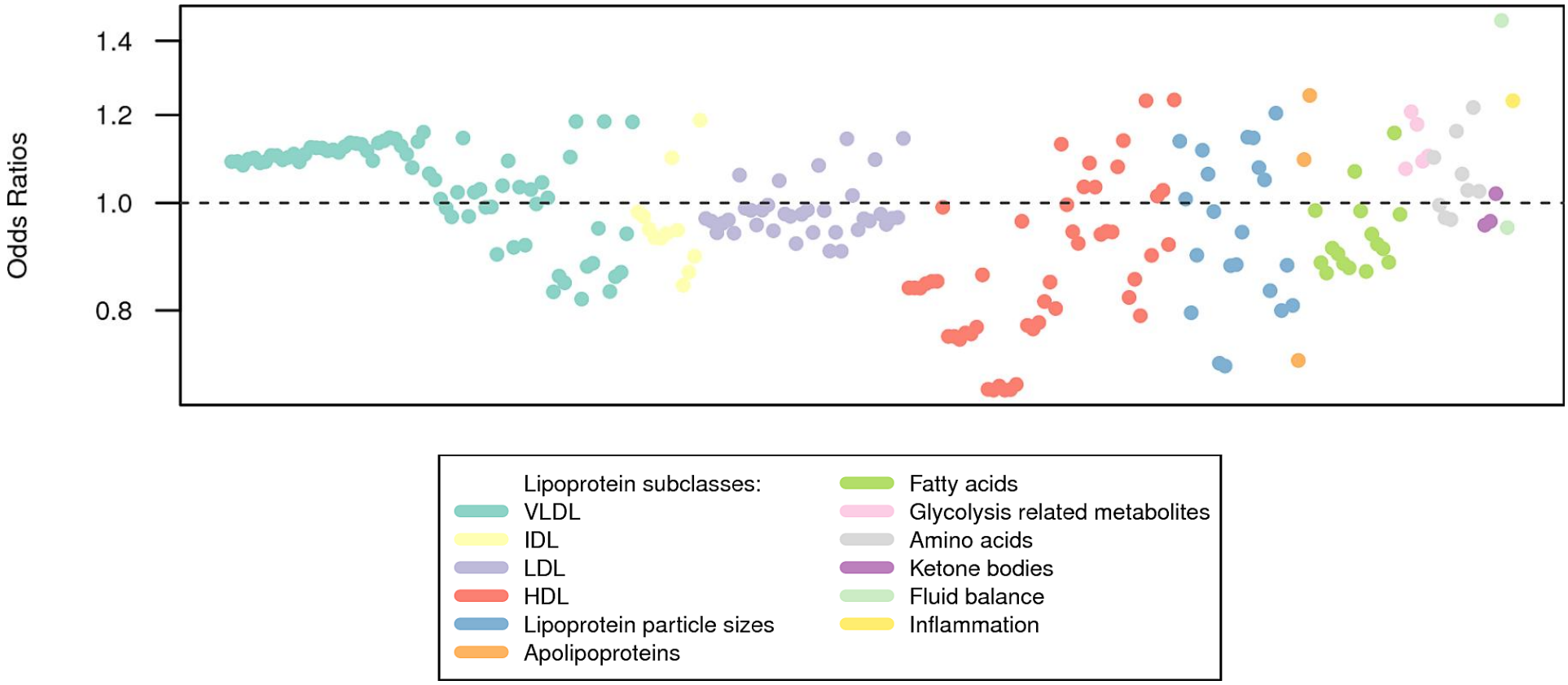


Figure 8-4 Odds ratios from Model 1 analysis of metabolites and cardiovascular disease in the combined UCLEB cohorts.

8.6.2 Associations between metabolites and CVD adjusted for cohort, age and sex

The next models fitted for the combined data were adjusted for age and sex as well as cohort. The results from these models are summarised in Figure 8-5 and Figure 8-6. After additional adjustment for age and sex, statistically significant associations between CVD and many of the same metabolites persisted, even at the Bonferroni correction level. The association between CVD and the inflammation metabolite glycoprotein appeared to be strengthened following this adjustment. The pattern of the ORs also remained very similar to the previous results. The largest effect sizes were observed for a number of the HDL particles, which were negatively associated with CVD, and creatinine, which had a positive association with CVD.

8.6.3 Associations between metabolites and CVD adjusted for cohort, age, sex and traditional cardiovascular risk factors

Figure 8-7 and Figure 8-8 summarise the results for the models which adjusted for HDL to total cholesterol ratio, sBP, smoking status, lipid lowering medication and anti-hypertensive lowering medication, in addition to age, sex and cohort. After additional adjustment for these key risk factors, none of the metabolites were found to have statistically significant associations with CVD at the 1% FDR threshold. Creatinine and a number of the HDL particles were statistically significantly associated with CVD at the 5% FDR threshold, but only one metabolite was statistically significant at the Bonferroni correction level – phospholipids in small HDL. The strong associations between a number of the lipoprotein particle sizes and the apolipoproteins were drastically reduced. Again, the pattern of the ORs remained consistent with the previous results, although the effect directions of the VLDL particles and some of the HDL particles were reversed. The largest effect sizes were still observed for a number of the HDL particles and creatinine.

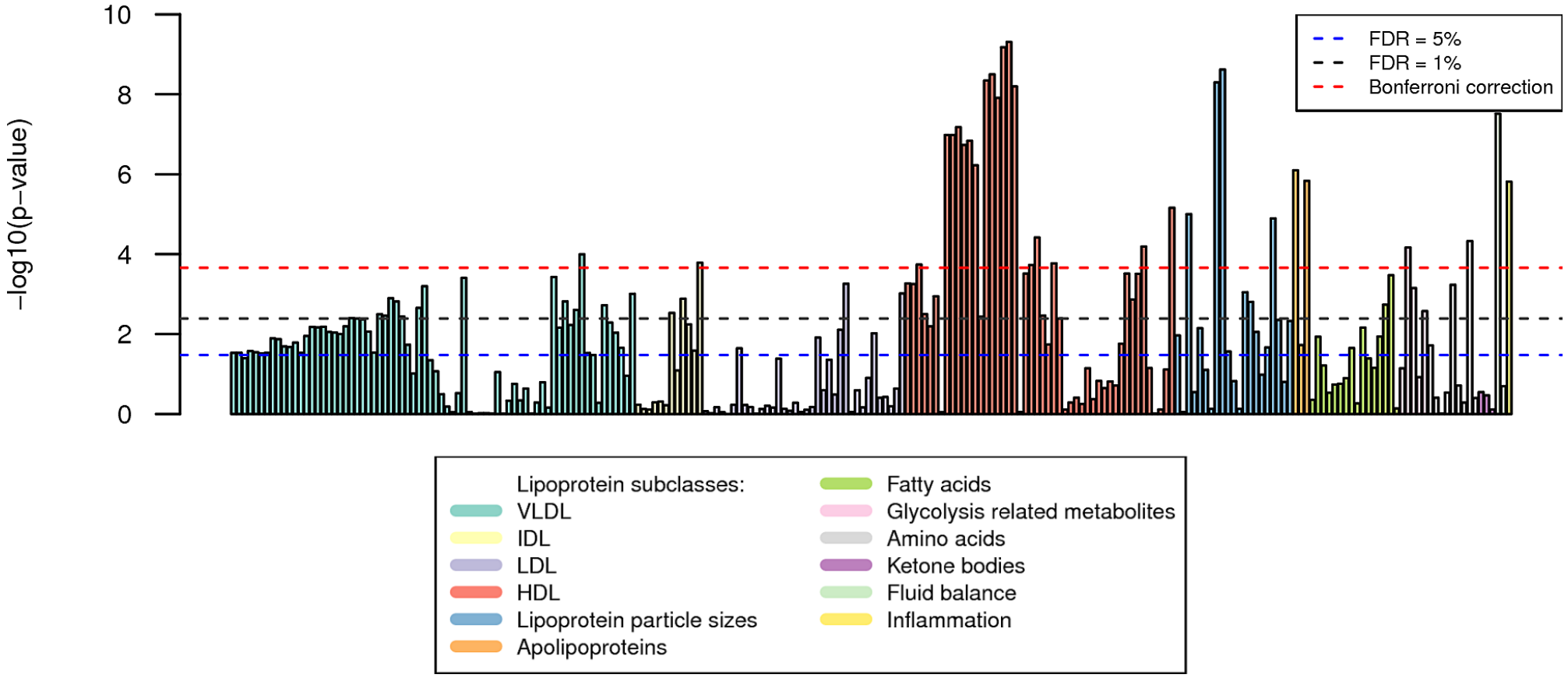


Figure 8-5 Bar plots of p-values for Model 2 analysis of metabolites and cardiovascular disease in the combined UCLEB cohorts.

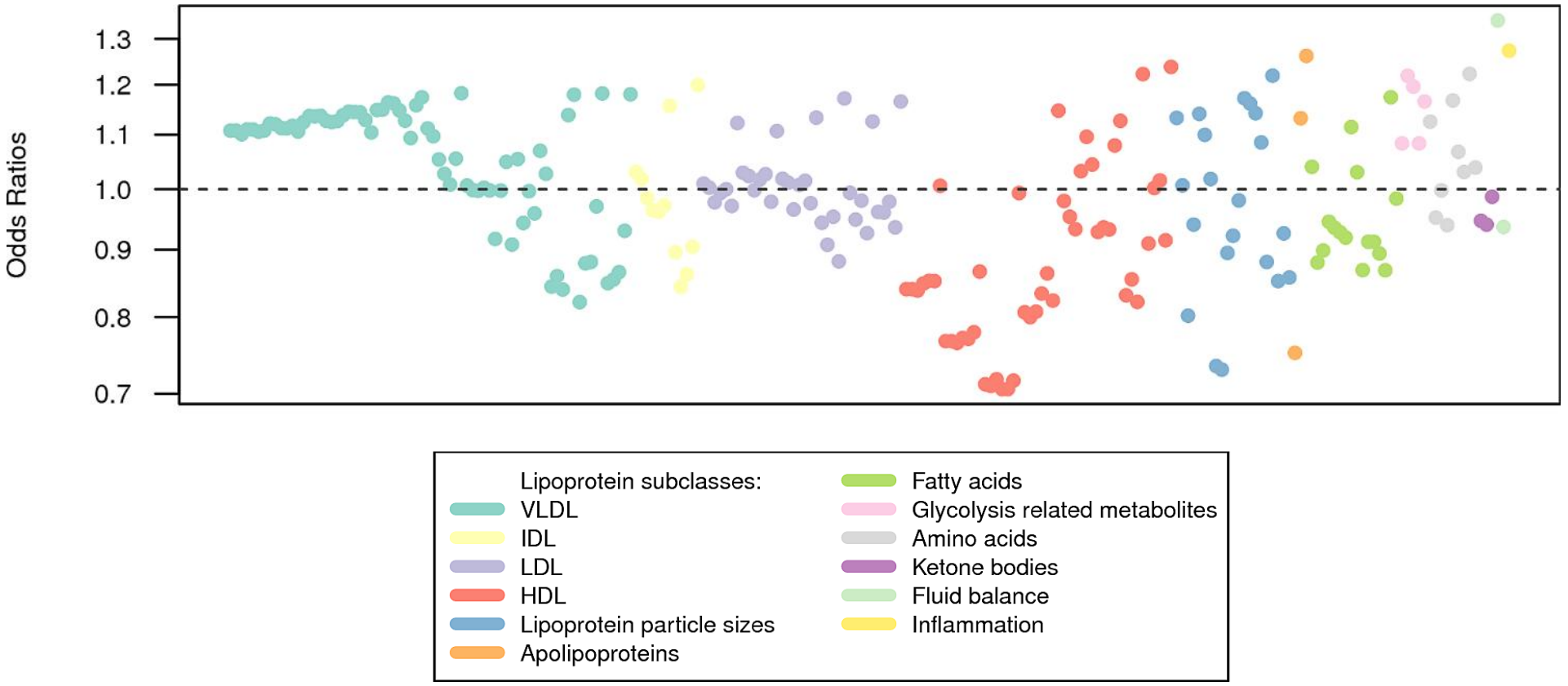


Figure 8-6 Odds ratios from Model 2 analysis of metabolites and cardiovascular disease in the combined UCLEB cohorts.

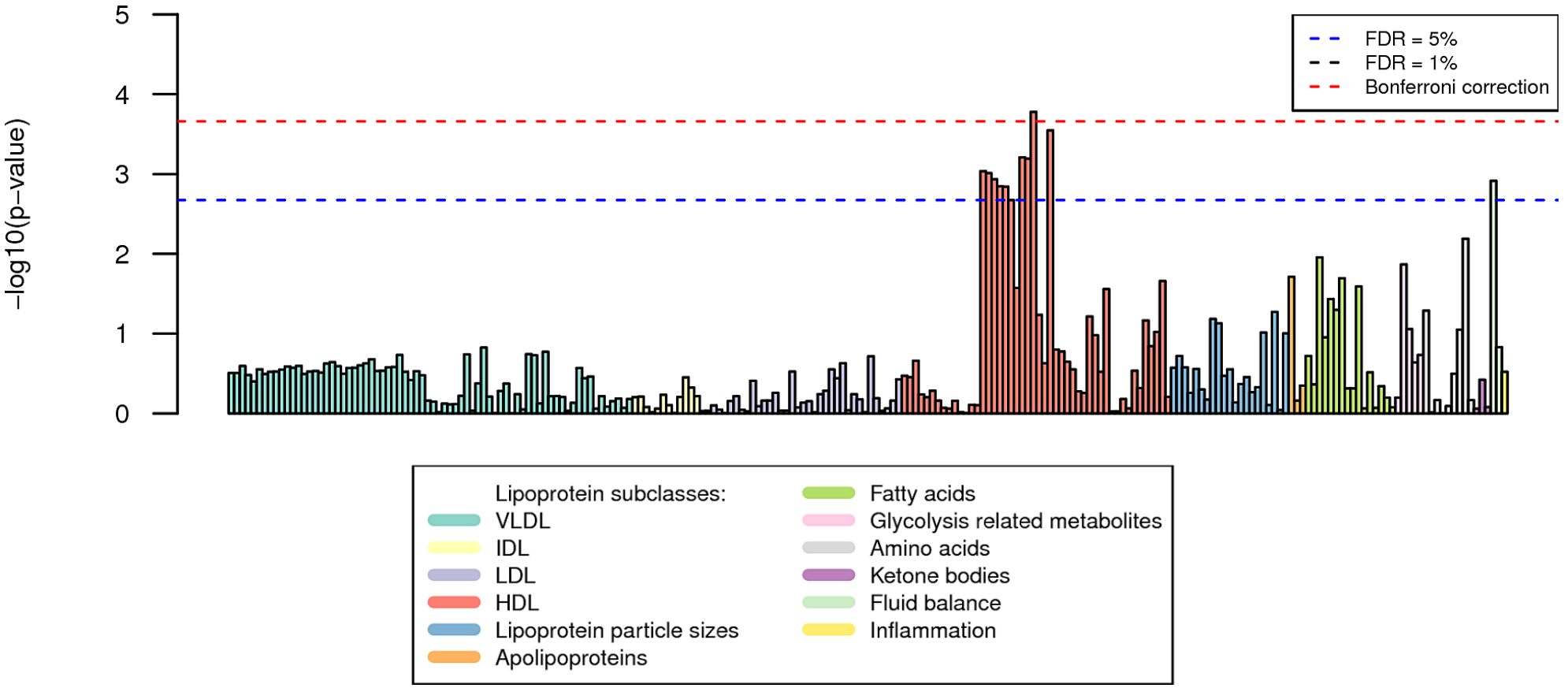


Figure 8-7 Bar plots of p-values for Model 3 analysis of metabolites and cardiovascular disease in the combined UCLEB cohorts.

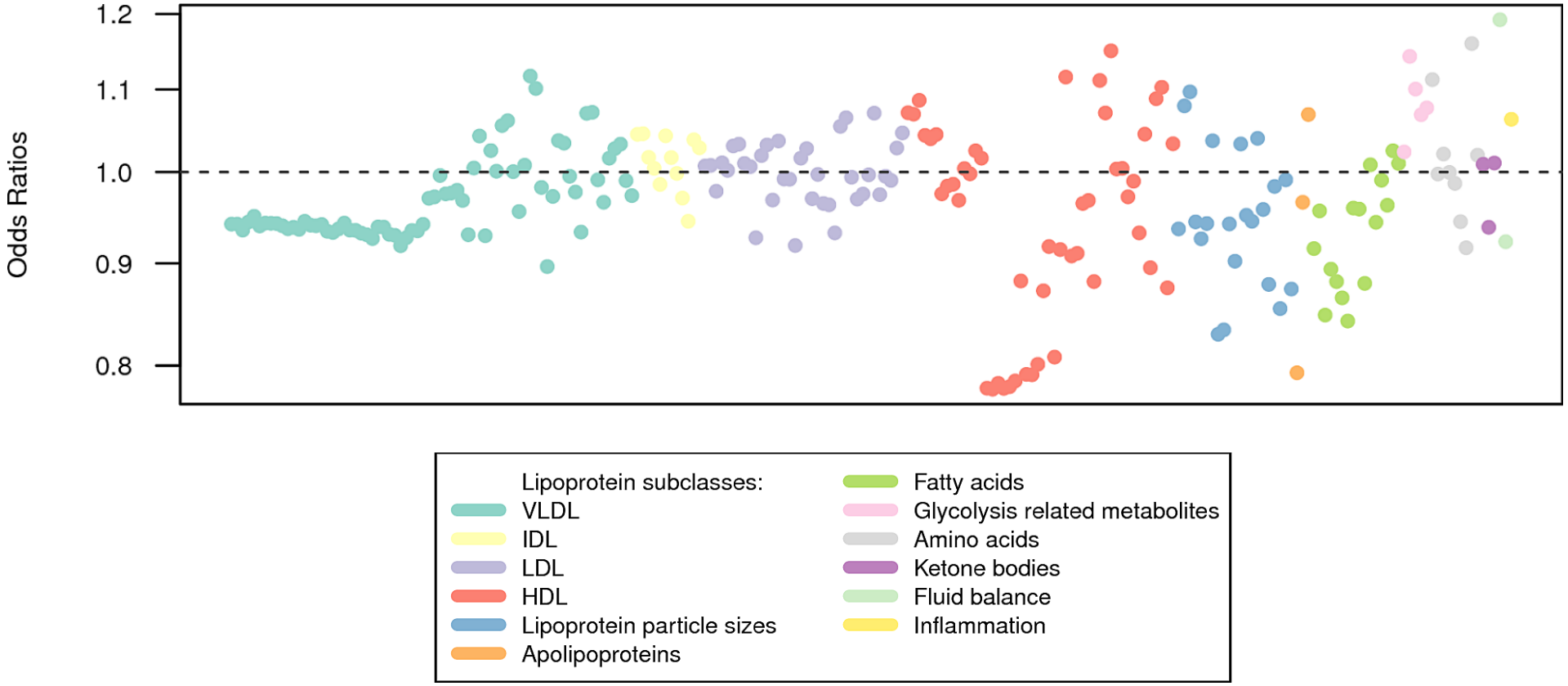


Figure 8-8 Odds ratios from Model 3 analysis of metabolites and cardiovascular disease in the combined UCLEB cohorts.

8.6.4 Associations between metabolites and CVD adjusted for cohort, age, sex, traditional cardiovascular risk factors, social status, BMI, eGFR and ethnicity

The results of the fully adjusted models which included all previous risk factors and additionally social status, BMI, eGFR and ethnicity are presented in Figure 8-9 and Figure 8-10. After adjustment for all risk factors, and consistent with the result for Model 3, none of the metabolites were found to have statistically significant associations with CVD at the 1% FDR threshold. Creatinine was no longer associated with CVD, an unsurprising result since Model 4 includes eGFR which is calculated using creatinine values. Four metabolites, all HDL particles, were statistically significant at the Bonferroni correction level – concentration of medium HDL particles, total lipids in medium HDL, phospholipids in medium HDL and phospholipids in small HDL. This result is perhaps more surprising given that Model 4 (and Model 3) included HDL to total cholesterol ratio as a covariate. The pattern of the ORs appeared almost identical to that from the Model 3 results. In order to check for heterogeneity among the five studies, interaction terms between cohort and each of the four metabolites most strongly associated with CVD were added to Model 3 (the most adjusted model which included all five studies). No statistically significant interactions were found, indicating that there is not significant evidence of inconsistency of metabolite effects across the five studies.

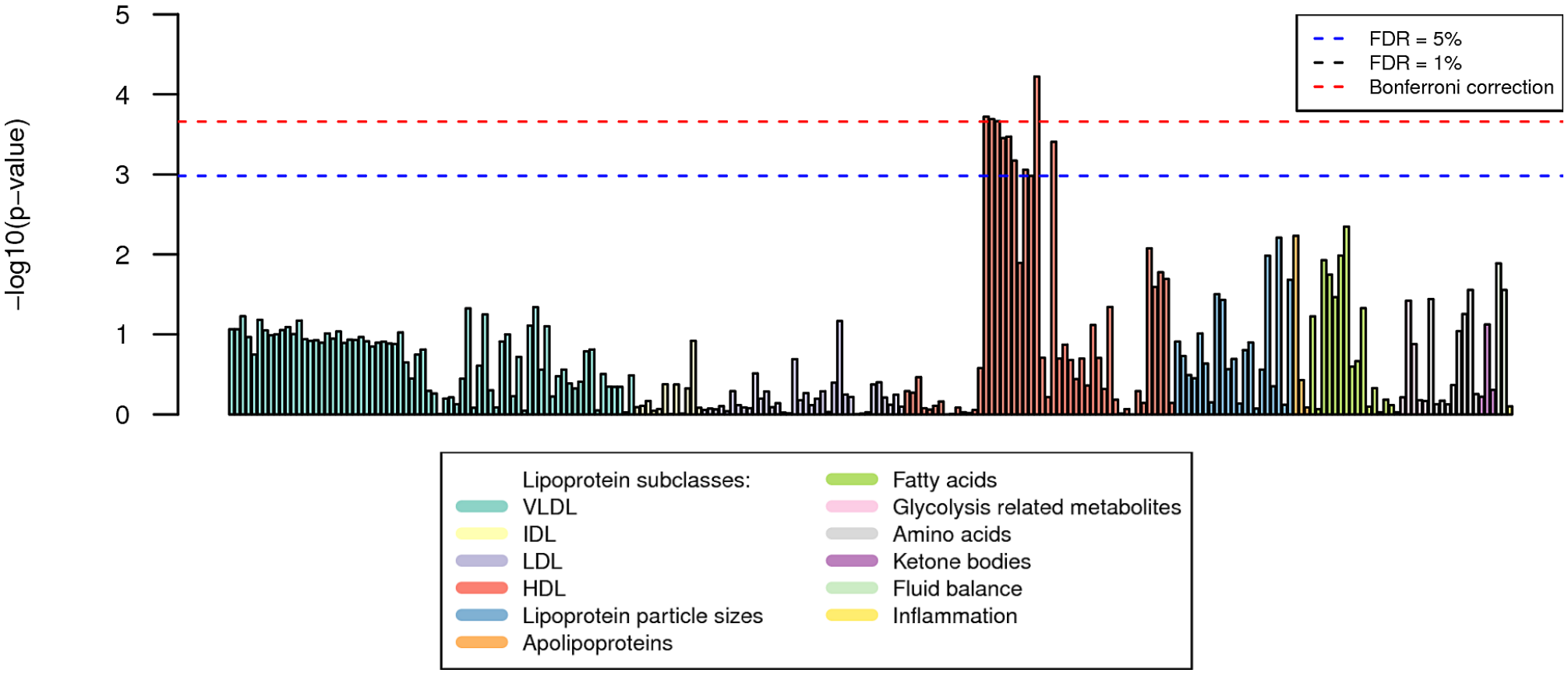


Figure 8-9 Bar plots of p-values for Model 4 analysis of metabolites and cardiovascular disease in the combined UCLEB cohorts.

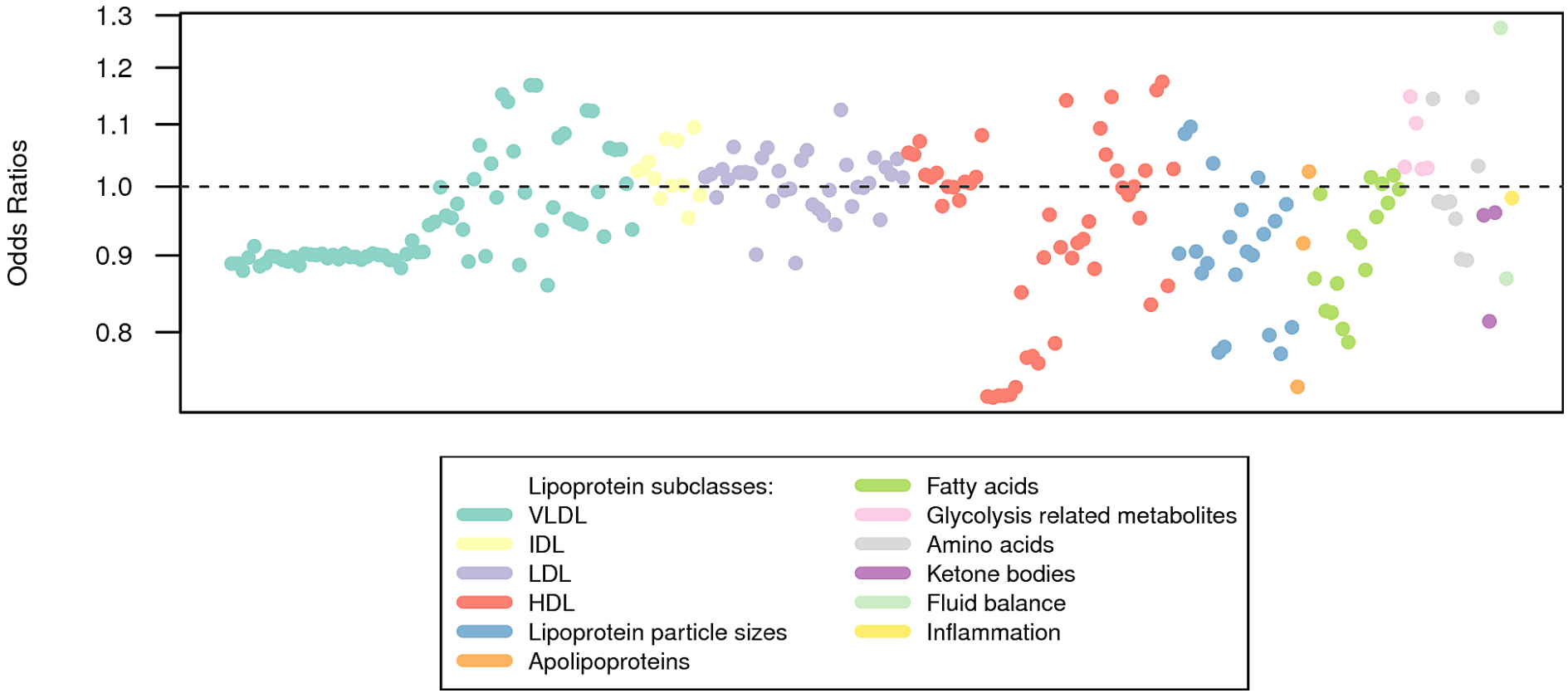


Figure 8-10 Odds ratios from Model 4 analysis of metabolites and cardiovascular disease in the combined UCLEB cohorts.

9 Discussion

This chapter summarises the key findings of the thesis, comparing them with previous studies. Strengths and limitations of the research are discussed, including the statistical methods used. The chapter concludes with recommendations for the direction of future research on the topic of improving cardiovascular risk scores using non-traditional biomarkers.

9.1 Key findings

9.1.1 Improving cardiovascular risk prediction using individual and combined biomarkers

In the ET2DS, a representative prospective study of older people with type 2 diabetes living in Scotland, it has been shown that three individual biomarkers, hs-cTnT, NT-proBNP and ABI, were significantly associated with the risk of experiencing an incident or recurrent cardiovascular event. The addition of each individual biomarker also improved the predictive ability of a basic model based on a carefully selected current cardiovascular risk score, QRISK2, with the greatest improvement found for hs-cTnT. GGT and an inflammation factor derived from CRP, IL-6 TNF-α and fibrinogen were not significantly associated with incident or recurrent cardiovascular events, despite both marginally improving model performance. When considered in combination, using all subsets regression, the greatest improvement in risk prediction was found for a model including ABI, hs-cTnT and GGT in addition to the QRISK2 predictors.

9.1.2 Associations between metabolomics data and CVD

In data from the UCLEB consortium of UK-based, prospective cohorts, a number of individual HDL particles out of a total of 228 metabolites measured using an NMR platform were significantly associated with CVD even after adjustment for a wide range of traditional cardiovascular risk factors (including age, sex, blood pressure, smoking and HDL to total cholesterol ratio). These metabolites were: concentration of medium HDL particles, total lipids in medium HDL, phospholipids in medium

HDL and phospholipids in small HDL. Creatinine was also significantly associated with CVD until adjustment for eGFR in the final stage of statistical modelling.

9.2 Improving cardiovascular risk prediction using individual and combined biomarkers

9.2.1 Strengths of the ET2DS

9.2.1.1 Recruitment and representativeness

One advantage of the study population of the ET2DS is that participants were recruited directly from a representative sampling frame of all individuals in Lothian with type 2 diabetes, the Lothian Diabetes Register, as opposed to a subgroup of people with diabetes identified from a larger study population which is the case in other studies on this topic (for example, van der Leeuw et al., 2016). The Lothian region in Scotland includes the city of Edinburgh and towns and rural locations in the surrounding areas, rather than recruitment from a single city or town which is also common in other cohort studies (Bruno et al., 2013; Kavousi et al., 2012). Furthermore, the study population included participants managing their type 2 diabetes through a range of treatment methods, from diet control to insulin treatment.

At baseline, participants in the ET2DS were broadly similar to those who chose not to participate, as shown by a comparison of key demographic characteristics and clinical features between study participants and non-respondents. Slightly more men than women were recruited into the study, and individuals from the least deprived group were overrepresented in the study, with those from the most deprived group underrepresented, suggesting that people from the least deprived group were more likely to agree to take part. However, these differences were small, and since it is expected that the risk factors will have similar relationships with cardiovascular outcomes across deprivation groups, this small degree of unrepresentativeness is not likely to be an issue. In general, differences in a wide range of characteristics indicated that the study population was largely representative of, and hence generalisable to, the target population.

9.2.1.2 Completeness and accuracy of data collection

A key strength of the ET2DS for this type of research is that the cohort has been extensively phenotyped so a wide range of risk factors and biomarkers were available for comprehensive analyses. Data were carefully collected, following appropriate protocols, ensuring a relatively small amount of missing data. Researchers involved in the ET2DS were given appropriate training in the collection of data in order to ensure that accurate information was recorded. Standard operating procedures were followed during clinic appointments in order to reduce observer bias. In the laboratories, strict quality control measures were used.

9.2.1.3 Cardiovascular follow up

Complete data on incident or recurrent cardiovascular events during follow-up was vital for valid analysis of the association between biomarkers and cardiovascular risk. In order to reduce loss to follow-up and avoid misdiagnosis, a thorough and systematic approach was taken to the follow-up data collection processes at both the four and eight year time points. Several sources were used to identify and confirm events at the four-year follow-up, including self-completed questionnaires, ECGs and ISD data linkage. At the eight year follow-up, ISD data linkage, which captures all hospital discharges for patients between pre-specified dates, was used in combination with clinical notes to corroborate further cardiovascular events. Identification of events was as comprehensive as possible, although if a participant had moved outside of Scotland between four and eight years after baseline then an event could have been missed.

9.2.1.4 Prospective design and sample size

The prospective design of the ET2DS is preferable for risk prediction analysis, as discussed in section 2.1 in Chapter 2, and allowed the thorough investigation of the relationship between baseline risk factors and subsequent development of CVD or experience of cardiovascular events. A further strength of this study is the relatively large sample size, as the power to detect associations in the data depends strongly on the number of observations. In the case of risk prediction, sample size requirements are often thought of in terms of events per variable (Steyerberg, 2009). It has been suggested that the minimum number of events per variable required for accurate

predictions is 10, and that less than 10 events per variable may lead to over fitting, though this rule should be taken as an approximate guide (Harrell et al., 1984, Peduzzi et al., 1995, Peduzzi et al., 1996). In the analysis in this thesis, the number of events which occurred during follow-up (208 events) was higher than the maximum number required according to this rule (full model with all risk factors and biomarkers included 18 predictors = 180 events).

9.2.2 Limitations of the ET2DS

9.2.2.1 Generalisability

The ET2DS was established, as described in Price et al., 2008, with the aim of investigating the relationships between potential risk factors and complications of type 2 diabetes such as cognitive decline, liver disease and CVD. Although, as discussed above, the study population was shown to be largely representative of the target population from which the study population was recruited, by design the participants recruited to the study were a group of older adults, aged between 60 and 75 years at baseline. Furthermore, the majority of the study participants were white, reflecting the limited ethnic variability in the Lothian region. These two features of the study population result in limited generalisability of the results to other populations such as younger adults and people of alternative ethnicities. Additionally, the potential impact of healthy survivor bias (the possibility that people who survive to age 60 years or above may be biologically different to those not surviving to this age) is unknown.

9.2.2.2 Prevalent CVD and prescription of lipid-lowering medication

A further potential limitation of the ET2DS is the high rates of both prevalent CVD in the study population at baseline (*n*=367, 35%) and participants prescribed with statins, or lipid-lowering medications, at baseline (*n*=912, 85.55%). These characteristics reflect the clinical status of this group of older adults with established type 2 diabetes, but they also raise questions about the impact of the results of this type of study on clinical treatment decisions. The QRISK2 score (Hippisley-Cox et al., 2010), which was used to build the basic model for the analyses in this thesis, was developed exclusively in people without previous diagnosis of CVD and the NICE clinical guidelines recommend not using the risk assessment tool for people

with pre-existing CVD (NICE CG181, 2016). However, excluding all participants in the ET2DS with prevalent CVD would have severely reduced the statistical power for the required analyses and is also not a realistic reflection of the characteristics of this particular group of patients. Similarly, the QRISK2 score was developed only in people who were not already prescribed statins, but excluding these participants in the ET2DS would again severely impact on statistical power and is not reflective of the treatment status of this high risk group. Furthermore, a change in the intensity of statin treatment may be beneficial for patients with type 2 diabetes (Stone et al., 2014) and future drug development may lead to new treatments that could be prescribed in addition to or instead of statins in this group.

9.2.3 Strengths of the analysis plan

Particular care was taken in developing the analysis plan, in order to ensure that the results were informative and clinically meaningful. The choice of risk score for the development of a basic model to which non-traditional biomarkers would be added was made according to a set of pre-specified characteristics (discussed in detail in the systematic review in Chapter 4). The key criterion was considered to be the requirement for recommendation of use of the score in current clinical guidelines. The addition of the non-traditional biomarkers to such a model therefore informs us of any added value of these biomarkers over and above the information that is currently used to make treatment decisions in clinical practice. Furthermore, the decision was taken to consider the full panel of additional biomarkers in combination, rather than focusing on one or two biomarkers which gave the strongest associations at the first stage of individual analysis. This avoided a prematurely narrow focus on any particular biomarker. Indeed, GGT was found in four of the five top combined models (including the top model) attained using all subsets regression, despite its weaker, non-significant association with cardiovascular events when added to the basic model alone. Although the underlying biological mechanisms explaining this phenomenon are unknown, the result highlights the importance of this unbiased approach to investigating new biomarkers.

9.2.4 Limitations of the analysis plan

The results of an analysis of new predictors are dependent on the specific basic model which is used. Unfortunately the coefficients for the QRISK2 score were unavailable and this led to difficulties in building the desired basic model. Furthermore, the exact definitions of a number of the variables in the QRISK2 score and the specific methods of missing data handling are not made public and the choice of deprivation measurement (the Townsend score) is restricted to people living in England and Wales. In order to overcome some of these issues, the definition of CKD used was chosen as the most accurate definition for the ET2DS and the SIMD was used as an alternative measure of deprivation (a detailed discussion of these challenges and how they were overcome can be found in section 5.1.7.1 of Chapter 5). It should be noted that the guidelines for reporting prediction models set out by the Enhancing the QUAlity and Transparency Of health Research (EQUATOR) network specify that all predictors should be clearly defined, missing data handling should be described including details of imputation methods and a full prediction model should be presented including all regression coefficients and model intercept or baseline survival function (Collins et al., 2015a). Strict adherence to these guidelines, particularly for risk scores which are recommended in clinical guidelines, would allow future studies to replicate prediction scores more accurately for this type of research.

9.2.5 Comparisons of findings with previous studies

9.2.5.1 NT-proBNP

In this thesis, high levels of NT-proBNP were strongly associated with increased risk of incident or recurrent cardiovascular events, independent of factors currently used to predict CVD, and the biomarker improved predictive performance. This finding is consistent with numerous previous studies in both the general population and diabetic populations.

In 2009, a systematic review and meta-analysis of over 87,000 participants from 40 long-term studies found a strong association between NT-proBNP and incident CVD, independent of conventional risk factors (Di Angelantonio et al., 2009). However, the studies used in this analysis were recruited from the general population, rather

than a subgroup with type 2 diabetes. Also, the choice of adjustment risk factors was not based on a single risk score and differed among the studies, although most included key factors such as age, sex, smoking status, diabetes, blood pressure and lipids. Only 14 studies reported a measure of model discrimination (the c-statistic) after the addition of NT-proBNP to a basic model, with increments in improvement ranging from 0.01 to 0.1. In 2012, a study of nearly 6000 participants from the Rotterdam Study investigated the added value of 12 new biomarkers (including NT-proBNP, CRP and fibrinogen) to the prediction of CHD and found that NT-proBNP improved risk prediction (Kavousi et al., 2012). The increase in the c-statistic compared to the basic model was 0.02. The basic model was based on the Framingham Risk Score, though again this study was carried out in the general population rather than in people with type 2 diabetes. More recent studies in the general population have found similar results, with increases in the c-statistic ranging between 0.017 and 0.03 (Welsh et al., 2016, van der Leeuw et al., 2016). The basic models used in these studies were based on a variety of cardiovascular risk scores: QRISK2, ASSIGN and UKPDS. One contrasting result was found by Welsh et al., 2016, in the Midspan Family Study (MFS) cohort where the addition of NT-proBNP resulted in no improvement to the basic model based on the ASSIGN score. Welsh et al., 2016, hypothesise that this result may be due to a combination of factors: poor statistical power, a low burden of subclinical CVD due to the young age of participants (aged 30 to 59 years at baseline) and a relatively high c-statistic for the basic model (0.752). van der Leeuw et al., 2016, also found that despite the improvement in the c-statistic, the number of patients successfully reclassified to a more appropriate risk category was limited, raising doubts about the clinical usefulness of adding NT-proBNP to current risk scores. Furthermore, participants who did not experience a cardiovascular event during follow-up were better reclassified using the extended model, but those who did experience an outcome event were given poorer reclassification, a result which is also observed in this thesis. This suggests that non-traditional biomarkers such as NT-proBNP may be more useful to assist in screening to rule out high cardiovascular risk.

In 2015, a study of over 8000 people with abnormal blood glucose levels found that, of 237 cardiometabolic biomarkers, the addition of 10 identified biomarkers

including NT-proBNP improved the prediction of a variety of cardiovascular outcomes (Gerstein et al., 2015). The c-statistic for the basic model based on the validated INTERHEART risk score was 0.64, increasing by 0.07 to 0.71 after the addition of the panel of 10 biomarkers, although the contribution of NT-proBNP alone was not investigated. Two studies have investigated the relationship between NT-proBNP and cardiovascular outcomes specifically in populations with type 2 diabetes. A study by Bruno et al., 2013, found that in nearly 2000 people with type 2 diabetes living in a town in northwest Italy, NT-proBNP was a strong predictor of cardiovascular death. Models were adjusted for a range of conventional risk factors, though these were not based on a specific risk score, and no measures of model improvement such as the c-statistic or net reclassification were presented. Finally, a study of nearly 4000 patients with type 2 diabetes from the ADVANCE trial, which investigated the ability of both NT-proBNP and hs-cTNT to improve the prediction of fatal and non-fatal cardiovascular events, found that NT-proBNP was significantly associated with outcome events and improved both classification and discrimination over and above traditional risk factors (Hillis et al., 2014). The c-statistic improved by an increment of 0.04, although again the basic model was not based on a specific risk cardiovascular risk score.

9.2.5.2 hs-cTnT

High levels of hs-cTnT were also strongly associated with increased risk of incident or recurrent CV events and predictive performance was improved, a finding which is consistent with some recent studies, but not all.

The study by Hillis et al., 2014, found that the addition of hs-cTnT improved the c-statistic by an increment of 0.024 compared to a basic model based on traditional risk factors in people with type 2 diabetes. This study also found that the combination of both hs-cTnT and NT-proBNP gave the optimal risk discrimination, suggesting that a combination of a few additional biomarkers may provide the best model improvement compared to individual biomarkers, as observed in the results in this thesis. By contrast, Welsh et al., 2016, found no improvement in risk models after the addition of hs-cTnT in both the MFS and BRHS cohorts, although as noted above

these studies were both carried out in the general population and the MFS cohort is potentially under-powered for this type of analysis.

9.2.5.3 ABI

In this thesis, ABI was negatively associated with cardiovascular risk, although the relationship was weaker than that for NT-proBNP or hs-cTnT. This result is consistent with a number of general population studies which have found modest improvements in risk prediction after the addition of ABI.

In 2010, a study of nearly 7000 participants from the Multi-Ethnic Study of Atherosclerosis (MESA) cohort with no previous history of CVD found that ABI was associated with CVD and significantly improved risk discrimination (Criqui et al., 2010). The basic model was adjusted for both conventional risk factors such as age, sex, smoking, blood pressure and lipids and for newer biomarkers including CRP, IL-6 and fibrinogen, and the increase in the c-statistic following the addition of ABI was 0.01. Subsequently, a meta-analysis including nearly 45,000 participants from 18 different studies suggested that, in the general population, measuring ABI may improve CV risk prediction beyond the Framingham Risk Score (Fowkes et al., 2014). A small increase in the c-statistic (+0.013) was observed for men, while a larger increase (+0.112) was found for women. Finally, a general population study of over 11,000 participants from the Atherosclerosis Risk in Communities Study (ARIC) cohort indicated that ABI has a small effect on cardiovascular risk with an increase of 0.002 in the c-statistic after the addition of ABI to a basic model based on available Framingham Risk Score variables (Murphy et al., 2012). Net reclassification improved, though non-significantly, and the authors note that ABI only improved risk prediction if the basic model was weak.

9.2.5.4 GGT

In the ET2DS, GGT was not significantly associated with outcome cardiovascular events, although the c-statistics did improve incrementally. The conclusions regarding GGT and cardiovascular risk have been mixed in the literature.

In 2005, a study of over 160,000 participants found that GGT was independently associated with cardiovascular mortality over and above conventional risk factors,

although no measures of model improvement were reported (Ruttmann et al., 2005). A recent general population cohort study of over 2500 patients with acute coronary syndrome found that GGT was associated with increased risk of all-cause mortality but not cardiac mortality (Ndrepepa et al., 2016). After the addition of GGT to a basic model including conventional risk factors such as age, sex, diabetes and smoking, the c-statistic improved by 0.002 and 0.007 for the outcomes of cardiac mortality and all-cause mortality respectively. Similarly, the PREVEND prospective cohort study reported that in nearly 7000 participants adding GGT to conventional cardiovascular risk factors did not improve the prediction of first-ever cardiovascular events in the general population (Kunutsor et al., 2015).

A recent experimental model carried out in mice has suggested that statins significantly decrease the expression of GGT in atherosclerotic plaques (Li et al., 2014). This could explain why in this thesis, where approximately 86% of participants were prescribed statins at baseline, GGT was not significantly associated with cardiovascular events when added to the basic model alone. In contrast, GGT was selected as one of a panel of biomarkers which gave the best improvement in prediction, although, as noted previously, more research is required in order to understand the biological mechanism underlying this result.

9.2.5.5 Inflammation factor

Finally, in the ET2DS, the inflammation factor *g* was not significantly associated with outcome cardiovascular events, although the c-statistic did improve marginally. A number of recent cohort studies and meta-analyses have investigated the role of the four inflammatory biomarkers used to create this factor (CRP, IL-6, TNF-α and fibrinogen) in cardiovascular risk prediction. These studies have shown that adding such biomarkers to current cardiovascular risk scores has only a moderate to weak effect in both general and diabetic populations.

In 2010, a study of over 2000 older adults (aged 70-79 at baseline) with no previous CVD found a moderate association between IL-6 and a range of cardiovascular outcomes, but weaker associations for both CRP and TNF-α in the general population (Rodondi et al., 2010). The c-statistic for the basic model based on the Framingham Risk Score was 0.631, increasing by 0.019, 0.007 and 0.016 for IL-6,

CRP and TNF-α respectively. Similar to the study of NT-proBNP by van der Leeuw et al., 2016, patients who did experience a cardiovascular event were given poorer reclassification after the addition of these biomarkers, whereas people who did not experience an event were better classified. van der Leeuw et al., 2016, also explored the role of CRP in cardiovascular risk prediction. However, they found that, although CRP was significantly associated with cardiovascular events, it lacked the ability to improve risk prediction in people with type 2 diabetes. Similarly, Bruno et al., 2013, found that CRP was not an independent predictor of cardiovascular mortality in people with type 2 diabetes.

