Differential networks (and other statistical issues) for the analysis of metabolomic data

David Macleod

Thesis submitted in accordance with the requirements for the degree of Doctor of Philosophy of the University of London
September 2016

Department of Medical Statistics
Faculty of Epidemiology and Population Health
London School of Hygiene and Tropical Medicine
Funded by the ESRC

Declaration

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

Abstract

Coronary heart disease (CHD) is the leading cause of death in the UK. Recent technological advances in metabolomics have the potential to contribute to further the understanding of CHD, especially because they are facilitating the collection of metabolomics data in large observational studies. However, the high dimensionality of this type of information and its strong interdependencies raise several analytical difficulties.

These difficulties were investigated, motivated by the study of 228 metabolites acquired from blood samples as part of the British Womens Heart and Health Study (BWHHS). Issues regarding transformations of the metabolomics data and their reliability were examined. Analytical methods typically adopted with high-dimensional data were reviewed, and then a more recently developed method, differential networks, was examined in detail.

When investigating differential networks using simulations of three alternative data generating scenarios, it was found that an edge between two nodes can be induced if the effect of one node on disease is modified by another node, or if the disease causes (or is associated with) a "breaking down" in the relationship between the two nodes. The simulations focused on simplified settings but exemplify the difficulties in interpreting differential networks and helped elucidate the sample sizes required.

Further algebraic examination of likely data generating mechanisms identified the potential pitfalls of relying on partial correlations in building differential networks. This shows that, when important nodes influencing the correlation structure are not measured, irrelevant edges may be selected, while relevant ones may be missed.

Analysis of the BWHHS metabolite data flagged a small number of metabolites that could potentially be associated with CHD, with small VLDL triglycerides being the strongest candidate. Comparisons were made with the results obtained using regression-based methods as these are more easily accessible to epidemiologists. The fact that there was little overlap in identified biomarkers is an indication of the complexity of this field of research.

Acknowledgements

I owe an huge debt of gratitude to my supervisor Bianca de Stavola, who supported me enourmously during my research, her generosity knows no bounds and I would not have finished without her. Thanks also to Anna Gottard, David Prieto and JP Casas who provided much needed data, help and feedback throughout. And thanks for the love and support provided by my partner Jenny.

Contents

1 Introduction **1**

2 British Women's Heart and Health Study **3**

2.1 Introduction 3

2.2 Participant Characteristics 4

2.2.1 Inclusion Criteria 5

2.2.2 Representativeness 6

2.2.3 Missing Data 7

2.2.4 Outcomes 7

3 NMR biomarkers **9**

3.1 Introduction 9

3.1.1 NMR biomarkers in the BWHHS 10

3.2 Data description 14

3.2.1 Missing Data 14

3.2.2 Biomarker Distributions 16

3.2.3 Zero values 16

3.2.4 NMR biomarker correlation structure 17

3.3 Analytical considerations 19

3.3.1 Transformations 19

3.4 Repeated Measures, Reliability and Agreement 21

3.4.1 Methods 22

3.4.2 Results 23

3.5 Summary 34

4 Methods for high-dimensional exposures **48**

4.1 Introduction 48

4.2 Overview of methods 48

4.2.1 Standard methods and multiple testing 48

4.2.2 Multiple testing adjustment 50

4.2.3 Other methods 51

4.3 Methods for BWHHS 54

4.3.1 Logistic regression 54

4.3.2 Cox regression 56

4.4 Data preparation 57
4.5 Results 59
4.5.1 Simple analyses, with adjustment for multiple testing . . 59
4.5.2 Principal Components Regression 64
4.5.3 Lasso Regression 67
4.6 Summary 70

5 Networks 74
5.1 Introduction 74
5.2 Network Basics 77
5.2.1 Nodes and Edges 77
5.2.2 Connectivity 78
5.2.3 Adjacency Matrix 79
5.2.4 Paths 80
5.2.5 Node centrality 80
5.2.6 Community Structure 84
5.3 Network analysis of BWHHS data 86
5.3.1 Methods 86
5.3.2 Results 90

6 Differential networks 94
6.1 Background 94
6.2 Correlation based differential networks 100
6.2.1 Edge selection and interpretation 101
6.3 Simulations 103
6.3.1 Scenarios 103
6.3.2 Methods 105
6.4 Results 111
6.4.1 Data generating model A 112
6.4.2 Data generating model B 127
6.4.3 Data generating model C 129
6.5 Summary and discussion 131

7 Overview and guidance for analysis of differential networks 135
7.1 Marginal or partial correlations? 135
7.1.1 Differential network based on marginal correlations 137
7.1.2 Differential network based on partial correlations 138
7.1.3 General results 145
7.1.4 Marginal or partial correlations in differential networks-which to use? 156
7.2 Differential centrality 156
7.3 Analysis of the BWHHS data 157
7.3.1 Data preparation 157
7.3.2 Differential network based on marginal correlations 157
7.3.3 Differential network based on partial correlations 165
7.4 Discussion 170

7.4.1 Differential network using marginal correlations 170
7.4.2 Differential network using partial correlations 172
7.4.3 Limitations and alternative strategies 172

8 Differential networks and time to event analysis 174
8.1 Methods 176
8.2 Results 176
8.2.1 Differential networks 176
8.2.2 Lasso regression 181
8.3 Discussion 182

9 Discussion 184
9.1 Metabolomic data 184
9.2 Differential Networks 187
9.2.1 Simulation study 188
9.2.2 Marginal vs. partial correlations 193
9.2.3 Differential networks using the BWHHS data 193
9.2.4 Results from the time to event analysis 199
9.3 Concluding comments 199

Chapter 1

Introduction

The amount of data accessible to medical researchers is increasing exponentially. Data collection techniques are advancing, and computer processing power is accelerating allowing more and more data to be acquired in a more cost effective manner. Once these data are collected there still remains the task of making sense of them. With such rich data available it is necessary to explore new methods, to complement more traditional methods of data analysis, in order to exploit this richness.

A dataset has been acquired from a cohort study of heart disease (the British Women's Heart and Health Study), where previously stored blood samples were analysed using ^{1}H-NMR spectroscopy providing over 200 new biomarkers which can be investigated to identify any association with the disease. The high-dimensionality of this dataset presents some difficulties in analysis, which provides the motivation for this thesis - can we exploit these data to gain a greater understanding of the cohort and its relationship with coronary heart disease? This leads to the two main aims of this thesis:

1. Describe the new metabolomic data and investigate its reliability
2. Investigate the use of a recently developed analysis method, differential networks, critically reviewing the suitability of the method and applying it to the cohort data available.

In chapter 2 the cohort study that provides the motivation for this research will be introduced and the characteristics of the cohort described. This is followed by a chapter describing in depth the metabolomic data that has been recently made available. It includes a description the data structure and any data issues arising, possible transformations of the data that could aid analyses, an investigation into how closely the metabolites measured using both ^{1}H-NMR spectroscopy and standard methods agree and an investigation into the short-term reliability of the metabolites measured using ^{1}H-NMR spectroscopy, addressing aim 1. Chapter 4 then provides an overview of a few typical methods currently used to analyse high-dimensional data, with each applied to the cohort data and interpreted.

The subsequent chapters are all dedicated to addressing aim 2, chapter 5 introduces the concept of network analysis and describes some of the basic terms and network statistics involved. Chapter 6 introduces the concept of differential network analysis and contains a short literature review on the topic, followed by a detailed simulation study investigating the potential interpretations of a differential network. Chapter 7 provides a more practical approach to differential network analysis, highlighting some of the difficulties involved with the process and describing its limitations before going on to apply the method to the cohort data described in chapter 3. Differential network analysis has not previously been attempted using time to event data, so chapter 8 expands the analysis from the previous chapter to investigate this possibility. The final chapter aims to provide an overview of the research described in this thesis and a discussion of the findings.

Chapter 2

British Women's Heart and Health Study

The motivation for this thesis is the investigation and development of methods to be used in the analysis of high-dimensional metabolomic data, which is becoming more widely available in epidemiological studies. The study that will be focussed on is the British Women's Heart and Health Study (BWHHS) which will be used as an applied example of the methods discussed, with the outcome of interest being CHD incidence. This chapter provides an overview of this study.

2.1 Introduction

The BWHHS is a prospective cohort study of heart disease, funded by the Department of Health and the British Heart Foundation [1]. 4286 British women between the ages of 60 and 79 were recruited from 23 British towns between 1999 and 2001 and the cohort subsequently followed up at regular intervals. The data currently available covers a 12 year period of follow up. The data from this cohort has been used in a wide range of research such as investigating associations between socioeconomic position and cardiovascular disease (CVD) [2, 3], geographical variations in CVD [4, 5], determinants of stroke and coronary heart disease (CHD) [6–8], physical activity in older women [9, 10] and also the investigation of novel risk factors for CHD [11, 12].

The study team visited each town involved for a two week period between 1999 and 2001 to collect baseline data. All women completed a questionnaire covering lifestyle factors and medical history, followed by an interview with a nurse, who asked in more detail about specific heart problems or symptoms the participant may have had as well as gathering information relating to any medication taken. A deprivation score (Carstairs Index) was derived for each participant based on their area of residence at the time of data collection. Next a medical exam with a second nurse took place where measurements of height, weight, blood pressure and lung function were made. A resting electrocardiogram was performed and a fasting blood sample was taken from consenting participants. Finally, all women interviewed during the first week of the two week period were asked if they would be willing to return the following week to have a second medical exam and provide another blood sample, for quality control purposes (neither the questionnaire nor the interview were required on the second visit).

A third nurse had the sole responsibility of processing the blood samples obtained. These samples were allowed to sit for 30 minutes prior to centrifuging at 3500rpm for 10 minutes to separate, and were then aliquotted into tubes. The aliquots were then snap frozen on dry ice and placed in a freezer at -20°C before being transferred for long term storage at -80°C.

From the collected blood samples, 36 biomarker measurements were quantified initially, and then in 2013 a further 228 biomarker measurements (specifically metabolomic data, discussed in chapter 3) were made available after all the serum samples were analysed using ^{1}H-Nuclear Magnetic Resonance (NMR) Spectroscopy by the computational medicine group based in Oulu, Finland. Throughout this thesis I will refer to the 36 original biomarkers as the *standard biomarkers* and the 228 biomarkers obtained via ^{1}H-NMR as the *NMR biomarkers* or *metabolites*.

Since baseline, the cohort has been followed up approximately every 3 years (2003, 2007, 2010 & 2013), where a self-administered questionnaire is completed by the individual along with reviews of GP records for non-fatal events and Office for National Statistics records for identifying deaths [13]. Although some individuals have not completed the repeat questionnaires, the CHD status has been ascertained for every individual in the study up to 12 years (with the linkage to status records being performed in 2013).

2.2 Participant Characteristics

There were a total of 4286 women recruited for the study, and although the aim was to recruit women from ages 60-79, the age of the women included in the study actually ranged from 59 to 80 at entry, with a mean age at entry of 68.9 years (SD 5.5) and a mean body mass index (BMI) of 27.6 kg/m^{2} (SD 5.0). Women were recruited from 23 different towns across the UK, with the number included from each town ranging from 140 (Bristol, Hartlepool) to 204 (Exeter). Mean systolic blood pressure was found to be 147.1 mmHg (SD 25.2) and mean diastolic blood pressure was 79.4 mmHg (SD 11.7). Smoking status was also measured for particpants with 2371 (55.6%) classed as never smokers, 1401 (32.8%) ex-smokers and 495 (11.6%) as current smokers (19 individuals were missing smoking status). There were 201 (4.7%) participants who had evidence of prevalent diabetes.

A CHD event is defined as an individual having had a confirmed myocardial infarction (fatal or non-fatal), or cardiac revascularization surgery (coronary artery bypass graft (CABG) or percutaneous transluminal coronary angioplasty (PTCA)). A participant is said to have prevalent CHD if their first CHD event occurred prior to the date of entry into the study, and incident CHD if an event occurred after entry into the study. A total of 159 (3.7%) study participants had evidence of prevalent CHD. The complete list of baseline characteristics is shown in table 2.1.

2.2.1 Inclusion Criteria

As the motivation for this thesis is to investigate methods to analyse metabolomic data in relation to CHD incidence, the analyses performed will only include women from whom a NMR metabolite profile has been obtained successfully.

A total of 4286 women were recruited in the BWHHS. No medical exam was performed on 291 of the participants, and 72 of those examined did not give consent to providing a blood sample, leaving 3923 women who provided a baseline blood sample. Of these, a further 146 were excluded from the dataset to be used for analysis, 136 because there was an insufficient set of blood samples to have a sample measured for the NMR biomarkers and 10 because of technical problems with the NMR assays, leaving 3777 women with a valid NMR metabolite profile.

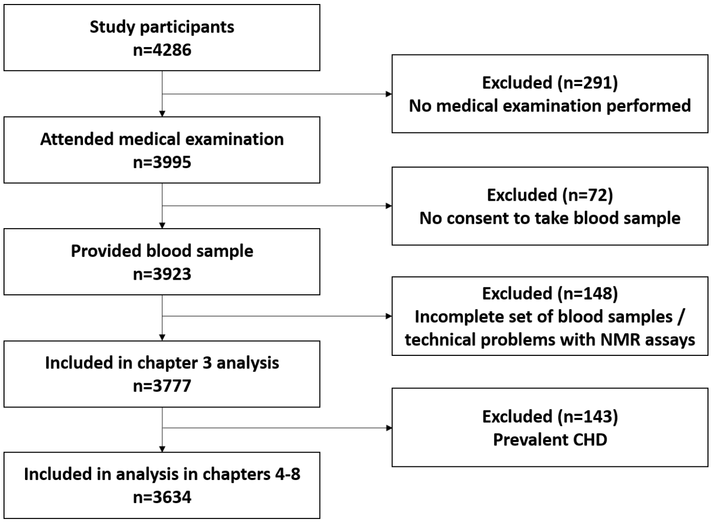

Study participants
n=4286
Excluded (n=291)
No medical examination performed
Attended medical examination
n=3995
Excluded (n=72)
No consent to take blood sample
Provided blood sample
n=3923
Excluded (n=148)
Incomplete set of blood samples /
technical problems with NMR assays
Included in chapter 3 analysis
n=3777
Excluded (n=143)
Prevalent CHD
Included in analysis in chapters 4-8
n=3634

Figure 2.1: Inclusion criteria

These 3777 women will all be used in the analysis in chapter 3, where the

NMR biomarkers are looked at in depth. However, in chapters 4, 7 and 8 CHD incidence will also be investigated, with respect to the NMR metabolite profiles of the participants. In these analyses, all individuals who have *prevalent* CHD at baseline will be excluded also. Out of the 3777 included women, 143 have prevalent CHD so will therefore be excluded from the analyses in chapters 4, 7 and 8.

2.2.2 Representativeness

To check that there was not a systematic difference between the women included in the analysis and the women excluded because of missing NMR metabolite profiles, the characteristics have been stratified by inclusion status in table 2.1. It is worth noting that one of the exclusion criteria was that no medical exam was performed, so the variables for BMI, waist-hip ratio and blood pressure are missing for 58% of the excluded participants, which could explain some of the differences identified. There were also large differences between the proportion excluded in each town, with three towns having over 30% of their participants excluded (Gloucester, Merthyr Tydfil and Mansfield) and 1 town (Ayr) having no excluded participants.

Table 2.1: Baseline patient characteristics for all participants, and stratified by whether they will be excluded from analysis due to a missing NMR metabolite profile. (BMI: Body Mass Index; SBP: Systolic Blood Pressure; DBP: Diastolic Blood Pressure; Deprivation: Percentage in top quintile of Carstairs index of deprivation)

	All (N=4286) Mean(SD) or %	Included (N=3777) Mean(SD) or %	Excluded (N=509) Mean(SD) or %
Age(years)	68.9 (5.5)	68.8 (5.5)	69.3 (5.7)
BMI (kg/m^2)	27.6 (5.0)	27.5 (4.9)	28.6 (5.3)
Waist-Hip Ratio	0.82 (0.07)	0.82 (0.07)	0.82 (0.07)
SBP(mmHg)	147.1 (25.2)	147.0 (25.2)	149.0 (25.3)
DBP (mmHg)	79.4 (11.7)	79.4 (11.7)	79.0 (12.2)
Deprivation	23.6%	22.6%	30.8%
Ever Smoker	44.4%	43.8%	49.0%
Prevalent Diabetes	4.7%	4.5%	6.1%
Prevalent CHD	3.7%	3.8%	3.1%

After performing a logistic regression using whether a participant was included or not as the outcome variable, there was some evidence to suggest women who were excluded from the study were more likely to have a higher BMI, be more deprived and more likely to be have ever been a smoker (although this evidence should be viewed with caution, given the large numbers of missing data for some of these variables). The effect of deprivation is strongly tied in with the differences found between towns, as deprivation is based on the participants' postcode, after adjusting for town the evidence of an association between deprivation and exclusion disappears. There appears to be no association between inclusion and prevalent CHD or diabetes.

2.2.3 Missing Data

The baseline data for the 3777 women to be included in the chapter 3 analyses are fairly complete. Table 2.2 shows the number of missing values for each of the baseline characteristics.

Table 2.2: Number of missing values for each of the baseline characteristics from the BWHHS (BMI: Body Mass Index; SBP: Systolic Blood Pressure; DBP: Diastolic Blood Pressure)

	Number of Missing values
Age(years)	0
BMI (kg/m^2)	33
Waist-Hip Ratio	43
SBP(mmHg)	24
DBP (mmHg)	24
Deprivation	15
Smoking Status	2
Prevalent Diabetes	0
Prevalent CHD	0

2.2.4 Outcomes

Of the 3777 women with a valid NMR metabolite profile, 143 had prevalent CHD at baseline (3.8%). Of the 3634 women who did not have CHD at baseline, 182 (5.0%) went on to have a CHD event in the 12 year follow up period while 705 (19.4%) died within the follow up period without experiencing a CHD event. The median time to CHD event was 5.3 years (IQR 2.7,8.2) and median time to death was 7.3 years (IQR 4.5,9.6). A curve plotting the cumulative incidence of both CHD and death is illustrated in figure 2.2.

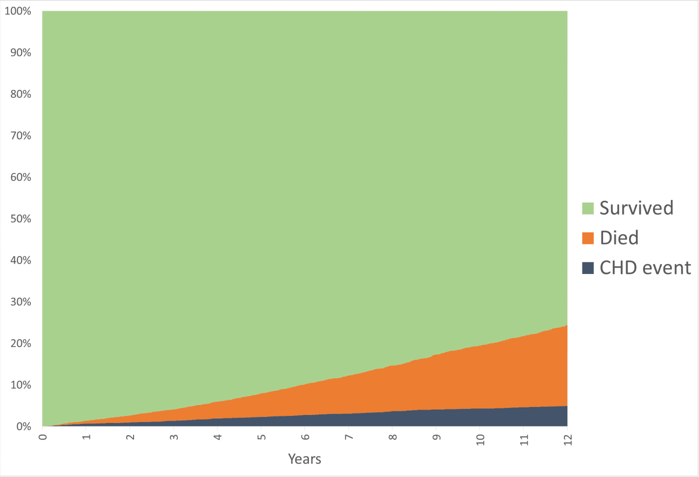


Figure 2.2: Cumulative incidence of non-CHD deaths (orange) and fatal/non-fatal CHD events (navy blue) among the 3634 women without prevalent CHD at baseline.

The focus of the next chapter will be the NMR metabolites measured on blood taken at baseline, providing a brief description of what they are followed by a detailed description of their observed distributions, before finally investigating their internal reliability and, where the same metabolite had been quantified previously using more traditional means, the agreement between the ^{1}H-NMR metabolite and the original measurement.

Chapter 3

NMR biomarkers

The aim of this chapter is to introduce the topic of metabolomics and to provide an in depth description of metabolomic data from the BWHHS, which includes an assessment of the short term reliability of the biomarkers measured as well as the agreement between biomarkers measured using two different techniques.

3.1 Introduction

Aetiological research in epidemiology is being increasingly complemented by metabolomics [14]. Metabolomics is the study of the metabolome, which is the collective name for the small molecules found in cells, tissues and biofluids [15]. These small molecules are referred to as metabolites. Measuring the metabolome of an individual is of interest because it is closely related to environmental and behavioural factors. i.e. the concentrations of certain metabolites will change depending on what an individual eats or drinks or if they are exposed to drugs. So measurement of the metabolome may help elucidate the causal pathways from behaviour to disease [16]. This could also be considered a downside, as it may be difficult to distinguish between transient effects (i.e. what the individual has just eaten) from more persistant exposures. It is also associated with the genome so its knowledge can help uncover the pathways from genotype to disease phenotype [17], and since metabolites are the end product of many cellular processes in the body they are considered to have a closer relationship to changes in the body due to disease than genomics, transcriptomics and proteomics [18]. Metabolomics is also generally cheaper per sample than proteomics and transcriptomics [16].

There are also two broad categories of metabolomics, one is 'discovery based' metabolomics, where the aim is to discover new metabolites that are potentially markers or risk factors for a disease phenotype [19] with the other category being the quantification of the concentrations of a specific set of known metabolites in a sample of tissue or fluid [20].

In recent years, technological developments have made it possible to measure greater numbers of metabolites in large epidemiological cohorts in a more cost effective and timely manner [21, 22]. The two main technologies used to do this are Mass Spectrometry [23] and Proton Nuclear Magnetic Resonance (^{1}H-NMR) Spectroscopy [22, 24]. The analysis performed on the BWHHS data used ^{1}H-NMR Spectroscopy on the participants' blood samples to quantify the concentration of a set of known metabolites. (Throughout this thesis often "NMR" is used as a shorthand for "^{1}H-NMR")

The protocol used to obtain the metabolite concentrations is described by Ala-Korpela et al and Soinenen et al. [20, 22] but very briefly the aim of ^{1}H-NMR spectroscopy is to determine the structure of tissue or fluid using a magnetic field. Subatomic particles have a characteristic spin [25] and by applying an external magnetic field, some particles, such as Hydrogen-1 (^{1}H), respond in a predictable manner. Each compound containing ^{1}H has its own unique response, known as its chemical shift [26]. A sample (in the BWHHS, a sample of plasma) is placed in the NMR spectrometer and a range of magnetic frequencies is applied, resulting in a spectrum of responses, which are then converted to metabolite concentrations. ^{1}H NMR spectroscopy is a non-destructive method [27], meaning that the sample can be preserved and potentially reused after analysis.

3.1.1 NMR biomarkers in the BWHHS

The values for 149 biomarkers were quantified directly using ^{1}H-NMR spectroscopy. Of these, 145 were the concentrations of metabolites found in the serum samples, 3 were measurements of the mean diameter of lipoprotein particles in the serum samples (measured in nm) and 1 was the estimated degree of unsaturation in total fatty acids. In addition to this, a further 79 biomarkers were provided, which were ratios derived from the directly quantified concentrations of these metabolites. The metabolite concentrations were provided in January 2015.

Table 3.5 illustrates the complete list of 228 biomarkers, including their units, their untransformed mean, their range of values, how many observations are missing and shorthand names for use in tables and figures throughout the thesis. The metabolites in this study can be categorised into 3 main groups, with the metabolites from each group arising from 1 of 3 molecular windows [22].

1. The LIPO window - Albumin, liporotein subclasses and derived measures (115 metabolites, 71 derived ratios)

2. The LMWM window - Amino acids and other low molecular weight metabolites (19 metabolites)
3. The LIPID window - Serum lipid extracts (15 metabolites, 8 derived ratios)

3.1.1.1 LIPO window

The LIPO window contains the most metabolites (115) and of these, 98 are measurements from lipoproteins that we class by their size and density. There is a naming convention for these lipoprotein subclasses that is used throughout this thesis. There are 4 different lipoproteins, named according to their density, these are (density lowest to highest) :

(i) Very Low Density Lipoprotein (VLDL)

(ii) Intermediate Density Lipoprotein (IDL)

(iii) Low Density Lipoprotein (LDL)

(iv) High Density Lipoprotein (HDL)

VLDLs are the largest particles, followed by IDL, LDL and finally the smallest set of particles among these are the HDLs. In addition to this, within each of these 4 lipoproteins there are sub-types, classified according to their size (XXL down to XS). So for example L–VLDL refers to the class of large, very low density lipoproteins. S–HDL refers to small, high density lipoproteins. These size classifications are not absolute, they are relevant within each lipoprotein, so S–VLDL would be expected to have larger particles than L–HDL.

A VLDL lipoprotein particle is illustrated in figure 3.1, to show how the different elements make up the overall structure. The core of the particle is made up from triglycerides and cholesterol esters, and is surrounded by a surface layer of free cholesterol, phospholipids and apolipoproteins.

For each lipoprotein subclass there are 7 measurements provided in the data, each referred to by a unique suffix. These suffixes are:

(i) –P : Particle concentration

(ii) –L : Total lipids

(iii) –PL : Phospholipids

(iv) –TG : Triglycerides

(v) –C : Total cholesterol

(vi) –FC : Free cholesterol

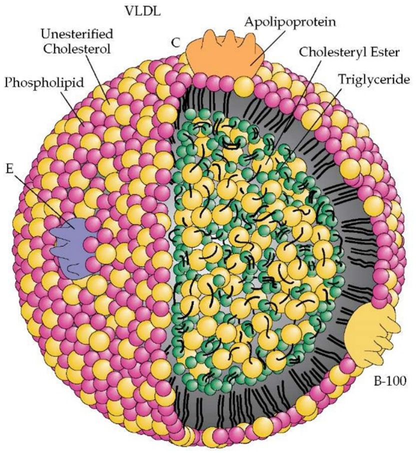


Figure 3.1: Structure of a VLDL particle [28]

(vii) –CE : Cholesterol esters

All measurements are made independently, however some of the metabolites are components within other metabolites. For example, the total lipids concentration within a particular lipoprotein subclass is made from adding together the concentrations of total cholesterol, phospholipids and triglycerides in that subclass. Total cholesterol is made up from free cholesterol and cholesterol esters. Figure 3.2 shows this hierarchy of measurements.

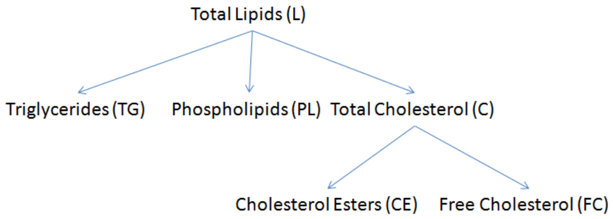


Figure 3.2: Hierarchy of the lipoprotein subclasses

The observations for each lipoprotein measurement are the concentration in the total serum sample. So for example an observation of 1 mmol/l of XL–HDL–C means that in one litre of serum there is 1 mmol of total cholesterol embedded in extra large HDL particles.

Within the remaining metabolites in the LIPO window (table 3.5) there are a number of other metabolites that are a composition of two or more other metabolites within the window. These are:

- VLDL triglycerides - The total VLDL triglyceride concentration in the serum sample. Equal to the sum of the 6 VLDL–TG concentrations from each of the VLDL subclass sizes (VLDL–TG = XXL–VLDL–TG+XL–VLDL–TG+L–VLDL–TG+M–VLDL–TG+S–VLDL–TG+XS–VLDL–TG)

- LDL triglycerides - The total LDL triglyceride concentration in the serum sample. Equal to the sum of the 3 LDL–TG concentrations from each of the LDL subclass sizes (LDL–TG = L–LDL–TG+M–LDL–TG+S–LDL–TG)

- HDL triglycerides - The total HDL triglyceride concentration in the serum sample. Equal to the sum of the 4 HDL–TG concentrations from each of the HDL subclass sizes (HDL–TG = XL–HDL–TG+L–HDL–TG+M–HDL–TG+S–HDL–TG)

- Serum triglycerides - made up from the above 3 triglyceride measurements plus IDL triglycerides (Serum-TG - VLDL–TG+IDL–TG+LDL–TG+HDL–TG)

- VLDL cholesterol esters- The total VLDL cholesterol esters concentration in the serum sample. Equal to the sum of the 6 VLDL–CE concentrations from each of the VLDL subclass sizes (VLDL–CE = XXL–VLDL–CE+XL–VLDL–CE+L–VLDL–CE+M–VLDL–CE+S–VLDL–CE+XS–VLDL–CE)

- LDL cholesterol esters- The total LDL cholesterol esters concentration in the serum sample. Equal to the sum of the 3 LDL–CE concentrations from each of the LDL subclass sizes (LDL–CE = L–LDL–CE+M–LDL–CE+S–LDL–CE)

- HDL cholesterol esters- The total HDL cholesterol esters concentration in the serum sample. Equal to the sum of the 4 HDL–CE concentrations from each of the HDL subclass sizes (HDL–CE = XL–HDL–CE+L–HDL–CE+M–HDL–CE+S–HDL–CE)

- VLDL free cholesterol - The total VLDL free cholesterol concentration in the serum sample. Equal to the sum of the 6 VLDL-FC concentrations from each of the VLDL subclass sizes (VLDL–FC = XXL–VLDL–FC+XL–VLDL–FC+L–VLDL–FC+M–VLDL–FC+S–VLDL–FC+XS–VLDL–FC)

- LDL free cholesterol - The total LDL free cholesterol concentration in the serum sample. Equal to the sum of the 3 LDL–FC concentrations from each of the LDL subclass sizes (LDL–FC = L–LDL–FC+M–LDL–FC+S–LDL–FC)

- HDL free cholesterol - The total HDL free cholesterol concentration in the serum sample. Equal to the sum of the 4 HDL–FC concentrations from each of the HDL subclass sizes (HDL–FC = XL–HDL–FC+L–HDL–FC+M–HDL–FC+S–HDL–FC)

- VLDL cholesterol - The total VLDL cholesterol concentration in the serum sample. Equal to the sum of the 6 VLDL–C concentrations from each of the VLDL subclass sizes (VLDL–C = XXL–VLDL–C+XL–VLDL–C+L–VLDL–C+M–VLDL–C+S–VLDL–C+XS–VLDL–C)
- LDL cholesterol - The total LDL cholesterol concentration in the serum sample. Equal to the sum of the 3 LDL–C concentrations from each of the LDL subclass sizes (LDL–C = L–LDL–C+M–LDL–C+S–LDL–C)
- HDL cholesterol - The total HDL cholesterol concentration in the serum sample. Equal to the sum of the 4 HDL–C concentrations from each of the HDL subclass sizes (HDL–C = XL–HDL–C+L–HDL–C+M–HDL–C+S–HDL–C)
  - HDL cholesterol is also the sum of HDL2 and HDL3 cholesterol (HDL–C = HDL2+HDL3)
- Remnant cholesterol - made up from VLDL and IDL cholesterol (Remnant–C = VLDL–C+IDL–C)
- Total serum cholesterol - made up from LDL, HDL and remnant cholesterol (Serum–C = LDL–C+HDL–C+Remnant–C)

3.1.1.2 LMWM window

The low molecular weight metabolite window contains 19 metabolites, and unlike the other 2 windows, these have no obvious hierarchy. The metabolites from within this window are not as strongly correlated with one another as metabolites from the other two windows are.

3.1.1.3 LIPID window

Of the 15 metabolites from the LIPID window 8 are related to fatty acids. As with the lipproteins these also have a hierarchy, as shown in figure 3.3. Total fatty acids is a sum of polyunsaturated, monounsaturated and saturated fatty acids and polyunsaturated is made up from both Omega-3 and Omega-6 fatty acids. 22:6, docosahexaenoic acid is a type of Omega-3 fatty acid and 18:2, lioleic acid is a type of Omega-6 fatty acid. When taking decisions in analysing these and the lipoprotein metabolites we need to consider their makeup to avoid introducing collinearity into any statistical models we use.

3.2 Data description

3.2.1 Missing Data

The 3777 observations from the women with NMR metabolite profiles were checked for any patterns of missing data. First only considering the 149 non-ratio metabolites we find that 3390 (89.8%) individuals have no missing values in

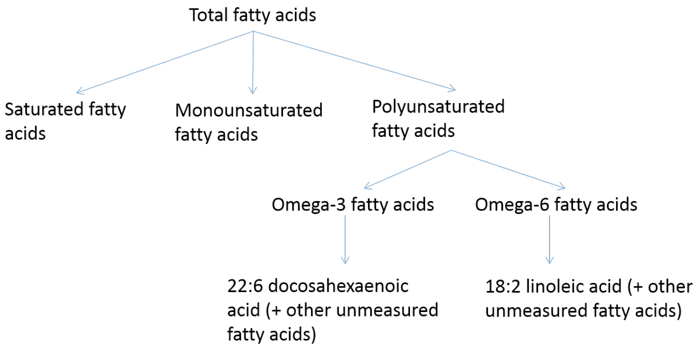

Total fatty acids
Saturated fatty acids
Monounsaturated fatty acids
Polyunsaturated fatty acids
Omega-3 fatty acids
Omega-6 fatty acids
22:6 docosahexaenoic acid (+ other unmeasured fatty acids)
18:2 linoleic acid (+ other unmeasured fatty acids)

Figure 3.3: Hierarchy of the fatty acids

these metabolites (i.e. they are complete records). There are no missing values in metabolites from the LIPO window. The majority of incomplete records are due to 2 of the amino acids - Creatinine, which has 308 missing values, and Glycerol, which has 78. Excluding these 2 metabolites, 3767 (99.7%) individuals had no other missing values. The 10 who did can be broken down as follows (again omitting Creatinine and Glycerol)

- 3 individuals are missing all values from the LMWM and LIPID windows (32 missing values)
- 4 individuals are missing all LIPID values (15)
- 1 individual is missing all LIPID values and Pyruvate (16)
- 1 individual is missing values for Pyruvate and Glutamine (2)
- 1 individual is missing a value for 3-hydroxybutyrate (1)

There are only 2787 (73.8%) complete records when focussing on the 78 ratio variables. However this is due to the fact that they are calculated from the observed metabolite concentrations and there are two scenarios that lead to missing data:

1. Since the fatty acid concentrations are missing in 8 individuals (those that are missing all their LIPID measurements), the fatty acid ratios are also therefore missing in those 8 individuals.
2. Some lipid ratios are missing for individuals who have a concentration of zero for a particular lipoprotein or lipoproteins.

So in both these cases it is correct that the ratio is undefined. The number of observations missing for each NMR metabolite is shown in the final column of table 3.5.

3.2.2 Biomarker Distributions

Many of the 149 (non-ratio) metabolites are highly skewed, particularly the low and very low density lipoproteins and the majority of these are right-skewed. The skewness ranges from -0.5 to 15.3 with a mean skewness of 1.36 and median skewness of 0.81, with 60 metabolites having a skewness greater than 1. A histogram of the skewness of the metabolites is shown in figure 3.4a. Now considering the 79 biomarkers that are ratios, we can observe that the distributions are different. The mean and median skewness is 0.07 and 0.17 respectively, with skewnesses ranging from -2.86 to 4.63. The histogram of skewness is shown in figure 3.4b. We do not see the extreme right skewness observed in the metabolite concentrations, we see more moderately skewed data, both left and right.

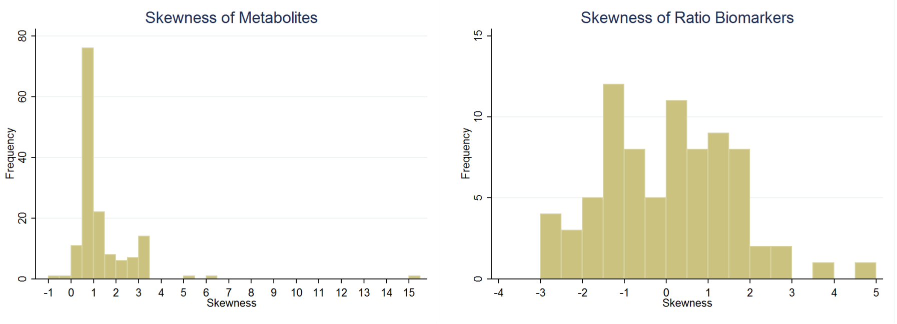


Figure 3.4: Histograms of skewness of a) 149 Metabolite concentrations b) 79 Ratios

In the LIPO window it is in particular the larger VLDL particles that exhibit the greatest skewness, a histogram of extremely large VLDL lipids illustrates a distribution that is typical of this lipoprotein class (figure 3.5a). It shows the majority of observations are equal to or close to zero but there is a very long tail in the distribution where there are a low number of observations with very high concentrations of this metabolite. In the smaller and denser lipoprotein subclasses the distributions are still skewed, but to a lesser extent as shown in figure 3.5b which illustrates the distribution of large HDL lipids. The metabolites from the LIPID and LMWM windows in general are less skewed than the lipoproteins, although the three metabolites with the greatest skewness were from the LMWM window: Acetate, Citrate and Glucose (15.33, 6.00 and 5.29 respectively).

3.2.3 Zero values

One thing is apparent when observing the histogram of large HDL lipids is that there is a peak at 0. Many of the lipoproteins have a number of observations equal to 0, which could be due to the fact that the individual had none of

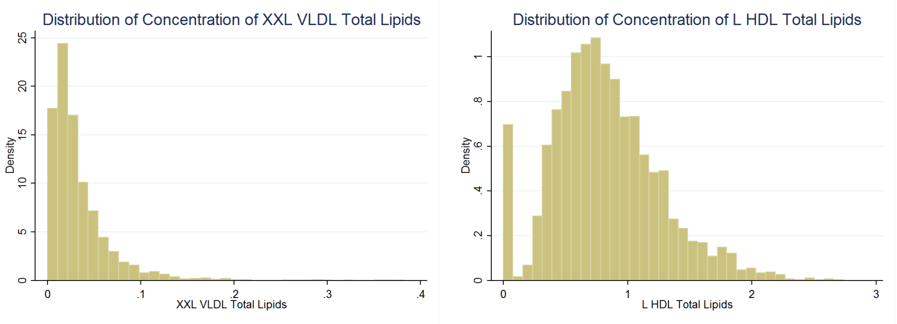


Figure 3.5: Histograms of the distribution of a) XXL VLDL total lipids and b) L HDL total lipids

that particular biomarker, or it could be due that the true concentration of the metabolite in the sample was so small as to be below the threshold of detection. If the particle concentration of concentration of total lipids of a lipoprotein subclass is equal to 0, all of the "children" (as per figure 3.2) of that lipoprotein are also equal to 0. For example if the concentration of total cholesterol for a particular lipoprotein subclass was measured to be 0, then both the free cholesterol and cholesterol esters for that subclass would also be equal to 0. Table 3.5 has a column noting the number of zero values that there were for each biomarker.

3.2.4 NMR biomarker correlation structure

With 228 biomarkers there are 228(227)/2=25878 possible pairwise correlations. In this section we will use Spearman correlations as our measure of association as it is not as affected by outliers as Pearson correlation. The histogram of pairwise Spearman correlations for the untransformed biomarkers is shown in figure 3.6 illustrating that a large number of the metabolites are highly correlated, with a median correlation of 0.14 (IQR -0.11,0.41) and a median *absolute* correlation of 0.28 (IQR 0.13,0.51). In fact if you split the biomarkers into the ratio and non-ratio metabolites it is possible to see that most of the metabolite concentrations are positively correlated, whereas the distribution ratio biomarker correlations are symmetric about 0 (figures 3.7a and b).
Figures 3.15 and 3.16 combine a dendrogram (based on hierarchical clustering using centroid linkage) and a heatmap, with the metabolites sorted into groups shown by the dendrogram at the edge of the grid, and the colour of each cell representing the Spearman correlation of each pair of variables. Red represents a positive correlation and blue a negative, with the intensity of colour representing the strength of association.

The first figure contains the 149 metabolites and two large groups of very strongly correlated metabolites are immediately obvious, the group in the top

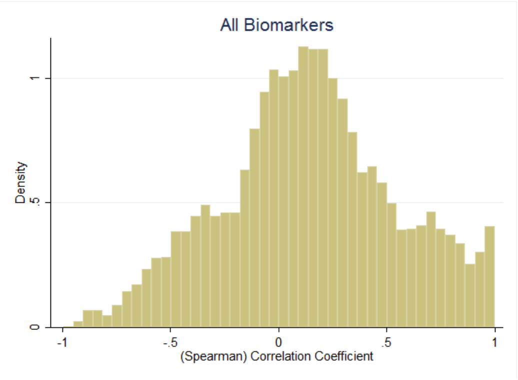


Figure 3.6: Histogram of the distribution of Spearman correlation coefficients for all NMR biomarkers

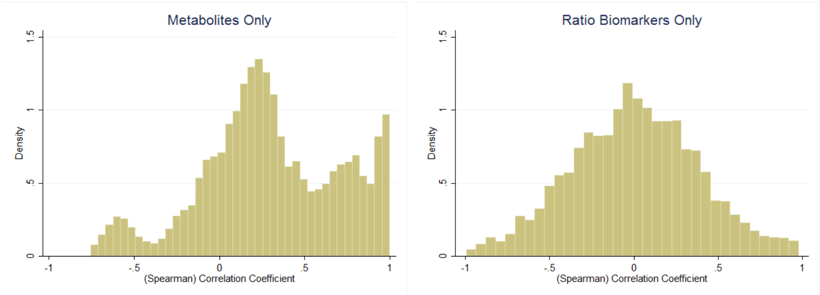


Figure 3.7: Histogram of the distribution of Spearman correlation coefficients for a) Metabolite concentrations and b) ratio biomarkers

left, mostly made up of fatty acids, IDL, LDL and XS VLDL metabolites, and the group in the centre, mostly made up from VLDL metabolites. Both these groups have strong correlations within the group, and a weaker correlation between groups. Another group of metabolites that strongly correlated are the HDL metabolites (towards the bottom right corner of the diagram) again these are strongly correlated with each other but they are negatively correlated with the VLDL group. The metabolites from the LMWM window tend not to be so strongly correlated with any of the other metabolites.

Looking at the heatmap of the ratios there are fewer extremely high correlations and the variables cluster as to their type (i.e. phospholipids, triglycerides, cholesterol) rather than by their lipoprotein subclass e.g. the proportion of triglycerides within the total lipids of each of VLDL, LDL, IDL and HDL are correlated with each other.

When the concentrations and ratios are combined in a single heatmap (figure 3.17) some interesting observations arise, although it can be difficult to see clearly on the printed diagram due to the size of the image required. The heatmap shows that the proportions of triglycerides are positively correlated with the concentrations of VLDL metabolites. This suggests those that have a higher concentration of VLDL metabolites also have a higher proportion of triglycerides within those (and other) metabolites.

So to summarize a few of the points of interest identified by inspecting the Spearman correlation matrix:

- Many of the metabolites are strongly positively correlated.
- Concentrations of cholesterol and phospholipids within lipoprotein subtypes are strongly positively associated i.e. those with higher LDL phospholipids are also likely to have higher LDL cholesterol (including cholesterol esters and free cholesterol).
- Concentrations of triglycerides are positively associated with the other metabolites within their lipoprotein subtype but are also positively correlated with triglycerides from other lipoprotein types and fatty acids. i.e. LDL triglycerides are positively associated with LDL phospholipids and cholesterol, but are also associated with HDL triglycerides.
- HDL metabolites are strongly positively associated with one another but negatively associated with the lower density lipoprotein metabolites.
- Metabolites from the LMWM window are less strongly associated with any of the other metabolites.
- Proportions of triglycerides are associated across lipoprotein types and are also associated with higher levels of VLDL lipoproteins i.e. those with higher concentrations of VLDL metabolites are likely to have higher proportions of triglycerides within any of the lipoprotein types.

3.3 Analytical considerations

3.3.1 Transformations

For some analyses it may be desirable to transform the skewed variables to make them closer to a normal distribution. For example, many of the methods used in chapters 5 and 6 are based around the correlation coefficient, with the Pearson correlation coefficient having a greater power than the Spearman correlation coefficient if the assumptions underlying it are correct, one of these assumptions being the variables are normally distributed.

First considering only the non-ratio metabolites which are mostly right-skewed, two transformations which may be appropriate for many of the variables are the $Y^{\frac{1}{3}}$ (cube root) and the $\log(Y)$ (log) transform. Either of these would be appropriate for a number of, if not all, the metabolites. However, when selecting a transformation, a decision must be made as to whether using the same transformation on all metabolites is an appropriate approach, or whether a transformation should be selected for each metabolite separately.

The former approach has the advantage that the metabolites will all be on the same transformed domain, retaining some level of interpretability. However this has the downside that it may not be an appropriate transformation for all metabolites and may, in some cases, make things worse. By comparing the different transformations for each metabolite and selecting the "best" we can avoid these problems, however it will mean that some variables will represent the concentration of a metabolite, some the log concentration and others the cube root concentration. This results in us changing the relationship between some metabolites and making interpretation more complicated.

Another issue to consider when transforming data is what happens when some of the observations are equal to zeros. There are 98 metabolites that have a number of observed values equal to zero, although in many of these there are only a few zero values which doesn't pose too much of a problem. However in the large, very large and extremely large VLDL particles there are non-trivial numbers of zero values (268, 585 and 356 zero observations in each, respectively). A zero is a valid observation, an individual can plausibly have none of a particular metabolite present in their serum sample, so we must consider this when applying any transformation. Also, the log of zero is undefined, so if using a log transform we will have to decide how to treat the zeroes. The zeroes could also be due to the metabolite being below the limit of detection.

There are a number of methods of dealing with values below the limit of detection. A review by Hewitt and Ganser [29] described and evaluated four different types of method based on maximum likelihood regression (MLE), log-probit regression (LPR), non-parametric methods and substitution methods. Although the MLE and LPR methods were the best performing in their review, the bias introduced by simple substitution was moderate when the proportion of observations below the limit of detection was less than 50% and the sample size was large, as is the case with our data (the variable with the greatest number of zeroes still only has 15%). As the substitution method is the easiest method to employ it is the method selected for use in this thesis. In this instance we substitute the zero values with a value equal to half the smallest observation [30].

We compared the distributions of each metabolite using the log-transform (using the above method of dealing with zeroes) with the untransformed and cube-root-transformed distributions. Of the 149 metabolites 59 are least skewed if a log transform is used, 64 if a cube-root transform is used and 26 are least skewed if no transform is performed. The large VLDL particles strongly favour the cube-root transform as the log transform tends to over compensate and the distribution becomes left-skewed. Also, for both the cube-root and log transform you get a peak at the left of the distribution, relating to the observations that are equal to zero in the untransformed data. The log and cube-root transforms were chosen as candidates as they are commonly used transformations for right-skewed data.

Although there are a similar number of metabolites that favour each of the log and cube-root transforms the metabolites that favour the log transform in general are only slightly less skewed than when they are cube-root transformed however some of the variables that favour the cube-root transform are not suitable for log-transformation. So if a single transformation was to be selected for all metabolites the cube-root-transform is a more appropriate choice under the criterion of reducing skewness. If per-metabolite transformations are selected (i.e. rather than choosing a single transformation for all metabolites, the "best" transformation is selected for each metabolite) this can have a consequence that the metabolites are on different scales and affect interpretability of any estimates obtained. However, in chapters 4-7 this strategy is selected, to maximise the chance of univariate and pairwise bivariate normality, required for the analysis in these chapters.

Finally, whether the metabolite is transformed or not, in our analyses we will standardise them by subtracting their mean value and dividing by the standard deviation, resulting in each of the standardized metabolites having a mean of 0 and standard deviation of 1. This allows all the NMR metabolites to exist on the same scale and the interpretation of a change of 1 unit to be equal to a change of 1 standard deviation of that metabolite (or of its transformed value).

The transformations discussed above are not suitable for left-skewed data. So in order to transform the 79 ratio variables, a different transform would be required, potentially the Y^2 or Y^3. This would mean that it would not be appropriate to use a single transformation across all 228 biomarkers (i.e. the concentrations and the ratios).

3.4 Repeated Measures, Reliability and Agreement

A subset of the women in the BWHHS gave a second blood sample a week after they provided their first, which provides us with an opportunity to assess the short term reliability of the biomarkers measured. By short term reliability we mean that we are investigating the consistency of biomarkers across different samples from the same individual, with the variation in the two samples arising from measurement error, sampling differences and biological changes. It was planned that in each of the 23 towns 10 women would be randomly selected among those who provided a baseline blood sample to provide a second sample

1 week later. However, recruitment was successful in only 11 towns, with a total of 45 women providing a second sample (19.6% of the desired total).

In addition to this, there were 6 biomarkers that were measured both using standard techniques and ^{1}H-NMR spectroscopy on all women. This also gives us an opportunity to check the agreement between the two measurement methods for those 6 biomarkers.

3.4.1 Methods

3.4.1.1 Representativeness

We compared the baseline attributes of women who provided repeat samples with those who did not provide it using univariable logistic regression (where the outcome was 1 if the participant provided two samples, 0 otherwise). The variables we compared were all measured at entry into the cohort and were: age (measured in years), body mass index (BMI; kg/m^2), smoking status (never/ever), coronary heart disease (no/yes), cardiovascular disease (no/yes), diabetes (no/yes), and systolic blood pressure (mmHg). The joint effect of these variables was not examined because of the low number of women who provided a second sample (to avoid small sample bias [31]).

3.4.1.2 Transformations

As discussed in section 3.3.1 there are both left and right skewed data across the 228 NMR biomarkers, so we need to assess the most appropriate method of dealing with these, as the methods we use to assess reliability are based on an assumption of normal distributions. In this analysis we are less concerned with the interpretation of individual biomarkers, only that their measurement is consistent across time. So to achieve this we measured the skewness of each metabolite in the measurements taken from the first sample for all participants. Four alternative transformations were considered to deal with right-skewed data: $Y^{\frac{1}{3}}$ (cube root) and $\log(Y)$ transform (where Y represents the original metabolite). For left skewed data the Y^2 (square) and Y^3 (cube) transforms were applied. To allow for values equal to zero in the log transform (since $\log(0)$ is undefined) as discussed above zeroes were replaced with values half the size of the minimum observation. The transformation that resulted in a skewness closest to 0 was selected for each NMR biomarker. The values were then internally standardized to allow comparisons between metabolites measured on different scales.

3.4.1.3 Short-term reliability

To assess the short term reliability of both NMR and standard biomarkers the intraclass correlation (ICC), often denoted by λ, was calculated [32]. This is to assess the size of the variability of each measure in the population (the between individual variance: σ_b^2) relative to the amount of variability seen within an individual (the within individual variance: σ_w^2). Hence

$$\lambda = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}$$

λ takes values between 0 and 1, with 0 meaning that the measurement is highly variable within an individual, and a value close to 1 means that the measurement is stable within an individual. Fleiss [33] suggests a rule of thumb that an ICC of less than 0.4 suggests poor reliability, 0.4 to 0.75 being fair to good and greater than 0.75 being excellent.

Letting Y_{i1} and Y_{i2} denote respectively the first and second measurements of the (possibly transformed) metabolite Y for participant i ($i = 1, 2, \ldots, n$), and $\overline{Y}$ the overall mean, an estimate of λ is obtained by first estimating its components:

$$\hat{\sigma}_w^2 = \frac{\sum_{i=1}^{n}(Y_{i1} - Y_{i2})^2}{2n}$$

and

$$\hat{\sigma}_b^2 = \hat{\sigma}^2 - \hat{\sigma}_w^2,$$

where,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{2}(Y_{ij} - \overline{Y})^2}{2n}.$$

Note that it is possible that $\hat{\sigma}_b^2$ takes negative values. This may happen when the variance between the participants' means is smaller than expected given $\hat{\sigma}_w^2$ [32]. In this instance $\hat{\lambda}$ is set equal to 0. Note also that the ICC estimated on replicates not taken at the same time can be interpreted as a measure of reliability only if the time interval between repeated samples can be reasonably assumed not to matter. To investigate this we also tested whether the mean difference between the first and second measurements (of women with both measurements) was significantly different from zero (using a paired-t test [34]).

3.4.1.4 Agreement

Agreement between measures of the same biomarker obtained by standard methods and ^{1}H-NMR Spectroscopy was assessed by comparison of the means via a paired t-test, inspection of Bland-Altman plots [35] and comparison of association with incident CHD of each of the biomarkers from both methods, by estimating the hazard ratio.

3.4.2 Results

3.4.2.1 Inclusions/Exclusions

Of the 45 repeat blood samples that were taken, 4 were incomplete, so 41 samples were sent for NMR analysis, all resulting in a successful NMR metabolite profile.

However, of these 41 samples, 4 did not have baseline NMR profiles (either due to an incomplete set of blood samples or technical problems). This left 37 women who had an NMR profile both from a sample taken at baseline and from a sample taken 1 week later and can therefore be used to estimate the short term reliability. The flowchart in figure 3.8 lays out the samples included and excluded.

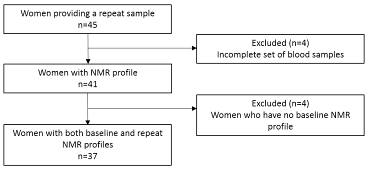

Women providing a repeat sample
n=45
Excluded (n=4)
Incomplete set of blood samples
Women with NMR profile
n=41
Excluded (n=4)
Women who have no baseline NMR profile
Women with both baseline and repeat NMR profiles
n=37

Figure 3.8: Numbers of repeat samples to be included in analysis

Summary values for the 3777 participants with valid baseline NMR metabolite profiles are shown in table 3.1, stratified by whether they provided two blood samples or not. The odds ratios (OR) of providing a second sample are also provided in the table, along with their related p-value. No individuals who had prevalent CHD or diabetes provided a repeat sample so the odds ratios could not be estimated, however a p-value from the χ^2 test is provided. No strong evidence was found of an association between any of the baseline characteristics and the odds of providing a second sample.

Table 3.1: Baseline characteristics stratified by whether participant provided 1 or 2 samples. (SBP: Systolic Blood Pressure; DBP: Diastolic Blood Pressure; Deprivation score: In top quintile of Carstairs deprivation score) † p-values from Fisher's exact test

Baseline characteristics	1 sample
	Mean(SD)
	n=3740
		2 samples
		Mean(SD)
		n=37
			Odds ratio
			(95% CI)
				p-value

Age(years)	68.8 (5.5)	67.9 (5.0)	0.97 (0.91,1.03)	0.30
BMI (kg/m^2)	27.6 (5.0)	26.7 (3.7)	0.96 (0.90,1.03)	0.30
Waist-Hip Ratio	0.82 (0.07)	0.80 (0.07)	0.02 (0.00,2.94)	0.12
SBP(mmHg)	147.1 (25.2)	141.2 (23.6)	0.99 (0.98,1.00)	0.16
DBP (mmHg)	79.4 (11.7)	79.6 (10.6)	1.00 (0.97,1.03)	0.94
Deprivation score	22.5%	35.1%	1.87 (0.95,3.69)	0.07
Ever Smoker	44.0%	32.4%	0.61 (0.31,1.22)	0.16
Prevalent Diabetes	4.6%	0.0%	N/A	0.41†
Prevalent CHD	3.8%	0.0%	N/A	0.40†

3.4.2.2 Transformations

Of the 228 NMR biomarkers 38 were least skewed if left untransformed, 80 were least skewed if a cube-root transform was applied. 78, 10 and 22 biomarkers were least skewed using the log, square and cube transforms respectively. The intraclass correlations were estimated using these transformed data. The transform selected for each biomarker is shown in table 3.6.

3.4.2.3 Short term reliability of the ^{1}H-NMR biomarkers

The ICC was estimated for all 228 NMR biomarkers, with estimates ranging between 0 (Histidine and proportions of cholesterol and cholesterol esters in extra large HDL) to 0.901 (apolipoprotein B by apolipoprotein A1). Details of the 10 highest and 10 lowest estimated ICCs are shown in Table 3.2 and a full list of all ICCs are provided in table 3.6. The distributions of the estimated ICCs for each of the three metabolite windows are shown in figure 3.9.