A meta-analysis carried out in 2012 investigated the association between first cardiovascular events and both CRP and fibrinogen (Emerging Risk Factors et al., 2014). In over 240,000 participants without previous CVD from 52 prospective studies, the c-statistic was found to increase by 0.0039 and 0.0027 after the addition of CRP and fibrinogen respectively to a basic model including predictors commonly used in standard risk scores. The study by Kavousi et al., 2012, discussed above in relation to NT-proBNP, also investigated CRP and fibrinogen and found that the improvement in the c-statistic following the addition of CRP was not significant (95% CI for change in c-statistic: -0.01, 0.00) and was only marginal following the addition of fibrinogen (CI: 0.00, 0.01).

Finally, a study of over 5500 patients with atrial fibrillation found that, after adjustment for clinical risk factors plus non-traditional biomarkers (including NT-proBNP and troponin), IL-6 was related to vascular mortality and CRP was associated with MI, but fibrinogen was not related to any cardiovascular outcomes (Aulin et al., 2015). IL-6 was found to be the strongest biomarker from this group of three inflammatory biomarkers, with an increase in the c-statistic of 0.067.

9.3 Associations between metabolomics data and CVD

9.3.1 Strengths of the UCLEB consortium studies

9.3.1.1 Sample size

Using combined data from the five UCLEB cohorts resulted in enhanced statistical power to test for associations across a large number of potential predictors, while

also adjusting for a range of important cardiovascular risk factors. The number of participants with type 2 diabetes available for analysis (*n*=2247) was over double the number that would have been available using just the ET2DS cohort alone (*n*=1058). The number of CVD outcomes was also large (*n*=1005), again over double the number of CVD outcomes in the ET2DS cohort (*n*=451). In terms of events per variable, this number was more than sufficient (full model with 12 risk factors plus one metabolite = minimum 130 events desirable). Multiple testing corrections were implemented in order to account for the number of tests performed.

9.3.1.2 Risk factors available

A wide range of baseline risk factors were available in all the UCLEB studies, including key cardiovascular risk factors such as smoking status, blood pressure, lipids, ethnicity and social status. This allowed for a final model which was adjusted for an extensive group of predictors which are commonly used in current cardiovascular risk scores.

9.3.1.3 Analysis plan

The exploratory association analysis carried out in this section of the thesis considered the full NMR metabolomics panel of 228 metabolites. Although initially I planned to undertake a prediction analysis using these data, limitations of the data (including numbers of cardiovascular events, inconsistent discrimination between prevalent, recurrent and incident events between studies and differing definitions used for key risk factors, as discussed below), meant that this was not ultimately possible. The analysis undertaken does however generate hypotheses for future work in this area and is a novel contribution to previous and on-going analyses based on metabolomics data, many of which have focused on one particular group of metabolites, or indeed a single metabolite.

9.3.2 Limitations of the UCLEB consortium data

9.3.2.1 Definition of outcome

The definition of the outcome for this analysis was given as CVD diagnosed at baseline or developed during the study follow-up period, in order to investigate the cross-sectional associations between metabolomics data and CVD. However, for the study of risk prediction, it is of course desirable to have an outcome based only on

incident cardiovascular events which are experienced during follow-up. It was not possible to carry out such an analysis in the UCLEB cohorts, since full event information was not available for all the studies – recurrent cardiovascular events were not shared for any study except ET2DS, so the outcome would have had to be restricted to truly incident events and people with prevalent CVD excluded. This would have drastically reduced the statistical power to investigate any associations.

9.3.2.2 Risk factor definitions and availability of variables

As discussed in the previous sections, a wide range of cardiovascular risk factors were available in the UCLEB cohorts used in this thesis. However, a measure of kidney function (eGFR) was not available in two of the cohorts: BRHS and WHII. Since it is well-established that impaired kidney function is an important risk factor for CVD (Di Angelantonio et al., 2010, Gansevoort et al., 2013) and as many cardiovascular risk scores designed specifically for people with type 2 diabetes include a measure of kidney function as a predictor (Folsom et al., 2003, Yang et al., 2008b, Hippisley-Cox et al., 2010), it was considered an important variable to include in the final model (Model 4). This resulted in the two cohorts, BRHS and WHII, for whom this variable was not available, being dropped from Model 4. Although these were not the largest contributing cohorts in terms of sample size, it is likely that this had an impact on the statistical power for the final stage of analysis.

Finally, the majority of the definitions for the available variables were harmonised across all UCLEB cohorts, with the exception of the social status variable. As discussed in detail in section 5.2.7 of Chapter 5, the social status variable was described according to a set of six occupation categories defined by UCLEB in three of the studies (BRHS, BWHHS and SABRE), five occupation categories in the ET2DS and three categories in WHII. Furthermore, these categories did not have particularly similar definitions between studies. In order to include social status in Model 4, social status was collapsed into three general groups (unskilled, skilled and professional) and individual categories in each study were matched as closely as possible with these groups. It is accepted that this simplifies the information captured by this variable and that assumptions have been made about the equivalence of some

groups. Ideally a uniformly defined variable for all UCLEB studies, with categories chosen in advance, would be used in this type of analysis.

9.3.3 Comparisons of findings with previous studies

In five studies from the UCLEB consortium, HDL particles were found to have the strongest association with CVD from a panel of 228 metabolites. This is consistent with previous studies which have suggested that novel HDL biomarkers may have strong relationships with CVD. HDL is one of five major groups of lipoproteins, complex particles consisting of both lipid and protein components. It is known to be atheroprotective and is non-uniform in structure, composition and function (Wurtz et al., 2011). Measures of HDL quality, for example particle size, subclass distribution and functionality can be obtained in addition to commonly measured components such as cholesterol (Camont et al., 2013). Furthermore, it has been proposed that HDL cholesterol may not be the best clinical summary of HDL (Würtz et al., 2012).

In particular, in this thesis, four HDL particles were found to be significantly associated with CVD even after adjustment for a wide range of cardiovascular risk factors commonly used in current risk scores including lipids and at a stringent threshold for multiple comparisons (the Bonferroni correction): concentration of medium HDL particles, total lipids in medium HDL and phospholipids in medium and small HDL. This finding requires further replication and investigation in order to understand the underlying biological mechanisms, although recent papers have suggested that HDL cholesterol is not the strongest component of HDL. In the HDL lipidome, it is phospholipids that are the strongest element, constituting between 40-60% of the total lipid weight (Kontush et al., 2013). This may explain why the phospholipids remained strongly associated with CVD despite adjustment for HDL cholesterol in both Models 3 and 4. Although further work is required in order to establish whether novel HDL biomarkers are superior to serum HDL cholesterol for risk prediction, a few recent papers indicate that this may well be the case (Rader and Hovingh, 2014, Mora et al., 2013, Würtz et al., 2012). In 2012, a study of 1595 participants from the Cardiovascular Risk in Young Finns Study found that, of 56 NMR metabolites, four improved the prediction of subclinical atherosclerosis when replacing total and HDL cholesterol in the Framingham Risk Score (Würtz et al.,

2012). These four biomarkers included NMR-determined medium HDL, although it was noted that no single metabolite improved risk discrimination alone and that the young age of the participants (aged 24-39 years at baseline) prevented the investigation of hard cardiovascular outcomes. Mora et al., 2013, found that in over 10,000 participants from the JUPITER trial, HDL particle number had a significant association with CVD which was stronger than that for HDL cholesterol, although it should be noted that criticisms have been made regarding the collection of cardiovascular event data and the short follow-up time (approximately 2 years) in this trial (de Lorgeril et al., 2010). Previous studies suggest that HDL particle number is less influenced by complex issues such as insulin resistance, abdominal obesity and inflammation, which strongly correlate with HDL cholesterol (Vergeer et al., 2010, Mackey et al., 2012), which suggests that HDL particle number may be a more useful biomarker for people with diabetes.

Finally, the NMR technique has been shown to provide more accurate measurements of commonly used biomarkers such as total cholesterol (Würtz et al., 2012), which makes this platform a promising tool for future risk prediction measurements.

9.4 Risk prediction methods

9.4.1 Impact of the choice of statistical method

The choice of statistical methods has an influence on the results presented in this thesis and this should be taken into account during the interpretation. In Chapter 7, the net reclassification was calculated according to pre-specified cardiovascular risk categories: 0-10%, 10-20% and >20% risk. These thresholds were selected based on the current and previous guidelines for statin prescription. Currently, it is recommended that patients with a 10% or greater 10-year risk of developing CVD should be offered statins (NICE CG181, 2016), although until 2014 this threshold was 20% (Rabar et al., 2014). It is possible that this could change again in the future, since concerns remain regarding issues such as unwanted side effects and the treatment adherence by patients who consider themselves healthy (Majeed, 2014). Risk prediction research should aim to reflect the current clinical practice in order for results to be clinically relevant, although this may require updating of reclassification measures should clinical guidelines change. Furthermore, recently Welsh et al., 2016,

noted that the clinical usefulness of new biomarkers added to current risk scores depends on the chosen risk thresholds and that this is particularly important in light of changing clinical guidelines since a reduction in risk thresholds leads to a decrease in specificity (the probability of predicting "no event" among those patients who do not experience the outcome of interest).

The choice of multiple correction adjustment method has a substantial impact on the interpretation of the type of analysis carried out in Chapter 8. Since this section of the thesis was an exploratory analysis, the choice of just one threshold for adjusted statistical significance was concerning as this would have split the metabolites into two distinct groups: "significant" and "not-significant". At this early stage of metabolomics analysis, this was not considered desirable, therefore three thresholds have been presented, ranging from a stringent control (the Bonferroni correction) to a more lenient control (a 5% FDR). Although four metabolites in particular have been discussed above (those that remained significant at the strictest adjustment threshold and after full risk factor adjustment) I would recommend that the full panel of metabolites is retained for further investigation.

Finally, in Chapter 7, an all subsets regression was carried out in order to select the top models from all possible combinations of additional biomarkers. The AIC was used as the pre-specified statistic to identify the top models, although, as discussed in Chapter 2, the choice of statistic can result in different "best" models (Steyerberg, 2009). For this reason, a number of top models were investigated further using more detailed model evaluation methods such as the c-statistic and net reclassification.

9.4.2 Model evaluation measures

Recently, the choice of model evaluation measures in risk prediction modelling has been debated (see Chapter 2). There is no clear consensus as to which measure, or measures, should be reported for new risk prediction models. Therefore, the decision was made to present a range of measures in this thesis, while reflecting on the various advantages and disadvantages of these methods. Modest improvements were observed for the c-statistic in Chapter 7, despite strong associations between some of the biomarkers and cardiovascular events even after adjustment for a current

cardiovascular risk score, in particular for hs-cTnT. This phenomenon of the insensitivity of the c-statistic to new biomarkers has been well-established in the literature and is particularly noticeable when the basic model includes strong predictors (Cook, 2008). Overall though, the c-statistic is still considered to provide useful information and is popular in publications, almost certainly because it is well-recognised and understood.

In order to evaluate the clinical usefulness of a new prediction model, more recently it has been suggested that a measure of reclassification should be reported in addition to the c-statistic. The NRI (or IDI) is a popular measure in recent publications, however this measure is difficult to interpret and may not be very clinically informative (Kerr et al., 2014). Instead of the combined NRI, two measures of net reclassification which are used to calculated the NRI have been reported in this thesis: the proportion of people given more accurate risk classification after the addition of one or more biomarkers for people who did not experience a cardiovascular event and the proportion of people given more accurate risk classification among those who did experience an event. These measures inform us of the number of people who would be given better or worse risk classification using the new model, which has clear implications for clinical practice. However this decision also made the results in this thesis difficult to compare with previous studies unless full reclassification tables had been reported, allowing for equivalent measures to be calculated. I would recommend that either the two measures of net reclassification or the full reclassification tables are reported for future studies on this subject.

9.4.3 Development of software for complex methods

Lastly, the development of software for complex statistical methods had an impact on the methods used in this thesis, and therefore the final results. For example, the use of multiple imputation methods for missing data may have been preferable in the analysis of the metabolomics data. However, as discussed in Chapter 8, the implementation of the required methods for this context of combined data (multi-level multiple imputations) is still under development and, if used incorrectly, such methods can bias the final results. For the analysis of metabolomics data, it is noted

in recent literature that technical issues such as statistical analysis still need to be overcome in order to fully understand the relationships between these biomarkers and CVD (Kontush et al., 2013).

9.5 Recommendations for future research

To conclude, my recommendations for future research on the topics studied in this thesis are as follows:

9.5.1 Improving cardiovascular risk prediction using individual and combined biomarkers

1. This thesis found that three non-traditional biomarkers (ABI, NT-proBNP and hs-cTnT) added value to a basic model based on a current cardiovascular risk score when incorporated individually, and that a combination of three biomarkers (ABI, hs-cTnT and GGT) added to such a model provided the best improvement. External validation of these results in larger studies is required before any of these promising biomarkers could be added to current risk scores. In order to expand the generalisability of future results, such cohorts should include younger adults, a wider variety of ethnicities and, if enough cardiovascular events are captured, restrict analyses to people with no previous CVD.
2. Further large studies are also required in order to establish the clinical significance of these results and the cost effectiveness of incorporating new biomarkers into cardiovascular risk scores.
3. It would also be interesting to investigate the associations and added predictive value of these biomarkers in subgroups of the CVD outcome, such as CHD (angina and MI) or cerebrovascular disease (stroke and TIA), since there may be differences between these types of CVD. This will require a much greater number of events for adequate statistical power than are available in the ET2DS.
4. The findings in this thesis highlight the need for thorough and consistent reporting of clinical risk prediction models and their evaluation. Although practice has improved over the last few decades and recent guidelines have been set for the publication of such scores (Moons et al., 2009, Collins et al., 2015b), these are not always followed. Adherence to such rules is vital for external

validation and the investigation of new biomarkers, particularly for risk scores which are used in clinical settings. Consistent use of model evaluation measures would also allow for more direct comparisons among studies.

9.5.2 Associations between metabolomics data and CVD

1. The research presented in this thesis found that certain metabolites, in particular from the HDL particle subclass, are strongly associated with CVD even after adjustment for risk factors used in cardiovascular risk scores and at a strict adjustment of statistical significance. However, the univariate method used in this thesis does not take into account the correlations between the metabolites, which can be strong, particularly among the same subclasses. Therefore, I would propose that the next step in this analysis should use a variable selection method such as LASSO regression to select a panel of metabolites which most strongly associate with CVD, while accounting for correlations between them. I would recommend LASSO regression since it is known to return a relatively large panel of predictors compared to other variable selection methods and this research is still in its preliminary stages.
2. This analysis then needs to be validated in larger studies, for example using cohorts in the COnsortium of METabolomics Studies (COMETS) of which UCLEB is a constituent member (COMETS, 2015). This will provide a much greater number of cardiovascular events and allow analysis to be restricted to incident events, which is preferable for risk prediction. A greater number of outcomes might also allow for the study of people with no previous CVD, although this will depend on statistical power.
3. Promising metabolites which are discovered through the previous suggested steps can then be investigated following a similar thorough approach as the one taken for the non-traditional biomarkers in Chapter 7 of this thesis, adding such metabolites to a current cardiovascular risk score and comparing them to other new biomarkers.
4. Since NMR spectroscopy is potentially more accurate at quantifying lipids than traditional lipid measurement, and is relatively cheap to carry out, it would also be worth exploring whether any metabolite on its own can produce the same, or even improved, levels of risk prediction as current scores.

5. Finally, once promising metabolites have been fully investigated it would be of interest to determine why they are predictive of CVD and whether the relationships can inform us about the underlying biological mechanisms of CVD.

10 Bibliography

AKAIKE, H. 2011. Akaike's Information Criterion. *In:* LOVRIC, M. (ed.) *International Encyclopedia of Statistical Science.* Berlin, Heidelberg: Springer Berlin Heidelberg.

ALCOCER, F., MUJIB, M., LOWMAN, B., PATTERSON, M. A., PASSMAN, M. A., MATTHEWS, T. C. & JORDAN, W. D. 2013. Risk scoring system to predict 3-year survival in patients treated for asymptomatic carotid stenosis. *Journal of Vascular Surgery,* 57**,** 1576-80.

ALLISON, M. A., HIATT, W. R., HIRSCH, A. T., COLL, J. R. & CRIQUI, M. H. 2008. A High Ankle-Brachial Index Is Associated With Increased Cardiovascular Disease Morbidity and Lower Quality of Life. *Journal of the American College of Cardiology,* 51**,** 1292-1298.

ALMAN, A. C., KINNEY, G. L., TRACY, R. P., MAAHS, D. M., HOKANSON, J. E., REWERS, M. J. & SNELL-BERGEON, J. K. 2013. Prospective Association Between Inflammatory Markers and Progression of Coronary Artery Calcification in Adults With and Without Type 1 Diabetes. *Diabetes Care,* 36**,** 1967-1973.

ALONSO, A., KRIJTHE, B. P., ASPELUND, T., STEPAS, K. A., PENCINA, M. J., MOSER, C. B., SINNER, M. F., SOTOODEHNIA, N., FONTES, J. D., JANSSENS, A. C., KRONMAL, R. A., MAGNANI, J. W., WITTEMAN, J. C., CHAMBERLAIN, A. M., LUBITZ, S. A., SCHNABEL, R. B., AGARWAL, S. K., MCMANUS, D. D., ELLINOR, P. T., LARSON, M. G., BURKE, G. L., LAUNER, L. J., HOFMAN, A., LEVY, D., GOTTDIENER, J. S., KAAB, S., COUPER, D., HARRIS, T. B., SOLIMAN, E. Z., STRICKER, B. H., GUDNASON, V., HECKBERT, S. R. & BENJAMIN, E. J. 2013. Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: the CHARGE-AF consortium. *Journal of the American Heart Association,* 2**,** e000102.

ALSSEMA, M., NEWSON, R. S., BAKKER, S. J., STEHOUWER, C. D., HEYMANS, M. W., NIJPELS, G., HILLEGE, H. L., HOFMAN, A., WITTEMAN, J. C., GANSEVOORT, R. T. & DEKKER, J. M. 2012. One risk assessment tool for cardiovascular disease, type 2 diabetes, and chronic kidney disease. *Diabetes Care,* 35**,** 741-8.

ALTMAN, D. G., VERGOUWE, Y., ROYSTON, P. & MOONS, K. G. 2009. Prognosis and prognostic research: validating a prognostic model. *BMJ,* 338**,** b605.

ANDERSEN, K. K. & OLSEN, T. S. 2011. One-month to 10-year survival in the Copenhagen stroke study: interactions between stroke severity and other prognostic indicators. *Journal of Stroke & Cerebrovascular Diseases,* 20**,** 117-23.

ANDERSON, K. M., ODELL, P. M., WILSON, P. W. & KANNEL, W. B. 1991a. Cardiovascular disease risk profiles. *Am Heart J,* 121**,** 293-8.

ANDERSON, K. M., WILSON, P. W., ODELL, P. M. & KANNEL, W. B. 1991b. An updated coronary risk profile. A statement for health professionals. *Circulation,* 83**,** 356-62.

ARIMA, H., YONEMOTO, K., DOI, Y., NINOMIYA, T., HATA, J., TANIZAKI, Y., FUKUHARA, M., MATSUMURA, K., IIDA, M. & KIYOHARA, Y.

2009. Development and validation of a cardiovascular risk prediction model for Japanese: the Hisayama study. *Hypertens Res,* 32**,** 1119-22.
ASIA PACIFIC COHORT STUDIES COLLABORATION 2006. Coronary risk prediction for those with and without diabetes. *Eur J Cardiovasc Prev Rehabil,* 13**,** 30-6.
ASSMANN, G., CULLEN, P. & SCHULTE, H. 2002. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Munster (PROCAM) study. *Circulation,* 105**,** 310-5.
ASSMANN, G., SCHULTE, H., CULLEN, P. & SEEDORF, U. 2007. Assessing risk of myocardial infarction and stroke: new data from the Prospective Cardiovascular Munster (PROCAM) study. *Eur J Clin Invest,* 37**,** 925-32.
AUDIGIER, V., WHITE, I. R., JOLANI, S., DEBRAY, T. P. A., QUARTAGNO, M., CARPENTER, J., VAN BUUREN, S. & RESCHE-RIGON, M. 2017. Multiple imputation for multilevel data with continuous and binary variables. *arXiv.org*.
AULIN, J., SIEGBAHN, A., HIJAZI, Z., EZEKOWITZ, M. D., ANDERSSON, U., CONNOLLY, S. J., HUBER, K., REILLY, P. A., WALLENTIN, L. & OLDGREN, J. 2015. Interleukin-6 and C-reactive protein and risk for death and cardiovascular events in patients with atrial fibrillation. *American Heart Journal,* 170**,** 1151-1160.
AXENTE, L., SINESCU, C. & BAZACLIU, G. 2011. Heart failure prognostic model. *Journal of Medicine & Life,* 4**,** 210-25.
BALKAU, B., HU, G., QIAO, Q., TUOMILEHTO, J., BORCH-JOHNSEN, K. & PYORALA, K. 2004. Prediction of the risk of cardiovascular mortality using a score that includes glucose as a risk factor. The DECODE Study. *Diabetologia,* 47**,** 2118-28.
BANNISTER, C. A., POOLE, C. D., JENKINS-JONES, S., MORGAN, C. L., ELWYN, G., SPASIC, I. & CURRIE, C. J. 2014. External validation of the UKPDS risk engine in incident type 2 diabetes: a need for new type 2 diabetes-specific risk equations. *Diabetes Care,* 37**,** 537-45.
BARALDI, A. N. & ENDERS, C. K. 2010. An introduction to modern missing data analyses. *J Sch Psychol,* 48**,** 5-37.
BASTIEN, M., POIRIER, P., LEMIEUX, I. & DESPRÉS, J.-P. 2014. Overview of Epidemiology and Contribution of Obesity to Cardiovascular Disease. *Progress in Cardiovascular Diseases,* 56**,** 369-381.
BEDENIS, R., PRICE, A. H., ROBERTSON, C. M., MORLING, J. R., FRIER, B. M., STRACHAN, M. W. & PRICE, J. F. 2014. Association Between Severe Hypoglycemia, Adverse Macrovascular Events, and Inflammation in the Edinburgh Type 2 Diabetes Study. *Diabetes Care*.
BENJAMINI, Y. & HOCHBERG, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological),* 57**,** 289-300.
BERG, A. H. & SCHERER, P. E. 2005. Adipose tissue, inflammation, and cardiovascular disease. *Circ Res,* 96**,** 939-49.
BERLIN, J. A. & COLDITZ, G. A. 1990. A meta-analysis of physical activity in the prevention of coronary heart disease. *Am J Epidemiol,* 132**,** 612-28.

BETHEL, M. A., CHACRA, A. R., DEEDWANIA, P., FULCHER, G. R., HOLMAN, R. R., JENSSEN, T., KAHN, S. E., LEVITT, N. S., MCMURRAY, J. J., CALIFF, R. M., RAPTIS, S. A., THOMAS, L., SUN, J. L. & HAFFNER, S. M. 2013. A novel risk classification paradigm for patients with impaired glucose tolerance and high cardiovascular risk. *American Journal of Cardiology,* 112**,** 231-7.

BETTENCOURT, N., OLIVEIRA, S., TOSCHKE, A. M., ROCHA, J., LEITE, D., CARVALHO, M., XARA, S., SCHUSTER, A., CHIRIBIRI, A., LEITE-MOREIRA, A., NAGEL, E., ALVES, H. & GAMA, V. 2011. Predictors of circulating endothelial progenitor cell levels in patients without known coronary artery disease referred for multidetector computed tomography coronary angiography. *Revista Portuguesa de Cardiologia,* 30**,** 753-60.

BLAND, J. M. & ALTMAN, D. G. 1995. Multiple significance tests: the Bonferroni method. *BMJ,* 310**,** 170.

BRUNO, G., LANDI, A., BARUTTA, F., GHEZZO, G., BALDIN, C., SPADAFORA, L., SCHIMMENTI, A., PRINZIS, T., CAVALLO PERIN, P. & GRUDEN, G. 2013. N-terminal probrain natriuretic peptide is a stronger predictor of cardiovascular mortality than C-reactive protein and albumin excretion rate in elderly patients with type 2 diabetes: the Casale Monferrato population-based study. *Diabetes Care,* 36**,** 2677-82.

BURJONROPPA, S. C., VAROSY, P. D., RAO, S. V., OU, F. S., ROE, M., PETERSON, E., SINGH, M. & SHUNK, K. A. 2011. Survival of patients undergoing rescue percutaneous coronary intervention: development and validation of a predictive tool. *Jacc: Cardiovascular Interventions,* 4**,** 42-50.

CAMERON, A. C. & WINDMEIJER, F. A. G. 1997. An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics,* 77**,** 329-342.

CAMONT, L., LHOMME, M., RACHED, F., LE GOFF, W., NEGRE-SALVAYRE, A., SALVAYRE, R., CALZADA, C., LAGARDE, M., CHAPMAN, M. J. & KONTUSH, A. 2013. Small, dense high-density lipoprotein-3 particles are enriched in negatively charged phospholipids: relevance to cellular cholesterol efflux, antioxidative, antithrombotic, anti-inflammatory, and antiapoptotic functionalities. *Arterioscler Thromb Vasc Biol,* 33**,** 2715-23.

CARROLL, S. J., PAQUET, C., HOWARD, N. J., ADAMS, R. J., TAYLOR, A. W. & DANIEL, M. 2014. Validation of continuous clinical indices of cardiometabolic risk in a cohort of Australian adults. *BMC Cardiovascular Disorders,* 14**,** 27.

CASTELLI, W. P. 1984. Epidemiology of coronary heart disease: the Framingham study. *Am J Med,* 76**,** 4-12.

CEDERHOLM, J., EEG-OLOFSSON, K., ELIASSON, B., ZETHELIUS, B., GUDBJORNSDOTTIR, S. & SWEDISH NATIONAL DIABETES, R. 2011. A new model for 5-year risk of cardiovascular disease in Type 1 diabetes; from the Swedish National Diabetes Register (NDR). *Diabetic Medicine,* 28**,** 1213-20.

CEDERHOLM, J., EEG-OLOFSSON, K., ELIASSON, B., ZETHELIUS, B., NILSSON, P. M. & GUDBJORNSDOTTIR, S. 2008. Risk prediction of cardiovascular disease in type 2 diabetes: a risk equation from the Swedish National Diabetes Register. *Diabetes Care,* 31**,** 2038-43.

CHADEAU-HYAM, M., CAMPANELLA, G., JOMBART, T., BOTTOLO, L., PORTENGEN, L., VINEIS, P., LIQUET, B. & VERMEULEN, R. C. 2013. Deciphering the complex: methodological overview of statistical models to derive OMICS-based biomarkers. *Environ Mol Mutagen,* 54**,** 542-57.

CHAHAL, H., BACKLUND, J. Y., CLEARY, P. A., LACHIN, J. M., POLAK, J. F., LIMA, J. A., BLUEMKE, D. A. & GROUP, D. E. R. 2012. Relation between carotid intima-media thickness and left ventricular mass in type 1 diabetes mellitus (from the Epidemiology of Diabetes Interventions and Complications [EDIC] Study). *American Journal of Cardiology,* 110**,** 1534-40.

CHAHAL, H., BLUEMKE, D. A., WU, C. O., MCCLELLAND, R., LIU, K., SHEA, S. J., BURKE, G., BALFOUR, P., HERRINGTON, D., SHI, P., POST, W., OLSON, J., WATSON, K. E., FOLSOM, A. R. & LIMA, J. A. 2015. Heart failure risk prediction in the Multi-Ethnic Study of Atherosclerosis. *Heart,* 101**,** 58-64.

CHIEN, K. L., SU, T. C., HSU, H. C., CHANG, W. T., CHEN, P. C., SUNG, F. C., CHEN, M. F. & LEE, Y. T. 2010. Constructing the prediction model for the risk of stroke in a Chinese population: report from a cohort study in Taiwan. *Stroke,* 41**,** 1858-64.

COLLINS, G. S., REITSMA, J. B., ALTMAN, D. G. & MOONS, K. G. 2015a. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Bmj,* 350**,** g7594.

COLLINS, G. S., REITSMA, J. B., ALTMAN, D. G. & MOONS, K. G. M. 2015b. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD StatementThe TRIPOD Statement. *Annals of Internal Medicine,* 162**,** 55-63.

COMETS. 2015. *COnsortium of METabolomics Studies (COMETS)* [Online]. Available: https://epi.grants.cancer.gov/comets/ [Accessed 2/2/17.

CONROY, R. M., PYORALA, K., FITZGERALD, A. P., SANS, S., MENOTTI, A., DE BACKER, G., DE BACQUER, D., DUCIMETIERE, P., JOUSILAHTI, P., KEIL, U., NJOLSTAD, I., OGANOV, R. G., THOMSEN, T., TUNSTALL-PEDOE, H., TVERDAL, A., WEDEL, H., WHINCUP, P., WILHELMSEN, L. & GRAHAM, I. M. 2003. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *Eur Heart J,* 24**,** 987-1003.

COOK, N. R. 2007. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation,* 115**,** 928-35.

COOK, N. R. 2008. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clin Chem,* 54**,** 17-23.

COUTINHO STORTI, F., MOFFA, P. J., UCHIDA, A. H., HUEB, W. A., MACHADO CESAR, L. A., FERREIRA, B. M., CAMARGO, P. A., JR. & CHALELA, W. A. 2011. New prognostic score for stable coronary disease evaluation. *Arquivos Brasileiros de Cardiologia,* 96**,** 411-8.

CRIQUI, M. H., MCCLELLAND, R. L., MCDERMOTT, M. M., ALLISON, M. A., BLUMENTHAL, R. S., ABOYANS, V., IX, J. H., BURKE, G. L., LIU, K. & SHEA, S. 2010. The ankle-brachial index and incident cardiovascular events in the MESA (Multi-Ethnic Study of Atherosclerosis). *J Am Coll Cardiol,* 56**,** 1506-12.

CROSS, D. S., MCCARTY, C. A., HYTOPOULOS, E., BEGGS, M., NOLAN, N., HARRINGTON, D. S., HASTIE, T., TIBSHIRANI, R., TRACY, R. P., PSATY, B. M., MCCLELLAND, R., TSAO, P. S. & QUERTERMOUS, T. 2012. Coronary risk assessment among intermediate risk patients using a clinical and biomarker based algorithm developed and validated in two population cohorts. *Current Medical Research & Opinion,* 28**,** 1819-30.

CUBBON, R. M., WOOLSTON, A., ADAMS, B., GALE, C. P., GILTHORPE, M. S., BAXTER, P. D., KEARNEY, L. C., MERCER, B., RAJWANI, A., BATIN, P. D., KAHN, M., SAPSFORD, R. J., WITTE, K. K. & KEARNEY, M. T. 2014. Prospective development and validation of a model to predict heart failure hospitalisation. *Heart,* 100**,** 923-9.

CUI, J., FORBES, A., KIRBY, A., MARSCHNER, I., SIMES, J., HUNT, D., WEST, M. & TONKIN, A. 2010. Semi-parametric risk prediction models for recurrent cardiovascular events in the LIPID study. *BMC Medical Research Methodology,* 10**,** 27.

D'AGOSTINO, R. B., SR., VASAN, R. S., PENCINA, M. J., WOLF, P. A., COBAIN, M., MASSARO, J. M. & KANNEL, W. B. 2008. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation,* 117**,** 743-53.

DAVIS, W. A., KNUIMAN, M. W. & DAVIS, T. M. 2010. An Australian cardiovascular risk equation for type 2 diabetes: the Fremantle Diabetes Study. *Intern Med J,* 40**,** 286-92.

DE LORGERIL, M., SALEN, P., ABRAMSON, J., DODIN, S., HAMAZAKI, T., KOSTUCKI, W., OKUYAMA, H., PAVY, B. & RABAEUS, M. 2010. Cholesterol lowering, cardiovascular diseases, and the rosuvastatin-JUPITER controversy: a critical reappraisal. *Arch Intern Med,* 170**,** 1032-6.

DI ANGELANTONIO, E., CHOWDHURY, R., SARWAR, N., ASPELUND, T., DANESH, J. & GUDNASON, V. 2010. Chronic kidney disease and risk of major cardiovascular disease and non-vascular mortality: prospective population based cohort study. *Bmj,* 341**,** c4986.

DI ANGELANTONIO, E., CHOWDHURY, R., SARWAR, N., RAY, K. K., GOBIN, R., SALEHEEN, D., THOMPSON, A., GUDNASON, V., SATTAR, N. & DANESH, J. 2009. B-type natriuretic peptides and cardiovascular risk: systematic review and meta-analysis of 40 prospective studies. *Circulation,* 120**,** 2177-87.

DI GIROLAMO, F., LANTE, I., MURACA, M. & PUTIGNANI, L. 2013. The Role of Mass Spectrometry in the "Omics" Era. *Current Organic Chemistry,* 17**,** 2891-2905.

DOLL, R., PETO, R., BOREHAM, J. & SUTHERLAND, I. 2004. Mortality in relation to smoking: 50 years' observations on male British doctors. *BMJ,* 328**,** 1519.

DONNAN, P. T., DONNELLY, L., NEW, J. P. & MORRIS, A. D. 2006. Derivation and validation of a prediction score for major coronary heart disease events in a U.K. type 2 diabetic population. *Diabetes Care,* 29**,** 1231-6.

DUVAL, S., MASSARO, J. M., JAFF, M. R., BODEN, W. E., ALBERTS, M. J., CALIFF, R. M., EAGLE, K. A., D'AGOSTINO, R. B., SR., PEDLEY, A., FONAROW, G. C., MURABITO, J. M., STEG, P. G., BHATT, D. L., HIRSCH, A. T. & INVESTIGATORS, R. R. 2012. An evidence-based score

to detect prevalent peripheral artery disease (PAD). *Vascular Medicine,* 17**,** 342-51.

ELLEY, C. R., ROBINSON, E., KENEALY, T., BRAMLEY, D. & DRURY, P. L. 2010. Derivation and validation of a new cardiovascular risk score for people with type 2 diabetes: the new zealand diabetes cohort study. *Diabetes Care,* 33**,** 1347-52.

EMERGING RISK FACTORS, C., DI ANGELANTONIO, E., GAO, P., KHAN, H., BUTTERWORTH, A. S., WORMSER, D., KAPTOGE, S., KONDAPALLY SESHASAI, S. R., THOMPSON, A., SARWAR, N., WILLEIT, P., RIDKER, P. M., BARR, E. L., KHAW, K. T., PSATY, B. M., BRENNER, H., BALKAU, B., DEKKER, J. M., LAWLOR, D. A., DAIMON, M., WILLEIT, J., NJOLSTAD, I., NISSINEN, A., BRUNNER, E. J., KULLER, L. H., PRICE, J. F., SUNDSTROM, J., KNUIMAN, M. W., FESKENS, E. J., VERSCHUREN, W. M., WALD, N., BAKKER, S. J., WHINCUP, P. H., FORD, I., GOLDBOURT, U., GOMEZ-DE-LA-CAMARA, A., GALLACHER, J., SIMONS, L. A., ROSENGREN, A., SUTHERLAND, S. E., BJORKELUND, C., BLAZER, D. G., WASSERTHEIL-SMOLLER, S., ONAT, A., MARIN IBANEZ, A., CASIGLIA, E., JUKEMA, J. W., SIMPSON, L. M., GIAMPAOLI, S., NORDESTGAARD, B. G., SELMER, R., WENNBERG, P., KAUHANEN, J., SALONEN, J. T., DANKNER, R., BARRETT-CONNOR, E., KAVOUSI, M., GUDNASON, V., EVANS, D., WALLACE, R. B., CUSHMAN, M., D'AGOSTINO, R. B., SR., UMANS, J. G., KIYOHARA, Y., NAKAGAWA, H., SATO, S., GILLUM, R. F., FOLSOM, A. R., VAN DER SCHOUW, Y. T., MOONS, K. G., GRIFFIN, S. J., SATTAR, N., WAREHAM, N. J., SELVIN, E., THOMPSON, S. G. & DANESH, J. 2014. Glycated hemoglobin measurement and prediction of cardiovascular disease. *JAMA,* 311**,** 1225-33.

EVERETT, B. M., COOK, N. R., MAGNONE, M. C., BOBADILLA, M., KIM, E., RIFAI, N., RIDKER, P. M. & PRADHAN, A. D. 2011. Sensitive cardiac troponin T assay and the risk of incident cardiovascular disease in women with and without diabetes mellitus: the Women's Health Study. *Circulation,* 123**,** 2811-8.

EVERITT, B. S., DUNN, G., EVERITT, B. S. & DUNN, G. 2013. Principal Components Analysis. *Applied Multivariate Data Analysis.* John Wiley & Sons, Ltd,.

FEINKOHL, I., KELLER, M., ROBERTSON, C. M., MORLING, J. R., MCLACHLAN, S., FRIER, B. M., DEARY, I. J., STRACHAN, M. W. J. & PRICE, J. F. 2015. Cardiovascular risk factors and cognitive decline in older people with type 2 diabetes. *Diabetologia,* 58**,** 1637-1645.

FERKET, B. S., VAN KEMPEN, B. J., HEERINGA, J., SPRONK, S., FLEISCHMANN, K. E., NIJHUIS, R. L., HOFMAN, A., STEYERBERG, E. W. & HUNINK, M. G. 2012. Personalized prediction of lifetime benefits with statin therapy for asymptomatic individuals: a modeling study. *PLoS Medicine / Public Library of Science,* 9**,** e1001361.

FERRARIO, M., CHIODINI, P., CHAMBLESS, L. E., CESANA, G., VANUZZO, D., PANICO, S., SEGA, R., PILOTTO, L., PALMIERI, L. & GIAMPAOLI, S. 2005. Prediction of coronary events in a low incidence population.

Assessing accuracy of the CUORE Cohort Study prediction equation. *Int J Epidemiol,* 34**,** 413-21.

FERRIE, J. E., SHIPLEY, M. J., DAVEY SMITH, G., STANSFELD, S. A. & MARMOT, M. G. 2002. Change in health inequalities among British civil servants: the Whitehall II study. *J Epidemiol Community Health,* 56**,** 922-6.

FISCHER, K., KETTUNEN, J., WURTZ, P., HALLER, T., HAVULINNA, A. S., KANGAS, A. J., SOININEN, P., ESKO, T., TAMMESOO, M. L., MAGI, R., SMIT, S., PALOTIE, A., RIPATTI, S., SALOMAA, V., ALA-KORPELA, M., PEROLA, M. & METSPALU, A. 2014. Biomarker profiling by nuclear magnetic resonance spectroscopy for the prediction of all-cause mortality: an observational study of 17,345 persons. *PLoS Med,* 11**,** e1001606.

FOLSOM, A. R., CHAMBLESS, L. E., DUNCAN, B. B., GILBERT, A. C. & PANKOW, J. S. 2003. Prediction of coronary heart disease in middle-aged adults with diabetes. *Diabetes Care,* 26**,** 2777-84.

FOWKES, F., MURRAY, G., BUTCHER, I., FOLSOM, A., HIRSCH, A., COUPER, D., DEBACKER, G., KORNITZER, M., NEWMAN, A., SUTTON-TYRRELL, K., CUSHMAN, M., LEE, A., PRICE, J., D'AGOSTINO, R., MURABITO, J., NORMAN, P., MASAKI, K., BOUTER, L., HEINE, R., STEHOUWER, C., MCDERMOTT, M., STOFFERS, H., KNOTTNERUS, J., OGREN, M., HEDBLAD, B., KOENIG, W., MEISINGER, C., CAULEY, J., FRANCO, O., HUNINK, M., HOFMAN, A., WITTEMAN, J., CRIQUI, M., LANGER, R., HIATT, W., HAMMAN, R. & COLLABORATION, A. B. I. 2014. Development and validation of an ankle brachial index risk model for the prediction of cardiovascular events. *European Journal of Preventive Cardiology,* 21**,** 310-320.

FOWKES, F. G., MURRAY, G. D., BUTCHER, I., HEALD, C. L., LEE, R. J., CHAMBLESS, L. E., FOLSOM, A. R., HIRSCH, A. T., DRAMAIX, M., DEBACKER, G., WAUTRECHT, J. C., KORNITZER, M., NEWMAN, A. B., CUSHMAN, M., SUTTON-TYRRELL, K., FOWKES, F. G., LEE, A. J., PRICE, J. F., D'AGOSTINO, R. B., MURABITO, J. M., NORMAN, P. E., JAMROZIK, K., CURB, J. D., MASAKI, K. H., RODRIGUEZ, B. L., DEKKER, J. M., BOUTER, L. M., HEINE, R. J., NIJPELS, G., STEHOUWER, C. D., FERRUCCI, L., MCDERMOTT, M. M., STOFFERS, H. E., HOOI, J. D., KNOTTNERUS, J. A., OGREN, M., HEDBLAD, B., WITTEMAN, J. C., BRETELER, M. M., HUNINK, M. G., HOFMAN, A., CRIQUI, M. H., LANGER, R. D., FRONEK, A., HIATT, W. R., HAMMAN, R., RESNICK, H. E., GURALNIK, J. & MCDERMOTT, M. M. 2008. Ankle brachial index combined with Framingham Risk Score to predict cardiovascular events and mortality: a meta-analysis. *JAMA,* 300**,** 197-208.

FOWLER, M. J. 2008. Microvascular and Macrovascular Complications of Diabetes. *Clinical Diabetes,* 26**,** 77-82.

FRAMINGHAM HEART STUDY 2002. Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation,* 106**,** 3143-421.

GANSEVOORT, R. T., CORREA-ROTTER, R., HEMMELGARN, B. R., JAFAR, T. H., HEERSPINK, H. J. L., MANN, J. F., MATSUSHITA, K. & WEN, C. P. 2013. Chronic kidney disease and cardiovascular risk: epidemiology, mechanisms, and prevention. *The Lancet,* 382**,** 339-352.

GEHLENBORG, N., O'DONOGHUE, S. I., BALIGA, N. S., GOESMANN, A., HIBBS, M. A., KITANO, H., KOHLBACHER, O., NEUWEGER, H., SCHNEIDER, R., TENENBAUM, D. & GAVIN, A. C. 2010. Visualization of omics data for systems biology. *Nat Methods,* 7**,** S56-68.

GENDERS, T. S., STEYERBERG, E. W., HUNINK, M. G., NIEMAN, K., GALEMA, T. W., MOLLET, N. R., DE FEYTER, P. J., KRESTIN, G. P., ALKADHI, H., LESCHKA, S., DESBIOLLES, L., MEIJS, M. F., CRAMER, M. J., KNUUTI, J., KAJANDER, S., BOGAERT, J., GOETSCHALCKX, K., CADEMARTIRI, F., MAFFEI, E., MARTINI, C., SEITUN, S., ALDROVANDI, A., WILDERMUTH, S., STINN, B., FORNARO, J., FEUCHTNER, G., DE ZORDO, T., AUER, T., PLANK, F., FRIEDRICH, G., PUGLIESE, F., PETERSEN, S. E., DAVIES, L. C., SCHOEPF, U. J., ROWE, G. W., VAN MIEGHEM, C. A., VAN DRIESSCHE, L., SINITSYN, V., GOPALAN, D., NIKOLAOU, K., BAMBERG, F., CURY, R. C., BATTLE, J., MAUROVICH-HORVAT, P., BARTYKOWSZKI, A., MERKELY, B., BECKER, D., HADAMITZKY, M., HAUSLEITER, J., DEWEY, M., ZIMMERMANN, E. & LAULE, M. 2012. Prediction model to estimate presence of coronary artery disease: retrospective pooled analysis of existing cohorts. *BMJ,* 344**,** e3485.

GERDS, T. A., KATTAN, M. W., SCHUMACHER, M. & YU, C. 2013. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Stat Med,* 32**,** 2173-84.

GERSTEIN, H. C., PARE, G., MCQUEEN, M. J., HAENEL, H., LEE, S. F., POGUE, J., MAGGIONI, A. P., YUSUF, S. & HESS, S. 2015. Identifying Novel Biomarkers for Cardiovascular Events or Death in People With Dysglycemia. *Circulation,* 132**,** 2297-304.

GERSZTEN, R. E. & WANG, T. J. 2008. The search for new cardiovascular biomarkers. *Nature,* 451**,** 949-52.

GRAHAM, J. W. 2009. Missing data analysis: making it work in the real world. *Annu Rev Psychol,* 60**,** 549-76.

GUZDER, R. N., GATLING, W., MULLEE, M. A., MEHTA, R. L. & BYRNE, C. D. 2005. Prognostic value of the Framingham cardiovascular risk equation and the UKPDS risk engine for coronary heart disease in newly diagnosed Type 2 diabetes: results from a United Kingdom study. *Diabet Med,* 22**,** 554-62.

HALTER, J. B., MUSI, N., MCFARLAND HORNE, F., CRANDALL, J. P., GOLDBERG, A., HARKLESS, L., HAZZARD, W. R., HUANG, E. S., KIRKMAN, M. S., PLUTZKY, J., SCHMADER, K. E., ZIEMAN, S. & HIGH, K. P. 2014. Diabetes and Cardiovascular Disease in Older Adults: Current Status and Future Directions. *Diabetes,* 63**,** 2578-2589.

HAND, D. J. 2006. Classifier Technology and the Illusion of Progress. 1-14.

HARRELL, F. E., JR., LEE, K. L., CALIFF, R. M., PRYOR, D. B. & ROSATI, R. A. 1984. Regression modelling strategies for improved prognostic prediction. *Stat Med,* 3**,** 143-52.

HARRELL, F. E., JR., LEE, K. L. & MARK, D. B. 1996. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med,* 15**,** 361-87.

HAYES, A. J., LEAL, J., KELMAN, C. W. & CLARKE, P. M. 2011. Risk equations to predict life expectancy of people with Type 2 diabetes mellitus following major complications: a study from Western Australia. *Diabetic Medicine,* 28**,** 428-35.

HIGGINS, J. P. T. & GREEN, S. 2011. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]*. [Online]. Available: www.handbook.cochrane.org

[Accessed 30/1/17.