Table 3.2: Highest and lowest ICCs

Biomarkers with highest ICC	Window	ICC (95% CI)
Apolipoprotein B by A1	LIPO	0.91 (0.85,0.96)
Mean Diameter of HDL Particles	LIPO	0.90 (0.83,0.95)
Large HDL Total Cholesterol	LIPO	0.89 (0.81,0.94)
Large HDL Cholesterol Esters	LIPO	0.89 (0.81,0.94)
Large HDL Free Cholesterol	LIPO	0.88 (0.80,0.94)
% of cholesterol esters in medium HDL	LIPO	0.85 (0.74,0.92)
Extra large HDL Phospholipids	LIPO	0.84 (0.73,0.92)
Large HDL Total Lipids	LIPO	0.84 (0.73,0.92)
Large HDL Particle Concentration	LIPO	0.83 (0.71,0.91)
% of monounsaturated fatty acids in total fatty acids	LIPID	0.83 (0.71,0.91)
Biomarkers with lowest ICC	**Window**	**ICC (95% CI)**
Histidine	LMWM	0.00 (0.00,1.00)
% of cholesterol esters in XL HDL	LMWM	0.00 (0.00,1.00)
% of cholesterol in XL HDL	LMWM	0.00 (0.00,1.00)
Albumin	LIPO	0.07 (0.00,0.82)
Small HDL Cholesterol	LIPO	0.10 (0.00,0.72)
Small HDL Cholesterol Esters	LIPO	0.14 (0.01,0.63)
% of phospholipids in extra small VLDL	LIPO	0.14(0.01,0.64)
Total Lipids in Small HDL	LIPO	0.17 (0.02,0.61)
Particle concentration of Small HDL	LIPO	0.18 (0.02,0.61)
Phospholipids in Medium HDL	LIPO	0.24 (0.05,0.61)

Under the Fleiss criteria, the analysis resulted in 25(11.0%) metabolites classed as having poor reproducibility, 135(59.2%) classed as fair to good and 68(29.8%) classed as excellent. Looking at the 3 separate windows, the mean estimated ICC for 186 metabolites in the LIPO window measures was 0.65, the mean ICC for the 19 metabolites in the LMWM window was 0.46 and the mean ICC for

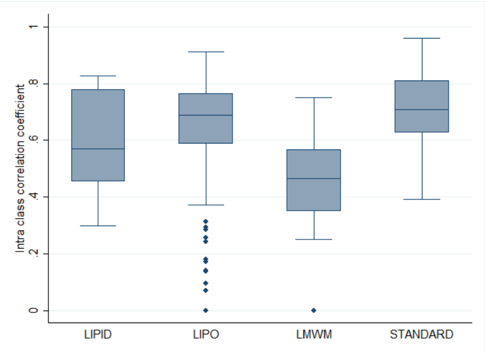


Figure 3.9: Distribution of intra-class correlation coefficients, by biomarker type

the 23 metabolites in the LIPID window was 0.59. These results are shown in table 3.3 along with their categories according to the Fleiss classification.

Table 3.3: Mean intra-class correlations for each group of biomarkers, along with classifications (using Fleiss' rule of thumb)

	LIPO	LMWM	LIPID	All NMR	Standard
	Window	Window	Window	Biomarkers	Biomarkers
	N=186	N=19	N=23	N=228	N=37
Mean ICC	0.65	0.46	0.59	0.63	0.71
Poor	15(8.1%)	5(26.3%)	5(21.7%)	25(11.0%)	1(2.7%)
Fair	112(60.2%)	14(73.7%)	9(39.1%)	135(59.2%)	21(56.8%)
Excellent	59(31.7%)	0(0.0%)	9(39.1%)	68(29.8%)	15(40.5%)

The mean ICC in the LMWM window was markedly lower than the estimated mean ICCs in the other two groups, and this difference persists even if Histidine, which had the lowest ICC in the LMWM window, was excluded from the analysis as a potential outlier. The poorer performance of biomarkers in the LMWM window can also be seen looking at the categorisation, as 5 out of 19 biomarkers (26.3%) from this window are classed as having poor reproducibility compared with 8.1% from the LIPO window respectively, and by viewing the distributions in figure 3.9 where it shows that the 75th percentile of the ICCs in the LMWM window is below the 25th percentile in the other two windows. Out of the 10 highest estimated ICCs for the 137 NMR biomarkers, 9 were from the LIPO window, 1 from the LIPID window and none from the LMWM window. Measurements relating to large HDL particles account for 8 out of the 10 highest ICCs, which was in contrast to small and medium sized HDL particles which were among the lowest.

3.4.2.4 Short term reliability of standard biomarkers

The mean estimated ICC for the 36 standard biomarkers was 0.71, with 16 (44.4%) being classed as excellent, 19 (52.7%) as fair to good and only 1 (2.8%) classed as poor (Phosphate). It can be seen in figure 3.9 that the ICCs are generally higher than those for any of the 3 categories of NMR metabolites, although the median estimated ICC of the LIPO window is close to that of the standard biomarkers. The highest ICC from the standard biomarker group is from HDL cholesterol (0.96 95% CI 0.92,0.98), with Gamma-Glutamyl Transpeptidase, Urate, Creatinine and LDL Cholesterol close behind. Again, the full list of ICCs can be found in table 3.6.

3.4.2.5 Agreement

Albumin, Creatinine, Glucose, HDL Cholesterol, LDL Cholesterol and Serum Triglycerides were the 6 biomarkers measured both by standard techniques and by NMR Spectroscopy. However for Albumin the reported unit in the two methods are different with the standard measurement in g/l and the NMR measurement unit described as “signal area”, with the two variables having very different observed values, and as a result showing very poor agreement, however the variables were retained for analysis as there was some association between the two. LDL cholesterol measured by standard methods includes IDL cholesterol, whereas the NMR biomarkers are separate. So for comparison the NMR measures of IDL and LDL cholesterol are added together. Table 3.4 displays the means and standard deviations of these biomarkers for both measurement methods, figure 3.10 shows a scatter plot of the standard and NMR observations of each biomarker and figure 3.11 shows the Bland-Altman plots (after transformation).

These illustrate a strong association between the standard and NMR measures (correlation coefficient, r = 0.74-0.95) in all metabolites except Albumin, where the association was weaker (r=0.34). In a previous publication [36], the association between standard biomarkers and NMR obtained was investigated, with LDL cholesterol found to have a Pearson correlation of 0.88 between standard and NMR and HDL cholesterol having a correlation of 0.93. These are slightly higher, but comparable to, the values obtained in our analysis of 0.81 and 0.82 respectively. However, no prior publications were found looking at the agreement in the absolute concentrations of metabolites using this platform.

A t-test to assess the null hypothesis that the observed mean of each variable is the same when measured by NMR and by standard methods yields a p-value of less than 0.0001 for 5 out of the 6 variables, with HDL cholesterol resulting in a p-value of 0.013. However, with a sample size of 3777 a small difference in means can lead to a very low p-value, also the significance of the test refers only to a shift in the mean which is just a single aspect of the distribution. The Pearson correlation of the NMR and standard measurements in Albumin

Table 3.4: Mean biomarker concentrations obtained using standard and ^{1}H-NMR techniques in the 3777 women who had a baseline NMR profile. To ensure fair comparisons when using transformed data we force each pair of measurements to have the same transform. The transformation that gives the lowest mean skewness is selected. Untransformed means are in mmol/l, aprt from Albumin which is measured in g/l for the standard measurement and signal area for the NMR. Hazard ratios are crude estimates.

Biomarker (Method)	Untransformed Mean(SD)	Transformed Mean(SD)	CHD Hazard Ratio for 1 SD change (95% CI)	ICC (95% CI)
Creatinine (Stand)	0.08(0.01)	-2.54(0.15)	1.34(1.17,1.53)	0.93(0.87,0.96)
Creatinine (NMR)	0.06(0.01)	-2.82(0.22)	1.31(1.12,1.53)	0.75(0.59,0.87)
Glucose (Stand)	6.05(1.64)	1.78(0.19)	1.06(0.92,1.22)	0.83(0.71,0.92)
Glucose (NMR)	4.94(1.40)	1.57(0.20)	0.98(0.84,1.15)	0.47(0.24,0.71)
HDL Cholesterol (Stand)	1.66(0.46)	1.17(0.11)	0.69(0.59,0.81)	0.96(0.93,0.98)
HDL Cholesterol (NMR)	1.67(0.45)	1.18(0.11)	0.75(0.65,0.87)	0.74(0.57,0.86)
LDL Cholesterol (Stand)	4.14(1.08)	1.59(0.14)	1.10(0.95,1.28)	0.92(0.86,0.96)
LDL Cholesterol (NMR)	3.38(1.01)	1.49(0.14)	1.09(0.94,1.27)	0.64(0.44,0.81)
Serum Triglycerides (Stand)	1.86(0.96)	0.51(0.45)	1.34(1.16,1.54)	0.78(0.63,0.88)
Serum Triglycerides (NMR)	1.68(0.84)	0.41(0.45)	1.35(1.17,1.56)	0.76(0.60,0.87)
Albumin (Stand)	43.97(2.51)	43.97(2.51)	0.95(0.82,1.11)	0.72(0.53,0.82)
Albumin (NMR)	0.99(0.01)	0.99(0.01)	1.01(0.87,1.17)	0.07(0.00,0.82)

is 0.34 which is much lower than the correlation in the other 5 metabolites, which range from 0.79 to 0.94). Another, potentially more informative method of assessing agreement is to inspect the Bland-Altman plots of the difference. From the plots we can see that there is better agreement between the standard and NMR measurements for HDL cholesterol and Triglycerides than there is for each of LDL cholesterol, Creatinine and Glucose where the standard measurements were greater than the NMR measurements. We also estimated the hazard ratios for CHD risk for each pair of measurements, using a Cox proportional hazards model, to confirm if the estimates were similar. We found in all pairs of measurements the estimates were relatively close and there was a large overlap in the confidence intervals (table 3.4).

3.4.2.6 Discussion

The estimated ICCs for the NMR biomarkers do seem to be comparable with the ICCs found in other studies on biomarkers obtained from more traditional methods of blood serum analysis [37, 38] and on biomarkers obtained via Mass Spectrometry [39, 40]. Two recent studies using mass spectrometry techniques resulted in lower intra class correlations than we observed (i.e. poorer reliability), however the time difference between observations was greater than in our study (4 months and 1-2 years compared to 1 week) so the results are not

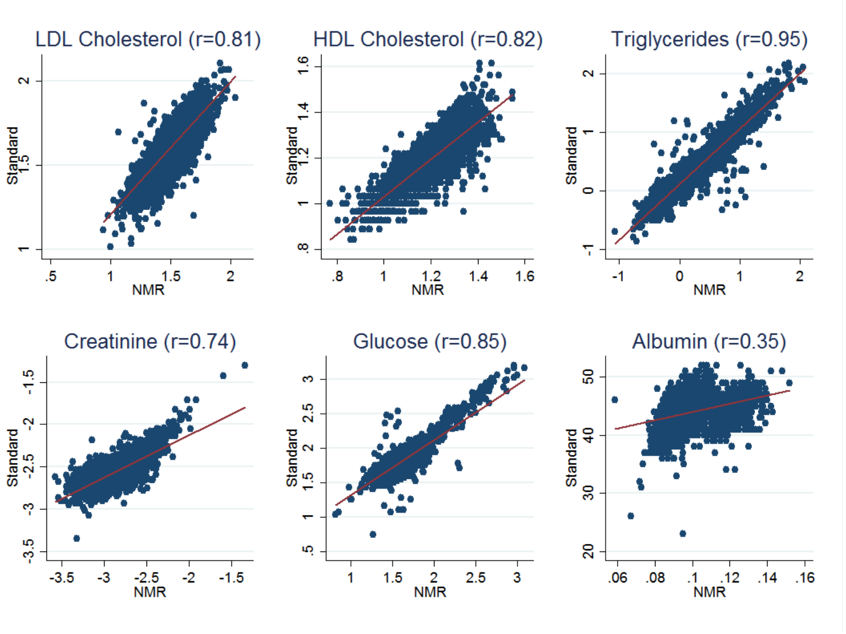


Figure 3.10: Scatter plots of the 6 untransformed biomarkers obtained using standard methods and ^{1}H-NMR Spectroscopy, red line is the line of best fit

directly comparable [41, 42].

Overall in this study, the estimated ICCs from NMR data were lower than the ICCs estimated for the biomarkers measured from the same women obtained using more traditional means, although in most cases there was not a large difference. However some of the NMR biomarkers had very low estimated short term reliability, Albumin, Histidine and the % of cholesterol and cholesterol esters in XL HDL being the 4 lowest. If these findings are replicated in other populations it will have implications on epidemiological findings based on these ^{1}H-NMR quantified biomarkers.

One difference to note is that the samples analysed using ^{1}H-NMR Spectroscopy had been in storage between 11 and 13 years, which was longer than for any of the samples used to estimate the concentrations of biomarkers using standard methods. It is possible that the additional storage time may have led to some extra degradation of the samples leading to greater variation within the samples. A future study would benefit from comparing samples that were analysed using the two methods at the same point in time.

There is some disagreement between the two measurement methods when comparing the 5 biomarkers measured using ^{1}H-NMR and standard methods. LDL

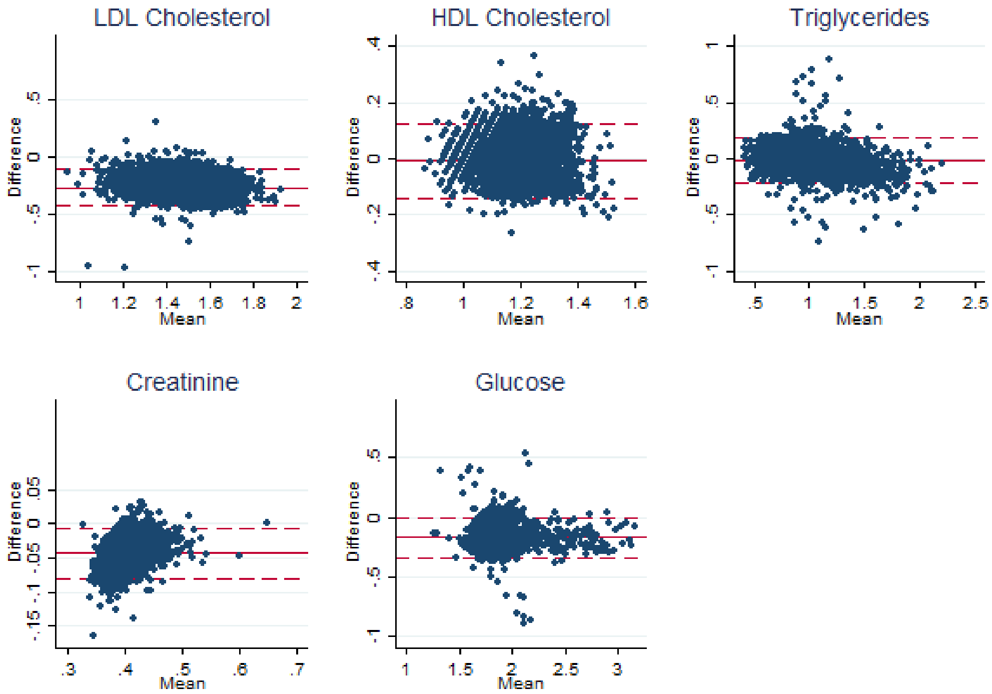


Figure 3.11: Bland-Altman plots of 5 biomarkers obtained (excluding Albumin as agreement between methods so poor) using standard methods and ^{1}H-NMR Spectroscopy (after transformation) The solid red line is the mean of the differences and the dashed red lines represent the 95% limits of agreement

cholesterol, Creatinine and Glucose have a small, but noticeable, difference. The agreement could have been affected by the difference in storage times mentioned above. Also a factor that could influence the observed disagreement in LDL cholesterol may be due to the fact that LDL cholesterol is not directly measured in the standard techniques, it is derived from 3 other measurements, whereas using NMR it is measured directly.

The lack of agreement is important when using a biomarker to make diagnosis of diseases where a threshold is used, however, when using a biomarker to identify an association with disease the differences are less relevant. In this scenario it is more relevant to ensure that both measurement methods have a similar association with the disease.

This contribution provides an indication of which metabolites can be reasonably quantified from a single sample measurement in a population of women aged over 50 and are minimally affected by measurement error. The sample size for estimation of the intraclass correlations is quite small resulting in imprecise estimates. If further contributions on this subject could be made with both genders, a wider age range and a range of sample storage times, it would allow a more precise picture of the biological variability of these biomarkers to be produced.

However, one issue that has not been covered in this chapter is that of measurement error due to batch effects. Unfortunately data as to when each sample was analysed using ^{1}H-NMR was not provided so an in depth analysis of batch effects was not possible. The closest proxy we had available to this was the town where the sample was obtained, as the samples from each town would have been collected in different weeks. This does not tell us anything about measurement error introduced during spectroscopy, but as the blood samples were collected on different weeks from each town it could be that different handling methods/times has introduced error. A principal component analysis was performed and a regression using Principal Component 1 as an outcome and town as a categorical exposure (Then repeated for PC2 as well), both of these result in a p-value <0.0001 for the association, suggesting strong evidence of an association between town and metabolite profile. The PCA score plots for each town are illustrated in figure 3.12. However, we would expect individuals' metabolite profiles to differ between towns, for example Harrogate is a very wealthy town and individuals from this town score highly on PC1 and PC2, whereas Falkirk is poorer and the scores are lower. We would fully expect variability between towns due to differing diets and behaviour, so this cannot necessarily be attributed to measurement error.

However, it was observed when using the Bland-Altman plots of triglycerides and glucose to check agreement that a number of observations were large outliers. If the observations that were outliers in triglycerides were the same as those which were outliers in glucose, it might suggest that this group were affected by measurement error. However, as can be seen in figures 3.13 and 3.14 it can be seen that those observations that are outliers in triglycerides are not outliers in glucose and vice versa. So it does not provide evidence to support the hypothesis that these outlying values are due to batch effects.

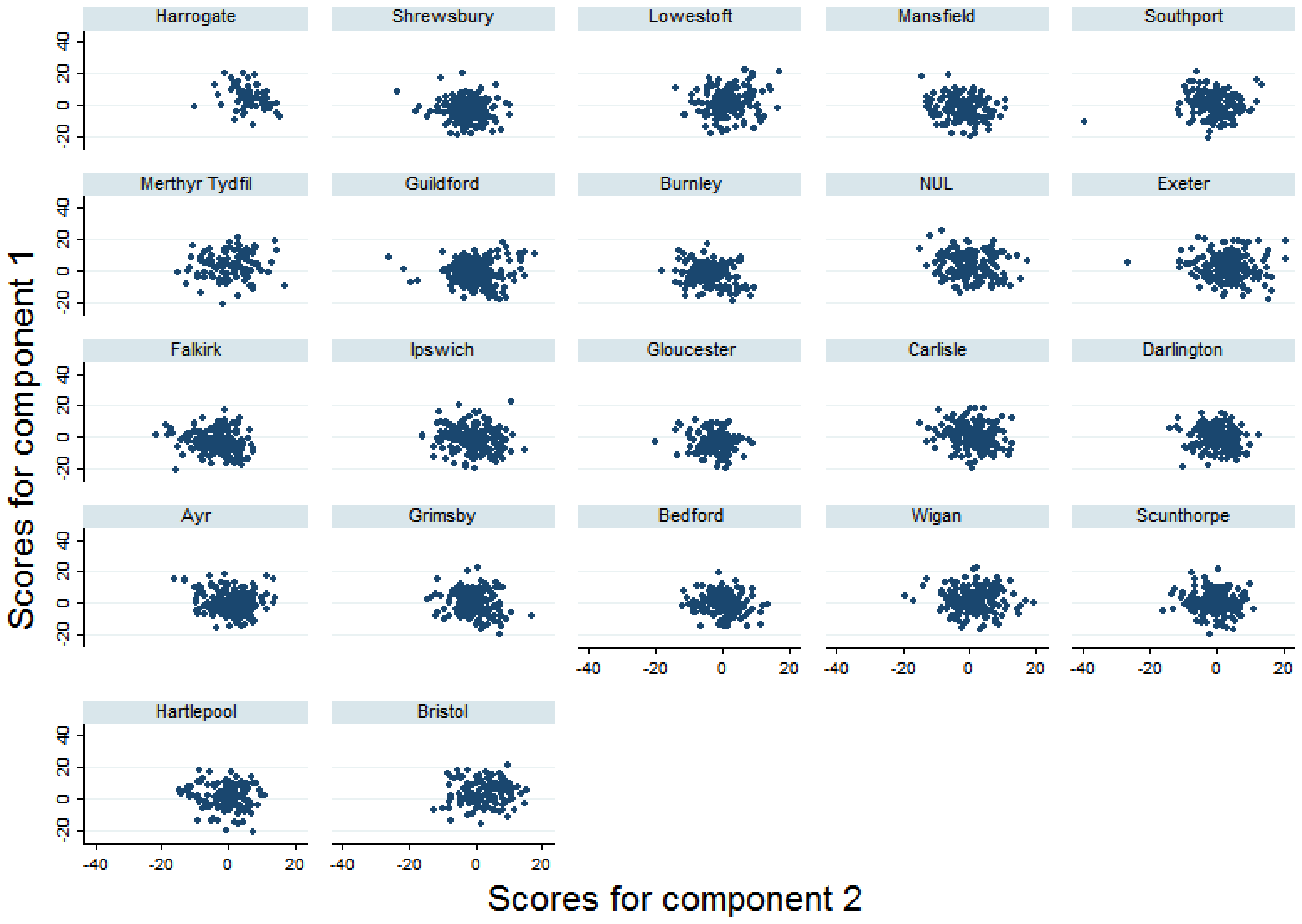


Figure 3.12: PCA score plots for each town in study

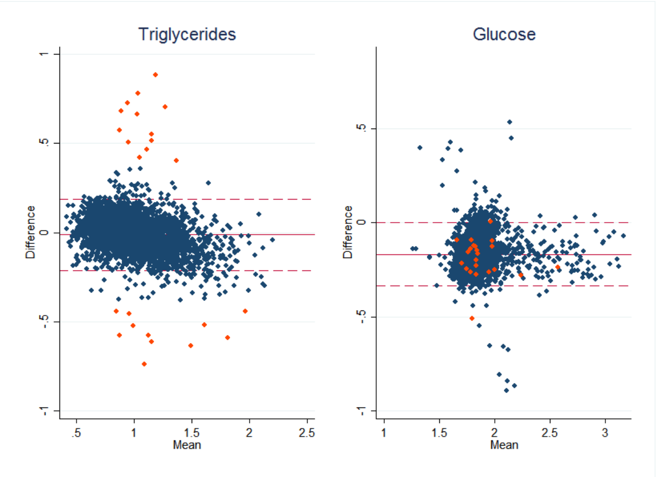


Figure 3.13: Bland-Altman plots of triglycerides and glucose comparing standard methods and ^{1}H-NMR Spectroscopy (after transformation) The solid red line is the mean of the differences and the dashed red lines represent the 95% limits of agreement. The 24 largest differences in triglycerides are highlighted orange in both plots.

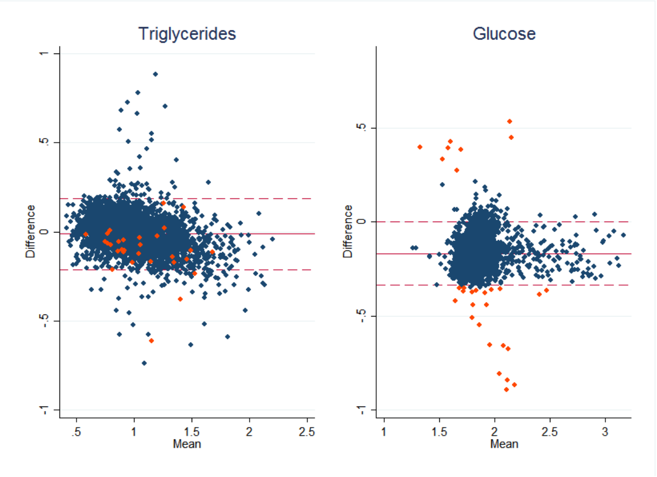


Figure 3.14: Bland-Altman plots of triglycerides and glucose comparing standard methods and ^{1}H-NMR Spectroscopy (after transformation) The solid red line is the mean of the differences and the dashed red lines represent the 95% limits of agreement. The 30 largest differences in glucose are highlighted orange in both plots.

3.5 Summary

In this chapter the metabolomic data that will be used in examples throughout this thesis has been introduced. It has been shown that there are very strong associations between many of the metabolite concentrations and the hierarchy of the metabolites has been described thoroughly. It is important to understand this hierarchy as it can inform decisions on variable inclusion/exclusion when perfoming analyses. It was identified that missing data was not a major problem in this dataset, although 2 metabolites, creatinine and glycerol, were missing in 8% and 2% of the sample - excluding these 2 variables meant that 99.7% of individuals had complete records.

Many of the metabolites were quite skewed, so suitable transformations were identified in the situation where an analysis method were to require a more normally distributed variable. It was also found that there were a number of zero values for some metabolites, so these may need to be handled carefully when performing analyses. The agreement between five metabolites measured using both ^{1}H-NMR Spectroscopy and standard methods was assessed finding excellent agreement in two metabolites and poorer agreement in the remaining three. Finally the reliability of the NMR metabolites were assessed using a subset of individuals who had repeat blood samples taken, with 89% of NMR biomarkers classed as having a good or excellent level of reliability and 97% of standard biomarkers.

The next chapter will focus on describing some of the methods that are more commonly applied to this type of metabolomic data, before applying them to the data described in this chapter. The data will then be revisited in chapters 7 and 8 where it will be used in an application of differential network analysis.

Table 3.5 - Summary of NMR biomarkers in the BWHHS sample

Type	Window	Variable	Description	Mean	SD	Min	Max	Skewness	Kurtosis	N Missing	N zeroes
METABOLITE CONCENTRATIONS	LIPO	xxl_vldl_p	Concentration of chylomicrons and XL VLDL particles (mol/l)	1.56E-10	1.60E-10	0.00E+00	1.77E-09	3.19	20.54	0	356
		xxl_vldl_l	Total lipids in chylomicrons and extremely large VLDL (mmol/l)	3.33E-02	3.46E-02	0.00E+00	3.83E-01	3.20	20.53	0	356
		xxl_vldl_pl	Phospholipids in chylomicrons and extremely large VLDL (mmol/l)	3.95E-03	4.36E-03	0.00E+00	4.65E-02	3.18	20.27	0	356
		xxl_vldl_c	Total cholesterol in chylomicrons and extremely large VLDL (mmol/l)	6.15E-03	6.67E-03	0.00E+00	7.83E-02	3.18	20.52	0	356
		xxl_vldl_ce	Cholesterol esters in chylomicrons and extremely large VLDL (mmol/l)	3.73E-03	3.92E-03	0.00E+00	4.67E-02	3.11	20.47	0	356
		xxl_vldl_fc	Free cholesterol in chylomicrons and extremely large VLDL (mmol/l)	2.42E-03	2.89E-03	0.00E+00	3.16E-02	3.22	20.34	0	356
		xxl_vldl_tg	Triglycerides in chylomicrons and extremely large VLDL (mmol/l)	2.32E-02	2.37E-02	0.00E+00	2.59E-01	3.21	20.84	0	356
		xl_vldl_p	Concentration of very large VLDL particles (mol/l)	8.00E-10	9.71E-10	0.00E+00	1.07E-08	3.06	18.64	0	585
		xl_vldl_l	Total lipids in very large VLDL (mmol/l)	7.84E-02	9.49E-02	0.00E+00	1.05E+00	3.08	18.78	0	585
		xl_vldl_pl	Phospholipids in very large VLDL (mmol/l)	1.39E-02	1.62E-02	0.00E+00	1.77E-01	3.02	18.54	0	585
		xl_vldl_c	Total cholesterol in very large VLDL (mmol/l)	1.57E-02	1.96E-02	0.00E+00	2.30E-01	3.25	20.85	0	585
		xl_vldl_ce	Cholesterol esters in very large VLDL (mmol/l)	8.03E-03	1.07E-02	0.00E+00	1.28E-01	3.34	21.65	0	585
		xl_vldl_fc	Free cholesterol in very large VLDL (mmol/l)	7.72E-03	9.05E-03	0.00E+00	1.02E-01	3.08	19.33	0	585
		xl_vldl_tg	Triglycerides in very large VLDL (mmol/l)	4.88E-02	5.96E-02	0.00E+00	6.44E-01	3.04	18.30	0	585
		l_vldl_p	Concentration of large VLDL particles (mol/l)	5.68E-09	5.57E-09	0.00E+00	5.85E-08	2.60	14.35	0	268
		l_vldl_l	Total lipids in large VLDL (mmol/l)	3.29E-01	3.24E-01	0.00E+00	3.41E+00	2.60	14.39	0	268
		l_vldl_pl	Phospholipids in large VLDL (mmol/l)	6.28E-02	5.91E-02	0.00E+00	6.16E-01	2.53	13.95	0	268
		l_vldl_c	Total cholesterol in large VLDL (mmol/l)	7.54E-02	7.67E-02	0.00E+00	8.33E-01	2.63	14.86	0	268
		l_vldl_ce	Cholesterol esters in large VLDL (mmol/l)	4.04E-02	3.96E-02	0.00E+00	4.33E-01	2.56	14.55	0	268
		l_vldl_fc	Free cholesterol in large VLDL (mmol/l)	3.50E-02	3.77E-02	0.00E+00	4.01E-01	2.72	15.43	0	268
		l_vldl_tg	Triglycerides in large VLDL (mmol/l)	1.91E-01	1.89E-01	0.00E+00	1.97E+00	2.62	14.46	0	268
		m_vldl_p	Concentration of medium VLDL particles (mol/l)	2.10E-08	1.46E-08	0.00E+00	1.44E-07	2.08	10.35	0	26
		m_vldl_l	Total lipids in medium VLDL (mmol/l)	7.08E-01	4.88E-01	0.00E+00	4.78E+00	2.06	10.20	0	26
		m_vldl_pl	Phospholipids in medium VLDL (mmol/l)	1.45E-01	9.40E-02	0.00E+00	9.06E-01	1.96	9.64	0	26
		m_vldl_c	Total cholesterol in medium VLDL (mmol/l)	2.09E-01	1.31E-01	0.00E+00	1.25E+00	1.88	9.34	0	26
		m_vldl_ce	Cholesterol esters in medium VLDL (mmol/l)	1.25E-01	7.15E-02	0.00E+00	7.02E-01	1.76	9.06	0	26
		m_vldl_fc	Free cholesterol in medium VLDL (mmol/l)	8.42E-02	6.14E-02	0.00E+00	5.82E-01	2.00	9.75	0	26
		m_vldl_tg	Triglycerides in medium VLDL (mmol/l)	3.53E-01	2.67E-01	0.00E+00	2.62E+00	2.16	10.80	0	26
		s_vldl_p	Concentration of small VLDL particles (mol/l)	3.97E-08	1.73E-08	0.00E+00	1.49E-07	1.15	5.30	0	4
		s_vldl_l	Total lipids in small VLDL (mmol/l)	7.86E-01	3.33E-01	0.00E+00	2.86E+00	1.09	5.07	0	4
		s_vldl_pl	Phospholipids in small VLDL (mmol/l)	1.80E-01	7.09E-02	0.00E+00	5.89E-01	0.87	4.47	0	4
		s_vldl_c	Total cholesterol in small VLDL (mmol/l)	3.08E-01	1.17E-01	0.00E+00	9.42E-01	0.79	4.21	0	4
		s_vldl_ce	Cholesterol esters in small VLDL (mmol/l)	1.97E-01	7.29E-02	0.00E+00	5.73E-01	0.75	4.19	0	4
		s_vldl_fc	Free cholesterol in small VLDL (mmol/l)	1.11E-01	4.70E-02	0.00E+00	3.69E-01	0.85	4.30	0	4
		s_vldl_tg	Triglycerides in small VLDL (mmol/l)	2.98E-01	1.59E-01	0.00E+00	1.33E+00	1.43	6.40	0	4
		xs_vldl_p	Concentration of very small VLDL particles (mol/l)	5.43E-08	1.51E-08	0.00E+00	1.29E-07	0.65	4.05	0	7
		xs_vldl_l	Total lipids in very small VLDL (mmol/l)	6.87E-01	1.91E-01	0.00E+00	1.68E+00	0.62	4.05	0	7
		xs_vldl_pl	Phospholipids in very small VLDL (mmol/l)	2.15E-01	6.05E-02	0.00E+00	5.59E-01	0.64	4.14	0	7
		xs_vldl_c	Total cholesterol in very small VLDL (mmol/l)	3.26E-01	9.37E-02	0.00E+00	8.76E-01	0.44	4.11	0	7
		xs_vldl_ce	Cholesterol esters in very small VLDL (mmol/l)	2.20E-01	6.23E-02	0.00E+00	6.02E-01	0.43	4.18	0	7
		xs_vldl_fc	Free cholesterol in very small VLDL (mmol/l)	1.06E-01	3.47E-02	0.00E+00	2.74E-01	0.34	3.90	0	7

xs_vldl_tg	Triglycerides in very small VLDL (mmol/l)	1.46E-01	5.56E-02	0.00E+00	4.42E-01	0.99	4.68	0	7
idl_p	Concentration of IDL particles (mol/l)	1.51E-07	3.89E-08	4.82E-08	3.97E-07	0.76	4.23	0	0
idl_l	Total lipids in IDL (mmol/l)	1.53E+00	3.99E-01	4.17E-01	4.08E+00	0.75	4.27	0	0
idl_pl	Phospholipids in IDL (mmol/l)	4.08E-01	1.03E-01	0.00E+00	1.06E+00	0.70	4.38	0	3
idl_c	Total cholesterol in IDL (mmol/l)	9.55E-01	2.67E-01	0.00E+00	2.70E+00	0.71	4.43	0	3
idl_ce	Cholesterol esters in IDL (mmol/l)	6.77E-01	1.89E-01	0.00E+00	1.91E+00	0.70	4.38	0	3
idl_fc	Free cholesterol in IDL (mmol/l)	2.78E-01	7.90E-02	0.00E+00	7.86E-01	0.67	4.44	0	3
idl_tg	Triglycerides in IDL (mmol/l)	1.63E-01	4.90E-02	0.00E+00	4.38E-01	1.05	4.82	0	3
l_ldl_p	Concentration of large LDL particles (mol/l)	2.58E-07	7.13E-08	0.00E+00	6.79E-07	0.77	4.13	0	1
l_ldl_l	Total lipids in large LDL (mmol/l)	1.83E+00	5.09E-01	0.00E+00	4.87E+00	0.76	4.15	0	1
l_ldl_pl	Phospholipids in large LDL (mmol/l)	4.38E-01	1.05E-01	0.00E+00	1.07E+00	0.73	4.17	0	1
l_ldl_c	Total cholesterol in large LDL (mmol/l)	1.25E+00	3.71E-01	0.00E+00	3.49E+00	0.74	4.18	0	1
l_ldl_ce	Cholesterol esters in large LDL (mmol/l)	9.09E-01	2.82E-01	0.00E+00	2.60E+00	0.75	4.13	0	1
l_ldl_fc	Free cholesterol in large LDL (mmol/l)	3.37E-01	9.01E-02	0.00E+00	8.97E-01	0.68	4.34	0	1
l_ldl_tg	Triglycerides in large LDL (mmol/l)	1.50E-01	4.60E-02	0.00E+00	4.09E-01	1.09	4.73	0	1
m_ldl_p	Concentration of medium LDL particles (mol/l)	2.13E-07	6.40E-08	0.00E+00	5.62E-07	0.79	4.07	0	1
m_ldl_l	Total lipids in medium LDL (mmol/l)	1.08E+00	3.22E-01	0.00E+00	2.85E+00	0.78	4.06	0	1
m_ldl_pl	Phospholipids in medium LDL (mmol/l)	2.70E-01	6.34E-02	0.00E+00	6.06E-01	0.75	4.00	0	1
m_ldl_c	Total cholesterol in medium LDL (mmol/l)	7.30E-01	2.39E-01	0.00E+00	2.08E+00	0.74	4.07	0	1
m_ldl_ce	Cholesterol esters in medium LDL (mmol/l)	5.44E-01	1.95E-01	0.00E+00	1.65E+00	0.74	4.08	0	1
m_ldl_fc	Free cholesterol in medium LDL (mmol/l)	1.86E-01	4.45E-02	0.00E+00	4.35E-01	0.73	4.12	0	1
m_ldl_tg	Triglycerides in medium LDL (mmol/l)	7.73E-02	2.83E-02	0.00E+00	2.31E-01	1.35	5.21	0	1
s_ldl_p	Concentration of small LDL particles (mol/l)	2.45E-07	7.25E-08	0.00E+00	6.21E-07	0.81	4.10	0	1
s_ldl_l	Total lipids in small LDL (mmol/l)	6.84E-01	2.02E-01	0.00E+00	1.74E+00	0.79	4.10	0	1
s_ldl_pl	Phospholipids in small LDL (mmol/l)	1.92E-01	4.71E-02	0.00E+00	4.16E-01	1.01	4.69	0	1
s_ldl_c	Total cholesterol in small LDL (mmol/l)	4.44E-01	1.45E-01	0.00E+00	1.25E+00	0.71	4.06	0	1
s_ldl_ce	Cholesterol esters in small LDL (mmol/l)	3.33E-01	1.18E-01	0.00E+00	9.99E-01	0.70	4.08	0	1
s_ldl_fc	Free cholesterol in small LDL (mmol/l)	1.12E-01	2.85E-02	0.00E+00	2.54E-01	0.76	4.10	0	1
s_ldl_tg	Triglycerides in small LDL (mmol/l)	4.82E-02	1.76E-02	0.00E+00	1.46E-01	1.15	4.70	0	1
xl_hdl_p	Concentration of very large HDL particles (mol/l)	5.13E-07	2.74E-07	0.00E+00	2.00E-06	0.78	4.41	0	167
xl_hdl_l	Total lipids in very large HDL (mmol/l)	5.19E-01	2.79E-01	0.00E+00	2.04E+00	0.77	4.38	0	167
xl_hdl_pl	Phospholipids in very large HDL (mmol/l)	2.59E-01	1.50E-01	0.00E+00	9.80E-01	0.73	3.97	0	167
xl_hdl_c	Total cholesterol in very large HDL (mmol/l)	2.43E-01	1.35E-01	0.00E+00	1.00E+00	0.68	4.09	0	167
xl_hdl_ce	Cholesterol esters in very large HDL (mmol/l)	1.77E-01	1.01E-01	0.00E+00	7.42E-01	0.66	3.85	0	167
xl_hdl_fc	Free cholesterol in very large HDL (mmol/l)	6.65E-02	3.84E-02	0.00E+00	2.62E-01	0.78	4.23	0	167
xl_hdl_tg	Triglycerides in very large HDL (mmol/l)	1.71E-02	9.55E-03	0.00E+00	8.04E-02	0.96	5.70	0	167
l_hdl_p	Concentration of large HDL particles (mol/l)	1.32E-06	6.80E-07	0.00E+00	4.29E-06	0.53	3.70	0	206
l_hdl_l	Total lipids in large HDL (mmol/l)	8.25E-01	4.34E-01	0.00E+00	2.74E+00	0.56	3.68	0	206
l_hdl_pl	Phospholipids in large HDL (mmol/l)	4.19E-01	1.98E-01	0.00E+00	1.28E+00	0.31	3.68	0	206
l_hdl_c	Total cholesterol in large HDL (mmol/l)	3.71E-01	2.28E-01	0.00E+00	1.40E+00	0.77	3.71	0	206
l_hdl_ce	Cholesterol esters in large HDL (mmol/l)	2.89E-01	1.74E-01	0.00E+00	1.07E+00	0.75	3.70	0	206
l_hdl_fc	Free cholesterol in large HDL (mmol/l)	8.19E-02	5.39E-02	0.00E+00	3.29E-01	0.81	3.71	0	206
l_hdl_tg	Triglycerides in large HDL (mmol/l)	3.51E-02	2.01E-02	0.00E+00	1.46E-01	0.70	4.22	0	206
m_hdl_p	Concentration of medium HDL particles (mol/l)	2.47E-06	5.70E-07	0.00E+00	5.20E-06	0.87	4.47	0	1
m_hdl_l	Total lipids in medium HDL (mmol/l)	1.04E+00	2.48E-01	0.00E+00	2.23E+00	0.85	4.42	0	1

	m_hdl_pl	Phospholipids in medium HDL (mmol/l)	4.90E-01	1.07E-01	0.00E+00	1.07E+00	0.83	4.49	0	1
	m_hdl_c	Total cholesterol in medium HDL (mmol/l)	4.93E-01	1.41E-01	0.00E+00	1.21E+00	0.78	4.34	0	1
	m_hdl_ce	Cholesterol esters in medium HDL (mmol/l)	3.94E-01	1.13E-01	0.00E+00	9.91E-01	0.79	4.41	0	1
	m_hdl_fc	Free cholesterol in medium HDL (mmol/l)	9.94E-02	2.93E-02	0.00E+00	2.42E-01	0.72	4.23	0	1
	m_hdl_tg	Triglycerides in medium HDL (mmol/l)	6.00E-02	2.06E-02	0.00E+00	1.95E-01	0.89	4.96	0	1
	s_hdl_p	Concentration of small HDL particles (mol/l)	5.58E-06	8.45E-07	0.00E+00	9.78E-06	0.67	5.24	0	2
	s_hdl_l	Total lipids in small HDL (mmol/l)	1.24E+00	1.90E-01	0.00E+00	2.17E+00	0.66	5.17	0	2
	s_hdl_pl	Phospholipids in small HDL (mmol/l)	6.25E-01	1.08E-01	0.00E+00	1.26E+00	0.71	4.98	0	2
	s_hdl_c	Total cholesterol in small HDL (mmol/l)	5.52E-01	1.02E-01	0.00E+00	9.76E-01	0.34	4.40	0	2
	s_hdl_ce	Cholesterol esters in small HDL (mmol/l)	4.35E-01	8.76E-02	0.00E+00	8.31E-01	0.26	4.38	0	2
	s_hdl_fc	Free cholesterol in small HDL (mmol/l)	1.17E-01	2.47E-02	0.00E+00	2.50E-01	0.26	4.50	0	2
	s_hdl_tg	Triglycerides in small HDL (mmol/l)	6.00E-02	2.09E-02	0.00E+00	1.86E-01	0.87	4.78	0	2
	serum_c	Serum total cholesterol (mmol/l)	5.98E+00	1.34E+00	2.17E+00	1.34E+01	0.73	4.08	0	0
	vldl_c	Total cholesterol in VLDL (mmol/l)	9.41E-01	3.86E-01	2.30E-02	3.59E+00	1.20	5.88	0	0
	remnant_c	Remnant cholesterol (non-HDL, non-LDL -cholesterol) (mmol/l)	1.90E+00	5.70E-01	4.40E-01	4.78E+00	0.77	4.01	0	0
	ldl_c	Total cholesterol in LDL (mmol/l)	2.42E+00	7.52E-01	2.89E-01	6.83E+00	0.74	4.12	0	0
	hdl_c	Total cholesterol in HDL (mmol/l)	1.67E+00	4.52E-01	4.57E-01	3.72E+00	0.58	3.54	0	0
	hdl2_c	Total cholesterol in HDL2 (mmol/l)	1.15E+00	4.21E-01	3.60E-02	3.03E+00	0.61	3.60	0	0
	hdl3_c	Total cholesterol in HDL3 (mmol/l)	5.20E-01	5.86E-02	1.89E-01	8.27E-01	-0.28	5.03	0	0
	serum_tg	Serum total triglycerides (mmol/l)	1.68E+00	8.45E-01	3.42E-01	7.98E+00	1.85	8.77	0	0
	vldl_tg	Triglycerides in VLDL (mmol/l)	1.07E+00	7.28E-01	1.10E-01	7.10E+00	2.12	10.55	0	0
	ldl_tg	Triglycerides in LDL (mmol/l)	2.76E-01	9.03E-02	9.72E-02	7.86E-01	1.17	4.79	0	0
	hdl_tg	Triglycerides in HDL (mmol/l)	1.75E-01	5.49E-02	4.88E-02	5.28E-01	1.00	5.09	0	0
	apoa1	Apolipoprotein A-I (g/l)	1.73E+00	2.75E-01	8.62E-01	3.12E+00	0.76	3.90	0	0
	apob	Apolipoprotein B (g/l)	1.15E+00	2.90E-01	4.91E-01	2.57E+00	0.82	3.95	0	0
	alb	Albumin (signal area)	9.92E-02	1.24E-02	5.86E-02	1.52E-01	1.10	3.86	0	0
	vldl_d	Mean diameter for VLDL particles (nm)	3.63E+01	1.38E+00	3.34E+01	4.26E+01	0.71	3.56	0	0
	ldl_d	Mean diameter for LDL particles (nm)	2.35E+01	1.57E-01	2.28E+01	2.44E+01	-0.54	4.63	0	0
	hdl_d	Mean diameter for HDL particles (nm)	9.98E+00	2.73E-01	9.26E+00	1.11E+01	0.43	3.15	0	0
LIPID	estc	Esterified cholesterol (mmol/l)	4.22E+00	9.74E-01	1.42E+00	9.58E+00	0.72	4.09	8	0
	freec	Free cholesterol (mmol/l)	1.77E+00	3.77E-01	7.47E-01	3.93E+00	0.73	4.04	8	0
	totpg	Total phosphoglycerides (mmol/l)	2.44E+00	5.15E-01	8.90E-01	5.09E+00	0.75	4.19	8	0
	pc	Phosphatidylcholine and other cholines (mmol/l)	2.49E+00	4.80E-01	1.04E+00	5.05E+00	0.82	4.31	8	0
	sm	Sphingomyelins (mmol/l)	6.05E-01	1.21E-01	2.66E-01	1.72E+00	1.02	6.50	8	0
	totcho	Total cholines (mmol/l)	2.95E+00	5.46E-01	1.34E+00	5.80E+00	0.80	4.17	8	0
	totfa	Total fatty acids (mmol/l)	1.34E+01	3.07E+00	4.58E+00	2.89E+01	0.98	4.63	8	0
	dha	22:6, docosahexaenoic acid (mmol/l)	3.02E-01	8.68E-02	0.00E+00	1.05E+00	1.26	6.98	8	1
	la	18:2, linoleic acid (mmol/l)	3.76E+00	8.26E-01	8.66E-01	7.73E+00	0.62	3.81	8	0
	faw3	Omega-3 fatty acids (mmol/l)	7.30E-01	2.15E-01	1.80E-01	2.21E+00	1.20	6.15	8	0
	faw6	Omega-6 fatty acids (mmol/l)	4.63E+00	9.67E-01	1.28E+00	8.82E+00	0.68	3.81	8	0
	pufa	Polyunsaturated fatty acids (mmol/l)	5.36E+00	1.10E+00	1.64E+00	1.02E+01	0.69	3.79	8	0
	mufa	Monounsaturated fatty acids; 16:1, 18:1 (mmol/l)	3.13E+00	1.14E+00	8.58E-01	1.01E+01	1.27	5.81	8	0
	sfa	Saturated fatty acids (mmol/l)	4.91E+00	1.22E+00	1.65E+00	1.26E+01	1.06	5.35	8	0
	unsat	Estimated degree of unsaturation	1.26E+00	8.49E-02	9.51E-01	1.66E+00	0.22	4.05	8	0
LMWM	glc	Glucose (mmol/l)	4.94E+00	1.40E+00	2.26E+00	2.20E+01	5.29	42.10	3	0

		lac	Lactate (mmol/l)	1.64E+00	5.66E-01	6.75E-01	6.35E+00	2.13	10.78	3	0
		pyr	Pyruvate (mmol/l)	9.77E-02	3.34E-02	3.36E-02	3.50E-01	1.38	6.15	5	0
		cit	Citrate (mmol/l)	1.21E-01	2.63E-02	3.86E-02	8.50E-01	6.00	157.98	3	0
		glol	Glycerol (mmol/l)	1.34E-01	4.72E-02	2.98E-02	3.94E-01	0.96	4.43	78	0
		ala	Alanine (mmol/l)	3.34E-01	5.27E-02	2.09E-01	6.11E-01	0.76	4.25	3	0
		gln	Glutamine (mmol/l)	4.94E-01	6.90E-02	2.11E-01	8.30E-01	0.40	3.68	4	0
		gly	Glycine (mmol/l)	3.04E-01	6.49E-02	1.63E-01	6.89E-01	1.04	4.57	3	0
		his	Histidine (mmol/l)	6.44E-02	1.04E-02	5.37E-03	1.05E-01	0.21	4.17	3	0
		ile	Isoleucine (mmol/l)	5.86E-02	1.75E-02	2.22E-02	1.79E-01	1.49	7.09	3	0
		leu	Leucine (mmol/l)	7.18E-02	1.62E-02	3.51E-02	1.84E-01	1.26	6.47	3	0
		val	Valine (mmol/l)	1.65E-01	3.38E-02	7.80E-02	3.56E-01	0.83	4.54	3	0
		phe	Phenylalanine (mmol/l)	8.39E-02	1.28E-02	5.03E-02	1.77E-01	0.77	4.48	3	0
		tyr	Tyrosine (mmol/l)	5.17E-02	1.18E-02	2.23E-02	1.34E-01	1.13	6.05	3	0
		ace	Acetate (mmol/l)	4.28E-02	1.56E-02	2.11E-02	4.97E-01	15.33	398.28	3	0
		acace	Acetoacetate (mmol/l)	6.66E-02	4.43E-02	0.00E+00	4.18E-01	1.63	7.32	3	1
		bohbut	3-hydroxybutyrate (mmol/l)	1.95E-01	1.02E-01	2.57E-02	1.28E+00	1.91	10.12	4	0
		crea	Creatinine (mmol/l)	6.08E-02	1.41E-02	2.81E-02	2.62E-01	2.03	19.99	308	0
		gp	Glycoprotein acetyls, mainly a1-acid glycoprotein (mmol/l)	1.60E+00	3.90E-01	8.54E-01	5.80E+00	1.69	10.05	3	0
METABOLITE RATIOS	LIPO	xxl_vldl_pl_pc	Phospholipids to total lipds ratio in chylomicrons and XL VLDL (%)	11.20	2.27	0.36	54.60	1.92	49.84	356	0
		xxl_vldl_c_pc	Total cholesterol to total lipids ratio in chylomicrons and XL VLDL (%)	18.00	4.59	2.18	81.10	0.89	15.72	356	0
		xxl_vldl_ce_pc	Cholesterol esters to total lipids ratio in chylomicrons and XL VLDL (%)	11.47	4.29	0.11	52.50	0.28	5.95	356	0
		xxl_vldl_fc_pc	Free cholesterol to total lipids ratio in chylomicrons and XL VLDL (%)	6.52	2.33	0.24	45.40	2.04	33.89	356	0
		xxl_vldl_tg_pc	Triglycerides to total lipids ratio in chylomicrons and XL VLDL (%)	70.82	5.16	10.70	96.10	-0.52	13.43	356	0
		xl_vldl_pl_pc	Phospholipids to total lipds ratio in very large VLDL (%)	18.67	4.16	0.19	74.20	2.76	25.47	585	0
		xl_vldl_c_pc	Total cholesterol to total lipids ratio in very large VLDL (%)	20.41	5.57	0.24	76.20	2.00	16.25	585	0
		xl_vldl_ce_pc	Cholesterol esters to total lipids ratio in very large VLDL (%)	9.87	4.14	0.12	49.30	1.33	13.07	585	0
		xl_vldl_fc_pc	Free cholesterol to total lipids ratio in very large VLDL (%)	10.54	3.70	0.11	59.20	2.80	23.59	585	0
		xl_vldl_tg_pc	Triglycerides to total lipids ratio in very large VLDL (%)	60.94	7.93	0.17	98.90	-1.83	12.64	585	0
		l_vldl_pl_pc	Phospholipids to total lipds ratio in large VLDL (%)	19.81	1.91	15.40	47.90	3.73	32.53	268	0
		l_vldl_c_pc	Total cholesterol to total lipids ratio in large VLDL (%)	22.29	4.24	2.63	59.30	-0.18	8.88	268	0
		l_vldl_ce_pc	Cholesterol esters to total lipids ratio in large VLDL (%)	12.54	3.75	0.18	38.70	0.17	5.75	268	0
		l_vldl_fc_pc	Free cholesterol to total lipids ratio in large VLDL (%)	9.76	2.15	0.89	31.70	-0.76	9.17	268	0
		l_vldl_tg_pc	Triglycerides to total lipids ratio in large VLDL (%)	57.91	4.44	6.34	79.40	-1.39	17.75	268	0
		m_vldl_pl_pc	Phospholipids to total lipds ratio in medium VLDL (%)	21.04	1.15	18.50	32.60	1.58	10.13	26	0
		m_vldl_c_pc	Total cholesterol to total lipids ratio in medium VLDL (%)	30.80	5.26	1.34	61.90	0.38	6.31	26	0
		m_vldl_ce_pc	Cholesterol esters to total lipids ratio in medium VLDL (%)	19.26	5.27	0.12	54.20	0.69	5.49	26	0
		m_vldl_fc_pc	Free cholesterol to total lipids ratio in medium VLDL (%)	11.54	1.20	1.10	20.20	-1.47	10.38	26	0
		m_vldl_tg_pc	Triglycerides to total lipids ratio in medium VLDL (%)	48.17	6.11	14.60	77.70	-0.60	5.85	26	0
		s_vldl_pl_pc	Phospholipids to total lipds ratio in small VLDL (%)	23.09	1.80	6.41	29.80	-1.04	7.61	4	0
		s_vldl_c_pc	Total cholesterol to total lipids ratio in small VLDL (%)	40.14	5.71	5.17	62.10	0.03	4.22	4	0
		s_vldl_ce_pc	Cholesterol esters to total lipids ratio in small VLDL (%)	26.02	5.26	1.66	58.30	0.33	4.17	4	0
		s_vldl_fc_pc	Free cholesterol to total lipids ratio in small VLDL (%)	14.12	1.36	1.72	18.20	-2.65	15.70	4	0
		s_vldl_tg_pc	Triglycerides to total lipids ratio in small VLDL (%)	36.79	6.42	14.70	77.60	0.31	4.11	4	0
		xs_vldl_pl_pc	Phospholipids to total lipds ratio in very small VLDL (%)	31.28	2.09	16.60	50.70	0.45	8.61	7	0

		xs_vldl_c_pc	Total cholesterol to total lipids ratio in very small VLDL (%)	47.46	4.70	4.89	60.00	-1.46	8.55	7	0
		xs_vldl_ce_pc	Cholesterol esters to total lipids ratio in very small VLDL (%)	32.21	4.16	3.00	45.30	-0.90	6.00	7	0
		xs_vldl_fc_pc	Free cholesterol to total lipids ratio in very small VLDL (%)	15.25	2.10	1.89	19.40	-2.65	10.90	7	0
		xs_vldl_tg_pc	Triglycerides to total lipids ratio in very small VLDL (%)	21.28	5.31	8.77	78.50	1.40	9.51	7	0
		idl_pl_pc	Phospholipids to total lipds ratio in IDL (%)	26.81	0.72	22.60	30.90	0.02	5.62	3	0
		idl_c_pc	Total cholesterol to total lipids ratio in IDL (%)	62.36	2.48	44.10	68.00	-1.11	5.51	3	0
		idl_ce_pc	Cholesterol esters to total lipids ratio in IDL (%)	44.24	2.13	30.90	52.80	-0.95	5.20	3	0
		idl_fc_pc	Free cholesterol to total lipids ratio in IDL (%)	18.12	1.15	4.32	20.70	-2.82	20.81	3	0
		idl_tg_pc	Triglycerides to total lipids ratio in IDL (%)	10.85	2.52	5.21	33.40	1.18	6.38	3	0
		l_ldl_pl_pc	Phospholipids to total lipds ratio in large LDL (%)	24.14	1.13	21.20	30.40	0.68	4.45	1	0
		l_ldl_c_pc	Total cholesterol to total lipids ratio in large LDL (%)	67.57	2.17	48.40	72.90	-1.46	8.04	1	0
		l_ldl_ce_pc	Cholesterol esters to total lipids ratio in large LDL (%)	49.10	2.18	35.60	55.00	-1.19	5.73	1	0
		l_ldl_fc_pc	Free cholesterol to total lipids ratio in large LDL (%)	18.47	1.10	2.34	21.40	-2.86	29.90	1	0
		l_ldl_tg_pc	Triglycerides to total lipids ratio in large LDL (%)	8.30	1.80	4.33	21.90	1.29	6.76	1	0
		m_ldl_pl_pc	Phospholipids to total lipds ratio in medium LDL (%)	25.60	2.34	19.00	49.40	1.94	12.93	1	0
		m_ldl_c_pc	Total cholesterol to total lipids ratio in medium LDL (%)	67.14	3.26	33.00	74.00	-2.18	14.39	1	0
		m_ldl_ce_pc	Cholesterol esters to total lipids ratio in medium LDL (%)	49.59	4.17	14.40	60.20	-1.85	10.29	1	0
		m_ldl_fc_pc	Free cholesterol to total lipids ratio in medium LDL (%)	17.56	1.41	13.30	25.80	0.80	4.73	1	0
		m_ldl_tg_pc	Triglycerides to total lipids ratio in medium LDL (%)	7.27	1.77	3.78	20.00	1.31	6.27	1	0
		s_ldl_pl_pc	Phospholipids to total lipds ratio in small LDL (%)	28.58	2.76	19.90	51.50	1.50	8.91	1	0
		s_ldl_c_pc	Total cholesterol to total lipids ratio in small LDL (%)	64.30	3.91	19.70	72.80	-2.38	17.48	1	0
		s_ldl_ce_pc	Cholesterol esters to total lipids ratio in small LDL (%)	47.77	4.56	2.65	60.10	-1.97	12.65	1	0
		s_ldl_fc_pc	Free cholesterol to total lipids ratio in small LDL (%)	16.54	1.14	12.40	22.50	0.76	4.94	1	0
		s_ldl_tg_pc	Triglycerides to total lipids ratio in small LDL (%)	7.13	1.90	3.60	29.70	2.26	17.24	1	0
		xl_hdl_pl_pc	Phospholipids to total lipds ratio in very large HDL (%)	49.22	9.49	1.17	87.60	-1.74	8.56	167	0
		xl_hdl_c_pc	Total cholesterol to total lipids ratio in very large HDL (%)	46.57	8.86	6.92	91.80	0.96	8.34	167	0
		xl_hdl_ce_pc	Cholesterol esters to total lipids ratio in very large HDL (%)	33.96	9.73	0.14	87.80	1.79	9.76	167	0
		xl_hdl_fc_pc	Free cholesterol to total lipids ratio in very large HDL (%)	12.61	2.24	1.13	26.90	-1.73	9.11	167	0
		xl_hdl_tg_pc	Triglycerides to total lipids ratio in very large HDL (%)	4.22	4.11	0.12	61.10	4.63	38.39	167	0
		l_hdl_pl_pc	Phospholipids to total lipds ratio in large HDL (%)	52.30	4.97	38.20	77.30	0.93	4.20	206	0
		l_hdl_c_pc	Total cholesterol to total lipids ratio in large HDL (%)	43.08	6.34	8.52	59.70	-1.01	4.35	206	0
		l_hdl_ce_pc	Cholesterol esters to total lipids ratio in large HDL (%)	33.80	4.52	6.49	48.00	-1.12	5.00	206	0
		l_hdl_fc_pc	Free cholesterol to total lipids ratio in large HDL (%)	9.28	2.09	1.23	14.20	-1.13	4.23	206	0
		l_hdl_tg_pc	Triglycerides to total lipids ratio in large HDL (%)	4.63	2.31	0.13	18.20	1.25	5.47	206	0
		m_hdl_pl_pc	Phospholipids to total lipds ratio in medium HDL (%)	47.22	1.85	39.50	69.50	0.38	10.64	1	0
		m_hdl_c_pc	Total cholesterol to total lipids ratio in medium HDL (%)	46.78	3.79	6.29	57.40	-1.19	8.65	1	0
		m_hdl_ce_pc	Cholesterol esters to total lipids ratio in medium HDL (%)	37.38	3.42	1.32	47.50	-0.92	7.69	1	0
		m_hdl_fc_pc	Free cholesterol to total lipids ratio in medium HDL (%)	9.40	0.77	2.47	11.10	-2.33	14.28	1	0
		m_hdl_tg_pc	Triglycerides to total lipids ratio in medium HDL (%)	6.01	2.49	0.85	25.20	1.75	9.48	1	0
		s_hdl_pl_pc	Phospholipids to total lipds ratio in small HDL (%)	50.55	4.07	29.70	73.10	0.20	4.32	2	0
		s_hdl_c_pc	Total cholesterol to total lipids ratio in small HDL (%)	44.59	4.39	15.20	64.60	-0.68	5.57	2	0
		s_hdl_ce_pc	Cholesterol esters to total lipids ratio in small HDL (%)	35.19	4.76	2.55	58.80	-0.52	5.53	2	0
		s_hdl_fc_pc	Free cholesterol to total lipids ratio in small HDL (%)	9.40	1.05	3.58	12.80	-1.23	5.93	2	0
		s_hdl_tg_pc	Triglycerides to total lipids ratio in small HDL (%)	4.87	1.68	0.66	19.10	1.40	8.11	2	0
		apob_apoa1	Ratio of apolipoprotein B to apolipoprotein A-I	0.68	0.17	0.29	1.41	0.56	3.40	0	0