HILLIS, G. S., WELSH, P., CHALMERS, J., PERKOVIC, V., CHOW, C. K., LI, Q., JUN, M., NEAL, B., ZOUNGAS, S., POULTER, N., MANCIA, G., WILLIAMS, B., SATTAR, N. & WOODWARD, M. 2014. The relative and combined ability of high-sensitivity cardiac troponin T and N-terminal pro-B-type natriuretic peptide to predict cardiovascular events and death in patients with type 2 diabetes. *Diabetes Care,* 37**,** 295-303.

HINNOUHO, G. M., CZERNICHOW, S., DUGRAVOT, A., NABI, H., BRUNNER, E. J., KIVIMAKI, M. & SINGH-MANOUX, A. 2015. Metabolically healthy obesity and the risk of cardiovascular disease and type 2 diabetes: the Whitehall II cohort study. *Eur Heart J,* 36**,** 551-9.

HIPPISLEY-COX, J., COUPLAND, C. & BRINDLE, P. 2013. 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,* 346**,** f2573.

HIPPISLEY-COX, J., COUPLAND, C. & BRINDLE, P. 2014. Validation of QRISK2 (2014) in patients with diabetes. Available: http://www.qrisk.org/index.php [Accessed 26/10/2015].

HIPPISLEY-COX, J., COUPLAND, C., ROBSON, J. & BRINDLE, P. 2010. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. *BMJ,* 341**,** c6624.

HIPPISLEY-COX, J., COUPLAND, C., VINOGRADOVA, Y., ROBSON, J., MINHAS, R., SHEIKH, A. & BRINDLE, P. 2008. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ,* 336**,** 1475-82.

HOSHINO, T., ISHIZUKA, K., SHIMIZU, S. & UCHIYAMA, S. 2013. CHADS2 score predicts functional outcome of stroke in patients with a history of coronary artery disease. *Journal of the Neurological Sciences,* 331**,** 57-60.

HOSMER, D. W., HOSMER, T., LE CESSIE, S. & LEMESHOW, S. 1997. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med,* 16**,** 965-80.

HSU, F.-C., KRITCHEVSKY, S. B., LIU, Y., KANAYA, A., NEWMAN, A. B., PERRY, S. E., VISSER, M., PAHOR, M., HARRIS, T. B., NICKLAS, B. J. & STUDY, F. T. H. A. 2009. Association Between Inflammatory Components and Physical Function in the Health, Aging, and Body Composition Study: A Principal Component Analysis Approach. *The*

Journals of Gerontology Series A: Biological Sciences and Medical Sciences, 64A**,** 581-589.
INTERNATIONAL DIABETES FEDERATION 2015. IDF Diabetes Atlas: Seventh Edition.
ISHIKAWA, S., MATSUMOTO, M., KAYABA, K., GOTOH, T., NAGO, N., TSUTSUMI, A. & KAJII, E. 2009. Risk charts illustrating the 10-year risk of stroke among residents of Japanese rural communities: the JMS Cohort Study. *J Epidemiol,* 19**,** 101-6.
JOLANI, S., DEBRAY, T. P. A., KOFFIJBERG, H., VAN BUUREN, S. & MOONS, K. G. M. 2015. Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. *Statistics in Medicine,* 34**,** 1841-1863.
JOUSILAHTI, P., RASTENYTE, D. & TUOMILEHTO, J. 2000. Serum gamma-glutamyl transferase, self-reported alcohol drinking, and the risk of stroke. *Stroke,* 31**,** 1851-5.
KAHN, S. E., HULL, R. L. & UTZSCHNEIDER, K. M. 2006. Mechanisms linking obesity to insulin resistance and type 2 diabetes. *Nature,* 444**,** 840-846.
KAMOUCHI, M., KUMAGAI, N., OKADA, Y., ORIGASA, H., YAMAGUCHI, T. & KITAZONO, T. 2012. Risk score for predicting recurrence in patients with ischemic stroke: the Fukuoka stroke risk score for Japanese. *Cerebrovascular Diseases,* 34**,** 351-7.
KANNEL, W. B. 1974. Role of blood pressure in cardiovascular morbidity and mortality. *Progress in Cardiovascular Diseases,* 17**,** 5-24.
KANNEL, W. B. 1996. Blood pressure as a cardiovascular risk factor: Prevention and treatment. *JAMA,* 275**,** 1571-1576.
KANNEL, W. B., HJORTLAND, M. C., MCNAMARA, P. M. & GORDON, T. 1976. Menopause and Risk of Cardiovascular DiseaseThe Framingham Study. *Annals of Internal Medicine,* 85**,** 447-452.
KAVOUSI, M., ELIAS-SMALE, S., RUTTEN, J. H., LEENING, M. J., VLIEGENTHART, R., VERWOERT, G. C., KRESTIN, G. P., OUDKERK, M., DE MAAT, M. P., LEEBEEK, F. W., MATTACE-RASO, F. U., LINDEMANS, J., HOFMAN, A., STEYERBERG, E. W., VAN DER LUGT, A., VAN DEN MEIRACKER, A. H. & WITTEMAN, J. C. 2012. Evaluation of newer risk markers for coronary heart disease risk classification: a cohort study. *Annals of Internal Medicine,* 156**,** 438-44.
KAZEMI-SHIRAZI, L., ENDLER, G., WINKLER, S., SCHICKBAUER, T., WAGNER, O. & MARSIK, C. 2007. Gamma glutamyltransferase and long-term survival: is it just the liver? *Clin Chem,* 53**,** 940-6.
KELLY, T. N., BAZZANO, L. A., FONSECA, V. A., THETHI, T. K., REYNOLDS, K. & HE, J. 2009. Systematic Review: Glucose Control and Cardiovascular Disease in Type 2 Diabetes. *Annals of Internal Medicine,* 151**,** 394-403.
KENGNE, A. P. 2013. The ADVANCE cardiovascular risk model and current strategies for cardiovascular disease risk evaluation in people with diabetes. *Cardiovascular Journal of Africa,* 24**,** 376-81.
KENGNE, A. P., PATEL, A., MARRE, M., TRAVERT, F., LIEVRE, M., ZOUNGAS, S., CHALMERS, J., COLAGIURI, S., GROBBEE, D. E., HAMET, P., HELLER, S., NEAL, B. & WOODWARD, M. 2011a.

Contemporary model for cardiovascular risk prediction in people with type 2 diabetes. *Eur J Cardiovasc Prev Rehabil,* 18**,** 393-8.
KENGNE, A. P., PATEL, A., MARRE, M., TRAVERT, F., LIEVRE, M., ZOUNGAS, S., CHALMERS, J., COLAGIURI, S., GROBBEE, D. E., HAMET, P., HELLER, S., NEAL, B., WOODWARD, M. & GROUP, A. C. 2011b. Contemporary model for cardiovascular risk prediction in people with type 2 diabetes. *European Journal of Cardiovascular Prevention & Rehabilitation,* 18**,** 393-8.
KERR, K. F., WANG, Z., JANES, H., MCCLELLAND, R. L., PSATY, B. M. & PEPE, M. S. 2014. Net reclassification indices for evaluating risk prediction instruments: a critical review. *Epidemiology,* 25**,** 114-21.
KEUN, H. C. & ATHERSUCH, T. J. 2011. Nuclear magnetic resonance (NMR)-based metabolomics. *Methods Mol Biol,* 708**,** 321-34.
KISTORP, C., RAYMOND, I., PEDERSEN, F., GUSTAFSSON, F., FABER, J. & HILDEBRANDT, P. 2005. N-terminal pro-brain natriuretic peptide, C-reactive protein, and urinary albumin levels as predictors of mortality and cardiovascular events in older adults. *JAMA,* 293**,** 1609-16.
KNUIMAN, M. W., VU, H. T. & BARTHOLOMEW, H. C. 1998. Multivariate risk estimation for coronary heart disease: the Busselton Health Study. *Aust N Z J Public Health,* 22**,** 747-53.
KONTUSH, A., LHOMME, M. & CHAPMAN, M. J. 2013. Unraveling the complexities of the HDL lipidome. *J Lipid Res,* 54**,** 2950-63.
KOOPMAN, L., VAN DER HEIJDEN, G. J. M. G., GROBBEE, D. E. & ROVERS, M. M. 2008. Comparison of Methods of Handling Missing Data in Individual Patient Data Meta-analyses: An Empirical Example on Antibiotics in Children with Acute Otitis Media. *American Journal of Epidemiology,* 167**,** 540-545.
KORO, C. E., LEE, B. H. & BOWLIN, S. J. 2009. Antidiabetic medication use and prevalence of chronic kidney disease among patients with type 2 diabetes mellitus in the United States. *Clinical Therapeutics,* 31**,** 2608-2617.
KOTHARI, V., STEVENS, R. J., ADLER, A. I., STRATTON, I. M., MANLEY, S. E., NEIL, H. A. & HOLMAN, R. R. 2002. UKPDS 60: risk of stroke in type 2 diabetes estimated by the UK Prospective Diabetes Study risk engine. *Stroke,* 33**,** 1776-81.
KRAMER, A. A. & ZIMMERMAN, J. E. 2007. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Crit Care Med,* 35**,** 2052-6.
KUMAR, R. G., RUBIN, J. E., BERGER, R. P., KOCHANEK, P. M. & WAGNER, A. K. 2016. Principal components derived from CSF inflammatory profiles predict outcome in survivors after severe traumatic brain injury. *Brain Behav Immun,* 53**,** 183-93.
KUNUTSOR, S. K., BAKKER, S. J., KOOTSTRA-ROS, J. E., GANSEVOORT, R. T. & DULLAART, R. P. 2015. Circulating gamma glutamyltransferase and prediction of cardiovascular disease. *Atherosclerosis,* 238**,** 356-64.
LAKIER, J. B. 1992. The Effects of Cigarette Smoking: A Global PerspectiveSmoking and cardiovascular disease. *The American Journal of Medicine,* 93**,** S8-S12.

LARIVE, C. K., BARDING, G. A. & DINGES, M. M. 2015. NMR Spectroscopy for Metabolomics and Metabolic Profiling. *Analytical Chemistry,* 87**,** 133-146.
LAWLOR, D. A., BEDFORD, C., TAYLOR, M. & EBRAHIM, S. 2003. Geographical variation in cardiovascular disease, risk factors, and their control in older women: British Women's Heart and Health Study. *J Epidemiol Community Health,* 57**,** 134-40.
LEE, D. H., SILVENTOINEN, K., HU, G., JACOBS, D. R., JR., JOUSILAHTI, P., SUNDVALL, J. & TUOMILEHTO, J. 2006a. Serum gamma-glutamyltransferase predicts non-fatal myocardial infarction and fatal coronary heart disease among 28,838 middle-aged men and women. *Eur Heart J,* 27**,** 2170-6.
LEE, E. T., HOWARD, B. V., WANG, W., WELTY, T. K., GALLOWAY, J. M., BEST, L. G., FABSITZ, R. R., ZHANG, Y., YEH, J. & DEVEREUX, R. B. 2006b. Prediction of coronary heart disease in a population with high prevalence of diabetes and albuminuria: the Strong Heart Study. *Circulation,* 113**,** 2897-905.
LEE, I. M., SHIROMA, E. J., LOBELO, F., PUSKA, P., BLAIR, S. N. & KATZMARZYK, P. T. 2012. Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *The Lancet,* 380**,** 219-229.
LEMESHOW, S. & HOSMER, D. W. 1982. A REVIEW OF GOODNESS OF FIT STATISTICS FOR USE IN THE DEVELOPMENT OF LOGISTIC REGRESSION MODELS. *American Journal of Epidemiology,* 115**,** 92-106.
LERNER, D. J. & KANNEL, W. B. 1986. Patterns of coronary heart disease morbidity and mortality in the sexes: A 26-year follow-up of the Framingham population. *American Heart Journal,* 111**,** 383-390.
LEVEY, A. S. & CORESH, J. 2012. Chronic kidney disease. *The Lancet,* 379**,** 165-180.
LI, G., WU, X. W., LU, W. H., AI, R., CHEN, F. & TANG, Z. Z. 2014. Effect of atorvastatin on the expression of gamma-glutamyl transferase in aortic atherosclerotic plaques of apolipoprotein E-knockout mice. *BMC Cardiovasc Disord,* 14**,** 145.
LINSSEN, G. C., BAKKER, S. J., VOORS, A. A., GANSEVOORT, R. T., HILLEGE, H. L., DE JONG, P. E., VAN VELDHUISEN, D. J., GANS, R. O. & DE ZEEUW, D. 2010. N-terminal pro-B-type natriuretic peptide is an independent predictor of cardiovascular morbidity and mortality in the general population. *Eur Heart J,* 31**,** 120-7.
LIU, J., HONG, Y., D'AGOSTINO, R. B., SR., WU, Z., WANG, W., SUN, J., WILSON, P. W., KANNEL, W. B. & ZHAO, D. 2004. Predictive value for the Chinese population of the Framingham CHD risk assessment tool compared with the Chinese Multi-Provincial Cohort Study. *JAMA,* 291**,** 2591-9.
LUMLEY, T., KRONMAL, R. A., CUSHMAN, M., MANOLIO, T. A. & GOLDSTEIN, S. 2002. A stroke prediction score in the elderly: validation and Web-based application. *J Clin Epidemiol,* 55**,** 129-36.
MACKEY, R. H., GREENLAND, P., GOFF, D. C., JR., LLOYD-JONES, D., SIBLEY, C. T. & MORA, S. 2012. High-density lipoprotein cholesterol and

particle concentrations, carotid atherosclerosis, and coronary events: MESA (multi-ethnic study of atherosclerosis). *J Am Coll Cardiol,* 60**,** 508-16.
MAHMOOD, S. S., LEVY, D., VASAN, R. S. & WANG, T. J. 2014. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *The Lancet,* 383**,** 999-1008.
MAINOUS, A. G., 3RD, KOOPMAN, R. J., DIAZ, V. A., EVERETT, C. J., WILSON, P. W. & TILLEY, B. C. 2007. A coronary heart disease risk score based on patient-reported information. *Am J Cardiol,* 99**,** 1236-41.
MAJEED, A. 2014. Statins for primary prevention of cardiovascular disease. *BMJ : British Medical Journal,* 348.
MARIONI, R. E., STRACHAN, M. W., REYNOLDS, R. M., LOWE, G. D., MITCHELL, R. J., FOWKES, F. G., FRIER, B. M., LEE, A. J., BUTCHER, I., RUMLEY, A., MURRAY, G. D., DEARY, I. J. & PRICE, J. F. 2010. Association between raised inflammatory markers and cognitive decline in elderly people with type 2 diabetes: the Edinburgh Type 2 Diabetes Study. *Diabetes,* 59**,** 710-3.
MATSUMOTO, M., ISHIKAWA, S., KAYABA, K., GOTOH, T., NAGO, N., TSUTSUMI, A. & KAJII, E. 2009. Risk charts illustrating the 10-year risk of myocardial infarction among residents of Japanese rural communities: the JMS Cohort Study. *J Epidemiol,* 19**,** 94-100.
MCDERMOTT, M. M., LIU, K., CRIQUI, M. H., RUTH, K., GOFF, D., SAAD, M. F., WU, C., HOMMA, S. & SHARRETT, A. R. 2005. Ankle-brachial index and subclinical cardiac and carotid disease: the multi-ethnic study of atherosclerosis. *Am J Epidemiol,* 162**,** 33-41.
MCGEECHAN, K., MACASKILL, P., IRWIG, L., LIEW, G. & WONG, T. Y. 2008. Assessing new biomarkers and predictive models for use in clinical practice: A clinician's guide. *Archives of Internal Medicine,* 168**,** 2304-2310.
MCGORRIAN, C., YUSUF, S., ISLAM, S., JUNG, H., RANGARAJAN, S., AVEZUM, A., PRABHAKARAN, D., ALMAHMEED, W., RUMBOLDT, Z., BUDAJ, A., DANS, A. L., GERSTEIN, H. C., TEO, K. & ANAND, S. S. 2011. Estimating modifiable coronary heart disease risk in multiple regions of the world: the INTERHEART Modifiable Risk Score. *Eur Heart J,* 32**,** 581-9.
MENOTTI, A., LANTI, M., AGABITI-ROSEI, E., CARRATELLI, L., CAVERA, G., DORMI, A., GADDI, A., MANCINI, M., MOTOLESE, M., MUIESAN, M. L., MUNTONI, S., MUNTONI, S., NOTARBARTOLO, A., PRATI, P., REMIDDI, S. & ZANCHETTI, A. 2005. Riskard 2005. New tools for prediction of cardiovascular disease risk derived from Italian population studies. *Nutr Metab Cardiovasc Dis,* 15**,** 426-40.
MENOTTI, A., LANTI, M., PUDDU, P. E., CARRATELLI, L., MANCINI, M., MOTOLESE, M., PRATI, P. & ZANCHETTI, A. 2002. The risk functions incorporated in Riscard 2002: a software for the prediction of cardiovascular risk in the general population based on Italian data. *Ital Heart J,* 3**,** 114-21.
MIQUEL, P. 2016. Manhattan Plot. 'Oxford University Press'.
MITCHELL, L. B., SOUTHERN, D. A., GALBRAITH, D., GHALI, W. A., KNUDTSON, M., WILTON, S. B. & INVESTIGATORS, A. 2014. Prediction of stroke or TIA in patients without atrial fibrillation using CHADS2 and CHA2DS2-VASc scores. *Heart,* 100**,** 1524-30.

MOONS, K. G., BOTS, M. L., SALONEN, J. T., ELWOOD, P. C., FREIRE DE CONCALVES, A., NIKITIN, Y., SIVENIUS, J., INZITARI, D., BENETOU, V., TUOMILEHTO, J., KOUDSTAAL, P. J. & GROBBEE, D. E. 2002. Prediction of stroke in the general population in Europe (EUROSTROKE): Is there a role for fibrinogen and electrocardiography? *J Epidemiol Community Health,* 56 Suppl 1**,** i30-6.

MOONS, K. G., ROYSTON, P., VERGOUWE, Y., GROBBEE, D. E. & ALTMAN, D. G. 2009. Prognosis and prognostic research: what, why, and how? *BMJ,* 338**,** b375.

MORA, S., GLYNN, R. J. & RIDKER, P. M. 2013. HDL cholesterol, size, particle number, and residual vascular risk after potent statin therapy. *Circulation,* 128.

MORLING, J. R., FALLOWFIELD, J. A., GUHA, I. N., NEE, L. D., GLANCY, S., WILLIAMSON, R. M., ROBERTSON, C. M., STRACHAN, M. W. & PRICE, J. F. 2014a. Using non-invasive biomarkers to identify hepatic fibrosis in people with type 2 diabetes mellitus: the Edinburgh type 2 diabetes study. *J Hepatol,* 60**,** 384-91.

MORLING, J. R., FALLOWFIELD, J. A., WILLIAMSON, R. M., NEE, L. D., JACKSON, A. P., GLANCY, S., REYNOLDS, R. M., HAYES, P. C., GUHA, I. N., STRACHAN, M. W. & PRICE, J. F. 2014b. Non-invasive hepatic biomarkers (ELF and CK18) in people with type 2 diabetes: the Edinburgh type 2 diabetes study. *Liver Int,* 34**,** 1267-77.

MORLING, J. R., FALLOWFIELD, J. A., WILLIAMSON, R. M., ROBERTSON, C. M., GLANCY, S., GUHA, I. N., STRACHAN, M. W. & PRICE, J. F. 2015. gamma-Glutamyltransferase, but not markers of hepatic fibrosis, is associated with cardiovascular disease in older people with type 2 diabetes mellitus: the Edinburgh Type 2 Diabetes Study. *Diabetologia,* 58**,** 1484-93.

MUKAMAL, K. J., KIZER, J. R., DJOUSSE, L., IX, J. H., ZIEMAN, S., SISCOVICK, D. S., SIBLEY, C. T., TRACY, R. P. & ARNOLD, A. M. 2013. Prediction and classification of cardiovascular disease risk in older adults with diabetes. *Diabetologia,* 56**,** 275-83.

MUNTNER, P., COLANTONIO, L. D., CUSHMAN, M., GOFF, D. C., JR., HOWARD, G., HOWARD, V. J., KISSELA, B., LEVITAN, E. B., LLOYD-JONES, D. M. & SAFFORD, M. M. 2014. Validation of the atherosclerotic cardiovascular disease Pooled Cohort risk equations. *JAMA,* 311**,** 1406-15.

MURPHY, T. P., DHANGANA, R., PENCINA, M. J. & D'AGOSTINO SR, R. B. 2012. Ankle–brachial index and cardiovascular risk prediction: An analysis of 11,594 individuals with 10-year follow-up. *Atherosclerosis,* 220**,** 160-167.

NDREPEPA, G., BRAUN, S., CASSESE, S., FUSARO, M., LAUGWITZ, K. L., SCHUNKERT, H. & KASTRATI, A. 2016. Relation of Gamma-Glutamyl Transferase to Cardiovascular Events in Patients With Acute Coronary Syndromes. *Am J Cardiol,* 117**,** 1427-32.

NEATH, A. A. & CAVANAUGH, J. E. 2012. The Bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics,* 4**,** 199-203.

NELSON, M. R., RAMSAY, E., RYAN, P., WILLSON, K., TONKIN, A. M., WING, L., SIMONS, L., REID, C. M. & SECOND AUSTRALIAN NATIONAL BLOOD PRESSURE MANAGEMENT, C. 2012. A score for

the prediction of cardiovascular events in the hypertensive aged. *American Journal of Hypertension,* 25**,** 190-4.
NICE CG181. 2016. *Cardiovascular disease: risk assessment and reduction, including lipid modification* [Online]. Available: https://www.nice.org.uk/guidance/cg181/chapter/1-Recommendations#identifying-and-assessing-cardiovascular-disease-cvd-risk-2 [Accessed 25/11/16.
NICE CG187. 2014. *Acute heart failure: diagnosis and management* [Online]. Available: https://www.nice.org.uk/guidance/cg187/chapter/1-recommendations.
NICE DG15. 2014. *Myocardial infarction (acute): Early rule out using high-sensitivity troponin tests (Elecsys Troponin T high-sensitive, ARCHITECT STAT High Sensitive Troponin-I and AccuTnI+3 assays)* [Online]. Available: https://www.nice.org.uk/guidance/dg15/chapter/1-Recommendations [Accessed 04/05/2016.
NICE NG28 2015. Type 2 diabetes in adults: management.
OFFICE FOR NATIONAL STATISTICS. 2010. *SOC2010 volume 3: The National Statistics Socio-economic classification* [Online]. Available: http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/guide-method/classifications/current-standard-classifications/soc2010/soc2010-volume-3-ns-sec--rebased-on-soc2010--user-manual/index.html [Accessed 02/12/16.
OFFICE OF POPULATION CENSUSES AND SURVEYS 1993. *Tabular list of the classification of surgical operations and procedures: Fourth revision consolidated version 1990*.
OLSEN, M. H., HANSEN, T. W., CHRISTENSEN, M. K., GUSTAFSSON, F., RASMUSSEN, S., WACHTELL, K., IBSEN, H., TORP-PEDERSEN, C. & HILDEBRANDT, P. R. 2007. N-terminal pro-brain natriuretic peptide, but not high sensitivity C-reactive protein, improves cardiovascular risk prediction in the general population. *Eur Heart J,* 28**,** 1374-81.
OSBORN, D. P., HARDOON, S., OMAR, R. Z., HOLT, R. I., KING, M., LARSEN, J., MARSTON, L., MORRIS, R. W., NAZARETH, I., WALTERS, K. & PETERSEN, I. 2015. Cardiovascular risk prediction models for people with severe mental illness: results from the prediction and management of cardiovascular risk in people with severe mental illnesses (PRIMROSE) research program. *JAMA Psychiatry,* 72**,** 143-51.
OVBIAGELE, B., GOLDSTEIN, L. B., AMARENCO, P., MESSIG, M., SILLESEN, H., CALLAHAN, A., 3RD, HENNERICI, M. G., ZIVIN, J., WELCH, K. M. & INVESTIGATORS, S. 2014. Prediction of major vascular events after stroke: the stroke prevention by aggressive reduction in cholesterol levels trial. *Journal of Stroke & Cerebrovascular Diseases,* 23**,** 778-84.
PARK, G. M., AN, H., LEE, S. W., CHO, Y. R., GIL, E. H., HER, S. H., KIM, Y. H., LEE, C. W., KOH, E. H., LEE, W. J., KIM, M. S., LEE, K. U., KANG, J. W., LIM, T. H., PARK, S. W., PARK, S. J. & PARK, J. Y. 2015. Risk score model for the assessment of coronary artery disease in asymptomatic patients with type 2 diabetes. *Medicine,* 94**,** e508.

PAVLOU, M., AMBLER, G., SEAMAN, S. R., GUTTMANN, O., ELLIOTT, P., KING, M. & OMAR, R. Z. 2015. How to develop a more accurate risk prediction model when there are few events. *BMJ,* 351.

PAYNTER, N. P., MAZER, N. A., PRADHAN, A. D., GAZIANO, J. M., RIDKER, P. M. & COOK, N. R. 2011. Cardiovascular risk prediction in diabetic men and women using hemoglobin A1c vs diabetes as a high-risk equivalent. *Archives of Internal Medicine,* 171**,** 1712-8.

PEDUZZI, P., CONCATO, J., FEINSTEIN, A. R. & HOLFORD, T. R. 1995. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epidemiol,* 48**,** 1503-10.

PEDUZZI, P., CONCATO, J., KEMPER, E., HOLFORD, T. R. & FEINSTEIN, A. R. 1996. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol,* 49**,** 1373-9.

PELLEGRINI, E., MAURANTONIO, M., GIANNICO, I. M., SIMONINI, M. S., GANAZZI, D., CARULLI, L., D'AMICO, R., BALDINI, A., LORIA, P., BERTOLOTTI, M. & CARULLI, N. 2011. Risk for cardiovascular events in an Italian population of patients with type 2 diabetes. *Nutrition Metabolism & Cardiovascular Diseases,* 21**,** 885-92.

PENCINA, M. J., D'AGOSTINO, R. B., SR., D'AGOSTINO, R. B., JR. & VASAN, R. S. 2008. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med,* 27**,** 157-72; discussion 207-12.

PENCINA, M. J., D'AGOSTINO, R. B., SR., LARSON, M. G., MASSARO, J. M. & VASAN, R. S. 2009. Predicting the 30-year risk of cardiovascular disease: the framingham heart study. *Circulation,* 119**,** 3078-84.

PENCINA, M. J., D'AGOSTINO, R. B. & VASAN, R. S. 2010. Statistical methods for assessment of added usefulness of new biomarkers. *Clin Chem Lab Med,* 48**,** 1703-11.

PEPE, M. S., FENG, Z., HUANG, Y., LONGTON, G., PRENTICE, R., THOMPSON, I. M. & ZHENG, Y. 2008. Integrating the predictiveness of a marker with its performance as a classifier. *Am J Epidemiol,* 167**,** 362-8.

PFISTER, R., CAIRNS, R., ERDMANN, E. & SCHNEIDER, C. A. 2013. A clinical risk score for heart failure in patients with type 2 diabetes and macrovascular disease: an analysis of the PROactive study. *International Journal of Cardiology,* 162**,** 112-6.

PIGNONE, M., SHERIDAN, S. L., LEE, Y. Z., KUO, J., PHILLIPS, C., MULROW, C. & ZEIGER, R. 2004. Heart to Heart: a computerized decision aid for assessment of coronary heart disease risk and the impact of risk-reduction interventions for primary prevention. *Prev Cardiol,* 7**,** 26-33.

POIRIER, P., GILES, T. D., BRAY, G. A., HONG, Y., STERN, J. S., PI-SUNYER, F. X. & ECKEL, R. H. 2006. Obesity and cardiovascular disease: pathophysiology, evaluation, and effect of weight loss. *Arterioscler Thromb Vasc Biol,* 26**,** 968-76.

POVEL, C. M., BEULENS, J. W., VAN DER SCHOUW, Y. T., DOLLE, M. E., SPIJKERMAN, A. M., VERSCHUREN, W. M., FESKENS, E. J. & BOER, J. M. 2013. Metabolic syndrome model definitions predicting type 2 diabetes and cardiovascular disease. *Diabetes Care,* 36**,** 362-8.

PRICE, J. F., REYNOLDS, R. M., MITCHELL, R. J., WILLIAMSON, R. M., FOWKES, F. G., DEARY, I. J., LEE, A. J., FRIER, B. M., HAYES, P. C. & STRACHAN, M. W. 2008. The Edinburgh Type 2 Diabetes Study: study protocol. *BMC Endocr Disord,* 8**,** 18.

PRIETO-MERINO, D., DOBSON, J., GUPTA, A. K., CHANG, C. L., SEVER, P. S., DAHLOF, B., WEDEL, H., POCOCK, S., POULTER, N. & INVESTIGATORS, A.-B. 2013. ASCORE: an up-to-date cardiovascular risk score for hypertensive patients reflecting contemporary clinical practice developed using the (ASCOT-BPLA) trial data. *Journal of Human Hypertension,* 27**,** 492-6.

PROSPECTIVE STUDIES COLLABORATION 2012. 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,* 370**,** 1829-1839.

PULGARON, E. & DELAMATER, A. 2014. Obesity and Type 2 Diabetes in Children: Epidemiology and Treatment. *Current Diabetes Reports,* 14**,** 1-12.

QIAO, Q., GAO, W., LAATIKAINEN, T. & VARTIAINEN, E. 2012. Layperson-oriented vs. clinical-based models for prediction of incidence of ischemic stroke: National FINRISK Study. *International Journal of Stroke,* 7**,** 662-8.

RABAR, S., HARKER, M., O'FLYNN, N. & WIERZBICKI, A. S. 2014. Lipid modification and cardiovascular risk assessment for the primary and secondary prevention of cardiovascular disease: summary of updated NICE guidance. *BMJ : British Medical Journal,* 349.

RADER, D. J. & HOVINGH, G. K. 2014. HDL and cardiovascular disease. *Lancet,* 384**,** 618-25.

RAPSOMANIKI, E., SHAH, A., PEREL, P., DENAXAS, S., GEORGE, J., NICHOLAS, O., UDUMYAN, R., FEDER, G. S., HINGORANI, A. D., TIMMIS, A., SMEETH, L. & HEMINGWAY, H. 2014. Prognostic models for stable coronary artery disease based on electronic health record cohort of 102 023 patients. *European Heart Journal,* 35**,** 844-52.

RECORD. 2015. *RECORD Checklist* [Online]. Available: http://record-statement.org/checklist.php [Accessed 24/11/16.

RETNAKARAN, R., CULL, C. A., THORNE, K. I., ADLER, A. I. & HOLMAN, R. R. 2006. Risk factors for renal dysfunction in type 2 diabetes: U.K. Prospective Diabetes Study 74. *Diabetes,* 55**,** 1832-9.

RIDKER, P. M., BURING, J. E., RIFAI, N. & COOK, N. R. 2007. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score. *JAMA,* 297**,** 611-9.

ROBINSON, T., ELLEY, C. R., WELLS, S., ROBINSON, E., KENEALY, T., PYLYPCHUK, R., BRAMLEY, D., ARROLL, B., CRENGLE, S., RIDDELL, T., AMERATUNGA, S., METCALF, P. & DRURY, P. L. 2012. New Zealand Diabetes Cohort Study cardiovascular risk score for people with Type 2 diabetes: validation in the PREDICT cohort. *Journal of Primary Health Care,* 4**,** 181-8.

RODONDI, N., MARQUES-VIDAL, P., BUTLER, J., SUTTON-TYRRELL, K., CORNUZ, J., SATTERFIELD, S., HARRIS, T., BAUER, D. C., FERRUCCI, L., VITTINGHOFF, E. & NEWMAN, A. B. 2010. Markers of

atherosclerosis and inflammation for prediction of coronary heart disease in older adults. *Am J Epidemiol,* 171**,** 540-9.

ROYSTON, P., MOONS, K. G., ALTMAN, D. G. & VERGOUWE, Y. 2009. Prognosis and prognostic research: Developing a prognostic model. *BMJ,* 338**,** b604.

RUTTMANN, E., BRANT, L. J., CONCIN, H., DIEM, G., RAPP, K. & ULMER, H. 2005. Gamma-glutamyltransferase as a risk factor for cardiovascular disease mortality: an epidemiological investigation in a cohort of 163,944 Austrian adults. *Circulation,* 112**,** 2130-7.

SAUNDERS, J. T., NAMBI, V., DE LEMOS, J. A., CHAMBLESS, L. E., VIRANI, S. S., BOERWINKLE, E., HOOGEVEEN, R. C., LIU, X., ASTOR, B. C., MOSLEY, T. H., FOLSOM, A. R., HEISS, G., CORESH, J. & BALLANTYNE, C. M. 2011. Cardiac troponin T measured by a highly sensitive assay predicts coronary heart disease, heart failure, and mortality in the Atherosclerosis Risk in Communities Study. *Circulation,* 123**,** 1367-76.

SCARBOROUGH, P., BHATNAGAR, P., KAUR, A., SMOLINA, K., WICKRAMASINGHE, K. & RAYNER, M. 2010. Ethnic Differences in Cardiovascular Disease: 2010 edition. *British Heart Foundation Health Promotion Research Group*.

SCHAU, B., BOYSEN, G., TRUELSEN, T., BODEN-ALBALA, B., CHENG, J., BABAMOTO, E., ZAHER, C. & SACCO, R. L. 2003. Development and validation of a model to estimate stroke incidence in a population. *J Stroke Cerebrovasc Dis,* 12**,** 22-8.

SHAH, S., CASAS, J. P., GAUNT, T. R., COOPER, J., DRENOS, F., ZABANEH, D., SWERDLOW, D. I., SHAH, T., SOFAT, R., PALMEN, J., KUMARI, M., KIVIMAKI, M., EBRAHIM, S., SMITH, G. D., LAWLOR, D. A., TALMUD, P. J., WHITTAKER, J., DAY, I. N. M., HINGORANI, A. D. & HUMPHRIES, S. E. 2013a. Influence of common genetic variation on blood lipid levels, cardiovascular risk, and coronary events in two British prospective cohort studies. *European Heart Journal,* 34**,** 972-981.

SHAH, T., ENGMANN, J., DALE, C., SHAH, S., WHITE, J., GIAMBARTOLOMEI, C., MCLACHLAN, S., ZABANEH, D., CAVADINO, A., FINAN, C., WONG, A., AMUZU, A., ONG, K., GAUNT, T., HOLMES, M. V., WARREN, H., DAVIES, T.-L., DRENOS, F., COOPER, J., SOFAT, R., CAULFIELD, M., EBRAHIM, S., LAWLOR, D. A., TALMUD, P. J., HUMPHRIES, S. E., POWER, C., HYPPONEN, E., RICHARDS, M., HARDY, R., KUH, D., WAREHAM, N., BEN-SHLOMO, Y., DAY, I. N., WHINCUP, P., MORRIS, R., STRACHAN, M. W. J., PRICE, J., KUMARI, M., KIVIMAKI, M., PLAGNOL, V., DUDBRIDGE, F., WHITTAKER, J. C., CASAS, J. P., HINGORANI, A. D. & THE, U. C. 2013b. Population Genomics of Cardiometabolic Traits: Design of the University College London-London School of Hygiene and Tropical Medicine-Edinburgh-Bristol (UCLEB) Consortium. *PLoS ONE,* 8**,** e71345.

SHAPER, A. G., POCOCK, S. J., WALKER, M., COHEN, N. M., WALE, C. J. & THOMSON, A. G. 1981. British Regional Heart Study: Cardiovascular Risk Factors In Middle-Aged Men In 24 Towns. *British Medical Journal (Clinical Research Edition),* 283**,** 179-186.

SIDAK, Z. 1967. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association,* 62**,** 626-633.
SIGAL, R. J., KENNY, G. P., WASSERMAN, D. H. & CASTANEDA-SCEPPA, C. 2004. Physical activity/exercise and type 2 diabetes. *Diabetes Care,* 27**,** 2518-39.
SIGN. 2006. Diagnosis and management of peripheral arterial disease: A national clinical guideline. Available: http://www.sign.ac.uk/pdf/sign89.pdf [Accessed 04/05/2016].
SIMES, R. J. 1986. An improved Bonferroni procedure for multiple tests of significance. *Biometrika,* 73**,** 751-754.
SIMMONS, R. K., COLEMAN, R. L., PRICE, H. C., HOLMAN, R. R., KHAW, K. T., WAREHAM, N. J. & GRIFFIN, S. J. 2009. Performance of the UK Prospective Diabetes Study Risk Engine and the Framingham Risk Equations in Estimating Cardiovascular Disease in the EPIC- Norfolk Cohort. *Diabetes Care,* 32**,** 708-13.
SINGER, D. E., CHANG, Y., BOROWSKY, L. H., FANG, M. C., POMERNACKI, N. K., UDALTSOVA, N., REYNOLDS, K. & GO, A. S. 2013. A new risk scheme to predict ischemic stroke and other thromboembolism in atrial fibrillation: the ATRIA study stroke risk score. *Journal of the American Heart Association,* 2**,** e000250.
SOININEN, P., KANGAS, A. J., WURTZ, P., SUNA, T. & ALA-KORPELA, M. 2015. Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circ Cardiovasc Genet,* 8**,** 192-206.
SONG, B., FANG, H., ZHAO, L., GAO, Y., TAN, S., LU, J., SUN, S., CHANDRA, A., WANG, R. & XU, Y. 2013. Validation of the ABCD3-I score to predict stroke risk after transient ischemic attack. *Stroke,* 44**,** 1244-8.
STEVEN, S., HOLLINGSWORTH, K. G., AL-MRABEH, A., AVERY, L., ARIBISALA, B., CASLAKE, M. & TAYLOR, R. 2016. Very-Low-Calorie Diet and 6 Months of Weight Stability in Type 2 Diabetes: Pathophysiologic Changes in Responders and Nonresponders. *Diabetes Care*.
STEVENS, R. J., KOTHARI, V., ADLER, A. I. & STRATTON, I. M. 2001. The UKPDS risk engine: a model for the risk of coronary heart disease in Type II diabetes (UKPDS 56). *Clin Sci (Lond),* 101**,** 671-9.
STEYERBERG, E. W. 2009. Clinical Prediction Models. Springer New York.
STEYERBERG, E. W., MOONS, K. G., VAN DER WINDT, D. A., HAYDEN, J. A., PEREL, P., SCHROTER, S., RILEY, R. D., HEMINGWAY, H. & ALTMAN, D. G. 2013. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med,* 10**,** e1001381.
STEYERBERG, E. W., VICKERS, A. J., COOK, N. R., GERDS, T., GONEN, M., OBUCHOWSKI, N., PENCINA, M. J. & KATTAN, M. W. 2010. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology,* 21**,** 128-38.
STONE, N. J., ROBINSON, J. G., LICHTENSTEIN, A. H., BAIREY MERZ, C. N., BLUM, C. B., ECKEL, R. H., GOLDBERG, A. C., GORDON, D., LEVY, D., LLOYD-JONES, D. M., MCBRIDE, P., SCHWARTZ, J. S., SHERO, S. T., SMITH, S. C., JR., WATSON, K. & WILSON, P. W. 2014. 2013

ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *J Am Coll Cardiol,* 63**,** 2889-934.

SUZUKI, S., SAGARA, K., OTSUKA, T., MATSUNO, S., FUNADA, R., UEJIMA, T., OIKAWA, Y., YAJIMA, J., KOIKE, A., NAGASHIMA, K., KIRIGAYA, H., SAWADA, H., AIZAWA, T. & YAMASHITA, T. 2012. A new scoring system for evaluating the risk of heart failure events in Japanese patients with atrial fibrillation. *American Journal of Cardiology,* 110**,** 678-82.

THE EMERGING RISK FACTORS COLLABORATION 2012. C-Reactive Protein, Fibrinogen, and Cardiovascular Disease Prediction. *N Engl J Med,* 367**,** 1310-20.

THE SCOTTISH GOVERNMENT. 2012. *SIMD Scottish Index of Multiple Deprivation* [Online]. Available: http://simd.scotland.gov.uk/publication-2012/ [Accessed 10/08/2016.

THOMSEN, T. F., DAVIDSEN, M., IBSEN, H., JORGENSEN, T., JENSEN, G. & BORCH-JOHNSEN, K. 2001. A new method for CHD prediction and prevention based on regional risk scores and randomized clinical trials; PRECARD and the Copenhagen Risk Score. *J Cardiovasc Risk,* 8**,** 291-7.

TILLIN, T., FOROUHI, N. G., MCKEIGUE, P. M. & CHATURVEDI, N. 2012. Southall And Brent REvisited: Cohort profile of SABRE, a UK population-based comparison of cardiovascular disease and diabetes in people of European, Indian Asian and African Caribbean origins. *Int J Epidemiol,* 41**,** 33-42.

TILLIN, T., HUGHES, A. D., WHINCUP, P., MAYET, J., SATTAR, N., MCKEIGUE, P. M. & CHATURVEDI, N. 2014. Ethnicity and prediction of cardiovascular disease: performance of QRISK2 and Framingham scores in a U.K. tri-ethnic prospective cohort study (SABRE--Southall And Brent REvisited). *Heart,* 100**,** 60-7.

TOWNSEND N., B. P., WILKINS E., WICKRAMASINGHE E., RAYNER M., 2015. CARDIOVASCULAR DISEASE STATISTICS, 2015. *British Heart Foundation: London.*

TOWNSEND, P., PHILLIMORE, P. & BEATTIE, A. 1988. *Health and Deprivation: Inequality and the North,* Croom Helm.

TRIPEPI, G., HEINZE, G., JAGER, K. J., STEL, V. S., DEKKER, F. W. & ZOCCALI, C. 2013. Risk prediction models. *Nephrol Dial Transplant,* 28**,** 1975-80.

TRIPEPI, G., JAGER, K. J., DEKKER, F. W. & ZOCCALI, C. 2010. Statistical methods for the assessment of prognostic biomarkers(part II): calibration and re-classification. *Nephrol Dial Transplant,* 25**,** 1402-5.

UKPDS. 2011. *UKPDS Risk Engine: FAQs* [Online]. Available: https://www.dtu.ox.ac.uk/riskengine/FAQ.php.

VAN DER HEIJDEN, A. A., ORTEGON, M. M., NIESSEN, L. W., NIJPELS, G. & DEKKER, J. M. 2009. Prediction of coronary heart disease risk in a general, pre-diabetic, and diabetic population during 10 years of follow-up: accuracy of the Framingham, SCORE, and UKPDS risk functions: The Hoorn Study. *Diabetes Care,* 32**,** 2094-8.

VAN DER LEEUW, J., BEULENS, J. W., VAN DIEREN, S., SCHALKWIJK, C. G., GLATZ, J. F., HOFKER, M. H., VERSCHUREN, W. M., BOER, J. M., VAN DER GRAAF, Y., VISSEREN, F. L., PEELEN, L. M. & VAN DER SCHOUW, Y. T. 2016. Novel Biomarkers to Improve the Prediction of Cardiovascular Event Risk in Type 2 Diabetes Mellitus. *J Am Heart Assoc,* 5.

VAN DIEREN, S., BEULENS, J. W., KENGNE, A. P., PEELEN, L. M., RUTTEN, G. E., WOODWARD, M., VAN DER SCHOUW, Y. T. & MOONS, K. G. 2012. Prediction models for the risk of cardiovascular disease in patients with type 2 diabetes: a systematic review. *Heart,* 98**,** 360-9.

VAN DIEREN, S., PEELEN, L. M., NOTHLINGS, U., VAN DER SCHOUW, Y. T., RUTTEN, G. E., SPIJKERMAN, A. M., VAN DER, A. D., SLUIK, D., BOEING, H., MOONS, K. G. & BEULENS, J. W. 2011. External validation of the UK Prospective Diabetes Study (UKPDS) risk engine in patients with type 2 diabetes. *Diabetologia,* 54**,** 264-70.

VERGEER, M., HOLLEBOOM, A. G., KASTELEIN, J. J. & KUIVENHOVEN, J. A. 2010. The HDL hypothesis: does high-density lipoprotein protect from atherosclerosis? *J Lipid Res,* 51**,** 2058-73.

VERONESI, G., GIANFAGNA, F., GIAMPAOLI, S., CHAMBLESS, L. E., MANCIA, G., CESANA, G. & FERRARIO, M. M. 2014. Improving long-term prediction of first cardiovascular event: the contribution of family history of coronary heart disease and social status. *Preventive Medicine,* 64**,** 75-80.

VERWOERT, G. C., ELIAS-SMALE, S. E., RIZOPOULOS, D., KOLLER, M. T., STEYERBERG, E. W., HOFMAN, A., KAVOUSI, M., SIJBRANDS, E. J., HOEKS, A. P., RENEMAN, R. S., MATTACE-RASO, F. U. & WITTEMAN, J. C. 2012. Does aortic stiffness improve the prediction of coronary heart disease in elderly? The Rotterdam Study. *Journal of Human Hypertension,* 26**,** 28-34.

WANG, T. J., LARSON, M. G., LEVY, D., BENJAMIN, E. J., LEIP, E. P., OMLAND, T., WOLF, P. A. & VASAN, R. S. 2004. Plasma natriuretic peptide levels and the risk of cardiovascular events and death. *N Engl J Med,* 350**,** 655-63.

WANNAMETHEE, S. G., SHAPER, A. G., LENNON, L., PAPACOSTA, O. & WHINCUP, P. 2016. Mild hyponatremia, hypernatremia and incident cardiovascular disease and mortality in older men: A population-based cohort study. *Nutrition, Metabolism, and Cardiovascular Diseases,* 26**,** 12-19.

WANNAMETHEE, S. G., WELSH, P., LOWE, G. D., GUDNASON, V., DI ANGELANTONIO, E., LENNON, L., RUMLEY, A., WHINCUP, P. H. & SATTAR, N. 2011. N-terminal pro-brain natriuretic Peptide is a more useful predictor of cardiovascular disease risk than C-reactive protein in older men with and without pre-existing cardiovascular disease. *J Am Coll Cardiol,* 58**,** 56-64.