	LIPID	tg_pg	Ratio of triglycerides to phosphoglycerides	0.52	0.24	0.04	2.66	1.66	8.44	8	0
		dha_fa	Ratio of 22:6 docosahexaenoic acid to total fatty acids (%)	2.28	0.54	0.00	9.33	1.44	13.19	8	1
		la_fa	Ratio of 18:2 linoleic acid to total fatty acids (%)	28.42	4.19	9.56	45.20	-0.06	3.43	8	0
		faw3_fa	Ratio of omega-3 fatty acids to total fatty acids (%)	5.53	1.42	1.47	19.60	1.37	9.05	8	0
		faw6_fa	Ratio of omega-6 fatty acids to total fatty acids (%)	34.93	4.34	14.20	51.40	-0.15	3.39	8	0
		pufa_fa	Ratio of polyunsaturated fatty acids to total fatty acids (%)	40.46	4.94	18.10	59.50	-0.15	3.43	8	0
		mufa_fa	Ratio of monounsaturated fatty acids to total fatty acids (%)	22.98	4.47	8.93	42.90	0.44	3.62	8	0
		sfa_fa	Ratio of saturated fatty acids to total fatty acids (%)	36.58	2.39	23.90	45.60	-0.18	3.71	8	0

Table 3.6 - Results from ICC Analysis

Type	Window	Variable	Description	Mean	SD	Min
METABOLITE CONCENTRATIONS	LIPO	xxl_vldl_p	Cube-root	0.67	0.66	0.49 (0.26 , 0.72)
		xxl_vldl_l	Cube-root	0.67	0.66	0.49 (0.26 , 0.72)
		xxl_vldl_pl	Cube-root	0.65	0.66	0.51 (0.28 , 0.73)
		xxl_vldl_c	Cube-root	0.66	0.72	0.54 (0.32 , 0.75)
		xxl_vldl_ce	Cube-root	0.66	0.74	0.56 (0.34 , 0.76)
		xxl_vldl_fc	Cube-root	0.63	0.68	0.54 (0.31 , 0.75)
		xxl_vldl_tg	Cube-root	0.68	0.64	0.47 (0.24 , 0.71)
		xl_vldl_p	Cube-root	0.51	0.87	0.74 (0.58 , 0.86)
		xl_vldl_l	Cube-root	0.52	0.87	0.74 (0.58 , 0.86)
		xl_vldl_pl	Cube-root	0.54	0.85	0.71 (0.54 , 0.85)
		xl_vldl_c	Cube-root	0.52	0.89	0.75 (0.59 , 0.87)
		xl_vldl_ce	Cube-root	0.50	0.91	0.77 (0.62 , 0.88)
		xl_vldl_fc	Cube-root	0.54	0.85	0.71 (0.54 , 0.85)
		xl_vldl_tg	Cube-root	0.51	0.87	0.75 (0.59 , 0.87)
		l_vldl_p	Cube-root	0.59	0.82	0.66 (0.46 , 0.82)
		l_vldl_l	Cube-root	0.59	0.82	0.66 (0.47 , 0.82)
		l_vldl_pl	Cube-root	0.61	0.81	0.64 (0.44 , 0.81)
		l_vldl_c	Cube-root	0.57	0.85	0.69 (0.50 , 0.83)
		l_vldl_ce	Cube-root	0.59	0.85	0.68 (0.49 , 0.83)
		l_vldl_fc	Cube-root	0.57	0.83	0.68 (0.49 , 0.83)
		l_vldl_tg	Cube-root	0.59	0.81	0.66 (0.46 , 0.82)
		m_vldl_p	Cube-root	0.46	0.84	0.77 (0.62 , 0.88)
		m_vldl_l	Cube-root	0.46	0.84	0.77 (0.61 , 0.88)
		m_vldl_pl	Cube-root	0.47	0.83	0.76 (0.61 , 0.87)
		m_vldl_c	Cube-root	0.48	0.81	0.74 (0.57 , 0.86)
		m_vldl_ce	Cube-root	0.47	0.79	0.74 (0.58 , 0.86)
		m_vldl_fc	Cube-root	0.50	0.85	0.75 (0.59 , 0.87)
		m_vldl_tg	Cube-root	0.45	0.85	0.78 (0.64 , 0.89)
		s_vldl_p	Cube-root	0.49	0.94	0.78 (0.64 , 0.89)
		s_vldl_l	Cube-root	0.50	0.93	0.78 (0.63 , 0.88)
		s_vldl_pl	Cube-root	0.48	0.91	0.78 (0.64 , 0.89)
		s_vldl_c	Cube-root	0.54	0.90	0.73 (0.57 , 0.86)
		s_vldl_ce	Cube-root	0.56	0.90	0.73 (0.56 , 0.85)
		s_vldl_fc	Cube-root	0.50	0.88	0.76 (0.60 , 0.87)
		s_vldl_tg	Cube-root	0.48	0.97	0.80 (0.67 , 0.90)
		xs_vldl_p	Log	0.53	0.92	0.75 (0.59 , 0.87)
		xs_vldl_l	Untransformed	0.61	1.08	0.76 (0.60 , 0.87)
		xs_vldl_pl	Untransformed	0.69	1.03	0.69 (0.51 , 0.84)
		xs_vldl_c	Untransformed	0.58	1.02	0.75 (0.60 , 0.87)
		xs_vldl_ce	Untransformed	0.58	1.00	0.75 (0.58 , 0.87)
		xs_vldl_fc	Untransformed	0.59	1.01	0.75 (0.59 , 0.87)
		xs_vldl_tg	Log	0.50	0.98	0.79 (0.65 , 0.89)
		idl_p	Log	0.69	0.95	0.66 (0.46 , 0.82)
		idl_l	Log	0.68	0.94	0.66 (0.46 , 0.82)
		idl_pl	Log	0.71	0.91	0.62 (0.42 , 0.80)
		idl_c	Log	0.65	0.94	0.68 (0.49 , 0.83)
		idl_ce	Log	0.65	0.95	0.68 (0.49 , 0.83)
		idl_fc	Untransformed	0.75	0.92	0.60 (0.39 , 0.78)
		idl_tg	Log	0.54	0.97	0.76 (0.61 , 0.87)
		l_ldl_p	Cube-root	0.71	0.88	0.61 (0.40 , 0.79)
		l_ldl_l	Cube-root	0.71	0.88	0.61 (0.4 0, 0.79)
		l_ldl_pl	Log	0.75	0.89	0.59 (0.37 , 0.78)
		l_ldl_c	Cube-root	0.70	0.89	0.62 (0.41 , 0.79)
		l_ldl_ce	Cube-root	0.70	0.89	0.62 (0.41 , 0.79)
		l_ldl_fc	Cube-root	0.69	0.86	0.61 (0.40 , 0.79)
		l_ldl_tg	Log	0.58	0.85	0.69 (0.50 , 0.83)
		m_ldl_p	Cube-root	0.69	0.87	0.62 (0.41 , 0.79)
		m_ldl_l	Cube-root	0.69	0.87	0.62 (0.41 , 0.79)
		m_ldl_pl	Log	0.74	0.89	0.59 (0.38 , 0.78)
		m_ldl_c	Cube-root	0.67	0.89	0.63 (0.43 , 0.80)
		m_ldl_ce	Cube-root	0.67	0.88	0.64 (0.44 , 0.81)
		m_ldl_fc	Log	0.69	0.93	0.64 (0.45 , 0.81)
		m_ldl_tg	Log	0.53	0.75	0.66 (0.47 , 0.82)
		s_ldl_p	Cube-root	0.64	0.87	0.65 (0.45 , 0.81)

		s_ldl_l	Cube-root	0.64	0.87	0.65 (0.46 , 0.81)
		s_ldl_pl	Cube-root	0.61	0.78	0.62 (0.42 , 0.80)
		s_ldl_c	Cube-root	0.63	0.89	0.67 (0.48 , 0.82)
		s_ldl_ce	Cube-root	0.64	0.88	0.66 (0.47 , 0.82)
		s_ldl_fc	Log	0.59	0.93	0.71 (0.54 , 0.85)
		s_ldl_tg	Log	0.57	0.80	0.67 (0.48 , 0.82)
		xl_hdl_p	Untransformed	0.43	0.86	0.80 (0.67 , 0.90)
		xl_hdl_l	Untransformed	0.43	0.86	0.80 (0.66 , 0.89)
		xl_hdl_pl	Untransformed	0.37	0.86	0.84 (0.73 , 0.92)
		xl_hdl_c	Untransformed	0.51	0.81	0.72 (0.54 , 0.85)
		xl_hdl_ce	Untransformed	0.52	0.76	0.68 (0.49 , 0.83)
		xl_hdl_fc	Untransformed	0.45	0.87	0.79 (0.64 , 0.89)
		xl_hdl_tg	Untransformed	0.67	0.61	0.45 (0.22 , 0.70)
		l_hdl_p	Untransformed	0.39	0.86	0.83 (0.71 , 0.91)
		l_hdl_l	Untransformed	0.37	0.87	0.84 (0.73 , 0.92)
		l_hdl_pl	Untransformed	0.45	0.84	0.78 (0.63 , 0.88)
		l_hdl_c	Untransformed	0.32	0.90	0.89 (0.81 , 0.94)
		l_hdl_ce	Untransformed	0.32	0.90	0.89 (0.81 , 0.94)
		l_hdl_fc	Untransformed	0.32	0.89	0.88 (0.80 , 0.94)
		l_hdl_tg	Untransformed	0.73	0.69	0.47 (0.24 , 0.71)
		m_hdl_p	Log	0.92	0.54	0.26 (0.06 , 0.61)
		m_hdl_l	Log	0.91	0.58	0.29 (0.08 , 0.62)
		m_hdl_pl	Log	0.92	0.52	0.24 (0.05 , 0.61)
		m_hdl_c	Cube-root	0.83	0.72	0.43 (0.20 , 0.69)
		m_hdl_ce	Cube-root	0.81	0.77	0.47 (0.24 , 0.71)
		m_hdl_fc	Cube-root	0.86	0.55	0.29 (0.08 , 0.62)
		m_hdl_tg	Cube-root	0.71	0.77	0.54 (0.32 , 0.75)
		s_hdl_p	Log	1.01	0.47	0.18 (0.02 , 0.61)
		s_hdl_l	Log	1.01	0.46	0.17 (0.02 , 0.61)
		s_hdl_pl	Log	0.82	0.75	0.46 (0.23 , 0.70)
		s_hdl_c	Untransformed	1.08	0.35	0.10 (0.00 , 0.72)
		s_hdl_ce	Untransformed	1.04	0.42	0.14 (0.01 , 0.63)
		s_hdl_fc	Untransformed	0.85	0.57	0.31 (0.10 , 0.63)
		s_hdl_tg	Cube-root	0.52	0.90	0.75 (0.59 , 0.87)
		serum_c	Log	0.74	0.87	0.58 (0.36 , 0.77)
		vldl_c	Cube-root	0.51	0.90	0.75 (0.60 , 0.87)
		remnant_c	Cube-root	0.62	0.96	0.70 (0.53 , 0.84)
		ldl_c	Cube-root	0.69	0.90	0.63 (0.43 , 0.80)
		hdl_c	Cube-root	0.51	0.86	0.74 (0.57 , 0.86)
		hdl2_c	Cube-root	0.47	0.87	0.77 (0.63 , 0.88)
		hdl3_c	Untransformed	0.64	0.67	0.52 (0.30 , 0.74)
		serum_tg	Log	0.54	0.95	0.76 (0.60 , 0.87)
		vldl_tg	Log	0.51	0.97	0.78 (0.64 , 0.88)
		ldl_tg	Log	0.57	0.82	0.68 (0.49 , 0.83)
		hdl_tg	Log	0.77	0.68	0.44 (0.21 , 0.69)
		apoa1	Log	0.79	0.69	0.43 (0.20 , 0.69)
		apob	Log	0.63	0.97	0.70 (0.52 , 0.84)
		alb	Log	1.23	0.34	0.07 (0.00 , 0.82)
		vldl_d	Log	0.47	0.84	0.76 (0.61 , 0.87)
		ldl_d	Cube	0.34	0.70	0.81 (0.68 , 0.90)
		hdl_d	Log	0.31	0.95	0.90 (0.83 , 0.95)
	LIPID	estc	Log	0.73	0.88	0.59 (0.38 , 0.78)
		freec	Log	0.77	0.84	0.55 (0.33 , 0.76)
		totpg	Log	0.81	0.55	0.31 (0.09 , 0.63)
		pc	Log	0.83	0.56	0.31 (0.09 , 0.63)
		sm	Log	0.80	0.73	0.45 (0.22 , 0.70)
		totcho	Log	0.85	0.57	0.31 (0.09 , 0.63)
		totfa	Log	0.80	0.65	0.40 (0.17 , 0.67)
		dha	Cube-root	0.73	0.71	0.49 (0.26 , 0.72)
		la	Cube-root	0.71	0.77	0.54 (0.32 , 0.75)
		faw3	Log	0.68	0.78	0.57 (0.35 , 0.77)
		faw6	Log	0.74	0.74	0.50 (0.27 , 0.73)
		pufa	Log	0.77	0.74	0.48 (0.25 , 0.72)
		mufa	Log	0.62	0.83	0.64 (0.44 , 0.81)
		sfa	Log	0.80	0.52	0.30 (0.08 , 0.62)
		unsat	Cube-root	0.40	0.75	0.78 (0.64 , 0.89)
	LMWM	glc	Log	0.58	0.54	0.47 (0.24 , 0.71)

		lac	Log	0.79	0.70	0.44 (0.21 , 0.70)
		pyr	Log	0.68	0.79	0.57 (0.36 , 0.77)
		cit	Log	0.68	1.00	0.68 (0.50 , 0.83)
		glol	Log	0.88	0.88	0.50 (0.27 , 0.73)
		ala	Log	0.86	0.70	0.40 (0.17 , 0.67)
		gln	Cube-root	0.79	0.85	0.53 (0.31 , 0.75)
		gly	Log	0.62	1.07	0.75 (0.59 , 0.87)
		his	Untransformed	0.95	0.00	N/A
		ile	Log	0.74	0.64	0.42 (0.19 , 0.68)
		leu	Log	0.84	0.58	0.32 (0.10 , 0.63)
		val	Log	0.75	0.73	0.48 (0.25 , 0.72)
		phe	Log	0.90	0.52	0.25 (0.05 , 0.61)
		tyr	Log	0.71	0.81	0.57 (0.35 , 0.77)
		ace	Log	0.63	0.61	0.48 (0.25 , 0.72)
		acace	Cube-root	0.70	0.51	0.35 (0.13 , 0.65)
		bohbut	Log	0.70	0.64	0.46 (0.23 , 0.70)
		crea	Log	0.48	0.83	0.75 (0.59 , 0.87)
		gp	Log	0.80	0.52	0.30 (0.09 , 0.63)
METABOLITE RATIOS	LIPO	xxl_vldl_pl_pc	Cube-root	0.52	0.50	0.49 (0.24 , 0.74)
		xxl_vldl_c_pc	Cube-root	0.45	0.62	0.65 (0.45 , 0.82)
		xxl_vldl_ce_pc	Untransformed	0.43	0.76	0.76 (0.59 , 0.88)
		xxl_vldl_fc_pc	Cube-root	0.71	0.67	0.48 (0.23 , 0.73)
		xxl_vldl_tg_pc	Untransformed	0.61	0.66	0.54 (0.31 , 0.76)
		xl_vldl_pl_pc	Cube-root	0.46	0.76	0.73 (0.53 , 0.87)
		xl_vldl_c_pc	Cube-root	0.52	0.61	0.58 (0.34 , 0.79)
		xl_vldl_ce_pc	Cube-root	0.78	0.60	0.37 (0.11 , 0.72)
		xl_vldl_fc_pc	Cube-root	0.38	0.71	0.78 (0.61 , 0.90)
		xl_vldl_tg_pc	Square	0.38	0.72	0.78 (0.62 , 0.89)
		l_vldl_pl_pc	Log	0.83	0.84	0.51 (0.27 , 0.75)
		l_vldl_c_pc	Untransformed	0.57	0.77	0.65 (0.43 , 0.82)
		l_vldl_ce_pc	Untransformed	0.82	0.65	0.38 (0.14 , 0.68)
		l_vldl_fc_pc	Untransformed	0.73	0.69	0.48 (0.22 , 0.74)
		l_vldl_tg_pc	Square	0.38	0.77	0.80 (0.65 , 0.90)
		m_vldl_pl_pc	Log	0.55	0.87	0.71 (0.54 , 0.85)
		m_vldl_c_pc	Untransformed	0.44	0.83	0.78 (0.64 , 0.88)
		m_vldl_ce_pc	Untransformed	0.44	0.93	0.81 (0.69 , 0.90)
		m_vldl_fc_pc	Square	0.74	0.61	0.40 (0.17 , 0.67)
		m_vldl_tg_pc	Square	0.40	0.82	0.81 (0.68 , 0.90)
		s_vldl_pl_pc	Cube	0.50	0.58	0.58 (0.36 , 0.77)
		s_vldl_c_pc	Untransformed	0.58	0.90	0.71 (0.53 , 0.84)
		s_vldl_ce_pc	Untransformed	0.59	0.95	0.72 (0.55 , 0.85)
		s_vldl_fc_pc	Cube	0.39	0.42	0.54 (0.31 , 0.75)
		s_vldl_tg_pc	Cube-root	0.58	0.91	0.71 (0.54 , 0.85)
		xs_vldl_pl_pc	Cube-root	0.67	0.27	0.14 (0.01 , 0.64)
		xs_vldl_c_pc	Cube	0.59	0.82	0.66 (0.47 , 0.82)
		xs_vldl_ce_pc	Square	0.64	0.82	0.62 (0.42 , 0.80)
		xs_vldl_fc_pc	Cube	0.39	0.47	0.59 (0.37 , 0.78)
		xs_vldl_tg_pc	Log	0.48	0.92	0.79 (0.64 , 0.89)
		idl_pl_pc	Untransformed	0.52	0.74	0.67 (0.48 , 0.82)
		idl_c_pc	Cube	0.41	0.89	0.82 (0.70 , 0.91)
		idl_ce_pc	Cube	0.48	0.83	0.75 (0.58 , 0.87)
		idl_fc_pc	Cube	0.47	0.74	0.71 (0.53 , 0.85)
		idl_tg_pc	Log	0.44	0.95	0.83 (0.70 , 0.91)
		l_ldl_pl_pc	Log	0.45	0.86	0.79 (0.65 , 0.89)
		l_ldl_c_pc	Cube	0.46	0.89	0.79 (0.65 , 0.89)
		l_ldl_ce_pc	Cube	0.52	0.90	0.75 (0.59 , 0.87)
		l_ldl_fc_pc	Cube	0.50	0.79	0.71 (0.54 , 0.85)
		l_ldl_tg_pc	Log	0.41	0.86	0.81 (0.68 , 0.90)
		m_ldl_pl_pc	Log	0.52	0.79	0.70 (0.52 , 0.84)
		m_ldl_c_pc	Cube	0.50	0.87	0.75 (0.59 , 0.87)
		m_ldl_ce_pc	Cube	0.57	0.87	0.70 (0.52 , 0.84)
		m_ldl_fc_pc	Log	0.63	0.77	0.59 (0.38 , 0.78)
		m_ldl_tg_pc	Log	0.40	0.82	0.81 (0.68 , 0.90)
		s_ldl_pl_pc	Log	0.61	0.84	0.66 (0.47 , 0.82)
		s_ldl_c_pc	Cube	0.51	0.85	0.73 (0.57 , 0.86)
		s_ldl_ce_pc	Cube	0.60	0.84	0.67 (0.48 , 0.82)
		s_ldl_fc_pc	Log	0.76	0.66	0.43 (0.20 , 0.69)

		s_ldl_tg_pc	Log	0.45	0.86	0.79 (0.64 , 0.89)
		xl_hdl_pl_pc	Square	0.51	0.56	0.55 (0.33 , 0.76)
		xl_hdl_c_pc	Cube-root	0.77	0.00	N/A
		xl_hdl_ce_pc	Cube-root	0.79	0.00	N/A
		xl_hdl_fc_pc	Square	0.59	0.56	0.48 (0.24 , 0.73)
		xl_hdl_tg_pc	Log	0.72	0.73	0.51 (0.27 , 0.74)
		l_hdl_pl_pc	Log	0.49	0.94	0.79 (0.64 , 0.89)
		l_hdl_c_pc	Cube	0.49	0.98	0.80 (0.66 , 0.90)
		l_hdl_ce_pc	Cube	0.50	0.99	0.80 (0.66 , 0.90)
		l_hdl_fc_pc	Cube	0.53	0.94	0.76 (0.59 , 0.87)
		l_hdl_tg_pc	Cube-root	0.60	0.96	0.72 (0.54 , 0.86)
		m_hdl_pl_pc	Log	0.56	0.77	0.65 (0.45 , 0.81)
		m_hdl_c_pc	Cube	0.49	0.99	0.81 (0.68 , 0.90)
		m_hdl_ce_pc	Cube	0.42	0.99	0.85 (0.74 , 0.92)
		m_hdl_fc_pc	Cube	0.76	0.65	0.42 (0.19 , 0.68)
		m_hdl_tg_pc	Log	0.51	1.04	0.81 (0.67 , 0.90)
		s_hdl_pl_pc	Cube-root	0.49	0.99	0.80 (0.67 , 0.90)
		s_hdl_c_pc	Square	0.57	0.82	0.68 (0.49 , 0.83)
		s_hdl_ce_pc	Square	0.56	0.88	0.71 (0.54 , 0.85)
		s_hdl_fc_pc	Cube	0.37	0.71	0.79 (0.64 , 0.89)
		s_hdl_tg_pc	Log	0.48	0.95	0.79 (0.66 , 0.89)
		apob_apoa1	Cube-root	0.34	1.10	0.91 (0.85 , 0.96)
	LIPID	tg_pg	Log	0.52	0.98	0.78 (0.63 , 0.88)
		dha_fa	Cube-root	0.39	0.83	0.82 (0.69 , 0.90)
		la_fa	Untransformed	0.48	0.86	0.76 (0.60 , 0.87)
		faw3_fa	Log	0.40	0.87	0.82 (0.70 , 0.91)
		faw6_fa	Untransformed	0.46	0.84	0.77 (0.62 , 0.88)
		pufa_fa	Untransformed	0.45	0.85	0.78 (0.64 , 0.89)
		mufa_fa	Cube-root	0.42	0.93	0.83 (0.71 , 0.91)
		sfa_fa	Square	0.43	0.80	0.78 (0.63 , 0.88)
STANDARD BIOMARKERS	N/A	White blood cell count	Log	0.53	0.71	0.64 (0.45 , 0.81)
		Haemoglobin	Cube	0.48	0.91	0.78 (0.63 , 0.88)
		Platelets	Cube-root	0.42	1.00	0.85 (0.74 , 0.92)
		Urea	Log	0.53	0.93	0.75 (0.59 , 0.87)
		Potassium	Cube-root	0.52	0.53	0.51 (0.29 , 0.74)
		Sodium	Square	0.52	0.58	0.55 (0.33 , 0.76)
		Creatinine	Log	0.26	0.92	0.93 (0.87 , 0.96)
		Urate	Log	0.27	0.95	0.93 (0.87 , 0.96)
		Magnesium	Untransformed	0.37	0.75	0.81 (0.67 , 0.90)
		Calcium	Log	0.40	0.75	0.78 (0.64 , 0.89)
		Phosphate	Cube-root	0.59	0.48	0.39 (0.16 , 0.67)
		Protein	Untransformed	0.67	0.64	0.48 (0.25 , 0.72)
		Albumin	Square	0.51	0.81	0.72 (0.53 , 0.86)
		Bilirubin	Log	0.62	0.86	0.66 (0.46 , 0.82)
		ALP level	Log	0.24	0.80	0.92 (0.85 , 0.96)
		Aspartate transaminase level	Log	0.53	0.85	0.72 (0.55 , 0.85)
		Alanine aminotransferase level	Log	0.52	0.78	0.69 (0.51 , 0.84)
		Gamma-glutamyl transpeptidase level	Log	0.18	0.82	0.95 (0.92 , 0.98)
		HDL cholesterol	Log	0.20	0.95	0.96 (0.92 , 0.98)
		LDL cholesterol	Cube-root	0.27	0.94	0.92 (0.86 , 0.96)
		Triglycerides	Log	0.46	0.86	0.78 (0.63 , 0.88)
		Glucose	Log	0.29	0.64	0.83 (0.71 , 0.92)
		Insulin	Log	0.57	0.61	0.53 (0.31 , 0.74)
		Fibrinogen clotting assay	Cube-root	0.66	0.73	0.55 (0.33 , 0.76)
		FVII	Untransformed	0.37	1.09	0.90 (0.82 , 0.95)
		FVIII	Untransformed	0.58	0.76	0.63 (0.43 , 0.80)
		FVIX	Untransformed	0.64	1.01	0.71 (0.53 , 0.85)
		Activated partial thromboplastin time	Log	0.33	0.47	0.66 (0.44 , 0.84)
		Activated partial thromboplastin time ratio	Cube-root	0.65	0.55	0.41 (0.15 , 0.72)
		Plasma viscosity	Log	0.53	0.70	0.64 (0.43 , 0.81)
		D-dimer	Log	0.56	0.70	0.61 (0.41 , 0.79)
		Tissue plasminogen activator	Cube-root	0.49	0.72	0.68 (0.49 , 0.83)
		Von Willebrand factor	Cube-root	0.38	0.78	0.81 (0.68 , 0.90)
		C-reactive protein	Log	0.47	0.82	0.75 (0.59 , 0.87)
		IL 6	Log	0.74	0.87	0.58 (0.36 , 0.78)
		Vitamin C	Untransformed	0.53	0.83	0.71 (0.52 , 0.85)
		Vitamin E	Cube-root	0.64	0.82	0.63 (0.40 , 0.82)

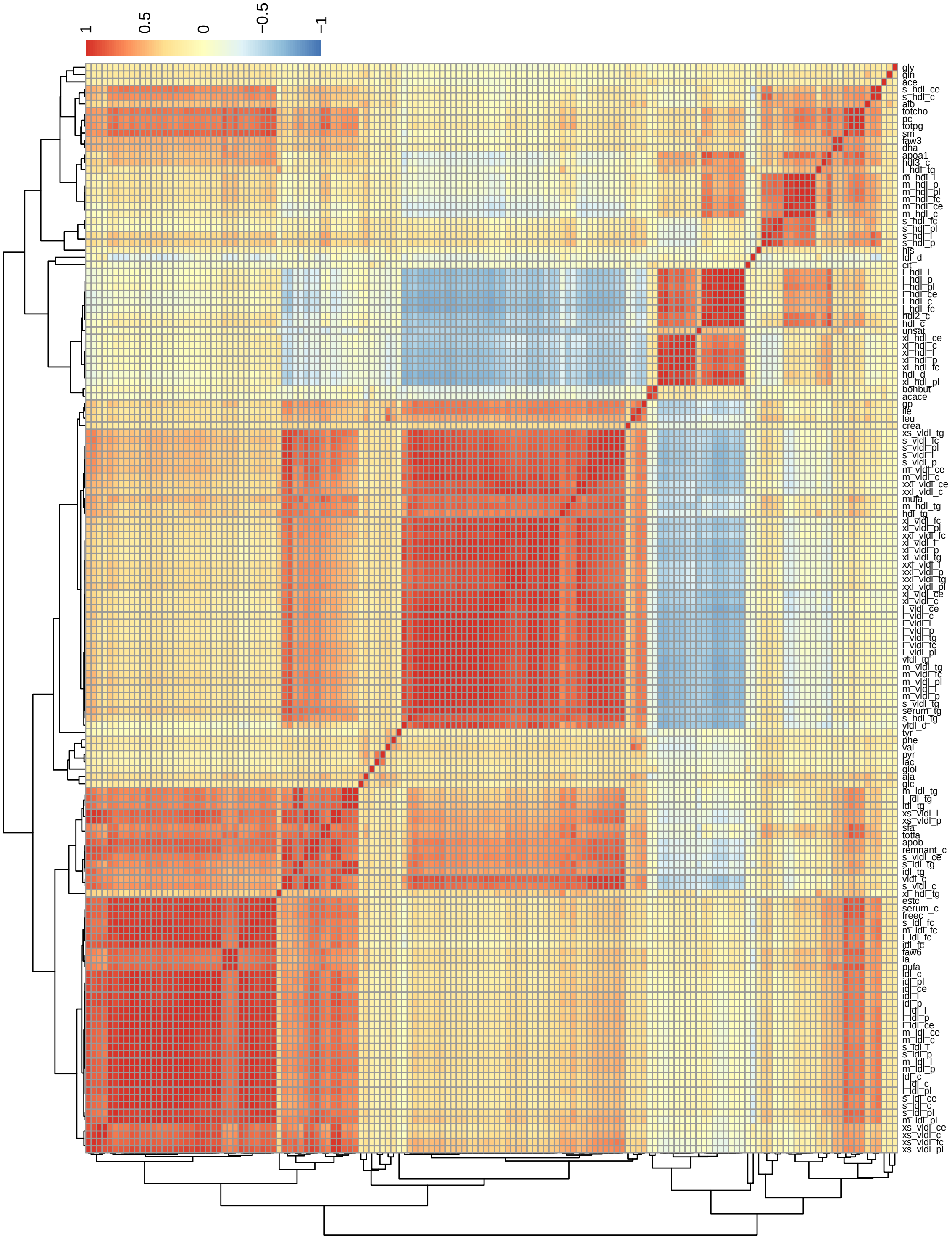


Figure 3.15: Heatmap showing the strength of Spearman correlations in the 149 metabolite concentrations. Red = positive correlation, blue = negative correlation

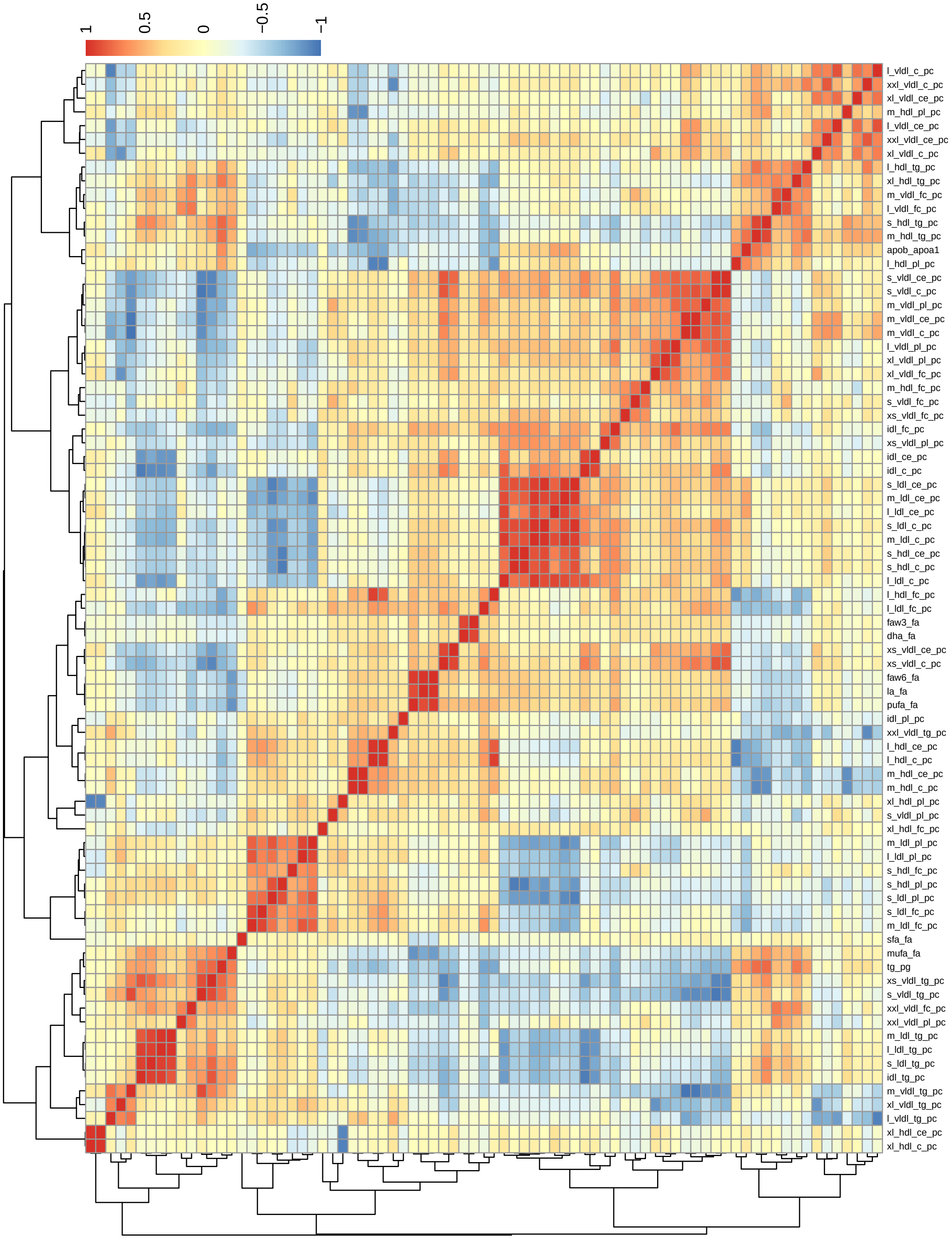


Figure 3.16: Heatmap showing the strength of Spearman correlations in the 79 metabolite ratios. Red = positive correlation, blue = negative correlation

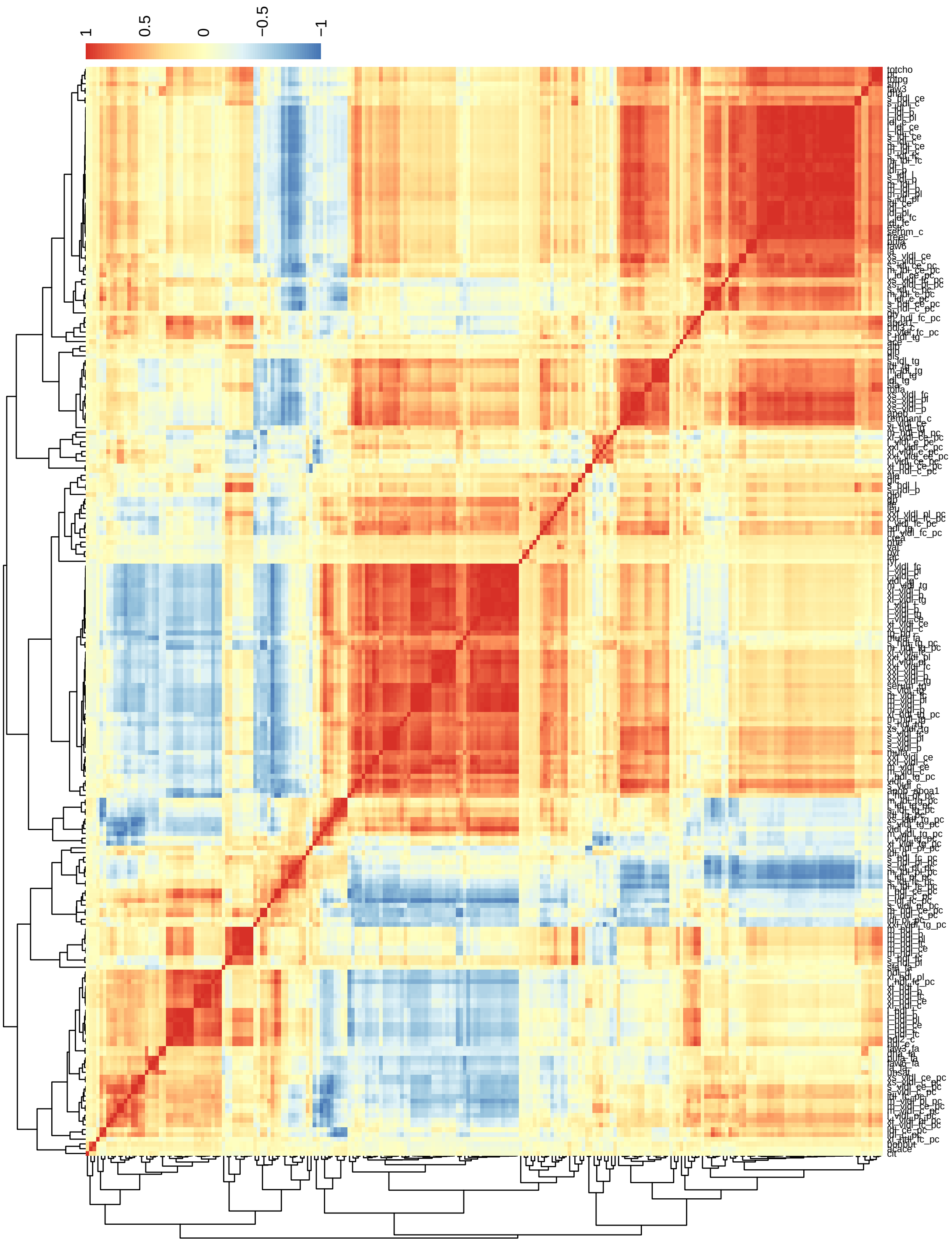


Figure 3.17: Heatmap showing the strength of Spearman correlations in the 228 NMR biomarkers. Red = positive correlation, blue = negative correlation

Chapter 4

Methods for high-dimensional exposures

4.1 Introduction

Before moving on to the use of network analysis methods which will form the core part of this thesis, we present a short review of some of the methods more commonly applied to high dimensional epidemiological data. For a more detailed source, Chadeau-Hyam et al [43] give a comprehensive overview of the statistical methods used for the analysis of -OMICS data and provide an excellent resource of references for this chapter. Similarly, Witten and Tibshirani [44] review statistical methods used in high dimensional situations, specifically applied to survival analysis.

One of the purposes of obtaining the concentrations of the NMR metabolites from the blood samples of the BWHHS cohort is to understand whether any of these metabolites are associated with rates of coronary heart disease. When analysing data with a large number of exposures we encounter a problem - multiple hypothesis testing [45]. In this chapter we describe this problem and how to account for it when drawing inferences. In addition to this, two alternative analysis methods commonly used for data of this type will be described - principal components regression and penalised regression. Each of these three approaches will be applied to the BWHHS dataset for illustration, using coronary heart disease by the end of the follow up period as the outcome of interest.

4.2 Overview of methods

4.2.1 Standard methods and multiple testing

The simplest approach to identifying which of the metabolites are associated with the outcome is to consider each metabolite individually. For each metabolite a linear regression (for a continuous outcome), a logistic regression (binary outcome) or a Cox proportional hazards model (time to event outcome) can be fitted to the data and a p-value for the association calculated. This p-value can be compared to a defined threshold (α) and if it is below that threshold then it can be said that evidence has been found to reject the null hypothesis that the metabolite is not associated with the outcome of interest. Or to say more simply, there is evidence of an association between the metabolite and the outcome of interest. However, when testing a high number of metabolites the typically chosen thresholds of $\alpha = 0.05$ or $\alpha = 0.01$ are often too lenient for determining "statistical significance" when multiple tests are viewed jointly.

For a single hypothesis test a p-value of 0.05 represents a 5% probability of rejecting the null hypothesis if the null hypothesis is true (i.e. there is no association between the metabolite and the outcome of interest). Incorrectly rejecting the null hypothesis is called a type I error or, alternatively, a false positive. If many tests are performed, and viewed jointly, the probability of obtaining a false positive increases, so the threshold at which we consider a p-value low enough to provide strong evidence of an association between predictor and outcome should be reduced. The two main methods of adjusting the threshold are to control the familywise error rate (FWER) or to control the false discovery rate (FDR). The FWER is the probability of at least 1 false positive occurring amongst all the tests that take place. The FDR is the expected proportion of false positives among all positive results obtained, which results in a less stringent threshold compared with FWER methods.

Table 4.1 illustrates a situation where m hypothesis tests have been performed and m_0 of the null hypotheses are true. Each cell contains the number of results observed for each possible scenario - U and S represent the ideal test results with U giving the number of tests where the null hypothesis has been correctly not rejected and S giving the number of tests where the null hypothesis has been correctly rejected. V represents the number of false positives (type I errors) and T the false negatives (type II errors). By reducing the threshold at which the null hypothesis is rejected the probability of false positives is reduced, but at the expense of potentially introducing some false negatives. From this table we can obtain the FWER and FDR - the FWER is the probability that V is greater than or equal to 1 and the FDR is the expected value of V/(V+S).

Table 4.1: Possible results from a set of m hypothesis tests [46]

	Null hypothesis not rejected	Null hypothesis rejected	
Null hypothesis is true	U	V	m_0
Null hypothesis is false	T	S	$m - m_0$
	m-R	R	m

4.2.2 Multiple testing adjustment

The most common method of adjusting the threshold is by a Bonferroni correction, which is achieved simply by dividing the original threshold (e.g. 0.05) by the number of tests performed [47]. Bonferroni is the most conservative threshold adjustment, it ensures that the FWER is less than or equal to 0.05. That is, of all the tests performed, there is at most a 5% chance that one or more of them results in a false positive, when the Bonferroni correction is applied.

When analysing data where the tests are not independent (i.e. the variables being tested are correlated) the Bonferroni adjustment is overly conservative. Alternative methods have been developed to take into account the dependence between tests. One of these methods was defined by Westfall and Young and is based on permutations [48]. It can be performed by randomly permuting the outcome variable z times, and following each permutation m hypothesis tests are performed and a p-value for each test is calculated. Under this scenario we know that there is no true association between the metabolites and the outcome, so if we were to set the threshold equal to the lowest p-value from the m tests we would say that this was the maximum threshold at which we detect no false positives. We then get the lowest p-value from each of the z permutations and we can identify the 5th percentile of this distribution and set this as our corrected threshold, i.e. in 5% of the z permutations performed a p-value was obtained that was lower than this threshold. If we use this as our corrected threshold we expect that 5% of the time we will obtain at least one false positive result among the m tests, i.e. our estimated FWER will now be 5%. If the tested hypotheses are strongly dependent it has been shown that use of this method gives a large gain in power over the Bonferroni correction and that the adjustment is asymptotically optimal as m increases [49].

The ultimate aim of performing a series of hypothesis tests is to correctly identify which variables are associated with the outcome of interest. As mentioned above, controlling the FWER reduces the probability of identifying a single false-positive amongst all tests performed. However, it may be the case that the aim of analysis is to identify a set of candidate variables from a much larger list and that finding some false positives is acceptable in order to be able to retain enough power to detect some true positives. In this case it may be preferable to adjust the FDR as opposed to the FWER.

Benjamini and Hochberg [46] defined the most commonly used method of controlling the FDR. The method is performed by ordering the p-values from the m hypothesis tests performed from the smallest to largest, p_1 to p_m. If we define α as the uncorrected threshold and k as the largest i for which

$$p_i \leq \frac{i}{m}\alpha$$

then we define α^* as

$$\alpha^* = \frac{k}{m}\alpha$$

and use this as our new threshold.

This method assumes that the tests are independent, and as mentioned previously this is often not the case. An updated method by Benjamini and Yekutieli [50] adjusts this method to allow for positive dependence among the tests performed.

4.2.3 Other methods

Alternative methods can be used in place of performing separate hypothesis tests for each of the possible exposures, adjusting for multiple testing. Methods for analysing high-dimensional data are usually split into two main categories [44]: supervised (where the data reduction takes place based on the outcome) and unsupervised (where the data reduction takes place before any consideration of the outcome is made) . In this thesis we are interested in analysing an outcome (the rate of CHD), so we will concentrate on supervised methods.

Within supervised methods there is a further distinction to be made between dimension reduction (or feature extraction) and variable selection. In dimension reduction methods the objective is to describe the high dimensional data using fewer dimensions, whereas with variable selection the objective is to choose a subset of the exposure variables. In this chapter we will focus on one method from each of these categories. An example of a variable selection method (lasso regression) is described below. One of the most commonly applied method of dimension reduction is principal component analysis (PCA), which is an unsupervised method. In this section we will describe its extension to a supervised method, principal component regression (PCR).

Table 4.2: Selected methods for analysing high-dimensional data

	Supervised	Unsupervised
Dimension reduction	Ridge regression Elastic net Principal component regression Partial least squares	Principal component analysis
Variable selection	Lasso Support vector machine	Factor analysis Hierarchical clustering

4.2.3.1 Principal Component Regression

Principal component regression consists of two steps, first a PCA of the explanatory variables is performed, followed by a regression on the outcome. The aim of PCA is to transform a set of correlated variables into a set of uncorrelated variables, called components. The new components are calculated so that the first component describes as much of the joint variance in the variables as possible, and subsequent components each describing gradually less and less of the variance. In doing this is it is often possible to describe most of the variation within a dataset using a smaller set of these components, reducing the dimensionality of the dataset and therefore simplifying any subsequent data analysis [51].

To carry out a PCA, the eigenvectors and eigenvalues of the covariance matrix of the variables of interest must be calculated [51]. The eigenvector that corresponds to each component defines the direction of that component with its terms called "loadings" capturing how the original values contribute to that component. The eigenvalue defines the amount of variation contained within that component and describes how much variability is explained by that component. By ordering the components from the one with the largest eigenvalue to the the one with the smallest, we derive how much of the overall is cumulatively due to them.

A decision that must be made in PCA is to choose how many components to select. Bartholomew et al. [52] suggest four possible criteria to be used when making this decision:

1. Keep enough components so that a large amount of the variation is retained (70-80%)
2. Retain components with an eigenvalue greater than 1
3. Examine a scree plot and retain components prior to the elbow of the plot
4. Identify if the components have any useful interpretation

In the second step, once the number of components to include has been chosen, these components can then be treated as any other variable would be in a multivariable regression analysis. One option is for the outcome of interest to be regressed upon all of the selected components, with those that are most strongly associated with the outcome selected using threshold for inclusion.

A difficulty with PCR is interpretation. When a component is found to be a risk or protective factor investigating its loadings to aid interpretation is rarely straightforward. So the method may be more useful for outcome prediction than for identifying which metabolites are most strongly associated with disease.

4.2.3.2 Lasso regression

Lasso regression is one of a group of methods known as penalised regression and belongs to the class of variable selection methods. The aim of variable selection is to select the subset of variables that are most strongly associated with the outcome of interest from a larger set of candidate variables. Standard forward and backwards stepwise selection includes or excludes variables based on their conditional association with the outcome, but the results can be unstable when there are a large number of predictor variables and particularly so if those variables are strongly correlated [53]. Ridge regression is a form of penalised regression, which penalises the estimated model coefficients by limiting the sum of the squared values of the regression coefficients [54], shrinking them towards zero. However, the original number of variables remain and so the model can be difficult to interpret. Lasso regression [55] combines the stability of ridge regression with the interpretability of subset selection because some of the coefficients are shrunk to exactly 0, giving a more parsimonious (and therefore interpretable) model. It does this by limiting the sum of the absolute (as opposed to the squared) values of the regression coefficients. For example, if the outcome variable is a continuous variable, a linear regression can be defined as:

$$E(y_i|x_{1i}, x_{2i}, ...x_{1K})) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} ... + \beta_K x_{iK} \tag{4.1}$$

Where K is the number of exposure variables, x_{i1}-x_{iK} are the observed exposure variables for individual i, β_1-β_K are the regression coefficients and y_i is the observed outcome variablefor individual i.
Ordinary least squares estimation of the parameters β_1-β_K is obtained by minimising the residual sum of squares:

$$\sum_{i=1}^{n}(y_i - \hat{y_i})^2 \tag{4.2}$$

Where n is the number of observations and $\hat{y_i}$ is the (model) predicted outcome variable for individual i (given the exposure variables) . However, the lasso adds another criterion to this:

$$\sum_{j=1}^{K}\left|\beta_j^s\right| \leq s \tag{4.3}$$

Where s is a tuning parameter and the β_j^s values are standardized. The smaller s is, the smaller the estimated β coefficients will be. If s is set very large, then the coefficient estimates will be the same as those estimated using the OLS method, and if s is set equal to zero then all β coefficients are set to 0. So selecting the value of s is key to how many parameters are estimated to be non-zero.

The value of s is often selected in a data driven manner and a common method of acquiring it is by cross-validation. The data are split into m subsamples, $m - 1$ of these subsamples are used to develop a lasso model each for a range of values of s. The resulting estimated parameters of each of these models are then used to predict the outcome on the unused subsample as a validation dataset and the mean square error corresponding to each value of s is calculated. This process is repeated m times, with each repetition using a different subsample as the validation set. The mean square errors are then averaged across the m folds to get the mean cross-validated error for each value of s. The value of s that provides the minimum mean cross-validated error can then be used as the selected s for the lasso regression.

4.3 Methods for BWHHS

We now apply the three methods described in the previous section to the BWHHS cohort data in order to highlight their main features. Each of the methods described are generalisable to any form of regression, however given the nature of our data (and the analysis presented in the next chapters) we use two different types of regression models: logistic regression and Cox (proportional hazard regression).

4.3.1 Logistic regression

Logistic regression is appropriate when the outcome of interest is binary in nature. In terms of our cohort data the outcome is defined as having suffered or not suffered a CHD event within the 12 year follow up period. Only women who either survived the whole 12 year follow up period or had a recorded CHD event at any time during the 12 years of follow-up are included in the analysis because of the bias that may be introduced by counting losses to follow-up as non-events (given the 12-year interval considered here). Those that died during follow up without having had a CHD event are excluded. This may introduce another source of bias, to deal with this the results will be interpreted as conditional on survival until the event or the end of the follow up period. Also, all observations are assumed to be independent.

The general form of the logistic regression model is

$$log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} ... + \beta_K x_{iK}$$

Where K is the number of exposure variables, π_i is the probability of an individual i having the outcome of interest (a CHD event within the 12 year follow up period), x_{i1}-x_{iK} are the observed exposure variables for individual i and β_1-β_K are the regression coefficients [56].Each coefficient can be interpreted as the expected increase in the log odds of the outcome of interest for a one unit increase of the exposure, holding constant all other variables in the model. Although the BWHHS is a cohort study and therefore a survival analysis is the most appropriate analytical approach, we also examine logistic regression here in order to draw some parallels with the differential network analysis studied in later chapters.

Method 1: Simple logistic regression, with adjustment for multiple testing A logistic regression model will be applied to the data using each of the 149 directly quantified metabolites in turn as our predictor variable. In addition to this, the model should be adjusted for potential confounders (excluding any other metabolites). In order to ensure that none of the potential confounders are on the causal pathway, only age will be adjusted for in this chapter. For each regression performed, the logistic model only including age as a predictor will be compared with a model including age, the metabolite and the metabolite squared (to allow for any non-linear effect of the metabolite) using a likelihood ratio test, with the null hypothesis being that the model with age alone describes the data as well as the model including the metabolite values.

The p-value for each hypothesis test is then compared to the Westfall & Young (controlling the family-wise error rate) and Benjamini & Yekutieli (controlling the false discovery rate) adjusted thresholds, as these both take into account the non-independence of the tests performed. The estimated odds ratios of the metabolites that result in a p-value below these thresholds will be reported.

Method 2: Principal components regression A principal components analysis will be performed on the BWHHS metabolite data, retaining all components where the eigenvalue is greater than 1. Then a logistic regression will be performed using the selected principal components as explanatory variables in the model with CHD events in the follow-up defining the outcome. Age will also be included in the model as an a priori confounder. The components that are most strongly associated with the outcome will be retained, using a threshold for inclusion of $p < 0.05$ (while age will forced to be retained in the model).

Method 3: Lasso regression Logistic regression with lasso will be performed using the all metabolites as predictor variables (along with age), with the level of penalization determined by cross-validation. The logistic regression model will include all the metabolites as well as all their pairwise interactions. The model with the interactions will be obtained using the GLINTERNET package from R. This package ensures that when an interaction effect is included in the model, the main effects are also included. However this package only allows the use of a logistic (or linear)regression, so when performing a Cox regression the pairwise interactions are obtained by multiplying each pair of metabolites together and using these products as normal in a lasso regression. This has the problem that it does not guarantee the inclusion of main effects when the interaction is selected. Therefore to allow some comparison between the logistic and Cox methods this method will also be done for the logistic lasso regression, providing two sets of results for the logistic analysis.

4.3.2 Cox regression

A better exploitation of the available data focusses on the time to CHD event as the outcome of interest, with individuals who died from other causes during follow-up, or were known to be alive and CHD free at the end of their follow-up, treated as censored events. Survival analysis can be implemented on these data with parametric and semi-parametric models as standard choices of analysis. In the following we will use Cox regression modelling of the hazard function, i.e. of the instantaneous conditional probability of the event.

The general form of the Cox regression model is

$$h_i(t) = h_{0i}(t)exp(\sum_{k=1}^{K} \beta_k x_{ki})$$

where $h_{0i}(t)$ is the baseline hazard for individual i, $h_i(t)$ is the time-varying hazard for individual i, K is the number of covariates and β_k is log hazard ratio corresponding to a unit increase in x_k controlling for the other variables in the model [57].
In a Cox regression model the baseline rate is allowed to change over time, however there is an assumption that the ratio of hazards between individuals with different exposure values is constant (proportional hazards assumption) and that individual observations are independent. For example the hazard of CHD may vary across the follow up period, but the ratio relating the hazard in smokers and the hazard in non-smokers is assumed to be constant across the period.