WELSH, P., DOOLIN, O., WILLEIT, P., PACKARD, C., MACFARLANE, P., COBBE, S., GUDNASON, V., DI ANGELANTONIO, E., FORD, I. & SATTAR, N. 2013. N-terminal pro-B-type natriuretic peptide and the prediction of primary cardiovascular events: results from 15-year follow-up of WOSCOPS. *Eur Heart J,* 34**,** 443-50.

WELSH, P., HART, C., PAPACOSTA, O., PREISS, D., MCCONNACHIE, A., MURRAY, H., RAMSAY, S., UPTON, M., WATT, G., WHINCUP, P., WANNAMETHEE, G. & SATTAR, N. 2016. Prediction of Cardiovascular Disease Risk by Cardiac Biomarkers in 2 United Kingdom Cohort Studies: Does Utility Depend on Risk Thresholds For Treatment? *Hypertension,* 67**,** 309-315.

WILSON, P. W., D'AGOSTINO, R., SR., BHATT, D. L., EAGLE, K., PENCINA, M. J., SMITH, S. C., ALBERTS, M. J., DALLONGEVILLE, J., GOTO, S., HIRSCH, A. T., LIAU, C. S., OHMAN, E. M., ROTHER, J., REID, C., MAS, J. L., STEG, P. G. & REGISTRY, R. 2012. An international model to predict recurrent cardiovascular disease. *American Journal of Medicine,* 125**,** 695-703.e1.

WILSON, P. W., D'AGOSTINO, R. B., LEVY, D., BELANGER, A. M., SILBERSHATZ, H. & KANNEL, W. B. 1998. Prediction of coronary heart disease using risk factor categories. *Circulation,* 97**,** 1837-47.

WOOD, D., DE BACKER, G., FAERGEMAN, O., GRAHAM, I., MANCIA, G. & PYORALA, K. 1998. Prevention of coronary heart disease in clinical practice: recommendations of the Second Joint Task Force of European and other Societies on Coronary Prevention. *Atherosclerosis,* 140**,** 199-270.

WOODWARD, M., BRINDLE, P. & TUNSTALL-PEDOE, H. 2007. Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). *Heart,* 93**,** 172-6.

WORLD HEALTH ORGANIZATION. 2011. *Use of glycated haemoglobin (HbA1c) in the diagnosis of diabetes mellitus* [Online]. Available: http://www.who.int/diabetes/publications/diagnosis_diabetes2011/en/.

WORLD HEALTH ORGANIZATION. 2014. *Diabetes Programme* [Online]. World Health Organization, Avenue Appia 20, 1211 Geneva 27, Switzerland: World Health Organization,. Available: http://www.who.int/diabetes/action_online/basics/en/ [Accessed September 12, 2014.

WORLD HEALTH ORGANIZATION 2015a. Cardiovascular diseases (CVDs).

WORLD HEALTH ORGANIZATION. 2015b. *ICD-10 Version: 2015* [Online]. Available: http://apps.who.int/classifications/icd10/browse/2015/en [Accessed 23/11/16.

WORLD HEALTH ORGANIZATION. 2016. *Global Report on Diabetes* [Online]. Available: http://apps.who.int/iris/bitstream/10665/204871/1/9789241565257_eng.pdf?ua=1&ua=1.

WU, Y., LIU, X., LI, X., LI, Y., ZHAO, L., CHEN, Z., LI, Y., RAO, X., ZHOU, B., DETRANO, R. & LIU, K. 2006. Estimation of 10-year risk of fatal and nonfatal ischemic cardiovascular diseases in Chinese adults. *Circulation,* 114**,** 2217-25.

WURTZ, P., HAVULINNA, A. S., SOININEN, P., TYNKKYNEN, T., PRIETO-MERINO, D., TILLIN, T., GHORBANI, A., ARTATI, A., WANG, Q., TIAINEN, M., KANGAS, A. J., KETTUNEN, J., KAIKKONEN, J., MIKKILA, V., JULA, A., KAHONEN, M., LEHTIMAKI, T., LAWLOR, D. A., GAUNT, T. R., HUGHES, A. D., SATTAR, N., ILLIG, T., ADAMSKI,

J., WANG, T. J., PEROLA, M., RIPATTI, S., VASAN, R. S., RAITAKARI, O. T., GERSZTEN, R. E., CASAS, J. P., CHATURVEDI, N., ALA-KORPELA, M. & SALOMAA, V. 2015. Metabolite profiling and cardiovascular event risk: a prospective study of 3 population-based cohorts. *Circulation,* 131**,** 774-85.

WÜRTZ, P., RAIKO, J. R., MAGNUSSEN, C. G., SOININEN, P., KANGAS, A. J., TYNKKYNEN, T., THOMSON, R., LAATIKAINEN, R., SAVOLAINEN, M. J., LAURIKKA, J., KUUKASJÄRVI, P., TARKKA, M., KARHUNEN, P. J., JULA, A., VIIKARI, J. S., KÄHÖNEN, M., LEHTIMÄKI, T., JUONALA, M., ALA-KORPELA, M. & RAITAKARI, O. T. 2012. High-throughput quantification of circulating metabolites improves prediction of subclinical atherosclerosis. *European Heart Journal,* 33**,** 2307-2316.

WURTZ, P., SOININEN, P., KANGAS, A. J., MAKINEN, V. P., GROOP, P. H., SAVOLAINEN, M. J., JUONALA, M., VIIKARI, J. S., KAHONEN, M., LEHTIMAKI, T., RAITAKARI, O. T. & ALA-KORPELA, M. 2011. Characterization of systemic metabolic phenotypes associated with subclinical atherosclerosis. *Mol Biosyst,* 7**,** 385-93.

YANG, X., MA, R. C., SO, W. Y., KONG, A. P., KO, G. T., HO, C. S., LAM, C. W., COCKRAM, C. S., TONG, P. C. & CHAN, J. C. 2008a. Development and validation of a risk score for hospitalization for heart failure in patients with Type 2 diabetes mellitus. *Cardiovasc Diabetol,* 7**,** 9.

YANG, X., SO, W. Y., KONG, A. P., HO, C. S., LAM, C. W., STEVENS, R. J., LYU, R. R., YIN, D. D., COCKRAM, C. S., TONG, P. C., WONG, V. & CHAN, J. C. 2007. Development and validation of stroke risk equation for Hong Kong Chinese patients with type 2 diabetes: the Hong Kong Diabetes Registry. *Diabetes Care,* 30**,** 65-70.

YANG, X., SO, W. Y., KONG, A. P., MA, R. C., KO, G. T., HO, C. S., LAM, C. W., COCKRAM, C. S., CHAN, J. C. & TONG, P. C. 2008b. Development and validation of a total coronary heart disease risk score in type 2 diabetes mellitus. *Am J Cardiol,* 101**,** 596-601.

YATSUYA, H., ISO, H., YAMAGISHI, K., KOKUBO, Y., SAITO, I., SUZUKI, K., SAWADA, N., INOUE, M. & TSUGANE, S. 2013. Development of a point-based prediction model for the incidence of total stroke: Japan public health center study. *Stroke,* 44**,** 1295-302.

YUDKIN, J. S. & CHATURVEDI, N. 1999. Developing risk stratification charts for diabetic and nondiabetic subjects. *Diabet Med,* 16**,** 219-27.

ZODPEY, S. P., KULKARNI, H. R., VASUDEO, N. D. & CHAUBEY, B. S. 1994. A risk scoring system for prediction of coronary heart disease based on multivariate analysis: development and validation. *Indian Heart J,* 46**,** 77-83.

11 Appendices

Appendix A Publications and presentations

Comparison of non-traditional biomarkers, and combinations of biomarkers, for vascular risk prediction in people with type 2 diabetes: The Edinburgh Type 2 Diabetes Study

Anna H. Price[1], MRes, Christopher J. Weir[1,2], PhD, Paul Welsh[3], PhD, Stela McLachlan[1], PhD, Mark W. J. Strachan[4], MD, Naveed Sattar[3], MD, Jackie F. Price[1], MD.

[1]Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, Scotland, UK
[2]Edinburgh Clinical Trials Unit, University of Edinburgh, Scotland, UK
[3]Glasgow Cardiovascular Research Centre, University of Glasgow, Glasgow, Scotland, UK
[4]Metabolic Unit, Western General Hospital, Edinburgh, Scotland, UK

Corresponding author:
Anna Price
Usher Institute of Population Health Sciences and Informatics
The University of Edinburgh
Medical School, Teviot Place
Edinburgh, EH8 9AG
Scotland, UK
Tel: +44 1316503038
Fax: +44 1316506909
E-mail: A.H.Price@ed.ac.uk

Alternate corresponding author:
Jackie Price
Usher Institute of Population Health Sciences and Informatics
The University of Edinburgh
Medical School, Teviot Place
Edinburgh, EH8 9AG
Scotland, UK
Tel: +44 1316503038
Fax: +44 1316506909
E-mail: Jackie.Price@ed.ac.uk

Abstract word count: 213
Main test word count: 4435
Number of tables: 3 Number of figures: 0

Abstract

Objective

To compare the impact of eight non-traditional biomarkers (ankle brachial pressure index (ABI), N-terminal pro-brain natriuretic peptide (NT-proBNP), high sensitivity cardiac troponin (hs-cTnT), gamma-glutamyl transpeptidase (GGT) and four markers of systemic inflammation), both individually and in combination, on cardiovascular risk prediction over and above the QRISK2 score in older people with type 2 diabetes.

Study design and setting

Prospective study of 1066 men and women aged 60-75 years with type 2 diabetes mellitus, living in Lothian, Scotland.

Results

After 8 years, 205 cardiovascular events occurred. Baseline hs-cTNT, NT-proBNP and ABI, but not GGT or an inflammatory factor on their own, independently improved cardiovascular risk prediction beyond QRISK2. Increases in C statistic (from 0.722; 95% CI 0.681, 0.763 for the basic QRISK2 model) were greatest for hs-cTnT (to 0.732; 0.690, 0.774)) and NT-proBNP (to 0.726; 0.685, 0.767). Models combining biomarkers had greater C statistics, with the highest for ABI, hs-cTnT and GGT combined (0.740; 0.699, 0.781).

Conclusions

Of a range of eight potentially useful biomarkers, NT-proBNP and hs-cTnT on their own appear to be the most promising in terms of improving vascular risk prediction in people with type 2 diabetes. Combining biomarkers adds further predictive value. Future studies should evaluate the clinical benefit versus cost of adding multiple biomarkers to existing risk scores.

Introduction

The risk of cardiovascular (CV) disease increases two-fold in people with type 2 diabetes (1). In the UK, the National Institute for Health and Care Excellence (NICE) clinical guidelines recommend the use of the QRISK2 score (2) to calculate 10-year CV risk; this score combines several traditional CV risk factors and has been validated in patients with and without type 2 diabetes (3). Numerous CV risk scores have been recommended worldwide, but in general all current scores appear to perform inadequately in people with type 2 diabetes, either under- or over-estimating risk of CV events (4-6). Although people with type 2 diabetes are routinely offered lifestyle advice and treatment with lipid-lowering agents after diagnosis, better risk stratification may allow targeted use of aggressive prevention strategies.

Increasing numbers of studies have suggested biomarkers which might improve vascular risk prediction scores in the general populations, and, to a lesser extent, in diabetic study populations (7-12). However, such studies have tended to look at the addition of single risk factors over-and-above a small panel of traditional risk factors (or an established risk score based on such traditional risk factors). A direct comparison of the value of different non-traditional biomarkers has not been well evaluated within the setting of a single epidemiological study. Similarly, the value of different combinations of the most promising biomarkers has not been well studied.

The aim of the current research was to compare the addition of a number of different biomarkers to a vascular risk score currently recommended for clinical use in people with diabetes in the UK (QRISK2), and to investigate the extent to which different combinations of these various biomarkers might improve prediction. The biomarkers selected included those identified in previous research, especially those which might be of particular importance in diabetes (such as inflammatory markers which are generally raised in people with diabetes as part of a pro-inflammatory state). Since the overall aim was to provide results which would be informative for potential application in a clinical setting, biomarkers selected were also restricted to those which can be relatively easily measured in a routine clinic setting, either by means of a blood test (N-terminal prohormone of brain natriuretic peptide (NT-proBNP), high-sensitivity cardiac troponin T (hs-cTnT), gamma-glutamyl transpeptidase (GGT) and markers of systemic inflammation such as C-reactive protein (CRP), interleukin-6 (IL-6), tumor necrosis factor alpha (TNF-α) and fibrinogen) or by means of a straightforward physical test (the ankle brachial pressure index (ABI).

Methods

Study population

The study population constituted the Edinburgh Type 2 Diabetes Study (ET2DS), a population-based, prospective cohort of 1066 men and women aged between 60 and 75 years with established type 2 diabetes mellitus living in the Lothian region of Scotland, UK. In 2006/2007, participants were recruited at random from the Lothian Diabetes Register (LDR), a registry of almost all people with type 2 diabetes living in Lothian, resulting in a cohort largely representative of this target population (13) and including patients attending both general practice and secondary care for routine diabetes healthcare. Recruitment and data collection at baseline have been described in detail previously (14). In 2011 (4 year follow-up) and 2015 (8 year follow up), all surviving participants were re-assessed for CV events. Use of routine data sources (record linkage to hospital discharge records and death certificates) and GP/hospital notes, as well as direct patient contact in 2011, ensured that follow-up included all ET2DS participants.

All study participants gave their informed consent and ethical approval was granted by the Lothian Medical Research Ethics Committee.

Baseline examination and data collection

At baseline research clinics, a questionnaire was used for self-reporting of age, sex, history of diabetes and CVD, other medical conditions (including atrial fibrillation and rheumatoid arthritis), medication and smoking habits. Height and weight (for calculation of body mass index, BMI), brachial BP and a 12-lead ECG were measured. A fasting blood sample was taken for measurement of total and HDL-cholesterol and creatinine. To measure the ABI, a sphygmomanometer cuff was placed around the arm and inflated to 30mmHg above the estimated systolic BP. The pressure was reduced at a rate of 2-3mmHg per second and the BP was recorded when the first clear sound was detected. This process was repeated in both arms and both ankles (dorsalis pedis and posterior tibial arteries) using a Doppler probe and subsequently ABI was calculated as the lowest ankle pressure divided by the highest brachial pressure. Data collected in the research clinics was supplemented by linkage to all medical and surgical discharge records from Scottish hospitals since 1981 (collated by Information Services Division (ISD) of National Health Service (NHS) Scotland), routine biochemistry

data extracted from the LDR (for diagnosis of chronic kidney disease (CKD)) and scrutiny of medical records by an expert clinician as required to confirm or refute clinical diagnoses.

Determination of circulating biomarkers

Plasma from fasting venous blood samples taken at baseline was frozen for storage. Plasma NT-proBNP and hs-cTNT were subsequently measured using the Elecsys 2010 electrochemiluminescence method (Roche Diagnostics, Burgess Hill, UK), and calibrated using the manufacturer's reagents. The manufacturer's controls were used with limits of acceptability defined by the manufacturer. GGT was analysed using a Vitros Fusion chemistry system (Ortho Clinical Diagnostics, High Wycombe, UK) at the Western General Hospital, Edinburgh, UK. Assays for plasma TNF-α, IL-6, CRP and fibrinogen were carried out in the University Department of Medicine, Glasgow Royal Infirmary. TNF-α and IL-6 antigen levels were determined using high-sensitivity ELISA kits (R&D Systems, Oxon, UK). CRP was assayed using a high-sensitivity immunonephelometric assay. Fibrinogen assays were performed using stored plasma anticoagulated with trisodium citrate and the automated Clauss assay (MDA-180 coagulometer, Organon Teknika).

Assessment of CVD events

At baseline, data were collected on self-reporting of a doctor's diagnosis of CVD, the WHO chest pain questionnaire and ECG findings. Data were also obtained from the Information Services Division (ISD) of National Health Service (NHS) Scotland on all medical and surgical discharge records from Scottish hospitals since 1981 and all ICD-10 codes for CVD were extracted. Using pre-defined criteria (13), these data were combined to assess prevalent CVD at baseline (MI, angina, transient ischaemic attack (TIA), stroke and coronary intervention).

Data on new CV events were collected four and eight years after recruitment. A combination of self-report and GP questionnaires plus ECG completed at year four, together with ISD record linkage for hospital discharge and death certificate data and review of clinical case notes at both four and eight years were used to define outcome events. Criteria for fatal and non-fatal events were as follows. MI: (1) ICD-10 code for new MI (I21-I23, I252) on discharge/death record, dated after baseline; codes confirmed by self-reported doctor diagnosis of MI, positive WHO chest pain questionnaire for MI, report of MI on GP questionnaire, or new ECG changes (for events by year 4) or by inspection of clinical notes (for events by year 8 where the relevant code was not a primary diagnosis or cause of death).

Angina: (1) ICD-10 code for angina (I20-I25) as primary diagnosis on hospital discharge record, dated after baseline, with no previous indication of angina; or (2) at least two of (a) self-reported doctor diagnosis of angina or new angina medication since baseline, (b) ECG codes for ischaemia that were not present at baseline and (c) positive WHO chest pain questionnaire. Fatal ischaemic heart disease (IHD): subject did not meet any of the criteria for fatal MI and had an ICD-10 code for IHD (I209, I249, I258, I259) as primary cause of death. Stroke: (1) ICD-10 code for stroke (I61, I63-I66, I679, I694) as primary diagnosis on discharge/death record, dated after baseline; or (2) self-report of stroke or non-primary ICD-10 discharge/death code for stroke dated after baseline, both confirmed on scrutiny of clinical notes. TIA: (1) ICD-10 code for TIA (G45, G659) as primary diagnosis on discharge record; or (2) self-report of stroke or non-primary ICD-10 discharge code for stroke or TIA dated after baseline, confirmed as TIA on scrutiny of clinical notes. Coronary intervention: OPCS operation code for coronary intervention (K40-K44, K49) on discharge record.

Statistical analysis

The distributions of ABI, NT-proBNP, hs-cTNT, Gamma-GT, TNF-α, CRP and IL-6 were skewed and therefore a log-transformation (using the natural logarithm) was used in all analyses. ABI has a reverse J-shaped relationship with CV risk and values greater than 1.4 measure medial arterial calcinosis rather than atherosclerosis, so in line with previous studies (29) participants with an ABI > 1 .4 (n=17) were omitted from analyses. For skewed biomarkers, medians with interquartile ranges (IQR) are given; all other continuous variables summarised using means and standard deviations (SD); categorical variables are given as total numbers with corresponding percentages. The Pearson correlation coefficient r and test of association were used to assess the relationships between the biomarkers.

The four inflammation biomarkers (TNF-α, CRP, IL-6 and fibrinogen) were combined into one general inflammation factor using an unrotated principal components analysis. All four markers loaded quite strongly onto the first principal component (0.44–0.80), which explained 49% of the total variability, and this was used to calculate the general inflammation factor, g.

An incident CV event was defined as the first CV event (fatal or non-fatal MI or stroke, fatal IHD, angina, TIA or coronary intervention) occurring after baseline. Baseline CV risk factors selected for model adjustment (age, sex, smoking, atrial fibrillation, rheumatoid arthritis, hypertension, CKD, BMI, sBP, total:HDL cholesterol and social status) were those included

in the QRISK2 score (2), except for family history of CV disease, which was not available in the ET2DS. The corresponding coefficients were estimated directly from the ET2DS data. Smoking was categorised as: non-smoker, ex-smoker, <10 cigarettes/day, 10-19 cigarettes/day and 20+ cigarettes/day; rheumatoid arthritis was recorded from a combination of self-report and linkage to ISD medical and surgical discharge records; atrial fibrillation (AF) was recorded if a subject self-reported use of digoxin, had the relevant hospital discharge code or AF was present on ECG; hypertension was defined as self-report of anti-hypertensive medication; CKD was defined as an estimated glomerular filtration rate (eGFR) <60ml/min on 2 of three consecutive measurements in the 12 to 24 months prior to baseline, to replicate doctor diagnosis of CKD used in QRISK2 Social status was categorised using the Scottish Index of Multiple Deprivation (SIMD), a composite index combining 38 indicators across seven domains (income, employment, health, education, skills and training, housing geographic access and crime), assigned according to post code (15). Baseline CVD status and lipid lowering medication were also included in the basic model, whereas in QRISK2, subjects with prior CVD or taking statins were excluded.

Binary logistic regression models were used to evaluate the relationships between each biomarker and CV events and results were summarised by odds ratios (OR) with corresponding 95% confidence intervals (CI) and p values. Logistic regression was chosen in favour of Cox regression to avoid invalid assumptions (proportional hazards) about the data, although a sensitivity analysis was carried out using Cox regression for the basic model and models incorporating the individual biomarker and the results were found to be consistent in terms of the statistical significance of individual hazards ratios and the sizes of the hazard ratio for each biomarker relative to the others. The added predictive value of including each biomarker in the model, over and above conventional predictors, was assessed. The C statistic was calculated for all models to provide a measure of model discrimination, the model's ability to distinguish between those who do or do not experience an event (ranging from 0.5, indicating no discriminative ability, to 1, indicating perfect discrimination). Corresponding 95% CIs are presented as an indication of statistical significance. The net reclassification index (NRI) was calculated, as well as the net reclassification (NR) separately for participants who did experience a CV event and those who did not. The NR compares two models (here, the basic model including only conventional CV risk factors and a new model incorporating one or more biomarkers) and gives the increase or decrease in the proportion of subjects correctly classified by the new model, according to pre-specified CV risk categories (0-10%,

10-20% and >20%). Calibration, the model's ability to correctly estimate the risk of a future event, was assessed using the Hosmer-Lemeshow test (null hypothesis assumes a well-calibrated model, therefore p-value > 0.05 indicates good calibration). All subsets regression was used to compare all possible combinations of biomarkers and obtain the best five models, according to a pre-specified statistical criterion (Akaike's Information Criterion (AIC) which measures the relative quality of a model while penalising for increasing numbers of predictors). In addition, a model was fitted which included conventional CV risk factors and the full panel of biomarkers.

A p value of <0.05 was taken to be statistically significant.

Results

Study characteristics at baseline and incident CV events

Due to low numbers of non-white participants (n=17), all analyses were restricted to Caucasian participants (n=1049; 515 women, 534 men). Mean age at baseline was 67.9 ± 2.4 years. Baseline prevalences of MI, angina, stroke, TIA and coronary intervention were 14.0% (n=147), 27.8% (n=292), 5.8% (n=61), 2.9% (n=30) and 10.1% (n=106) respectively. Full baseline characteristics of the study population, including median levels of ABI and circulating biomarkers, are shown in Table 1. A total of 205 first incident CV events (61 fatal/non-fatal MI, 38 angina, 53 stroke, 11 TIA, 24 coronary intervention and18 fatal IHD) occurred during the eight year follow-up period (19.5% of study population).

Associations between biomarkers at baseline

At baseline, moderate to strong relationships were observed between most of the biomarkers. In particular, the group of inflammatory markers (TNF-α, IL-6, CRP and fibrinogen) were positively correlated with each other (Table 2), and these associations were found to be strongly statistically significant (p<0.001). By design, the general inflammation factor g was strongly correlated with all inflammatory markers. Moderate correlations were found between NT-proBNP and both ABI and hs-cTnT (r = -0.21 and 0.38 respectively; both p<0.001). ABI and hs-cTNT were weakly negatively associated (r = -0.10, p<0.01). GGT correlated with three of the inflammatory markers (TNF-α, IL-6 and CRP; r = 0.08, 0.16, 0.24 respectively).

Adding individual biomarkers to the basic QRISK2 model

Increased levels of individual circulating biomarkers and the inflammatory factor were associated with an increased incidence of CV events over-and above the basic QRISK model, but only the associations for NT-proBNP and hs-cTnT were statistically significant (Table 3). The strongest association was observed for hs-cTnT (OR for 1 SD increase 1.35; 95% CI, 1.13, 1.61). A lower ABI was associated with a higher incidence of events (OR 0.86, 95% CI 0.73, 1.00).

The basic QRISK2 model had a C statistic of 0.722 (0.681, 0.763) and was well-calibrated (Hosmer-Lemeshow p=0.97) (Table 3). Addition of each individual biomarker increased the C statistic, with the greatest increases seen for hs-cTnT (C statistic increased by 0.01 from 0.722, 95% CI 0.681, 0.763 to 0.732, 95% CI 0.690, 0.774). Addition of individual biomarkers also improved the risk classification for participants who did not experience a CV event, although this generally resulted in poorer risk classification for participants who did experience a CV event (Table 3). The addition of hs-cTnT resulted in poorer risk classification by 1.6% for participants who experienced a CV event, but improved risk classification by 2.2% for those who did not. All the models were shown to be well-calibrated (Hosmer-Lemeshow $p > 0.05$).

Adding combinations of biomarkers to the basic model

An all subsets regression was carried out and identified the top five models according to a pre-specified statistical criterion, after adjusting for conventional risk factors, from all possible combinations of biomarkers. All five models (Table 3) included hs-cTnT and none included the general inflammation factor g. The best model selected using this method added ABI, hs-cTnT and GGT to the set of conventional CV risk factors. This model was well-calibrated and had a C statistic of 0.740 (0.699, 0.781), an increase of 0.018 compared to the basic model. The addition of the three biomarkers resulted in slightly poorer risk classification by 1.1% for participants who experienced a CV event, but improved risk classification by 4.4% for those who did not. The second best model was well-calibrated and showed the same increase in the C statistic as the top model, but the NR was poorer for participants who experienced a CV event (-2.7%s) and for those who did not (3.4%). For comparison, the full model including all biomarkers is also shown in Table 3. The C statistic showed the same increase as the top model. The addition of all biomarkers resulted in poorer risk classification by 1.6% for participants who experienced a CV event, but improved risk classification by 5.2% for those who did not.

Due to the high proportion of participants with prevalent CVD at baseline, removing these subjects from analysis resulted in poor statistical power (n=643 subjects, n=83 events). Despite this, the increase in c-statistic found on addition of the individual biomarkers to the basic QRISK2 model was in the same direction as in table 3 (though the size of the increase was, as expected, much smaller). For the best model of combined biomarkers (ABI, hs-cTnT and GGT), the c-statistic improved by a greater extent than for any of the individual biomarkers (from 0.680 for the basic model to 0.700), consistent with findings in the full study population.

Discussion

In older people with type 2 diabetes, higher levels of hs-cTnT and NT-proBNP were most strongly associated with increased risk of incident CV events, independent of factors currently used to predict CVD, and both improved predictive performance. NT-proBNP is released by the heart in response to increased pressure on the ventricular wall with low levels used in clinical practice to rule out heart failure (16), while cardiac troponin levels increase in response to clinical and subclinical myocardial ischaemia and is currently used to aid the diagnosis of myocardial infarction (17). In both general and diabetic population studies, NT-proBNP and hs-cTnT have been associated with the risk of CVD and may add predictive value independent of conventional risk factors (8; 9; 18; 19). The findings from our study, which enabled the comparison of a number of different biomarkers, are consistent with a very recent study in patients with type 2 diabetes in which, of a panel of 23 novel biomarkers, NT-proBNP was one of only three biomarkers which improved CV risk prediction beyond traditional risk factor (20).

In our study, ABI was negatively associated with CV risk, but the relationship was weaker than that for NT-proBNP or hs-cTnT. A reduced ABI (ratio of systolic blood pressure (BP) in the ankle to that in the arm) is used in the diagnosis of peripheral arterial disease and is a marker of generalized atherosclerosis (7). In 2008, a meta-analysis capturing over 480,000 person years follow-up suggested that in the general population, measuring ABI may improve CV risk prediction beyond the Framingham Risk Score (7). More recently, two general population studies indicated that ABI had a small effect on CV risk and only improved risk prediction if the basic model was weak (21; 22). Evidence on the ABI as a predictor of vascular events specifically in diabetes has previously been lacking, despite the known strong

association between diabetes and the development of peripheral arterial disease of the lower limbs.

There is previous evidence suggesting that a group of proteins used to assess levels of systemic inflammation (CRP, IL-6, TNF-α and fibrinogen) may add predictive value independent of conventional risk factors in both general and diabetic populations, but this evidence is inconsistent (11; 20; 23-27). Given that these four inflammatory markers are highly correlated, it has been suggested that they may best be combined into one general factor which describes the overall inflammatory burden (12; 28-30). This was the approach we chose for our study. However, when added individually to QRISK, the inflammation factor was not significantly associated with outcome CV events, although the C statistic did improve incrementally. Similarly, when the liver function test, GGT, was added on its own, there was no evidence of a statistically significant association, although again, the C statistic improved marginally. GGT has previously been associated with CVD in two large general population studies (31; 32) and in people with type 2 diabetes, although it did not improve CV prediction beyond traditional risk factors (33). Similarly, a recent general population cohort study of 2500 patients with acute coronary syndrome found that GGT was associated with increased risk of all-cause mortality but not cardiac mortality (34) and the PREVEND prospective cohort study suggested that adding GGT to conventional CV risk factors did not improve the prediction of first-ever CV events in the general population (35). Interestingly, although GGT on its own did not seem to add predictive value in the ET2DS, it was retained in the best combined model. This may indicate the importance of not pre-selecting biomarkers according to the statistical significance of any association with events prior to the inclusion of multiple biomarkers in a risk prediction model. Overall, a combination of risk factors improved risk prediction beyond that possible with any single biomarker, although an upper limit to model performance was suggested by the same C statistic value (0.740) for the two best combined models and the full model.

One of the strengths of the current study was the use of the risk score currently recommended for use in type 2 diabetes for risk prediction in the UK (the QRISK2 score, currently recommended by NICE clinical guidelines). However, replicating the QRISK2 score in the ET2DS proved challenging. Family history of CVD was not available in the ET2DS and the SIMD was used as a measure of social status rather than the Townsend index, which is only applicable to England and Wales. The definition of CKD in QRISK2 is a clinical diagnosis of CKD, but the list of corresponding clinical codes is not readily available. A similar doctor-

diagnosis definition of CKD created for the ET2DS only affected 1.7% of the cohort, much lower than anticipated in an elderly diabetic population (36; 37). A new variable for CKD, based on an eGFR <60ml/min (equivalent to Stage 3-5 CKD) identified 24.6% of the cohort and, as this was considered to be a more accurate definition of CKD, was used in subsequent analysis. Finally, QRISK2 excluded participants with previous CVD or taking statins. Because a very large proportion of the ET2DS had prevalent CVD at baseline or were taking lipid lowering medication (representing the situation in the target population of elderly people with type 2 diabetes) we included all subjects in our analyse, subsequently including prevalence of CVD and lipid lowering medication as additional covariates. Our model therefore has the advantage of being potentially applicable to all people with type 2 diabetes, including those both with and without clinically-diagnosed CVD, all of whom could benefit from more accurate CV risk prediction. However, whilst our sensitivity analysis of key results suggested that results were likely to have been similar in a study population free from CVD at baseline, future larger analysis should consider these groups of patients both combined and separately to be most informative.

The C statistic for the model including only traditional CV risk factors in this study was similar to those found by previous studies in type 2 diabetes, which used a variety of risk factors and/or CV risk scores as their basic model (9; 27; 38). The size of improvements in the C statistic following the addition of various biomarkers was also consistent with previous studies. Although the increases in C statistic were small, it should be noted that the C statistic can be insensitive when adding a new predictor to a model, even though such a predictor may make an independent and statistically significant contribution to the model (39). This phenomenon is particularly noticeable when the baseline model includes strong predictors and has a large C statistic. In order to evaluate the clinical usefulness of our models, we also considered the NR as a measure of reclassification. This suggested that, in general, the risk classification improved after the addition of a biomarker for people who did not experience a CV event, but slightly worsened for people who did experience an event. Further large studies are needed to validate this conclusion and to ascertain whether any improvements are clinically significant.

This study benefited from the representativeness of the type 2 diabetes population, the relatively long term follow up for CV events and the thorough and systematic approach for assessing incident CV events which ensured loss to follow-up was minimal. The wide variety of biomarkers available allowed for the inclusion of a large panel of potential predictors, both

individually and in combination. The study also has limitations. In addition to the insensitivity of the C statistic, the NR is dependent on the choice of risk thresholds. The continuous net reclassification index can be used to avoid this decision, but this is less clinically relevant. The NR should therefore be considered as a descriptive tool to demonstrate what would happen to risk scores in a clinical setting if the new model was used with the chosen risk categories (0-10%, 10-20%, >20%). Data were missing for some of the predictor variables and the complete case analysis performed can produce biased estimates or reduce statistical power. However, since missing data was less than 5%, and an analysis of subjects with missing data versus those without indicated that missing data rates did not depend on the outcome or key predictor variables, these risks were considered to be negligible.

In general, previous studies into biomarkers have focused on the general population and it often remains uncertain whether they contribute to risk prediction in a diabetic population. In particular, the potential value of one biomarker compared with others in the same study population has rarely been addressed. In attempting to address this gap in knowledge, we have shown that hs-cTnT, NT-proBNP and ABI, but not GGT or an inflammatory factor on their own, are able to independently improve CV risk prediction beyond traditional risk factors in patients with type 2 diabetes. Of these, NT-proBNP and hs-cTnT appeared to be the most promising biomarkers, in terms of the extent to which they improve prediction when added individually, but ABI has the advantage of not requiring a blood test. Our results also indicate that a combination of biomarkers results in further improvement to risk prediction compared with one strong biomarker alone, and that biomarkers which on their own may not appear to add predictive value, may do so when added in combination with others. Future studies should explore the balance between the clinical benefit of adding multiple biomarkers to a risk score versus the cost of doing so.

Author Contributions

JFP and MWJS conceived and designed the ET2DS and oversaw the acquisition and analysis of data. For the current paper, AHP, JFP and CW conceived the idea and designed the analysis, which was performed by AHP. AHP, JFP and CW wrote the paper. All authors contributed to data collection, interpretation of findings and preparation of the final manuscript, including commenting on the final draft. JFP and AHP are the guarantors of this work, and as such had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. The authors thank all patients and research staff involved in the ET2DS.

Acknowledgements

The sponsor for the ET2DS was the University of Edinburgh. The study was funded by the Medical Research Council (UK), the Chief Scientist Office of the Scottish Executive, Pfizer plc. and Diabetes UK. The funders had no other role in the design, analysis or writing of this manuscript. CJW was supported in this work by NHS Lothian via the Edinburgh Clinical Trials Unit.

No potential conflicts of interest relevant to this article were reported.

Study participants gave informed consent and ethical permission was obtained from the Lothian Medical Research Ethics Committee.

References

1. The Emerging Risk Factors C: Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. Lancet 2010;375:2215-2222
2. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, Brindle P: Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ (Clinical research ed) 2008;336:1475-1482
3. Hippisley-Cox J, Coupland C, Brindle P: Validation of QRISK2 (2014) in patients with diabetes. http://www.qrisk.org/index.php, QRISK, 2014
4. Simmons RK, Coleman RL, Price HC, Holman RR, Khaw KT, Wareham NJ, Griffin SJ: Performance of the UK Prospective Diabetes Study Risk Engine and the Framingham Risk Equations in Estimating Cardiovascular Disease in the EPIC- Norfolk Cohort. Diabetes care 2009;32:708-713
5. van der Heijden AA, Ortegon MM, Niessen LW, Nijpels G, Dekker JM: Prediction of coronary heart disease risk in a general, pre-diabetic, and diabetic population during 10 years of follow-up: accuracy of the Framingham, SCORE, and UKPDS risk functions: The Hoorn Study. Diabetes care 2009;32:2094-2098
6. van Dieren S, Beulens JW, Kengne AP, Peelen LM, Rutten GE, Woodward M, van der Schouw YT, Moons KG: Prediction models for the risk of cardiovascular disease in patients with type 2 diabetes: a systematic review. Heart (British Cardiac Society) 2012;98:360-369
7. Fowkes FG, Murray GD, Butcher I, Heald CL, Lee RJ, Chambless LE, Folsom AR, Hirsch AT, Dramaix M, deBacker G, Wautrecht JC, Kornitzer M, Newman AB, Cushman M, Sutton-Tyrrell K, Fowkes FG, Lee AJ, Price JF, d'Agostino RB, Murabito JM, Norman PE, Jamrozik K, Curb JD, Masaki KH, Rodriguez BL, Dekker JM, Bouter LM, Heine RJ, Nijpels G, Stehouwer CD, Ferrucci L, McDermott MM, Stoffers HE, Hooi JD, Knottnerus JA, Ogren M, Hedblad B, Witteman JC, Breteler MM, Hunink MG, Hofman A, Criqui MH, Langer RD, Fronek A, Hiatt WR, Hamman R, Resnick HE, Guralnik J, McDermott MM: Ankle brachial index combined with Framingham Risk Score to predict cardiovascular events and mortality: a meta-analysis. JAMA : the journal of the American Medical Association 2008;300:197-208
8. Welsh P, Doolin O, Willeit P, Packard C, Macfarlane P, Cobbe S, Gudnason V, Di Angelantonio E, Ford I, Sattar N: N-terminal pro-B-type natriuretic peptide and the prediction of primary cardiovascular events: results from 15-year follow-up of WOSCOPS. European heart journal 2013;34:443-450
9. Hillis GS, Welsh P, Chalmers J, Perkovic V, Chow CK, Li Q, Jun M, Neal B, Zoungas S, Poulter N, Mancia G, Williams B, Sattar N, Woodward M: The relative and combined ability of high-sensitivity cardiac troponin T and N-terminal pro-B-type natriuretic peptide to predict cardiovascular events and death in patients with type 2 diabetes. Diabetes care 2014;37:295-303
10. Kengne AP, Czernichow S, Stamatakis E, Hamer M, Batty GD: Gamma-glutamyltransferase and risk of cardiovascular disease mortality in people with and without diabetes: pooling of three British Health Surveys. Journal of Hepatology 2012;57:1083-1089
11. Emerging Risk Factors C, Kaptoge S, Di Angelantonio E, Pennells L, Wood AM, White IR, Gao P, Walker M, Thompson A, Sarwar N, Caslake M, Butterworth AS, Amouyel P, Assmann G, Bakker SJ, Barr EL, Barrett-Connor E, Benjamin EJ, Bjorkelund C, Brenner H, Brunner E, Clarke R, Cooper JA, Cremer P, Cushman M, Dagenais GR, D'Agostino RB, Sr., Dankner R, Davey-Smith G, Deeg D, Dekker JM, Engstrom G, Folsom AR, Fowkes FG, Gallacher J, Gaziano JM, Giampaoli S, Gillum RF, Hofman A, Howard BV, Ingelsson E, Iso H, Jorgensen T, Kiechl S, Kitamura A, Kiyohara Y, Koenig W, Kromhout D, Kuller LH, Lawlor DA, Meade TW, Nissinen A, Nordestgaard BG, Onat A, Panagiotakos DB, Psaty

BM, Rodriguez B, Rosengren A, Salomaa V, Kauhanen J, Salonen JT, Shaffer JA, Shea S, Ford I, Stehouwer CD, Strandberg TE, Tipping RW, Tosetto A, Wassertheil-Smoller S, Wennberg P, Westendorp RG, Whincup PH, Wilhelmsen L, Woodward M, Lowe GD, Wareham NJ, Khaw KT, Sattar N, Packard CJ, Gudnason V, Ridker PM, Pepys MB, Thompson SG, Danesh J: C-reactive protein, fibrinogen, and cardiovascular disease prediction. New England Journal of Medicine 2012;367:1310-1320
12. Kumar RG, Rubin JE, Berger RP, Kochanek PM, Wagner AK: Principal components derived from CSF inflammatory profiles predict outcome in survivors after severe traumatic brain injury. Brain, behavior, and immunity 2016;53:183-193
13. Marioni RE, Strachan MW, Reynolds RM, Lowe GD, Mitchell RJ, Fowkes FG, Frier BM, Lee AJ, Butcher I, Rumley A, Murray GD, Deary IJ, Price JF: Association between raised inflammatory markers and cognitive decline in elderly people with type 2 diabetes: the Edinburgh Type 2 Diabetes Study. Diabetes 2010;59:710-713
14. Price JF, Reynolds RM, Mitchell RJ, Williamson RM, Fowkes FG, Deary IJ, Lee AJ, Frier BM, Hayes PC, Strachan MW: The Edinburgh Type 2 Diabetes Study: study protocol. BMC endocrine disorders 2008;8:18
15. SIMD Scottish Index of Multiple Deprivation [article online], 2012. Available from http://simd.scotland.gov.uk/publication-2012/. Accessed 10/08/2016
16. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JGF, Coats AJS, Falk V, González-Juanatey JR, Harjola V-P, Jankowska EA, Jessup M, Linde C, Nihoyannopoulos P, Parissis JT, Pieske B, Riley JP, Rosano GMC, Ruilope LM, Ruschitzka F, Rutten FH, van der Meer P, Filippatos G, McMurray JJV, Aboyans V, Achenbach S, Agewall S, Al-Attar N, Atherton JJ, Bauersachs J, John Camm A, Carerj S, Ceconi C, Coca A, Elliott P, Erol Ç, Ezekowitz J, Fernández-Golfín C, Fitzsimons D, Guazzi M, Guenoun M, Hasenfuss G, Hindricks G, Hoes AW, Iung B, Jaarsma T, Kirchhof P, Knuuti J, Kolh P, Konstantinides S, Lainscak M, Lancellotti P, Lip GYH, Maisano F, Mueller C, Petrie MC, Piepoli MF, Priori SG, Torbicki A, Tsutsui H, van Veldhuisen DJ, Windecker S, Yancy C, Zamorano JL, Zamorano JL, Aboyans V, Achenbach S, Agewall S, Badimon L, Barón-Esquivias G, Baumgartner H, Bax JJ, Bueno H, Carerj S, Dean V, Erol Ç, Fitzsimons D, Gaemperli O, Kirchhof P, Kolh P, Lancellotti P, Lip GYH, Nihoyannopoulos P, Piepoli MF, Ponikowski P, Roffi M, Torbicki A, Vaz Carneiro A, Windecker S, Sisakian HS, Isayev E, Kurlianskaya A, Mullens W, Tokmakova M, Agathangelou P, Melenovsky V, Wiggers H, Hassanein M, Uuetoa T, Lommi J, Kostovska ES, Juillière Y, Aladashvili A, Luchner A, Chrysohoou C, Nyolczas N, Thorgeirsson G, Marc Weinstein J, Di Lenarda A, Aidargaliyeva N, Bajraktari G, Beishenkulov M, Kamzola G, Abdel-Massih T, Čelutkienė J, Noppe S, Cassar A, Vataman E, Abir-Khalil S, van Pol P, Mo R, Straburzyńska-Migaj E, Fonseca C, Chioncel O, Shlyakhto E, Otasevic P, Goncalvesová E, Lainscak M, Díaz Molina B, Schaufelberger M, Suter T, Yılmaz MB, Voronkov L, Davies C: 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC)
Developed with the special contribution of the Heart Failure Association (HFA) of the ESC 2015;
17. NICE CG95: Chest pain of recent onset: assessment and diagnosis. 2010;
18. Saunders JT, Nambi V, de Lemos JA, Chambless LE, Virani SS, Boerwinkle E, Hoogeveen RC, Liu X, Astor BC, Mosley TH, Folsom AR, Heiss G, Coresh J, Ballantyne CM: Cardiac troponin T measured by a highly sensitive assay predicts coronary heart disease, heart failure, and mortality in the Atherosclerosis Risk in Communities Study. Circulation 2011;123:1367-1376

19. Welsh P, Hart C, Papacosta O, Preiss D, McConnachie A, Murray H, Ramsay S, Upton M, Watt G, Whincup P, Wannamethee G, Sattar N: Prediction of Cardiovascular Disease Risk by Cardiac Biomarkers in 2 United Kingdom Cohort Studies: Does Utility Depend on Risk Thresholds For Treatment? Hypertension 2016;67:309-315
20. Wannamethee SG, Welsh P, Lowe GD, Gudnason V, Di Angelantonio E, Lennon L, Rumley A, Whincup PH, Sattar N: N-terminal pro-brain natriuretic Peptide is a more useful predictor of cardiovascular disease risk than C-reactive protein in older men with and without pre-existing cardiovascular disease. J Am Coll Cardiol 2011;58:56-64
21. Murphy TP, Dhangana R, Pencina MJ, D'Agostino Sr RB: Ankle–brachial index and cardiovascular risk prediction: An analysis of 11,594 individuals with 10-year follow-up. Atherosclerosis 2012;220:160-167
22. Fowkes F, Murray G, Butcher I, Folsom A, Hirsch A, Couper D, DeBacker G, Kornitzer M, Newman A, Sutton-Tyrrell K, Cushman M, Lee A, Price J, D'Agostino R, Murabito J, Norman P, Masaki K, Bouter L, Heine R, Stehouwer C, McDermott M, Stoffers H, Knottnerus J, Ogren M, Hedblad B, Koenig W, Meisinger C, Cauley J, Franco O, Hunink M, Hofman A, Witteman J, Criqui M, Langer R, Hiatt W, Hamman R, Collaboration ABI: Development and validation of an ankle brachial index risk model for the prediction of cardiovascular events. European Journal of Preventive Cardiology 2014;21:310-320
23. Berg AH, Scherer PE: Adipose tissue, inflammation, and cardiovascular disease. Circ Res 2005;96:939-949
24. Bettencourt N, Oliveira S, Toschke AM, Rocha J, Leite D, Carvalho M, Xara S, Schuster A, Chiribiri A, Leite-Moreira A, Nagel E, Alves H, Gama V: Predictors of circulating endothelial progenitor cell levels in patients without known coronary artery disease referred for multidetector computed tomography coronary angiography. Revista Portuguesa de Cardiologia 2011;30:753-760
25. Olsen MH, Hansen TW, Christensen MK, Gustafsson F, Rasmussen S, Wachtell K, Ibsen H, Torp-Pedersen C, Hildebrandt PR: N-terminal pro-brain natriuretic peptide, but not high sensitivity C-reactive protein, improves cardiovascular risk prediction in the general population. European heart journal 2007;28:1374-1381
26. Kengne AP, Czernichow S, Stamatakis E, Hamer M, Batty GD: Fibrinogen and future cardiovascular disease in people with diabetes: aetiological associations and risk prediction using individual participant data from nine community-based prospective cohort studies. Diabetes & Vascular Disease Research 2013;10:143-151
27. van der Leeuw J, Beulens JW, van Dieren S, Schalkwijk CG, Glatz JF, Hofker MH, Verschuren WM, Boer JM, van der Graaf Y, Visseren FL, Peelen LM, van der Schouw YT: Novel Biomarkers to Improve the Prediction of Cardiovascular Event Risk in Type 2 Diabetes Mellitus. J Am Heart Assoc 2016;5
28. Alman AC, Kinney GL, Tracy RP, Maahs DM, Hokanson JE, Rewers MJ, Snell-Bergeon JK: Prospective Association Between Inflammatory Markers and Progression of Coronary Artery Calcification in Adults With and Without Type 1 Diabetes. Diabetes care 2013;36:1967-1973
29. Hsu F-C, Kritchevsky SB, Liu Y, Kanaya A, Newman AB, Perry SE, Visser M, Pahor M, Harris TB, Nicklas BJ, Study ftHA: Association Between Inflammatory Components and Physical Function in the Health, Aging, and Body Composition Study: A Principal Component Analysis Approach. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences 2009;64A:581-589
30. Bedenis R, Price AH, Robertson CM, Morling JR, Frier BM, Strachan MW, Price JF: Association Between Severe Hypoglycemia, Adverse Macrovascular Events, and Inflammation in the Edinburgh Type 2 Diabetes Study. Diabetes care 2014;