The other assumption required for Cox regression is that of non-informative censoring. That is, all individuals who are lost to follow up prior to the end of the follow up and have not had the event of interest are considered to have a similar chance of survival as those that remain in the study. If those who are lost to follow up are different in some way to those who remain in the study, the interpretation of the model's parameters needs to be refined. In this study, all the individuals do not have a CHD event and leave the study prior to the end of the follow up period are lost due to (non-CHD related) death thus it is unlikely that this assumption would hold. The estimated hazard ratios reported below should therefore be interpreted as being conditional on surviving other causes of death.

Method 1: Simple Cox regression, with adjustment for multiple testing The process for this analysis is exactly the same as for the analyses performed for the logistic regression, except a Cox regression model is used. The assumption of proportional hazards will be checked for each fitted model by performing a Schoenfeld test [58].

Methods 2 and 3: Principal components regression and lasso regression These will be performed in exactly the same manner as the analyses using logistic regressions, but using time to CHD event as the outcome in a Cox regression model instead of the logistic regression model. Also only one set of results will be provided for the Cox lasso model as the GLINTERNET package is not available for time to event data.

4.4 Data preparation

Variable selection For this, and subsequent chapters, we will concentrate on the use of the metabolite concentrations, rather than the ratio biomarkers, meaning 149 metabolites are eligible for inclusion in our analysis. For the univariable analysis, all 149 metabolites are analysed. However, as described in section 3.1.1, many of the 149 biomarkers are exact sums of others, which would cause collinearity problems in the PCA and lasso regressions, where all metabolites are analysed together. For example, within lipoprotein classes, total lipids are equal to the sum of cholesterol, triglycerides and phospholipids. So if the association between total lipids and cholesterol was estimated, adjusting for triglycerides and phospholipids, we would get a resulting partial correlation of 1, because if phospholipids and triglycerides are held constant, any rise in cholesterol would raise the total lipids by the exact same amount. Similarly, the partial correlation of cholesterol and triglycerides adjusting for phospholipids and total lipids would be equal to -1, because for any increase in cholesterol, there would have to be a corresponding decrease in triglycerides to keep the other variables constant.

So given total cholesterol is the sum of cholesterol esters and free cholesterol, its measurements will be excluded, along with the measurements of total lipids as discussed above. In section 3.1.1 there is a list of metabolites whose concentrations are exact sums of other metabolites. In each case the metabolite that is a combination of the others will be excluded, while retaining the lower level metabolites. This results in a total of 94 variables to be eligible for analysis.

Also, as both Creatinine and Glycerol have large numbers of missing values (292 and 73 respectively) these are also excluded from the PCA and lasso analyses, as any individuals with an incomplete record of data will be excluded, so including these variables will lead to a large reduction in the sample size. Finally, in the previous chapter, the intraclass correlation of each of the metabolites were estimated using the repeat samples and a number of metabolites were identified as having poor reliability. If there is little agreement between the concentrations of a metabolite in an individual measured 2 weeks apart, it is unlikely to be a reliable feature for indicating CHD risk in a 12 year follow up period, so a decision was made to exclude any metabolite that exhibited an ICC of less than 0.4 (classified as poor reliability), which led to a further 14 metabolites being excluded. So a final set of 78 metabolites were selected to be used in the PCA and lasso regressions. Since the simple analysis treats all the metabolites separately, the analysis can be performed on the metabolites with poor reliability, however the results should be interpreted bearing their reliability in mind.

Transformations As many of the metabolites are strongly right-skewed, the same transformations applied to the 149 metabolite concentrations for the reliability analysis (section 3.4.1) will be applied in the analyses in this chapter (i.e. where the best transformation from cube-root, square-root, log, square and cube transforms is selected for each metabolite, "best" meaning the transformation resulting in the distribution with the skewness closest to 0). For details see table 3.5.

Study size As per figure 2.1, all women with prevalent CHD at baseline (n=143) were excluded from the analyses in this chapter (and subsequent chapters), resulting in a sample size of 3634 women.

For the logistic regression analyses, only women who had a CHD event in the 12-year follow-up or survived until the end of follow-up were included, so only 2929 women were included in these analyses. For the Cox regression analyses all 3634 women who were CHD free at baseline were included. Censoring was defined as a non-CHD related death prior to the end of the 12-year follow-up period, or being event free at the end of follow-up.

Complete case analysis was carried out, leading in particular to the exclusion of nine (out of the 3634) women who had missing values in at least one metabolite. Hence in the Cox analyses the data were restricted to 3625 participants. Seven of these nine women with missing values were in the sample for the logistic regression, so the sample size for that was restricted to 2922.

The PCA and lasso regression analyses also require a complete set of observations, so for both these analyses the sample size was also 3625 in the Cox analyses and 2922 in the logistic analyses. The numbers involved with the separate simple analyses vary by the completeness of the observed data for each metabolite, with the number of observations missing for each metabolite available in table 3.5. Age was not affected by missingness.

4.5 Results

4.5.1 Simple analyses, with adjustment for multiple testing

Logistic regression The estimated odds ratios, each estimated separately, for the linear effects of the 149 metabolites ranged from 0.67 to 1.44. For the square term in each of the tests (included to allow or non-linear effects), the odds ratios ranged from 0.90 to 1.14 (which can not be interpreted independently of the odds ratios from the linear effects). Of the 149 hypothesis tests, testing the null hypothesis that including a linear and quadratic effect of a metabolite provided no improvement over a model only including age, 92 (62%) resulted in a p value less than the uncorrected threshold of 0.05. The high proportion of low p-values suggests that there is some association between the metabolome and the odds of CHD during the follow-up period. The distribution of p-values can be seen in figure 4.1.

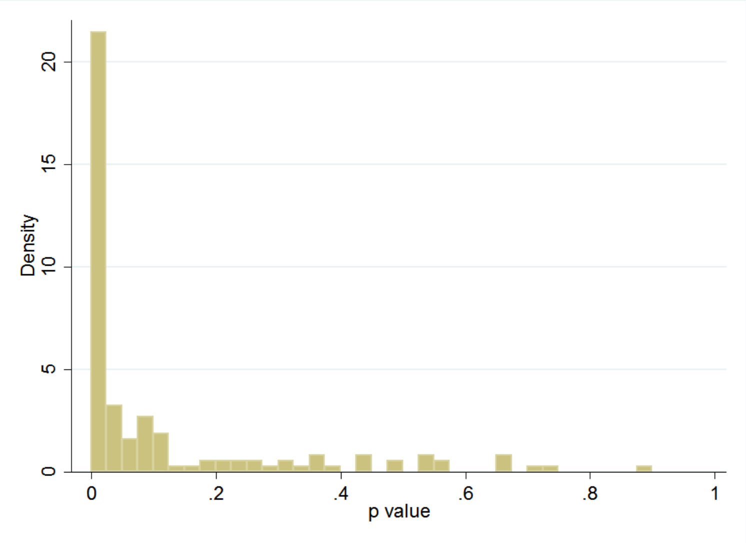


Figure 4.1: Distribution of p-values from simple logistic regression of 149 NMR metabolites

For the 149 tests we performed on the data the Bonferroni corrected threshold is 0.00033. There are 27 metabolites that have a p-value less than 0.00033, so one way of presenting these results would be to say that there is evidence of an association between each of these 27 metabolites and the odds of a CHD event in a 12-year follow-up period (after adjusting for age), significant at the Bonferroni adjusted threshold of $p < 0.00033$ and conditional on survival until the CHD event or the end of the follow up period. However, as this procedure does not take account of the dependence between tests it is likely to be an overly conservative adjustment. To that end a new threshold can be identified based on the Westfall and Young method which leads to a new corrected threshold of p <0.00095, leading to a further 11 metabolites being identified as having a significant association. So it can be said that there is evidence supporting the association of 38 metabolites to the odds of CHD in the follow up period, significant at the Westfall-Young corrected threshold of p <0.00095, which controls the family wise error rate to 5%. These metabolites are listed in table 4.3 along with their estimated odds ratios and the joint p-value for the likelihood ratio test of the inclusion of including both linear and quadratic terms (adjusted for age) compared with only age in the model.

From inspection of the results, it appears that the concentration of large HDL cholesterol metabolites are associated with a reduced odds of CHD in the follow up period with monounsaturated fatty acids, triglycerides and very large VLDL metabolites being associated with increased odds. If instead of the FWER, the FDR was to be controlled for, the Benjamini-Yekutieli corrected threshold is 0.0038, with 64 metabolites falling below this threshold.

Cox regression The estimated hazard ratios, each estimated separately, for the linear effects of the 149 metabolites ranged from 0.69 to 1.38. For the square term in each of the tests, the hazard ratios ranged from 0.92 to 1.12. Of the 149 hypothesis tests, testing the null hypothesis that including a linear and quadratic effect of a metabolite provided no improvement over a model only including age, 81 (54%) resulted in a p value less than the uncorrected threshold of 0.05. The high proportion of low p-values suggests that there is some association between the metabolome and the hazard of CHD. The distribution of p-values can be seen in figure 4.2.

As before the Bonferroni corrected threshold is 0.00033, with 15 metabolites having a p-value below this threshold. The Westfall and Young corrected threshold was this time found to be 0.00088, resulting in an additional 11 metabolites identified as significant. So there is evidence supporting the association of 26 metabolites to time to CHD event, significant at the Westfall-Young corrected threshold of 0.00088 (table 4.4). If instead of the FWER, the FDR was controlled for, the Benjamini-Yekutieli corrected threshold is 0.0028, with 47 metabolites falling below this threshold.

The Schoenfeld test, which tests the null hypothesis that the hazards are proportional, was performed after each Cox regression. No evidence was found to reject the null hypothesis for the metabolites that were found to be associated with the outcome so we can say that the assumption of proportional hazards is consistent with the results observed. There was also little evidence to suggest that any of the quadratic terms included in the models had any effect on the outcome. All the metabolites identified as associated with CHD hazard in the

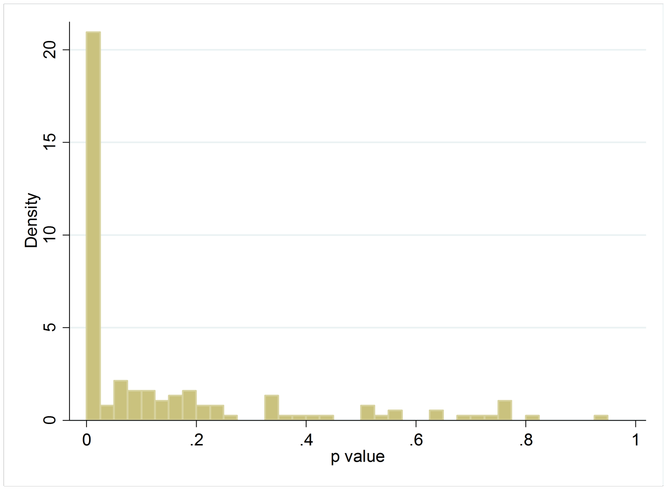


Figure 4.2: Distribution of p-values from simple Cox regression of 149 NMR metabolites

Cox regression analysis had previously been identified as being associated with odds of CHD in the logistic regression analysis.

Table 4.3: Metabolites associated with the odds of CHD in the follow up period, statistically significant at the Westfall-Young corrected threshold of p <0.00095, ordered by p-value of the joint test of significance of the main and squared terms in the model. Metabolites have been standardized, so odds ratios relate to a one standard deviation change in the metabolite concentration.

	Main term	Square term	Joint p-value
Variable	Odds Ratio (95% CI)	Odds Ratio (95% CI)	
Free cholesterol in large HDL	0.67 (0.56,0.79)	1.14 (1.03,1.26)	<0.0001
Total cholesterol in large HDL	0.67 (0.57,0.79)	1.13 (1.02,1.25)	<0.0001
Cholesterol esters in large HDL	0.68 (0.58,0.80)	1.12 (1.02,1.24)	<0.0001
Phospholipids in XL HDL	0.69 (0.58,0.80)	1.10 (1.01,1.21)	<0.0001
Monounsaturated fatty acids	1.36 (1.16,1.59)	1.05 (0.95,1.15)	<0.0001
Total lipids in large HDL	0.71 (0.61,0.83)	1.08 (0.98,1.19)	<0.0001
Creatinine	1.19 (1.02,1.37)	1.12 (1.04,1.20)	<0.0001
Concentration of large HDL	0.72 (0.62,0.84)	1.08 (0.98,1.18)	<0.0001
Total cholesterol in HDL2	0.72 (0.61,0.86)	1.01 (0.93,1.09)	<0.0001
Total cholesterol in HDL	0.73 (0.62,0.85)	1.03 (0.93,1.14)	0.0001
Mean diameter for HDL particles	0.73 (0.63,0.85)	1.08 (0.97,1.19)	0.0001
Serum total triglycerides	1.39 (1.17,1.65)	1.00 (0.90,1.12)	0.0001
Triglycerides in very small VLDL	1.41 (1.18,1.68)	0.99 (0.89,1.10)	0.0001
Free cholesterol in XL HDL	0.71 (0.60,0.83)	1.09 (1.00,1.19)	0.0001
Triglycerides in VLDL	1.37 (1.17,1.61)	1.02 (0.92,1.14)	0.0001
Triglycerides in small VLDL	1.39 (1.17,1.65)	0.99 (0.90,1.10)	0.0001
Triglycerides in medium VLDL	1.41 (1.18,1.68)	0.98 (0.88,1.08)	0.0002
Concentration of XL HDL particles	0.72 (0.62,0.84)	1.08 (0.99,1.17)	0.0002
Phospholipids in large HDL	0.74 (0.63,0.86)	1.04 (0.94,1.14)	0.0002
Triglycerides in IDL	1.37 (1.15,1.64)	0.99 (0.90,1.10)	0.0002
Total lipids in XL HDL	0.72 (0.62,0.84)	1.08 (0.99,1.17)	0.0002
Triglycerides in small LDL	1.27 (1.09,1.48)	1.08 (0.97,1.19)	0.0002
Triglycerides in large VLDL	1.40 (1.18,1.65)	1.01 (0.92,1.11)	0.0002
Estimated degree of unsaturation	0.79 (0.68,0.91)	1.06 (0.98,1.15)	0.0003
Concentration of medium VLDL	1.40 (1.17,1.68)	0.97 (0.88,1.08)	0.0003
Concentration of large VLDL	1.39 (1.18,1.64)	1.01 (0.92,1.11)	0.0003
Total lipids in large VLDL	1.39 (1.18,1.64)	1.01 (0.92,1.11)	0.0003
Bonferroni threshold			
Phospholipids in large VLDL	1.39 (1.18,1.64)	1.02 (0.93,1.12)	0.0003
Total lipids in medium VLDL	1.40 (1.17,1.67)	0.97 (0.88,1.08)	0.0004
Free cholesterol in medium VLDL	1.39 (1.17,1.66)	0.98 (0.88,1.08)	0.0004
Phospholipids in medium VLDL	1.39 (1.16,1.66)	0.98 (0.89,1.08)	0.0004
Triglycerides in XL VLDL	1.33 (1.14,1.55)	1.06 (0.96,1.18)	0.0005
Glycoprotein acetyls	1.44 (1.20,1.74)	0.92 (0.83,1.03)	0.0005
Free cholesterol in large VLDL	1.36 (1.16,1.60)	1.01 (0.92,1.12)	0.0005
Concentration of XL VLDL	1.32 (1.13,1.54)	1.06 (0.96,1.18)	0.0006
Triglycerides in large HDL	0.78 (0.67,0.91)	1.13 (1.05,1.21)	0.0006
Total lipids in XL VLDL	1.32 (1.13,1.54)	1.06 (0.96,1.18)	0.0006
Total cholesterol in large VLDL	1.36 (1.16,1.60)	1.01 (0.91,1.11)	0.0008

Table 4.4: Metabolites associated in survival analysis with CHD hazard, statistically significant at the Westfall-Young corrected threshold of p <0.00088, ordered by p-value of the joint test of significance of the main and squared terms in the model. Metabolites have been standardized, so odds ratios relate to a one standard deviation change in the metabolite concentration.

	Main term	Square term	Joint p-value
Variable	Hazard Ratio (95% CI)	Hazard Ratio (95% CI)	
Total cholesterol in large HDL	0.70 (0.60,0.82)	1.11 (1.01,1.23)	<0.0001
Free cholesterol in large HDL	0.69 (0.59,0.81)	1.12 (1.02,1.24)	<0.0001
Cholesterol esters in large HDL	0.70 (0.60,0.82)	1.11 (1.00,1.22)	<0.0001
Creatinine	1.20 (1.03,1.40)	1.05 (1.00,1.10)	0.0001
Monounsaturated fatty acids	1.32 (1.14,1.54)	1.04 (0.94,1.14)	0.0001
Phospholipids in very large HDL	0.72 (0.62,0.84)	1.08 (0.98,1.19)	0.0001
Total lipids in large HDL	0.74 (0.63,0.85)	1.07 (0.97,1.17)	0.0001
Concentration of large HDL	0.74 (0.64,0.86)	1.06 (0.97,1.17)	0.0001
Total cholesterol in HDL2	0.75 (0.63,0.88)	1.00 (0.93,1.08)	0.0002
Triglycerides in very small VLDL	1.38 (1.17,1.63)	1.00 (0.89,1.11)	0.0002
Serum total triglycerides	1.36 (1.15,1.61)	0.99 (0.89,1.10)	0.0002
Mean diameter for HDL	0.75 (0.65,0.87)	1.06 (0.96,1.18)	0.0003
Triglycerides in VLDL	1.34 (1.15,1.57)	1.01 (0.91,1.13)	0.0003
Triglycerides in small VLDL	1.37 (1.15,1.62)	0.99 (0.89,1.09)	0.0003
Total cholesterol in HDL	0.75 (0.64,0.88)	1.02 (0.92,1.12)	0.0003
Bonferroni threshold			
Triglycerides in medium VLDL	1.38 (1.16,1.65)	0.97 (0.88,1.07)	0.0004
Concentration of medium VLDL	1.38 (1.15,1.65)	0.97 (0.88,1.07)	0.0005
Triglycerides in large VLDL	1.36 (1.16,1.60)	1.01 (0.92,1.10)	0.0005
Triglycerides in IDL	1.34 (1.13,1.57)	1.00 (0.90,1.10)	0.0006
Total lipids in medium VLDL	1.38 (1.15,1.65)	0.97 (0.88,1.07)	0.0006
Phospholipids in large HDL	0.75 (0.65,0.88)	1.03 (0.93,1.13)	0.0006
Free cholesterol in medium VLDL	1.37 (1.15,1.63)	0.97 (0.88,1.08)	0.0006
Concentration of large VLDL	1.36 (1.16,1.59)	1.01 (0.92,1.10)	0.0007
Phospholipids in medium VLDL	1.37 (1.15,1.64)	0.98 (0.88,1.07)	0.0007
Phospholipids in large VLDL	1.36 (1.16,1.59)	1.01 (0.93,1.10)	0.0007
Total lipids in large VLDL	1.35 (1.15,1.59)	1.01 (0.92,1.10)	0.0007

4.5.2 Principal Components Regression

The PCA was performed on the 78 metabolites measured in 3625 women with complete records included in this analysis, generating 78 components. The first component explains 39.1% of the variation in the data, the second component explains a further 22.6%. So the first 2 components alone explain just under 62% of the variation in the dataset. Overall 12 terms had an eigenvalue greater than 1 (figure 4.3), accounting for just over 90% of the total variation. Following Bartholomew's second criterion, these were retained as predictor variables in the logistic and Cox regressions.

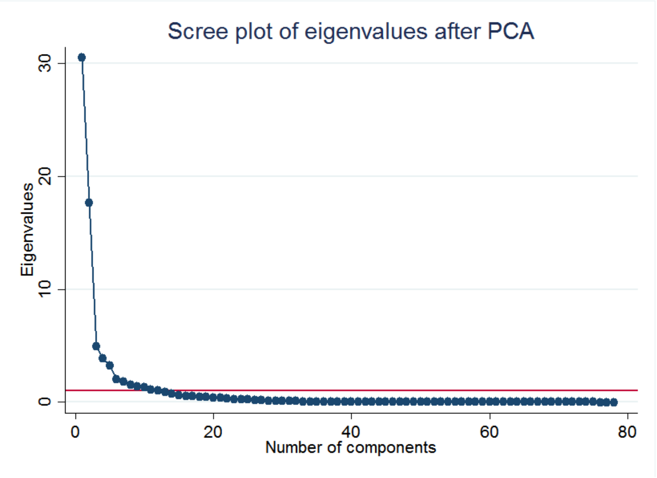


Figure 4.3: Scree plot of eigenvalues against each component

Logistic regression The selected components were used as predictor variables, alongside participants' age, in a multivariable logistic regression using CHD event as the outcome variable. Linear effects were assumed and no interactions were included. Those components with the strongest association with the odds of CHD by 12 years were retained in the model, using the threshold for inclusion of p <0.05. It should be noted that the components are independent of one another by definition, although are controlled for age. The results from the final selected model can be seen in table 4.5. There were 4 components that were retained in the final model, the first 3 components as well as component 7.

Controlling for age, component 1 was associated with increased odds of a CHD event (estimated odds ratio of 1.05; 95% CI 1.02,1.08) as was component 7 (estimated OR 1.15; 1.05,1.29). Components 2 and 3 were associated with reduced odds of a CHD event with ORs of 0.96 (0.92,0.99) and 0.92 (0.86,0.99) respectively. There was no strong evidence identified of a non-linear effect of any of

Table 4.5: Results from final multivariable logistic regression model, chosen by including any component with p <0.05 out of the 12 principal components, N=2922

Variable	Odds Ratio (95% CI)	p-value
Component 1	1.05 (1.02, 1.08)	<0.001
Component 2	0.96 (0.92, 0.99)	0.016
Component 3	0.92 (0.86, 0.99)	0.026
Component 7	1.15 (1.02, 1.29)	0.018

the components.

Components do not necessarily have a straightforward interpretation. However, the component loadings can be inspected to provide some insight as to what they represent. Figure 4.4 shows a chart of each of the 4 components loading profile, with each bar in the chart representing the loading attributed to one of the 78 metabolites considered in these analyses. There is no particular order of the metabolites, however they have been grouped by their window, and within the LIPO window the lipoproteins have been grouped by density.

Component 1 has consistently high positive loadings on the VLDL and LDL lipoproteins and was identified as a risk factor for CHD, adjusting for age. Component 2 has negative loadings of most of the VLDL components and positive loadings for the other metabolites from the LIPO and LIPID windows and was identified as a protective factor for CHD. This could possibly be interpreted as individuals who have higher concentrations of LDL/HDL/other lipoproteins and low VLDL metabolite concentrations have a reduced odds of CHD. Component 3 is also estimated to be a protective factor for CHD, and suggests that those who have high concentrations of HDL metabolites relative to the concentration of LDL metabolites are of lower risk of CHD. It is not at first clear from inspecting the profile of the loadings of component 7 what it represents, but in fact the individual large "spikes" seen in the LDL section of the chart, i.e. the large positive loadings, are all measures of LDL triglycerides concentration. The large green peaks are lactate, pyruvate and citrate. This suggests that individuals who have higher concentrations of LDL triglycerides as well as these amino acids, relative to the concentrations of all other lipoproteins, have increased odds of CHD.

Cox regression As would be expected, the results from the Cox regression are very similar to that from the logistic regression, but are presented here for completeness. The 12 components were again used as predictor variables, alongside participants' age, in a multivariable Cox regression on time to first CHD event, retaining the components with the strongest association with CHD hazard (using the threshold p <0.05). The results from the final selected model can

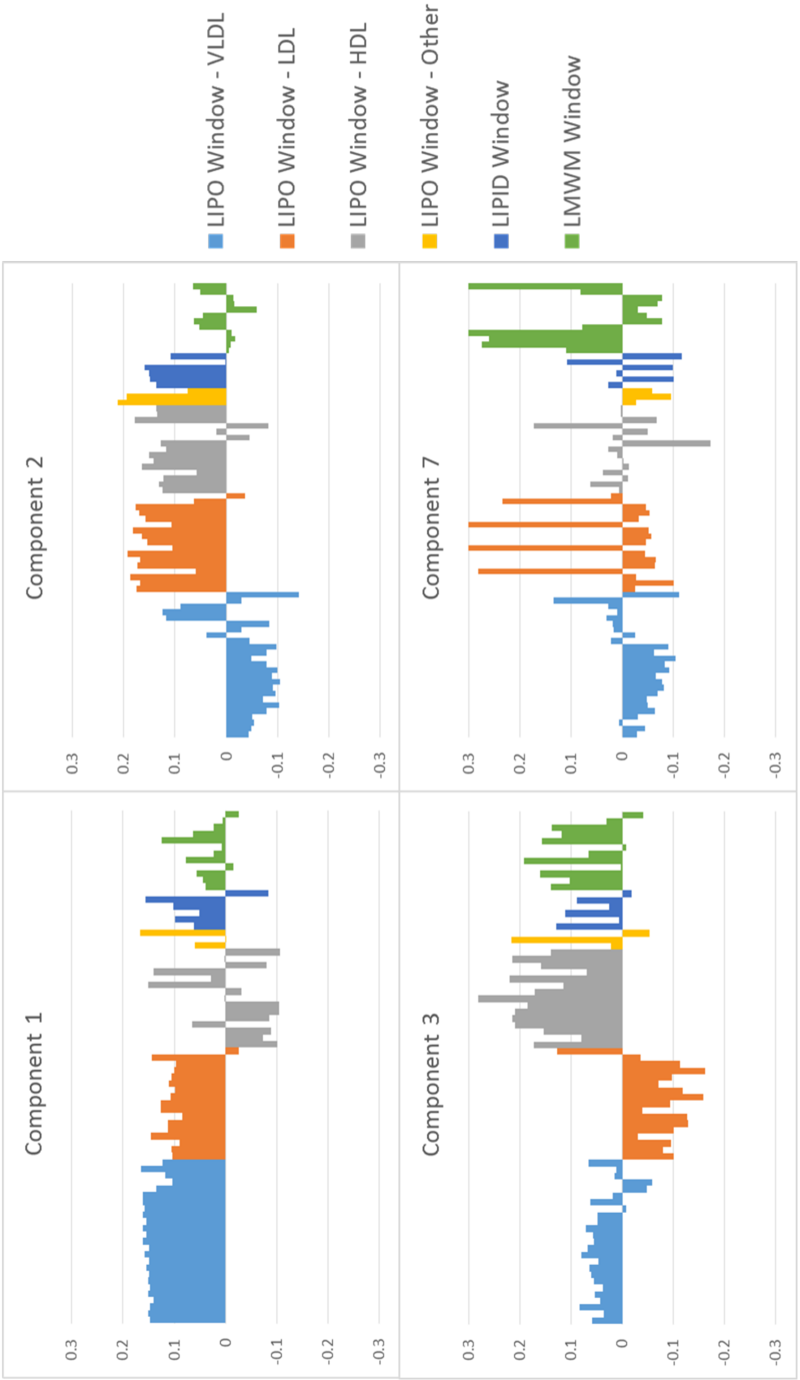


Figure 4.4: Component loadings for the first, second, third and seventh principal components from PCA of the metabolite dataset. Each bar represents the component loading attributable to a single metabolite, the metabolites are grouped by colour.

be seen in table 4.6. The same results as the logistic regression were obtained, although the evidence for the inclusion of component 7 in the model was slightly weaker (p=0.06) so did not quite meet the criteria for inclusion. However, it has been included in the table of results for comparison of the effect estimates with those from the logistic regression (including or excluding it has little effect on the other estimates). Again, no evidence was identified of a non-linear association between any of the components and the outcome.

Table 4.6: Results from final multivariable Cox regression model, chosen by including any component with p <0.05 out of the 12 principal components, N=3625

Variable	Hazard Ratio (95% CI)	p-value
Component 1	1.05 (1.02, 1.08)	<0.001
Component 2	0.96 (0.93, 1.00)	0.045
Component 3	0.92 (0.86, 0.99)	0.019
Component 7	1.11 (0.99, 1.25)	0.061

Principal component regression can be a useful tool in predicting the odds or hazard in individuals when a large number of predictor variables are available and most of the information stored within them can be summarised by a few components. The interpretation of the results of the regression are not necessarily straightforward though, so it may not be as informative about the aetiology of disease as some of the other methods described in this chapter. We made some speculative assessments about what each of the components of the regression may represent, and the components identified did seem to have some sensible interpretations, however that is not always the case. In the simple analysis we were able to assess how each predictor variable was associated with the outcome, however the results were open to confounding by the other metabolites. In the PC regression we were able to use all the provided data to perform a single logistic or Cox regression, however the interpretation of the predictors is not straightforward.

4.5.3 Lasso Regression

Logistic regression A logistic regression with lasso was performed using the 78 metabolites (and age), resulting in 15 metabolites where the estimated regression coefficient was estimated as non-zero. These are listed in table 4.7 along with the estimated penalised odds ratios for each, ordered by the direction of effect and effect size. Age was also estimated as a non-zero coefficient (1.07) but is excluded from the table of results.

There are some overlaps with the results from the simple analysis, with monounsaturated fatty acids and VLDL triglycerides being identified as risk factors for

Table 4.7: Metabolites with non-zero effect estimates from lasso logistic regression, ordered by direction of effect and effect size, where 78 metabolites and age were used as candidate predictor variables

Variable	Odds Ratio
Risk factors	
18:2, linoleic acid	1.37
Free cholesterol in IDL	1.24
Triglycerides in large VLDL	1.12
Monounsaturated fatty acids; 16:1, 18:1	1.09
Lactate	1.04
Triglycerides in medium VLDL	1.04
Protective factors	
Apolipoprotein A-I	0.82
Cholesterol esters in medium HDL	0.83
Cholesterol esters in XXL VLDL	0.84
Triglycerides in small LDL	0.87
Triglycerides in large HDL	0.87
Glycine	0.91
Glutamine	0.93
Pyruvate	0.93
Alanine	0.95

CHD. Interestingly, in the simple analysis all triglycerides were identified as risk factors for CHD by 12 years, however, in the lasso multivariable model, small LDL and large HDL triglycerides have been identified as potentially protective factors. This may be a chance finding, or may illustrate that after adjusting for all other metabolites, it is in fact the VLDL triglycerides that are associated with the increased odds of a CHD event in the follow up period i.e. amongst people with similar levels of triglycerides, those whose triglycerides are mostly contained within VLDL particles are at greater risk of CHD than those whose triglycerides are mostly found in small HDL and large HDL.

Following this, the analysis was repeated, this time including all interaction terms (which may be useful when drawing parallels to the differential network method in later chapters). Using the GLINTERNET package in R, 24 metabolites were included as main effects and 13 interactions were identified 4.9. Of the metabolites identified as main effects three were risk factors for CHD, monounsaturated fatty acids, glucose and IDL triglycerides, and HDL cholesterol 2 was identified as a protective factor (estimated OR 0.85). The other main effects had estimated ORs very close to 1 (in the range 0.99-1.01).

The table 4.10 is also included to provide a paralell to the Cox regression, rather than using the GLINTERNET package, the interaction terms were created manually by calculating the product of each pair of metabolites. Only 3 main effects were identified and 20 interaction terms (11 not included previously, 9 matching). Four interaction terms were identified using GLINTERNET but not here. Of the three metabolites identified as main effects two were risk factors for CHD, monounsaturated fatty acids and IDL triglycerides, and HDL cholesterol 2 was identified as a protective factor (estimated OR 0.85). These were also previously picked up as associated with the outcome in the simple analysis. Note that although all the interaction terms have no associated main effect reported, in fact the main effects are included in the model, but they have an estimated coefficient equal exactly to 0 (an OR of 1).

Cox regression A Cox regression with lasso was performed using the 78 selected metabolites (and age) as predictor variables resulting in 12 metabolites with non-zero coefficients. The selected predictors are listed in table 4.8 along with the estimated hazard ratios for each.

Table 4.8: Metabolites with non-zero effect estimates from lasso Cox regression, ordered by direction of effect and effect size, where 78 metabolites and age were used as candidate predictor variables

Variable	Hazard Ratio
Risk factors	
18:2, linoleic acid	1.17
Free cholesterol in IDL	1.16
Monounsaturated fatty acids; 16:1, 18:1	1.02
Protective factors	
Triglycerides in small LDL	0.91
Apolipoprotein A-I	0.91
Cholesterol esters in medium HDL	0.92
Glycine	0.93
Glutamine	0.93
Phospholipids in small HDL	0.95
Pyruvate	0.98
Tyrosine	0.98
Triglycerides in large HDL	0.99

As might be expected these results are very close to the results from the logistic regression with lasso, with lineolic acid and free cholesterol in IDL being identified as the strongest risk factors for time to CHD in the model where only main effects were included. Similar results were also obtained when all pairwise interactions were included in the model as well (table 4.11).

4.6 Summary

In this chapter, three different methods of analysis were adopted to illustrate typical approaches to the analysis of large metabolomic datasets. Each method was applied using both a logistic regression model and a Cox model, although the latter model is more appropriate for analysis of the BWHHS data, we have used both to provide a reference for the differential network analysis that is the topic of later chapters. Overall, and reassuringly, the results from the two approaches were similar. Adopting the first method we identified a number of metabolites that had evidence of an association with either the odds of a CHD event by 12 years (among the survivors) or time to CHD event. However, as this was a minimally adjusted analysis (only adjusting for age) these results may include spurious associations between metabolites and the outcome either due to correlation with the other metabolites or other CHD risk factors, such as BMI. We have not controlled for such health factors, however, because of concerns they may lie on the causal pathway between metabolites and the outcome.
In these analyses we have only considered age as a source of confounding, BMI was not included as a confounder due to concerns it may lie on the causal pathway between metabolites to the outcome.

An alternative approach we have adopted consisted of a 2-stage analysis where first PCA was applied to the metabolomic data, then the components from this PCA were used as predictors in regression models. This used all the information provided by the measured metabolites in the model, however the model was difficult to interpret although it may be more useful for predicting risk.

Finally one-stage multivariable analyses were performed, allowing all metabolites to be included in the model, with the lasso method used to select a subset of the metabolites as being associated with time to CHD event (Cox regression) or the odds of a CHD event in the follow up period (logistic regression).

Each of these methods are based on the same set of information (although the simple multiple hypothesis testing was performed on a greater number of metabolites). In the logistic regression analyses, there were three metabolites identified as risk factors in both the simple and lasso analyses; triglycerides in large VLDL, triglycerides in medium VLDL and monounsaturated fatty acids, the two triglyceride metabolites were also important contributors to the components selected in the PCR. There was one metabolite identified as a protective factor in the simple and lasso analyses - triglycerides in large HDL. When the Cox model was used only monounsaturated fatty acids was identified as associated with the outcome in both the simple and lasso analyses, again as a risk factor.

Table 4.9: Results from logistic regression with lasso, including all pairwise interactions using the GLINTERNET R package ensuring main effects are included when a metabolite is involved in a selected interaction.

Variable Main terms		Odds Ratio
Triglycerides in IDL		1.08
Total cholesterol in HDL2		0.88
Monounsaturated fatty acids; 16:1, 18:1		1.07
Acetate		0.99
Glucose		1.12
Cholesterol esters in XXL VLDL		1.00
Free cholesterol in small LDL		1.00
Citrate		1.00
Triglycerides in large VLDL		1.00
Glycine		1.00
Cholesterol esters in medium LDL		1.00
Free cholesterol in medium LDL		1.00
Triglycerides in very large HDL		1.00
Cholesterol esters in small LDL		1.00
Triglycerides in small LDL		1.00
Valine		1.00
Phospholipids in very large HDL		0.97
Tyrosine		0.97
Estimated degree of unsaturation		0.99
Cholesterol esters in very large HDL		1.00
Alanine		1.00
Pyruvate		1.00
18:2 lineolic acid		0.99
Mean diameter for VLDL particles		1.01
Interaction Terms		
Cholesterol esters in XXL VLDL	Free cholesterol in small LDL	1.01
Cholesterol esters in XXL VLDL	Citrate	1.00
Triglycerides in large VLDL	Glycine	1.00
Cholesterol esters in medium LDL	Monounsaturated fatty acids; 16:1, 18:1	0.99
Free cholesterol in medium LDL	Triglycerides in very large HDL	1.01
Cholesterol esters in small LDL	Monounsaturated fatty acids; 16:1, 18:1	1.01
Triglycerides in small LDL	Valine	1.00
Phospholipids in very large HDL	Tyrosine	1.04
Phospholipids in very large HDL	Estimated degree of unsaturation	1.01
Cholesterol esters in very large HDL	Alanine	1.01
Total cholesterol in HDL2	Pyruvate	1.00
18:2 lineolic acid	Valine	0.99
Tyrosine	Mean diameter for VLDL particles	0.98

Table 4.10: Results from logistic regression with lasso, including all pairwise interactions. Table included as it uses the same method as in the Cox lasso regression (main effects not necessarily included if interaction term is)

Variable Main terms		Odds Ratio
Triglycerides in IDL		1.03
Total cholesterol in HDL2		0.85
Monounsaturated fatty acids; 16:1, 18:1		1.10
Interaction Terms		
Cholesterol esters in XXL VLDL	Free cholesterol in small LDL	1.06
Cholesterol esters in XXL VLDL	Citrate	0.95
Triglycerides in XXL VLDL	Mean diameter for LDL particles	0.98
Free cholesterol in very large VLDL	Valine	0.99
Free cholesterol in very large VLDL	Tyrosine	0.96
Triglycerides in large VLDL	Glycine	0.97
Cholesterol esters in small VLDL	Glucose	0.96
Free cholesterol in medium LDL	Triglycerides in very large HDL	1.04
Phospholipids in small LDL	Mean diameter for LDL particles	0.99
Triglycerides in small LDL	Valine	0.99
Phospholipids in very large HDL	Tyrosine	1.01
Phospholipids in very large HDL	Mean diameter for LDL particles	1.01
Phospholipids in very large HDL	Estimated degree of unsaturation	1.02
Cholesterol esters in very large HDL	Triglycerides in very large HDL	1.02
Cholesterol esters in very large HDL	Alanine	1.05
Omega-6 fatty acids	Valine	0.99
Glucose	Citrate	1.04
Citrate	Tyrosine	0.97
Glycine	Isoleucine	0.99
Tyrosine	Mean diameter for VLDL particles	0.98

Table 4.11: Results from Cox regression with lasso, including all pairwise interactions.

Variable		Odds Ratio
Main terms		
Triglycerides in IDL		1.05
Total cholesterol in HDL2		0.87
Monounsaturated fatty acids; 16:1, 18:1		1.08
Interaction Terms		
Cholesterol esters in XXL VLDL	Free cholesterol in small LDL	1.07
Cholesterol esters in XXL VLDL	Citrate	0.94
Triglycerides in XXL VLDL	Mean diameter for LDL particles	0.98
Free cholesterol in very large VLDL	Tyrosine	0.94
Triglycerides in large VLDL	Glycine	0.99
Cholesterol esters in small VLDL	Glucose	0.93
Free cholesterol in IDL	Triglycerides in very large HDL	1.01
Phospholipids in small LDL	Valine	0.98
Phospholipids in very large HDL	Mean diameter for LDL particles	1.01
Phospholipids in very large HDL	Estimated degree of unsaturation	1.03
Cholesterol esters in very large HDL	Triglycerides in very large HDL	1.01
Cholesterol esters in very large HDL	Alanine	1.05
Citrate	Tyrosine	0.99
Glycine	Isoleucine	0.99
Tyrosine	Mean diameter for VLDL particles	0.98

Chapter 5

Networks

In the previous chapter we looked at some of the more commonly applied techniques for analysis of high-dimensional data. In this chapter we will concentrate on the use of *networks* as an analytic tool for exploring metabolomic data, which is a method used less frequently within epidemiology. In this chapter network theory will be introduced along with some of the basic terminology and features used to describe networks. This will be followed in chapters 6, 7 and 8 by an exploration of the emerging method *differential network analysis*.

5.1 Introduction

Network science is not a new discipline, the use of networks within the mathematical branch of graph theory has dated back from the 18th century [59], but has become particularly popular since the 1990s [60]. A network, in its most general form, is a graphical representation of a set of objects that are connected in some way [61]. A network (sometimes also called a graph) is made up of two key elements

- Nodes (sometimes called vertices) - the objects represented in the network.
- Edges (sometimes called links) - the connections between pairs of nodes.

In graph theory the elements are usually referred to as vertices and edges, in network science they are usually referred to as nodes and links [59] however in this thesis we will use the terms node and edge to reflect the terminology often used in epidemiological literature [62].

Figure 5.1 shows a classic example of a social network commonly used to introduce the concept of networks. It is a network of karate club members who were studied by Zachary in the 1970s [63], where each node represents an individual from the karate club with an edge drawn between a pair of nodes if the two individuals had a friendship outside of the karate club (friendship being determined by the number of common activities that pairs of individuals took part in).

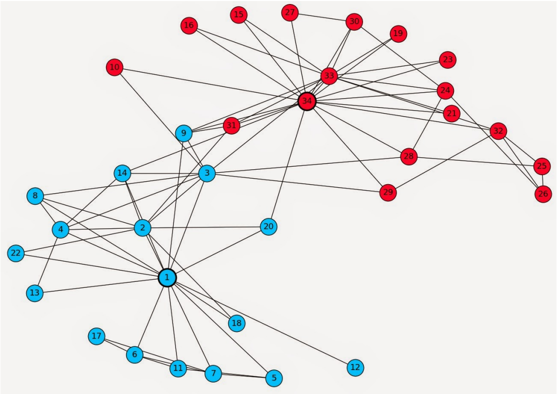


Figure 5.1: Zachary's karate club. Numbers inside each node are just node IDs, colours represent two groups identified.

From inspection of the network, it is possible to see that there are a few nodes that are highly connected (1, 2, 33, 34), and that most of the other nodes are connected to these nodes. In network terminology these nodes are said to be *central* to the network and are often called *hubs*. It also appears that two groups (coloured blue and red) have been identified centred around nodes 1 and 34. By a group we mean that there appears to be a set of nodes that are highly connected within each other but not with other groups. This is a concept called *modularity*. In Zachary's study, there was a disagreement within the club resulting in it splitting up into two separate karate clubs - the individuals that went to each of the two resulting clubs were split exactly as the groupings identified by the network analysis, barring 1 individual.

Another example network is that characterized by Goh et al [64], which defined a network of related diseases, with diseases connected by an edge if the diseases are associated with mutations in a common gene. In performing this network analysis, Goh found that the genetic origin of many diseases are shared, with the network modules illustrating groups of diseases with similar origins. A subset of this network is shown in figure 5.2.

A final example network (figure 5.3) is a correlation network of patient data from 4197 participants in the FinnDiane study [65]. Each of the nodes represents one of 39 patient characteristics, an edge is drawn between nodes if the Spearman correlation between the two variables is strong enough. The blue edges represent a negative correlation and the red a positive correlation, with the intensity of the colour representing the strength of that correlation.

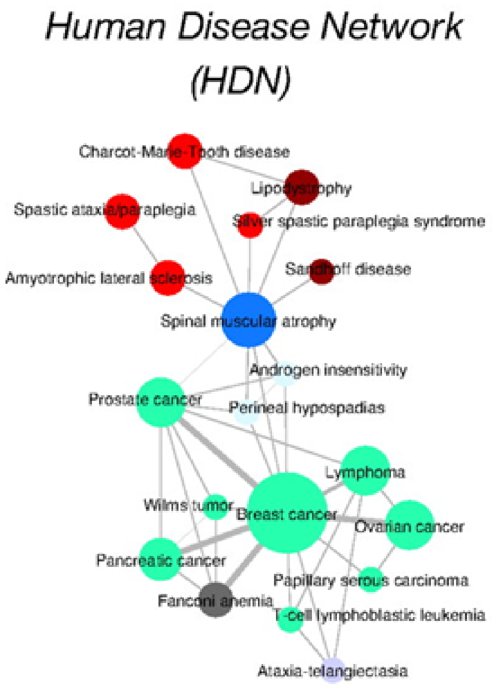


Figure 5.2: Subset of human disease network [64], colours representing the network modules

This network helps highlight the patient characteristics that tend to be more closely associated with one another. On the right hand side of the network the lipoproteins are all highly connected, at the top the weight, waist, hip and BMI nodes are all connected. This is similar information as to what can be found in a correlation heatmap, but is potentially in a format that would be easier for some readers to interpret, and also allows the network analysis methods described in this chapter to be applied, providing a different insight into the data.

These three examples have shown that networks can be used to summarise a range of different data types. For example in Zachary's network the nodes represented the 34 individuals in the study and the edges represented associations between those individuals. Within epidemiology, networks of this type are often used to investigate the spread of diseases between individuals [66–68]. The human disease network described used diseases as the nodes, and an edge represented when those diseases were associated with a common genetic mutation.

However in the FinnDiane study example, the nodes represented variables in the study, *not* the participants, with the edges representing associations be-

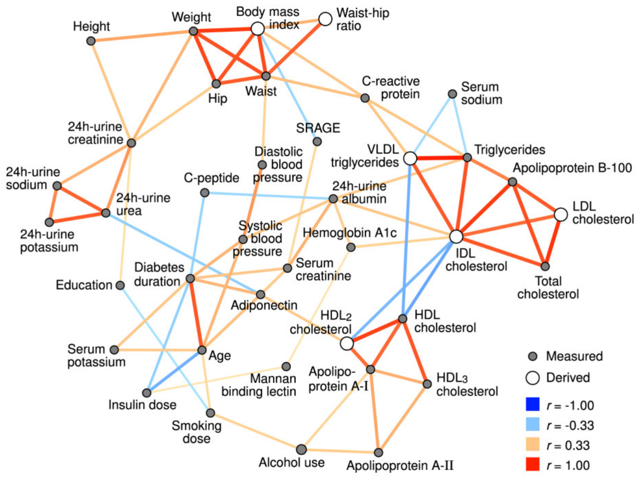


Figure 5.3: Network of patient characteristics from a study of 4,197 type-1 diabetes patients [65]

tween those variables (i.e. the Spearman correlation of the pair of variables in the 4197 participants). It is networks of this type, where variables are the nodes of interest, that we will be focussing on in this thesis. Much of the epidemiological literature on this type of network is from genetic epidemiology [69, 70], with some more recent literature covering their use in metabolomics [71, 72], which is our specific area of interest.

We will use network analysis on our dataset of 228 metabolites to develop our understanding of the metabolome and investigate associations between the metabolome and disease. But first I will describe some of the basic terms and features that are used in the analysis of networks.

5.2 Network Basics

5.2.1 Nodes and Edges

So far two new terms have been introduced as the basis for any network analysis: the node and the edge. Edges can either be directed or undirected, they are directed if the association between the nodes has a specific direction, however we will only be focussing on undirected networks. Edges can also have a weight attached to them, relating to the strength of the association between the nodes connected by a particular edge. For example in figure 5.3 the edges represented a Spearman correlation, so the magnitude of this correlation could be used as a weight for the edge. Unweighted networks can be considered a special case of a weighted network, with all edges having a weight equal to 1. We will look at both unweighted and weighted networks, but unweighted edges will be used unless specified. Figure 5.4 shows a basic, small, unweighted network that we will use as an example to introduce some network concepts.

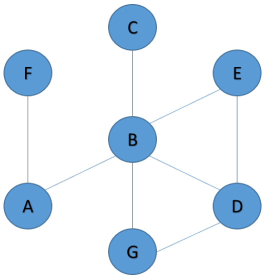


Figure 5.4: Example network 1, with nodes identified by the letters A to G

5.2.2 Connectivity

The simplest feature that can be used to describe a network is the network *size*, which is just the number of nodes in the network. The size of the network in figure 5.4 is 7. For a network of size p the maximum number of edges possible (in an undirected network) is

$$E_{max} = \frac{p(p-1)}{2}$$

If every node was connected to every other node we would call it a *fully connected* network. The network in figure 5.4 has 8 edges out of a possible 21 which can be used to calculate the *network density*, D. The network density is the ratio of edges in a network compared to total possible edges. So the density of the network in figure 5.4 is $8/21 = 0.38$. More formally the network density can be defined by :

$$D = \frac{2E}{p(p-1)}$$

where E is the number of edges in the network and p is the network size. The network density gives an idea of how "connected" the network is. A density of 1 means the network is fully connected, a density equal to 0 means that there are no edges connecting any nodes.

The number of edges connected to a node I is referred to as its *degree*, denoted by k_i. In the network in figure 5.4 node A has a degree of 2 i.e. $k_A = 2$, also $k_B = 5$ and $k_C = 1$. Two other network characteristics related to the degree can be calculated. The *average degree* is the mean number of edges each node is connected to, which is simply the number of edges divided by the number of nodes, multipled by 2 (since each edge is connected to 2 nodes):

$$\bar{k} = \frac{2E}{p}$$

The degree distribution describes the proportions of nodes within a network having each particular degree. *P(k)* is the probability that a randomly selected node in the network has a degree k [73]. Table 5.1 shows each of the degrees in the network shown in figure 5.4, along with how many nodes exhibit that degree and therefore the proportion of nodes in the network. The degree distribution can then be plotted (figure 5.5).

Degree	Nodes	Proportion
0		0/7
1	C, F	2/7
2	A, E, G	3/7
3	D	1/7
4		0/7
5	B	1/7

Table 5.1: Degree distribution of example network from figure 5.4

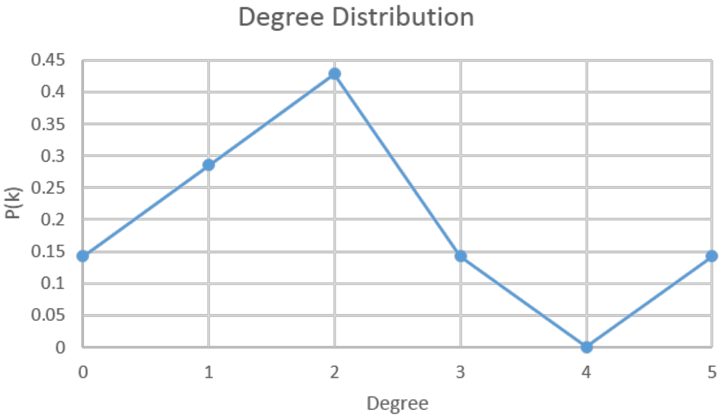


Figure 5.5: Plot of degree distribution of example network from figure 5.4

5.2.3 Adjacency Matrix

If a pair of nodes in a network are connected via an edge they are said to be *adjacent* and an adjacency matrix is a means of representing the edges in a network. If the network is of size p, then a $p \times p$ matrix is used, with the rows and columns referring to the nodes, a 1 representing an edge between a pair of nodes and a 0 representing the absence of an edge. The adjacency matrix for the network in figure 5.4 can be seen in figure 5.6. The matrix is symmetric as if A has an edge to B then B also has an edge to A. However in the example of a directed network it would be possible to use one half of the matrix to represent the edges directed one way and the other half to represent the edges directed the other way. Also, in the case of a weighted network, the matrix cells could represent the strength of the edge rather than just representing whether there is an edge or not.

$$\begin{array}{cc} & \begin{array}{ccccccc} A & B & C & D & E & F & G \end{array} \\ \begin{array}{c} A \\ B \\ C \\ D \\ E \\ F \\ G \end{array} & \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

Figure 5.6: Adjacency matrix for network in figure 5.4

5.2.4 Paths

A path is a series of edges that connect 2 nodes to one another, with the path's *length* being defined as the number of edges traversed [74]. For example in figure 5.4 there is a path from E to F that goes E-B-A-F which has a length of 4. The *shortest path* between a pair of nodes is defined as being the path between nodes that has the shortest length, for example there is a path from A to D (A-B-D) of length 3 and also paths (A-B-E-D) and (A-B-G-D) of length 4, with the path of length 3 being the shortest path.

A network is called *connected* if a path exists between every pair of nodes in the network [59], such as in the network in figure 5.4. In figure 5.7 the network is *disconnected* as there are no paths connecting nodes A and F to the other nodes.

5.2.5 Node centrality

The *centrality* of a node is a measure of its relative "importance" in a network [75]. Consider the network in figure 5.4. Node F is only connected to one node so we might say it is not particularly integral to the network. Node B on the other hand is connected to 5 other nodes so we might consider it an important node in the network. Calculating a node's centrality is a formal method of assigning it an importance, and there are different measures of centrality that represent different characteristics of the network.

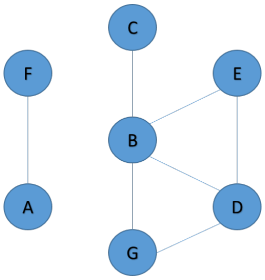


Figure 5.7: Example network 2 - example of a disconnected network

5.2.5.1 Degree Centrality

The most straightforward measure of a nodes centrality in a network is *degree centrality*. Each node's centrality is simply set equal to its degree, as defined in section 5.2.2. These can be standardised to between 0 and 1 by dividing it by the maximum degree observed in the network.

5.2.5.2 Betweenness Centrality

Betweenness centrality is another measure of a node's centrality within a network, based on the concept of shortest paths. A node's betweenness centrality is equal to the proportion of shortest paths that go through that node. It is estimated using the following steps for all nodes in the network:

1. Select a node that we want to calculate the betweenness centrality for (we'll refer to it as the node of interest)

2. For all other pairs of nodes in the network identify the shortest path between each pair

3. For each of the shortest paths that travel through the node of interest, add 1 to the betweenness centrality score

4. If there are two or more paths that are equally short between a pair of nodes, add the proportion of these paths that go through the node of interest to the betweenness centrality score. e.g. the shortest path between E and G is of length 2, and there are 2 different routes of that length (E-B-G) and (E-D-G), if node D was our node of interest we would add 1/2 to the betweenness centrality score

5. Scale result to between 0 and 1 by dividing the score computed in steps 1-4 by the number of possible node pairs involved. So the denominator here is $((p-1)(p-2)/2)$

This sounds a little more complicated than it actually is so to make it clearer we apply it to network 1 from figure 5.4 to calculate the betweenness centrality of node A. There are 5 pairs of nodes that have a shortest path through A, these are (B,F), (C,F), (D,F), (E,F) and (F,G). There are 6 nodes in the graph, excluding A, which means $\frac{6\times5}{2} = 15$ pairs of nodes. So betweenness centrality of node A is $\frac{5}{15} = 0.33$. Node B has a betweenness centrality of 0.77 and nodes C, F, E and G have a betweenness centrality of 0.

Node	No. Shortest Paths	Betweenness Centrality
A	5	0.33
B	11.5	0.77
C	0	0
D	0.5	0.03
E	0	0
F	0	0
G	0	0

Table 5.2: Betweenness centrality of example network in figure 5.4

5.2.5.3 Closeness Centrality

Closeness centrality is a measure of how close each node is to all other nodes. If we describe the distance between a pair of nodes as the length of the shortest path between those two nodes then we can calculate the mean distance from a node to all other nodes in the network by summing all the distances together and dividing by the number of nodes. We want to measure the *closeness* rather than the distance, so we take the reciprocal of this value which will provide a value between 0 and 1.

One limitation of closeness centrality is that it can only be calculated for connected networks [76]. For example in network 2 nodes A and F are not connected to any other nodes, so their distance from every other node is infinity. Therefore the mean distance for all nodes will be infinity, and the closeness will be 0. We can calculate the closeness centrality for each node in the example network from figure 5.4. Table 5.3 shows the distance between each node and provides the closeness centrality for each node. As might be expected node B has the largest closeness centrality, node C has a greater closeness centrality than node F despite having the same degree, so this feature captures the fact that F is more isolated than C.

5.2.5.4 Eigenvector Centrality

Eigenvector centrality is based on the principle that a central node in the network will not only itself be highly connected but also be connected to other more central nodes, so the eigenvector centrality of a node is a function of the

Node	Distance							Mean Distance	Closeness Centrality
	A	B	C	D	E	F			
A		1	2	2	2	1	2	1.67	0.60
B	1		1	1	1	2	1	1.17	0.86
C	2	1		2	2	3	2	2.00	0.50
D	2	1	2		1	3	1	1.67	0.60
E	2	1	2	1		3	2	1.83	0.55
F	1	2	3	3	3		3	2.50	0.40
G	2	1	2	1	2	3		1.83	0.55

Table 5.3: Closeness centrality of example network from figure 5.4

centrality of its adjacent nodes. To calculate the eigenvector centrality of all nodes in a network we use the following iterative process:

1. As a starting point the eigenvector centrality of each node can be initially set to be equal to its degree centrality
2. An updated eigenvector centrality is calculated for each node by adding together all the centralities set in step 1 from adjacent nodes
3. This new centrality is scaled by dividing it by the largest centrality in the network, resulting in centralities for all nodes between 0 and 1
4. Repeat steps 2 and 3 until all the centralities stabilise (i.e. stop changing according to a preset amount, say $\tau = 0.001$)

A table of the eigenvector centralities for example network 1 (figure 5.4) is shown in table 5.4, alongside each of the other centralities calculated for comparison. Nodes E and G have the same centralities in all cases as they both have identical edges (2 edges, connected to nodes B and D). Node F has the lowest (or joint lowest) centrality in each case, which is understandable from inspection of the network, where it only has 1 edge and is not connected to the hub node B.
Comparing nodes A and G is interesting as they have the same degree centrality (both have a degree of 2), but G has a higher eigenvector centrality whereas A has higher betweenness and closeness centralities. Both are connected to node B so those effects cancel each other out, but the other edge of G is connected to D which is a fairly central node in the network, and he other edge for A is connected to node F, the least central node in the network. However, in terms of betweenness, any path that connects node F with any other node must pass through node A, whereas node G has no shortest paths passing through it, hence A has a higher betweenness centrality. Also, as F is more “distant” from G than A is from any other node, the closeness centrality of G is therefore smaller.

Node	Eigenvector centrality	Degree centrality	Betweenness centrality	Closeness centrality
A	0.40	0.40	0.33	0.60
B	1.00	1.00	0.77	0.86
C	0.35	0.20	0.00	0.50
D	0.80	0.60	0.03	0.60
E	0.64	0.40	0.00	0.55
F	0.14	0.20	0.00	0.40
G	0.64	0.40	0.00	0.55

Table 5.4: Each of the 4 centrality measures for example network 1 (figure 5.4)

5.2.6 Community Structure

Within networks it is common to see groups of nodes where members densely connected with each other, but not with members of other groups [77]. This property is referred to as the *community structure* or *modularity* of the network. Figure 5.8 illustrates a network where there are 3 sets of nodes densely interconnected within their groups, but only a few connections between groups. We will refer to a group of nodes like this as a network *module*.

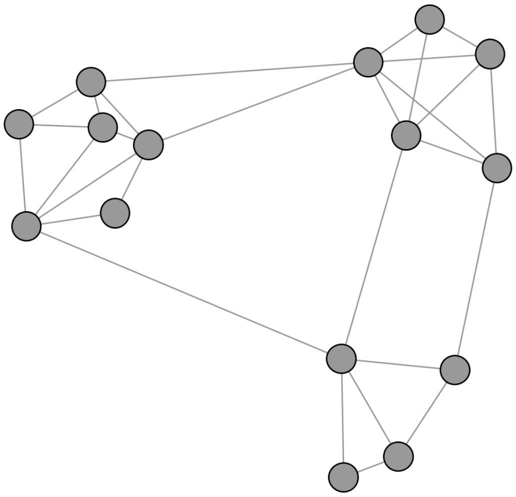


Figure 5.8: Example of a network with 3 modules

Identification of these modules could be useful e.g identification of groups of metabolites that act together confirming some a priori knowledge or identification of new, previously unknown structures. There are parallels to hierarchical cluster analysis, where closely related nodes are grouped together.

The basis of most algorithms that aim to identify community structure is the modularity, Q. This feature quantifies the proportions of edges that exist within communities compared with the number of edges that would be expected within a community if the edges were randomly distributed. A value of Q close to 1 (high modularity) would mean that the groups are very densely connected, with very few edges linking different communities, a value close to 0 (low modularity) would mean that the connections within a community are as dense as you would expect if all the edges were randomly allocated across the network. The formula for Q was defined by Newman and Girvan [77] and is as follows:

$$Q = \sum_{c_i \in C} \left[\frac{\left|E_{c_i}^{in}\right|}{|E|} - \left(\frac{2\left|E_{c_i}^{in}\right| + \left|E_{c_i}^{out}\right|}{2\,|E|} \right)^2 \right]$$

Where c_i indexes the ith community and C the set of all communities. E is the total number of edges in the network, $E_{c_i}^{in}$ is the number of edges that exist between all the nodes within community c_i and $E_{c_i}^{out}$ is the number of edges that exist from nodes within community c_i to nodes in other communities.

There are many procedures used to identify network modules, with the main focus of recent algorithms being to reduce the computation time, as the procedure tends to be computationally intensive [78]. In this thesis we will not review different methods of detecting network modules, I will use the method that is built into the visualisation tool, Gephi, used to draw the network diagrams in this thesis. The algorithm is described in the paper by Blondel et al [79], a brief overview is given here. The aim of the algorithm is to maximise the Q-statistic defined above and is an iterative algorithm with the following steps:

1. Initially each node is defined as an individual community, so the number of communities is equal to the number of nodes and Q is calculated.

2. Each node in the network is then considered in turn. It is sequentially moved into each of its neighbours' communities, and for each of these new set of communities the resulting change in modularity (ΔQ), if any is calculated.

3. If any of these moves results in a positive increase in modularity, then the node is allocated into the community that resulted in the largest increase in modularity.

4. Steps 2 and 3 are repeated until the communities are stable and another pass of the process results in no change to the community structure.

5. Then, the identified communities are collapsed down to create a new network, with each node in the new network representing a community identified in the previous iteration. The weight of the edges between nodes is equal to the number of edges connecting those communities in the original network.

6. Steps 2,3 and 4 are carried out on the new collapsed network.

7. This process is repeated until the maximum value of Q is found.

5.3 Network analysis of BWHHS data

5.3.1 Methods

Having given an introduction to a few of the basic terms involved in a network analysis we now want to use the data from the BWHHS to allow us to revisit these concepts in a real setting, illustrating practical difficulties and the statistical issues encountered. However, before we can describe a network we need to generate the network in the first place.

The basic steps required to generate a network are as follows:

1. Define what the nodes are going to be, and select which ones are to be included.

2. Choose the metric of association which will define whether an edge exists (or the weight of the edge) between two nodes. e.g. Pearson's correlation or Spearman correlation, marginal or partial correlation *etc.*

3. Estimate the metrics of association from the available data and draw network.

4. Sometimes the network will be too interconnected to do any reasonable inference. In this case it may help to make the network more sparse by 'thinning out' the network, retaining only the more important edges.

5. Present and discuss the network.

Our primary interest is in creating a network that represents the associations between the NMR biomarkers. We define the nodes in the network as the NMR biomarkers and an edge between each pair of nodes by a correlation. We will discuss the different methods of correlation that can be used in section 5.3.1.1. We also define the criteria for inclusion of an edge in the network via the p-value of the selected correlation coefficient (with the null hypothesis being that the correlation between the pair of variables is 0) used to determine inclusion. The method of calculating the p-value will depend on the measure of association chosen, but for example if the Pearson correlation is chosen then the p-value would be calculated using equations 5.5 and 5.6. The threshold at which an edge is included in the network will be the p-value at which we can control the FDR to 5%. This is chosen ahead of controlling the FWER to ensure we include most of the "true" edges since we are prepared to accept that, in expectation, 5% of the edges generated to be false positives.

5.3.1.1 Defining an edge

As discussed in the introduction, edges can be defined using alternative measures of association among nodes. In the following we will consider only Pearson correlations to define such measures. The observations drawn however extend directly to Spearman correlations (since the latter are Pearson's correlation calculated on ranks).

Correlation The Pearson correlation coefficient is a well established statistical method of measuring a linear association between a pair of variables x and y [80]. It is defined as:

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

where $\bar{x}$ and $\bar{y}$ are the variables population means and N is the population size.

Assumptions underlying the Pearson's correlation coefficient are

1. There is a linear relationship between the variables
2. The variables should be approximately normally distributed
3. The distribution of each variable has constant variance for given values of the other (homoscedasticity)

If we use the Pearson correlation coefficient as a measure of association between pairs of biomarkers (i.e. as edges in the network) we measure the overall association between the biomarkers. These edges then represent both direct and indirect associations between variables and, since metabolomic datasets are generally highly correlated [81], there will be many indirect associations included. This would result in an extremely dense network, providing little insight into the structure of the metabolome. By instead using the partial correlation coefficient as a measure of association, we focus on the direct association between pairs of biomarkers, i.e. the association between a pair of biomarkers *conditional* on all other variables in the network.

The partial correlation is a way of measuring the correlation *adjusted* for all other variables in the network [80]. One way to estimate the partial correlation of a pair of nodes A and B, adjusting for a vector of variables **Z**, is to perform a linear regression of A on **Z** and also of B on **Z**, then calculate a standard Pearson correlation on the estimated set of residuals from each regression. The resulting correlation coefficient will be the partial correlation coefficient relating A and B adjusting for **Z**.

It is also possible to estimate the partial correlations of a set of variables from their covariance matrix, Σ, provided the covariance matrix is invertible. This method requires less computational time if the number of variables is large. If we define the inverse of the covariance matrix as Σ^{-1} and each element within that inverse covariance matrix as s_{ij} and the partial correlation between variable i and variable j as p_{ij}, then the partial correlations are related to the inverse covariance matrix using equation (5.1) [80],

$$p_{ij} = -s_{ij}/\sqrt{s_{ii}s_{jj}} \tag{5.1}$$

Shrinkage Method When the number of variables is high relative to the number of observations the covariance matrix is estimated imprecisely. Also, if the number of variables is greater than the number of observations, the covariance matrix cannot be inverted. Strimmer and Schäfer propose an improved estimation of the covariance matrix using shrinkage methods [82], which addresses these problems.

This method is based on combining two estimates of the covariance matrix. The first, **U**, is the empirical estimate of the covariance matrix (i.e. calculated from the observed data), which is unbiased but has a high variance. The second is a proposed, constrained estimate of the covariance matrix, **T**, which will have a lower variance but will potentially be a biased estimate of the true covariance matrix (it can be thought of as a 'prior' covariance matrix). A weighted combination of these provides a new (and hopefully improved) estimate of the covariance matrix, **U***

$$\mathbf{U}^* = \lambda\mathbf{T} + (1-\lambda)\mathbf{U} \tag{5.2}$$

λ is the shrinkage intensity, so if $\lambda = 0$ then the estimated covariance matrix will be equal to the empirical covariance matrix, and if $\lambda = 1$ the estimated covariance matrix will be equal to the constrained covariance matrix. As the shrinkage intensity increases, more bias is introduced, but it is offset by the reduction in the variance of **U***.

So there are two steps involved in the shrinkage. First choose the constrained matrix, then calculate the shrinkage intensity. In the original paper, 6 potential constrained matrices are suggested for use in common situations. We are actually only interested in shrinking the covariances (we're not interested in the variances), so we select the constrained covariance matrix where the covariances are equal to 0 and the variances are equal to the sample variances, with the target matrix **T** shown below. The resulting matrix **U*** will have covariances that have been shrunk towards 0.

$$\mathbf{T} = \begin{pmatrix} v_{11} & 0 & \cdots & 0 \\ 0 & v_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & v_{KK} \end{pmatrix} \tag{5.3}$$

Where v_{ii} is the variance of the i-th variable and K is the number of variables.

The shrinkage intensity λ should be chosen to give an appropriate balance between reducing the variance of the estimated covariance matrix while trying to retain as little bias as possible. The derivation of an appropriate value of λ is provided in detail by Ledoit and Wolf [83], with the aim being to reduce the sum of the squared differences between each element of the sample and true covariance matrices. The algebra simplifies down to show that selection of λ is based on the variance of the sample covariances. The formula is:

$$\lambda = \frac{\sum_{i \neq j} \widehat{Var}(s_{ij})}{\sum_{i \neq j} s_{ij}^2} \tag{5.4}$$

So if the sum of the variances of the estimated covariances is large compared to the value of the sum of the estimated covariances (squared) then the shrinkage intensity, λ will be large. If the sum of the variance of the estimated covariances is relatively small, the shrinkage intensity will be small. Or to put it more simply, if the number of observations, n, is much larger than the number of variables, p, then λ will be very small. As p increases in size, relative to n, then λ will increase. A detailed explanation of how to calculate the variance of the correlation coefficients is found in the paper by Schafer [82].

We can also calculate the p-value for the partial correlations using Fisher's z-transformation [84] using equations:

$$z_{ij} = \frac{1}{2} ln \left(\frac{1 + p_{ij}}{1 - p_{ij}} \right) \tag{5.5}$$

$$Z = \frac{z_{ij}}{\sqrt{\frac{1}{N-p-3}}} \tag{5.6}$$

Packages to estimate these statistics are available in R and for the purposes of the analyses performed in this thesis, functions were written in Stata.

Node inclusion In chapter 3 the NMR biomarkers were described and it was identified that many of these are highly collinear and some are composites of other biomarkers. The same approach as was taken for the PCA and lasso analyses in chapter 4 will be applied here, removing those metabolites that are exact sums of other metabolites, removing metabolites that have a large proportion of missing data and removing those metabolites identified as having poor reliability in chapter 3. This results in a final set of 78 metabolites to be used.

Network Visualisation The software Gephi [85] has been used to visualise the networks in this thesis. Gephi requires the user to provide it with the list of nodes and a list of edges that are to be included in the network, and also to select the algorithm which generates the network layout. The algorithm used for generating the networks in figures 5.9, and 5.10 is the *Force Atlas 2* algorithm [86]. Briefly, it works by starting in a random configuration, then setting a repulsive force to act between all nodes, but if an edge exists between a pair of nodes an attractive force works in the opposite direction to this repulsion (if the edge is weighted the attractive force is relative to the weight of the edge). The algorithm then works iteratively, moving unconnected nodes apart, while pulling connected nodes closer together. This has the effect that nodes more closely co-located on the network graph are more closely related to one another, either because they have a strong connection or they are both connected to a similar set of nodes. This method of laying out the graph allows us to observe if the nodes cluster and can give some visual intuition as to the structure of the data. The method may result in different final configurations depending on the starting positions of the nodes, however when the algorithm was applied in practice on the BWHHS data, the starting positions made no material difference to the final structure.

5.3.2 Results

Figure 5.9 shows the network that is generated using partial Pearson correlation coefficients (the partial correlation being the correlation between a pair of nodes adjusted for all other nodes in the network) with unweighted edges, using the 78 metabolite concentrations (using the same selection process as in chapter 4) as the network nodes, no other pre-filtering of nodes was performed,. As described earlier we estimate the partial correlation of each pair of nodes and retain an edge between a pair if the p-value of that partial correlation is less than that required to keep the FDR at 5%.

Very little can be inferred from visually inspecting this network, it is dense (density $D = 56\%$, average degree $\bar{k} = 42.9$), although by running the modularity algorithm 5 modules are detected - the green module is mainly extremely large VLDL lipoproteins, the yellow module mainly the small to large VLDL lipoproteins and some triglycerides, the orange module contains mainly LDL metabolites and the blue mainly HDL lipoproteins. The red module is a bit more detached from the others and contains the amino acids and fatty acid metabolites. Edges representing a positive partial correlation are depicted in black, with negative correlations shown in red, however, given the density of this network it is not easy to see these clearly as it just appears as a large mesh of edges.

Applying the shrinkage method described in section 5.3.1 we find the network shown in figure 5.10. There are fewer edges ($D = 43\%, \bar{k} = 32.7$) with 4 modules identified, and all VLDL particles are combined into a single module.

The networks above can be useful for describing a dataset and gaining a greater insight into the associations between metabolites, however they do not provide any information relating to an outcome/disease of interest. To investigate this we can begin to look at the concept of *differential networks*, discussed in the next chapter.

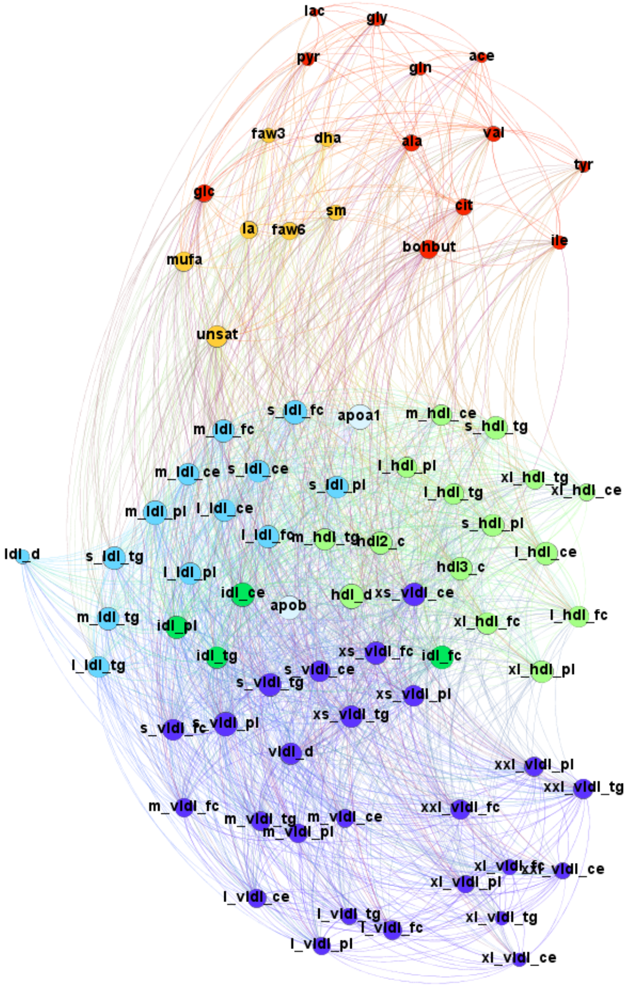


Figure 5.9: Network of 78 metabolites from the BWHHS defined using the Pearson partial correlation coefficient as a measure of association. Node size is proportional to node degree. Node colour represents the different groups of metabolites (LIPO window VLDL = purple, LIPO window LDL = blue, LIPO window IDL = dark green, LIPO window HDL = light green, LIPO window other = grey, LMWM window = red, LIPID window = orange.)

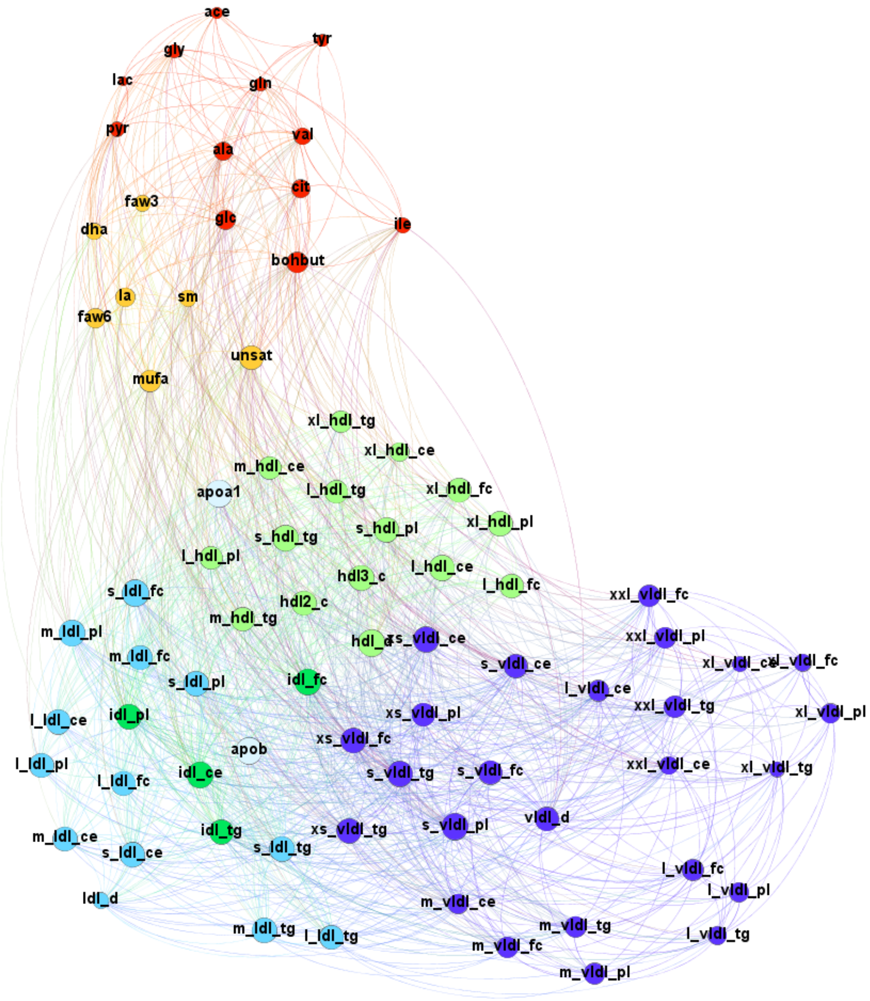


Figure 5.10: Network of 94 metabolites from the BWHHS defined using the shrunk Pearson partial correlation coefficient as a measure of association, after shrinkage of the covariance matrix. Node size is proportional to node degree. Node colour represents the different groups of metabolites (LIPO window VLDL = purple, LIPO window LDL = blue, LIPO window IDL = dark green, LIPO window HDL = light green, LIPO window other = grey, LMWM window = red, LIPID window = orange.)

Chapter 6

Differential networks

In this chapter the topic of differential networks will be introduced, including a brief literature review on the topic. Then a detailed simulation study will be described, illustrating potential scenarios that can give rise to an edge in the network.

6.1 Background

The term “differential network” refers to a method of analysis that compares two, or possibly more, networks, which represent two, or more, subgroups allowing a description of how the two networks differ to be made. It does not refer to a single standard method, rather it is an umbrella term for any method that compares networks derived from different groups. For example we can compare the metabolomic network formed by a set of diseased individuals with the network formed by a set of disease-free individuals.

Since 2005 there have been academic papers using the term differential networks to mean a number of different types of analyses. However, in the literature review provided here, only research that uses differential networks to compare the differences between networks generated from data representing two (or more) different biological states will be reviewed.

One of the earliest examples of the use of differential networks in this way was described by Fuller et al in 2007 [87]. In this instance they used genotype data from 135 mice to create a gene coexpression network, with the nodes in the network representing gene expressions and an edge between nodes representing a Pearson correlation between a pair of gene expressions. To construct the differential network they took the 30 leanest mice and 30 heaviest mice and constructed a gene coexpression matrix for each set of mice. For each node in the network the weighted degree centrality was estimated and the difference between this in the lean and heavy mouse groups was estimated. A total of 61 genes were identified as being differentially connected (in terms of their expression), and of those 12 had previously been identified as causal factors for obesity in mice. The information from the difference in degree centrality was also combined with the results of a standard set of t-tests comparing the gene expression levels between the two groups to further refine the group of genes to be identified as potentially associated with mouse weight.

This was an early, informal, example of network comparison and the authors were careful not to over interpret their results, and were mainly used to strengthen the evidence already found from using the more standard differential expression analysis (univariate t-tests). If a gene was found to be both differentially expressed (on average) AND differentially connected (in the network) then greater evidence that this gene is associated with mouse weight would be provided than if only differential expression was identified.

The first paper that provided a formal approach to differential network analysis was by Gill et al [88] "A statistical framework for differential network analysis". They expanded on the idea of comparing how the network structure of genetic microarray data changes between two (or more) biological states by proposing three novel statistical tests to address the following questions:

- Does the network structure differ overall across different states?
- Does the network structure of a particular network module differ across states?
- Does the connectivity of a single gene differ across different states?

The proposed method involves generating a network of "healthy" individuals and a network of "diseased" individuals, and testing their difference using the three tests corresponding to the bullet points above. Within each network a node represents a gene and an edge between the nodes represents a weighted measure of association. They propose three measures (correlation, partial correlation and partial least squares) as possible statistics for quantifying the association, although in their simulations they concentrate on correlations. To test the difference in connectivity of a single gene, every weighted edge (weighted using the coefficient of the chosen association measure) in one network is subtracted from the corresponding weighted edge in the second network, then for every node, the mean value of the difference in edges surrounding it is calculated. This gives the node a differential connectivity score. Similarly this score can be estimated for a network module, or the network as a whole.

To assess the significance of these statistics, a permutation test is performed. The observed data are randomly assigned to the "healthy" or "diseased" groups and the scores recalculated for each gene, module and the overall network. This can be repeated a number of times (in the paper 1000 times) and a distribution of scores created. Then the original scores from the observed data can be compared to the permuted distributions and p-values obtained for each.

To investigate the method's potential, the authors perform a series of simulations and report the results, although only for the final of the 3 proposed statistical tests (connectivity of single genes). The simulations are performed for a network size of 20 and of 100, and in each case 10 genes are defined as being "important", that is they are associated with each other in the "case" network and not associated in the "control" network. So the control network is based on a network where no genes are thought to be associated and the case network is one where 10 genes are associated with one another and the rest are not. The tests are performed using a cut-off for "significance" at the unadjusted threshold of 5% and also adjusted using the Benjamini-Hochberg adjustment [46]. The results from the simulations suggest that the method is promising, in terms of sensitivity, specificity, true discovery rate and true non-discovery rate which in most scenarios are found to be adequate, although the sensitivity of the test reduces to 34% when 100 genes are used, on a sample size of 50 and the correlations of the "important" genes are set to the lowest level tested, 0.5.

This paper provided a rigorous, formal procedure for testing differences in networks, and in performing a series of simulations it attempted to assess the quality of the method in achieving its stated aims, however it is difficult to simulate a comprehensive range of scenarios, given the complex nature of the data that they were trying to simulate. The control network they chose was a set of fully independent genes, and the case network had 10 highly associated genes (the smallest correlation they used was $\rho = 0.5$). A wider range of simulations may have provided a greater understanding of benefits and limitations of the method, however the method used was very computationally expensive so this also constrained the ability to investigate further scenarios. Nevertheless this paper was able to produce evidence that the method worked as expected on a small number of idealised scenarios. The researchers were also careful to note that these statistics should only be used as exploratory method of analysis, and recommended it as a potential first step to filter out less interesting genes and provide a subset of genes that can be further explored. More recently in 2015, Kujala et al [89] expanded these methods to be able to deal with missing observations by performing multiple imputation. The methods were then applied to an observational study comparing a group of patients who suffered a fatal CVD event to those who had not.

In 2010 de la Fuente [90] produced a review piece describing the move from differential gene expression to differential networking and it seems at around this time the concept of differential networking was becoming a wider research topic. Ideker and Krogan [91] performed a review of differential network biology two years later where they argue that "differential network mapping, which allows for the interrogation of previously unexplored interaction spaces, will become a standard mode of network analysis in the future".

In 2011 Valcarcel et al [71] applied a differential network method to metabolomic data, in this instance the outcome of interest was pre-clinical diabetes, measured by raised (cases) or normal (non-cases) fasting glucose levels. Separate networks were estimated for the cases and non-cases, with the network nodes being 60 metabolites, and the edges defined by the partial correlation between pairs of metabolites. To estimate the differential network, the difference between each edge in the two networks was calculated and a permutation test used to derive a p-value for each differential edge. A cut-off of p <0.01 was used for inclusion of an edge in the differential network. This method was later applied a study of Caenorhabditis elegans DAF-2 mutants by Castro et al [92]. Also in a subsequent analysis in 2014 Valcarcel et al combined their earlier method with a genome-wide correlation analysis, to identify if any of the differential correlations detected were associated with any genetic variants [72].

Walley et al [93] and Chu et al [94] propose similar methods to the above, with the former using Kendall's tau correlation as the measure of association and the latter using a Bayesian approach and estimating a posterior odds ratio of connectivity for each pair of nodes in the network.

Odibat and Reddy [95] expanded the methods previously suggested by moving beyond what they term "differential connectivity" (differential degree centrality) to "differential centrality" (differential *betweenness* centrality). These are illustrated in figures 6.1 and 6.2 respectively. The examples described so far look at all the edges emanating from a node and how they compare across the two networks from each biological state, which is closer to the definition of "differential connectivity" in this context. Whereas the differential centrality proposed by Odibat and Reddy estimates a statistic measuring the "importance" of each node within the overall network.

Bockmayr et al [96] take a similar approach, defining two protocols (which they call DCloc and DCglob, short for "Differential Connectivity Local" and "Differential Connectivity Global") which compare the local and global topologies of the two correlation networks, however their specific methods of doing so are quite different from those proposed previously. In DCglob, first the disease-specific networks are derived, using the Fisher-transformed Pearson correlation of each pair of genes as the measure of association. Rather than using an arbitrary cut-off for inclusion of an edge, 200 cut-off levels are used, ranging from 0 to the maximum value of Fisher-transformed Pearson correlation (which was 2.5), to create 200 potential networks for both the cases and non-cases (i.e. when the threshold was 0, every possible edge was included in the network, when the threshold was 2.5, there were no edges). Then at each of these 200 cut-offs the case and non-case networks were compared. Any genes that are connected to at least 2 other genes in both networks are excluded from the analysis, as are genes not connected to at least two other genes(i.e. connected to one or zero

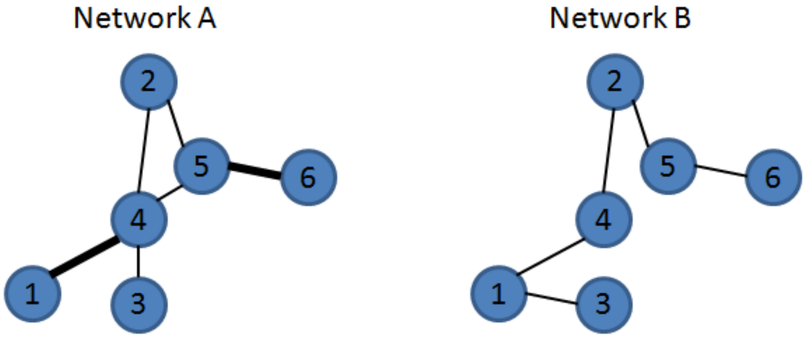


Figure 6.1: Illustration of differential connectivity. Node 4 has 4 edges in network A and 2 in network B while all other nodes have at most 3 edges in A and at most 2 in B, therefore node 4 has a high differential connectivity score compared to the other nodes in the network.

other genes) in both networks, leaving the set of genes that are determined to be “differentially connected”. The definition of a gene being differentially connected is that in one group (either cases or non-cases) the gene is connected to two or more different genes and in the other group the gene is connected to fewer than two genes (either one or zero). The set of differentially connected genes can then be defined for each of the 200 potential cut-offs.

In DCloc, the process is the same up to the point where the 200 comparison networks have been obtained. At this point each gene is taken in turn and at each threshold level the difference in its degree is measured. The average difference in degree is taken as the measure of differential local connectivity.

These methods proposed by Bockmayr et al have an advantage over previous ones in that the threshold at which an edge is included is not arbitrarily set. However the results are difficult to interpret, which the authors acknowledge themselves in their discussion, and also the methods have not been statistically evaluated, but only applied to a real dataset.

Gambardella et al [97] take a practical approach to the problem, developing a procedure called DINA (DIfferential Network Analysis) aimed at testing known pathways and identifying whether they differ across different disease states, rather than identifying novel pathways.

Danaher et al [98], Zhao et al [99] and Xia et al [100] have investigated methods of increasing the efficiency of the estimation of differential networks, with the first paper investigating the use of the joint graphical lasso and the latter two

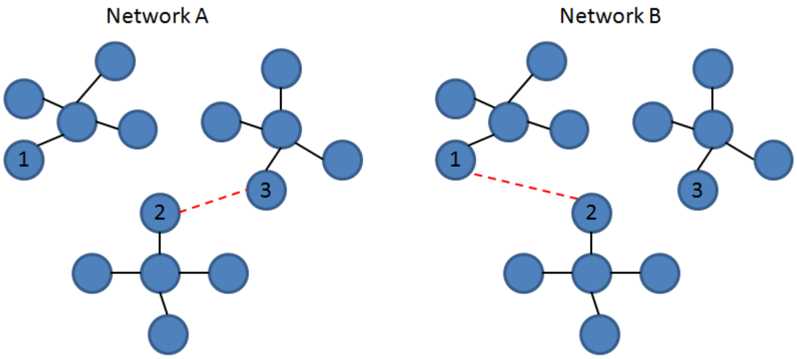


Figure 6.2: Illustration of differential centrality (The red dashed line represents the edge that differs in the two networks). Node 3 is central to network A, but in network B it is less central. It would have a high differential centrality score.

proposing their own novel methods of estimation, directly identifying the differential network rather than estimating two separate networks to be compared.

Each of these authors agree that quantifying the differences in associations between variables across biological states can potentially provide useful information, although all are cautious when interpreting the results. No-one is proposing differential network analysis is yet ready to be a first port of call when doing an analysis of high dimensional data, rather that it may be a useful process to carry out when performing exploratory analysis on a suitable dataset. A number of different methods are proposed, with different measures of association, measures of difference and statistical tests used in each.

However, one element that is lacking from the literature to date is an assessment of the scenarios that give rise to a difference in the associations in the networks. A few of the papers show simple examples of what correlations in the cases and non-cases look like when they differ, such as the one shown in figure 6.3, with no mention of the data generating model that is assumed that would give rise to such data. Without understanding this, the interpretation of what a differential network is doing is difficult. In the rest of this chapter we aim to take one of these methods, the method proposed by Valcarcel [71] (reviewed above), and investigate what joint distributions of data might induce an edge between a pair of variables in a differential network. This method was chosen as it appeared to be the method with the simplest definition of an edge in the differential network - a difference in the partial correlations (or shrunk partial correlation) of a pair of variables between the two networks. We will refer to this as “correlation based differential networks”.

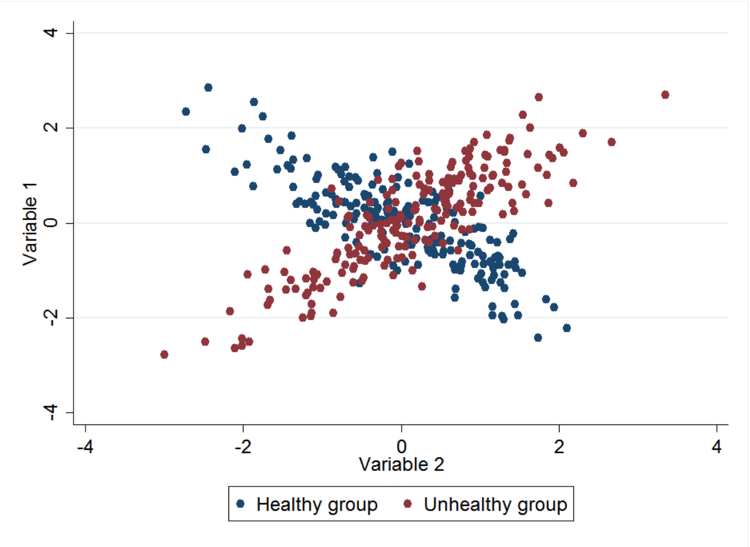


Figure 6.3: Example of a scatter plot of observations from a pair of variables, with data points coloured by groups - blue: healthy individuals, red: unhealthy

6.2 Correlation based differential networks

For each pair of variables in the differential network a difference defined in terms of the partial correlations of the two original networks is estimated along with a p-value testing the null hypothesis that the difference is equal to 0. A network diagram can then be created, by setting a threshold for the p-value and including only those edges that have a p-value below that threshold. Going into more detail on the process for estimating an edge between a pair of nodes i and j, which we will define as δ_{ij}, the process defined in [71] is as follows:

Step 1 Estimating δ_{ij}

- Estimate the (shrunk) Pearson partial correlation in the cases only
- Estimate the (shrunk) Pearson partial correlation in the non-cases only
- Subtract the partial correlation coefficient in the non-cases from the coefficient in the cases to give the estimated difference, $\widehat{\delta}_{ij}$

Step 2 Non-parametric inference for $\widehat{\delta}_{ij}$

- Randomly permute the observations to be cases or non-cases (with the same number of cases and non-cases as in the original data) a preselected number of times.
- For each permutation $p = 1, 2...P$ repeat step 1 to get a permuted value for δ_{ij}, denoted δ_{ij}^{p}

- Calculate the proportion of the permuted values where δ_{ij}^{p} is greater in magnitude than the estimated $\widehat{\delta}_{ij}$
- This proportion is the p-value for the null hypothesis that $\delta_{ij} = 0$

Using this process a differential network was generated for the BWHHS data using the women with no prevalent CHD at baseline and a complete metabolite profile. The inclusion criteria for metabolites and individuals in the network was the same as in the PCR/lasso sections in chapter 4. with the outcome of interest specified as having a CHD event in the 12 year follow up, and those who died of a non-CHD cause are excluded, resulting in a sample size of 2922. There were 182 CHD events, so two networks were created, one using the 2740 women who did not have an event and the other using the 182 women who experienced an event. The same 78 metabolites were used as per the inclusion criteria in chapter 4. The differential network obtained is shown in figure 6.4, using a p-value of 0.01 as the threshold for edge inclusion as per the Valcarcel paper.

Each edge could represent a feature of the data that is associated with disease. So in figure 6.4 we see that 44 edges reach the p-value threshold of p <0.01, and investigating further the data that have given rise to these edges could provide some insight into the onset of the disease. For example, there is an edge between Valine (val) and Isoleucine (ile), so it may be that the relationship between these two metabolites could provide information about CHD risk. Also, nodes that have many edges could be important variables in this process because their association with a number of other variables changes across disease states. For example, the nodes with the highest degree are small VLDL triglycerides (s_vldl_tg), large LDL cholesterol esters (l_ldl_ce), small LDL free cholesterol (s_ldl_fc) and VLDL diameter (vldl_d) which all have a degree of 5, so it could be that these lipoprotein subclasses could have an important associations with CHD. However, in practice it is difficult to interpret what this network actually represents, as we don't have a detailed understanding about what an edge represents and the threshold of p <0.01 for edge inclusion is arbitrary. These results will be revisited in chapter 7.

6.2.1 Edge selection and interpretation

The Benjamini-Hochberg FDR adjustment (setting FDR to 5%) was also used to identify significant edges but yielded a network with no edges, suggesting that the power to detect edges in this study is very low (or there are no true edges). In the original Valcarcel paper an unadjusted threshold of p <0.01 was used, albeit in a smaller network size (60 nodes) and a larger sample size (4931). If we use this threshold we identify 44 edges in the differential network (which is the network illustrated in figure 6.4). So one issue identified with applying this method is a lack of power to detect differences in correlations. The power to detect edges is based on the overall sample size, and specifically on the sample

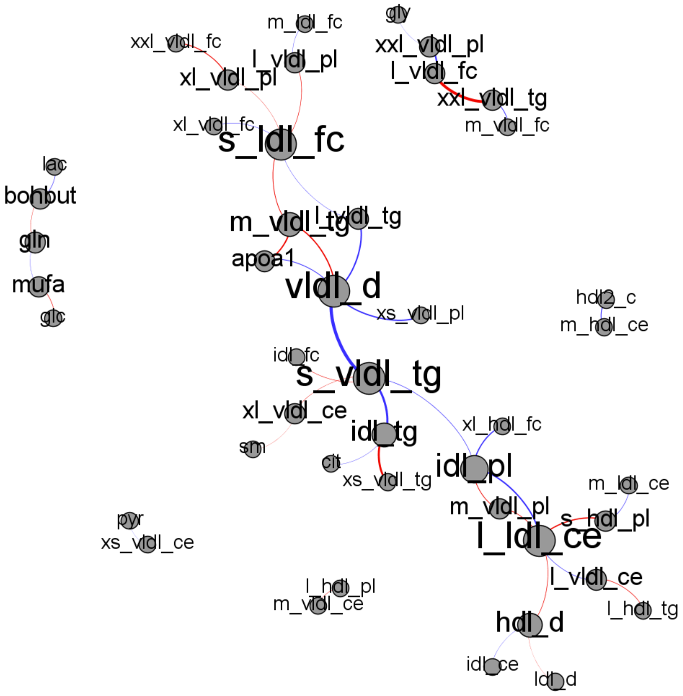


Figure 6.4: Differential network of 78 metabolites from the BWHHS defined using the difference in the Pearson partial correlation coefficient as a measure of association and a cut-off threshold for edge inclusion of p <0.01. Node size is relative to the degree of the node in the differential network, larger nodes indicate higher degree. Edge thickness is relative to $\widehat{\delta}_{ij}$. Nodes with no edges are excluded from the diagram.

size of the diseased and non-diseased groups. In this instance there were only 182 women in the diseased group, so the correlations between pairs of variables in that group are imprecisely estimated leading to low power in the final results.

A practical method used within the field of Gaussian graphical modelling to select edges is the SINful procedure [101, 102], where a conservative network is produced using the usual stringent threshold and another set of edges are identified, deemed as having an "intermediate" significance (defined by the user), so the more conservative network or the more liberal can be selected, depending on the which is more appropriate to the analysis being performed.

As well as the issue with power, there are questions about the interpretation of the results. What does a difference in correlation between a pair of variables mean? We can see there are a few pairs of nodes that appear to be differentially correlated, and it may be that this indicates a difference in the way these metabolites interact in diseased and non-diseased individuals. However, conceptually it is quite difficult to interpret what the meaning of an edge in a differential network is, even before going into what it might mean biologically. In the overall networks an edge was a measure of association between nodes, but in differential networks an edge carries several levels of information.

These two issues will be addressed in the rest of this chapter, where simulations will be run to

1. Generate data using a known model and investigate the results that applying differential networks to those data will obtain, with the purpose of aiding the interpretation of an edge
2. Investigate the sample sizes required to give the method sufficient power to detect true edges

6.3 Simulations

The literature to date makes little specific reference as to what a difference in the partial correlations represents, so the aim of this chapter is to simulate data that would arise from 3 different settings and in each of these describe the results obtained from a differential network analysis. Each of these settings has a set of nodes X_1, X_2, X_3, representing the metabolome, a node D representing a binary disease indicator, and, depending on the setting/scenario, nodes U and Z representing other possible causes of disease and potential sources of correlation of the metabolites.

The three scenarios considered are :

(A) X_1 and X_2 are joint causes of disease

(B) Disease modifies the joint distribution of X_1 and X_2

(C) There is a common cause of the disease and of the joint distribution of X_1 and X_2

These scenarios were chosen because they are the most likely data generating processes leading to associations between the metabolome and the disease.

6.3.1 Scenarios

6.3.1.1 Scenario A

This scenario is consistent with a cohort study where metabolomics data are measured at baseline, cohort members are followed up until a later point in time when a disease may or may have been diagnosed. We consider the metabolites

(X_1, X_2 and X_3) to be causes of disease (with potential effect modification existing between them) and to possibly be correlated because of a common set of causes **U**, which may also be a cause of D. We then consider a range of settings and investigate whether edges would be potentially identified in a differential network.

The use of differential networks in this context could help suggest causal pathways to disease, by identifying the difference in the associations between metabolites in those individuals that go on to develop a disease (cases) and those that do not (non-cases) by a selected time point. This model is depicted in figure 6.5, where letters represent variables and arrows causal effects.

At this point it is important to be clear that we are not proposing that identification of an edge in a differential network would imply that there was a causal effect, rather we are investigating the reverse, whether a causal effect from a pair of variables leads to identification of an edge in a differential network.

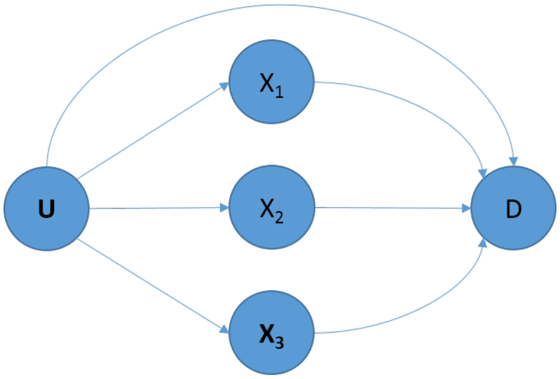


Figure 6.5: Hypothesized scenario A

6.3.1.2 Scenario B

Rather than X_1 and X_2 being causes of disease (together with X_3 and U), the relationship may be the other way round, the disease may influence the joint distribution of X_1 and X_2. More specifically, and for simplicity, we consider the setting where X_1 causes X_2 and X_3 and disease influences the strength of the $X_1 - X_2$ relationship.

This would correspond to a situation where there is a process in the body that influences an individual's level of a certain pair of metabolites. This process leads to these metabolites being correlated in a group of healthy individuals, however it may be that this association is distorted somehow by the presence of the disease with the possible extreme scenario being that the original correlation disappears in diseased individuals. In this case an edge between X_1 and X_2 may be identified in a differential network. This data generating model is shown in figure 6.6.

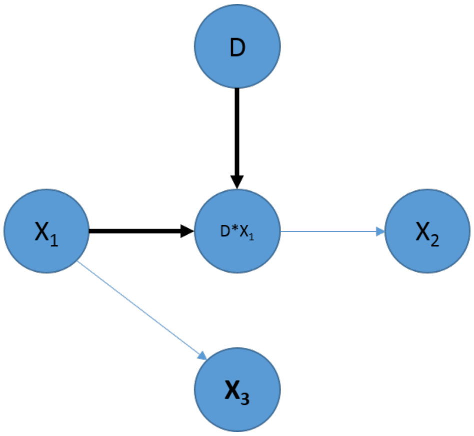


Figure 6.6: Hypothesized scenario B. The D*X_1 node represents the product of D and X_1, so when D=0 there is no association between X_1 and X_2 and when D=1 there is an association between X_1 and X_2.

6.3.1.3 Scenario C

In this scenario X_1 is a cause of X_2 and X_3, and there is an unmeasured variable Z which influences the association between X_1 and X_2 and also causes D. This scenario is very similar to B, but does not make the assumption that an individuals disease status is already set. This model is shown in figure 6.7.

In the following sections we will describe each of the scenarios in detail and then the steps used to simulate data corresponding to these scenarios, specifying the corresponding parametric models and then discuss which scenarios would lead to discovery of an edge (for a given sample size) in the differential network. The parameter values in each of the data generating models are specific to each scenario.

6.3.2 Methods

6.3.2.1 Data generation - Scenario A

We shall simulate data of this type based on the data generation model shown in figure 6.8. We take all variables involved to be univariate and sequentially

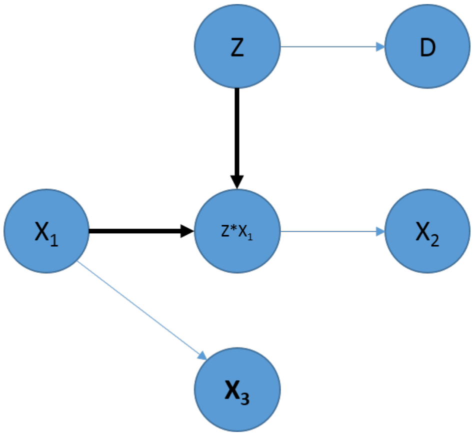


Figure 6.7: Hypothesized scenario C - The Z*X_1 node represents the product of Z and X_1, so when Z=0 there is no association between X_1 and X_2 and when Z=1 there is an association between X_1 and X_2

generate U, then X_1-X_3, then D. U is generated as standardized normal; X_j, for $j = 1, 2, 3$, are generated as normally distributed variables according to

$$X_j = \lambda_j U + e_j$$

where e_j is a normally distributed variable generated with mean 0 and standard deviation equal to $\sqrt{1 - {\lambda_j}^2}$. This choice of standard deviation guarantees that X_1-X_3 are also standardized normal (i.e. have mean 0 and standard deviation 1).

Since D is a binary variable we generate it according to a logistic distribution with

$$\begin{aligned} logit\{\text{Prob}\,(D = 1)\} = \alpha_D + \beta_u U + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \\ \gamma_{12} X_1 X_2 + \gamma_{13} X_1 X_3 + \gamma_{23} X_2 X_3 \end{aligned} \quad (6.1)$$

Given that all the explanatory variables in this equation are standardized normal, the intercept α_D represents the log-odds of disease when all variables are at their mean value.

In our simulations we will vary the values of λ_1, λ_2, λ_3, β_1, β_2, β_3, γ_{12}, γ_{13} and γ_{23}, then describe the results obtained from a differential network analysis in each setting, as well as investigating the impact of different sample sizes. It should be noted that these simulations only includes such a low number of metabolites (3) to aid the interpretation of the results (which is the aim of this chapter); there would be no reason in practice to use the differential network

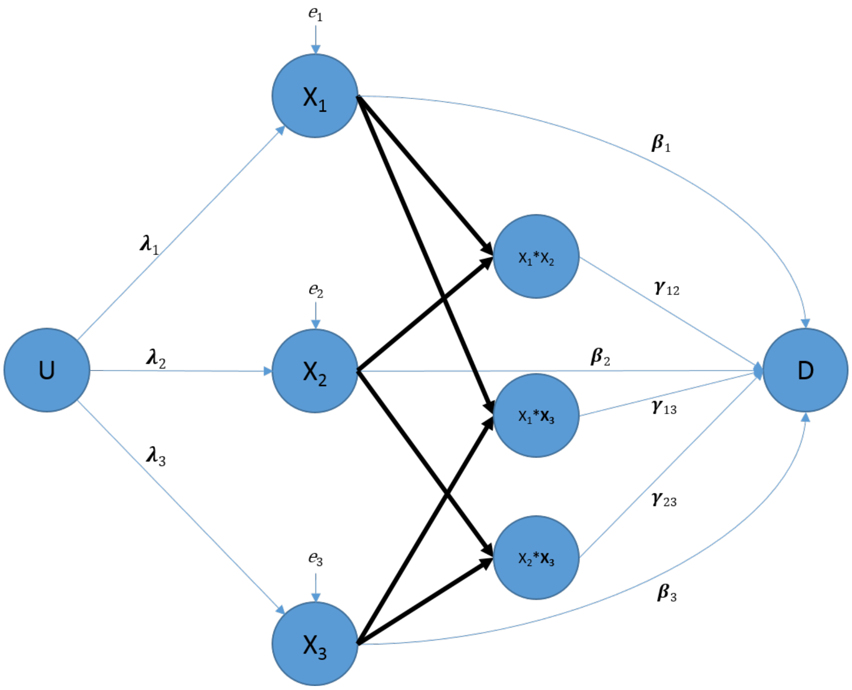


Figure 6.8: Data generation model, corresponding to scenario A. The bold black lines represent deterministic effects (because the nodes representing the interaction terms are the direct product of two random variables)

method with such a low number of variables, however the principles hold when extending the data generating model to include many more metabolites, which is a more realistic situation.

For a given simulated dataset of a selected sample size, we estimate the Pearson partial correlation of X_1 and X_2 within the cases and separately within the non-cases and then draw inference with regards to their difference, $\hat{\delta}_{12}$ as described in section 6.1. We do not use the shrinkage method in estimating the partial correlation coefficients, because we are only using 3 variables.

The steps we follow are:

1. Specify the parameters of the data generating model (i.e. λ_j, β_j and γ_{kl} for $j = 1, 2, 3$ and $kl = 12, 13, 23$) and the sample size
2. Generate realizations of the data generating model
3. Estimate δ_{12}

4. Calculate the p-value corresponding to a test of $\widehat{\delta}_{12} = 0$. The method for calculating the p-value is given in the section 6.3.2.4
5. Repeat steps 1-4 500 times
6. Count how many times the p-value is less than 0.01. This critical value was selected to reflect the p-value used in the original differential network paper by Valcarcel et al [71].

Primarily we will investigate plausible ranges for the data generating parameters, however we will also consider more extreme, unrealistic scenarios, as this can aid our understanding as to what is happening. The plausible ranges of values used in the simulations are:

- α_D will range from -4 to 0 (equivalent to an expected risk of disease of 1.7% to 50%, when all other variables = 0)
- β_1 and β_2 will range from -0.5 to 0.5 (equivalent to odds ratios of 0.6-1.6), moving in steps of 0.05
- γ_{12} will range from -0.5 to 0.5 (equivalent to odds ratios of 0.6-1.6), moving in steps of 0.05
- the sample size N will be set to 10,000 unless specified otherwise

Note that in these simulations X_3 represents all other possible metabolites in the model. If X_3 is not associated with X_1 and X_2 at all, then the partial correlation of X_1 and X_2 will be equal to their marginal correlation. If X_3 is strongly associated with both X_1 and X_2 then the partial correlation of X_1 and X_2 will be substantially different from their marginal correlation.

6.3.2.2 Data Generation - Scenario B

We shall simulate data of this type based on the data generation model shown in figure 6.9. Again D is the disease indicator (taking a value of 0 for non-cases and 1 for cases), however the proportion of individuals who have $D = 1$ is set first using the very simple generating model.

$$logit\{\text{Prob}\,(D = 1)\} = \alpha_D \tag{6.2}$$

X_1 is generated as a standardized normal variable then X_2 is generated as:

$$\begin{aligned} X_2 &= \lambda_1 X_1 + e_2 \qquad \text{if } D = 0 \\ X_2 &= e_2 \qquad\qquad\quad\; \text{if } D = 1 \end{aligned} \tag{6.3}$$

where e_2 is a normally distributed variable generated with mean 0 and standard deviation equal to $\sqrt{1 - {\lambda_1}^2}$ if $D = 0$ and equal to 1 if $D = 1$. This choice of standard deviation guarantees that X_2 is also standardized normal within each disease group. This in effect means if an individual has $D = 0$ their value of X_2 is picked at random from a distribution associated with X_1, however if $D = 1$ their value of X_2 is picked at random from a standard normal distribution.

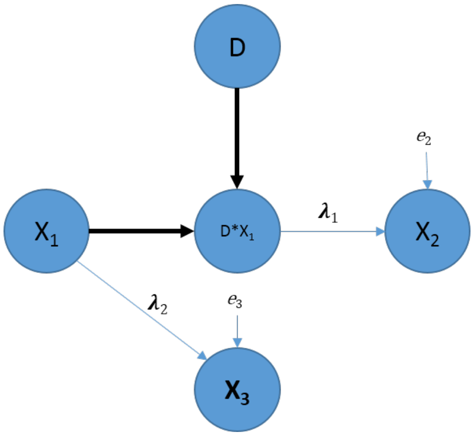


Figure 6.9: Data generation model for scenario B. The bold black lines represent deterministic effects (because the nodes representing the interaction term are the direct product of two random variables)

As before X_3 is a variable representing all other metabolites, it is associated with X_1 (and hence X_2) and is generated as:

$$X_3 = \lambda_2 X_1 + e_3 \tag{6.4}$$

where e_3 is a normally distributed variable generated with mean 0 and standard deviation equal to $\sqrt{1 - \lambda_2{}^2}$. When simulating this model, and model C, the process for analysing the results will be the same as for model A.

The ranges of values for the parameters will be:

- λ_1 will range between 0 - 0.8 in steps of 0.04
- λ_2 will be set to 0 and 0.8
- Total sample size will be set to 500, 1000 and 10,000
- Number of cases will be set to 50, 100 and 500

6.3.2.3 Data Generation - Scenario C

Scenario C is very similar to B, except that rather than D itself switching on and off the path from X_1 to X_2, a binary indicator Y does this, with the distribution of Y depending on a continuous variable Z which is also a cause of D, with the relationships described as follows:

$$\begin{aligned} logit\{\text{Prob}\,(Y=1)\} &= \alpha_Y + \beta_Y Z \\ logit\{\text{Prob}\,(D=1)\} &= \alpha_D + \beta_D Z \end{aligned} \tag{6.5}$$

Due to the similarities between scenario B and C, scenario C will only be briefly investigated. As before X_3 is a variable representing all other metabolites, it is associated with X_1 (and hence X_2) and is generated as:

$$X_3 = \lambda_2 X_1 + e_3 \tag{6.6}$$

where e_3 is a normally distributed variable generated with mean 0 and standard deviation equal to $\sqrt{1-\lambda_2{}^2}$. So in practice the data generating models B and C are similar, although in B the disease "switches off" the association between X_1 and X_2 and it precedes it, whereas in C there is no temporal ordering and the relationship is associational, those with the disease are *more likely* to have the association between X_1 and X_2 "switched off" (figure 6.10).

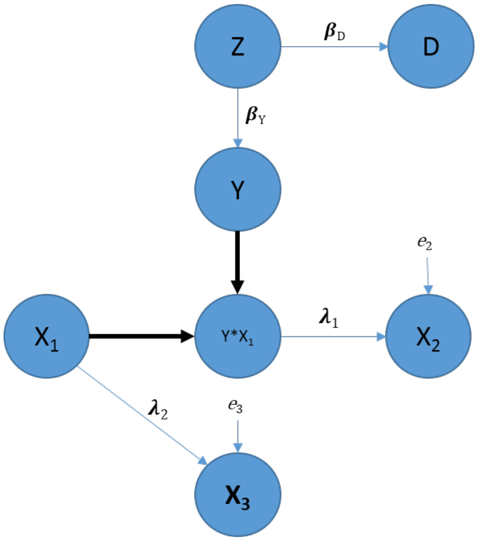


Figure 6.10: Data generation model for scenario C. The bold black lines represent deterministic effects (because the nodes representing the interaction term are the direct product of two random variables)

The same steps and sample sizes as those described for the simulations of scenario A are followed here with parameters taking values:

- λ_1 will range between 0 - 0.8 in steps of 0.04
- λ_3 will be set to 0 and 0.8

- Sample size will be set to 10,000
- α_Y and α_D will be set to -2 and -3
- β_Y and β_D will be set to 1 and 2

6.3.2.4 A note on p-values

Although the method we are proposing using for the simulations is that specified by Valcarcel et al [71], rather than use their permutation method of calculating the p-value for each differential edge (which is computationally intensive) we are going to apply a Z-test to the Fisher transformed partial correlations [90] because of the number of simulations we perform. So the partial correlation (r_{ijk}) calculated in each disease group needs to be transformed using the following function:

$$z_{ijk} = \frac{1}{2} ln \left| \frac{1 + r_{ijk}}{1 - r_{ijk}} \right| \tag{6.7}$$

where k indicates the subgroup (cases = 1, non-cases = 0) then the Z-statistic is found:

$$Z_{ij} = \frac{|z_{ij1} - z_{ij0}|}{\sqrt{\frac{1}{n_0 - 3} + \frac{1}{n_1 - 3}}} \tag{6.8}$$

where n_0 and n_1 are the sample sizes of the two subgroups. The Z-statistic is then compared to the normal cumulative distribution function to obtain the corresponding p-value.