31. Lee DH, Silventoinen K, Hu G, Jacobs DR, Jr., Jousilahti P, Sundvall J, Tuomilehto J: Serum gamma-glutamyltransferase predicts non-fatal myocardial infarction and fatal coronary heart disease among 28,838 middle-aged men and women. European heart journal 2006;27:2170-2176
32. Jousilahti P, Rastenyte D, Tuomilehto J: Serum gamma-glutamyl transferase, self-reported alcohol drinking, and the risk of stroke. Stroke 2000;31:1851-1855
33. Kengne AP, Batty GD, Hamer M, Stamatakis E, Czernichow S: Association of C-reactive protein with cardiovascular disease mortality according to diabetes status: pooled analyses of 25,979 participants from four U.K. prospective cohort studies. Diabetes care 2012;35:396-403
34. Ndrepepa G, Braun S, Cassese S, Fusaro M, Laugwitz KL, Schunkert H, Kastrati A: Relation of Gamma-Glutamyl Transferase to Cardiovascular Events in Patients With Acute Coronary Syndromes. The American journal of cardiology 2016;117:1427-1432
35. Kunutsor SK, Bakker SJ, Kootstra-Ros JE, Gansevoort RT, Dullaart RP: Circulating gamma glutamyltransferase and prediction of cardiovascular disease. Atherosclerosis 2015;238:356-364
36. Retnakaran R, Cull CA, Thorne KI, Adler AI, Holman RR: Risk factors for renal dysfunction in type 2 diabetes: U.K. Prospective Diabetes Study 74. Diabetes 2006;55:1832-1839
37. Koro CE, Lee BH, Bowlin SJ: Antidiabetic medication use and prevalence of chronic kidney disease among patients with type 2 diabetes mellitus in the United States. Clinical Therapeutics 2009;31:2608-2617
38. 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 (Clinical research ed) 2013;346:f2573
39. Cook NR: Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. Clinical chemistry 2008;54:17-23

Table 1: Baseline characteristics of the ET2DS population	
Variable	
Age (years)	67.9 ± 4.2
Sex (female)	515 (49.1)
Lipid-lowering medication	896 (85.4)
Hypertension	858 (81.8)
Smoking status	
Non-smoker	411 (39.2)
Ex-smoker	491 (46.8)
Current smoker – light (<10 cigarettes/day)	31 (3.0)
Current smoker – moderate (10-19 cigarettes/day)	47 (4.5)
Current smoker – heavy (20+ cigarettes/day)	69 (6.6)
Atrial fibrillation	69 (6.6)
Chronic kidney disease	258 (24.6)
Rheumatoid arthritis	39 (3.7)
SIMD	
Quintile 1 (most deprived)	127 (12.1)
Quintile 2	205 (19.5)
Quintile 3	185 (17.6)
Quintile 4	192 (18.3)
Quintile 5 (least deprived)	340 (32.4)
BMI (kg/m^2)	31.5 ± 5.7
sBP (mmHg)	133.3 ± 16.5
Total cholesterol : HDL cholesterol	3.5 ± 1.1
CVD at baseline[a]	
MI	147 (14.0)
Angina	292 (27.8)
Stroke	61 (5.8)
TIA	30 (2.9)
CI	106 (10.1)
ABI	1.0 (0.9, 1.1)
NT-proBNP (pg/ml)	76 (38, 172)
hs-cTnT (ng/l)	9.6 (6.9, 13.8)
GGT (U/L)	18 (11, 32)
TNF-α (pg/ml)	1.1 (0.7, 1.6)
IL-6 (pg/ml)	2.9 (2.0, 4.5)
CRP (mg/l)	1.9 (0.9, 4.4)
Fibrinogen (g/L)	3.6 ± 0.7

Data are presented as means ± SD, *n* (%) or median (lower IQR, upper IQR)
[a] Note that there is overlap among these subgroups
Maximum *n* = 1049

Table 2: Correlation coefficients between biomarkers at baseline (max n = 1032)[a]									
	ABI < 1.4	**NT-proBNP**	**hs-cTnT**	**GGT**	**TNF-α**	**IL-6**	**CRP**	**Fibrinogen**	**g**
ABI < 1.4	1	-0.21***	-0.10**	-0.05	-0.07*	-0.12***	-0.13***	-0.18***	-0.18***
NT-proBNP		1	0.38***	-0.03	0.16***	0.18***	0.11***	0.21***	0.23***
hs-cTnT			1	0.03	0.19***	0.17***	-0.03	0.05	0.12***
GGT				1	0.08*	0.16***	0.24***	-0.06	0.16***
TNF-α					1	0.31***	0.12***	0.12***	0.43***
IL-6						1	0.42***	0.34***	0.75***
CRP							1	0.54***	0.80***
Fibrinogen								1	0.76***
g									1

[a] Missing data ranges from 8 to 39 data points
* Pearson correlation test p-value < 0.05
** Pearson correlation test p-value < 0.01
*** Pearson correlation test p-value <0.001

Table 3: Adding biomarkers to the basic model (n = 989)*

Model	Predictors in the model, additional to conventional risk factors[a]	OR for a one SD increase in biomarker (95% CI)	C statistic (95% CI)	NR – event[b] (%)	NR – no event[b] (%)	NRI	Hosmer-Lemeshow p value
Basic model		-	0.722 (0.681, 0.763)	-	-	-	0.97
+ ABI < 1.4	ABI	0.86 (0.73, 1.00)	0.725 (0.684, 0.766)	-2.2	2.0	0.015	0.83
+ NT-proBNP	NT-proBNP	1.23 (1.02, 1.49)	0.726 (0.685, 0.767)	-2.2	1.5	-0.007	0.81
+ Troponin	hs-cTnT	1.35 (1.13, 1.61)	0.732 (0.690, 0.774)	-1.6	2.2	0.006	0.09
+ Gamma-GT	GGT	1.16 (0.98, 1.37)	0.726 (0.685, 0.766)	-2.7	1.1	-0.016	0.40
+ *g*	*g*	1.07 (0.90, 1.27)	0.724 (0.683, 0.765)	0.5	1.2	0.018	0.90
Top five models chosen using all-subsets regression selection							
1	ABI + hs-cTnT + GGT	-	0.740 (0.699, 0.781)	-1.1	4.4	0.033	0.15
2	ABI + hs-cTnT + GGT + NT-proBNP	-	0.740 (0.699, 0.780)	-2.7	3.5	0.008	0.34
3	hs-cTnT + GGT + NT-proBNP	-	0.738 (0.697, 0.779)	-1.6	5.1	0.035	0.47
4	ABI + hs-cTnT	-	0.735 (0.694, 0.776)	-3.2	5.4	0.021	0.35
5	hs-cTnT + GGT	-	0.738 (0.697, 0.778)	-1.1	3.9	0.028	0.21
Full model							
	ABI + hs-cTnT + GGT + NT-proBNP + *g*	-	0.740 (0.699, 0.781)	-1.6	5.2	0.036	0.39

*A complete case analysis was carried out, n = 989

[a] Conventional risk factors: age, sex, smoking, atrial fibrillation, chronic kidney disease, arthritis, hypertension, BMI, sBP, total:HDL cholesterol, social status, baseline CVD status (MI, angina, TIA and stroke) and lipid lowering medication

[b] n = 186 for event, n = 803 for no event

European Association for the Study of Diabetes Annual Meeting 2016:

Adding novel biomarkers to current cardiovascular risk scores for people with Type 2 diabetes: the Edinburgh Type 2 Diabetes Study (ET2DS)

AH Price[1], C Weir[1], MWJ Strachan[2], N Sattar[3], S McLachlan[1] and JF Price[1]

[1]*Centre for Population Health Sciences, University of Edinburgh, Edinburgh, UK,* [2] *Metabolic Unit, Western General Hospital, Edinburgh, UK,* [3] *Glasgow Cardiovascular Research Centre, University of Glasgow, Glasgow, UK*

Background and aims: Increasing evidence suggests novel biomarkers may improve cardiovascular (CV) risk prediction in the general population. Whether they could improve current CV risk scores in people with Type 2 diabetes is uncertain.

Materials and methods: Conventional CV risk factors and novel biomarkers were measured in 1066 adults (48.7% female) aged 60-74 years with Type 2 diabetes participating in the population-based ET2DS. Seven year follow-up for incident CV events used clinical examination, hospital admission record and death certificate linkage. Predictors in the QRISK2 cardiovascular risk score (age, sex, smoking, atrial fibrillation, chronic kidney disease, rheumatoid arthritis, hypertension, BMI, sBP and total:HDL cholesterol) were considered for the basic model, which was also adjusted for prevalent CV disease and lipid-lowering drugs. The following novel biomarkers were then added to this starting model individually to assess their added value to the model: Ankle Brachial Index (ABI), N-terminal pro brain natriuretic peptide (NT-proBNP), troponin, gamma-glutamyl transpeptidase (Gamma-GT) and an inflammatory factor (*g*) combining C-reactive protein, interleukin-6, tumor necrosis factor alpha and fibrinogen using principal components analysis.

Results: 208 (19.5%) subjects had an incident CV event (first fatal or non-fatal myocardial infarction, new onset angina, fatal or non-fatal stroke, transient ischaemic attack or other fatal ischaemic hear disease). Results showed baseline ABI, NT-proBNP and troponin were significantly associated with risk of CV events over-and-above QRISK2 (odds ratios for 1 standard deviation increase in biomarker 0.81 (95% CI 0.69, 0.96), 1.24 (1.02, 1.52) and 1.48 (1.22, 1.80) respectively). No significant association was found for Gamma-GT and *g*. C statistics improved from 0.729 (basic model) to 0.735, 0.731, 0.745 and 0.730 for ABI, NT-proBNP, troponin and Gamma-GT respectively and all models had good calibration, assessed using the Hosmer-Lemeshow goodness-of-fit test (p-values all > 0.05). Only troponin provided a net improvement in correctly reclassifying people who both did and did not experience a CV event (net reclassification improvement for people who did suffer an event was 1.5% and net reclassification improvement for people who did not suffer an event was 3.8%).

Conclusion: These preliminary results suggest moderate potential for selected novel biomarkers to add value to current CV risk scores. Further investigation will be carried out on the biomarkers in combination to discover the panel of biomarkers which best predict CV disease.

Table 1: adding novel biomarkers to the basic model, with model summary measures						
	Basic model	**+ ABI**	**+ NT-pro BNP**	**+ Trop-onin**	**+ Gamma-GT**	**+ *g***
OR for a 1 SD increase in biomarker (95% CI)	-	0.81 (0.69, 0.96)	1.24 (1.02, 1.52)	1.48 (1.22, 1.80)	1.18 (0.99, 1.39)	1.07 (0.89, 1.29)
Model summary measures:						
C statistic	0.729	0.735	0.731	0.745	0.730	0.725
Net reclassification improvement – event (%)	-	-1.0	-1.5	1.5	-0.5	1.0
Net reclassification improvement – no event (%)	-	2.7	0.6	3.8	1.6	-0.6
Hosmer-Lemeshow p-value	0.878	0.636	0.974	0.269	0.070	0.334

Diabetes Professional Conference 2016:

Improving cardiovascular (CV) risk scores with novel biomarkers in people with Type 2 diabetes: the Edinburgh Type 2 Diabetes Study (ET2DS)

AH Price[1], C Weir[1], MWJ Strachan[2], N Sattar[3], S McLachlan[1], CM Robertson[1]and JF Price[1]

[1]C*entre for Population Health Sciences, University of Edinburgh, Edinburgh, UK,* [2] *Metabolic Unit, Western General Hospital, Edinburgh, UK,* [3] *Glasgow Cardiovascular Research Centre, University of Glasgow, Glasgow, UK*

Aims: Increasing evidence suggests novel biomarkers may improve CV risk prediction in the general population. Whether they could improve current CV risk scores in people with Type 2 diabetes is uncertain.

Methods: Conventional CV risk factors and novel biomarkers were measured in 1,066 adults (48.7% female) aged 60□74 years with Type 2 diabetes participating in the population-based ET2DS. Seven year follow-up for incident CV events used clinical examination and hospital admission record and death certificate linkage. Predictors in the QRISK2 score (age, sex, ethnicity, smoking, atrial fibrillation, chronic kidney disease, rheumatoid arthritis, hypertension, body mass index, systolic blood pressure and total:high-density lipoprotein cholesterol) were considered for the starting model, which was also adjusted for prevalent CV disease and lipid-lowering drugs.

Results: 208 (19.5%) subjects had an incident CV event (first fatal or non-fatal myocardial infarction, new onset angina, fatal or non- fatal stroke, transient ischaemic attack or other fatal ischaemic heart disease). A more accurate definition of chronic kidney disease and incorporating social class produced slight improvements in model performance. Initial results showed baseline NT-proBNP and troponin were significantly associated with risk of CV events over and above QRISK2 [hazard ratios for one standard deviation increase in biomarker 1.25 (95% confidence interval 1.05, 1.49) and 1.37 (1.20, 1.60) respectively]; c indexes improved from 0.709 (starting model) to 0.712 and 0.725 for NT-proBNP and troponin respectively. No significant association was found for C-reactive protein.

Conclusions: These preliminary results suggest the potential for selected novel biomarkers to add value to current CV risk scores. Further investigation will be carried out on a wider range of novel biomarkers, in combination, in the ET2DS.

Appendix B PAC application form

NHS National Services Scotland Privacy Advisory Committee

Application for Privacy Advisory Committee Approval

Guidance

Introduction

THE PRIVACY ADVISORY COMMITTEE (PAC) is an advisory committee to NHS National Services Scotland (NSS) and the Registrar General. The PAC advises on the correct balance between protecting personal data and making data available for research, audit and other important uses and ensures that any information releases are carefully controlled.

NSS follows the Proportionate Governance Approach favoured by Scottish Government (Joined up Data for better Decisions 2012). NSS Information Governance Team assesses all applications for access to data under the control of NSS or National Records of Scotland (NRS) that have the potential to be person-identifiable, and in respect of any new record linkages. The views of PAC are sought in relation to requests which the team assess as carrying high privacy or other risk. Further information is available on the PAC Website.

Where the data you wish to access are not in the control of NSS or NRS you may need to consider applying to another advisory body - see 'Appendix One: Advisory Bodies within NHSScotland'.

Assistance with PAC

NSS Research Coordinators are a team of analysts and data specialists based within the Information Services Division (ISD) of NSS. The NSS Research Coordinators can explain what information is available and help you to decide which variables would be useful for your study. They can help you to define your requirements for data linkage, preparation of data extracts, and any analyses required. They can also advise when you may need to approach another advisory body.

When to Complete a Privacy Advisory Application Form

The PAC Application Form must be completed for data requests that involve:

access to identifiable or potentially identifiable information;

circumstances where NSS or National Records of Scotland (NRS) have indicated their intention to seek guidance from PAC; and/or

record linkage of previously unlinked datasets involving data from more than one Health Board.

Our aim is to make data for research available through the ScottisH Informatics Programme (SHIP) infrastructure. All applicants to PAC will be expected to use the SHIP National Safe Haven as a means of accessing data except in very exceptional circumstances.

ScottisH Informatics Programme (SHIP)

SHIP is a Scotland-wide research platform for the collation, management, dissemination and analysis of Electronic Patient Records (EPRs). The programme brings together the Universities of Dundee, Edinburgh, Glasgow and St Andrews with the Information Services Division (ISD) of NSS.

The SHIP programme will provide a platform for Scottish record linkage that will drive EPR research throughout the UK and abroad. The SHIP National Safe Haven is the analytical platform for this. This includes the following.

Provision of a record linkage service where personal identifiers are kept separate from the payload/content data.

Provision of a secure environment for researchers to analyse anonymised patient level or summarised records.

Provision of a Secure File Transfer service to support the transmission of data between data providers and researchers.

Anyone wishing to access data through the SHIP National Safe Haven requires to be a member of an appropriate institution and to be able to demonstrate having successfully completed approved training in Information Governance within the last three years.

Research coordinators hold a list of SHIP approved training courses. Anyone proposing to or having completed another course should provide our Research Coordinator with details including the name of the course, the name of the institution providing the training, and the content of the course.

Further Advice on National Records of Scotland Service and NHS Central Registry

To find out more about NRS data including those relating to the NHS Central Registry visit http://www.gro-scotland.gov.uk/national-health-service-central-register/index.html or e-mail: dumf-uhb.NHSCR-Scotland-Medical-Research@nhs.netmailto:dumf-uhb.NHSCR-Scotland-Medical-Research@nhs.net

For All Other Requests

For all other requests or advice please contact the NSS Research Coordinators at nss.eDRIS@nhs.net

Application Checklist

Before you submit your application, you should include the following items and ensure that the application has been signed by the appropriate individuals.

Your application should be typed, not handwritten.

Items to support application

Where applicable, you should include the following:

Study protocol

Information provided to study participants and/or the wider public

Participant consent forms

Draft correspondence (if the data generated through your study will be used to contact any individuals)

Evidence of ethical approval

Evidence of approval from other Data Controllers eg Caldicott Guardians or CHIAG

Local Information Governance/security policies and procedures (if you are not using the SHIP National Safe Haven)

The list of variables you require in the file for analysis (if not included under question 10)

Details of each individual accessing the data (where there are more than the five allowed on this form)

Content of Information Governance training course undertaken if it is not already SHIP approved.

Please ensure that your application has been signed by the:

main study contact

study information custodian (if you are not using the SHIP National Safe Haven).

You must also ensure that individuals named on the form have read and approved this submission.

After completing your application form you must save it as a PDF file before sending it with relevant appendices by email to your research coordinator. If you do not have a research coordinator please email nss.eDRIS@nhs.net. The research coordinators/eDRIS will forward your application to PAC after checking it.

Application for

Privacy Advisory Committee Approval

Application Title	**Third Wave of Record Linkage for the Edinburgh Type 2 Diabetes Study**
Date Submitted	**5/11/14**
NSS Research Co-ordinator Name	**Rose Sisk**
NSS Study Number (Your NSS Research Coordinator will provide this information)	**XRB14155**

Please Note:

The information contained in this application form will be regarded as confidential whilst it goes through the scrutiny process but it is the duty of applicants to point out any information within the application that they consider to be particularly sensitive, confidential or commercially sensitive.

As NHS National Services Scotland is a public authority, it is subject to the Freedom of Information (Scotland) Act.

PAC applications are kept by NHS National Services Scotland for a minimum period of 15 years from date of application or date of the last linkage undertaken in relation to the application.

▪ People Involved

The names of all the individuals involved in and responsible for the design and analysis of the study should be included here. It is expected that those who will have access to the data supplied by NSS have adequate and regular updated knowledge and skills in the secure and confidential handling of health data.

You must ensure that everyone who will access the data provided is a member of an appropriate institution and is able to demonstrate having successfully completed approved training in Information Governance within the last three years.

Research coordinators hold a list of SHIP approved training courses. If you propose to or have completed another course, please provide your Research Coordinator with details including the name of the course, the name of the institution providing the training, and the content of the course.

You must ensure that all staff taking part in this study have appropriate contracts in place containing clauses that clearly identify their duties and responsibilities for confidentiality, data protection, and data security.

You do not need to include clerical and secretarial support staff.

○ **Head of Department responsible for project/study or the Principal Investigator** *This should be the name of the person who will take overall responsibility for the study*	
Title	Professor
Forename	Jackie
Surname	Price
Position	Professor of Molecular Epidemiology
Qualifications	BSc (Hons), MBChB, MD, FFPHM
Professional Registration Number (if applicable) eg General Medical Council (GMC)	3659775
Organisation name	Centre for Population Health Science (CPHS)
Address	Medical School, Teviot Place
Postcode	EH8 9AG
Telephone number	01316503240
Email	Jackie.Price@ed.ac.uk
Complete the following question if you will access the individual level data requested in this application	
Provide details of the most recent Information Governance training undertaken.	Name of Course: Information Governance Training Tool Link to content (of course) if available: https://www.igte-learning.connectingforhealth.nhs.uk/igte/index.cfm
	Institution: HSCIC
	Date attended: To be attended in the next few months (will be completed before the data is supplied by ISD)

○ **Main Contact** *Researcher responsible for day-to-day running of project (to whom all correspondence will be sent) - if different from the person named in □o.*	
Title	Miss
Forename	Anna
Surname	Price
Position	PhD Student
Qualifications	BSc (Hons), MRes
Professional Registration Number (if applicable) eg GMC	
Organisation name	CPHS
Address	Medical School, Teviot Place
Postcode	EH8 9AG
Telephone number	07944616765
Email	s1356205@sms.ed.ac.uk
Complete the following question if you will access the individual level data requested in this application	
Provide details of the most recent Information Governance training undertaken.	Name of Course: SHIP: Information Governance Link to content (of course) if available: http://www.law.ed.ac.uk/teaching/online_distance_learning/cpd_courses/ship_information_governance/course_overview
	Institution: Edinburgh Law School, University of Edinburgh
	Date attended: Online course begun 26/9/14 – will be completed before data is received from ISD

○ **Information Custodian** *The Information Custodian is the person taking responsibility for safeguarding the confidentiality of the data. This is likely to be the Head of Department responsible for the project. An Information Custodian is ONLY required if the SHIP National Safe Haven will NOT be used. The information custodian should have training in Information Governance*	
Title	Professor
Forename	Jackie
Surname	Price
Position	See section 1.1
Qualifications	See section 1.1
Professional Registration Number (if applicable) eg GMC	See section 1.1
Organisation name	See section 1.1
Address	See section 1.1
Postcode	See section 1.1
Telephone number	See section 1.1
Email	See section 1.1
Provide details of the most recent Information Governance training undertaken.	Name of Course: See section 1.1 Link to content (of course) if available: See section 1.1
	Institution: See section 1.1
	Date attended: See section 1.1

Please provide the details of all additional people (if any) who will access the individual level data requested in this application. There is space here to provide details of three people. If there are more than three, please append the additional information with your application.

○ **Access Individual Level Data – 1**	
Title	Dr
Forename	Joanne
Surname	Morling
Position	Clinical Training Fellow/Specialist Registrar in Public Health
Qualifications	BSc (Hons), MBChB, MSc, MRCP, MFPH
Professional Registration Number (if applicable) eg GMC	6097377
Organisation name	CPHS
Address	Medical School, Teviot Place
Postcode	EH8 9AG
Telephone number	0131 6503244
Email	J.Morling@ed.ac.uk
Provide details of the most recent Information Governance training undertaken.	Name of Course: SHIP Information Governance Link to content (of course) if available: http://www.law.ed.ac.uk/teaching/online_distance_learning/cpd_courses/ship_information_governance/course_overview
	Institution: Edinburgh Law School, University of Edinburgh
	Date attended: Online course begun 08/10/14 – will be completed before data is received from ISD

○ Access Individual Level Data - 2	
Title	Dr
Forename	Stela
Surname	McLachlan
Position	Data Manager
Qualifications	PhD
Professional Registration Number (if applicable) eg GMC	
Organisation name	CPHS
Address	Medical School, Teviot Place
Postcode	EH8 9AG
Telephone number	0131 650 6193
Email	stela.mclachlan@ed.ac.uk
Provide details of the most recent Information Governance training undertaken.	Name of Course: Information Governance Training Tool Link to content (of course) if available: https://www.igte-learning.connectingforhealth.nhs.uk/igte/index.cfm
	Institution: HSCIC
	Date attended: To be attended in the next few months (will be completed before the data is supplied by ISD)

○ **Access Individual Level Data – 3**	
Title	
Forename	
Surname	
Position	
Qualifications	
Professional Registration Number (if applicable) eg GMC	
Organisation name	
Address	
Postcode	
Telephone number	
Email	
Provide details of the most recent Information Governance training undertaken.	Name of Course: Link to content (of course) if available:
	Institution:
	Date attended:

○ **Other People Involved** Please list the names of any people not listed above that have had significant input into the design and content of this study, or will be involved in the study hereafter, but who will not access the individual level data requested in the application.		
Name (Forename/Surname)	**Organisation**	**Role**
Mark Strachan	Western General Hospital, Edinburgh	Consultant Physician & Honorary Professor
Rebecca Reynolds	Centre for Cardiovascular Science, University of Edinburgh	Professor of Metabolic Medicine and Honorary Consultant Physician
Ian Deary	University of Edinburgh Centre for Cognitive Ageing and Cognitive Epidemiology (CCACE)	Professor of Psychology & Director, CCACE

○ **Previous publications** Please list up to three publications which members of the research team have produced or been involved in that demonstrate relevant experience in the use of administrative data for research.			
Authors	**Title of publication**	**Journal**	**Citation** (year: volume; pages)
Jie Ding, Mark W.J. Strachan, Rebecca M. Reynolds, Brian M. Frier, Ian J. Deary, F. Gerald R. Fowkes, Amanda J. Lee, Janet McKnight, Patricia Halpin, Ken Swa, and Jackie F. Price	Diabetic retinopathy and cognitive decline in older people with type 2 diabetes: the Edinburgh Type 2 Diabetes Study	Diabetes	2010: 59; 2883-2889
Riccardo E. Marioni, Ian J. Deary, Gordon D. Murray, Gordon D. O. Lowe, Snorri B. Rafnsson, Mark W. J. Strachan, Michelle Luciano, Lorna M. Houlihan. Alan J. Gow, Sarah E. Harris, Marlene C. Stewart, Ann Rumley, F. Gerry R. Fowkes, Jackie F. Price	Genetic variants associated with altered plasma levels of C-reactive protein are not associated with with late-life cognitive ability in four Scottish samples	Behavioural Genetics	2010: 40; 3-11
Riccardo E. Marioni, Mark W.J. Strachan, Rebecca M. Reynolds, Gordon D.O. Lowe, Rory J. Mitchell, F. Gerry R. Fowkes, Brian M. Frier, Amanda J. Lee, Isabella Butcher, Ann Rumley, Gordon D. Murray, Ian J. Deary and Jackie F. Price	Association between raised inflammatory markers and cognitive decline in elderly people with type 2 diabetes: the Edinburgh Type 2 Diabetes Study	Diabetes	2010: 59; 710-713

▪ Study Overview

In order to help the PAC assess your application, you are required to provide an overview of your study. It is important that the following section is completed with information accessible and comprehensible to a lay reader, and any acronyms are expressed in full when first used. You should include your study protocol with your application.

○ **What is the background to the study?**
Risk factors underlying the development and progression of some of the less well-recognised complications of type 2 diabetes, including cognitive impairment and non-alcoholic fatty liver disease, are poorly understood. The Edinburgh Type 2 Diabetes Study was established in 2006 in order to investigate the role of potential risk factors in these complications, as well as to further investigate mechanisms underlying the development and progression of micro and macrovascular disease in type 2 diabetes.
○ **Why is the study needed?**
Type 2 diabetes currently affects around two million people in the UK and approximately 10% of people aged over 65 years. The prevalence of the condition is predicted to double over the next 20 years, with a particular increase in elderly people. Much has been done over the last decade to try to prevent and treat the well-recognised micro- and macrovascular complications of diabetes, including atherosclerotic cardiovascular disease, diabetic retinopathy, nephropathy and peripheral neuropathy. However, morbidity and mortality from vascular disease remains high in older people with type 2 diabetes. Detailed information on potential risk factors is crucial to identify causal and modifiable risk factors that can be targeted for the development of appropriate preventive and therapeutic interventions, in addition to helping to identify patients who are at increased risk of developing complications.

○ What are the aims and objectives of the study?
The aims of this project are: 1. To determine the association between potentially modifiable risk factors (including microvascular disease, inflammatory mediators and hormones of the hypothalamic-pituitary-adrenal (HPA) axis) and cognitive decrements in people with type 2 diabetes. 2. To determine, in older people with type 2 diabetes, (i) the prevalence of Non-Alcoholic Fatty Liver Disease (NAFLD), (ii) clinical factors that might permit early detection of people at increased risk of developing NAFLD and (iii) potentially causal risk factors for the development and progression of NAFLD. 3. To identify circulating biomarkers and other risk factors which (i) predict the development of symptomatic and asymptomatic micro- and macrovascular disease, (ii) are associated with progression of these complications and/or (iii) have a potentially causal role in their development. 4. To establish a well-characterised and compliant population sample with extensive phenotyping and the potential for genotyping, which can be used as the sampling frame for subsequent nested case control studies (including neuroimaging), and as a replication population for findings arising from genome-wide association studies.
○ **Give a brief outline of the study design and data sources involved**
Subjects for this study have been recruited through the Lothian Diabetes Register (LDR). Invitations were sent out to a random sample of individuals on the register, by LDR staff, and details of individuals who returned a reply slip stating that they were interested in participating were then passed on to the Edinburgh Type 2 Diabetes Study team. The one-year recruitment and baseline data collection phase ended in August 2007 and a follow-up data collection phase was carried out in 2011, which included Record Linkage data from ISD. The subjects for this study comprise approximately 1066 individuals with type 2 diabetes aged between 60 and 75.

○ **Describe your study sample (inclusion/exclusion criteria eg involvement in trial/survey, health event, relevant date range, requirement for a matched control cohort, etc)**
With the permission of the Lothian Diabetes Services Advisory Group and the Caldicott Guardian for NHS Lothian, patients recorded as having type 2 diabetes were selected from the Lothian Diabetes Register (LDR). The LDR is a computerised database, which was established in 2001, and contains clinical details on over 20,000 patients with known type 2 diabetes living in Lothian, Scotland. Comparison of age-sex specific prevalences of diabetes recorded on the LDR with those from other data sources in Scotland suggests that the LDR captures almost everyone with diagnosed diabetes in Lothian (Prof Sarah Wild, personal communication). Any subject in whom it was not possible to confirm a clinical diagnosis of type 2 diabetes by review of hospital and/or GP records was excluded. Other exclusion criteria were, (i) non-English speakers (since fluent English is required for some of the cognitive tasks), (ii) corrected visual acuity worse than 6/36 for distance vision or unable to read large print text (as at least moderate visual function is required to complete some of the cognitive tasks), (iii) unwilling to give consent (or judged by clinical research staff to be unable to give fully-informed consent) (iv) physically unable to complete the clinical and cognitive examination.
○ **Indicate whether this study has any implications for sensitive groups or vulnerable populations (see Appendix Two for details).**
There are no implications for sensitive groups or vulnerable populations in this study.
○ **Describe envisaged benefits of your study either to patients or the wider public**
This study will provide a wealth of epidemiological and biomarker data that should be invaluable in the identification of potentially modifiable, causal risk factors for diabetes-related cognitive impairment, liver dysfunction and vascular disease, which can be targeted for the development of preventive and therapeutic interventions.
○ **Do you anticipate that the data being requested will be used to develop products for the purpose of profit?**
No.

	Please embolden **Yes** or **No** as appropriate	
○ **Is this application an extension to/update of a previous PAC application?**	**Yes**	No
If yes, please provide the PAC reference number(s)	3707	

○ **To your knowledge, has any external dataset you wish to use been linked to NSS data in the past?**	Yes	**No**
If yes, please provide details, including relevant PAC reference number(s) if available		
○ **Do you have funding in place for your study?**	**Yes**	No
If yes, please provide the name of all funding bodies	The ET2DS is funded by the Medical Research Council and Pfizer Ltd	
Has the assessment for funding included peer review?	**Yes**	No
○ **Are any of the funding bodies commercial, for profit, organisations?**	**Yes**	No
○ **Do you intend to access the data requested through the SHIP National Safe Haven?** *Researchers based in Scotland will be expected to use the SHIP national safe haven. If you are using the safe haven you do not need to complete the majority of □ which relates to information governance and security. Question 5.1 requires to be completed by all applicants.*	Yes	**No**
If no, please explain your reason why:		
The ET2DS is an active and on-going research project based on a fully consented cohort of patients who have agreed to allow access to their medical records and healthcare data for the purposes of the research. The cohort has been linked to ISD data twice previously and on neither occasion was it necessary to use a Safe Haven to analyse the data. This would have entailed moving the entire ET2DS dataset into a Safe Haven, linking to a relatively small amount of ISD data, and then having several different researchers accessing the resultant dataset on a daily basis for 6 years from a remote location. Rather, we hold the full ET2DS dataset on a secure server as described in the rest of this form, where we are able to access it directly. Experience with collaborators on other projects has demonstrated that regular analysis of a dataset held by someone other than the researchers themselves is problematic and time consuming. Recent discussion (between staff in CPHS who hold several of their own, similarly linked datasets, Janet Murray and colleagues from the Safe Haven), have indicated to us that there is no intention for use of the Safe Haven to impede on-going research, so we were led to believe that we would be granted access to the ISD data again this time around, for subsequent merging with our secure ET2DS dataset.		

▪ Data Requests

This section of the form requests further detail regarding the use of data to meet the objectives of the study.

National Records of Scotland (NRS) Data

	Please embolden **Yes** or **No** as appropriate	
○ **Does your study involve data to be provided by NRS? If not, go to question □o**	Yes	**No**
○ **Does your study require use of NHS Central Registry (NHSCR) as a sampling frame for study controls?**	Yes	**No**
○ **Does your study involve flagging of individuals on the NHSCR? If yes, please answer the following questions.**	Yes	**No**
Is the flagging of individuals:		
To help trace and contact individuals throughout the UK?	Yes	**No**
To be informed of fact and cause of death?	Yes	**No**
To be informed of cancer registrations?	Yes	**No**
To be informed of emigrations prospectively and retrospectively?	Yes	**No**
○ **Does your study require the provision of any other service from NRS? If so detail below.**	Yes	**No**

National Services Scotland (NSS) Data

	Please embolden **Yes** or **No** as appropriate	
○ **Does your study involve NSS data?**	**Yes**	No

If yes, please tick the dataset(s) involved below and indicate the relevant time period:		
	Tick	Time period
SMR00 - Outpatients		
SMR01 - Inpatients and Day Cases	✓	Last linkage was done on 01/01/2011. Any new SMR01 records since this date – present date
SMR02 - Maternity		
SMR04 - Mental Health		
SMR06 - Cancer Registration		
SMR11/SBR - Neonatal/Scottish Birth Record (please specify which)		
CHSP-PS/CHSP-S/SIRS - Child Health Surveillance and Immunisation (please specify which)		
A&E - Accident and Emergency		
PIS - Prescribing Information		
National Audits and Disease Registries eg Surgical Mortality, Renal Registry (please specify which)		
Birth, Stillbirth or Death Records (NRS) (please specify which)	✓ (Death)	Last linkage was done up to the end of the 2009/2010 period. Any new death records since the end of this period – present date
Scottish Drugs Misuse Database (SDMD)		

PTI (Primary Care Data)		
Other (please list below)		

○ **Indicate by ticking all the box(es) that apply whether the information provided by NSS Scotland will be used to make direct contact with the following**			
	Make Contact		
	By Letter	By Telephone	Other method - please specify
Hospital consultants			
Other hospital staff			
General Practitioners			
Study members or patients			
Relatives of study members or patients - please specify			
Some other party - please specify			
Please explain why you will contact each group and provide copies of the relevant correspondence:			

In order to establish whether or not subjects within the ET2DS suffered from an event (e.g. a CV event) a variety of sources were used as evidence in combination (see below for specific examples of the criteria used to define CV events at both baseline and 4-year follow-up). Therefore, we do not rely solely on data from ISD to define an event. There may be occasions where we don't already have the evidence required to confirm an event shown by the ISD data and in these instances it may be necessary to contact GPs/study members in order to verify. It should be noted that study participants have already given consent for their medical records to be consulted and their GPs to be contacted.

Baseline event criteria: The following criteria were used to define MI: *1*) subject recall of a doctor's diagnosis of MI, *2*) positive WHO chest pain questionnaire for MI, *3*) ECG evidence of ischemia (Minnesota codes 1.1–1.3, 4.1– 4.2, 5.1–5.3 or 7.1), and *4*) prior hospital discharge code for MI (ICD-10 codes I21–I23, I252). MI was recorded if two of the first three criteria were met or if both the first and last criteria were met. Equivalent criteria for angina were: *1*) subject recall of a doctor's diagnosis of the condition or being on regular medication for angina, *2*) positive WHO chest pain questionnaire for angina, *3*) ECG evidence of ischemia, and *4*) prior hospital discharge code for ischemic heart disease (ICD-10 codes I20 –I25). Angina was recorded if two of the first three criteria were met or if both the first and last criteria were met. Stroke was recorded if two of three of the following criteria were met: *1*) subject recall of a doctor's diagnosis of stroke, *2*) prior hospital discharge code consistent with stroke (ICD-10 codes I61, I63–I66, I679, I694), and *3*) confirmation by review of clinical notes that the event was not due to a transient ischemic attack.

Four-year follow-up event criteria: Four years after recruitment, participants were followed-up for new CV events using a combination of repeat self-completion questionnaire, repeat ISD record linkage for hospital discharge and death certificate data and review of clinical case notes as required. Criteria for fatal and non-fatal events were as follows. Myocardial Infarction (MI): (1) ICD-10 code for new MI on hospital discharge/death record, dated after baseline, plus either subject report of a doctor diagnosis of MI, positive WHO chest pain questionnaire for MI, report of MI on GP questionnaire (provided all the dates were consistent with ICD-10 coded event) or ECG codes for MI which were not present at baseline; or (2) clinical criteria for MI met following scrutiny of hospital and/or GP notes. New angina (in subjects without a diagnosis of angina at baseline): (1) ICD-10 code for angina as primary diagnosis on hospital discharge record, dated after baseline; or (2) at least 2 of (a) subject self-report of a doctor diagnosis of angina or of starting angina medication since baseline, (b) ECG codes for ischaemia which were not present at baseline, and (c) positive WHO chest pain questionnaire; or (3) clinical diagnosis of angina on scrutiny of hospital notes. Fatal IHD: ICD-10 codes for IHD (other than MI) as underlying cause of death from death certification data. Stroke: (1) ICD-10 code for stroke as primary diagnosis on hospital discharge/death record, dated after baseline; or (2) clinical criteria for stroke met on scrutiny of clinical notes in subjects with either self-report of stroke or with non-primary ICD-10 hospital discharge/death code for stroke. Transient Ischaemic Attack (TIA): (1) ICD-10 code for TIA as primary diagnosis on hospital discharge record; or (2) clinical criteria for TIA met on scrutiny of clinical notes in subjects with either-self-report of stroke or with non-primary ICD-10 hospital discharge code for stroke or TIA.

.

	Please embolden **Yes** or **No** as appropriate	
○ **Does your application request NSS to facilitate communication with individuals in the study sample?**	Yes	**No**

○ **Does your study involve use of NSS data as a sampling frame?**	Yes	**No**

○ **Will all analysis be done in NSS by NSS staff ie you only require aggregate output?**	Yes	**No**

Non-NRS/NSS Datasets

	Please embolden **Yes** or **No** as appropriate	
○ **Does your study involve linkage to non-NRS/NSS data? If yes, you must provide information on each dataset. If no go to question □o.**	**Yes**	No

Non NRS/NSS data that may be involved include:

Data held by another NHS Board (eg treatment or audit data)

Data held by GPs

Research dataset eg clinical trial

Survey dataset eg national social surveys

Data relating to social care

Data relating to education

Other data from a public authority

Other (eg data from a commercial organisation)

Please provide the information requested for each of the non-NRS/NSS datasets that you will provide to ISD. There is space here to provide details for three non-NRS/NSS datasets. If there are more than three involved, please append the additional information with your application.

○ **Non NRS/NSS Dataset - 1**		
What is the name of the dataset?	ET2DS database	
The purpose for which it was collected	To investigate the role of potential risk factors in complications such as cognitive impairment, as well as to further investigate mechanisms underlying the development and progression of micro and macrovascular disease in type 2 diabetes.	
Describe the content of the dataset	The study constitutes 1066 men and women aged 60 to 75 years with established type 2 diabetes, living in the Lothian region of central Scotland. At baseline, subjects underwent detailed cognitive and physical examination, the latter including measures of micro- and macro-vascular disease, glycaemic control, body fat composition and plasma inflammatory markers, cortisol, lipids and liver function tests. Participants were re-examined after one year with hepatic ultrasonography and additional measures of vascular disease. Record Linkage has been used to determine cardiovascular outcomes at a four year follow-up stage.	
The time period to which it pertains	2006-2011	
What is the name of the data controller	Professor Jackie Price	
Describe how patients have been informed of this use of their data.	All subjects gave written informed consent at the baseline clinic.	
The identifying variables which will be provided to ISD to enable linkage (please tick all that apply)	Forename	✓
	Middle name	✓
	Surname	✓
	CHI Number	
	Postcode	✓
	Date of Birth	✓
	Gender	✓

	UK NHS Birth Registration Number	
	Other (please specify below): ✓	
Study ID Number		

○ **Non NRS/NSS Dataset - 2**		
What is the name of the dataset?		
The purpose for which it was collected		
Describe the content of the dataset		
The time period to which it pertains		
What is the name of the data controller		
Describe how patients have been informed of this use of their data.		
The identifying variables which will be provided to ISD to enable linkage (please tick all that apply)	Forename	
	Middle name	
	Surname	
	CHI Number	
	Postcode	
	Date of Birth	
	Gender	
	UK NHS Birth Registration Number	
	Other (please specify below):	

○ **Non NRS/NSS Dataset - 3**		
What is the name of the dataset?		
The purpose for which it was collected		
Describe the content of the dataset		
The time period to which it pertains		
What is the name of the data controller		
Describe how patients have been informed of this use of their data.		
The identifying variables which will be provided to ISD to enable linkage (please tick all that apply)	Forename	
	Middle name	
	Surname	
	CHI Number	
	Postcode	
	Date of Birth	
	Gender	
	UK NHS Birth Registration Number	
	Other (please specify below):	

Output File for Analysis

The risk of inadvertent identification/disclosure of individuals increases with increased level of detail contained in the dataset. Only variables required to meet study objectives should be requested.

Identifiable data includes variables such as name and date of birth. In general, access to personal identifiers will not be provided. Exceptional requests for access may be considered taking account of Information Governance principles.

	Please embolden **Yes** or **No** as appropriate	
○ **Do you require any patient identifiers in the output file for analysis?**	**Yes**	No
If yes, please identify which patient identifiers you require in the output file for analysis		
Forename		
Surname		
CHI Number		
NHS Number		
Full Postcode		
Full Date of Birth		
Full Date of Death	Yes	
Full Date of Admission	Yes	
Full Date of Discharge	Yes	
Other (please list)		
Provide justification for why you need these patient identifiers. This information can only be provided where a clear need is shown.		
These full dates were provided in the previous record linkages and were used to calculate time to event in days for statistical analysis. For consistency we would like to do the same again.		

Data without identifiable variables may retain the potential to identify an individual. This increases with the level of detail included, particularly where the denominator population is small, for example when rare conditions or low level geographies are involved.

Our Research Co-ordinators will be able to advise whether 'derived variables' may be provided to reduce the risk of identifying individuals, for example month of admission and length of stay rather than dates of admission and discharge.

Output files may include a study index number where recognition of individual records is necessary.

○ **Identify which, if any, of the following variables are required in the output file.**	
Partial dates:	
Partial Date of Birth (eg month and year – please specify)	
Partial Date of Death	
Partial Date of Admission	
Partial Date of Discharge	
Gender:	
Gender	
Geographical variables:	
NHS Board area	
Local Authority area	
Datazone	
Partial Postcode (please specify)	
Other (please specify)	
Clinical variables:	
Information regarding rare conditions (please list the conditions)	
Provide justification for why you need these variables. Variables can only be provided where a clear need is shown.	

Please list here or in an appendix all the other variables (not already listed) which are required in the output file, identifying which variables come from which data source.

○ **All other variables to be provided in the output file**	
Variable	Data Source
Study ID Number	Provided by us
Record type	Provided by ISD
Date of admission	Provided by ISD
Data of discharge	Provided by ISD
Hospital name	Provided by ISD
All diagnoses (main and other)	Provided by ISD
All operations (main and other)	Provided by ISD
Principal cause of death	Provided by ISD
Secondary causes of death	Provided by ISD

Duration of the Study

○ **What is the expected duration of the study?**
5 years

Updates and Retention

	Please embolden **Yes** or **No** as appropriate	
○ **Does your study require access to a regular update of data?**	Yes	**No**
If yes, please explain the reason for this and the frequency of updates required. Please note approval is for 5 years. Updates after that time require a new application.		
○ **For how long will you either keep the data or require it to be retained in the safe haven (including the updates) after the study is complete?** Please note the standard archive time for safe haven is 2 years.		
At least 10 years within the ET2DS database.		
○ **Provide justification for why you need to retain the data for that length of time.**		
In order to undertake substantial analysis.		