Using the asymptotic distribution of the differences to obtain p-values for each edge makes the simulation study much more efficient than if the permutation method is used, and given the time and computing constraints the number of simulations performed could not have been done using the permutation method. However, using the asymptotic distribution requires the underlying assumption of bivariate normality between each pair of variables and is also sensitive to large outliers. However, in the simulation study the data are generated as normally distributed so this is not such an issue. Therefore the simulations looking at the power of the method are based on the asymptotic distribution. The permutation method can be used when we are less confident that the assumptions will hold, however the power of this method will be reduced, compared to the power observed in the simulations in this chapter.

6.4 Results

For each scenario, 500 simulations were run with results showing mean estimates from the 500 simulations.

6.4.1 Data generating model A

In this section we describe the results from the simulations performed using data generating model A, starting with the simplest setting, when D is generated only as a function of X_1 and X_2 without interaction (i.e. $\beta_3 = \beta_U = \gamma_{12} = \gamma_{13} = \gamma_{23} = 0$) and the sample size is fixed at 10,000. After this we will investigate the effect of including an interaction between X_1 and X_2 (i.e. $\gamma_{12} \neq 0$), before finally moving on to investigating the effect of sample size and disease prevalence on the results. The effect of varying β_3 and λ_3 is not examined in detail in this chapter.

6.4.1.1 Metabolites are direct causes of disease (no interaction)

First we consider the situation where either X_1 or X_2 or both X_1 and X_2 are causes of disease. We vary the strength of their causal effect by altering β_1 and β_2. To simplify matters, for the time being β_u, β_3 and λ_3 are set to 0. So in these scenarios we are considering the outcome to be generated by:

$$logit\{\text{Prob}\,(D = 1)\} = \alpha_D + \beta_1 X_1 + \beta_2 X_2 \tag{6.9}$$

So the only parameters that we will be varying at this point are λ_1 and λ_2, which affect the correlation of the metabolites, and β_1 and β_2, with β_i representing the increase in log odds of disease for an increase of 1 standard deviation in the concentration of X_i and λ_i determining the strength of the association between U and X_i. We start with a fixed sample size of 10,000 observations and a fixed value of -2 for α_D (α_D represents the log odds of disease when the concentrations of all metabolites are equal to their mean value), which relates to a risk of disease of 12% amongst individuals who have a mean value of each metabolite. This is slightly higher than, but of a similar order to, the proportion of CHD cases in the BWHHS dataset of 5%.

X_1 and X_2 are uncorrelated We started by considering the setting where X_1 and X_2 are uncorrelated, normally distributed metabolites (λ_1 and λ_2 both are set equal to zero). At each level of β_1 and β_2, ranging from -0.5 to 0.5 in increments of 0.05. We also simulated data in the range -3 to 3 to get a view as to what happens at more extreme values.

Figure 6.13a shows a contour plot illustrating the magnitude and direction of the mean estimate of $\widehat{\delta}_{12}$ (to be given the notation $\overline{\widehat{\delta}}_{12}$) from the simulations, for a range of values of β_1 and β_2. The white area represents where $\overline{\widehat{\delta}}_{12}$ is close to between -0.05 and 0.05, i.e. close to zero. The blue area represents a negative $\overline{\widehat{\delta}}_{12}$ less than -0.05 with the intensity increasing as the magnitude of $\overline{\widehat{\delta}}_{12}$ increases. The red area represents a positive $\overline{\widehat{\delta}}_{12}$ greater than 0.05, again with the intensity increasing as the magnitude increases.

The first thing to note is, within the ranges β_1 = -0.5,0.5 and β_2 = -0.5,0.5, the plot is white, so within the plausible region defined in the methods section $\bar{\hat{\delta}}_{12}$ remains at (or very close to) 0, whether X_1 and or X_2 are causes of D. However, to aid our understanding of the method we have widened the plot to include values beyond the plausible range, up to β_1 and $\beta_2 = 3$. It is possible to see from the plot that no matter how large an effect X_1 or X_2 has on D, if only 1 of them is a cause of D the value of $\bar{\hat{\delta}}_{12}$ remains equal to, or close to, 0. However when both X_1 and X_2 are strong causes of disease, when β_1 and β_2 reach values of about 1, we begin to see $\bar{\hat{\delta}}_{12}$ moving away from 0, with $\bar{\hat{\delta}}_{12}$ going negative when β_1 and β_2 are the same sign and positive when β_1 and β_2 are different signs.

We can explain why this happens by looking closely at one instance of the data. When X_1 and X_2 are uncorrelated their distribution looks like that in figure 6.11a-d. In figure 6.11a neither X_1 and X_2 are associated with disease, so the cases (depicted in red) and controls (blue) are distributed evenly among the observations. So the overall correlation is approximately equal to 0, and the correlation in the cases will be approximately 0 and the correlation in the controls will also be approximately 0. So there is no difference in the correlations, leading to a $\bar{\hat{\delta}}_{12}$ of approximately 0. As we move from figure 6.11a through to figure 6.11d, the strength of association between X_1 and D increases, which can be seen in the figures as the cases appear more in the upper half of the plots. However, the correlation of X_1 and X_2 remains at 0 both within the cases and within the non-cases. This is what was observed in the simulations, if X_1 and X_2 are uncorrelated (ignoring disease status) and only 1 of them was associated with D, then no matter how strong that association was, $\bar{\hat{\delta}}_{12}$ would remain equal to 0.

We now look at the example when X_1 and X_2 are both positively associated with D, as illustrated in figure 6.12. We can see that as the coefficients β_1 and β_2 increase, the cases move into the upper right quadrant of the scatter plot(fig. 6.12d). This results in a negative correlation between X_1 and X_2 within the cases. At the same time, the correlation within the non-cases also goes negative, since there are fewer non-cases in the upper right quadrant. However, the negative correlation induced in the non-cases is smaller in magnitude than the negative correlation in the cases, and remembering δ_{12} is defined as the partial correlation of X_1 and X_2 in the cases minus the partial correlation in the non-cases, therefore $\bar{\hat{\delta}}_{12}$ has a negative value.

So far it has been illustrated how the value of $\bar{\hat{\delta}}_{12}$ is affected by the values of β_1 and β_2. We are also interested in testing whether the estimated $\widehat{\delta}_{12}$ is different from 0, which can be done by testing the null hypothesis of $\delta_{12} = 0$ against the alternative hypothesis of $\delta_{12} \neq 0$ using the method described in section 6.3.2.4.

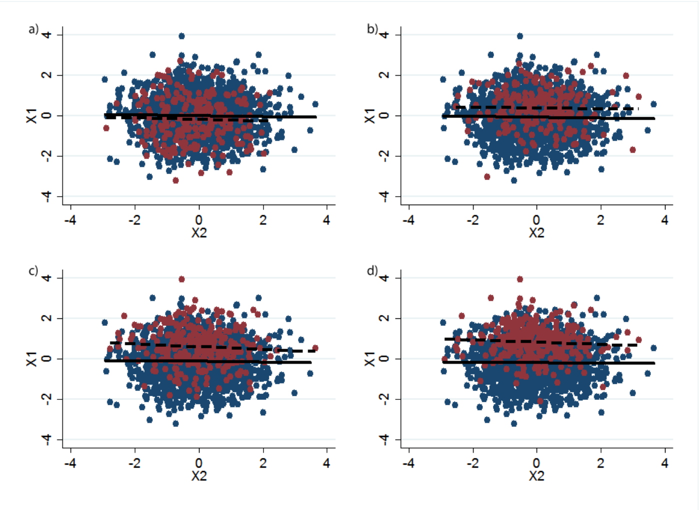


Figure 6.11: Four scatter plots from a single randomly selected simulation where cases = red non-cases =blue. In each plot the observations of X_1 and X_2 are the same, but the generation of D differs because it is dependent on the values of β_1 and β_2. In a) $\beta_1 = 0$, $\beta_2 = 0$ b) $\beta_1 = 0.4$, $\beta_2 = 0$ c) $\beta_1 = 0.8$, $\beta_2 = 0$ d) $\beta_1 = 1.2$, $\beta_2 = 0$. The solid line is the line of best fit relating X_1 to X_2 in the non-cases, the dashed line in the cases. $\alpha_D = -2$.

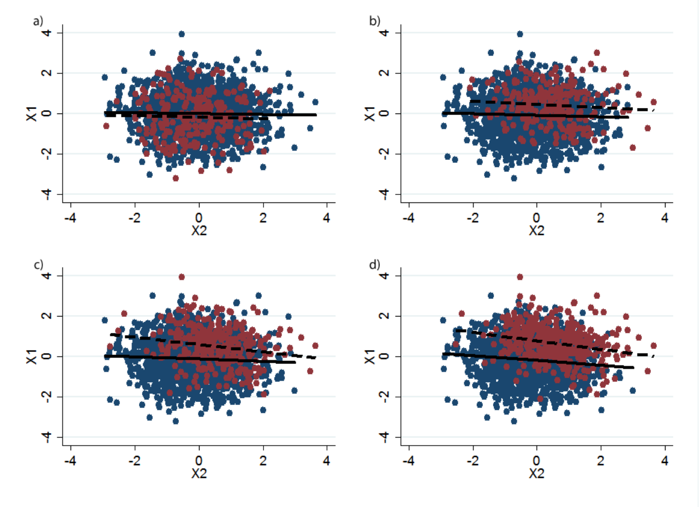


Figure 6.12: Four scatter plots from a randomly selected simulation where cases = red non-cases =blue. In each plot the observations of X_1 and X_2 are the same, but the generation of D differs because it is dependent on the values of β_1 and β_2. In a) $\beta_1 = \beta_2 = 0$ b) $\beta_1 = \beta_2 = 0.4$ c) $\beta_1 = \beta_2 = 0.8$ d) $\beta_1 = \beta_2 = 1.2$. The solid line is the line of best fit relating X_1 to X_2 in the non-cases, the dashed line in the cases. $\alpha_D = -2$.

Figure 6.13b is another contour chart, this time illustrating the proportion of simulations that yielded a p-value of less than 0.01 from the above hypothesis test. Once both β_1 and β_2 had an absolute magnitude of 1 (odds ratio 2.7 or 0.37) then about 50% of simulations resulted in a p-value less than 0.01. Once both β_1 and β_2 have an absolute magnitude of 1.5 (odds ratio 4.5 or 0.22), almost 100% of simulations yielded a p-value of <0.01. So it takes an extremely large causal effect from both variables to induce a $\widehat{\delta}_{12}$ that is different from 0, at the significance level of 1%.

X_1 and X_2 are correlated Now if we look at the situation where X_1 and X_2 are correlated (i.e. both λ_1 and λ_2 are non-zero, therefore inducing a correlation between X_1 and X_2) we can plot the same surface chart for the same ranges of β_1 and β_2 as before. Figure 6.13c shows the values of $\overline{\widehat{\delta}}_{12}$ across the ranges of β_1 and β_2 when the partial correlation of X_1 and X_2 (ignoring case status) is 0.49 and in figure 6.13e where this partial correlation is 0.81. As before, blue represents a negative $\overline{\widehat{\delta}}_{12}$, white a $\overline{\widehat{\delta}}_{12}$ of close to 0 and red a positive $\overline{\widehat{\delta}}_{12}$. Again, in the plausible range of values for β_1 and β_2, $\overline{\widehat{\delta}}_{12}$ remains close to 0.

However, it is possible to see in the wider plot that the situations that give rise to a $\overline{\widehat{\delta}}_{12}$ different from 0 changes, now that X_1 and X_2 are correlated. If β_1 or β_2 are of opposite signs we do not see a positive $\overline{\widehat{\delta}}_{12}$ arising, instead the value of $\overline{\widehat{\delta}}_{12}$ remains close to 0. When β_1 and β_2 are the same sign, we find a negative value of $\overline{\widehat{\delta}}_{12}$ (as before). Also, in this situation compared with the uncorrelated case, smaller values of β_1 or β_2 are required to induce the same $\overline{\widehat{\delta}}_{12}$. It is also possible, if β_1 or β_2 were large enough, that a non-zero $\overline{\widehat{\delta}}_{12}$ could be obtained if only one of X_1 or X_2 were a cause of D. For example, inspecting the plot in figure 6.13e, when $\beta_1 = 0$ once β_2 moves above 1 or below -1 a negative value of $\overline{\widehat{\delta}}_{12}$ is induced, represented by the light blue shading.

We can again explain these results using one instance of the generated data, where figures 6.14a-d are obtained with correlated X_1 and X_2 (with a partial correlation of 0.81 obtained by setting λ_1 and $\lambda_2 = 0.9$). In figure 6.14a, both β_1 and β_2 are set equal to 0, resulting in the partial correlation of X_1 and X_2 within the cases and within the controls both being approximately equal to the partial correlation in the whole dataset, therefore resulting in a $\widehat{\delta}_{12}$ of 0. As X_1 becomes a stronger cause of disease (by increasing β_1) the observations in the upper half of the plot are more likely to be cases, and since X_1 and X_2 are so strongly correlated this leads to the cases being constrained into the top right quadrant of the plot. So as β_1 increases, the partial correlation within the cases reduces, but the partial correlation in the non-cases remains close to the overall partial correlation of 0.81. This leads to a negative value of $\widehat{\delta}_{12}$. However it should be noted that it is only in situations where the strength of the association

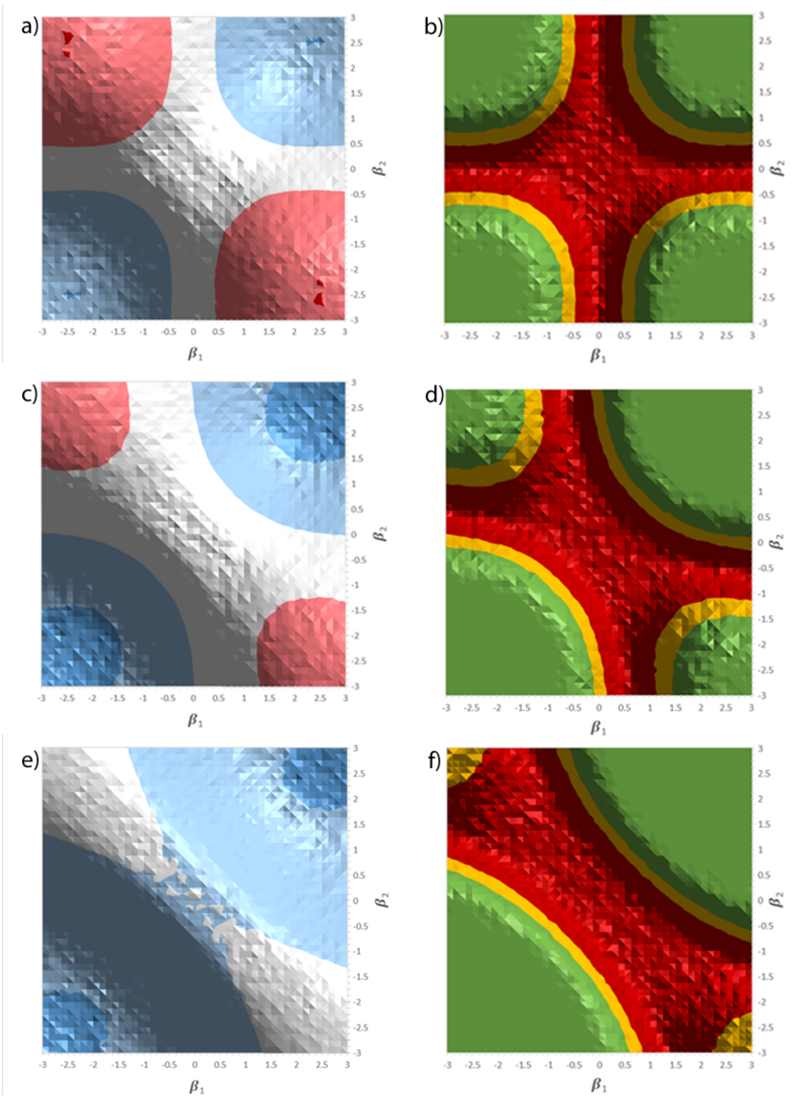


Figure 6.13: Each of the 3 contour plots on the left hand side represents $\bar{\bar{\hat{\delta}}}_{12}$ across range from -3 to 3 of β_1 and β_2. Red areas represent a positive $\bar{\bar{\hat{\delta}}}_{12}$, blue a negative value. The 3 contour plots on the right hand side represent the proportion of simulations that result in a p-value <0.01 across the range from -3 to 3 of β_1 and β_2. In the top plots X_1 and X_2 are uncorrelated, in the middle plots the correlation of X_1 and X_2 is 0.49 and in the bottom plots it is 0.81. Green areas represent greater than 80% of simulations resulting in a p-value <0.01, yellow 40-80% and red <40%. The sample size for each simulation was N=10,000. The number of simulations at each point was 500. $\alpha_D = -2$.

between X_1 and D is extremely strong and the partial correlation between X_1 and X_2 is very high that we would observe this occurring.

When both X_1 and X_2 are causes of D then the logic is very similar to the above example, except the effect is more pronounced. As β_1 and β_2 increase, the cases move into the upper right quadrant, with the partial correlation of X_1 and X_2 within the cases reducing from the overall partial correlation, and the partial correlation in the non-cases remaining steady (or reducing at a slower rate). Again this results in negative value of $\widehat{\delta}_{12}$.

Once both β_1 and β_2 had, in absolute terms, a magnitude of 1 (odds ratio 2.7 or 0.37) then about 50% of simulations resulted in a p-value less than 0.01. If we again plot a contour chart of the proportions of simulations that lead to a p-value for $\widehat{\delta}_{12}$ of less than 0.01 we get a different picture to the uncorrelated setting. Figure 6.13d illustrates the results from simulations where X_1 and X_2 have a positive partial correlation of magnitude 0.49 ($\lambda_1 = \lambda_2 = 0.7$). The values of $\widehat{\delta}_{12}$ are lower when the variables are correlated, but the threshold at which a p-value is deemed to be different from zero is also lower. Again, in the range of plausible values from -0.5 to 0.5 almost no simulations result in a p-value for $\widehat{\delta}_{12}$ that is less than 0.01. However for more extreme values of β_1 and β_2 we observe a different pattern to before. Where in the setting with X_1 and X_2 were uncorrelated, 50% of simulations yielded a p-value <0.01 when $\beta_1 = \beta_2 = 1$, now 90% of the simulations result in this. However, in the previous uncorrelated example, the results were completely symmetrical, so when $\beta_1 = 1$ and $\beta_2 = -1$ we still saw 50% of simulations resulting in an edge detection, now we see only 3%. So as the partial correlation of X_1 and X_2 increases, we see a $\widehat{\delta}_{12}$ more readily picked up as being different from 0 if X_1 and X_2 are both positively or both negatively associated with D, but we see a reduction in edges detected if X_1 and X_2 affect D in opposite directions.

The effect is even more marked when we increase the values of λ_1 and λ_2 to 0.9, resulting in a partial correlation of 0.81 between X_1 and X_2 in the overall dataset (figure 6.13f). In this situation we see 99% of simulations resulting in a p-value for $\widehat{\delta}_{12}$ of less than 0.01 when $\beta_1 = \beta_2 = 1$. We can also see another effect now, if the magnitude of either one of β_1 or β_2 is large enough (≈ 2), then this will lead to detection of an edge, even if the other coefficient is equal to 0. The impact of altering α_D is discussed in section 6.4.1.4.

6.4.1.2 Metabolites are direct causes of disease (with interaction)

Here we describe the results from simulations where the association between X_1 and D is modified by X_2 (and vice versa), γ_{12} defines the strength of this modification.

$$logit\{\text{Prob}\,(D = 1)\} = \alpha_D + \beta_1 X_1 + \beta_2 X_2 + \gamma_{12} X_1 X_2 \tag{6.10}$$

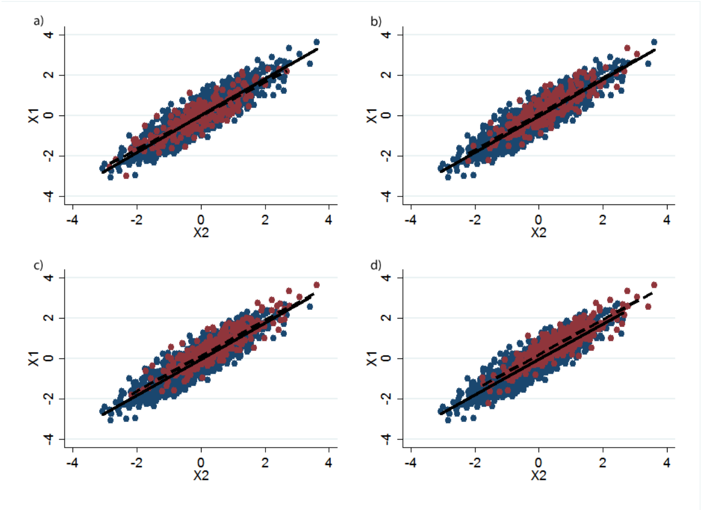


Figure 6.14: Four scatter plots from a randomly selected simulation where X_1 and X_2 are correlated, with a partial correlation of 0.81. Cases = red non-cases =blue. In each plot the observations of X_1 and X_2 are the same, but the generation of D differs because it is dependent on the values of β_1 and β_2. In a) $\beta_1 = 0$, $\beta_2 = 0$ b) $\beta_1 = 0.4$, $\beta_2 = 0$ c) $\beta_1 = 0.8$, $\beta_2 = 0$ d) $\beta_1 = 1.2$, $\beta_2 = 0$. The solid line is the line of best fit relating X_1 to X_2 in the non-cases, the dashed line in the cases. $\alpha_D = -2$.

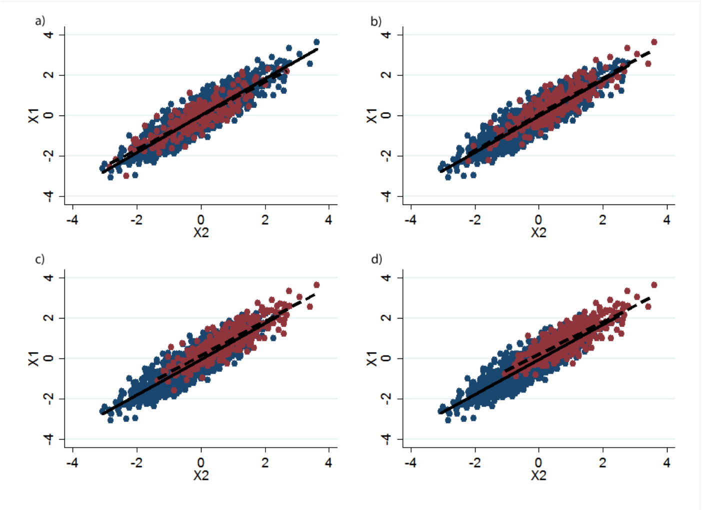


Figure 6.15: Four scatter plots from a randomly selected simulation simulation where X_1 and X_2 are correlated, with a partial correlation of 0.81. Cases = red non-cases =blue. In each plot the observations of X_1 and X_2 are the same, but the generation of D differs because it is dependent on the values of β_1 and β_2. In a) $\beta_1 = \beta_2 = 0$ b) $\beta_1 = \beta_2 = 0.4$ c) $\beta_1 = \beta_2 = 0.8$ d) $\beta_1 = \beta_2 = 1.2$. The solid line is the line of best fit relating X_1 to X_2 in the non-cases, the dashed line in the cases. $\alpha_D = -2$.

However, initially, to aid clarity we set β_1 and β_2 equal to 0. We shall again fix α_D to be -2 in these analyses. So the only parameters that we will be varying at this point are λ_1 and λ_2, which affect the correlation of the metabolites, and γ_{12} which determines the strength of the effect of the interaction of X_1 and X_2 on the outcome. As before we will use a sample size of 10,000 observations and 500 sets of simulations.

X_1 and X_2 are uncorrelated When X_1 and X_2 are uncorrelated (λ_1 = λ_2 = 0) in the overall dataset, the difference in the partial correlations in the cases and non-cases ($\widehat{\delta}_{12}$) is equal to the interaction term (γ_{12}) within the range -0.2 <(γ_{12}) <0.2, as seen by the red line on the chart in figure 6.16. As the correlation of X_1 and X_2 increases, the estimated $\widehat{\delta}_{12}$ shrinks for the same value of γ_{12} but the relationship is still linear within this range, as seen by the red line in figure 6.16. The reason that the linearity only holds within this range is because δ_{12} is constrained to be between 2 and -2 so as δ_{12} approaches the boundaries the linear relationship will not hold.

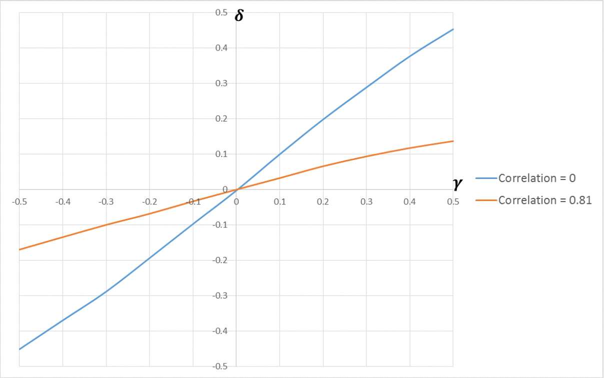


Figure 6.16: Chart showing the estimated value of δ_{12} against the value of γ_{12} for the situation where X_1 and X_2 are uncorrelated (blue) and when the partial correlation of X_1 and X_2 in the overall dataset is 0.81 (red) (Sample size N=10,000)

As with the previous examples, we will describe how the data appear when an interaction is included and how that affects the values of $\widehat{\delta}_{12}$. Figure 6.17a shows the data when γ_{12} is set equal to 0, which, since the correlation of X_1 and X_2 within the cases and the controls is 0, leaves $\widehat{\delta}_{12} = 0$. As γ_{12} increases positively, the cases are more likely to be found in the upper right and lower left quadrants, leading to a positive correlation within the cases. The non-cases are more likely to appear in the upper left and lower right quadrants, so will become slightly negatively correlated. These together result in a positive value of $\widehat{\delta}_{12}$.

X_1 and X_2 are correlated In the situation where X_1 and X_2 are correlated, the logic is very similar to the last example. In this scenario, when $\gamma_{12} = 0$ then X_1 and X_2 are correlated within both the cases and non-cases, but this correlation is the same in both groups so again $\widehat{\delta}_{12}$ is 0 (figure 6.18a) . As above, as γ_{12} increases positively (figures 6.18b-d) the correlation of X_1 and X_2 within the cases increases, and the correlation of X_1 and X_2 within the non-cases decreases slightly. However, this happens to a lesser extent than if the data were uncorrelated, so for the same value of γ_{12}, $\widehat{\delta}_{12}$ is smaller than compared with the scenario where X_1 and X_2 are uncorrelated, as is seen in the chart in figure 6.16.

In the uncorrelated scenario, a γ_{12} of 0.1 (equivalent to an odds ratio of 1.1) is sufficient to result in 75% of simulations having a p-value <0.01, with a γ_{12} of 0.15 (equivalent OR = 1.16) resulting in 99% of simulations with p <0.01. Although as the correlation increases, the estimated value of δ_{12} decreases for the same value of γ_{12}, the proportion of p-values for $\widehat{\delta}_{12}$ that fall below 0.01 is unaffected. So an edge is detected between a pair of variables if the interaction of the pair of variables is a cause of D. A plot of the proportion of simulations that resulted in p <0.01 is shown in figure 6.19, with one line each representing 3 levels of partial correlation of X_1 and X_2 (0, 0.49 and 0.81). The three lines all appear on top of one another, so the proportion of simulations identifying a value of δ_{12} with statistically significant (at the 1% level) difference from 0 is independent of the correlation of X_1 and X_2 in the overall dataset.

6.4.1.3 Summary

In the previous two sections it was shown that a significant difference in partial correlations in cases and non-cases is likely to be found when:

- Either X_1 or X_2 are extremely strong causes of disease and are partially correlated
- Both X_1 and X_2 are strong causes of disease
- The effect of each of the metabolites on disease is modified by the other (i.e. there is an interaction between X_1 and X_2)

There were some variants around these three situations, that depened on the number of cases and the partial correlation between X_1 and X_2, but broadly these are the three settings that lead to a significant $\widehat{\delta}_{12}$.

However, we only considered effect modification when the main effects were set to 0, a more realistic scenario would be when an interaction effect and main effects exist then this can lead to difficulties.

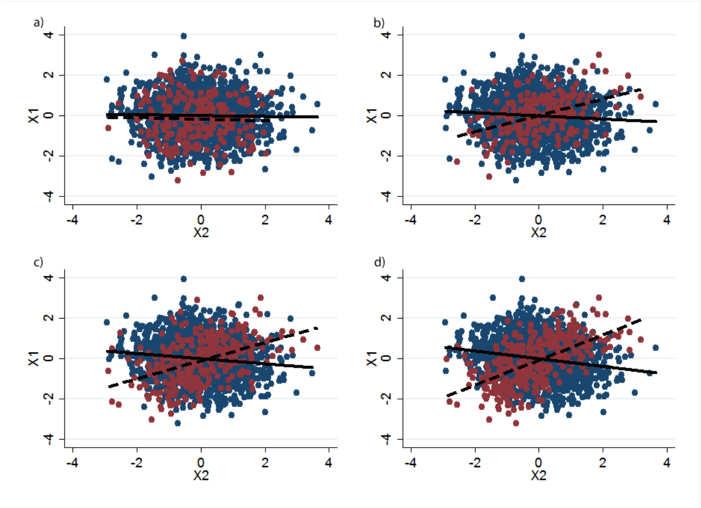


Figure 6.17: Four scatter plots from a randomly selected simulation where X_1 and X_2 are uncorrelated. Cases = red non-cases =blue. In each plot the observations of X_1 and X_2 are the same, but the generation of D differs because it is dependent on the values of β_1 and β_2. In a) $\gamma_{12} = 0$ b) $\gamma_{12} = 0.4$ c) $\gamma_{12} = 0.8$ d) $\gamma_{12} = 1.2$. In all, $\beta_1 = \beta_2 = \beta_3 = \gamma_{13} = \gamma_{23} = 0$. The solid line is the line of best fit relating X_1 to X_2 in the non-cases, the dashed line in the cases. $\alpha_D = -2$.

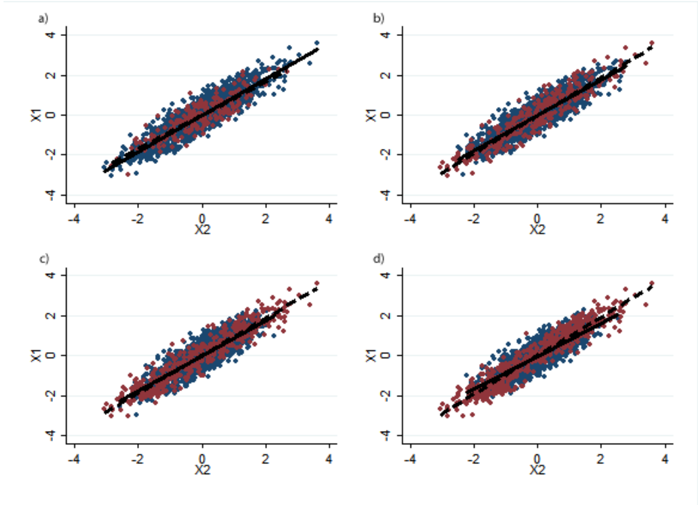


Figure 6.18: Four scatter plots from a randomly selected simulation where X_1 and X_2 are correlated, with a partial correlation of 0.81. Cases = red non-cases =blue. In each plot the observations of X_1 and X_2 are the same, but the generation of D differs because it is dependent on the values of β_1 and β_2. In a) $\gamma_{12} = 0$ b) $\gamma_{12} = 0.4$ c) $\gamma_{12} = 0.8$ d) $\gamma_{12} = 1.2$. In all, $\beta_1 = \beta_1 = 0$. The solid line is the line of best fit relating X_1 to X_2 in the non-cases, the dashed line in the cases. $\alpha_D = -2$.

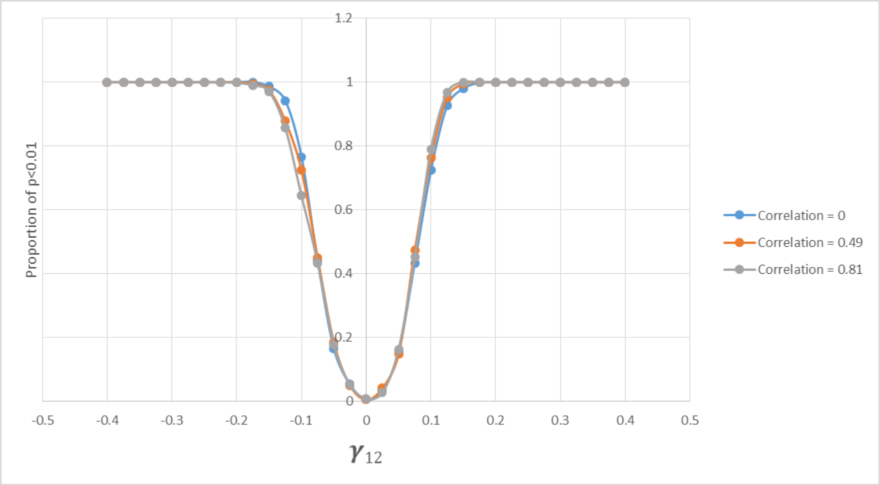


Figure 6.19: Line chart plotting proportion of p-values for testing $\delta_{12} = 0$ that are less than 0.01 against the value of γ_{12} used in generating the data. Sample size N=10,000, number of simulations = 500, $\beta_1 = \beta_2 = 0$, $\alpha_D = -2$.

If X_1 and X_2 are risk factors for disease, (i.e. a higher value leads to an increased risk of being a case, β_1 and β_2 are positive) then this results in a negative value of $\widehat{\delta}_{12}$ (correlation in the non-cases is greater than in the cases). If the interaction term γ_{12} is positive, this leads to a positive value of $\widehat{\delta}_{12}$. So if all three of β_1, β_2 and γ_{12} are positive this can lead to a canceling effect. In a situation where we may have identified a p-value for $\widehat{\delta}_{12}$ of less than 0.01 due to an interaction in the absence of main effects, we might not pick up this effect if β_1 and β_2 are large and positive as this would "dilute" the effect that the interaction has on the difference in partial correlations.

In fact, if the main effects were strong enough, in comparison to the interaction term, there would be a point where the difference would "cross over" and a $\widehat{\delta}_{12}$ of an opposite sign would appear.

6.4.1.4 Proportion of cases

Varying the value of α_D when generating the data for the simulations changes the underlying prevalence of the disease, D. The parameter α_D is equal to the *logit* of the probability that an individual, with a value of 0 for all other causal variables, will have the disease. Until now we have fixed α_D to -2, however the prevalence of disease is an important factor in the determination of differential correlations, as if one group is particularly small then the partial correlation within that group will be imprecisely estimated.

To illustrate the effect of varying α_D we will examine two model specifications. First we fix γ_{12} to 0.1, with $\beta_1 = \beta_2 = 0$, which was found to result in approximately 75% of simulations giving a p-value of less than 0.01 for testing the hypothesis that $\delta_{12} = 0$ (section 6.4.1.2, figure 6.18).

The proportions of p-values that are less than 0.01 at different levels of α_D against the number of cases in the sample is shown in figure 6.20 (from a sample size of 10,000 and 500 simulations). When the number of cases is at 500 (5%) approximately 30% of simulations result in a p-value <0.01. As the number of cases increases towards 5000 (i.e. α_D increases towards 0), the proportion of p-values also increases, with 95% of simulations resulting in p-values <0.01 once the number of cases reaches 2300 (i.e. α_D = -1.2 leading to a disease prevalence of 23% among those with mean levels of X_1 and X_2). This suggests as the size of the smaller group (the cases in these simulations) increases, the power of the method also increases, and this association is independent of the marginal correlation between X_1 and X_2.

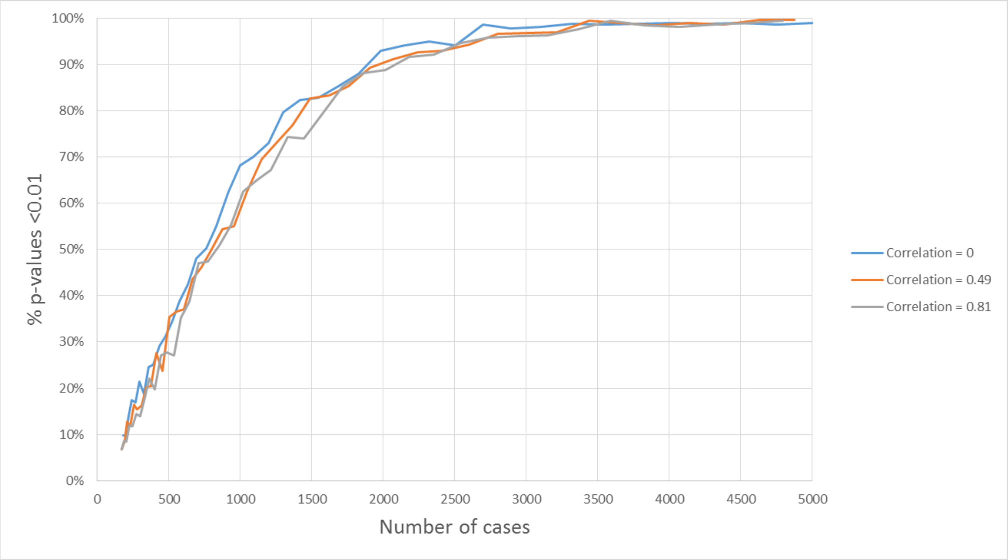


Figure 6.20: Line chart plotting proportion of p-values for the hypothesis test of δ_{12}=0 that are less than 0.01 against the number of cases (from a sample of 10,000). $\gamma_{12} = 0.1$, $\beta_1 = \beta_2 = 0$. α_D is varied to alter the number of cases.

It should be noted, the figure only shows the values up to $\alpha_D = 0$ (equivalent to number of cases = 5000), the picture is symmetric beyond this point (i.e. we would expect the same % of p-values <0.01 when there are 6000 cases as when there are 4000, similarly 9000 and 1000), because once α_D >0, the cases become a larger group than the non-cases, but the inference is the same as it was when the groups were the other way round. Finally, this process was re-run for situations where X_1 and X_2 were correlated, and as seen in section 6.4.1.2, there was no change in terms of the proportion of edges detected.

Now we fix $\beta_1 = \beta_2 = 1$ and $\gamma_{12} = 0$, which previously resulted in approximately 50% of simulations giving a p-value of less than 0.01 (section 6.4.1.1, figure 6.13b), although that proportion increased as the partial correlation of X_1 and X_2 increased (figures 6.13d and f). If we investigate the effect of α_D in this scenario, as before, α_D increases from its lowest level the proportion of p-values <0.01 also increases. This peaks at the point where the number of cases is approximately 2000. However, as the number of cases approaches 5000 (i.e. α_D approaches 0), this proportion drops off dramatically with virtually 0% of edges detected once $\alpha_D = 0$ (figure 6.21). As the partial correlation between X_1 and X_2 increases the power to detect an edge also increases.

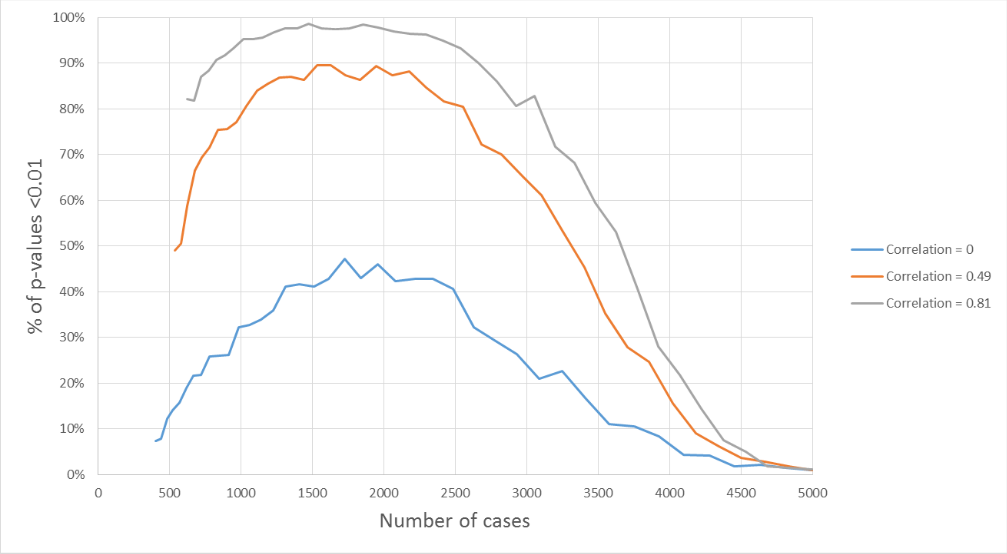


Figure 6.21: Line chart plotting proportion of p-values for $\delta_{12} = 0$ that are less than 0.01 against the number of cases (from a sample of 10,000). β_1=β_2=1, γ_{12}=0. α_D is varied to alter the number of cases.

6.4.1.5 Sample size

In the previous section, we took a fixed sample size and observed what occurred if we varied the number of cases within that fixed sample size of 10,000. We saw that when the smaller group reached a size of 2300, 95% of the simulations resulted in a p-value of less than 0.01 when $\gamma_{12} = 0.1$ (and $\beta_1 = \beta_2 = 0$). We repeated this with an increased sample size of 100,000, with figure 6.22

illustrating the results, showing the proportion of simulations that result in a p-value <0.01 by the number of cases (when the cases are the smaller group) when $\gamma_{12} = 0.1$. When the sample size was 10,000, it took 2300 cases to result in 95% of simulations detecting an edge, when the sample size was 100,000, 2000 cases were required.

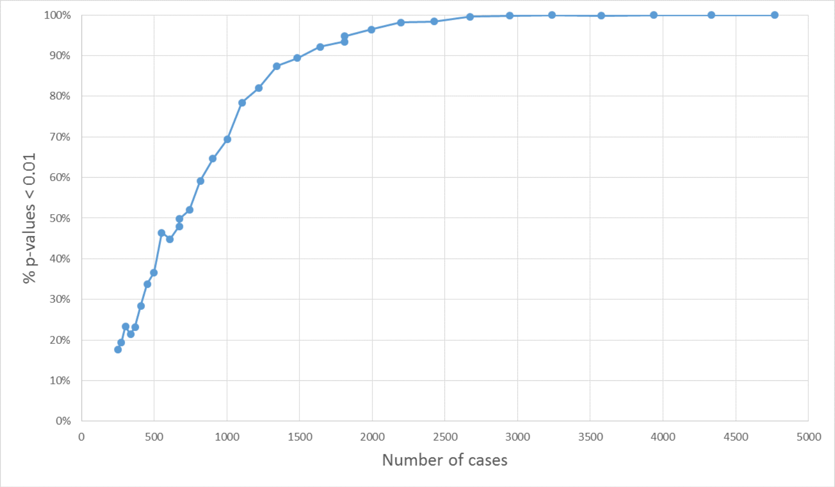


Figure 6.22: Line chart plotting proportion of p-values for $\delta_{12} = 0$ that are less than 0.01 against the number of cases (from a sample of 100,000). $\gamma_{12} = 0.1$, $\beta_1 = \beta_2 = 0$, 500 simulations. α_D is varied to alter the number of cases.

It is the size of the smaller group that determines the ability of the differential network method to determine an edge, since the variance of $\widehat{\delta}_{12}$ is based on the variance of the partial correlation of X_1 and X_2 in both the cases and in the non-cases, and these variances are in turn based on the size of the respective groups. If one of the groups is very small, its partial correlation will be imprecisely estimated and the variance of $\widehat{\delta}_{12}$ will therefore be high, even if the other group is extremely large and therefore its partial correlation precisely estimated. These latest simulations show us that we need a minimum of 2000 cases to ensure that we detect an edge in this scenario (95% of the time), with the number of cases required increasing, with decreasing sample size. Once the overall sample size drops to 7000, 3500 (50%) cases are required. So the minimum sample size required to detect difference in partial correlations of this magnitude is 7000.

6.4.1.6 Comparison with logistic regression

Given the data generating model and the results showing the agreement between γ_{12} and δ_{12}, we now investigate whether focusing our analysis on estimating γ_{12} is as or more informative as performing a differential network analysis.

For each set of simulations performed, where $\beta_1 = \beta_2 = 0$ and $\gamma_{12} \neq 0$, a logistic regression model was also fitted with explanatory variables X_1, X_2 and their interaction, reflecting how the data were generated.

$$logit\{\text{Prob}\,(D = 1)\} = \alpha_D^* + \beta_1^* X_1 + \beta_2^* X_2 + \gamma_{12}^* X_1 X_2 \tag{6.11}$$

The significance of the estimated regression coefficient for the interaction term was then evaluated. Figure 6.23 shows the proportion of p-values <0.01 for the interaction between X_1 and X_2 when the sample size is 10,000. The dashed lines represent the results from when a logistic regression has been used, the solid lines are the results previously presented in figure 6.19, representing the inferred results for δ_{12}. It can be seen from the chart that when the correlation of X_1 and X_2 is set to 0, that the differential network and logistic regression identify a similar proportion of p-values <0.01 across the values of γ_{12}. However, as the correlation of X_1 and X_2 strengthens, the logistic regression performs better at identifying the interaction, where the differential network stays the same. These results are also reflected when we investigate the effect of changing the size of the cases group (figure 6.24).

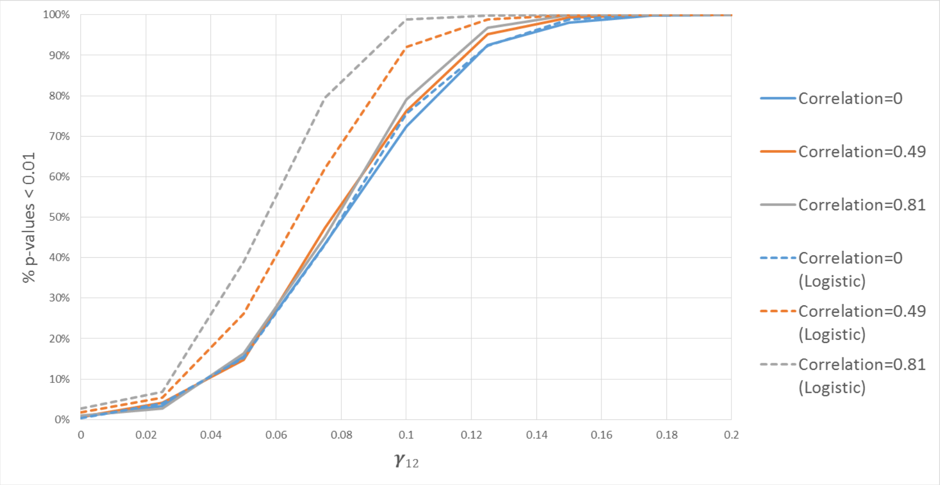


Figure 6.23: Line chart plotting proportion of p-values that are less than 0.01 across a range of values of γ_{12}^*. Dashed lines represent the results from a logistic regression testing the null hypothesis γ_{12}=0, solid lines represent the results from the test of $\delta_{12} = 0$. Sample size = 10,000, simulations = 500, $\beta_1 = \beta_2 = 0$, $\alpha_D = -2$

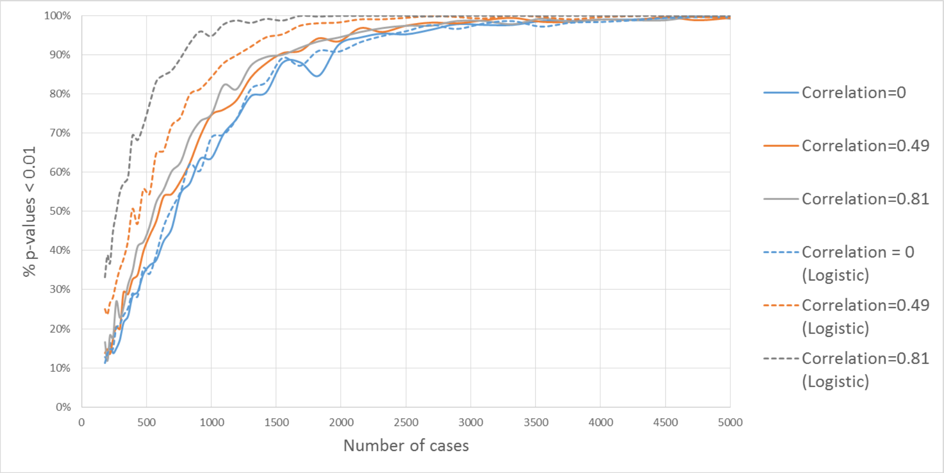


Figure 6.24: Line chart plotting proportion of p-values that are less than 0.01 across a range of sizes of the cases group, with γ_{12} set equal to 0.1. Dashed lines represent the results from a logistic regression testing the null hypothesis γ^*_{12}=0, solid lines represent the results from the test of $\delta_{12} = 0$. Sample size = 10,000, simulations = 500, $\beta_1 = \beta_2 = 0$. α_D is varied to alter the number of cases.

6.4.1.7 Comments

It may seem obvious that a logistic regression would be more efficient than the differential network when using this data generation model, as the model is exactly that assumed for a logistic regression. i.e. when identifying an interaction, the logistic regression is testing whether the interaction term is equal to 0, whereas the differential network is testing whether the difference in partial correlations, induced by a non-zero interaction term, is equal to 0. So it is perhaps logical that the direct test would be more efficient than a more indirect test. However, under a different data generating model, we may see different results.

6.4.2 Data generating model B

The simulations performed for data generating model B are simpler, due to it involving fewer parameters. In this scenario we vary the sample size, proportion of cases in the sample and the strength of association between X_1 and X_2 in the non-cases (defined as λ_1). Recall that this scenario would arise when pre-existing disease modifies the correlation between the metabolites.

Again, starting with a sample size of 10,000 we can plot the proportion of simulations that result in a p-value <0.01 for the signifcance testing of $\delta_{12} = 0$ (figure 6.25). Here we observe that with this sample size and 500 cases, 95% of simulations yield a p-value <0.01 once λ_1 reaches 0.2 (implying a correlation among the non-cases of 0.2 and a correlation of 0 in the cases, resulting in a δ_{12} of -0.2). When the number of cases reduces to 100 and 50 respectively the λ_1 required increases to 0.4 and 0.55 respectively. It should be noted that these are fairly large differences in correlations in the two groups, the cancelling out of a correlation of 0.2 or more is quite large effect. If the difference was reduced to 0.1, the power with 500 cases is reduced to 18%.

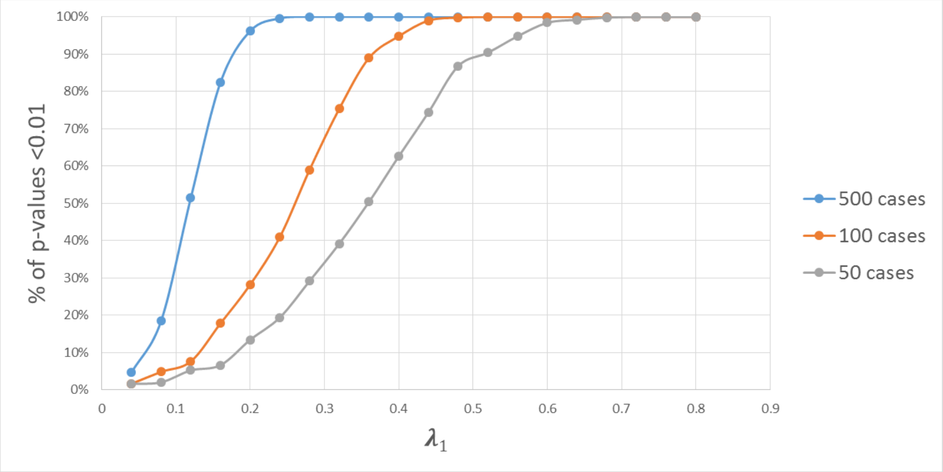


Figure 6.25: Proportion of p-values <0.01, across values of λ_1 from 0 to 0.9 (which corresponds to the correlation in the non-cases) for 3 different numbers of cases. Total 10,000 observations, 500 simulations

.

By fixing the number of cases and varying the overall sample size we can observe that the overall sample size has little effect on the proportion of simulations where the p-value for δ_{12} that are <0.01. Figure 6.26 shows the proportion of simulations that yield a p-value <0.01 when the number of cases is fixed at 100 and the total sample size changes from 10,000 to 1000 to 500. It can be seen that there is very little difference between them, suggesting it is the size of the smaller group, rather than the overall sample size, that is the main influence for the power to detect an edge in the differential network.

In each of the simulation sets described so far the value of λ_2 has been set to 0, so X_3 is not associated with X_1. However, when this is increased to a large value (0.8) we can observe that the power to detect an edge is reduced (figure 6.27). This is because of the strong correlation of X_1 and X_3 means that the association between X_1 and X_2 after adjusting for X_3 is reduced, so the partial correlation between X_1 and X_2 in the non-cases is reduced.

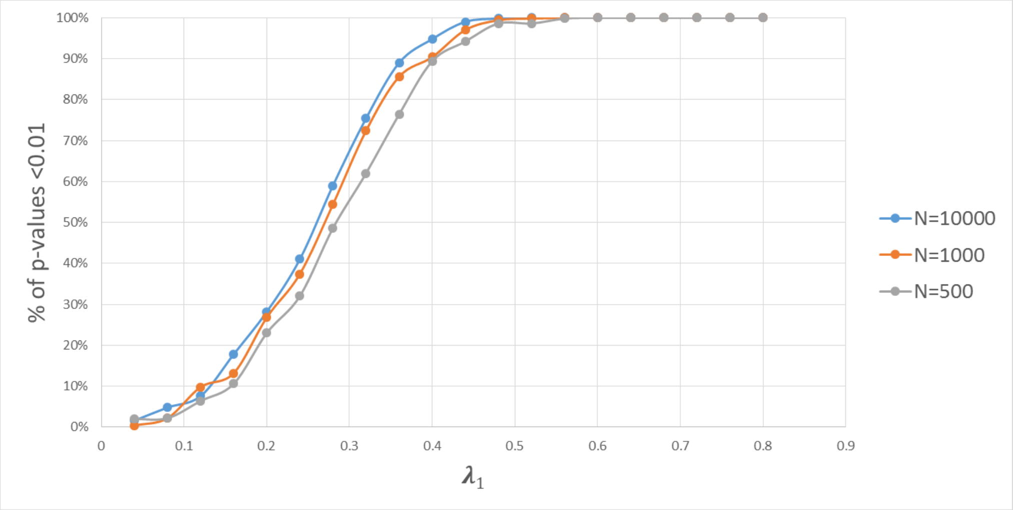


Figure 6.26: Proportion of p-values <0.01, across values of λ_1 for 3 different total sample sizes. Number of cases is 100 in each set of 500 simulations

.

6.4.2.1 Comparison with logistic regression

In all the simulations illustrated so far, a logistic regression has also been performed, and the proportion of p-values <0.01 for the interaction of X_1 and X_2 found, using the same associational model as when comparing scenario A to logistic regression.

$$logit\{\text{Prob}\,(D=1)\} = \alpha_D^* + \beta_1^* X_1 + \beta_2^* X_2 + \gamma_{12}^* X_1 X_2 \qquad (6.12)$$

In situations where $\lambda_2 = 0$ the differences in results from the logistic regression and the differential network were negligible. When λ_2 is set to a non-zero value however, the differential network approach has more power to detect a difference in correlations than the logistic regression has to detect a significant interaction. Figure 6.28 shows the comparison between the two approaches when there are 1000 observations and 100 cases and $\lambda_2 = 0.8$.

In this scenario an edge in the differential network is detected 95% of the time when λ_1 is increased to 0.48, but using the logistic regression an interaction is detected only 57% of the time at the same level. By the time λ_1 is increased to 0.8 the logistic regression picks up an interaction approximately 95% of the time.

6.4.3 Data generating model C

This data generating model is very similar to data model B, other than the association between D and the "switching off" of the association between X_1 and

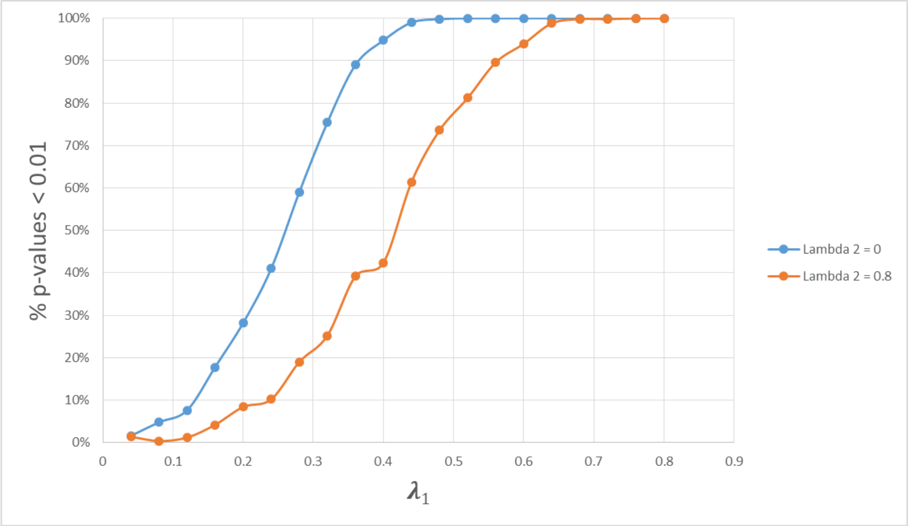


Figure 6.27: Proportion of p-values <0.01, across range of λ_1 for two different values of λ_2. There are 10,000 observations and 100 cases in each set of 500 simulations

.

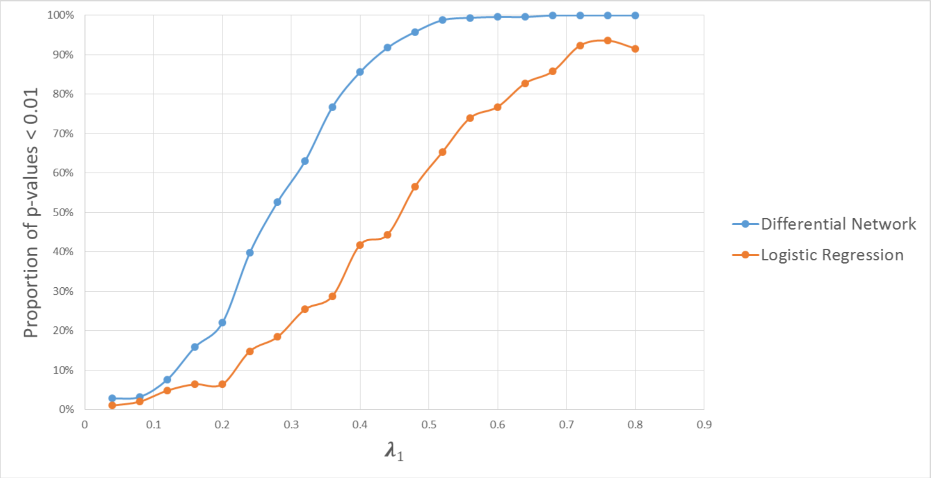


Figure 6.28: Proportion of p-values <0.01, across range of λ_1 for the test of a) $\delta_{12} = 0$ and b) $\gamma^*_{12} = 0$. There are 1,000 observations and 100 cases in each set of 500 simulations. $\lambda_2 = 0.8$

X_2 is associational rather than causal. In this case therefore there is no time ordering between metabolites and disease. If β_Y and β_D are set high enough then D $\approx$ Y and the results from analysing the two scenarios will be very similar, even if the relationships between D, X_1 and X_2 are not causal.

As for scenario A we can illustrate the effect of number of cases on the results, by varying the parameter α_D. Two examples are illustrated in figure 6.29, where α_D has been set equal to -2 and -3, which in a sample size of 10,000 is equivalent to approximately 2250 and 1300 cases respectively.

Also similar to before, when λ_2 is increased, the differential edge δ_{12} is less likely to be picked up because the partial correlations are reduced in absolute size (figure 6.30).

6.4.3.1 Comparison with logistic regression

As before, the differential network approach outperforms the logistic regression approach, with figure 6.31 illustrating the difference in performance. It shows that when λ_1 is equal to 0.72, 96% of tests for $\delta_{12} = 0$ result in a p-value less than 0.01, compared with 64% of tests for an interaction between X_1 and X_2.

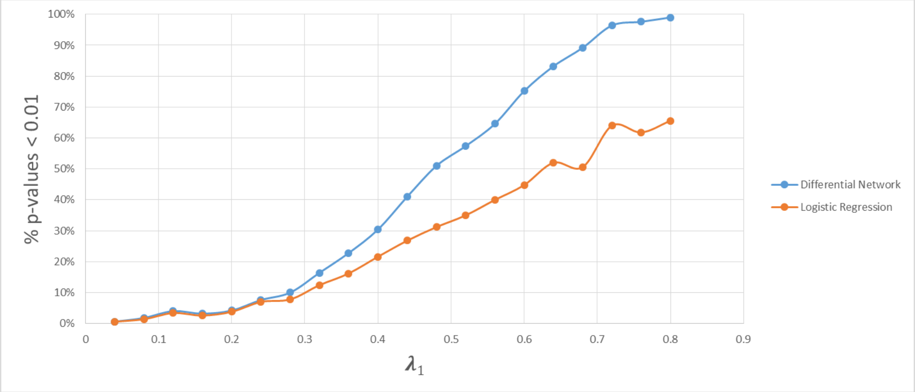


Figure 6.31: Proportion of p-values <0.01, across range of λ_1 for the test of a) $\delta_{12} = 0$ and b) $\gamma_{12} = 0$. There are 1000 observations in each simulation. $\alpha_D = -2$, $\beta_Y = \beta_D = 2$, $\lambda_2 = 0$

6.5 Summary and discussion

Differential networks have been proposed as exploratory methods for highlighting possible changes in associations among metabolites experienced when in different biological states. In this chapter we applied the method to the BWHHS

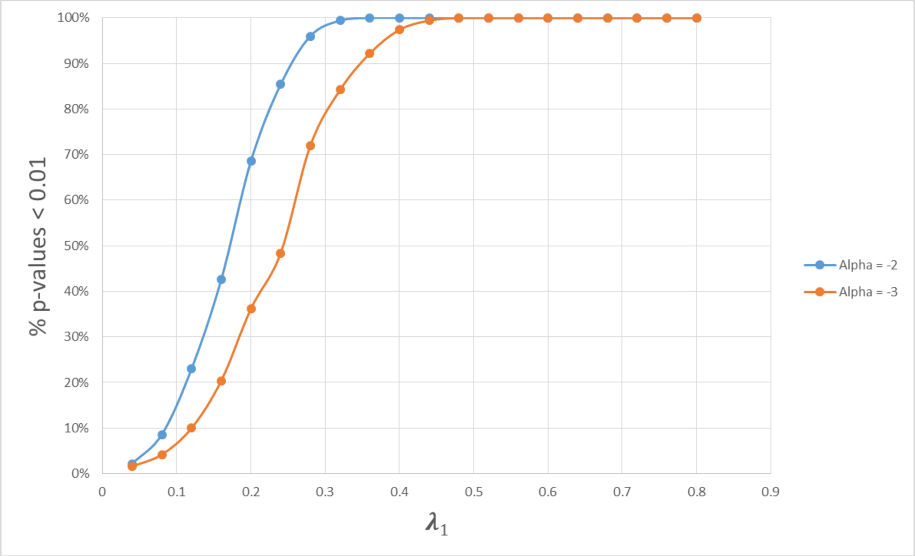
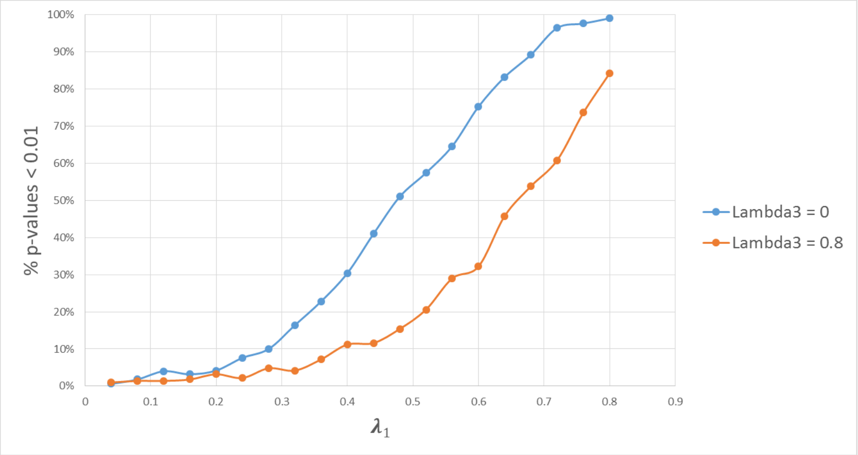


Figure 6.29: Proportion of p-values <0.01, across range of λ_1 for a set of 500 simulations where a) $\alpha_D = -2$ and b) $\alpha_D = -3$. There are 10,000 observations in each simulation. $\lambda_2 = 0$ $\beta_D = \beta_Y = 2$



.

Figure 6.30: Proportion of p-values <0.01, across range of λ_1 for two different values of λ_2. There are 10,000 observations in each simulation. $\alpha_D = -2$ $\beta_D = \beta_Y = 2$

.

data, creating a differential network using the partial correlations of 78 metabolites, but were unable to provide a suitable interpretation of what the network represented. As a result we simulated data from three different possible data generating models and a range of settings within each model. In doing this we were able to investigate which situations would lead to the inclusion of an edge in a differential network (defined in terms of partial correlation). The three scenarios were chosen because they are likely to capture likely causal models: scenario A has the metabolites as causes of disease, scenario B has the disease causing the behaviour of the metabolites and scenario C has disease and the metabolites sharing a common cause.

While examining scenario A, we have found that a difference in partial correlations can be found in situations where the effect of X_1 on disease is modified by X_2 (and vice-versa). It can also be found where both X_1 and X_2 are strong independent causes of disease, or if just one is an extremely strong cause of disease and the two metabolites are correlated. If there is an interaction AND one of these other situations, the two effects can cancel each other out to give a null finding. In practice, the main effects required to induce a difference in correlations are so large that their effect on δ_{12} can be considered negligible compared to any interaction effect.

We have also found that to detect a difference in partial correlations, induced by a relatively moderate interaction (OR = 1.1) between the two variables considered, with no main effects, a minimum sample size of 7000 observations was required, if the disease prevalence was 50%. A lower prevalence would require a larger sample size, but irrespective of the overall sample size, the smaller group must have at least 2000 observations. These are all based on the assumption that there are no causes of D unaccounted for (i.e. U does not have a direct effect on D).

When examining scenarios B and C it was found that if the disease (or another cause of disease) modifies the relationship between a pair of metabolites, this can lead to an edge arising in the differential network, again assuming no other causes of D (i.e. X_3 is not a cause of D).

For each set of simulations estimating the difference in partial correlations and its associated p-value, a logistic regression (for experiencing the event) was performed for comparison, with the regression coefficient for the interaction between the two metabolites being the focus of inference. In every scenario, for data generating model A, the logistic regression was more powerful at picking up the interaction, with the added advantage of being able to identify the main effects of the metabolites. In scenarios B and C the differential network was equally or more powerful in all settings examined.

These findings are perhaps obvious, since in scenario A the data were generated by specifying γ_{12}, which induces a non-zero value of δ_{12} so it would be expected that testing $\gamma_{12}^* = 0$ would be more powerful than testing $\delta_{12} = 0$. In scenarios B and C we effectively specify δ_{12}, which induces a non-zero value of γ_{12} so again it follows that testing $\delta_{12} = 0$ would be more powerful than testing $\gamma_{12}^* = 0$.

Once a differential edge is identified it is not possible to know for sure the data generating model that has led to it (or if the edge was identified by chance) but a some recommendations as to what to do if one is encountered are as follows:

- Investigate the scatter plot of the two variables involved in the differential edge. What has led to the difference in correlations? (e.g. the cases are correlated, the non-cases are uncorrelated).
- Have the two variables involved been picked up as associated with the outcome in a univariable analysis?
- If you perform a logistic regression with both variables and their interaction, is there evidence of effect modification?
- Does the difference in correlations exist when not adjusted for any other variable? If not the addition of which variables in the network lead to the difference observed, try stepping through each variable in turn (this process is explained in section 7.1.2).
- Consider the biological plausibility that the correlation between this pair of variables differs in the cases and non-cases.

In this chapter we looked in depth at the most basic element of a differential network, the edge between two nodes. We discovered that an edge can be induced in a differential network if the effect of one node on disease is modified by another node, or if the disease causes (or is associated with) a “breaking down” in the relationship between the two nodes. We looked at a simplified model, and only examined the edge of interest, not investigating any knock on effects throughout a larger network. This provided us with a grounding in the building blocks of a differential network, and illustrated that the data generating model affected the results.

Chapter 7

Overview and guidance for analysis of differential networks

In this chapter the aim is to consider the implications of what has been observed so far in applying differential networks in practice. First we will investigate the choice between marginal and partial correlation as a measure of association between a pair of nodes, then we will discuss the role of variable selection on the results of analysis. Finally the methods described will be applied to the BWHHS dataset and the results described.

7.1 Marginal or partial correlations?

Let us start by considering a data generating model, which is an extension of the models used in the previous chapter, where there are three unmeasured latent factors (U_1, U_2 and U_3), which are correlated with one another. Each of these latent factors, in healthy individuals, strongly influences the concentration levels of 10 metabolites, denoted respectively X_1-X_{10}, Y_1-Y_{10} and Z_1-Z_{10}. However, in unhealthy individuals the concentration of the X metabolites is not influenced by U_1. This is different to scenario A in the previous chapter as rather than the values of X influencing D it is the disease status D that influences the values of X. This could be happening directly (as in data generating model B) or via an association with another variable (as in model C). Figure 7.1 illustrates the proposed data generation models for healthy and unhealthy groups separately, which is equivalent to data generating model B from the previous chapter. Note that in this set up only the metabolites $X_1 - X_{10}$, $Y_1 - Y_{10}$, $Z_1 - Z_{10}$ are observed while U_1, U_2 and U_3 are latent so that only the former will be contributing to network analyses.

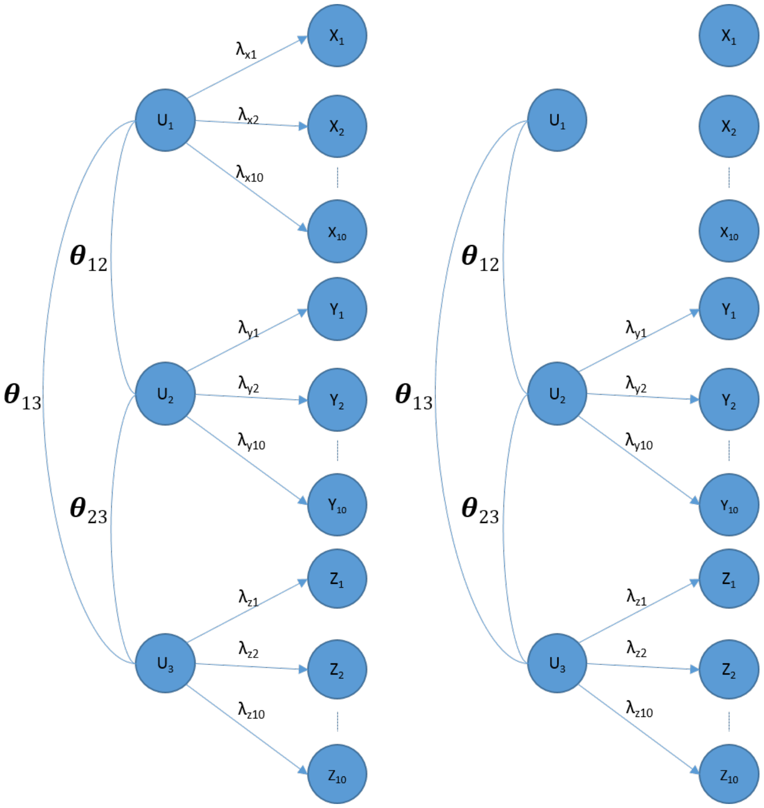


Figure 7.1: Data generating model for healthy (left) and unhealthy (right) individuals, error terms not shown. (Data generating model 1)

The correlations between the latent variables are specified in general as θ_{12}, θ_{13} and θ_{23}. The latent variables U_1, U_2 and U_3 are distributed $N(0,1)$ and the equations used to generate the metabolites are, for i=1,2,...,10:

$$\begin{aligned} X_i &= \lambda_{xi}U_1 - \lambda_{xi}U_1D + \epsilon_{xi} \\ Y_i &= \lambda_{yi}U_2 + \epsilon_{yi} \\ Z_i &= \lambda_{zi}U_3 + \epsilon_{zi} \end{aligned} \tag{7.1}$$

where $\epsilon_i^j \sim N(0, \sigma_{ji}^2)$ (where j = x,y,z and i = 1,2...10), $D = 0$ in healthy individuals and $D = 1$ in unhealthy individuals. This is a special case where when $D = 1$ the two $\lambda_{xi}U_1$ terms cancel so that $X_i = \epsilon_{xi}$.

The aim of this chapter is not to further investigate the effect of sample size, the aim is to study the values of the δ parameters expected to be observed in the differential network. So when performing the analysis we will use a large, fixed sample size of 1,000,000, and a disease prevalence of 10%. A sample of 1,000,000 is generated first, with 100,000 individuals randomly (and independently of the other variables) assigned as having the disease. Once the λ, θ and σ values are selected, the values of X, Y and Z are then generated for each individual. For simplicity, in the following we let each of the λ values be equal, similarly all the σ values are set to be equal, and as described above all the θ values will be set equal as well. Using the generating model proposed in figure 7.1, the impact of using marginal or partial correlations as the basis of the differential network will be examined.

7.1.1 Differential network based on marginal correlations

Using the data model described, the metabolomic network among healthy individuals we would expect to observe would be made of 3 groups, with nodes strongly associated within groups and members of each group also correlated with members of the other groups, albeit more loosely.

In this section we will use some specific terminology that must be defined:

- Parent - in figure 7.1 U_3 is a **parent** of Z_1, because U_3 is a cause of Z_1
- Child - in figure 7.1 Z_1 is a **child** of U_3
- Sibling - in figure 7.1 Z_1 and Z_2 are **siblings** as they have a common parent
- Cousin - in figure 7.1 Y_1 and Z_1 are **cousins** as they have parents who are siblings

Using the above terminology we expect the marginal correlation between siblings, amongst healthy individuals to be:

$$\rho_{x1x2} = \frac{\lambda_{x1}\lambda_{x2}}{\sqrt{(\lambda_{x1}^2 + \sigma_x^2)(\lambda_{x2}^2 + \sigma_x^2)}} \tag{7.2}$$

and the correlation between cousins to be:

$$\rho_{x1y1} = \frac{\theta\lambda_{x1}\lambda_{y1}}{\sqrt{(\lambda_{x1}^2 + \sigma_y^2)(\lambda_{y1}^2 + \sigma_y^2)}} \tag{7.3}$$

and similarly for each other pair of siblings or cousins. For illustrative purposes, we can draw the marginal correlation network (for healthy individuals only) and we would expect to see a network as shown in figure 7.2. For this example all values of λ are set to be equal to 0.9 and, for this figure, $\sigma = 1$ and $\theta = 0.5$. The edges in the network (i.e. the marginal correlations) between sibling nodes are equal to 0.44, and between cousins they are equal to 0.22.

In this illustrative example, in the network for unhealthy individuals the X variables are by design all uncorrelated to the other X variables and also to all of the Y and Z variables. Therefore the correlation between the Xs is equal to 0 as is the correlation between each X and their cousins. In contrast the Ys and Zs still form modules, with a marginal correlation again of 0.44 between siblings and 0.22 between cousins, but the Xs are all isolated nodes with a marginal correlation of 0 between any X node and any other node in the network (figure 7.3).

Finally we consider the differential network that this scenario would induce when calculations are based on these marginal correlations. Among the Ys and the Zs the correlation between siblings does not change between healthy and unhealthy groups, so there would be no differential edges within these groups, i.e. $\delta = 0$, where δ denotes the edge of the differential network as in the previous chapter. However the Xs are all correlated in healthy individuals and uncorrelated in unhealthy individuals so an edge would be generated between each of the X variables. The marginal correlation between Xs in healthy individuals was 0.44 and between Xs in unhealthy individuals was 0, so the value of δ in this example is 0.44. Also, between each X and its cousins there is a marginal correlation in healthy individuals of 0.22 but not amongst unhealthy individuals, so again an edge would be generated between each X and all of its cousins with a value of $\delta = 0.22$. This would result in a differential network as shown in figure 7.4.

An important fact to note is that for a given data generating model these edges are unaffected by inclusion or exclusion of any other nodes in the network, meaning that when using marginal correlations to define a differential network we get a stable result whatever set of metabolites one studies. So finding (or not) an edge between two nodes in a differential network based on marginal correlations is independent of how many other nodes are included in the network.

7.1.2 Differential network based on partial correlations

We can now consider the networks generated if partial correlations are used instead of marginal correlations. Because all the variables in our example are positively correlated, by adjusting for other variables in the network we reduce the partial correlation between each pair of nodes. By introducing additional sibling nodes we reduce the strength of the correlation by a greater amount than by introducing additional cousins.

Because of the simplicity of the example discussed here, the healthy and unhealthy networks actually look the same when partial correlations are used as when marginal correlations are used, although the strengths of the edges are weaker. So in the network formed by healthy individuals, the network topology is the same as that shown in figure 7.2 however, estimating the partial correlations using a sample size of 1,000,000, the strength of an edge between siblings is estimated to be 0.093 and between cousins it is 0.005 (as opposed to 0.44 and

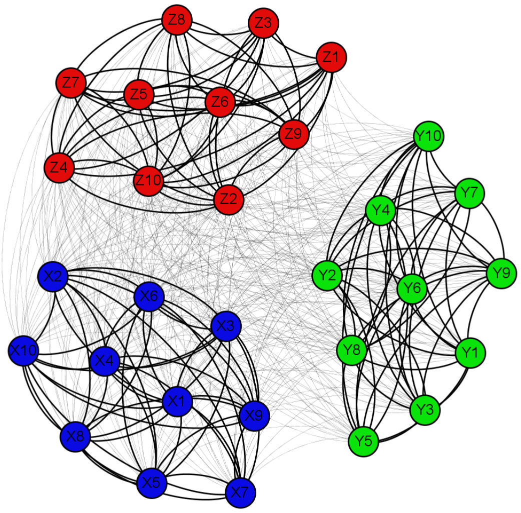


Figure 7.2: Network of the 30 observed variables in healthy individuals. Data generated using equation 7.1, with $\lambda = 0.9, \sigma = 1, \theta = 0.5, D = 0$.

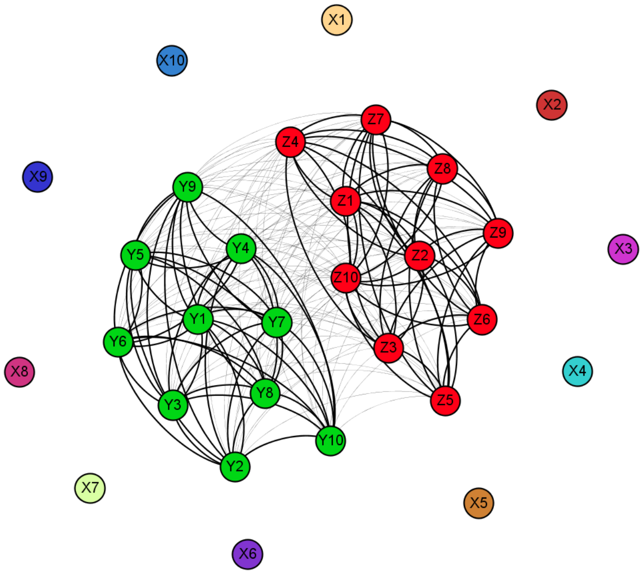


Figure 7.3: Network of the 30 observed variables in unhealthy individuals. Data generated using equation 7.1, with $\lambda = 0.9, \sigma = 1, \theta = 0.5, D = 1$.

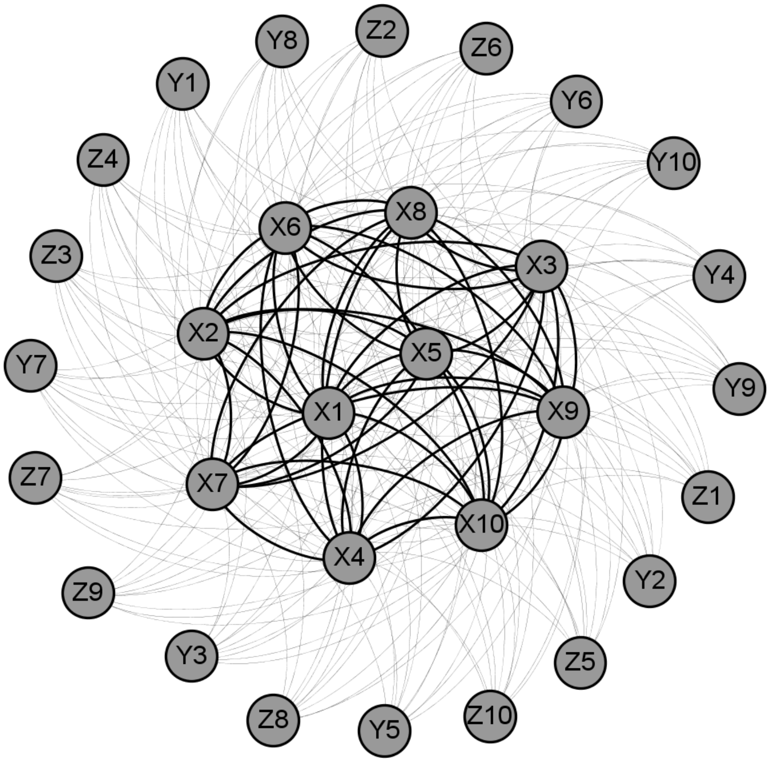


Figure 7.4: Differential network using marginal correlation coefficients

0.22 in the network formed using marginal correlations). In the partial correlation network formed by unhealthy individuals, the topology is again the same as in the network using marginal correlations shown in figure 7.3. The X variables are isolated nodes as they are uncorrelated with any of the other variables, the partial correlation between siblings (within the Y and Z variables) is 0.095 and and between cousins in the Y and Z variables is 0.006 (compared with 0.44 and 0.22 in the network formed using marginal correlations).

So in the differential network formed using partial correlations, the observed value of δ between each pair of X variables is -0.093 and between each X variable and its cousins the value of δ is -0.005. There are also very small, non-zero values of δ between siblings and cousins in the Y and Z variables ($\delta = -0.0001$).

The setting defined by the data generating model describes a situation where the association between U_1 and $X_1 - X_{10}$ changes between a group of healthy individuals and a group of unhealthy individuals. As U_1 is unobserved, what we might expect from a differential network analysis performed on this setting as defined by the data generating model is a differential network that shows edges from each of the X nodes to all other nodes in the network. We would also expect all other pairs of nodes to have an expected δ of 0 between them. This is what was found when performing the differential network analysis using marginal correlations. It was also identified using partial correlations, although there were tiny non-zero values of δ between the Y and Z nodes as well. In this setting these were so small that no hypothesis test would identify them as a non-zero edge unless the sample size was enormous, however in section 7.1.2.1 this is expanded upon with a discussion about where these edges may pose a problem.

The other point to note is that when using marginal correlations, the value of δ estimated between a pair of nodes will be the same, irrespective of what other nodes are included in the analysis, for a given data generating model. For example, the δ between X_1 and X_2 will remain the same if there are 10 other nodes in the network or if there are 100. When using partial correlations this is not the case, each additional node added to the analysis contributes to the estimate of every δ in the network, which will also be shown in section 7.1.2.1.

7.1.2.1 An additional issue when using partial correlations in differential networks

Consider healthy individuals, and the scenario where only 2 metabolites are observed from each class X_1, X_2, Y_1, Y_2, Z_1 and Z_2, but the data generating model is the same as in the previous section. With the specifications described in the previous section the partial correlation of X_1 and X_2 is 0.38 (and is the same as the partial correlation of Y_1,Y_2 and Z_1,Z_2). The partial correlation between cousins is equal to 0.07 for the same values of θ and σ as before (figure 7.5).

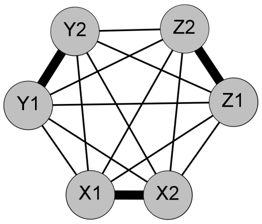


Figure 7.5: Association network in healthy individuals using partial correlation coefficients, with two variables observed (instead of ten as per marginal network) from each set of cousins. Data generated using equation 7.1, with $\lambda = 0.9, \sigma = 1, \theta = 0.5, D = 0$.

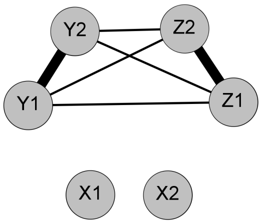


Figure 7.6: Association network in unhealthy individuals using partial correlation coefficients, with two variables observed (instead of ten as per marginal network) from each set of cousins. Data generated using equation 7.1, with $\lambda = 0.9, \sigma = 1, \theta = 0.5, D = 1$.

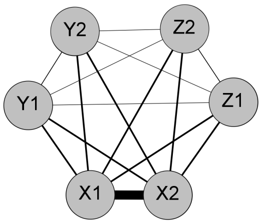


Figure 7.7: Differential network using partial correlation coefficients, corresponding to figures 7.5 and 7.6

Now in unhealthy individuals the partial correlation between the Xs is equal to 0, and also the partial correlation between Xs and their cousins are equal to zero. The partial correlation between Y_1 and Y_2 however is now increased to 0.41, because X_1 and X_2 no longer "explain away" any of the correlation between Y_1 and Y_2 so adjusting for these variables now has no effect (this is true for Z_1 and Z_2). The partial correlations between Ys and Zs are also slightly increased to 0.09 (figure 7.6).

When we look at the differential edges we see a strong edge between X_1 and X_2, with $\delta = -0.38$, and weaker edges between X_1 and X_2 and their cousins ($\delta = -0.07$), as we might expect. However, we now also see weak edges appearing between Y_1/Y_2 and Z_1/Z_2 ($\delta = -0.03$) and also between the Ys and the Zs ($\delta = -0.02$) with the differential network illustrated in figure 7.7.

As we observe more variables, the results change, and depending on which variables are included we observe different changes. For example, as further X variables are added this reduces the partial correlations between all pairs of variables in the healthy population (as everything is positively correlated), however it reduces the partial correlation between pairs of X variables than it does between Y or Z variables, or between cousins. Therefore the partial correlation of siblings among the Xs will be reduced in the healthy population and will remain at 0 among the unhealthy, so the difference (δ) will be reduced. It will have a similar effect throughout the network although not as pronounced.

The impact of increasing the number of observed X variables, for a given data generating model, on the difference in partial correlations can be viewed graphically in figure 7.8. If we take data generating model 1 (figure 7.1) and we assume only a single Y variable and a single Z variable is observed and we start with two X variables X_1 and X_2 observed.

When there are only two X variables observed, the edge connecting X_1 and X_2 in the differential network is of a much greater magnitude than the edge connecting X_1 and Y_1, the edge connecting X_1 and Z_1 and the edge connecting Y_1 and Z_1. However, as more X variables are observed, the differences between partial correlations in the Xs reduces, while the difference in partial correlation between Y_1 and Z_1 increases. Once there are greater than four X variables, the edge between Y_1 and Z_1 is stronger than any of the edges between either of them and each of the X variables, once there are greater than 11 X variables, the differential edges between Xs are smaller than those between Y_1 and Z_1.

This illustrates a potential problem with differential networks using partial correlations, unexpected edges can be introduced between a pair of variables not closely related to the true data generating process, when the nodes are strongly correlated within subgroups (as is the case with the BWHHS metabolomic data).

If instead of increasing the number of Xs we increase the number of Y variables included, the edge between X_1 and X_2 remains strong, as does the edge between X_1 and Z_1. All edges involving Y variables tend towards 0 (figure 7.9).

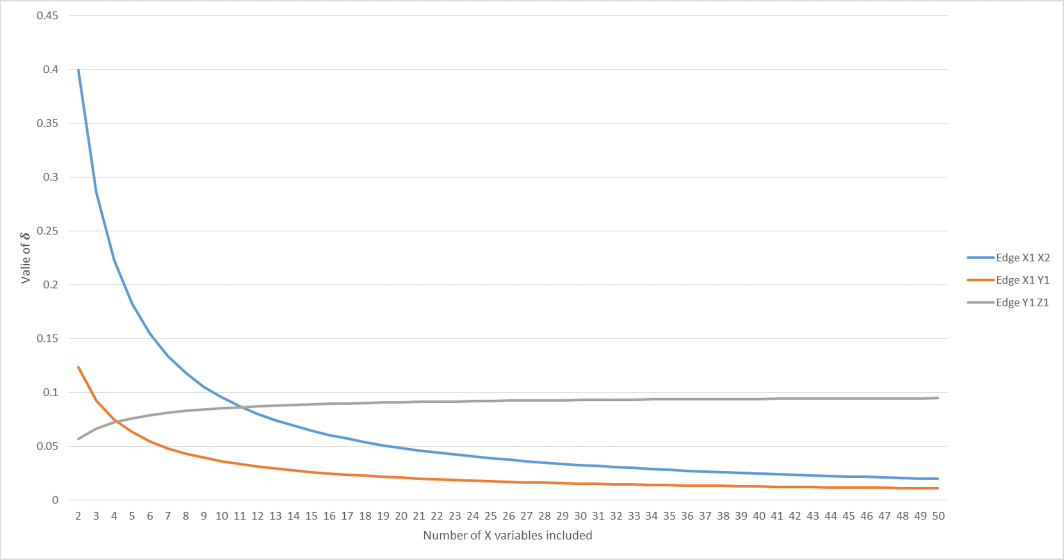


Figure 7.8: Edge strength for 3 selected pairs of variables against number of X variables included in network ($\theta = 0.5, \lambda = 0.9, \sigma = 1$) - the edge connecting X_1 and Z_1 is not shown as it is of an identical strength to the X_1/Y_1 edge

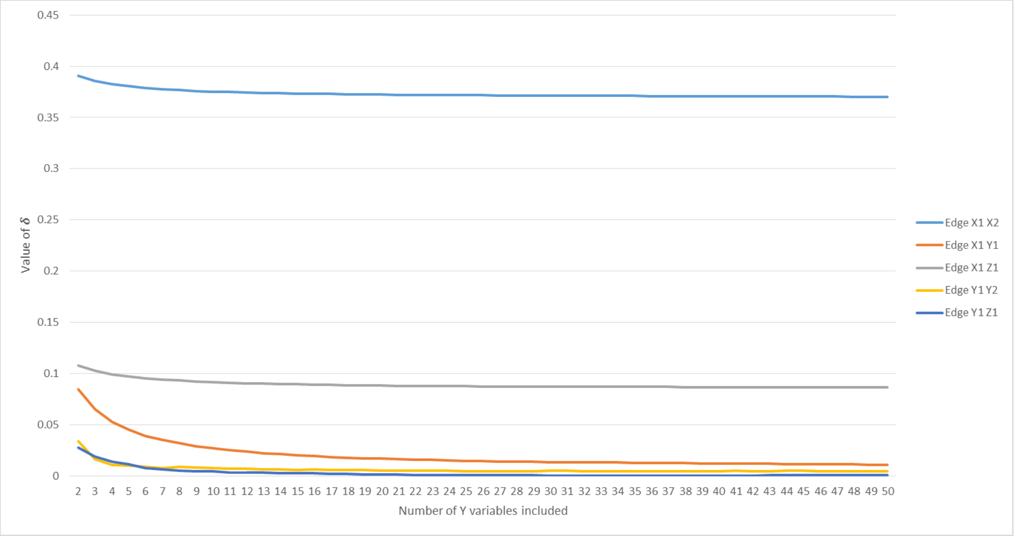


Figure 7.9: Edge strength for 5 selected pairs of variables against number of X variables included in network ($\theta = 0.5, \lambda = 0.9, \sigma = 1$)

7.1.3 General results

To try to be more analytic about why edges appear in the differential network, we will describe the mathematics behind the results by applying the methodology initially described by Wermuth and Cox [103] and applied previously by de Stavola et al [104]. To simplify the algebra used in this section, a second data generating model is proposed, which illustrates the issues described above.

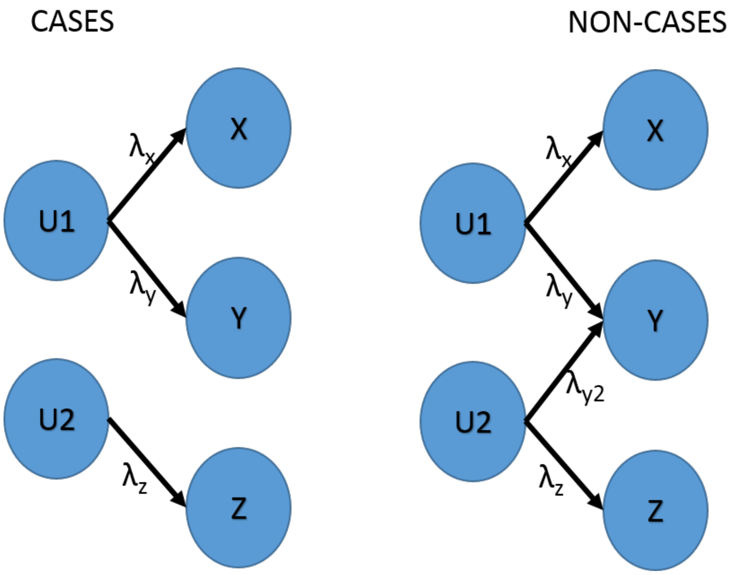


Figure 7.10: Separate data generating models for cases and non-cases. (Data generating model 2)

If we consider the setting described in data generating model 2 (figure 7.10), there is a variable U_1 that is a common cause of X and Y. Also there is a variable U_2 that is a cause of Y in the non-cases but not amongst the cases. U_2 is also a cause of Z in both cases and non-cases. We assume for simplicity that there are no background confounders. We can describe the marginal and partial differential network in general terms in this very simple scenario, assuming linear relationships among these variables, with no interactions and uncorrelated errors. First we can describe the data generating model for the cases:

Cases

$$\begin{cases} U_1 &= \epsilon_{u1} \\ U_2 &= \epsilon_{u2} \\ X &= \lambda_x U_1 + \epsilon_x \\ Y &= \lambda_y U_1 + \epsilon_y \\ Z &= \lambda_z U_2 + \epsilon_z \end{cases} \tag{7.4}$$

where $\tilde{\boldsymbol{\epsilon}} = (\epsilon_{u1}, \epsilon_{u2}, \epsilon_x, \epsilon_y, \epsilon_z)$ are error terms with mean 0 and variance covariance matrix Σ,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{u1}^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_{u2}^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_x^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_y^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_z^2 \end{pmatrix}$$

Rearranging these equations so that they are expressed in terms of the error terms:

$$\begin{cases} \epsilon_{u1} &= U_1 \\ \epsilon_{u2} &= U_2 \\ \epsilon_x &= X - \lambda_x U_1 \\ \epsilon_y &= Y - \lambda_y U_1 \\ \epsilon_z &= Z - \lambda_z U_2 \end{cases} \tag{7.5}$$

Let $\tilde{\mathbf{R}} = (U_1, U_2, X, Y, Z)^T$ and

$$\tilde{\mathbf{A}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ -\lambda_x & 0 & 1 & 0 & 0 \\ -\lambda_y & 0 & 0 & 1 & 0 \\ 0 & -\lambda_z & 0 & 0 & 1 \end{pmatrix}$$

then model (x.x) can be written in matrix notation as $\tilde{\mathbf{A}}\tilde{\mathbf{R}} = \tilde{\boldsymbol{\epsilon}}$

To marginalise this model with respect to $\boldsymbol{U} = (U_1, U_2)$, the two unobserved variables, we first use partial inversion of $\tilde{\boldsymbol{A}}$, as described in Wermuth and Cox [103] with respect to $\boldsymbol{U}$. This consists of first partitioning $\tilde{\boldsymbol{A}}$ into components that involve/do not involve $\boldsymbol{U}$:

$$\tilde{\mathbf{A}} = \left(\begin{array}{c|c} \tilde{\mathbf{A}}_{UU} & \tilde{\mathbf{A}}_{U\bar{U}} \\ \hline \tilde{\mathbf{A}}_{\bar{U}U} & \tilde{\mathbf{A}}_{\bar{U}\bar{U}} \end{array}\right).$$

where

$$\tilde{\mathbf{A}}_{\bar{U}\bar{U}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\tilde{\mathbf{A}}_{U\bar{U}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$\tilde{\mathbf{A}}_{\bar{U}U} = \begin{pmatrix} -\lambda_x & 0 \\ -\lambda_y & 0 \\ 0 & -\lambda_z \end{pmatrix}$$

and

$$\tilde{\mathbf{A}}_{UU} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

This leads to the new matrix $\boldsymbol{B}$

$$\begin{aligned} \mathbf{B} &= \text{inv}_U \tilde{\mathbf{A}} \\ &= \begin{pmatrix} \tilde{\mathbf{A}}_{UU}^{-1} & -\tilde{\mathbf{A}}_{UU}^{-1}\tilde{\mathbf{A}}_{U\bar{U}} \\ \tilde{\mathbf{A}}_{\bar{U}U}\tilde{\mathbf{A}}_{UU}^{-1} & \tilde{\mathbf{A}}_{\bar{U}\bar{U}} - \tilde{\mathbf{A}}_{\bar{U}U}\tilde{\mathbf{A}}_{UU}^{-1}\tilde{\mathbf{A}}_{U\bar{U}} \end{pmatrix} \end{aligned} \tag{7.6}$$

which equates as

$$\mathbf{B} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ -\lambda_x & 0 & 1 & 0 & 0 \\ -\lambda_y & 0 & 0 & 1 & 0 \\ 0 & -\lambda_z & 0 & 0 & 1 \end{pmatrix}$$

We will next refer to the elements of this matrix and of the matrix $\mathbf{\Sigma}$ using the same partitioning notation as for $\tilde{\boldsymbol{A}}$, that is

$$\tilde{\mathbf{B}} = \left(\begin{array}{c|c} \tilde{\mathbf{B}}_{UU} & \tilde{\mathbf{B}}_{U\bar{U}} \\ \hline \tilde{\mathbf{A}}_{\bar{U}U} & \tilde{\mathbf{B}}_{\bar{U}\bar{U}} \end{array}\right).$$

$$\tilde{\mathbf{\Sigma}} = \left(\begin{array}{c|c} \tilde{\mathbf{\Sigma}}_{UU} & \tilde{\mathbf{B}}_{U\bar{U}} \\ \hline \tilde{\mathbf{A}}_{\bar{U}U} & \tilde{\mathbf{\Sigma}}_{\bar{U}\bar{U}} \end{array}\right).$$

Additionally let $\boldsymbol{X} = (X, Y, Z)^T$ and $\boldsymbol{\epsilon} = (\epsilon_x, \epsilon_y, \epsilon_z)^T$. We are now ready to write the marginalised model in terms of these new elements, using Lemma 1 in Wermuth and Cox [103],

$$\boldsymbol{\eta} = \boldsymbol{B}_{UU}\boldsymbol{X}$$

where $\boldsymbol{\eta} = \boldsymbol{\epsilon} - \boldsymbol{B}_{U\bar{U}}\boldsymbol{U}$ and the variance covariance matrix of these error terms, K, is

$$\mathbf{K} = \mathbf{\Sigma}_{\bar{U}\bar{U}} + \mathbf{B}_{\bar{U}U}\mathbf{\Sigma}_{UU}\mathbf{B}_{\bar{U}U}^T$$

So to obtain $\mathbf{K}$,

$$\mathbf{B}_{\bar{U}U} = \begin{pmatrix} -\lambda_x & 0 \\ -\lambda_y & 0 \\ 0 & -\lambda_z \end{pmatrix}$$

$$\mathbf{\Sigma}_{UU} = \begin{pmatrix} \sigma_{u1}^2 & 0 \\ 0 & \sigma_{u2}^2 \end{pmatrix}$$

$$\mathbf{B}_{\bar{U}U}\mathbf{\Sigma}_{UU}\mathbf{B}_{\bar{U}U}^T = \begin{pmatrix} \lambda_x^2\sigma_{u1}^2 & \lambda_x\lambda_y\sigma_{u1}^2 & 0 \\ \lambda_x\lambda_y\sigma_{u1}^2 & \lambda_y^2\sigma_{u1}^2 & 0 \\ 0 & 0 & \lambda_z^2\sigma_{u2}^2 \end{pmatrix}$$

$$\mathbf{K} = \begin{pmatrix} \sigma_x^2 + \lambda_x^2\sigma_{u1}^2 & \lambda_x\lambda_y\sigma_{u1}^2 & 0 \\ \lambda_x\lambda_y\sigma_{u1}^2 & \sigma_y^2 + \lambda_y^2\sigma_{u1}^2 & 0 \\ 0 & 0 & \sigma_z^2 + \lambda_z^2\sigma_{u2}^2 \end{pmatrix}$$

So amongst the cases the covariance between X and Z is 0 and also between Y and Z. However the covariance between X and Y is

$$\mathbf{cov}(\mathbf{X}, \mathbf{Y}|\mathbf{case}) = \lambda_x\lambda_y\sigma_{u1}^2$$

So to get the correlation between X and Y (amongst the cases) we need to divide by the standard deviations of X and Y. The matrix $\mathbf{K}$ has the variances on the diagonal, so we must take the square root of these

$$\rho_{\mathbf{xy}.\mathbf{case}} = \frac{\lambda_x\lambda_y\sigma_{u1}^2}{\sqrt{\sigma_x^2 + \lambda_x^2\sigma_{u1}^2}\sqrt{\sigma_y^2 + \lambda_y^2\sigma_{u1}^2}}$$

$$\rho_{\mathbf{xz}.\mathbf{case}} = 0$$

$$\rho_{\mathbf{yz}.\mathbf{case}} = 0$$

The above are the marginal correlations in the cases, we are also interested in the partial correlations. If we define the covariance matrix $\mathbf{K}$ as:

$$\mathbf{K} = \begin{pmatrix} k_{11} & k_{12} & \cdots & k_{1J} \\ k_{21} & k_{22} & \cdots & k_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ k_{I1} & k_{I2} & \cdots & k_{IJ} \end{pmatrix}$$

and the inverse of this covariance matrix is $\tilde{\mathbf{K}}$:

$$\tilde{\mathbf{K}} = \begin{pmatrix} \tilde{k}_{11} & \tilde{k}_{12} & \cdots & \tilde{k}_{1J} \\ \tilde{k}_{21} & \tilde{k}_{22} & \cdots & \tilde{k}_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{k}_{I1} & \tilde{k}_{I2} & \cdots & \tilde{k}_{IJ} \end{pmatrix}$$

Then the partial correlation of a pair of variables i and j , adjusting for all other variables **(**Q) can be defined as $\rho_{ij.Q}$:

$$\rho_{ij.Q} = \frac{-\tilde{k}_{ij}}{\sqrt{\tilde{k}_{ii}}\sqrt{\tilde{k}_{jj}}}$$

So if we take the covariance matrix for the cases and invert it, the coefficients can be used to obtain partial correlation coefficients.
We can invert the covariance matrix for the cases. At this point to keep the algebra simple we assume that all σ terms are equal to 1 (unit variance), also the denominator from each term is excluded from each matrix element as it is very unwieldy and cancels out anyway when obtaining the partial correlation coefficients, which is the aim of this step:

$$\tilde{\mathbf{K}} = \begin{pmatrix} (1+\lambda_y^2)(1+\lambda_z^2) & -\lambda_x\lambda_y(1+\lambda_z^2) & 0 \\ -\lambda_x\lambda_y(1+\lambda_z^2) & (1+\lambda_x^2)(1+\lambda_z^2) & 0 \\ 0 & 0 & (1+\lambda_x^2)(1+\lambda_y^2) - \lambda_x^2\lambda_y^2 \end{pmatrix}$$

Only $\tilde{k}_{xy}$ is non-zero, so the only non-zero partial correlation will be the one between x and y. And the solution to that simplifies down to:

$$\rho_{\mathbf{xy.z,case}} = \frac{\lambda_x\lambda_y}{\sqrt{1+\lambda_x^2}\sqrt{1+\lambda_y^2}}$$

$$\rho_{\mathbf{xz.y,case}} = 0$$

$$\rho_{\mathbf{yz.x,case}} = 0$$

Non-cases If we now go on to repeat the above exercise for the non-cases, we can define the 5 variables as follows:

$$\begin{cases} U_1 &= \epsilon_{u1} \\ U_2 &= \epsilon_{u2} \\ X &= \lambda_x U_1 + \epsilon_x \\ Y &= \lambda_y U_1 + \lambda_{y2} U_2 + \epsilon_y \\ Z &= \lambda_z U_2 + \epsilon_z \end{cases} \tag{7.7}$$

and rearranging the formulae to describe the error terms gives us:

$$\begin{cases} \epsilon_{u1} &= U_1 \\ \epsilon_{u2} &= U_2 \\ \epsilon_x &= X - \lambda_x U_1 \\ \epsilon_y &= Y - \lambda_y U_1 - \lambda_{y2} U_2 \\ \epsilon_z &= Z - \lambda_z U_2 \end{cases} \tag{7.8}$$

Following the same process as before (omitting the intermediate steps) this results in a matrix **B** of:

$$\mathbf{B} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ -\lambda_x & 0 & 1 & 0 & 0 \\ -\lambda_y & -\lambda_{y2} & 0 & 1 & 0 \\ 0 & -\lambda_z & 0 & 0 & 1 \end{pmatrix}$$

and a resulting variance-covariance matrix **K**:

$$\mathbf{K} = \begin{pmatrix} \sigma_x^2 + \lambda_x^2 \sigma_{u1}^2 & \lambda_x \lambda_y \sigma_{u1}^2 & 0 \\ \lambda_x \lambda_y \sigma_{u1}^2 & \sigma_y^2 + \lambda_y^2 \sigma_{u1}^2 + \lambda_{y2}^2 \sigma_{u2}^2 & \lambda_{y2} \lambda_z \sigma_{u2}^2 \\ 0 & \lambda_{y2} \lambda_z \sigma_{u2}^2 & \sigma_z^2 + \lambda_z^2 \sigma_{u2}^2 \end{pmatrix}$$

So amongst the non-cases the covariance between X and Z is 0. However the covariance between X and Y is the same as before:

$$\mathbf{cov}(\mathbf{X}, \mathbf{Y}) = \lambda_x \lambda_y \sigma_{u1}^2$$

and now there is a non-zero covariance between Y and Z.

$$\mathbf{cov}(\mathbf{Y}, \mathbf{Z}) = \lambda_{y2} \lambda_z \sigma_{u2}^2$$

So to get the correlation between X and Y we need to divide by the standard deviations of X and Y. The matrix **K** has the variances on the diagonal, so we must take the square root of these. Now the variance of Y is different to before (As it is caused by both U_1 and U_2), so there is a slightly different correlation

$$\rho_{\mathbf{xy.non-case}} = \frac{\lambda_x \lambda_y \sigma_{u1}^2}{\sqrt{\sigma_x^2 + \lambda_x^2 \sigma_{u1}^2} \sqrt{\sigma_y^2 + \lambda_y^2 \sigma_{u1}^2 + \lambda_2^2 \sigma_{u2}^2}}$$

$$\rho_{\mathbf{xz.non-case}} = 0$$

$$\rho_{\mathbf{yz.non-case}} = \frac{\lambda_{y2} \lambda_z \sigma_{u2}^2}{\sqrt{\sigma_z^2 + \lambda_z^2 \sigma_{u2}^2} \sqrt{\sigma_y^2 + \lambda_y^2 \sigma_{u1}^2 + \lambda_2^2 \sigma_{u2}^2}}$$

Again to get the partial correlation coefficients we need to invert the covariance matrix for the non-cases. As before, to keep the algebra simple we assume that all σ terms are equal to 1 (unit variance), and the denominator is excluded:

$$\tilde{\mathbf{K}} = \begin{pmatrix} (1+\lambda_y^2+\lambda_{y2}^2)(1+\lambda_z^2) - \lambda_{y2}^2\lambda_z^2 & -\lambda_x\lambda_y(1+\lambda_z^2) & \lambda_x\lambda_y\lambda_{y2}\lambda_z \\ -\lambda_x\lambda_y(1+\lambda_z^2) & (1+\lambda_x^2)(1+\lambda_z^2) & -(1+\lambda_x^2)\lambda_{y2}\lambda_z \\ \lambda_x\lambda_y\lambda_{y2}\lambda_z & -(1+\lambda_x^2)\lambda_{y2}\lambda_z & (1+\lambda_x^2)(1+\lambda_y^2+\lambda_{y2}^2) - \lambda_x^2\lambda_y^2 \end{pmatrix}$$

This results in a non-zero partial correlation between each pair of metabolites. The partial correlation of x and y in the non-cases is:

$$\rho_{\mathbf{xy.z,non-case}} = \frac{\lambda_x\lambda_y(1+\lambda_z^2)}{\sqrt{(1+\lambda_y^2+\lambda_{y2}^2)(1+\lambda_z^2) - \lambda_{y2}^2\lambda_z^2}\sqrt{(1+\lambda_x^2)(1+\lambda_z^2)}}$$

The partial correlation of x and z in the non-cases is:

$$\rho_{\mathbf{xz.y,non-case}} = -\frac{\lambda_x\lambda_y\lambda_{y2}\lambda_z}{\sqrt{(1+\lambda_x^2)(1+\lambda_y^2+\lambda_{y2}^2) - \lambda_x^2\lambda_y^2}\sqrt{(1+\lambda_y^2+\lambda_{y2}^2)(1+\lambda_z^2) - \lambda_{y2}^2\lambda_z^2}}$$

The partial correlation of y and z in the non-cases is:

$$\rho_{\mathbf{yz.x,non-case}} = \frac{(1+\lambda_x^2)\lambda_{y2}\lambda_z}{\sqrt{(1+\lambda_x^2)(1+\lambda_z^2)}\sqrt{(1+\lambda_x^2)(1+\lambda_y^2+\lambda_{y2}^2) - \lambda_x^2\lambda_y^2}}$$

Differences So we can now calculate the differences (δ) for each pair of variables in both the marginal and partial differential networks, by subtracting the correlation in the non-cases from the correlation in the cases. First the marginal differences:

$$\delta_{\mathbf{xy}} = \frac{\lambda_x\lambda_y\sigma_{u1}^2}{\sqrt{\sigma_x^2+\lambda_x^2\sigma_{u1}^2}}\left(\frac{1}{\sqrt{\sigma_y^2+\lambda_y^2\sigma_{u1}^2}} - \frac{1}{\sqrt{\sigma_y^2+\lambda_y^2\sigma_{u1}^2+\lambda_2^2\sigma_{u2}^2}}\right)$$

$$\delta_{\mathbf{xz}} = 0$$

$$\delta_{\mathbf{yz}} = \frac{\lambda_{y2}\lambda_z\sigma_{u2}^2}{\sqrt{\sigma_z^2+\lambda_z^2\sigma_{u2}^2}\sqrt{\sigma_y^2+\lambda_y^2\sigma_{u1}^2+\lambda_2^2\sigma_{u2}^2}}$$

And for the partial correlations:

$$\delta_{\mathbf{xy}} = \frac{\lambda_x \lambda_y}{\sqrt{1+\lambda_x^2}\sqrt{1+\lambda_y^2}} - \frac{\lambda_x \lambda_y (1+\lambda_z^2)}{\sqrt{(1+\lambda_y^2+\lambda_{y2}^2)(1+\lambda_z^2) - \lambda_{y2}^2\lambda_z^2}\sqrt{(1+\lambda_x^2)(1+\lambda_z^2)}}$$

$$\delta_{\mathbf{xz}} = \frac{\lambda_x \lambda_y \lambda_{y2} \lambda_z}{\sqrt{(1+\lambda_x^2)(1+\lambda_y^2+\lambda_{y2}^2) - \lambda_x^2\lambda_y^2}\sqrt{(1+\lambda_y^2+\lambda_{y2}^2)(1+\lambda_z^2) - \lambda_{y2}^2\lambda_z^2}}$$

$$\delta_{\mathbf{yz}} = -\frac{(1+\lambda_x^2)\lambda_{y2}\lambda_z}{\sqrt{(1+\lambda_x^2)(1+\lambda_z^2)}\sqrt{(1+\lambda_x^2)(1+\lambda_y^2+\lambda_{y2}^2) - \lambda_x^2\lambda_y^2}}$$

This illustrates, when using partial correlations, that differences can be induced between variables that are not directly involved in the changed pathways between healthy and unhealthy individuals.