▪ Permissions to Use Data

For each dataset not under the control of NRS or NSS, you must seek authority to access those data.

An application to the CHI Advisory Group (CHIAG) is necessary where the study requires access to information from the CHI dataset.

An application to the NHSScotland Caldicott Guardian Forum is necessary where the study requires access to information datasets held by multiple NHS Boards.

Approval from the National Research Ethics Service should be sought in the following situations:

Where the study involves linkage of a research dataset to another dataset.

Where the study involves use of identifiable data.

Where study involves use of highly disclosive data eg information regarding rare conditions or at a small area level.

It is Scottish Government policy that patients are informed regarding the use of their data ('Protecting patient confidentiality' 2002). Applicants must demonstrate that the proposed use of the datasets they wish to link is in line with the information provided to patients and provide copies of relevant information such as participant information leaflets and consent forms, posters, links to websites etc.

Please provide the information requested for each non NRS/NSS dataset involved in your application.

○ **Evidence that use of the data is authorised by the Data Controller(s).** *This may include authorisation from CHIAG, NHS Caldicott Guardians Forum or other. This can be attached to your application.*
Caldicott Guardian approval was provided in Nov 2005 for random sampling of ET2DS participants from the Lothian Diabetes Register.
○ **Describe the methods used to inform study participants and/or the wider public regarding the use of their data in this way.**
An information sheet was given to all participants at first contact (the baseline visit), and written informed consent was then obtained.
○ **Provide copies of the information provided to study participants and/or the wider public regarding the use of their data in this way.** *This can be attached to your application.*
Attached.
○ **Provide the participant consent form for research studies or surveys.** *This can be attached to your application.*
Attached.
○ **Where no consent for proposed use has been obtained from data subjects, please provide justification below explaining why there is use without consent.** *For example, please explain why consent has not been obtained and explain how this proposed use relates to the original purpose of data collection.*
N/A
○ **Have any members of the public/lay people been involved in the study design?** *Please provide information.*
No.

▪ Information Governance

NSS must ensure that any data approved for release will be adequately protected against inappropriate access and use during the study, and will be securely disposed of once the study is completed.

Researchers will be expected to access NSS data using the SHIP National Safe Haven. Any alternative to this will require justification.

You may need to consult your organisation's Information Governance Lead and IT service supplier when completing this part of the form.

○ Information Governance Incident Reporting

Your organisation(s) should have Information Governance incident reporting procedures available and these should be accessible to and used by all staff on this study.

NSS should be notified immediately of any information governance breaches that have occurred involving NSS supplied data during this study. Please confirm you will notify **nss.pac@nhs.net** of any such incidents by ticking here ☒

The rest of this section does not need to be completed for studies in which data will be accessed only through the SHIP National Safe Haven. Please complete Questions □o to □o for all other applications ie studies requiring release of data directly to the applicant.

○ Local Information Governance Policies and Procedures

Your organisation(s) should have Information Governance policies and procedures available and these should be accessible to and used by all staff on this study. Please provide copies of the local Information Governance policy/policies that apply in each of the organisations/locations where the data provided for this study will be held. Provide these files along with your application or provide URLs below if the policies are available online.

URL(s) if policies are available online: http://www.ed.ac.uk/schools-departments/records-management-section/data-protection/guidance-policies/research/research

○ Data Protection Registration

Please provide the Data Protection Registration Number of each of the organisation(s) where data will be held.

Organisation(s) Name/Data Storage location	Data Protection Registration Number
University of Edinburgh	Z6426984

○ ISO 27001

If your organisation(s) have adopted ISO27001 - Information Security - Security Techniques - Information Security Management Systems, please provide your certification number.

Organisation Name / Data Storage location	ISO 27001 Certification Number

○ NSS Data Transfer and Storage Policies

NSS requires that all sensitive and person identifiable data are encrypted during transfer and whilst stored on mobile data storage devices and desktop and laptop computers.

NSS prefers that any data provided are stored on secure networked drives as part of a secure managed server. If mobile data storage devices have to be used, you must implement adequate protection against device loss or theft, unauthorised interception and access.

Where NSS supplied data are being stored on mobile data storage devices (for example but not limited to: USB 'sticks' and USB data storage drives, desktop or laptop computer) these devices must be fully encrypted to FIPS 140-2/CESG CAPS certified level of security protection. Where devices cannot be encrypted (for example: CDs, DVDs) then the data must be encrypted to FIPS 140-2/CESG CAPS certified level of security prior to storage on the mobile data storage device.

NSS can only send sensitive and person identifiable data to other users using NHSmail or secure file transfer protocol (SFTP). Please discuss methods of transfer with your Research Coordinator.

There are risks associated with using any email services for the transfer of data including sending the communication to the wrong email address. Please ensure that the NHSmail email address you provide is the correct address to be used. NSS will only send data to individual user (eg named) email addresses and not to generic email addresses. Please note that it is not possible to send or receive password encrypted attachments via NHSmail.

Please confirm that you have read and understood the details regarding NSS Data Transfer and Storage Policies by ticking here ☒

○ Data Storage: Locations

Please provide details on where you will store the data supplied by NSS. If data are being stored in more than one location then this section needs to be completed for each location.

At what location(s) within Scotland will data be stored? *Please list.*	CPHS, University of Edinburgh
At what location(s) outside Scotland will data be stored? *Please list.* *Specific considerations will apply where data is stored outside of the European Union.*	None

○ Data Storage: Devices and Formats

Please provide details on how you will store the data supplied by NSS.

Storage Device	Please tick all that apply and specify for each, the location at which the data will be stored.	
	Confirm	Location
Networked server disk drive	✓	CPHS, University of Edinburgh
Networked desktop PC*/ laptop* **delete as appropriate*		
Standalone desktop PC*/ laptop* **delete as appropriate*		
Mobile device		

Storage Format	Please tick all that apply and specify for each, the location at which the data will be stored.	
	Confirm	Location
Database		
Oracle database		
Microsoft Access database	✓	CPHS, University of Edinburgh
Microsoft SQL server		
IBM DB2		
MySQL		

Flat file eg Excel spreadsheet, comma delimited file.		

- **Backup**

Please confirm that your back-up schedule is subject to appropriate security measures, eg only appropriate staff have access to the back-up media; the back-up media are stored in a secure restricted location ☒ (check box to confirm)

If not please provide details of your backup here.

- **Other Encryption or Anonymisation Procedures**

NSS Data Transfer and Storage Policies require that all sensitive and person identifiable data are encrypted during transfer and whilst stored on mobile data storage devices and desktop and laptop computers to the standards outlined earlier. Please provide details of any other encryption or anonymisation procedures that may be used and at what stage.

Any other encryption or anonymisation procedures used	At what stage

- **Data Transfer-In**

If you are providing NSS with a copy of data for linkage and/or analytical purposes please detail how this data will be transferred and what security, complying with NSS Policy, will be used to protect the data from interception and inappropriate access.

Please note that by sending sensitive or personal data you will be responsible for ensuring the data are adequately protected against inappropriate access and tampering during the transfer. Data must not be sent via fax services.

The NSS Research coordinator dealing with your application will discuss requirements for usernames and/or passwords for the data transfer and/or encryption process. You should not provide this information to PAC either via this form or to the PAC e-mail address.

	Please embolden **Yes** or **No** as appropriate	
NSS SFTP service (recommended)	**Yes**	No
If no, please specify the methods to be used		
Mobile data storage device, eg CD, USB, data stick		
FTP URL		
Other SFTP URL		
Email address from which data will be sent		
Other data transfer method		

- **User Access**

Please provide details on user access and account management policies that you have in place to limit or prevent inappropriate access to the data supplied by NSS.

	Please **embolden** as appropriate	
Will those accessing data, access it through individual or shared accounts?	**Individual**	Shared
Are 'complex' passwords (a mixture of alpha, numeric, upper/lower case, special characters) used on all accounts?	**Yes**	No
How often are users required to change their passwords?	Monthly	
	Quarterly	
	Bi-annually	
	Annually	
	Other please specify	
Are procedures in place to regularly review user access to sensitive and potentially identifiable personal data?	**Yes**	No
Are procedures in place to revoke user access to sensitive and potentially identifiable personal data when the user no longer requires this access?	**Yes**	No
Will the data be accessed by staff working off site eg staff working from home?	Yes	**No**
If yes, please detail how this access will be secured		
Provide any additional details of how data provided by NSS will be protected from unauthorised access.		

- **Hardware Security**

Describe the physical security arrangements for the location where the data is to be stored eg this could be your computer department if the data is stored on a networked server, or may be where the PC/laptop holding the data is physically located.
The data will be stored on the University of Edinburgh network server.
Describe the physical security arrangements for the location where the data is to be processed eg this is where your PC/laptop is located or wherever you are accessing the data from.
The data can only be accessed by people working in a specific office in CPHS. There is a pin code to enter the room, and then an Edinburgh University username and password is required to log onto a networked computer. Furthermore, permission is required from Jackie Price to access the networked folder containing the record-level data. Similarly, access to aggregated and anonymised data resulting from the ISD data requires an Edinburgh University username and password, plus permission granted by Jackie Price to view the relevant folders.
Detail any protection that is implemented against the introduction of malicious software (eg computer viruses) in the areas where the data will be stored and processed.
University virus protection software.

Do your hardware replacement agreement(s) address how data are handled when hardware under warranty fails?	Yes	No
If yes, would the hardware be returned to the supplier if there was a fault(s)?	Yes	No

Explain below how your organisation(s) dispose of hardware that they no longer require, that are faulty or covered by warranty.
Since no files are saved locally, there is no data stored on any particular piece of hardware. The data is only stored on the Edinburgh University networked server. For this reason, the sections regarding hardware have not been completed as it does not apply in our case.
If the data is being held in long-term archive(s) please explain how this data will be secured against further unauthorised access.
Similarly, we do not keep archived copies of the data – we only store data which is in use. The data will be used for years to come – at least a decade, as stated elsewhere on the form, if not longer, in order to carry out extensive analysis.
Who will have data management responsibilities for the data whilst in archive(s)?

What procedures are in place to retrieve the data from the archive(s)?

- **Data Retention and Disposal**

Data should not be kept any longer than is necessary.

Give details of your data retention policy for each of the organisations(s) holding the data, including any back-up copies.
Data will be stored for at least 10 years. In the future, anonymised data may be placed in a public archive accompanied by appropriate confidentiality and access safeguards.
Give details of how the data, and any back-up copies, will be securely disposed of at the appropriate time by each of the organisation(s) holding the data.
No paper copies of the data will be kept. After the project is complete, the electronic data files will be destroyed.

▪ Declaration

I DECLARE THAT this application is accurate, and that any health data made accessible to me, should it be successful, will be used for no other purpose, and in no other way, than as described above.

I UNDERSTAND THAT NHS National Services Scotland will refuse any future applications by me, or my employing or sponsoring organisation, should I use any health data made accessible to me for any other purpose or in any other way than that described above.

I CERTIFY THAT all staff who have access to health data are aware of the requirements of confidentiality and understand that its breach (eg disclosure of confidential information to a person not authorised to receive it) constitutes grounds for disciplinary action, which might result in dismissal.

I GUARANTEE THAT no publication will appear in any form in which an individual may be identified unless the written permission of that individual has been obtained, and that I will follow the ISD Statistical Disclosure Control Protocol when planning publications involving the data requested.

To be signed by the applicant

Applicant Signature: Anna Price	Date: 5/11/14
Name (in Capitals): ANNA PRICE	

To be signed by the Information Custodian named in Part One where the Information Custodian is not the applicant.

I DECLARE THAT (the applicant named above) is a bona fide worker engaged in a reputable project and that the data he/she asks for can be entrusted to him/her in the knowledge that he/she will conscientiously discharge his/her obligations, including in regard to confidentiality of the data, as stated in the declaration above.

Information Custodian Signature: 	Date: 5/11/14
Name (in Capitals): JACKIE PRICE	

Appendix One: Advisory Bodies within NHSScotland

NHS Scotland Caldicott Guardian Forum

NHS Scotland Caldicott Guardian Forum was established in 2010 comprising all NHS Scotland Caldicott Guardians. It has established a process for scrutiny of applications for access to Board-wide personal health information for health research or audit purposes. Applicants who wish to access data under the control of more than one NHS Board (excluding NSS) in support of their study, should make a separate application to the Caldicott Guardian Forum. More information on the Caldicott Guardian Forum can be found at:
http://www.knowledge.scot.nhs.uk/caldicottguardians/caldicott-forum.aspx.

The Community Health Index Advisory Group (CHIAG)

The role of the Community Health Index Advisory Group (CHIAG) is to advise the Chief Medical Officer (CMO) and the Directors of Public Health in Scotland on access to the Community Health Index (CHI) for various purposes including operational management of the NHS, audit and research.

Applicants who wish to make use of data processed by the Community Health Index in relation to their study should make a separate application to CHIAG. More information on CHIAG can be found at http://www.chiadvisorygroup.scot.nhs.uk/.

National Research Ethics Service

The National Research Ethics Service (NRES) reviews the ethical standards of research. Approval from NRES should be sought for studies involving linkage to a research dataset or for studies involving access to data containing identifying variables. Advice from an NRES Scientific Advisor should be sought where the study involves access to data which has high potential for identification of individuals where rare conditions or low level geographies are involved. The NSS Caldicott Guardian reserves the right to request that NRES review is sought for studies about which they have ethical concerns.

Other Governing Bodies

Some established research databases are governed by bodies which have been delegated authority by the relevant clinician or Caldicott Guardian. Examples include Scottish Diabetes Research Network, Aberdeen Maternal and Neonatal Databank and Aberdeen Children of the 1950's.

Appendix Two: Sensitive Data and Vulnerable Populations

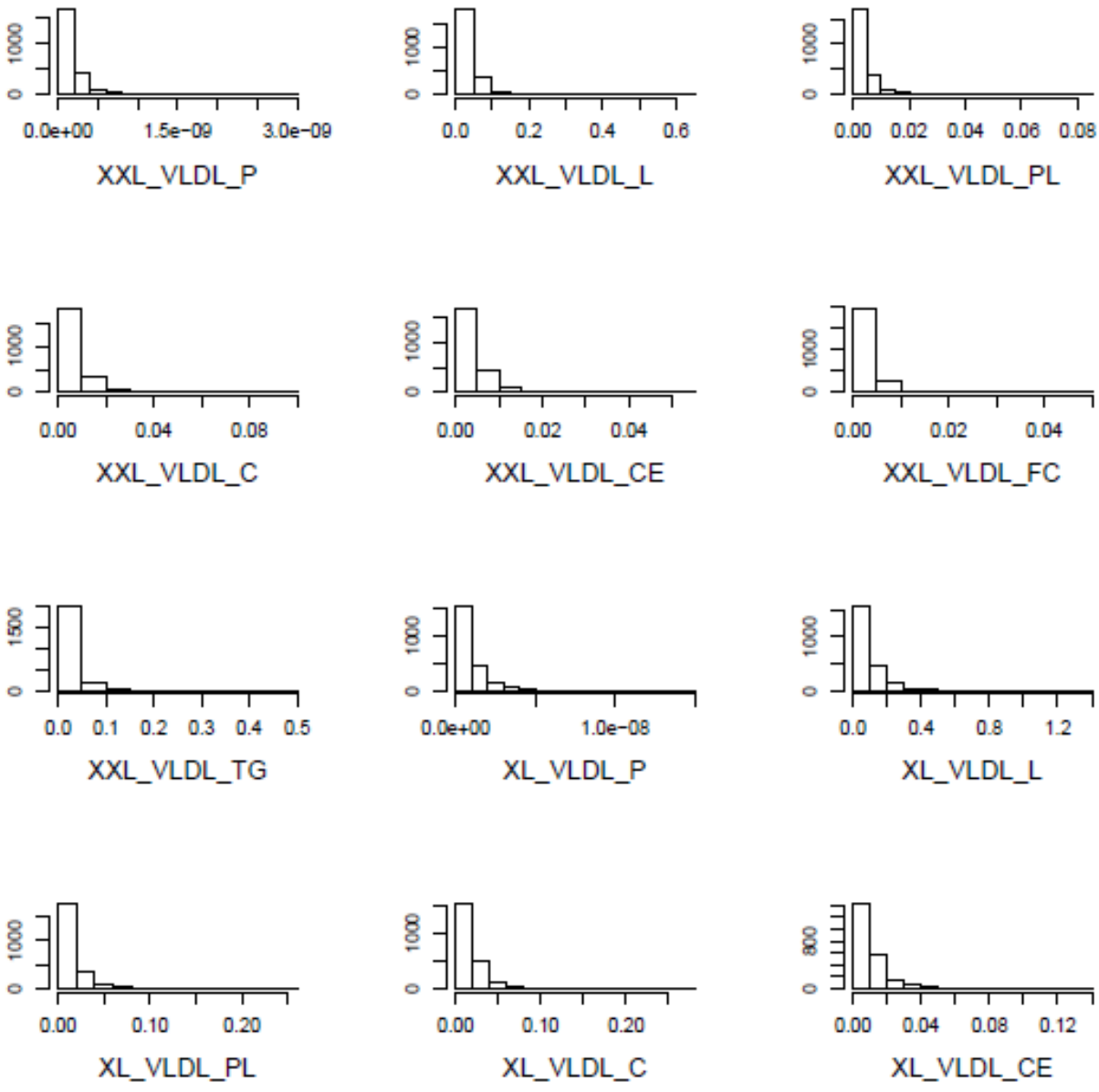
Some data carry a higher risk of harm to individuals who may be identified because they are perceived as particularly sensitive or they are part of a vulnerable population. A non-exclusive list is provided below.

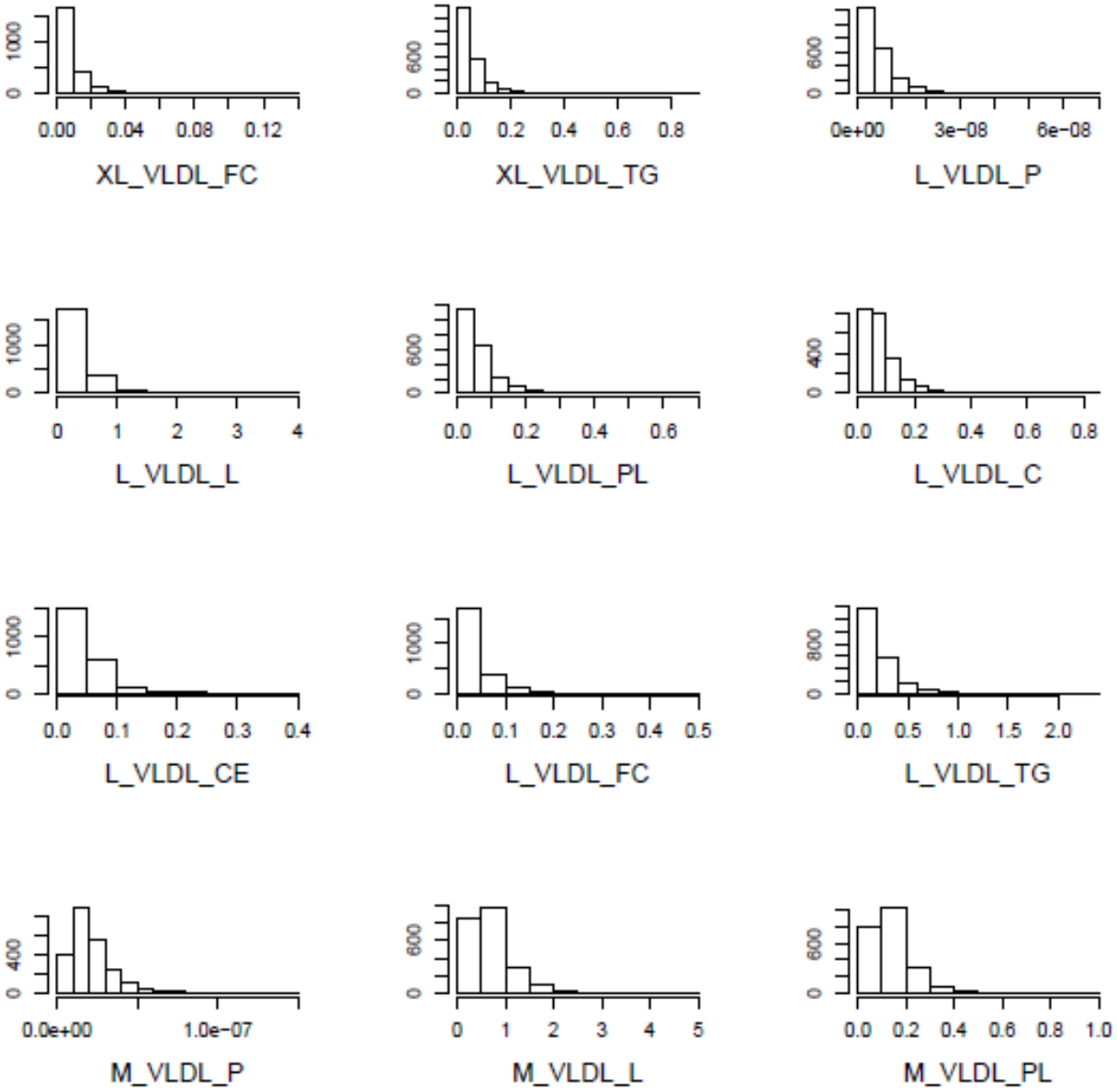
Sensitive Data pertains to:
Abortion
Pregnancy in age < 16 years
Sexually transmitted disease
Mental health
Drugs and alcohol misuse
Suicide
Mental health
Contraceptives
Crime related statistics
Ethnicity

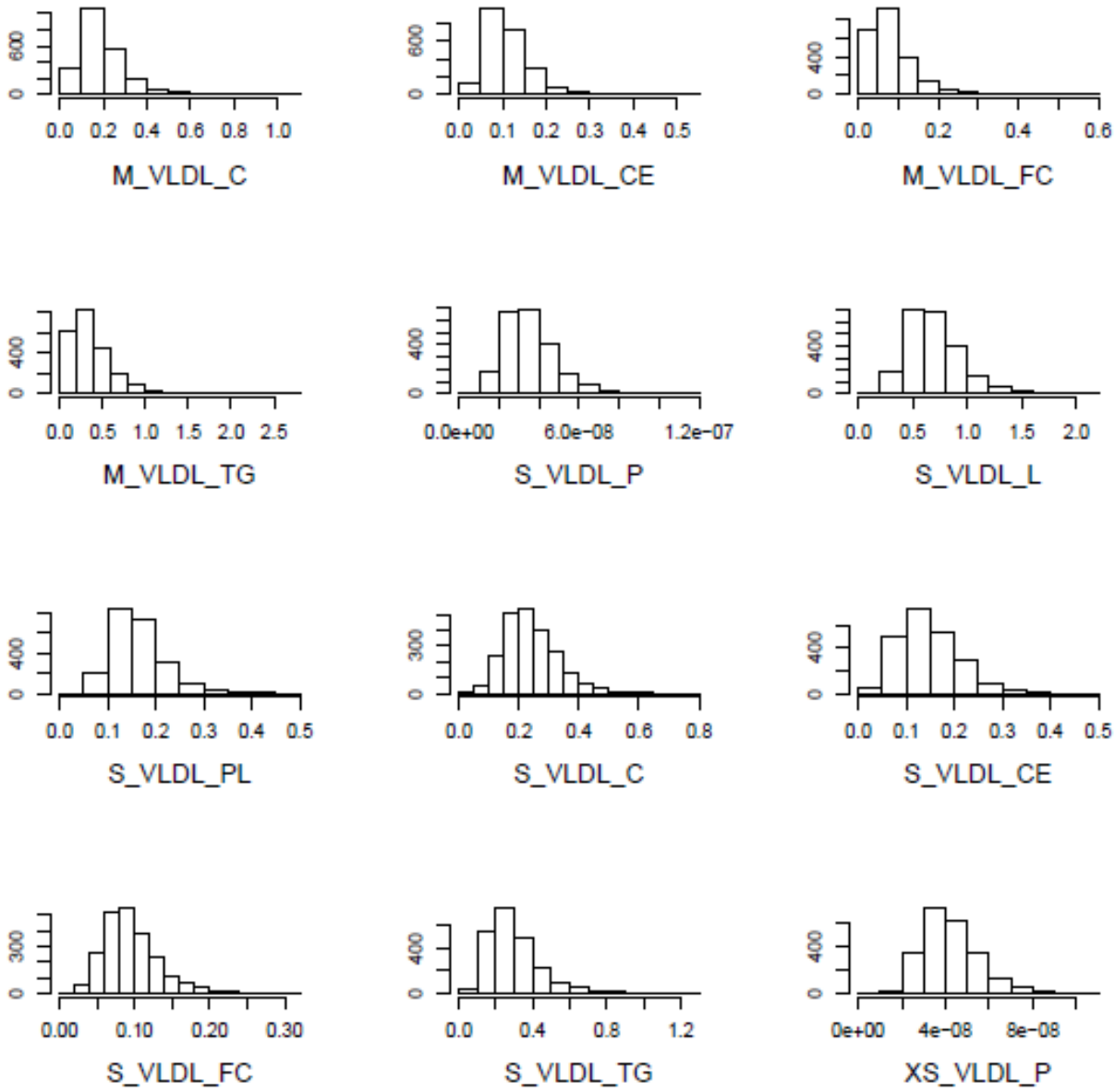
Vulnerable Populations pertains to:
Adults with Incapacity
Minority ethnic groups
Drug users

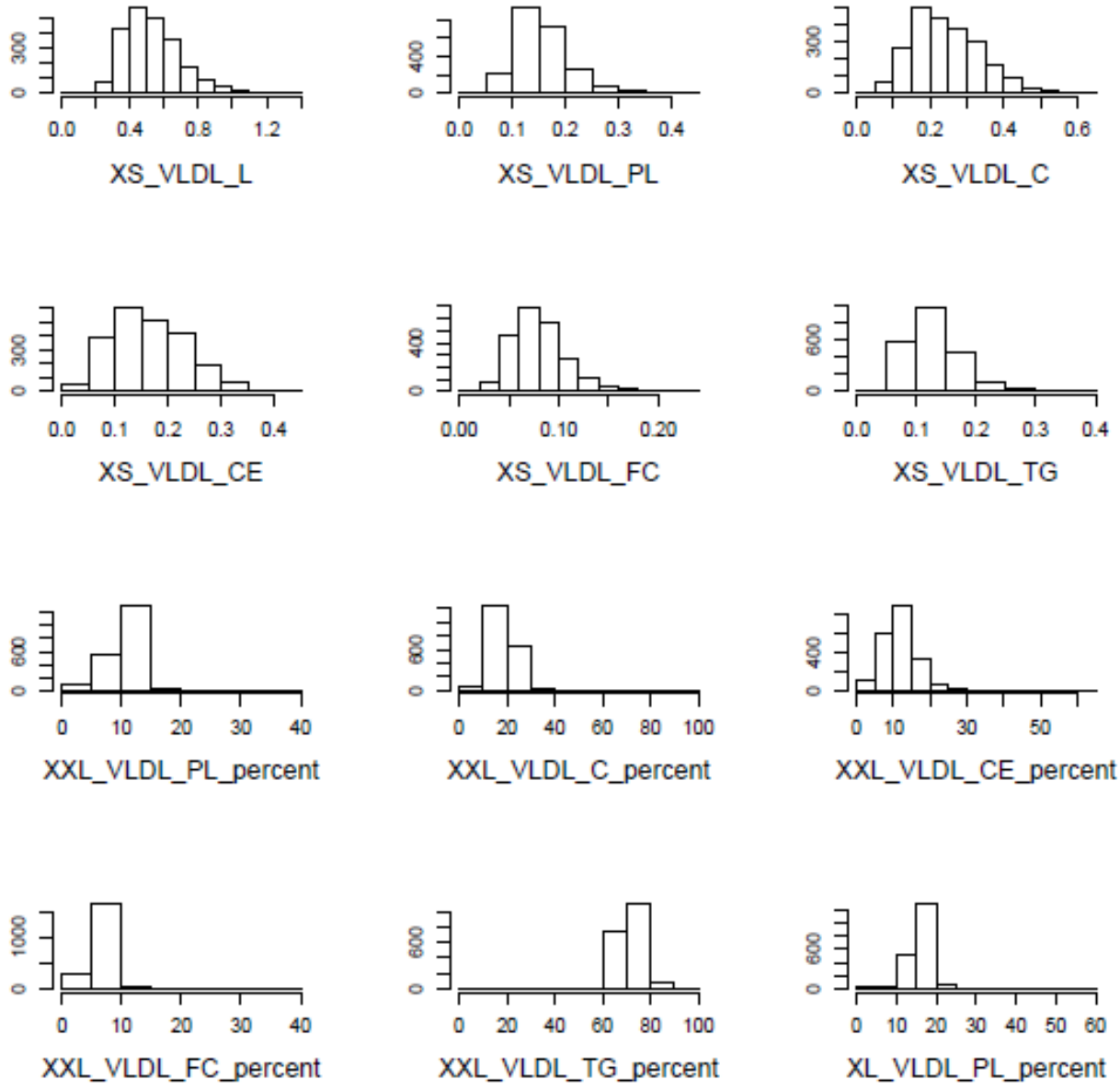
Appendix C Descriptive statistics of metabolites

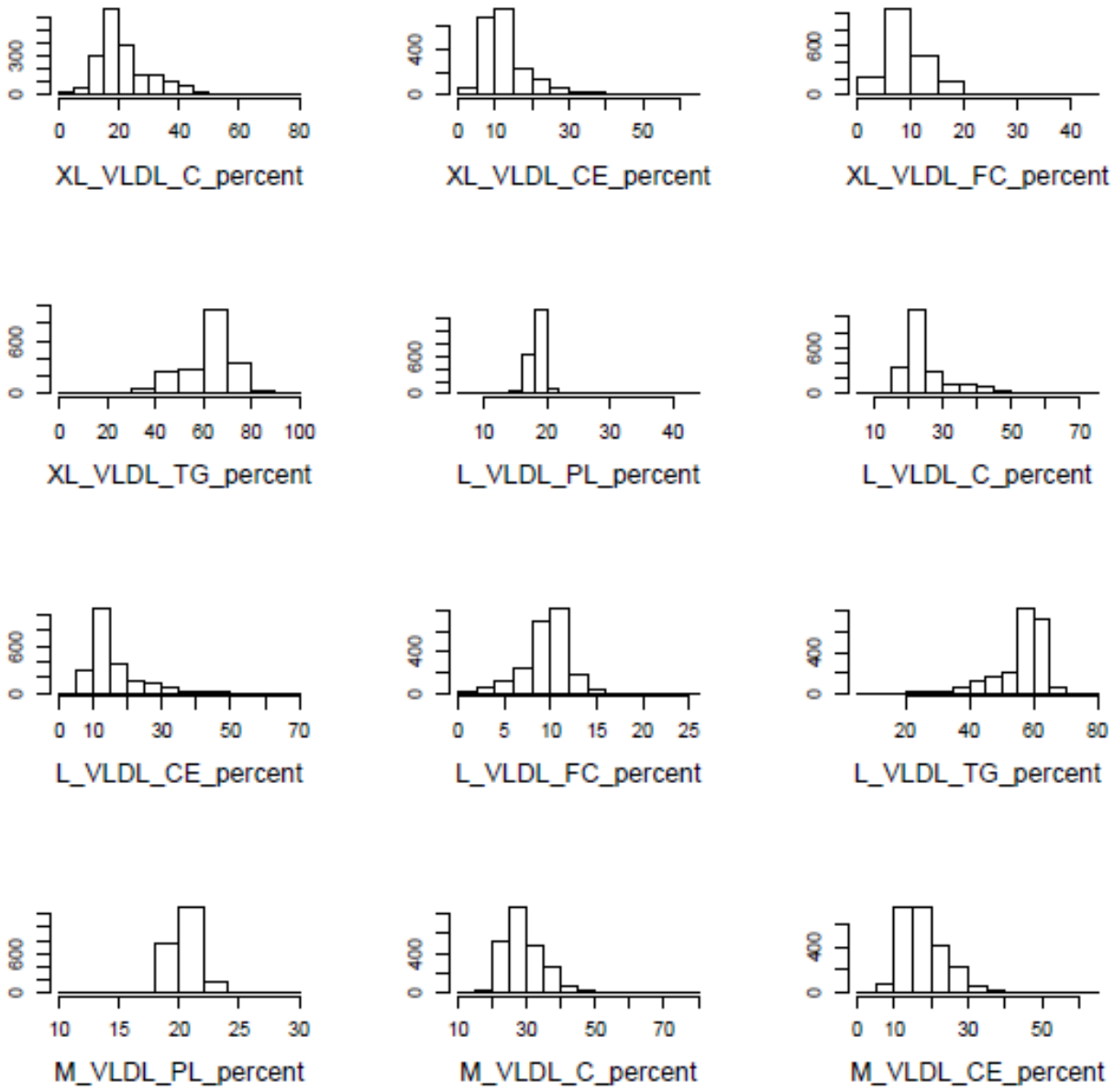
Figure C-1 Histograms of individual metabolites from the combined UCLEB data (full names of metabolites can be found in Table C-2)

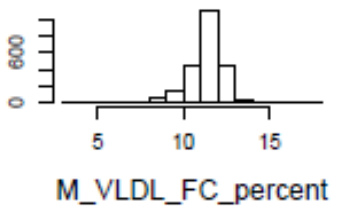


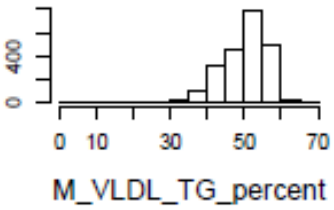

XL_VLDL_FC
XL_VLDL_TG
L_VLDL_P
L_VLDL_L
L_VLDL_PL
L_VLDL_C
L_VLDL_CE
L_VLDL_FC
L_VLDL_TG
M_VLDL_P
M_VLDL_L
M_VLDL_PL

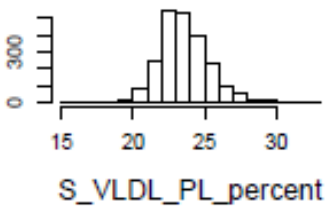

M_VLDL_C
M_VLDL_CE
M_VLDL_FC
M_VLDL_TG
S_VLDL_P
S_VLDL_L
S_VLDL_PL
S_VLDL_C
S_VLDL_CE
S_VLDL_FC
S_VLDL_TG
XS_VLDL_P

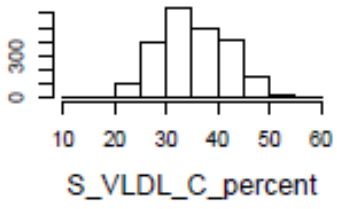

0 300
0.0 0.4 0.8 1.2
XS_VLDL_L
0 400
0.0 0.1 0.2 0.3 0.4
XS_VLDL_PL
0 300
0.0 0.2 0.4 0.6
XS_VLDL_C
0 300
0.0 0.1 0.2 0.3 0.4
XS_VLDL_CE
0 400
0.00 0.10 0.20
XS_VLDL_FC
0 600
0.0 0.1 0.2 0.3 0.4
XS_VLDL_TG
0 600
0 10 20 30 40
XXL_VLDL_PL_percent
0 600
0 20 40 60 80 100
XXL_VLDL_C_percent
0 400
0 10 30 50
XXL_VLDL_CE_percent
0 1000
0 10 20 30 40
XXL_VLDL_FC_percent
0 600
0 20 40 60 80 100
XXL_VLDL_TG_percent
0 600
0 10 20 30 40 50 60
XL_VLDL_PL_percent

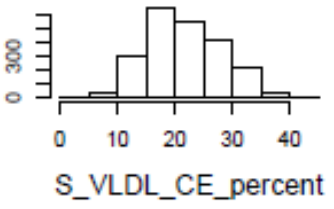

0 300
0 20 40 60 80
XL_VLDL_C_percent
0 400
0 10 30 50
XL_VLDL_CE_percent
0 600
0 10 20 30 40
XL_VLDL_FC_percent
0 600
0 20 40 60 80 100
XL_VLDL_TG_percent
0 600
10 20 30 40
L_VLDL_PL_percent
0 600
10 30 50 70
L_VLDL_C_percent
0 600
0 10 30 50 70
L_VLDL_CE_percent
0 400
0 5 10 15 20 25
L_VLDL_FC_percent
0 400
20 40 60 80
L_VLDL_TG_percent
0 600
10 15 20 25 30
M_VLDL_PL_percent
0 400
10 30 50 70
M_VLDL_C_percent
0 400
0 10 30 50
M_VLDL_CE_percent

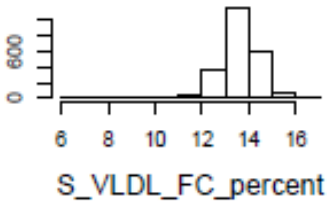

0 600
5 10 15
M_VLDL_FC_percent

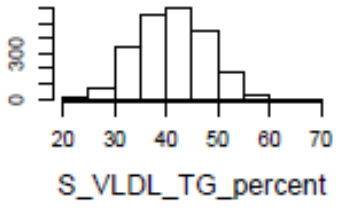

0 400
0 10 30 50 70
M_VLDL_TG_percent

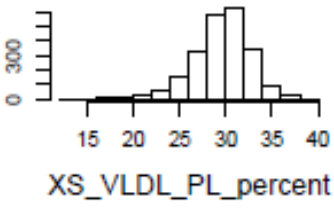

0 300
15 20 25 30
S_VLDL_PL_percent

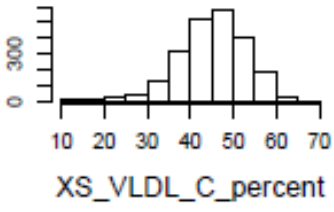

0 300
10 20 30 40 50 60
S_VLDL_C_percent

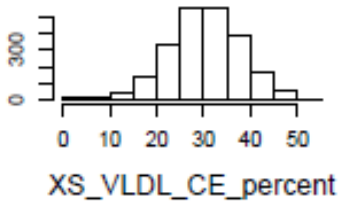

0 300
0 10 20 30 40
S_VLDL_CE_percent

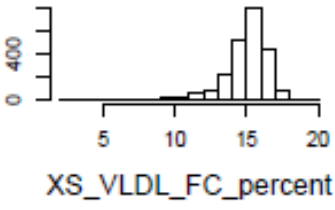

0 600
6 8 10 12 14 16
S_VLDL_FC_percent

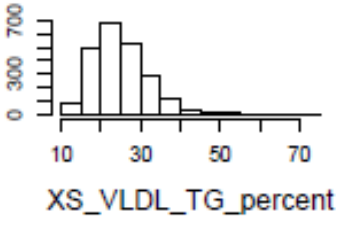

0 300
20 30 40 50 60 70
S_VLDL_TG_percent

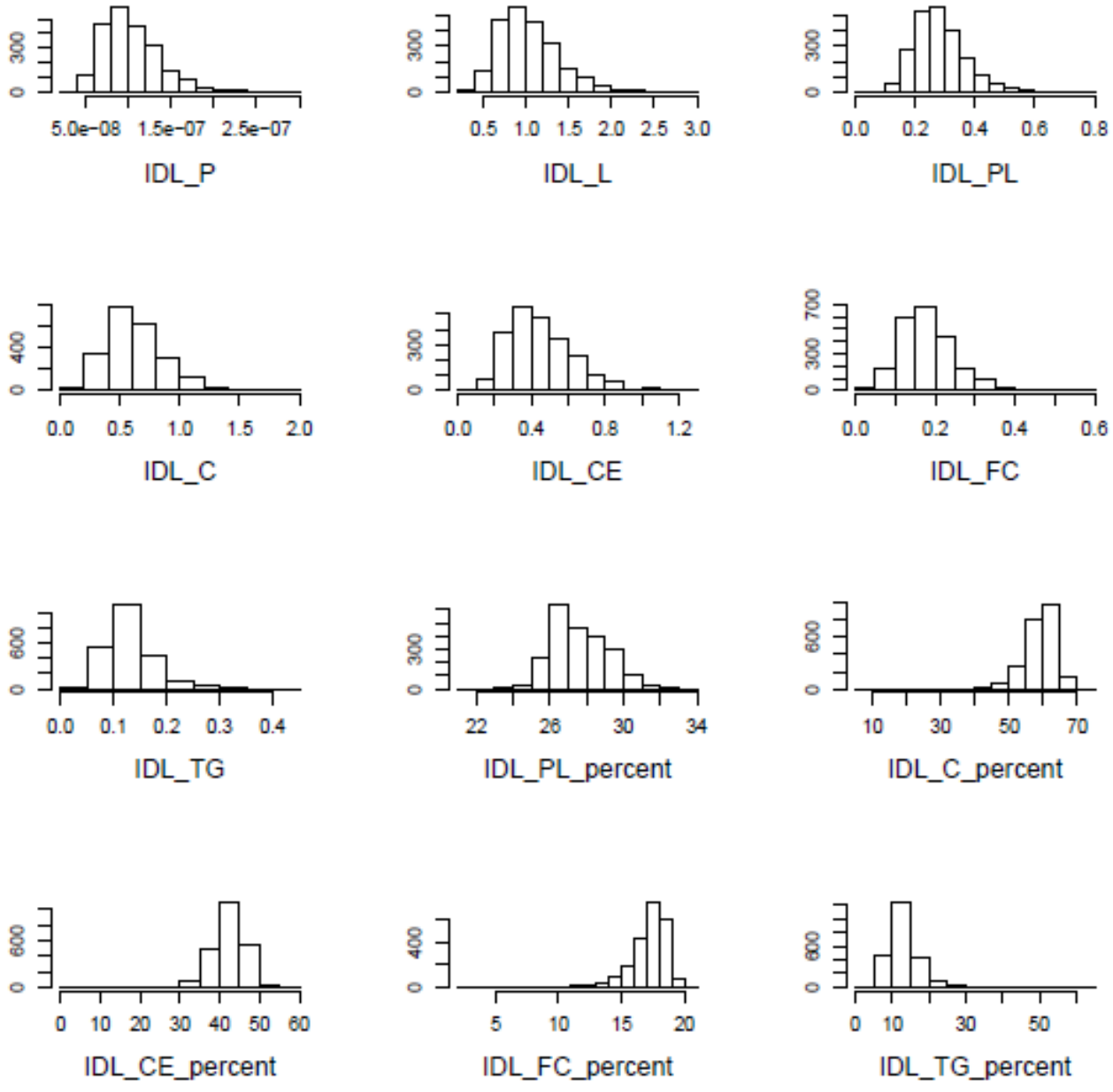

0 300
15 20 25 30 35 40
XS_VLDL_PL_percent

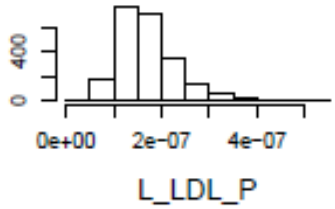

0 300
10 20 30 40 50 60 70
XS_VLDL_C_percent


0 300
0 10 20 30 40 50
XS_VLDL_CE_percent

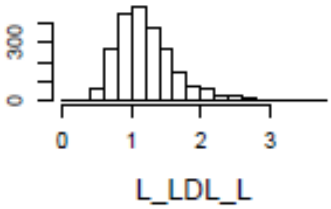

0 400
5 10 15 20
XS_VLDL_FC_percent


0 300 700
10 30 50 70
XS_VLDL_TG_percent

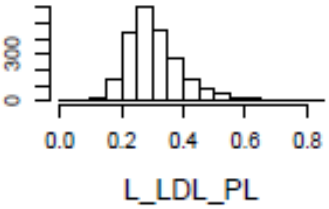

0 300
5.0e-08 1.5e-07 2.5e-07
IDL_P
0 300
0.5 1.0 1.5 2.0 2.5 3.0
IDL_L
0 300
0.0 0.2 0.4 0.6 0.8
IDL_PL
0 400
0.0 0.5 1.0 1.5 2.0
IDL_C
0 300
0.0 0.4 0.8 1.2
IDL_CE
0 300 700
0.0 0.2 0.4 0.6
IDL_FC
0 600
0.0 0.1 0.2 0.3 0.4
IDL_TG
0 300
22 26 30 34
IDL_PL_percent
0 600
10 30 50 70
IDL_C_percent
0 600
0 10 20 30 40 50 60
IDL_CE_percent
0 400
5 10 15 20
IDL_FC_percent
0 600
0 10 30 50
IDL_TG_percent



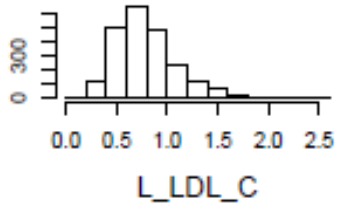
L_LDL_P



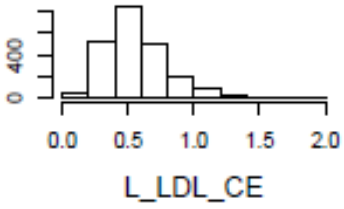
L_LDL_L



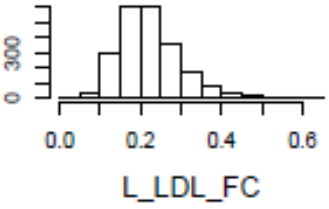
L_LDL_PL



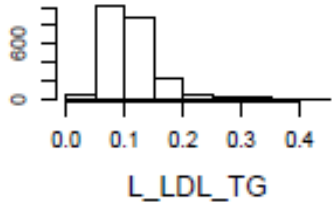
L_LDL_C



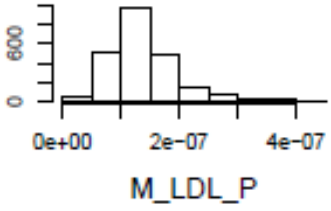
L_LDL_CE



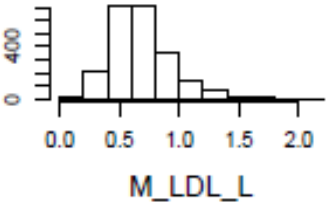
L_LDL_FC



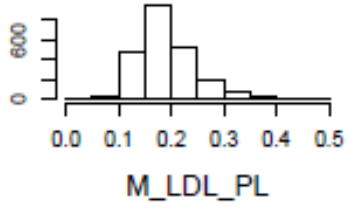
L_LDL_TG



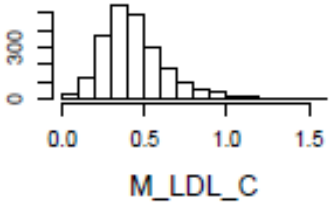
M_LDL_P



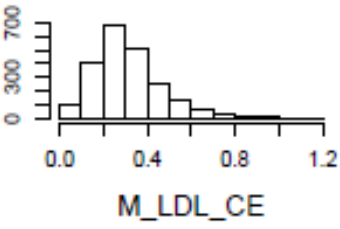
M_LDL_L



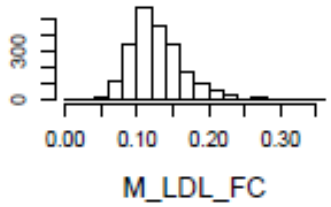
M_LDL_PL

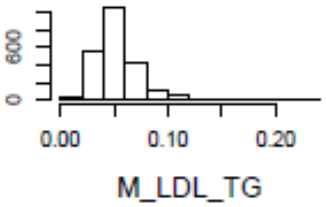


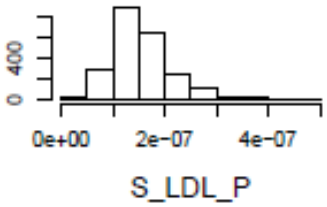
M_LDL_C

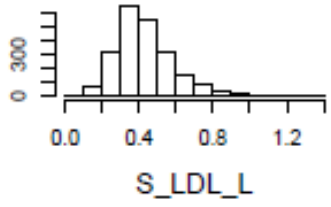


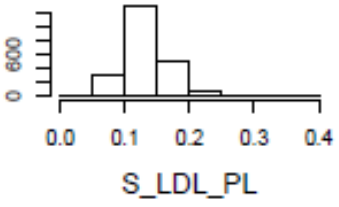
M_LDL_CE

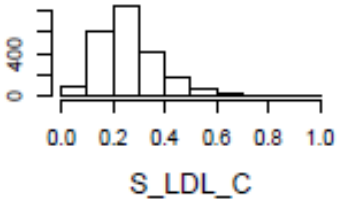

0
300
0.00
0.10
0.20
0.30
M_LDL_FC

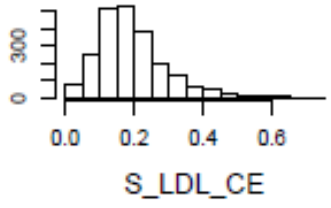

0
600
0.00
0.10
0.20
M_LDL_TG

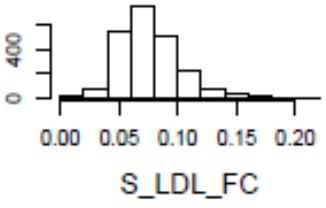

0
400
0e+00
2e−07
4e−07
S_LDL_P

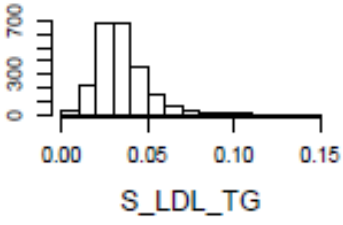

0
300
0.0
0.4
0.8
1.2
S_LDL_L

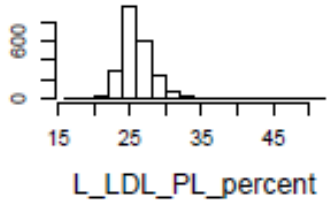

0
600
0.0
0.1
0.2
0.3
0.4
S_LDL_PL

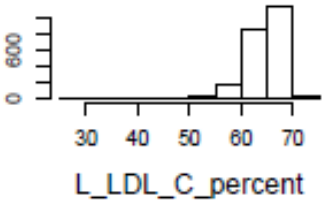

0
400
0.0
0.2
0.4
0.6
0.8
1.0
S_LDL_C

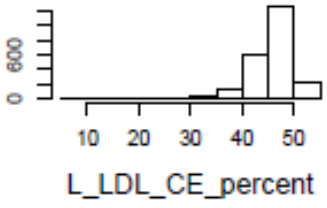

0
300
0.0
0.2
0.4
0.6
S_LDL_CE

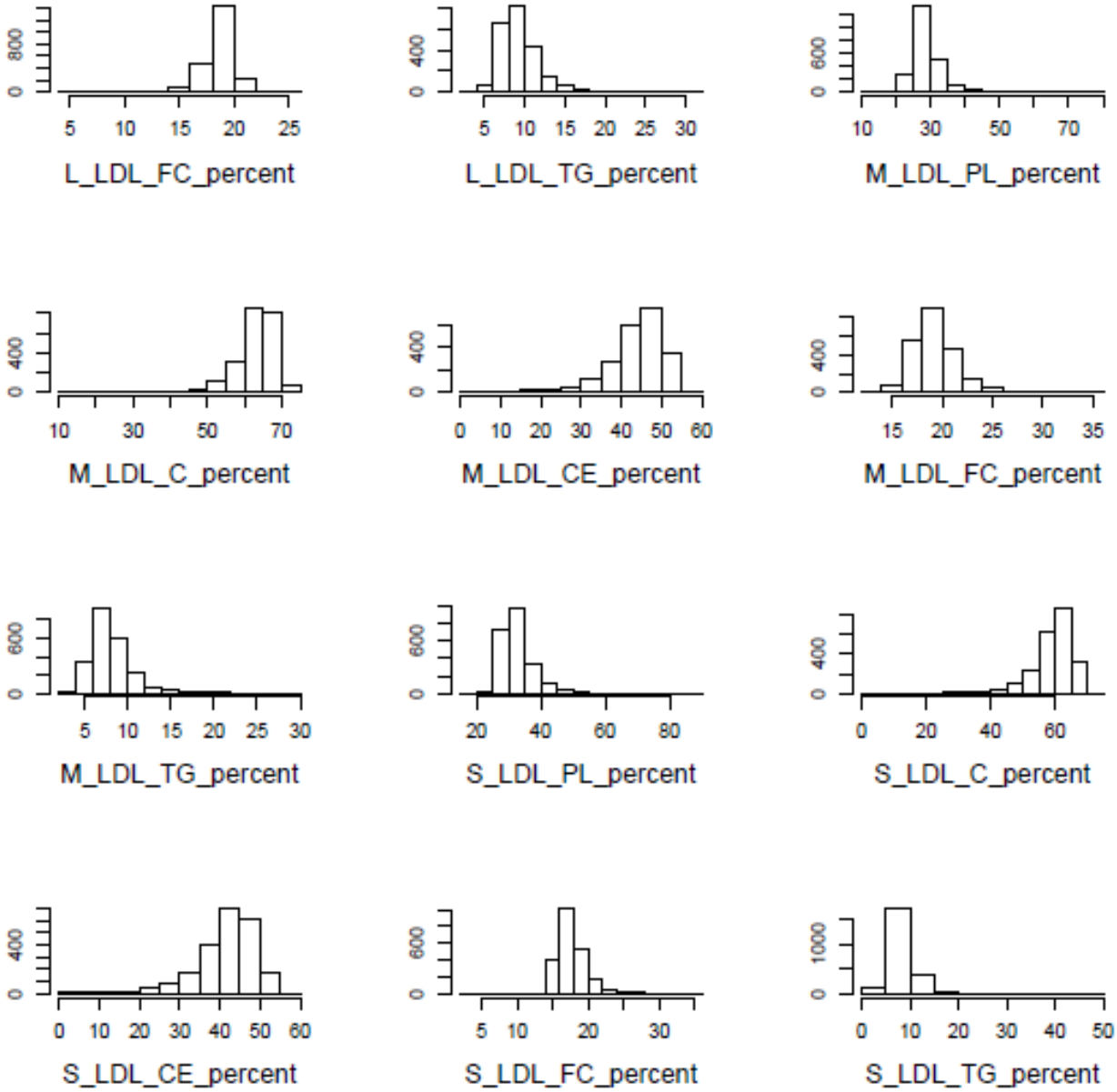

0
400
0.00
0.05
0.10
0.15
0.20
S_LDL_FC

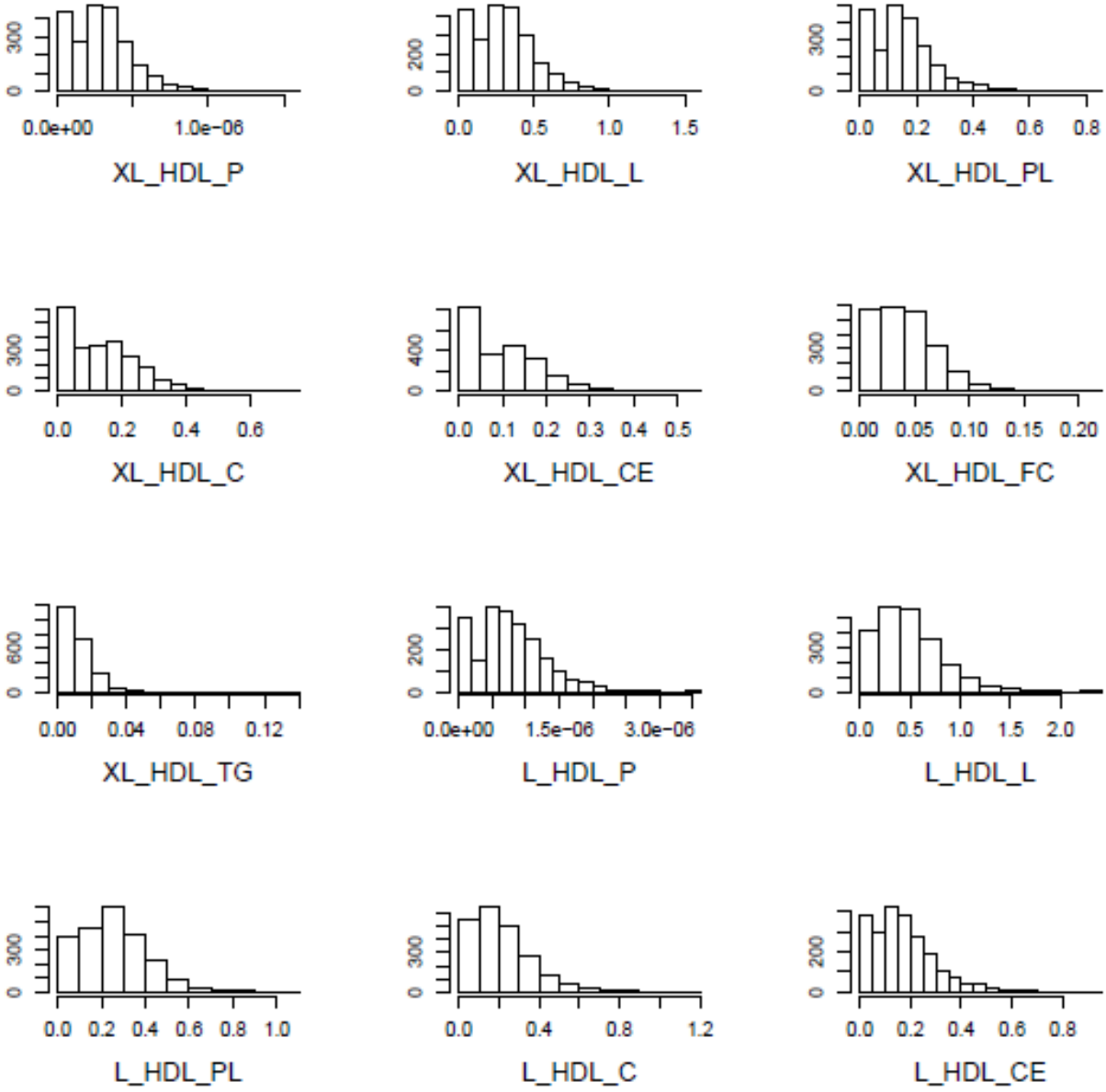

0
300
700
0.00
0.05
0.10
0.15
S_LDL_TG

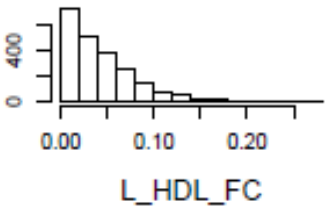

0
600
15
25
35
45
L_LDL_PL_percent

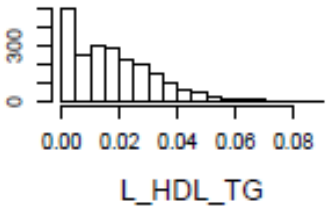

0
600
30
40
50
60
70
L_LDL_C_percent

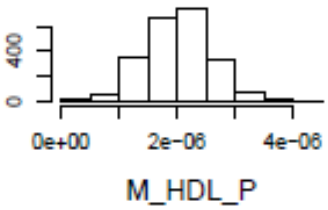

0
600
10
20
30
40
50
L_LDL_CE_percent

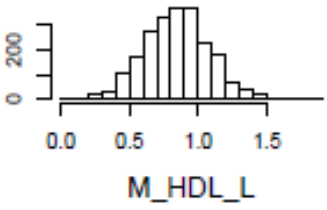

0 800
5 10 15 20 25
L_LDL_FC_percent
0 400
5 10 15 20 25 30
L_LDL_TG_percent
0 600
10 30 50 70
M_LDL_PL_percent
0 400
10 30 50 70
M_LDL_C_percent
0 400
0 10 20 30 40 50 60
M_LDL_CE_percent
0 400
15 20 25 30 35
M_LDL_FC_percent
0 600
5 10 15 20 25 30
M_LDL_TG_percent
0 600
20 40 60 80
S_LDL_PL_percent
0 400
0 20 40 60
S_LDL_C_percent
0 400
0 10 20 30 40 50 60
S_LDL_CE_percent
0 600
5 10 20 30
S_LDL_FC_percent
0 1000
0 10 20 30 40 50
S_LDL_TG_percent

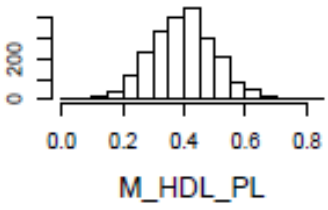

0
300
0.0e+00
1.0e−06
XL_HDL_P
0
200
0.0
0.5
1.0
1.5
XL_HDL_L
0
300
0.0
0.2
0.4
0.6
0.8
XL_HDL_PL
0
300
0.0
0.2
0.4
0.6
XL_HDL_C
0
400
0.0
0.1
0.2
0.3
0.4
0.5
XL_HDL_CE
0
300
0.00
0.05
0.10
0.15
0.20
XL_HDL_FC
0
600
0.00
0.04
0.08
0.12
XL_HDL_TG
0
200
0.0e+00
1.5e−06
3.0e−06
L_HDL_P
0
300
0.0
0.5
1.0
1.5
2.0
L_HDL_L
0
300
0.0
0.2
0.4
0.6
0.8
1.0
L_HDL_PL
0
300
0.0
0.4
0.8
1.2
L_HDL_C
0
200
0.0
0.2
0.4
0.6
0.8
L_HDL_CE

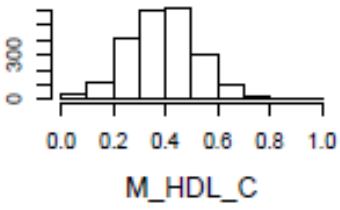

0
400
0.00
0.10
0.20
L_HDL_FC

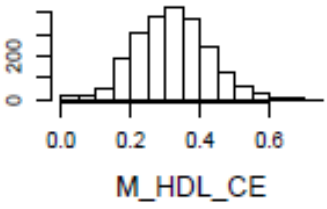

0
300
0.00
0.02
0.04
0.06
0.08
L_HDL_TG

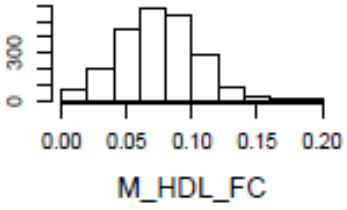

0
400
0e+00
2e−06
4e−06
M_HDL_P

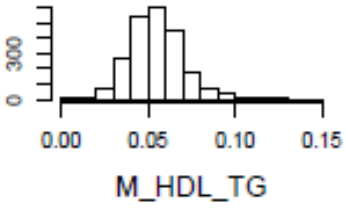

0
200
0.0
0.5
1.0
1.5
M_HDL_L

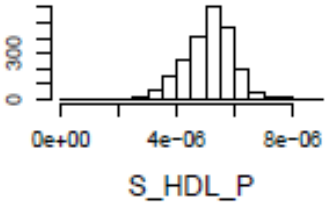

0
200
0.0
0.2
0.4
0.6
0.8
M_HDL_PL

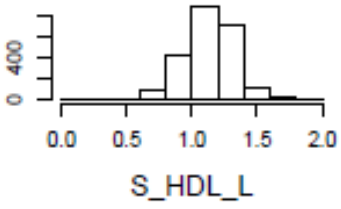

0
300
0.0
0.2
0.4
0.6
0.8
1.0
M_HDL_C

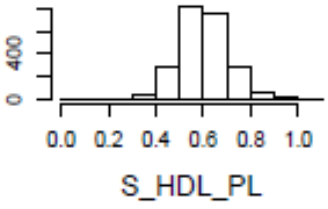

0
200
0.0
0.2
0.4
0.6
M_HDL_CE

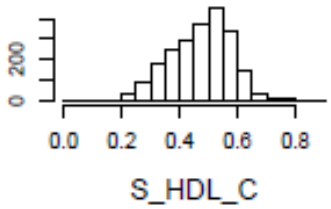

0
300
0.00
0.05
0.10
0.15
0.20
M_HDL_FC

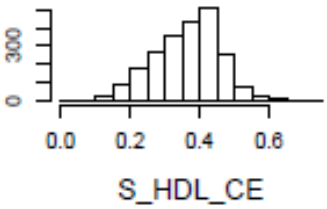

0
300
0.00
0.05
0.10
0.15
M_HDL_TG

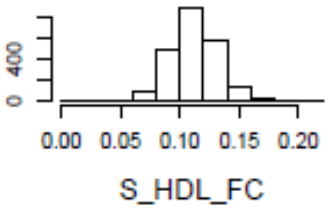

0
300
0e+00
4e−06
8e−06
S_HDL_P

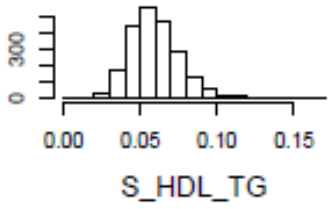

0
400
0.0
0.5
1.0
1.5
2.0
S_HDL_L

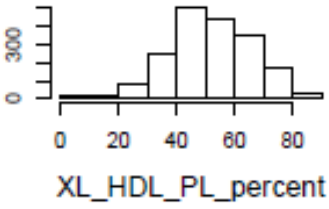

0
400
0.0
0.2
0.4
0.6
0.8
1.0
S_HDL_PL

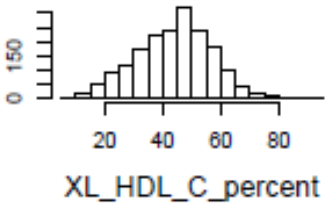

0 200
0.0 0.2 0.4 0.6 0.8
S_HDL_C

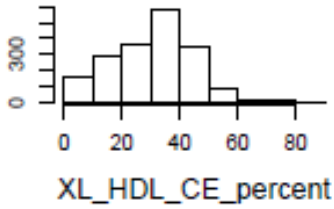

0 300
0.0 0.2 0.4 0.6
S_HDL_CE

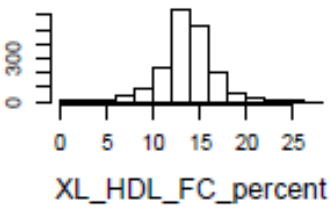

0 400
0.00 0.05 0.10 0.15 0.20
S_HDL_FC

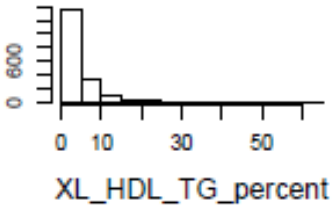

0 300
0.00 0.05 0.10 0.15
S_HDL_TG

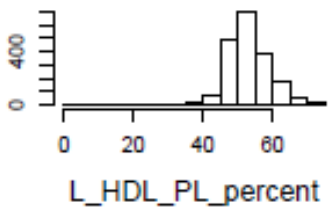

0 300
0 20 40 60 80
XL_HDL_PL_percent

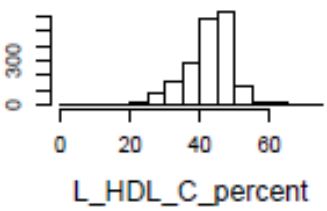

0 150
20 40 60 80
XL_HDL_C_percent

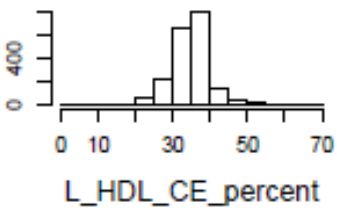

0 300
0 20 40 60 80
XL_HDL_CE_percent

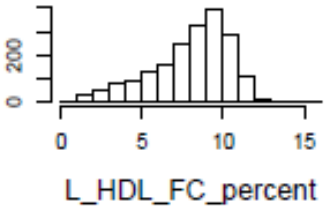

0 300
0 5 10 15 20 25
XL_HDL_FC_percent

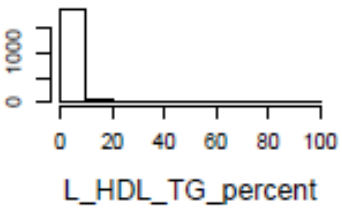

0 600
0 10 30 50
XL_HDL_TG_percent

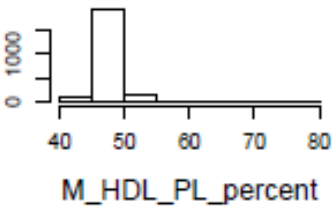

0 400
0 20 40 60
L_HDL_PL_percent

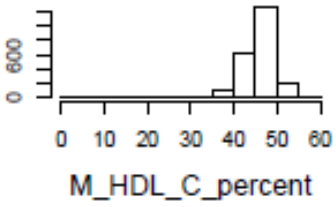

0 300
0 20 40 60
L_HDL_C_percent

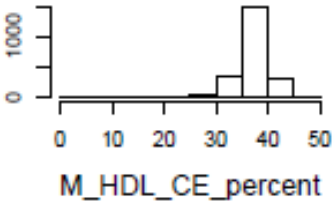

0 400
0 10 30 50 70
L_HDL_CE_percent

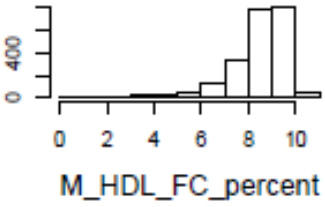

0
200
0
5
10
15
L_HDL_FC_percent

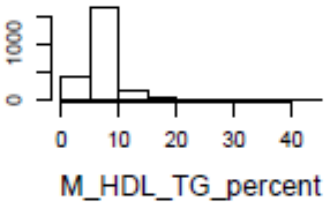

0
1000
0
20
40
60
80
100
L_HDL_TG_percent

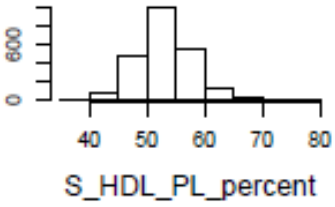

0
1000
40
50
60
70
80
M_HDL_PL_percent

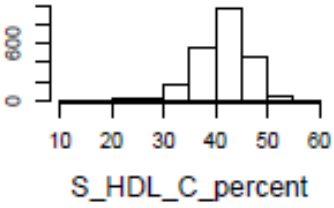

0
600
0
10
20
30
40
50
60
M_HDL_C_percent

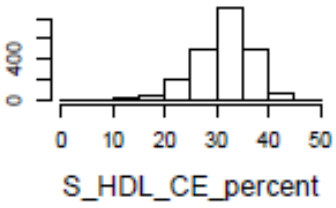

0
1000
0
10
20
30
40
50
M_HDL_CE_percent

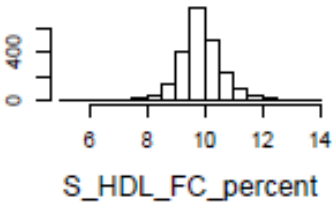

0
400
0
2
4
6
8
10
M_HDL_FC_percent

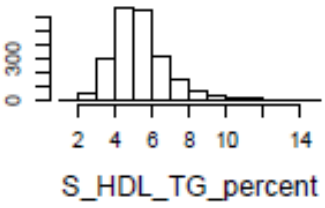

0
1000
0
10
20
30
40
M_HDL_TG_percent

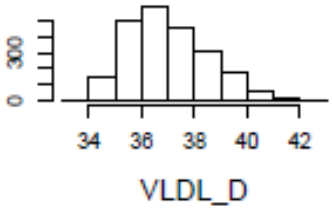

0
600
40
50
60
70
80
S_HDL_PL_percent

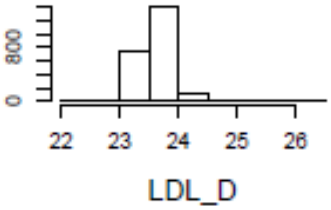

0
600
10
20
30
40
50
60
S_HDL_C_percent

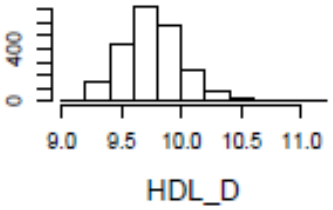

0
400
0
10
20
30
40
50
S_HDL_CE_percent

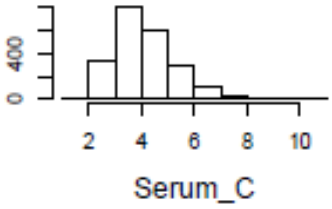

0
400
6
8
10
12
14
S_HDL_FC_percent

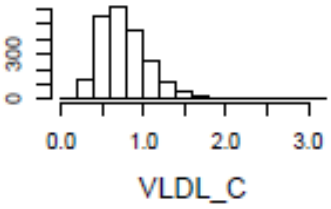

0
300
2
4
6
8
10
14
S_HDL_TG_percent

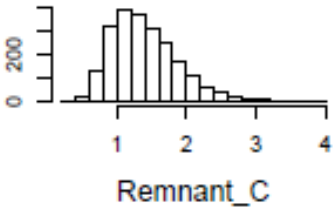

0
300
34
36
38
40
42
VLDL_D

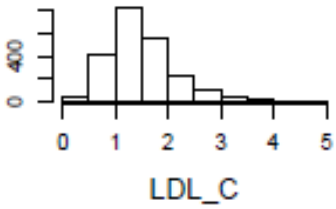

0
800
22
23
24
25
26
LDL_D

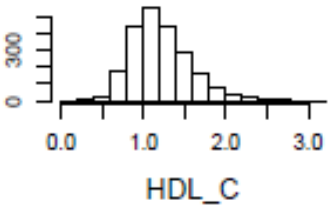

0
400
9.0
9.5
10.0
10.5
11.0
HDL_D

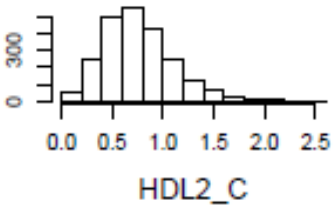

0
400
2
4
6
8
10
Serum_C

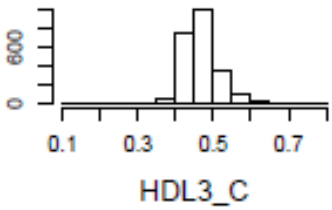

0
300
0.0
1.0
2.0
3.0
VLDL_C

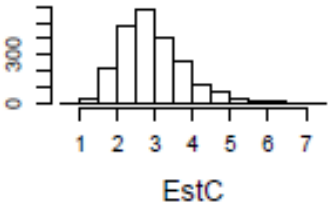

0
200
1
2
3
4
Remnant_C

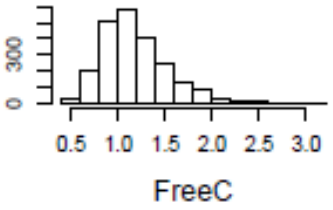

0
400
0
1
2
3
4
5
LDL_C

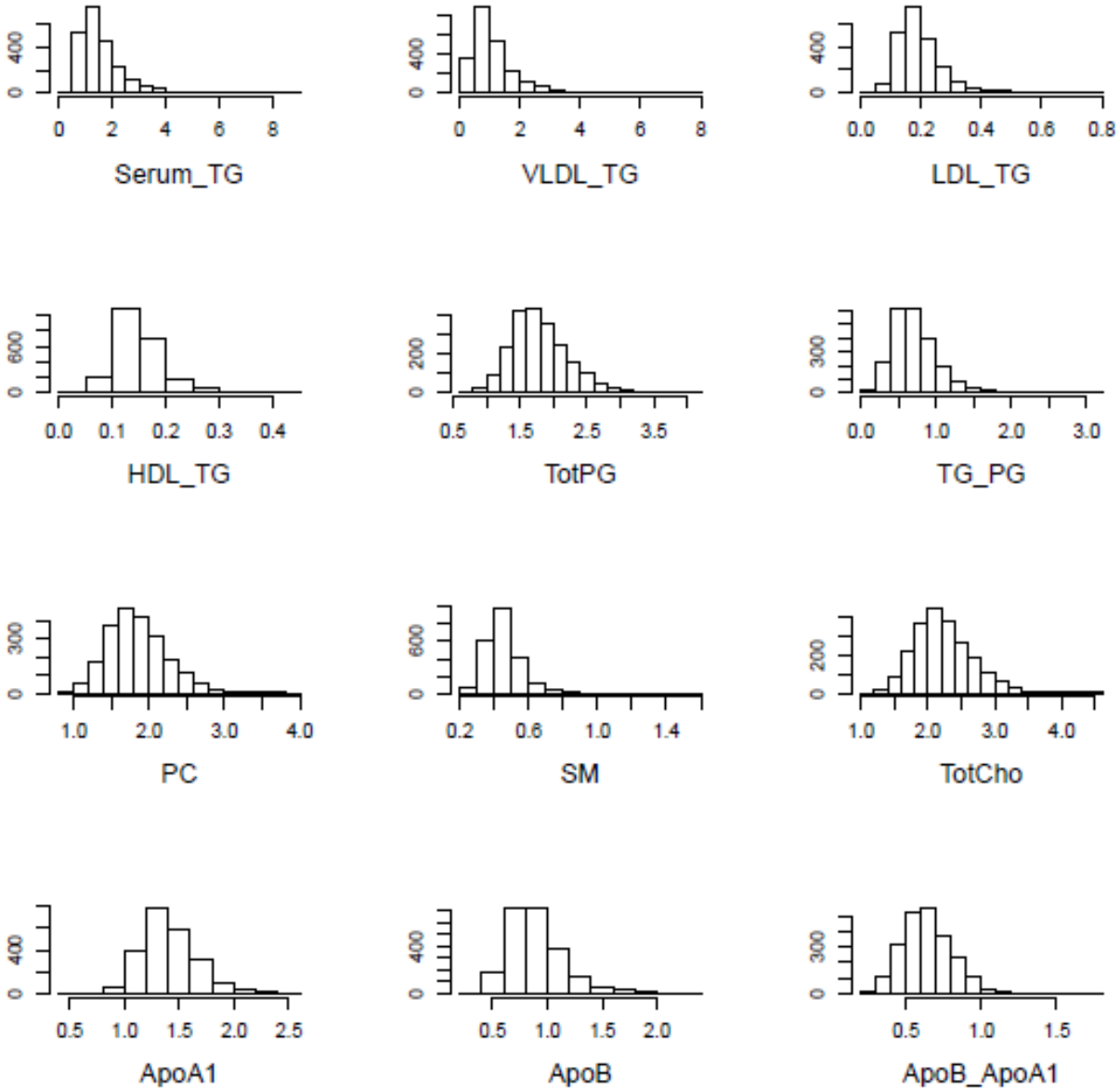

0
300
0.0
1.0
2.0
3.0
HDL_C

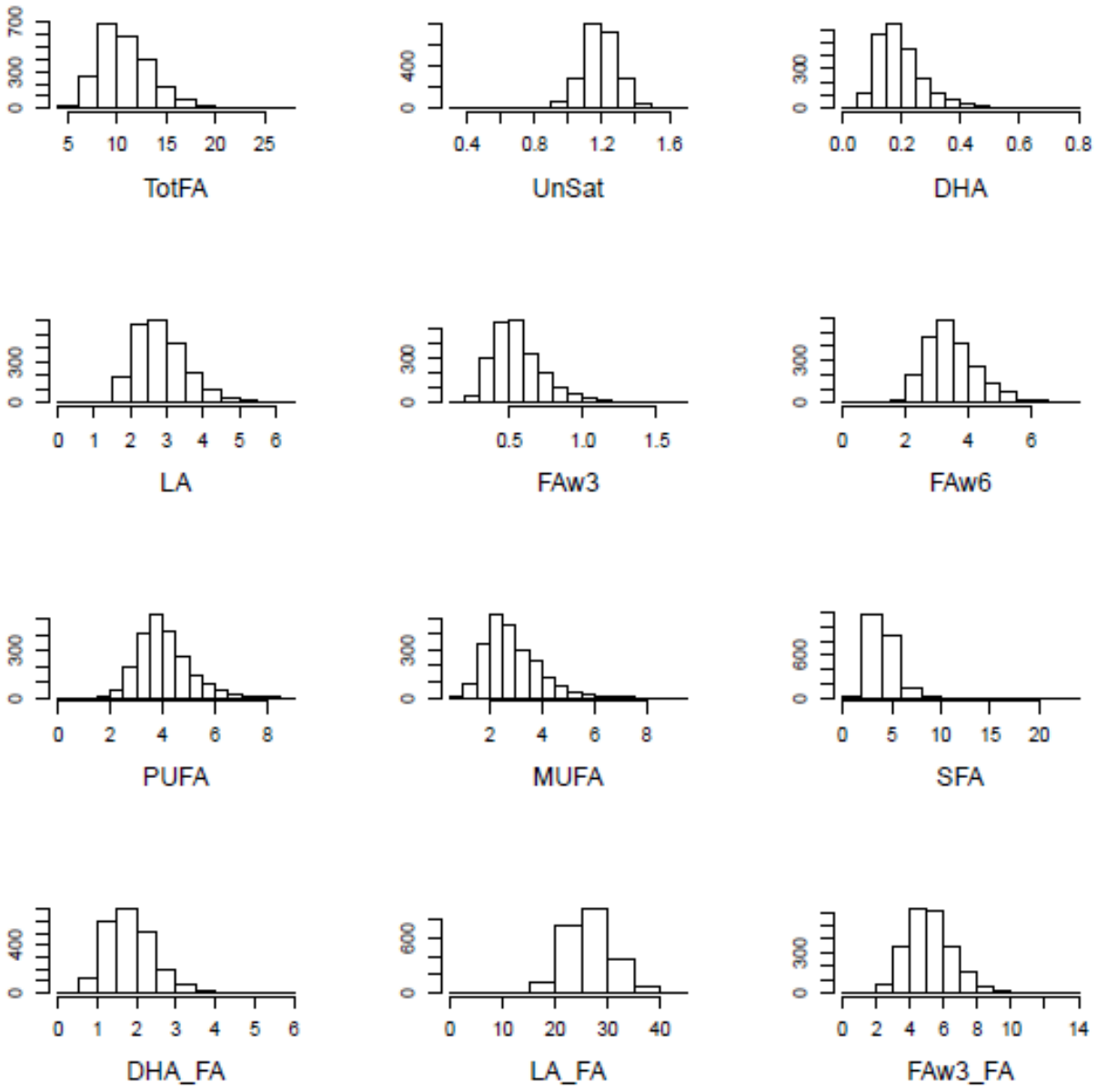

0
300
0.0
0.5
1.0
1.5
2.0
2.5
HDL2_C

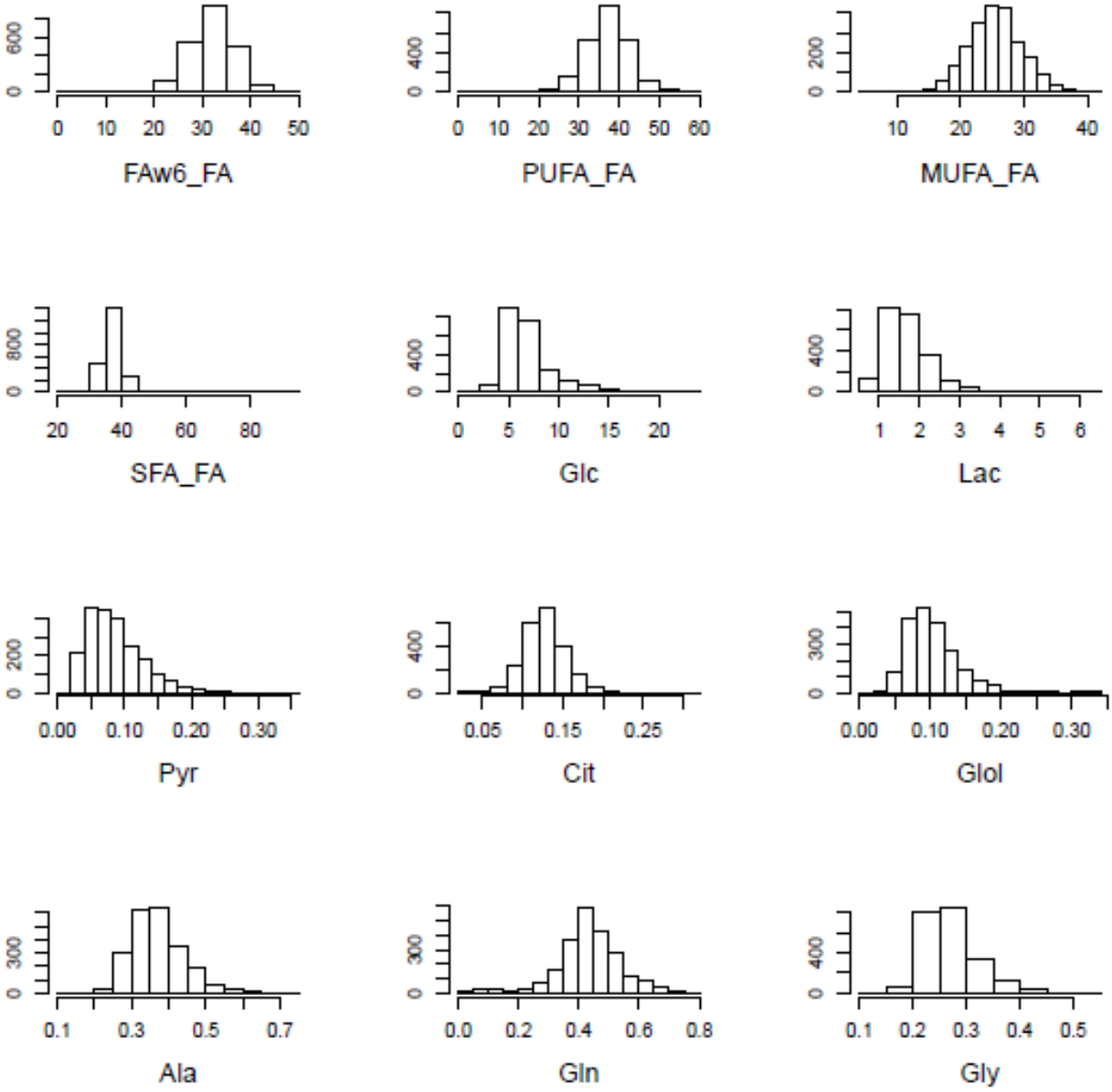

0
600
0.1
0.3
0.5
0.7
HDL3_C


0
300
1
2
3
4
5
6
7
EstC


0
300
0.5
1.0
1.5
2.0
2.5
3.0
FreeC


0 400
0 2 4 6 8
Serum_TG
0 400
0 2 4 6 8
VLDL_TG
0 400
0.0 0.2 0.4 0.6 0.8
LDL_TG
0 600
0.0 0.1 0.2 0.3 0.4
HDL_TG
0 200
0.5 1.5 2.5 3.5
TotPG
0 300
0.0 1.0 2.0 3.0
TG_PG
0 300
1.0 2.0 3.0 4.0
PC
0 600
0.2 0.6 1.0 1.4
SM
0 200
1.0 2.0 3.0 4.0
TotCho
0 400
0.5 1.0 1.5 2.0 2.5
ApoA1
0 400
0.5 1.0 1.5 2.0
ApoB
0 300
0.5 1.0 1.5
ApoB_ApoA1


0 300 700
5 10 15 20 25
TotFA
0 400
0.4 0.8 1.2 1.6
UnSat
0 300
0.0 0.2 0.4 0.6 0.8
DHA
0 300
0 1 2 3 4 5 6
LA
0 300
0.5 1.0 1.5
FAw3
0 300
0 2 4 6
FAw6
0 300
0 2 4 6 8
PUFA
0 300
2 4 6 8
MUFA
0 600
0 5 10 15 20
SFA
0 400
0 1 2 3 4 5 6
DHA_FA
0 600
0 10 20 30 40
LA_FA
0 300
0 2 4 6 8 10 14
FAw3_FA


FAw6_FA
PUFA_FA
MUFA_FA
SFA_FA
Glc
Lac
Pyr
Cit
Glol
Ala
Gln
Gly

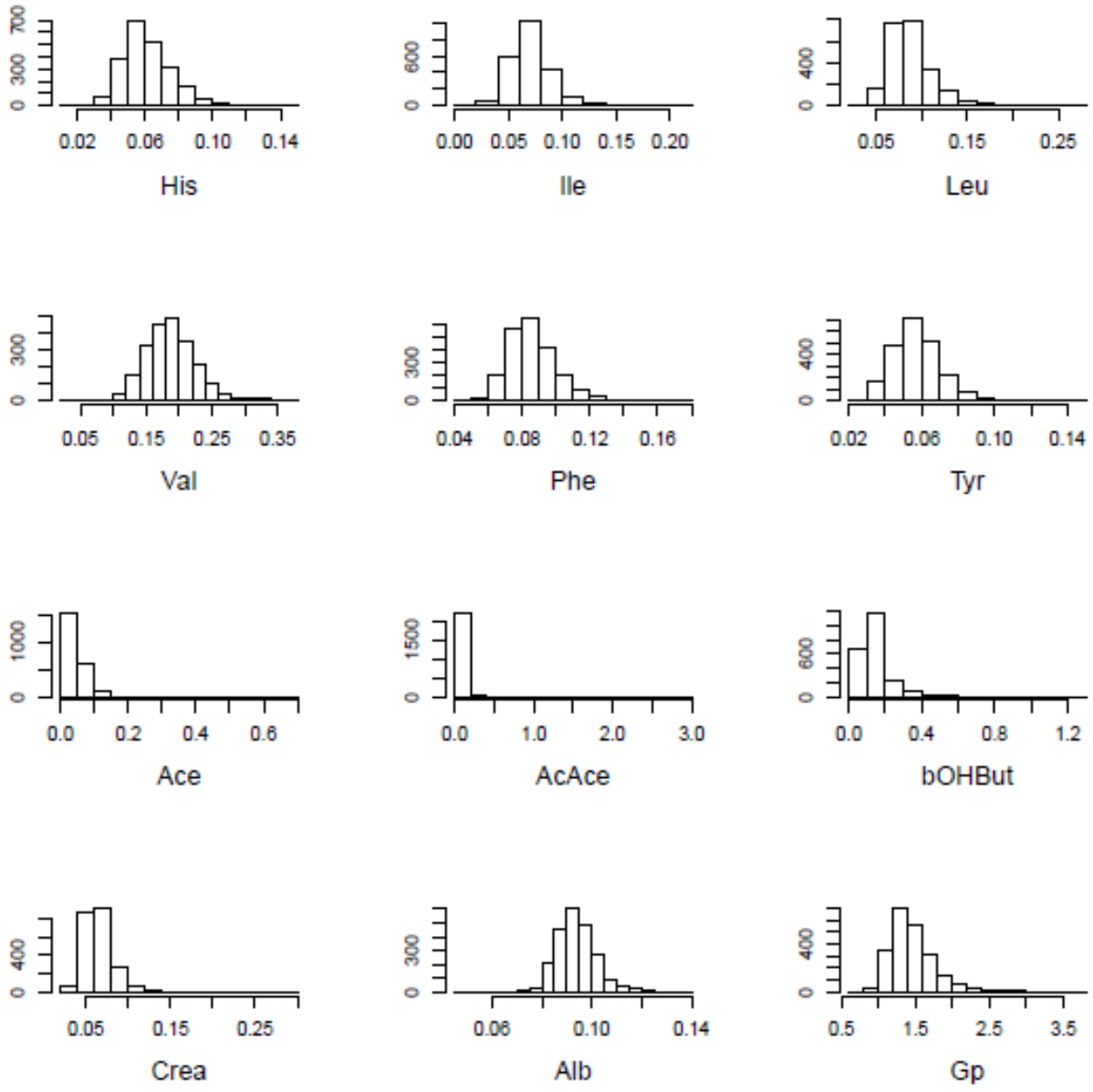

0 300 700
0.02 0.06 0.10 0.14
His
0 600
0.00 0.05 0.10 0.15 0.20
Ile
0 400
0.05 0.15 0.25
Leu
0 300
0.05 0.15 0.25 0.35
Val
0 300
0.04 0.08 0.12 0.16
Phe
0 400
0.02 0.06 0.10 0.14
Tyr
0 1000
0.0 0.2 0.4 0.6
Ace
0 1500
0.0 1.0 2.0 3.0
AcAce
0 600
0.0 0.4 0.8 1.2
bOHBut
0 400
0.05 0.15 0.25
Crea
0 300
0.06 0.10 0.14
Alb
0 400
0.5 1.5 2.5 3.5
Gp

Table C-1 Medians, interquartile ranges (IQR) and missing values (NA; blank cell indicates no missing values) for 228 metabolites in the five individual UCLEB studies (full names of metabolites can be found in Table C-2)

		BRHS			BWHHS			ET2DS			SABRE			WH2		
Metabolite	Units	Median	IQR	NA	Median	IQR	NA	Median	IQR	NA	Median	IQR	NA	Median	IQR	NA
XXL VLDL P	mol/L	0.000	0.000		0.000	0.000		0.000	0.000		0.000	0.000		0.000	0.000	
XXL VLDL L	mmol/L	0.043	0.062		0.037	0.043		0.015	0.030		0.034	0.030		0.028	0.025	
XXL VLDL PL	mmol/L	0.005	0.008		0.005	0.006		0.001	0.003		0.003	0.004		0.003	0.003	
XXL VLDL C	mmol/L	0.008	0.010		0.007	0.009		0.003	0.005		0.007	0.006		0.005	0.004	
XXL VLDL CE	mmol/L	0.005	0.006		0.004	0.005		0.002	0.003		0.005	0.004		0.003	0.002	
XXL VLDL FC	mmol/L	0.003	0.005		0.003	0.004		0.001	0.002		0.002	0.002		0.002	0.002	
XXL VLDL TG	mmol/L	0.030	0.046		0.025	0.030		0.010	0.021		0.023	0.020		0.019	0.018	
XL VLDL P	mol/L	0.000	0.000		0.000	0.000		0.000	0.000		0.000	0.000		0.000	0.000	
XL VLDL L	mmol/L	0.119	0.149		0.104	0.149		0.051	0.091		0.054	0.060		0.054	0.064	
XL VLDL PL	mmol/L	0.018	0.025		0.018	0.025		0.007	0.014		0.010	0.011		0.009	0.010	
XL VLDL C	mmol/L	0.022	0.029		0.020	0.027		0.008	0.015		0.018	0.017		0.012	0.013	
XL VLDL CE	mmol/L	0.013	0.015		0.010	0.014		0.005	0.009		0.011	0.008		0.007	0.007	
XL VLDL FC	mmol/L	0.009	0.013		0.009	0.013		0.003	0.007		0.008	0.008		0.006	0.006	
XL VLDL TG	mmol/L	0.076	0.098		0.066	0.093		0.036	0.064		0.025	0.032		0.032	0.041	
L VLDL P	mol/L	0.000	0.000		0.000	0.000		0.000	0.000		0.000	0.000		0.000	0.000	
L VLDL L	mmol/L	0.472	0.451		0.418	0.489		0.290	0.316		0.155	0.146		0.239	0.241	
L VLDL PL	mmol/L	0.085	0.083		0.079	0.090		0.053	0.057		0.028	0.028		0.043	0.042	
L VLDL C	mmol/L	0.104	0.104		0.092	0.106		0.062	0.065		0.056	0.045		0.058	0.056	
L VLDL CE	mmol/L	0.056	0.055		0.048	0.054		0.036	0.031		0.038	0.025		0.034	0.026	
L VLDL FC	mmol/L	0.047	0.055		0.045	0.057		0.026	0.035		0.017	0.022		0.024	0.028	
L VLDL TG	mmol/L	0.285	0.275		0.240	0.293		0.174	0.192		0.071	0.075		0.138	0.139	
M VLDL P	mol/L	0.000	0.000		0.000	0.000		0.000	0.000		0.000	0.000		0.000	0.000	

M VLDL L	mmol/L	0.862	0.636		0.836	0.765		0.636	0.490		0.414	0.254		0.605	0.388	
M VLDL PL	mmol/L	0.171	0.119		0.173	0.143		0.128	0.092		0.088	0.053		0.123	0.077	
M VLDL C	mmol/L	0.220	0.170		0.228	0.177		0.165	0.112		0.148	0.091		0.187	0.106	
M VLDL CE	mmol/L	0.120	0.093		0.131	0.094		0.092	0.054		0.101	0.055		0.114	0.051	
M VLDL FC	mmol/L	0.099	0.079		0.104	0.094		0.072	0.060		0.047	0.035		0.072	0.052	
M VLDL TG	mmol/L	0.482	0.379		0.435	0.417		0.339	0.287		0.177	0.120		0.302	0.222	
S VLDL P	mol/L	0.000	0.000		0.000	0.000		0.000	0.000		0.000	0.000		0.000	0.000	
S VLDL L	mmol/L	0.748	0.336		0.872	0.469		0.617	0.290		0.581	0.232		0.734	0.291	
S VLDL PL	mmol/L	0.172	0.065		0.197	0.094		0.150	0.059		0.131	0.047		0.171	0.063	
S VLDL C	mmol/L	0.249	0.110		0.317	0.155		0.192	0.078		0.250	0.099		0.283	0.091	
S VLDL CE	mmol/L	0.148	0.070		0.192	0.094		0.106	0.048		0.169	0.069		0.177	0.054	
S VLDL FC	mmol/L	0.098	0.040		0.121	0.065		0.084	0.040		0.078	0.033		0.105	0.042	
S VLDL TG	mmol/L	0.325	0.188		0.362	0.233		0.276	0.155		0.202	0.093		0.280	0.145	
XS VLDL P	mol/L	0.000	0.000		0.000	0.000		0.000	0.000		0.000	0.000		0.000	0.000	
XS VLDL L	mmol/L	0.513	0.180		0.648	0.244		0.428	0.144		0.551	0.182		0.634	0.166	
XS VLDL PL	mmol/L	0.151	0.058		0.200	0.080		0.133	0.048		0.145	0.052		0.183	0.048	
XS VLDL C	mmol/L	0.235	0.086		0.298	0.121		0.175	0.066		0.300	0.101		0.319	0.077	
XS VLDL CE	mmol/L	0.159	0.056		0.199	0.078		0.110	0.047		0.221	0.069		0.219	0.054	
XS VLDL FC	mmol/L	0.076	0.029		0.099	0.042		0.065	0.025		0.081	0.028		0.100	0.024	
XS VLDL TG	mmol/L	0.126	0.045		0.162	0.077		0.117	0.053		0.105	0.047		0.136	0.056	
XXL VLDL PL %	%	11.600	1.280	7	12.000	1.300	17	10.520	3.229	265	10.080	2.199		11.450	1.270	3
XXL VLDL C %	%	16.750	2.950	7	18.000	4.800	17	17.910	5.700	265	20.950	3.058		17.240	1.230	3
XXL VLDL CE %	%	10.100	2.968	7	10.600	4.930	17	11.050	6.657	265	14.005	3.623		10.170	1.457	3
XXL VLDL FC %	%	7.090	1.147	7	7.700	1.600	17	7.034	2.623	265	7.075	1.295		7.277	1.101	3
XXL VLDL TG %	%	71.550	3.170	7	70.200	4.750	17	71.820	5.440	265	69.090	3.410		71.300	1.630	3
XL VLDL PL %	%	16.100	1.280	7	17.000	2.100	21	14.880	2.265	243	17.570	2.450		16.810	1.515	13

XL VLDL C %	%	19.100	1.920	7	18.900	4.000	21	15.950	5.355	243	34.300	9.680		22.170	4.900	13
XL VLDL CE %	%	10.600	1.648	7	9.380	3.080	21	9.606	3.440	243	19.590	7.695		11.750	3.120	13
XL VLDL FC %	%	8.460	1.252	7	9.220	1.930	21	6.341	2.737	243	14.255	2.676		10.400	2.070	13
XL VLDL TG %	%	64.600	2.470	7	63.700	5.250	21	69.080	6.590	243	47.785	9.000		61.230	6.150	13
L VLDL PL %	%	18.100	0.450	2	18.800	1.200	7	18.240	0.640	57	17.883	1.593		18.060	0.750	7
L VLDL C %	%	21.900	1.800	2	22.000	3.500	7	21.360	3.300	57	36.440	9.760		24.610	3.130	7
L VLDL CE %	%	11.900	2.050	2	11.400	3.500	7	12.530	4.080	57	25.285	11.687		14.150	3.780	7
L VLDL FC %	%	10.000	1.380	2	10.600	1.100	7	9.084	2.364	57	11.435	3.599		10.530	1.349	7
L VLDL TG %	%	60.100	2.000	2	59.000	3.700	7	60.270	3.780	57	45.810	9.343		57.470	3.400	7
M VLDL PL %	%	19.700	0.900		20.250	1.300	4	20.240	1.300		21.030	0.950		20.270	0.860	1
M VLDL C %	%	25.600	3.800		27.900	5.300	4	26.050	4.710		35.200	4.970		30.520	4.610	1
M VLDL CE %	%	14.100	3.800		15.900	5.350	4	14.855	5.335		24.060	5.566		18.840	5.430	1
M VLDL FC %	%	11.500	0.700		12.000	0.700	4	11.270	1.030		11.330	1.592		11.850	0.920	1
M VLDL TG %	%	54.700	4.400		51.900	6.330	4	53.660	5.680		43.705	5.352		49.190	5.630	1
S VLDL PL %	%	22.800	1.500		22.700	2.050	1	24.300	2.130		22.480	1.860		23.120	1.120	1
S VLDL C %	%	33.000	6.000		36.800	7.850	1	31.000	6.000		42.560	4.920		38.920	6.130	1
S VLDL CE %	%	19.600	5.700		22.800	7.200	1	17.485	5.817		29.055	5.323		24.840	5.880	1
S VLDL FC %	%	13.200	0.800		14.100	1.100	1	13.520	0.840		13.440	0.850		14.200	0.510	1
S VLDL TG %	%	43.900	6.500		40.400	8.850	1	44.590	7.030		34.830	4.850		37.590	7.240	1
XS VLDL PL %	%	29.200	3.000		30.600	2.500	3	31.180	2.790	2	26.300	2.800		29.090	1.920	1
XS VLDL C %	%	46.000	4.400		45.700	6.800	3	41.340	7.700	2	54.410	5.140		50.320	4.460	1
XS VLDL CE %	%	30.800	4.100		30.300	6.400	3	26.000	7.580	2	39.740	5.488		34.600	4.000	1
XS VLDL FC %	%	15.000	1.600		15.600	1.600	3	15.340	1.494	2	14.490	1.467		15.780	1.040	1
XS VLDL TG %	%	24.800	6.400		23.600	7.700	3	27.360	8.500	2	19.160	4.570		20.760	5.570	1
IDL P	mol/L	0.000	0.000		0.000	0.000	1	0.000	0.000		0.000	0.000		0.000	0.000	1
IDL L	mmol/L	1.010	0.365		1.390	0.505	1	0.825	0.281		1.060	0.358		1.245	0.307	1

IDL PL	mmol/L	0.271	0.100		0.369	0.131		0.238	0.073		0.281	0.095		0.330	0.077	
IDL C	mmol/L	0.634	0.244		0.856	0.344		0.475	0.184		0.660	0.241		0.779	0.211	
IDL CE	mmol/L	0.460	0.157		0.615	0.251		0.331	0.124		0.476	0.179		0.558	0.152	
IDL FC	mmol/L	0.170	0.080		0.244	0.103		0.143	0.057		0.182	0.067		0.224	0.061	
IDL TG	mmol/L	0.117	0.036		0.161	0.068		0.113	0.041		0.114	0.055		0.138	0.042	
IDL PL %	%	26.700	1.120	1	26.650	0.900	2	28.800	1.560		26.540	0.920		26.340	1.080	2
IDL C %	%	61.650	3.460	1	61.950	3.600	2	57.510	4.640		62.595	4.327		62.830	3.050	2
IDL CE %	%	44.700	3.220	1	44.000	3.540	2	40.295	3.992		45.285	3.650		44.870	2.520	2
IDL FC %	%	16.850	2.520	1	17.700	1.770	2	17.310	1.830		17.335	1.217		18.050	1.100	2
IDL TG %	%	11.450	3.220	1	11.500	3.800	2	13.625	4.472		10.850	3.915		10.835	3.130	2
L LDL P	mol/L	0.000	0.000		0.000	0.000		0.000	0.000		0.000	0.000		0.000	0.000	
L LDL L	mmol/L	1.210	0.479		1.660	0.680		0.980	0.353		1.148	0.442		1.397	0.394	
L LDL PL	mmol/L	0.302	0.095		0.397	0.135		0.264	0.072		0.287	0.094		0.349	0.081	
L LDL C	mmol/L	0.801	0.351		1.110	0.484		0.626	0.252		0.752	0.321		0.928	0.287	
L LDL CE	mmol/L	0.578	0.254		0.819	0.376		0.444	0.187		0.532	0.246		0.670	0.218	
L LDL FC	mmol/L	0.221	0.087		0.301	0.119		0.183	0.065		0.219	0.074		0.260	0.068	
L LDL TG	mmol/L	0.097	0.035		0.142	0.064		0.090	0.033		0.102	0.053		0.116	0.028	
M LDL P	mol/L	0.000	0.000		0.000	0.000		0.000	0.000		0.000	0.000		0.000	0.000	
M LDL L	mmol/L	0.709	0.283		0.983	0.424		0.567	0.214		0.601	0.256		0.773	0.230	
M LDL PL	mmol/L	0.196	0.052		0.255	0.083		0.165	0.042		0.174	0.060		0.214	0.048	
M LDL C	mmol/L	0.472	0.216		0.669	0.323		0.356	0.167		0.376	0.185		0.506	0.184	
M LDL CE	mmol/L	0.341	0.176		0.490	0.268		0.246	0.135		0.252	0.151		0.360	0.153	
M LDL FC	mmol/L	0.133	0.036		0.171	0.057		0.109	0.029		0.123	0.039		0.145	0.032	
M LDL TG	mmol/L	0.049	0.017		0.072	0.036		0.046	0.018		0.047	0.028		0.053	0.014	
S LDL P	mol/L	0.000	0.000		0.000	0.000		0.000	0.000		0.000	0.000		0.000	0.000	
S LDL L	mmol/L	0.458	0.166		0.630	0.263		0.360	0.132		0.376	0.169		0.496	0.136	

S LDL PL	mmol/L	0.142	0.033		0.177	0.056		0.119	0.029		0.120	0.041		0.152	0.029	
S LDL C	mmol/L	0.288	0.123		0.403	0.195		0.211	0.101		0.226	0.115		0.306	0.103	
S LDL CE	mmol/L	0.207	0.103		0.302	0.161		0.149	0.082		0.149	0.091		0.218	0.084	
S LDL FC	mmol/L	0.080	0.022		0.103	0.036		0.061	0.018		0.074	0.027		0.089	0.020	
S LDL TG	mmol/L	0.035	0.013		0.048	0.024		0.030	0.013		0.028	0.019		0.034	0.012	
L LDL PL %	%	25.200	1.750	1	24.200	1.550	1	26.950	2.480		25.090	1.800		24.920	1.160	2
L LDL C %	%	66.800	3.530	1	67.100	2.800	1	63.670	3.910		65.610	4.010		66.680	2.650	2
L LDL CE %	%	48.550	2.900	1	49.200	2.750	1	45.090	4.110		46.200	4.652		47.840	2.720	2
L LDL FC %	%	18.200	1.620	1	18.100	1.400	1	18.730	1.390		19.430	1.200		18.860	0.980	2
L LDL TG %	%	7.955	1.938	1	8.530	2.605	1	9.188	2.758		9.051	3.564		8.255	2.180	2
M LDL PL %	%	27.500	4.000	1	25.600	3.250	1	28.990	4.660		29.180	4.330		27.480	2.780	1
M LDL C %	%	65.700	4.600	1	66.800	3.750	1	62.750	5.590		62.550	7.050		65.570	4.190	1
M LDL CE %	%	47.550	6.720	1	49.600	5.100	1	43.445	7.628		41.820	9.500		46.630	5.620	1
M LDL FC %	%	18.500	2.000	1	17.400	1.800	1	19.210	2.480		20.580	3.010		18.760	1.680	1
M LDL TG %	%	6.640	1.668	1	7.340	2.690	1	8.002	2.499		7.955	3.734		6.815	1.883	1
S LDL PL %	%	30.700	4.620	1	28.700	3.700	1	33.120	5.780	1	32.100	5.030		30.560	3.490	1
S LDL C %	%	61.850	6.250	1	63.600	4.750	1	58.660	6.880	1	60.240	6.997		62.720	4.690	1
S LDL CE %	%	44.750	6.820	1	47.100	5.800	1	41.670	8.040	1	40.010	9.230		44.610	5.830	1
S LDL FC %	%	17.300	1.400	1	16.300	1.600	1	16.900	1.750	1	19.880	3.017		18.110	1.570	1
S LDL TG %	%	7.540	2.427	1	7.630	2.510	1	8.209	2.805	1	7.527	3.850		6.669	2.176	1
XL HDL P	mol/L	0.000	0.000		0.000	0.000		0.000	0.000		0.000	0.000		0.000	0.000	
XL HDL L	mmol/L	0.274	0.181		0.361	0.314		0.180	0.306		0.369	0.141		0.415	0.217	
XL HDL PL	mmol/L	0.108	0.087		0.164	0.168		0.124	0.195		0.147	0.068		0.196	0.122	
XL HDL C	mmol/L	0.144	0.086		0.169	0.161		0.050	0.107		0.210	0.098		0.206	0.102	
XL HDL CE	mmol/L	0.117	0.065		0.122	0.121		0.021	0.062		0.159	0.075		0.148	0.070	
XL HDL FC	mmol/L	0.029	0.025		0.044	0.045		0.029	0.046		0.051	0.027		0.058	0.030	

XL HDL TG	mmol/L	0.017	0.012		0.018	0.013		0.003	0.008		0.012	0.010		0.015	0.007	
L HDL P	mol/L	0.000	0.000		0.000	0.000		0.000	0.000		0.000	0.000		0.000	0.000	
L HDL L	mmol/L	0.316	0.527		0.594	0.514		0.420	0.395		0.343	0.225		0.720	0.322	
L HDL PL	mmol/L	0.175	0.263		0.324	0.237		0.243	0.182		0.168	0.104		0.367	0.149	
L HDL C	mmol/L	0.115	0.235		0.236	0.268		0.164	0.204		0.162	0.108		0.329	0.163	
L HDL CE	mmol/L	0.097	0.188		0.189	0.206		0.132	0.155		0.134	0.081		0.259	0.122	
L HDL FC	mmol/L	0.019	0.047		0.047	0.069		0.031	0.050		0.026	0.027		0.071	0.040	
L HDL TG	mmol/L	0.024	0.032		0.029	0.029		0.012	0.018		0.012	0.013		0.026	0.014	
M HDL P	mol/L	0.000	0.000		0.000	0.000		0.000	0.000		0.000	0.000		0.000	0.000	
M HDL L	mmol/L	0.682	0.177		0.962	0.300		0.890	0.247		0.588	0.230		0.988	0.212	
M HDL PL	mmol/L	0.326	0.088		0.452	0.129		0.416	0.106		0.284	0.103		0.458	0.100	
M HDL C	mmol/L	0.305	0.099		0.445	0.177		0.419	0.141		0.259	0.116		0.480	0.125	
M HDL CE	mmol/L	0.249	0.076		0.357	0.139		0.340	0.113		0.215	0.088		0.388	0.098	
M HDL FC	mmol/L	0.053	0.026		0.089	0.038		0.078	0.029		0.045	0.025		0.094	0.026	
M HDL TG	mmol/L	0.052	0.022		0.063	0.025		0.055	0.016		0.043	0.016		0.058	0.016	
S HDL P	mol/L	0.000	0.000		0.000	0.000		0.000	0.000		0.000	0.000		0.000	0.000	
S HDL L	mmol/L	1.040	0.145		1.230	0.220		1.210	0.165		0.893	0.154		1.125	0.166	
S HDL PL	mmol/L	0.559	0.068		0.629	0.130		0.619	0.109		0.498	0.096		0.646	0.108	
S HDL C	mmol/L	0.435	0.105		0.533	0.133		0.532	0.082		0.339	0.081		0.430	0.075	
S HDL CE	mmol/L	0.337	0.094		0.416	0.116		0.414	0.071		0.245	0.079		0.311	0.060	
S HDL FC	mmol/L	0.100	0.014		0.119	0.027		0.116	0.020		0.092	0.020		0.120	0.023	
S HDL TG	mmol/L	0.065	0.023		0.068	0.029		0.061	0.018		0.049	0.016		0.050	0.017	
XL HDL PL %	%	43.200	15.500	11	49.100	10.220	42	64.040	10.940	347	40.510	10.538		48.260	7.010	2
XL HDL C %	%	51.800	10.350	11	46.500	7.250	42	32.870	11.300	347	56.270	9.670		48.120	5.920	2
XL HDL CE %	%	41.850	10.040	11	33.500	7.180	42	17.740	11.890	347	42.550	8.630		34.390	4.950	2
XL HDL FC %	%	10.350	2.690	11	12.800	2.080	42	14.760	2.800	347	13.610	2.826		13.735	1.800	2

XL HDL TG %	%	6.265	5.651	11	4.470	5.489	42	1.924	3.084	347	3.261	2.544		3.614	2.175	2
L HDL PL %	%	51.550	5.320	41	52.900	6.700	63	55.300	7.830	221	49.330	5.650		50.530	3.160	1
L HDL C %	%	42.000	8.850	41	42.200	9.100	63	41.710	8.020	221	47.060	6.695		45.750	3.600	1
L HDL CE %	%	34.000	5.430	41	33.700	6.400	63	33.460	5.340	221	38.990	5.370		36.220	2.390	1
L HDL FC %	%	8.035	3.613	41	8.740	3.400	63	8.337	3.288	221	7.825	3.524		9.745	1.398	1
L HDL TG %	%	6.295	4.198	41	5.170	3.780	63	3.014	2.207	221	3.578	3.170		3.428	1.367	1
M HDL PL %	%	47.500	1.940	3	47.300	2.150	1	46.800	1.630	1	48.390	2.860		46.270	1.820	1
M HDL C %	%	44.800	5.000	3	46.200	4.750	1	46.940	3.470	1	44.090	4.717		48.000	3.710	1
M HDL CE %	%	37.000	3.780	3	37.000	4.100	1	38.140	2.780	1	36.700	4.270		38.600	3.180	1
M HDL FC %	%	7.760	1.959	3	9.230	1.050	1	8.840	0.927	1	7.628	1.699		9.408	0.606	1
M HDL TG %	%	7.695	4.627	3	6.750	3.440	1	6.239	2.041	1	7.309	2.594		5.715	1.708	1
S HDL PL %	%	52.900	4.880	3	51.100	5.300	1	51.510	4.290		56.220	5.400		57.100	3.780	1
S HDL C %	%	41.100	5.800	3	43.600	5.600	1	43.660	4.320		38.060	5.780		38.290	4.440	1
S HDL CE %	%	31.800	6.300	3	33.800	5.850	1	33.940	4.570		27.970	6.425		27.830	4.800	1
S HDL FC %	%	9.485	0.772	3	9.710	0.895	1	9.693	0.577		10.295	0.955		10.610	0.570	1
S HDL TG %	%	6.205	2.177	3	5.550	2.275	1	4.957	1.400		5.550	1.574		4.473	1.381	1
VLDL D	nm	38.000	2.400		37.100	2.300	1	37.360	2.120		35.900	0.860		36.320	1.610	1
LDL D	nm	23.500	0.200		23.500	0.200	1	23.550	0.220		23.800	0.230		23.660	0.110	1
HDL D	nm	9.650	0.280		9.810	0.355	1	9.651	0.309		9.792	0.213		9.968	0.238	1
Serum C	mmol/L	4.140	1.130		5.480	1.750	1	3.529	1.001		3.773	1.288		4.965	1.057	1
VLDL C	mmol/L	0.855	0.467		0.996	0.508	1	0.612	0.284		0.793	0.360		0.864	0.302	1
Remnant C	mmol/L	1.520	0.560		1.850	0.710	1	1.093	0.397		1.452	0.570		1.663	0.457	1
LDL C	mmol/L	1.550	0.700		2.200	0.985	1	1.190	0.516		1.360	0.604		1.737	0.581	1
HDL C	mmol/L	1.020	0.397		1.410	0.580	1	1.184	0.389		0.969	0.260		1.446	0.345	1
HDL2 C	mmol/L	0.541	0.355		0.897	0.556	1	0.737	0.356		0.500	0.232		0.972	0.331	1
HDL3 C	mmol/L	0.473	0.045		0.513	0.069	1	0.453	0.038		0.462	0.048		0.482	0.050	1

EstC	mmol/L	2.840	0.840		3.860	1.265	1	2.508	0.750	1	2.757	0.946	12	3.544	0.778	6
FreeC	mmol/L	1.220	0.330		1.630	0.485	1	1.013	0.289	1	1.037	0.349	12	1.372	0.315	6
Serum TG	mmol/L	1.810	1.050		1.990	1.245	1	1.393	0.831		1.021	0.512		1.407	0.735	1
VLDL TG	mmol/L	1.330	0.999		1.300	1.105	1	0.962	0.736		0.602	0.376		0.918	0.631	1
LDL TG	mmol/L	0.180	0.058		0.263	0.125	1	0.167	0.062		0.178	0.101		0.203	0.050	1
HDL TG	mmol/L	0.164	0.062		0.185	0.066	1	0.136	0.043		0.118	0.044		0.151	0.042	1
TotPG	mmol/L	1.690	0.410		2.320	0.620	1	1.661	0.426	1	1.559	0.421	13	2.070	0.425	6
TG/PG	mmol/L	0.873	0.537		0.596	0.432	1	0.676	0.373	1	0.688	0.273	13	0.686	0.411	6
PC	mmol/L	1.690	0.390		2.360	0.575	1	1.765	0.421	1	1.588	0.422	12	2.038	0.409	6
SM	mmol/L	0.386	0.114		0.539	0.157	1	0.403	0.094	1	0.475	0.065	14	0.482	0.105	6
TotCho	mmol/L	2.010	0.470		2.760	0.665	1	2.094	0.459	1	2.095	0.486	12	2.421	0.448	6
ApoA1	g/L	1.310	0.190		1.620	0.315	1	1.361	0.232		1.189	0.220		1.577	0.213	1
ApoB	g/L	0.978	0.277		1.150	0.405	1	0.762	0.231		0.821	0.268		0.962	0.222	1
ApoB/ApoA1		0.752	0.261		0.714	0.268	1	0.567	0.189		0.690	0.183		0.625	0.152	1
TotFA	mmol/L	11.400	3.300		12.900	3.950	1	9.548	2.648	1	9.898	3.126	12	12.980	3.120	6
UnSat		1.170	0.100		1.230	0.130	1	1.220	0.119	1	1.104	0.128	12	1.197	0.099	6
DHA	mmol/L	0.143	0.055		0.286	0.107	1	0.188	0.074	1	0.136	0.051	12	0.188	0.085	6
LA	mmol/L	2.800	0.790		3.550	1.080	1	2.475	0.590	1	2.775	0.818	12	3.355	0.861	6
FAw3	mmol/L	0.426	0.159		0.697	0.276	1	0.545	0.186	1	0.448	0.159	13	0.520	0.181	6
FAw6	mmol/L	3.480	0.830		4.310	1.265	1	3.100	0.724	1	3.167	0.924	19	4.173	0.989	6
PUFA	mmol/L	3.920	0.850		5.040	1.440	1	3.661	0.877	1	3.601	1.086	12	4.720	1.020	6
MUFA	mmol/L	3.090	1.440		3.360	1.440	1	2.435	1.037	1	2.364	0.968	12	3.281	1.204	6
SFA	mmol/L	4.220	1.330		4.680	1.650	1	3.462	1.070	1	3.864	1.283	18	4.888	1.291	6
DHA/FA	%	1.250	0.380		2.210	0.645	1	1.940	0.695	1	1.372	0.461	12	1.460	0.624	6
LA/FA	%	24.800	5.600		27.000	5.700	1	25.560	5.370	1	28.030	6.605	12	26.450	4.780	6
FAw3/FA	%	3.580	0.920		5.210	1.830	1	5.707	1.526	1	4.565	1.099	13	3.989	1.459	6

FAw6/FA	%	31.000	5.400		33.200	6.350	1	31.940	6.000	1	31.970	7.030	19	32.480	4.580	6
PUFA/FA	%	35.200	5.700		38.500	7.650	1	37.680	6.440	1	35.860	7.475	12	36.830	4.890	6
MUFA/FA	%	27.300	5.900		25.100	6.200	1	25.910	5.750	1	24.430	4.090	12	25.230	4.750	6
SFA/FA	%	37.700	2.600		36.100	3.700	1	36.290	3.250	1	38.610	3.850	18	37.850	2.210	6
Glc	mmol/L	6.160	3.710		6.310	3.065	1	6.226	1.879	1	6.318	3.803	1	6.169	2.930	3
Lac	mmol/L	1.700	0.640		1.830	0.770	1	1.356	0.484		1.745	0.508	1	2.070	0.751	
Pyr	mmol/L	0.106	0.041	1	0.121	0.053	1	0.063	0.043	21	0.065	0.035	47	0.099	0.049	2
Cit	mmol/L	0.112	0.021		0.121	0.035	1	0.136	0.027		0.107	0.027	4	0.129	0.024	
Glol	mmol/L	0.088	0.028	18	0.126	0.055	9	0.090	0.038	3	0.094	0.042	81	0.111	0.049	4
Ala	mmol/L	0.429	0.091		0.365	0.074	1	0.344	0.072		0.349	0.080	1	0.447	0.094	1
Gln	mmol/L	0.428	0.096	1	0.469	0.102	1	0.434	0.078		0.347	0.168	2	0.589	0.099	
Gly	mmol/L	0.252	0.049	1	0.280	0.066	1	0.249	0.053		0.281	0.057	13	0.271	0.075	11
His	mmol/L	0.058	0.014		0.063	0.014	1	0.052	0.011	1	0.074	0.020	1	0.075	0.012	3
Ile	mmol/L	0.067	0.030		0.068	0.028	1	0.071	0.020		0.064	0.019	1	0.068	0.026	
Leu	mmol/L	0.089	0.027		0.079	0.025	1	0.077	0.021		0.100	0.030	1	0.095	0.026	
Val	mmol/L	0.179	0.047		0.186	0.054	1	0.179	0.041		0.193	0.053	1	0.213	0.051	
Phe	mmol/L	0.077	0.012		0.088	0.017	1	0.080	0.014		0.097	0.021	1	0.086	0.016	
Tyr	mmol/L	0.055	0.014		0.056	0.018	1	0.053	0.016	1	0.062	0.016	1	0.060	0.014	
Ace	mmol/L	0.040	0.013		0.039	0.010	1	0.037	0.010		0.066	0.043	1	0.066	0.026	1
AcAce	mmol/L	0.057	0.031		0.048	0.036	1	0.040	0.025		0.033	0.028	1	0.048	0.045	1
bOHBut	mmol/L	0.140	0.053		0.145	0.077	1	0.121	0.074	3	0.101	0.047	6	0.126	0.095	1
Crea	mmol/L	0.067	0.017		0.060	0.018	28	0.063	0.021	1	0.056	0.014	1	0.076	0.017	2
Alb	Signal area	0.087	0.007		0.096	0.012		0.095	0.007		0.088	0.007		0.101	0.008	
Gp	mmol/L	1.420	0.370		1.720	0.565	1	1.383	0.301		1.276	0.281	1	1.540	0.316	

Blank cells indicate no missing data

BRHS: British Regional Heart Study; BWHHS: British Women's Health and Heart Study; ET2DS: Edinburgh Type 2 Diabetes Study; IQR: interquartile range; NA: missing value; SABRE: Southall and Brent Revisited Study; WHII: Whitehall-II Study

Table C-2 Data dictionary for full metabolite names

Metabolite abbreviation	Full metabolite description
XXL_VLDL_P	Concentration of chylomicrons and extremely large VLDL particles
XXL_VLDL_L	Total lipids in chylomicrons and extremely large VLDL
XXL_VLDL_PL	Phospholipids in chylomicrons and extremely large VLDL
XXL_VLDL_C	Total cholesterol in chylomicrons and extremely large VLDL
XXL_VLDL_CE	Cholesterol esters in chylomicrons and extremely large VLDL
XXL_VLDL_FC	Free cholesterol in chylomicrons and extremely large VLDL
XXL_VLDL_TG	Triglycerides in chylomicrons and extremely large VLDL
XL_VLDL_P	Concentration of very large VLDL particles
XL_VLDL_L	Total lipids in very large VLDL
XL_VLDL_PL	Phospholipids in very large VLDL
XL_VLDL_C	Total cholesterol in very large VLDL
XL_VLDL_CE	Cholesterol esters in very large VLDL
XL_VLDL_FC	Free cholesterol in very large VLDL
XL_VLDL_TG	Triglycerides in very large VLDL
L_VLDL_P	Concentration of large VLDL particles
L_VLDL_L	Total lipids in large VLDL
L_VLDL_PL	Phospholipids in large VLDL
L_VLDL_C	Total cholesterol in large VLDL
L_VLDL_CE	Cholesterol esters in large VLDL
L_VLDL_FC	Free cholesterol in large VLDL
L_VLDL_TG	Triglycerides in large VLDL
M_VLDL_P	Concentration of medium VLDL particles
M_VLDL_L	Total lipids in medium VLDL
M_VLDL_PL	Phospholipids in medium VLDL
M_VLDL_C	Total cholesterol in medium VLDL
M_VLDL_CE	Cholesterol esters in medium VLDL
M_VLDL_FC	Free cholesterol in medium VLDL
M_VLDL_TG	Triglycerides in medium VLDL
S_VLDL_P	Concentration of small VLDL particles
S_VLDL_L	Total lipids in small VLDL
S_VLDL_PL	Phospholipids in small VLDL
S_VLDL_C	Total cholesterol in small VLDL
S_VLDL_CE	Cholesterol esters in small VLDL
S_VLDL_FC	Free cholesterol in small VLDL
S_VLDL_TG	Triglycerides in small VLDL
XS_VLDL_P	Concentration of very small VLDL particles
XS_VLDL_L	Total lipids in very small VLDL
XS_VLDL_PL	Phospholipids in very small VLDL

XS_VLDL_C	Total cholesterol in very small VLDL
XS_VLDL_CE	Cholesterol esters in very small VLDL
XS_VLDL_FC	Free cholesterol in very small VLDL
XS_VLDL_TG	Triglycerides in very small VLDL
XXL_VLDL_PL_.	Phospholipids to total lipids ratio in chylomicrons and extremely large VLDL
XXL_VLDL_C_.	Total cholesterol to total lipids ratio in chylomicrons and extremely large VLDL
XXL_VLDL_CE_.	Cholesterol esters to total lipids ratio in chylomicrons and extremely large VLDL
XXL_VLDL_FC_.	Free cholesterol to total lipids ratio in chylomicrons and extremely large VLDL
XXL_VLDL_TG_.	Triglycerides to total lipids ratio in chylomicrons and extremely large VLDL
XL_VLDL_PL_.	Phospholipids to total lipids ratio in very large VLDL
XL_VLDL_C_.	Total cholesterol to total lipids ratio in very large VLDL
XL_VLDL_CE_.	Cholesterol esters to total lipids ratio in very large VLDL
XL_VLDL_FC_.	Free cholesterol to total lipids ratio in very large VLDL
XL_VLDL_TG_.	Triglycerides to total lipids ratio in very large VLDL
L_VLDL_PL_.	Phospholipids to total lipids ratio in large VLDL
L_VLDL_C_.	Total cholesterol to total lipids ratio in large VLDL
L_VLDL_CE_.	Cholesterol esters to total lipids ratio in large VLDL
L_VLDL_FC_.	Free cholesterol to total lipids ratio in large VLDL
L_VLDL_TG_.	Triglycerides to total lipids ratio in large VLDL
M_VLDL_PL_.	Phospholipids to total lipids ratio in medium VLDL
M_VLDL_C_.	Total cholesterol to total lipids ratio in medium VLDL
M_VLDL_CE_.	Cholesterol esters to total lipids ratio in medium VLDL
M_VLDL_FC_.	Free cholesterol to total lipids ratio in medium VLDL
M_VLDL_TG_.	Triglycerides to total lipids ratio in medium VLDL
S_VLDL_PL_.	Phospholipids to total lipids ratio in small VLDL
S_VLDL_C_.	Total cholesterol to total lipids ratio in small VLDL
S_VLDL_CE_.	Cholesterol esters to total lipids ratio in small VLDL
S_VLDL_FC_.	Free cholesterol to total lipids ratio in small VLDL
S_VLDL_TG_.	Triglycerides to total lipids ratio in small VLDL
XS_VLDL_PL_.	Phospholipids to total lipids ratio in very small VLDL
XS_VLDL_C_.	Total cholesterol to total lipids ratio in very small VLDL
XS_VLDL_CE_.	Cholesterol esters to total lipids ratio in very small VLDL
XS_VLDL_FC_.	Free cholesterol to total lipids ratio in very small VLDL
XS_VLDL_TG_.	Triglycerides to total lipids ratio in very small VLDL
IDL_P	Concentration of IDL particles
IDL_L	Total lipids in IDL
IDL_PL	Phospholipids in IDL
IDL_C	Total cholesterol in IDL
IDL_CE	Cholesterol esters in IDL
IDL_FC	Free cholesterol in IDL
IDL_TG	Triglycerides in IDL

IDL_PL_.	Phospholipids to total lipids ratio in IDL
IDL_C_.	Total cholesterol to total lipids ratio in IDL
IDL_CE_.	Cholesterol esters to total lipids ratio in IDL
IDL_FC_.	Free cholesterol to total lipids ratio in IDL
IDL_TG_.	Triglycerides to total lipids ratio in IDL
L_LDL_P	Concentration of large LDL particles
L_LDL_L	Total lipids in large LDL
L_LDL_PL	Phospholipids in large LDL
L_LDL_C	Total cholesterol in large LDL
L_LDL_CE	Cholesterol esters in large LDL
L_LDL_FC	Free cholesterol in large LDL
L_LDL_TG	Triglycerides in large LDL
M_LDL_P	Concentration of medium LDL particles
M_LDL_L	Total lipids in medium LDL
M_LDL_PL	Phospholipids in medium LDL
M_LDL_C	Total cholesterol in medium LDL
M_LDL_CE	Cholesterol esters in medium LDL
M_LDL_FC	Free cholesterol in medium LDL
M_LDL_TG	Triglycerides in medium LDL
S_LDL_P	Concentration of small LDL particles
S_LDL_L	Total lipids in small LDL
S_LDL_PL	Phospholipids in small LDL
S_LDL_C	Total cholesterol in small LDL
S_LDL_CE	Cholesterol esters in small LDL
S_LDL_FC	Free cholesterol in small LDL
S_LDL_TG	Triglycerides in small LDL
L_LDL_PL_.	Phospholipids to total lipids ratio in large LDL
L_LDL_C_.	Total cholesterol to total lipids ratio in large LDL
L_LDL_CE_.	Cholesterol esters to total lipids ratio in large LDL
L_LDL_FC_.	Free cholesterol to total lipids ratio in large LDL
L_LDL_TG_.	Triglycerides to total lipids ratio in large LDL
M_LDL_PL_.	Phospholipids to total lipids ratio in medium LDL
M_LDL_C_.	Total cholesterol to total lipids ratio in medium LDL
M_LDL_CE_.	Cholesterol esters to total lipids ratio in medium LDL
M_LDL_FC_.	Free cholesterol to total lipids ratio in medium LDL
M_LDL_TG_.	Triglycerides to total lipids ratio in medium LDL
S_LDL_PL_.	Phospholipids to total lipids ratio in small LDL
S_LDL_C_.	Total cholesterol to total lipids ratio in small LDL
S_LDL_CE_.	Cholesterol esters to total lipids ratio in small LDL
S_LDL_FC_.	Free cholesterol to total lipids ratio in small LDL
S_LDL_TG_.	Triglycerides to total lipids ratio in small LDL

XL_HDL_P	Concentration of very large HDL particles
XL_HDL_L	Total lipids in very large HDL
XL_HDL_PL	Phospholipids in very large HDL
XL_HDL_C	Total cholesterol in very large HDL
XL_HDL_CE	Cholesterol esters in very large HDL
XL_HDL_FC	Free cholesterol in very large HDL
XL_HDL_TG	Triglycerides in very large HDL
L_HDL_P	Concentration of large HDL particles
L_HDL_L	Total lipids in large HDL
L_HDL_PL	Phospholipids in large HDL
L_HDL_C	Total cholesterol in large HDL
L_HDL_CE	Cholesterol esters in large HDL
L_HDL_FC	Free cholesterol in large HDL
L_HDL_TG	Triglycerides in large HDL
M_HDL_P	Concentration of medium HDL particles
M_HDL_L	Total lipids in medium HDL
M_HDL_PL	Phospholipids in medium HDL
M_HDL_C	Total cholesterol in medium HDL
M_HDL_CE	Cholesterol esters in medium HDL
M_HDL_FC	Free cholesterol in medium HDL
M_HDL_TG	Triglycerides in medium HDL
S_HDL_P	Concentration of small HDL particles
S_HDL_L	Total lipids in small HDL
S_HDL_PL	Phospholipids in small HDL
S_HDL_C	Total cholesterol in small HDL
S_HDL_CE	Cholesterol esters in small HDL
S_HDL_FC	Free cholesterol in small HDL
S_HDL_TG	Triglycerides in small HDL
XL_HDL_PL_.	Phospholipids to total lipids ratio in very large HDL
XL_HDL_C_.	Total cholesterol to total lipids ratio in very large HDL
XL_HDL_CE_.	Cholesterol esters to total lipids ratio in very large HDL
XL_HDL_FC_.	Free cholesterol to total lipids ratio in very large HDL
XL_HDL_TG_.	Triglycerides to total lipids ratio in very large HDL
L_HDL_PL_.	Phospholipids to total lipids ratio in large HDL
L_HDL_C_.	Total cholesterol to total lipids ratio in large HDL
L_HDL_CE_.	Cholesterol esters to total lipids ratio in large HDL
L_HDL_FC_.	Free cholesterol to total lipids ratio in large HDL
L_HDL_TG_.	Triglycerides to total lipids ratio in large HDL
M_HDL_PL_.	Phospholipids to total lipids ratio in medium HDL
M_HDL_C_.	Total cholesterol to total lipids ratio in medium HDL
M_HDL_CE_.	Cholesterol esters to total lipids ratio in medium HDL

M_HDL_FC_.	Free cholesterol to total lipids ratio in medium HDL
M_HDL_TG_.	Triglycerides to total lipids ratio in medium HDL
S_HDL_PL_.	Phospholipids to total lipids ratio in small HDL
S_HDL_C_.	Total cholesterol to total lipids ratio in small HDL
S_HDL_CE_.	Cholesterol esters to total lipids ratio in small HDL
S_HDL_FC_.	Free cholesterol to total lipids ratio in small HDL
S_HDL_TG_.	Triglycerides to total lipids ratio in small HDL
VLDL_D	Mean diameter for VLDL particles
LDL_D	Mean diameter for LDL particles
HDL_D	Mean diameter for HDL particles
Serum_C	Serum total cholesterol
VLDL_C	Total cholesterol in VLDL
Remnant_C	Remnant cholesterol (non-HDL, non-LDL -cholesterol)
LDL_C	Total cholesterol in LDL
HDL_C	Total cholesterol in HDL
HDL2_C	Total cholesterol in HDL2
HDL3_C	Total cholesterol in HDL3
EstC	Esterified cholesterol
FreeC	Free cholesterol
Serum_TG	Serum total triglycerides
VLDL_TG	Triglycerides in VLDL
LDL_TG	Triglycerides in LDL
HDL_TG	Triglycerides in HDL
TotPG	Total phosphoglycerides
TG/PG	Ratio of triglycerides to phosphoglycerides
PC	Phosphatidylcholine and other cholines
SM	Sphingomyelins
TotCho	Total cholines
ApoA1	Apolipoprotein A-I
ApoB	Apolipoprotein B
ApoB/ApoA1	Ratio of apolipoprotein B to apolipoprotein A-I
TotFA	Total fatty acids
UnSat	Estimated degree of unsaturation
DHA	22:6, docosahexaenoic acid
LA	18:2, linoleic acid
FAw3	Omega-3 fatty acids
FAw6	Omega-6 fatty acids
PUFA	Polyunsaturated fatty acids
MUFA	Monounsaturated fatty acids; 16:1, 18:1
SFA	Saturated fatty acids
DHA/FA	Ratio of 22:6 docosahexaenoic acid to total fatty acids

LA/FA	Ratio of 18:2 linoleic acid to total fatty acids
FAw3/FA	Ratio of omega-3 fatty acids to total fatty acids
FAw6/FA	Ratio of omega-6 fatty acids to total fatty acids
PUFA/FA	Ratio of polyunsaturated fatty acids to total fatty acids
MUFA/FA	Ratio of monounsaturated fatty acids to total fatty acids
SFA/FA	Ratio of saturated fatty acids to total fatty acids
Glc	Glucose
Lac	Lactate
Pyr	Pyruvate
Cit	Citrate
Glol	Glycerol
Ala	Alanine
Gln	Glutamine
Gly	Glycine
His	Histidine
Ile	Isoleucine
Leu	Leucine
Val	Valine
Phe	Phenylalanine
Tyr	Tyrosine
Ace	Acetate
AcAce	Acetoacetate
bOHBut	3-hydroxybutyrate
Crea	Creatinine
Alb	Albumin
Gp	Glycoprotein acetyls, mainly a1-acid glycoprotein