7.1.3.1 A potential benefit of using partial correlations in differential networks

In the previous sections we considered the scenario where there were unobserved latent variables responsible for the observed X, Y and Z variables. Now if we consider a scenario where we return to using model (7.1) and observe everything, i.e. the U variables are also observed, we can discuss how this changes the results.

In this scenario we will use the parameters used before for illustration ($\theta = 0.5, \lambda = 0.9, \sigma = 1$). The marginal differential network will pick up all the edges identified before, as well as edges between each of the Xs and the Us (The strongest edges being those between U_1 and each of the Xs - see figure 7.11). This provides us more information than before as the strongest edges are between U_1 and its children ($\delta = 0.67$), followed by the edges between the Xs ($\delta = 0.45$), the edges from the Xs to their “uncles/aunts” i.e. U_2 and U_3 ($\delta = 0.33$) and finally from the Xs to their cousins ($\delta = 0.22$)

However, when we perform a differential network analysis based on partial correlations we see a different picture to before. This time the network among healthy individuals has a strong association between each of the U variables and their children. We see no association between siblings or cousins (as adjusting for the parent ensures that siblings or cousins are not associated). So we see 3 distinct groups, linked via the 3 U nodes (figure 7.12).

In the network formed by unhealthy individuals U_1 is no longer connected to the X nodes and, the strength of the association between U_1 and the other two U nodes is greater than in the healthy network, as adjusting for all the X variables

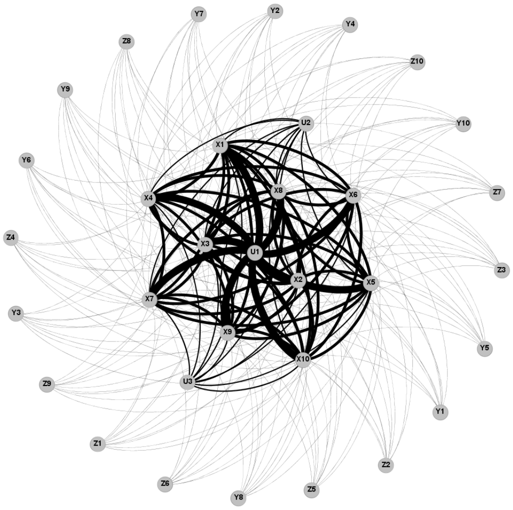


Figure 7.11: Differential network using **marginal** correlations and all variables observed ($\theta = 0.5, \lambda = 0.9, \sigma = 1$) calculated for data generating model 1

no longer affects the partial correlation between U_1/U_2 and U_1/U_3 (figure 7.13).

So the edges that a differential network will identify as non-zero will be all the edges between the Xs and U_1 which are exactly the edges which have been altered between the two states and also weaker edges between U_1/U_2 and U_1/U_3 (figure 7.14). So when all variables contributing to the data generating model are observed a differential network using partial correlations can be thought of as being more *specific* than a differential network using marginal correlations. Indeed this is the ideal situation for which differential networks were devised.

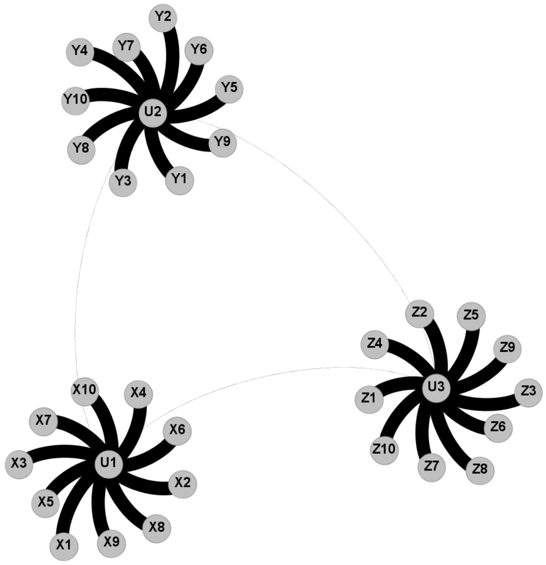


Figure 7.12: Network of healthy individuals using partial correlations obtained for data generating model 1 when all variables observed ($\theta = 0.5, \lambda = 0.9, \sigma = 1$), the thickness of the edge represents the magnitude of δ

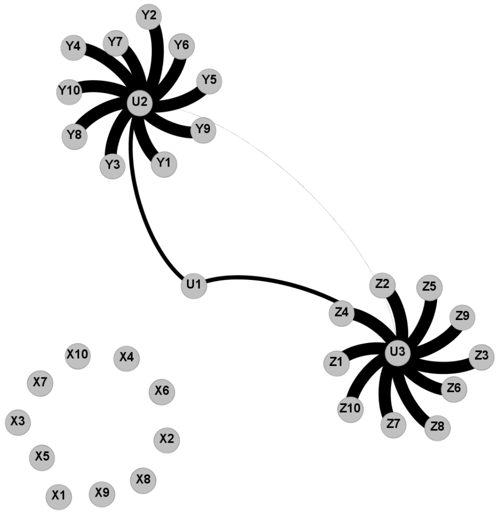


Figure 7.13: Network of unhealthy individuals using partial correlations obtained for data generating model 1 when all variables observed ($\theta = 0.5, \lambda = 0.9, \sigma = 1$), the thickness of the edge represents the magnitude of δ

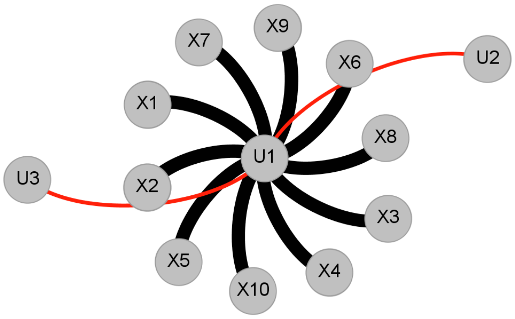


Figure 7.14: Differential network formed using partial correlations obtained for data generating model 1 when all variables observed, the thickness of the edge represents the magnitude of δ, black edges represent a positive δ, red a negative

7.1.4 Marginal or partial correlations in differential networks- which to use?

Marginal correlations have greater power, and the edge estimated between a pair of nodes is not affected by what other nodes are included or excluded in the network. They have the disadvantage in that they do not differentiate between proximal associations and associations that are mediated via other metabolites. The marginal differential network does not provide any information that could not be identified via a series of pairwise logistic regressions, using each possible pair of metabolites in turn. However, visualising this information in the differential network form may enlighten us as to some patterns that may have been difficult to infer from the individual analyses.

In partial correlations it is the reverse, they have the advantage that they are comparing the associations between variables adjusted for the other metabolites in the network, but the edges are dependent on what other nodes are included in the network. As we can see from the examples above, the differential network based on partial correlations is effective in the setting where all relevant nodes have been included. However, where important nodes are omitted, edges are likely to be be identified that are unrelated to the true differences that we are trying to uncover.

7.2 Differential centrality

In chapter 6, some of the methods described in the literature review discuss differential centrality. In this and the previous chapter the main focus has been on the interpretation of a single edge in a differential network, which is important as it is a fundamental building block of the network and needs to be understood. However, as described, there is little power to detect whether or not a specific edge in the differential network has a δ different from 0. So in practice, the benefits of a differential network may be better realised when we consider the network as a whole, incorporating some of the network statistics introduced in chapter 5 into the differential network analysis.

In the interest of keeping the analysis simple and interpretable we suggest using the most basic measure of centrality, degree centrality, as a key network statistic in order to identify nodes of interest in the differential network. That is, the more edges a node has, the more "important" we consider it to be, and it may be a variable that is worth exploring further in terms of its relationship with the disease of interest. So for the estimated differential network, the centrality of each node should be calculated. This will identify nodes which are differentially correlated with the greatest number of other nodes, which could indicate that the node has an important role to play in the development, or diagnosis, of disease.

7.3 Analysis of the BWHHS data

Here we will apply a differential network analysis to the BWHHS data to illustrate this method. A differential network will be estimated using both marginal and partial correlations, the strongest edges identified in each and the nodes with the largest differential centrality highlighted. The nodes and edges of interest will then be explored in more detail to gain an explanation as to why they have been identified. Differential centrality with the differential network will also be considered. For our two comparison groups we choose to compare those who develop CHD in a 12-year follow up period with those who survive to the end of the 12 year follow up without developing CHD (as in previous chapters). Prior to this analysis, we need to prepare the data appropriately.

7.3.1 Data preparation

The data preparation for the BWHHS differential network analysis is exactly the same as that done for the PCR and lasso regressions preformed in chapter 4, with a detailed description found in section 4.4. The sample size was 2922 individuals and 78 metabolites were included in the analysis.

7.3.2 Differential network based on marginal correlations

The differential network formed using marginal correlations is shown in figure 7.15. There are 38 edges with a p-value <0.01 involving 33 different nodes. The average degree among the nodes included is 2.3, with the 3 nodes with the highest differential degree centrality being Tyrosine (*tyr* in the diagram), Valine (*val*) and Glucose (*glc*), with 15, 10 and 7 edges respectively. The 5 edges relating to the 5 largest differences in correlation (δ) are displayed in table 7.1.

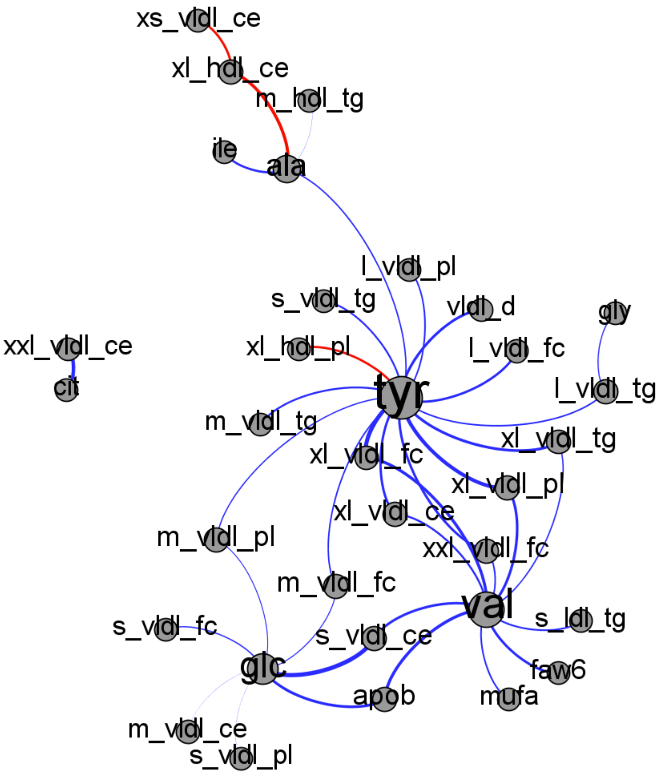


Figure 7.15: Differential network of the BWHHS data formed using marginal correlations - blue = negative δ, red = positive δ. Node size (and label size) proportional to degree centrality of the node.

Table 7.1: Strongest edges in BWHHS differential network using marginal Pearson correlations ($\hat{\delta} = \hat{\rho}_{cases} - \hat{\rho}_{non-cases}$) N=2922

Node 1	Node 2	$\hat{\delta}$
Glucose	Small VLDL Cholesterol Esters	-0.241
Tyrosine	XL VLDL Free Cholesterol	-0.238
Tyrosine	XL VLDL Phospholipids	-0.231
Citrate	XXL VLDL Cholesterol Esters	-0.230
Alanine	XL HDL Cholesterol Esters	0.223

First, we will examine each of these 5 edges in greater detail, starting with the edge with the largest $\hat{\delta}$, which connected Glucose and Small VLDL cholesterol esters which had an observed $\hat{\delta}$ of -0.241. A scatter plot of all the observations of these two variables is shown in figure 7.16, with the individuals who had a

CHD event in the follow up period coloured red and those that did not coloured blue. The estimated Pearson correlation coefficient in the healthy group is 0.17 and in the diseased group it is -0.07 (leading to the $\hat{\delta}$ of -0.24). This difference is not obvious from visual inspection of the plot, but there does appear to be a few outliers due to 3 low value observations of small VLDL cholesterol which are all non-cases. A sensitivity analysis was performed excluding these observations which led to no material change in the estimate of δ suggesting that the large value of $\hat{\delta}$ was not due to these observations. So it appears that among individuals who survived the 12-year follow up period, there was a positive association between Glucose and Small VLDL cholesterol esters, however this association disappeared (or in fact turned to a small negative association) within individuals who went on to develop the disease.

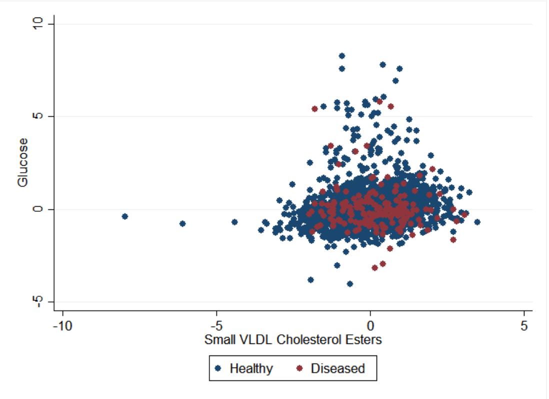


Figure 7.16: Scatter plots of Glucose and Small VLDL cholesterol esters from the BWWHS data with observations from the surviving group in blue, and those who had a CHD event in red. N=2922, $\hat{\delta} = -0.241$

A similar examination was performed on the other 4 edges highlighted in table 7.1 with the scatter plots shown in figure 7.17

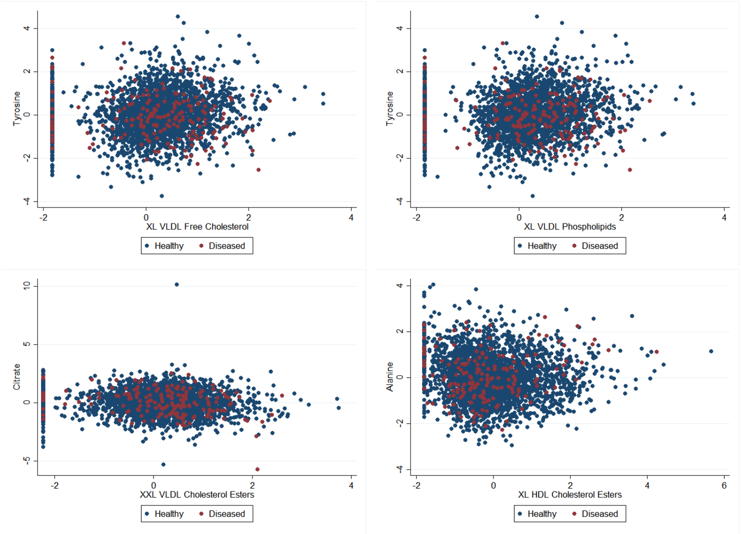


Figure 7.17: Scatter plots of the 2nd, 3rd, 4th and 5th strongest edges in the BWHHS differential network formed using marginal correlation coefficients (N=2922). Observations from the surviving group are in blue, and those who had a CHD event in red

The most striking thing from each of these scatter plots is the column of points on the far left of the plot. This is due to the number of observations equal zero in the cholesterol and phospholipid observations. These have then been transformed using a log transformation where the zeroes are replaced with a value of half the smallest observation, which in these 4 variables equates to an observation ranging from -1.79 in XL HDL cholesterol esters, through to -2.23 in XXL VLDL cholesterol esters.

Starting with the edge between Tyrosine and XL VLDL free cholesterol the Pearson correlation in the disease group is -0.09 and in the surviving group it is 0.15 (leading to a $\hat{\delta}$ of -0.24). To check that the large number of zero values was not having an undue influence on this relationship the correlation coefficients were recalculated excluding them, giving a coefficient of -0.08 in the diseased group and 0.16 in the surviving group (leading to an unchanged $\hat{\delta}$ of -0.24), so the zero values did not appear to have an undue influence on the observed results. The results of the edge between Tyrosine and XL VLDL phospholipids is almost identical, since XL VLDL phospholipids and XL VLDL free cholesterol are extremely highly correlated ($\rho = 0.99$), so the observed $\hat{\delta}$ of -0.231 is made up of a correlation of 0.153 in the healthy group and -0.078 in the diseased. Again, there is no material difference when the zero values are excluded. So, similar to the relationship between Glucose and small VLDL cholesterol esters, it appears that among individuals who survived the 12-year follow up period, there was a positive association between Tyrosine and both XL VLDL free cholesterol and phospholipids, however this association does not exist (or is estimated to be a small negative correlation) within individuals who went on to develop the disease.

The edge between Citrate and XXL VLDL cholesterol esters is again negative $\hat{\delta} = -0.230$, however it differs from the previous 3 edges in that it is made up from a correlation of -0.04 in the healthy group and a correlation of -0.27 in the diseased group. So this time it appears that there is little association between Citrate and XXL VLDL cholesterol esters in those who survive the 12-year follow up period but a moderate negative association between these metabolites in those who have a CHD event. Again, excluding the zero values leads to no material difference in the observed results, but in this instance there is 1 very small observation of Citrate in the diseased group, when this is excluded the $\hat{\delta}$ reduces to -0.196. This estimated δ would still be large enough to be included in the differential network so our interpretations can still remain the same, however it would be among the “borderline” edges included in the network, rather than amongst the strongest.

Finally the only edge resulting from a positive $\hat{\delta}$ is that between Alanine and XL HDL cholesterol esters. There is a small negative association in the healthy group ($\hat{\rho}_{healthy} = -0.094$) but a positive association amongst the diseased group ($\hat{\rho}_{diseased} = 0.129$) leading to a $\hat{\delta}$ of 0.223. This suggests that within the group that have a CHD event in the follow up period, there is a positive association between Alanine and XL HDL cholesterol esters and a negative association within the group that survived the 12-year follow up without having a CHD event. When performing a sensitivity analysis excluding the zero values the observed $\hat{\delta}$ increases to 0.321, due to an increase in the correlation estimated within the diseased group (up to $\hat{\rho}_{diseased} = 0.283$). So the estimated δ does not appear to have been induced by the large number of zero values, in fact it appears that excluding them increases the estimated δ. However, as can be seen on the scatter plot for this edge, the zero values are not really outlying as they are much closer to the observed non-zero values, and there are fewer of them.

Now considering differential network degree centrality we observe that there appear to be three “important” nodes; Tyrosine, Valine and Glucose which have 15, 10 and 7 edges respectively. Of the 15 edges connected to Tyrosine, 13 of them are VLDL metabolites, which are all strongly correlated with one another. So it might be expected that if a metabolite (in this case Tyrosine) was to be differentially correlated with one of the VLDL metabolites, then it would be differentially correlated with many of the other VLDL metabolites that are strongly correlated with one another. Valine is also connected to a number of VLDL metabolites, but also has edges connecting it to monunsaturated fatty acids, Omega-6 fatty acids and apolipoprotein-B, and 6 out of the 7 edges of glucose connect it to VLDL metabolites. So although the differential degree centrality measure may help us identify potentially “important” nodes, the interpretation has to be informed by the inclusion criteria for nodes, if there are a large group of very highly correlated nodes, then it is likely that if one node from within this group has an edge to a node outside the group, then many will. So in terms of interpreting a node with high differential degree centrality as being “important” it is necessary to review, for a differential network based on marginal correlations, the nodes to which it is connected.

7.3.2.1 Adjusting for age and BMI

Adjusting for age had little effect on the results from the differential network, 36 edges were identified as opposed to 38, the 5 strongest edges were the same in both and Tyrosine, Valine and Glucose were still the 3 nodes with the highest differential degree centrality, with 15, 10 and 5 edges respectively. Once BMI was adjusted for, there was a large impact on the observed differential network, with only 2 edges remaining, both involving Citrate. Citrate had an edge in this network to both XXL VLDL cholesterol esters (δ = -0.24, p=0.004) and glucose (δ = 0.20, p=0.008).

7.3.2.2 Comparison with logistic regression

The similarities between the results from estimating the significance of an edge using a differential network approach based on marginal correlations and using the interaction term in a logistic regression were discussed in the previous chapter, where simulations were run comparing the two. Using the BWHHS data, a series of simple pairwise logistic regressions were run, using disease status (at 12 years) as the outcome and each pair of metabolites and their interaction as predictor variables. Similar to the differential network a cut off of p <0.01 was used to identify a “significant” edge.

Many more interactions resulted in a p <0.01 than edges in the differential network analysis, with 121 interactions identified compared to the 38 edges identified in the differential network analysis. Out of the 5 strongest interactions (table 7.2), 4 were also in the top 5 edges from the differential network, which is to be expected given the very close relationship between the coefficient of the interaction term and δ. Figure 7.18 shows the strong association between the two coefficients, calculated for all 78 metabolites. Their estimated Pearson correlation coefficient is 0.85.

Table 7.2: Largest 5 estimated interaction coefficients from pairwise logistic regressions

Metabolite 1	Metabolite 2	Interaction term
Alanine	XL HDL Cholesterol Esters	0.298
Glucose	Small VLDL Cholesterol Esters	-0.265
Tyrosine	XL VLDL Free Cholesterol	-0.265
Tyrosine	XL VLDL Phospholipids	-0.260
Tyrosine	XL HDL Phospholipids	0.260

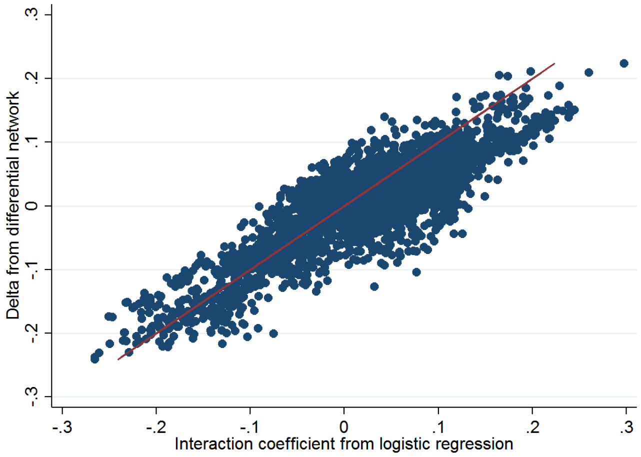


Figure 7.18: Scatter plot of the estimated interaction coefficients from pairwise logistic regressions compared with the $\hat{\delta}$ from the differential network analysis using marginal correlation coefficients on the BWHHS data. Red line is a line of equality. N=2922

Looking at the variables involved in the greatest number of interactions with p <0.01, there were 20 interactions involving XL HDL triglycerides (table 7.3), which was not previously identified as an important node in the differential network analysis, but the node with the 2nd highest number of interactions (15) identified was Tyrosine, which was the node with the highest differential network degree centrality. Glucose and Valine, which were the nodes with the 2nd and 3rd highest degree centralities in the differential network analysis, were 15th and 38th in the list of variables involved in interactions with p <0.01, with 6 and 2 respectively. The top four nodes with the highest differential network degree centrality and the top ten metabolites involved in the most interactions with p <0.01 are listed in table 7.3.

Table 7.3: List of the four nodes with the highest differential degree centrality from the marginal differential network and the ten nodes which were involved in the most interactions with $p < 0.01$ in the pairwise logistic regression analysis.

Rank	**Degree centrality**		**Number of interactions**	
	Metabolite	**N**	**Metabolite**	**N**
1	Tyrosine	15	XL HDL TG	20
2	Valine	11	Tyrosine	15
3	Glucose	7	XXL VLDL CE	13
4	Alanine	4	XXL VLDL TG	10
5	+	11 tied on 2	XXL VLDL FC	10
6			S LDL CE	10
7			S LDL FC	10
8			XL HDL CE	10
9			XXL VLDL PL	9
10			L LDL FC	9

7.3.3 Differential network based on partial correlations

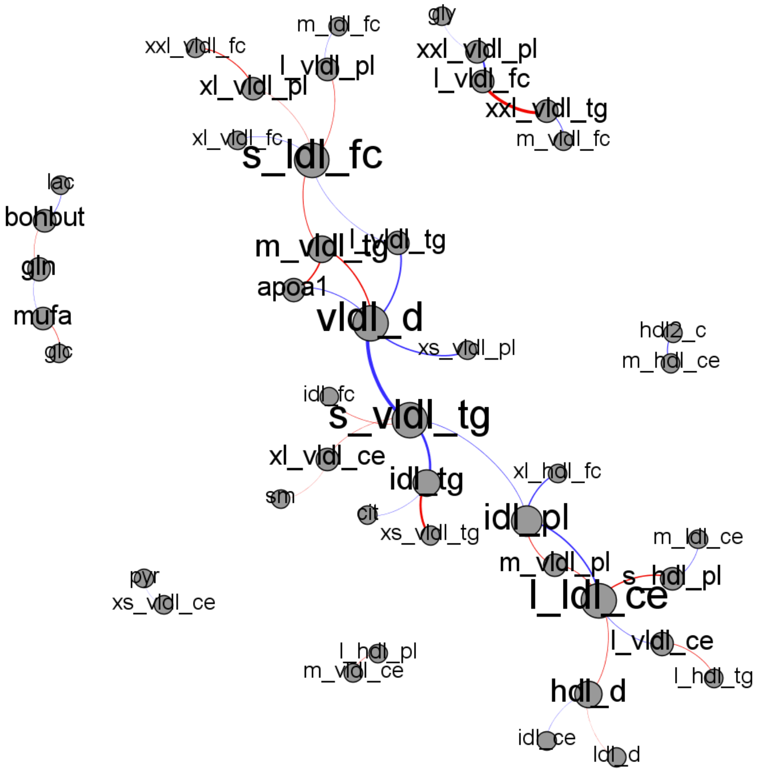


Figure 7.19: Differential network of the BWHHS data formed using partial correlations - blue = negative δ, red = positive δ Node size (and label size) proportional to degree centrality of the node. N=2922

The differential network formed using partial correlations is shown in figure 7.19. There are 42 edges with a p-value <0.01 involving 45 different nodes. The average degree among the nodes included is 1.9, with 4 nodes equally having the highest degree (5): Small VLDL triglycerides, large LDL cholesterol esters, small LDL free cholesterol and VLDL diameter. The 5 edges relating to the 5 largest differences in correlation ($\hat{\delta}$) are displayed in table 7.4.

Table 7.4: Strongest edges in differential network using partial Pearson correlations ($\delta = \hat{\rho}_{cases} - \hat{\rho}_{non-cases}$)

Node 1	Node 2	$\hat{\delta}$
Small VLDL triglycerides	Mean diameter for VLDL	-0.733
XXL VLDL triglycerides	Large VLDL free cholesterol	0.677
Small VLDL triglycerides	IDL triglycerides	-0.578
XS VLDL triglycerides	IDL triglycerides	0.553
XXL VLDL phospholipids	Large VLDL free cholesterol	-0.488

Taking the strongest edge to investigate further, we can perform a regression of small VLDL triglycerides on the other 76 metabolites (excluding VLDL mean diameter) and also a regression of VLDL mean diameter on the same 76 metabolites separately for the diseased and healthy groups. Then by taking the residuals from each regression we can create a scatter plot to investigate what has led to the large difference in partial correlations. As can be seen in figure 7.20 there is a clear association between the residuals of each metabolite among the healthy group ($\hat{\rho} = 0.39$). However among the unhealthy group, the individuals seem to be clustered in the middle of the plot, with the estimated correlation in the residuals (and therefore the partial correlation) being $\hat{\rho} = -0.34$, leading to a $\hat{\delta}$ of -0.73.

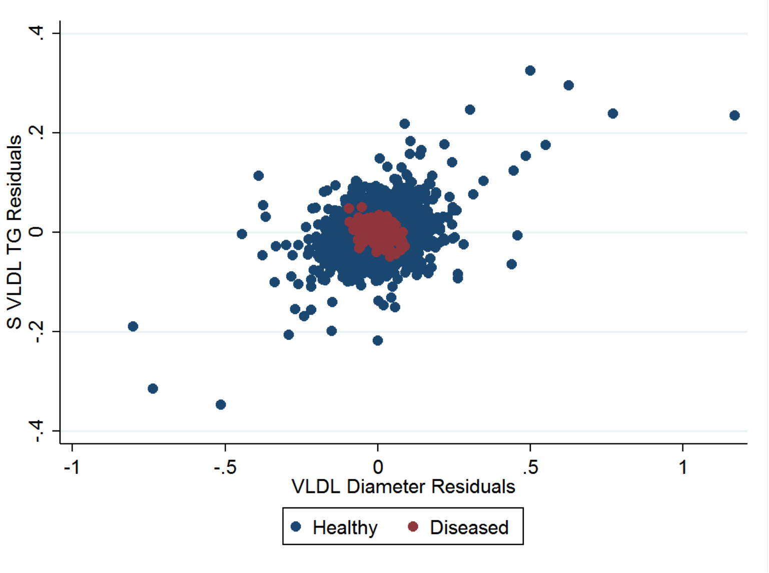


Figure 7.20: Scatter plot of the residuals, after regressing for all other metabolites, of small VLDL triglycerides and VLDL diameter from the surviving group in blue, and those who had a CHD event in red. N=2922

This is a rather strange looking scatter plot so it would be interesting to investigate further what has led to this distribution. First the marginal joint distribution in the healthy and unhealthy groups can be checked, which is illustrated in figure 7.21. It is clear from the plot that marginally the two metabolites are strongly positively associated, in the healthy group the marginal Pearson correlation is 0.86 and in the diseased group it is 0.85.

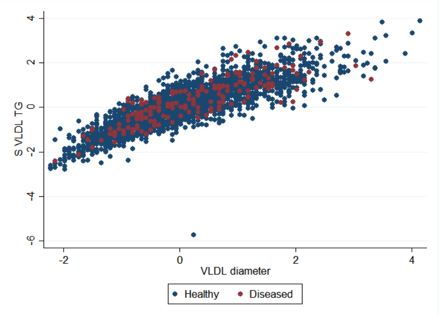


Figure 7.21: Scatter plot of the observations of small VLDL triglycerides and VLDL diameter from the surviving group in blue, and those who had a CHD event in red. N=2922

To try to understand how this relationship changes when other metabolites are added a process will be applied, much like that used in typical epidemiological model building, where we estimate the change in $\hat{\delta}$ after adding each of the other metabolites one by one to the model as a covariate. The metabolite that leads to the largest change in $\hat{\delta}$ is then selected and included in the model and the process is repeated for all the remaining metabolites, each time the metabolite leading to the largest change being included in the model. So the process is as follows:

- For selected edge, estimate $\hat{\delta}$ using marginal correlations, denoted by $\hat{\delta}_{base}$
- Estimate $\hat{\delta}$ for the same edge adjusting for each of the other metabolites one at a time, and identify the metabolite that leads to the largest absolute change from $\hat{\delta}_{base}$
- Estimate $\hat{\delta}$ using partial correlations, only adjusting for the metabolite identified in the previous step. This new adjusted $\hat{\delta}$ is denoted by $\hat{\delta}_1$
- Repeat the process by estimate $\hat{\delta}$ for the same edge adjusting for both the first selected confounder and each of the remaining metabolites one at a time, and identify the metabolite that leads to the largest absolute change in $\hat{\delta}_1$
- Estimate $\hat{\delta}$ using partial correlations, only adjusting for the metabolite identified in the previous 2 steps. This new adjusted $\hat{\delta}$ can be denoted by $\hat{\delta}_2$

- Repeat this process until all metabolites are added into the model and the edge from the full partial differential network is now estimated.

For each interim model the partial correlation in the healthy and unhealthy groups can be estimated as well as $\hat{\delta}$. These estimates for the strongest edge (small VLDL triglycerides and VLDL diameter) are plotted in figure 7.22, with the blue line representing the estimate $\hat{\delta}$, the grey line the partial correlation among healthy individuals and the orange line the partial correlation among unhealthy individuals. The first point on the x-axis represents the estimates from the differential network using marginal correlations, the final point on the x-axis are the estimates obtained using a full partial differential network. (The explanations of the metabolite abbreviations used in the graph can be found in table 3.5).

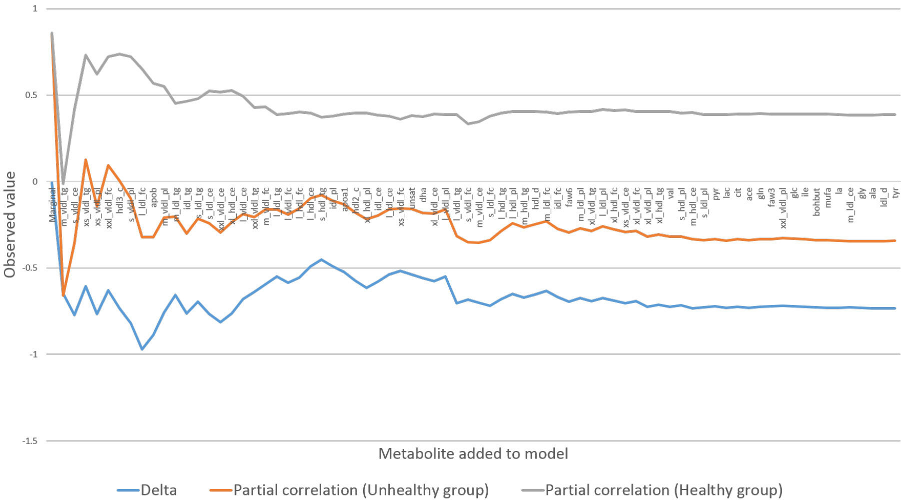


Figure 7.22: Line chart illustrating the changes to $\hat{\delta}$ and the group specific partial correlations as each additional metabolite is added to the differential network, for the edge between small VLDL triglycerides and VLDL diameter.

So, as described earlier, when marginal correlations are used the correlation is high (approx 0.85) in both the healthy and unhealthy groups so $\hat{\delta}$ is close to 0. The metabolite that, when adjusted for, leads to the biggest change in the estimate of $\hat{\delta}$ is medium VLDL triglycerides. This leads to a partial correlation in the healthy group of 0 and of -0.65 in the unhealthy group, giving a $\hat{\delta}$ of -0.65. In addition to plotting the estimated partial correlations and $\hat{\delta}$ at each intermediate model, the scatter plots of the residuals can also be examined. Figure 7.23 shows this for the first 15 metabolites added to the model. From this it can be seen that after adjusting for medium VLDL triglycerides, there appears to be a (small) negative association between small VLDL triglycerides and VLDL diameter in the unhealthy group. However, in the healthy group there are a number of large residuals, leading to a partial correlation close to 0. Investigating these “outliers” shows that the majority of them are observations where medium VLDL triglycerides are equal to 0. Among those individuals who have an observation of 0, there is no-one who went on to develop CHD in the follow up period, so this could be influencing the strength of the adjusted $\hat{\delta}$.

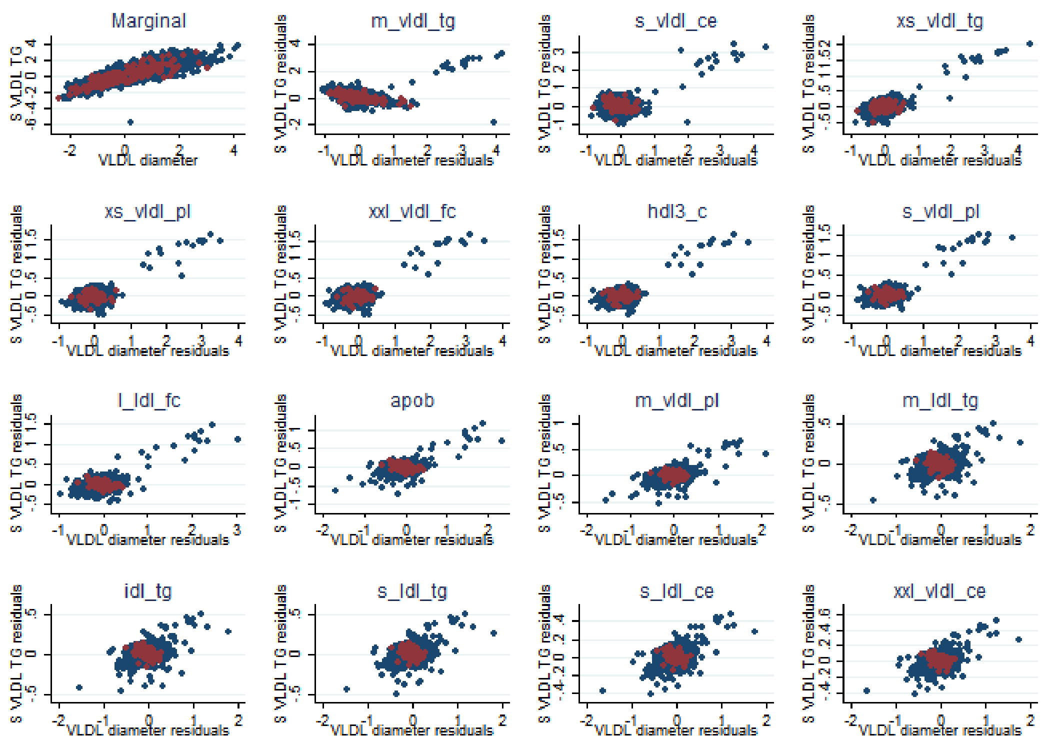


Figure 7.23: Scatter plots of the residuals of small VLDL triglycerides and VLDL diameter in the healthy (blue) and unhealthy (red) groups. The first plot is that from the marginal differential network, and each subsequent plot is adjusted for an additional metabolite (shown in the plot title). So the second plot is the residuals after adjusting for medium VLDL triglycerides, the third plot is after adjusting for both medium VLDL triglycerides and small VLDL cholesterol esters etc.

7.3.3.1 Adjusting for age and BMI

As with the marginal differential network, adjusting for age does not have a great deal of material effect on the results, although there is an increase in the number of edges detected, increasing from 42 to 51, with the strongest edges similar on both analyses. After additionally adjusting for BMI there is also little change. Given we are already adjusting for the other 76 metabolites for each edge in the network, adjusting for BMI over and above that changes little in the analysis, as mentioned previously many of the metabolites are closely associated with BMI so additionally adjusting for it has little impact.

7.3.3.2 Comparison with lasso logistic regression

In our simulations it was identified that effect modification between a pair of metabolites was one scenario which may lead to an edge in a differential network based on marginal correlations, and also that a difference in the marginal correlations in healthy and diseased populations can lead to detection of interaction in a logistic regression. These comparisons were carried out in the very simplified setting with only 3 metabolites. We were able to compare the differential network with marginal correlations to pairwise logistic regressions with interactions, however with partial correlations the edges are adjusted for all other metabolites, so a more plausible comparison is to that of a multivariable logistic regression including all interactions with lasso, as performed in chapter 4. However, there was no overlap at all between the set of edges identified in the differential network based on partial correlations and the set of non-zero interactions in the lasso logistic regression.

7.4 Discussion

7.4.1 Differential network using marginal correlations

From the unadjusted marginal differential network, 3 nodes were identified as potentially important - Tyrosine, Valine and Glucose. If we take Tyrosine, for example, it may be that Tyrosine is a cause of CHD and its edges are due to a δ being induced between it and other marginally associated metabolites. If this was the case we might be able to identify this in a simple univariable logistic regression of Tyrosine on the odds of disease (The results of this give an OR of 0.9, 95% CI 0.77,1.05, p=0.17 - suggesting little evidence of a marginal association between Tyrosine and the odds of developing CHD in the 12 year follow up period). Or it may be that Tyrosine modifies the effect that a number of metabolites have on disease, this would correspond to scenario A in chapter 6, a little weight may be lent to this hypothesis given that a large number of the pairwise logistic regressions with interactions involving Tyrosine resulted in p-values less than 0.01. It also may be that individuals predisposed to develop CHD in the subsequent 12 years have a breakdown (or increase) in the association between Tyrosine and a number of other metabolites, which is the hypothesised situation in scenario C in chapter 6. In any case this analysis would suggest that Tyrosine may be an interesting metabolite to investigate further in its relationship with CHD.

Valine is an interesting case as it was the node that resulted in the second highest number of edges in the differential network (10), but in the pairwise logistic regressions involving Valine only 2 interactions resulted in p <0.01. This perhaps suggests that the relationship between Valine and the disease is potentially more likely to be that described in scenario C than in scenario A in chapter 6.

Adjustment for age makes little difference in the analysis, this may be due to the relatively small age range of women in the study (60-79) and/or may be due to the fact that the metabolites included the study are not strongly associated with age. However adjusting for BMI in the age-adjusted differential network based on marginal correlations makes a large difference to the analysis, resulting in only 2 edges with a p-value <0.01. Many of the metabolites measured are strongly associated with BMI, so by adjusting for BMI in the analysis, much of the within metabolite variation has already been described by BMI amongst both the healthy and the diseased groups. This is likely to move the values of $\hat{\delta}$ towards the null resulting in very few edges being included in the network. This suggests that most of the variability between the groups can be described by BMI. Including a differential network analysis (using marginal correlations, albeit adjusted for BMI) over and above this adds little information. The two edges included in this adjusted network based on marginal correlations both involve Citrate, so it may be worth investigating this metabolite a little further to uncover whether this metabolite is of interest.

However, when interpreting all these observations it is important to remember the cut off p-value of 0.01 is an arbitrary threshold. We are testing 3003 potential edges so if we were to use a Benjamini-Hochberg adjusted threshold a more appropriate p-value cut-off would be .000017. If we impose this threshold we do not see any edges reaching statistical significance, so the evidence of the edges described is very weak, however, given this is an exploratory analysis being used to identify candidates for further investigation it may be acceptable to use such a liberal threshold.

On the whole the marginal analysis could be recreated by performing a series of pairwise logistic regressions with interaction terms, so the benefit of a differential network using marginal correlations is limited, since the pairwise regressions are a more established method and their interpretation more understandable to audiences. However, it may be that by using the differential centrality measure on a marginal differential network that an important node could be picked up as it modifies the effect of a number of other metabolites. So as a first step it will be useful to perform a differential network using marginal correlations, to pick up anything missed from a prior standard analysis, before going on to perform the differential network using partial correlations, which adds a new element to the analysis.

7.4.2 Differential network using partial correlations

From the differential network based on partial coefficients we observed a greater number of edges in the network than in the marginal differential network but with fewer obviously central nodes, so the network could be described as being much more dispersed. More of the VLDL metabolites are involved in this network than in the marginal differential network, suggesting that, after adjustment for all other metabolites, these VLDL variables may be modifying the effect on disease of other variables in the network. Or they could be strong, independent risk factors for disease and the estimated δs are echoes of this strong effect. It is also possible that the association between these pairs of variables is modified by an unmeasured variable that is associated with the risk of an individual developing CHD in the subsequent 12 years.

However, given the nature of the BWHHS data it may be that we are observing the unexpected effects described in section 7.1.2.1. The BWHHS data has many highly correlated metabolites, many of which are known to be associated with CHD. Given that we are therefore adjusting for a number of highly correlated metabolites we may be not picking up some "true" edges ("true" in the sense that they really do provide some information about the aetiology of CHD), and we may be picking up some spurious edges (spurious in the sense that they are only very loosely related to any differences in pathways between the diseased and healthy groups).

When investigating the single edge between small VLDL triglycerides and VLDL diameter it was interesting to note the joint distributions (of the residuals) in the diseased and healthy groups. In the healthy group there was a wider range of residuals and a strong positive correlation. In the diseased individuals, the residuals were constrained to a very small range.

Again when interpreting the findings of this differential network the role of chance is the most plausible explanation for any extreme results observed. Given the high number of tests performed and the lack of multiple testing adjustment we would expect to observe a number of edges by chance. We have chosen a liberal threshold for our cut off to allow us to explore potential candidates of nodes and edges that may be related to CHD, but it may just be that there is no signal within this data to uncover.

7.4.3 Limitations and alternative strategies

We chose the metabolite transformations by choosing the "best" transformation out of 3 options, however this may not be a suitable strategy when performing a differential network. By transforming metabolites in different manners we are potentially making it less plausible that there is a linear association between the pair of variables, it may be unlikely for there to be a linear association between the cube root transform of one variable and the log transform of another. So it might be a better strategy to choose the single most effective (on average) transform to be used for the whole metabolite set. Another solution to this problem could be to use Spearman correlations instead of Pearson correlations as the measure of association. This would have the additional benefit of reducing the impact of data where there a high number of zero values.

There are problems with both of the two methods of differential networks proposed (using marginal or partial correlations). When using marginal correlations it is possible the the edges observed are explained via other nodes in the network, so the edge identified may relate to a distant association via other variables rather than a direct association between the two nodes. At the other extreme we use partial correlations to measure the association between a pair of nodes, adjusting for every other node in the network. When we have the situation where there are a high number of nodes in a network and a large proportion are strongly correlated with one another we have little power to identify an edge. Two potential strategies for dealing with this could be:

1. More stringent variable selection criteria
2. Limit the number of variables adjusted for

The first of these two methods is the simpler approach, it is simply to use the partial correlation as our measure of association but to exclude variables that introduce high collinearity into the model. Previously only variables that introduced perfect collinearity were excluded (i.e. when one variable was a sum of two or more others) but it may have been a better strategy to set a pragmatic threshold for excluding variables that are extremely correlated with one another.

The second is to perform an exploration of the data between the two extremes described. Rather than adjusting for none (as in marginal), or all (as in partial), of the other nodes in the network, we could adjust for a subset of representative nodes. A possible way to select this set of representative nodes could be to generate a network using all our data. Then we find the network modules defined by this overall network, we could select the most central node within each network module as a representative set of nodes from the network and adjust for only those nodes.

Finally, due to the long follow up time and the age of the participants in the study, this analysis is subject to bias due to competing risks (704 participants died of non-CHD related causes during follow up and were therefore excluded from the analysis). In the following chapter I will describe a potential method that will take into account the time to the CHD event, which may go some way to alleviating this problem.

Chapter 8

Differential networks and time to event analysis

In chapter 7, the BWHHS cohort was analysed using a differential network analysis comparing the group of individuals who had a CHD event in the 12 year follow up period to those who survived to the end of the 12 years without a CHD event. This meant that 703 individuals who were lost to follow up (i.e. died of a non-CHD related cause in the follow up, without having had a CHD event) were excluded from the analysis altogether. This could lead to bias in the analysis as those that were excluded may have been more likely to have a CHD event and also may have had quite a different metabolomic profile to the included groups, so the differential network analysis may have been based on comparing groups, one of which was selectively depleted of frailer individuals.

Table 8.1: Number of women who had suffered a CHD event, died or survived by the end of each year.

Year	At risk	CHD	Died	Survived
1	3625	26	25	3574
2	3574	36	60	3529
3	3529	50	100	3475
4	3475	71	146	3408
5	3408	85	204	3336
6	3336	101	268	3256
7	3256	114	332	3179
8	3179	134	400	3091
9	3091	149	480	2996
10	2996	158	549	2918
11	2918	170	621	2834
12	2834	182	703	2740

Table 8.1 shows the cumulative number of women who have had a CHD event,

died or survived by the end of each year of follow up, with figure 8.1 showing the estimated cumulative incidence of death and CHD. These illustrate that death is a much more common outcome than CHD, so the competing risk of death in this analysis may have a significant influence on the results obtained.

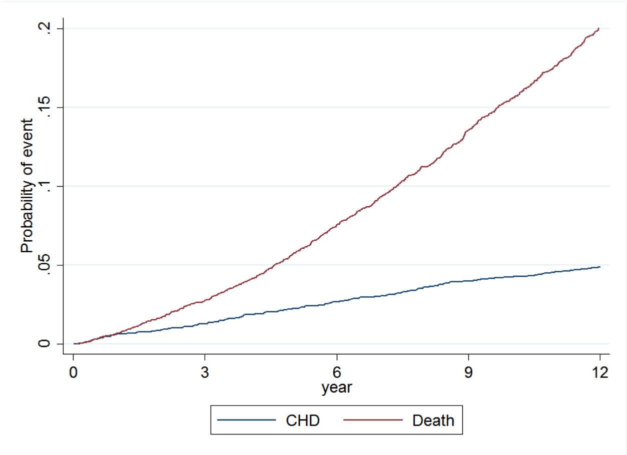


Figure 8.1: Plot across time of the probability of an individual having died or having experienced a CHD event.

Furthermore, because metabolite concentrations can change substantially over time, it may be that the concentrations of these metabolites measured at baseline are not good indicators of CHD risk over a 12-year period, but may provide some indication of the risk over a shorter period of time.

The objective of this chapter is to investigate the possibility of adapting the differential network method to also include the information about the time to CHD event in the analysis, this may be a more appropriate method of using differential networks with cohort data. The proposed method is to perform a differential network analysis at yearly intervals and "important" metabolites identified. These metabolites will be compared to those identified by the Cox regression with lasso in chapter 4 to check if the two methods highlight the same variables associated with CHD. So the 2 aims of this chapter are:

1. To incorporate information on time to event into the differential network analysis
2. To compare the results from the time to event differential network analysis to results from the Cox/logistic regressions with lasso, to identify if similar sets of important metabolites or important interactions are identified.

8.1 Methods

As before all individuals with prevalent CHD at baseline will be excluded from the analysis. At each yearly time point individuals will be categorised into those that have had a CHD event, those that have survived to that time point and those lost to follow up. Those lost to follow up at each time point will be excluded from the analysis and a differential network analysis using partial correlations will be performed. The differential network will be age-adjusted by regressing each metabolite on age and using the residuals from that regression as the covariates for the analysis. Again, a p-value of 0.01 will be used as the threshold for inclusion of an edge.

In the first 4 years of follow up, there are fewer events than there are nodes in the network (78) so the partial correlation coefficients cannot be estimated without using the "shrinkage" method described earlier in section 5.3.1.1. Because of this, the analysis will be limited to estimating the networks formed at 5 years through to 12 years.

Also, for another comparison, lasso logistic regressions will be performed at each of the same timepoints, with the outcome of each being whether an individual has had an event by the same follow up times as above. In the lasso regression age will also be included, as well as all metabolites and all pairwise interactions between metabolites. No conditions will be attached to whether a coefficient is included so it is possible that an interaction may be selected as an non-zero coefficient in the model but the main effect of the metabolites involved may be set to zero. This is done as it is potentially more comparable to the differential network method being proposed than the Cox regression is. The methods for the Cox and logistic lasso analyses are described in section 4.3

8.2 Results

8.2.1 Differential networks

The main characteristics from the networks estimated at years 5-12 (8 differential networks) are described in table 8.2. At 5 years only 3 edges were included in the differential network, this is because there were only 81 events after 5 years of follow up, and with 78 nodes in the network the partial correlation of the cases would be very imprecisely estimated, so it would require a very large $\hat{\delta}$ to attain a p-value <0.01. However from 6 years onwards there were a greater number of edges included in the differential networks, with the numbers of edges ranging from 24 (at year 11) up to 72 (at year 6). The average degree (among nodes included in the networks) ranged from 1.41 in year 10 up to 2.73 in year 7.

Table 8.2: Characteristics of the 8 differential networks estimated.

Year	CHD group (N)	Survived group (N)	Nodes Included	Edges Included	Average Degree
5	85	3336	5	3	1.20
6	101	3256	54	72	2.67
7	114	3179	49	67	2.73
8	134	3091	40	47	2.35
9	149	2996	34	27	1.59
10	158	2918	37	26	1.41
11	170	2834	32	24	1.50
12	182	2740	39	29	1.49

Of the 3003 possible edges that could be included in each network, 2836 were not included in any of the 8 networks, leaving 167 edges that were included in 1 or more networks and 106 of these were only included in 1 network. No edge was identified in all of the networks (due to the fact that almost no edges were identified in the 5 year network) but 1, the edge between XXL VLDL triglycerides and large VLDL free cholesterol was found in the 7 networks from year 6 onwards. The correlations within the cases and the non-cases along with $\hat{\delta}$ are tabulated in table 8.3. It is also possible to plot the size of $\hat{\delta}$ for each of the 167 edges included in any network to identify any trends across time, which is done in figure 8.2 with the 3 edges described above highlighted.

Table 8.3: Table of correlations in cases, non-cases and $\hat{\delta}$ for each of the three edges identified as the strongest in the time to event differential network analysis

		Year							
		5	6	7	8	9	10	11	12
XXL VLDL TG / L VLDL FC	ρ_{case}	0.16	0.64	0.43	0.51	0.51	0.49	0.47	0.47
	$\rho_{non-case}$	-0.21	-0.22	-0.22	-0.22	-0.22	-0.21	-0.22	-0.22
	$\hat{\delta}$	0.37	0.86	0.65	0.73	0.72	0.71	0.69	0.68
L VLDL CE / M VLDL FC	ρ_{case}	-0.71	-0.73	-0.65	-0.51	-0.34	-0.34	-0.34	-0.36
	$\rho_{non-case}$	0.08	0.09	0.09	0.09	0.07	0.08	0.08	0.08
	$\hat{\delta}$	-0.79	-0.81	-0.74	-0.60	-0.42	-0.42	-0.42	-0.43
S VLDL TG / VLDL D	ρ_{case}	-0.17	-0.32	-0.41	-0.42	-0.47	-0.45	-0.47	-0.45
	$\rho_{non-case}$	0.23	0.25	0.25	0.26	0.32	0.31	0.30	0.26
	$\hat{\delta}$	-0.40	-0.57	-0.67	-0.67	-0.78	-0.76	-0.77	-0.71

It can be observed that the estimated values of $\hat{\delta}$ in general tend towards the null as time progresses, although there are two edges that clearly persist and that are larger in magnitude than all the other observed edges. These are the edge between small VLDL triglycerides and VLDL diameter (the large negative δ) and the edge between XXL VLDL triglycerides and large VLDL free cholesterol (the large positive δ). Both these were picked up as strong edges in the original 12-year differential network analysis, but there is also one edge that appears strongly in years 5,6 and 7 before becoming less strong in the later years - this is the edge between large VLDL cholesterol esters and medium VLDL free cholesterol.

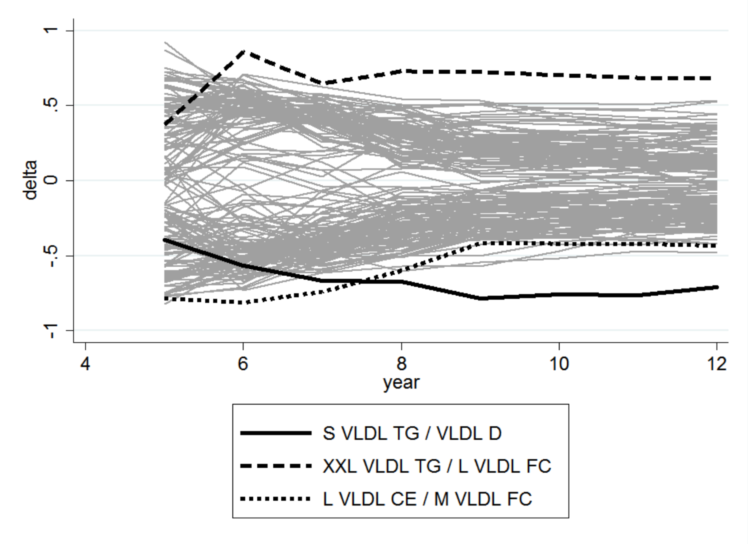


Figure 8.2: Each line on the plot represents an edge included in the network in at least 1 timepoint, the x-axis represents time and the y-axis the estimated $\hat{\delta}$.

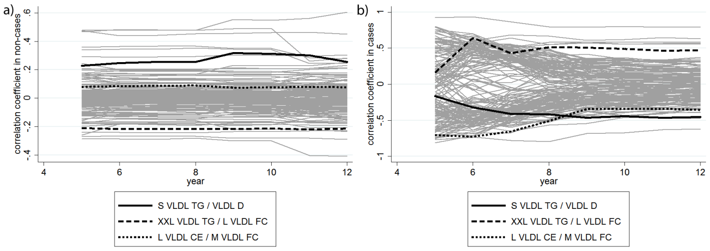


Figure 8.3: Each line on the plot represents a pair of nodes in the network, with the observations in sub-figure a) representing the Pearson correlation coefficients in the non-cases and in sub-figure b) the Pearson correlation coefficients in the cases.

If we look at the correlations within the cases and the non-cases across the years it is apparent that the estimates of the correlations within the non-cases are stable across the years, whereas it is the estimate of the correlation in the cases that drives the changes in $\hat{\delta}$ (figure 8.3).

Some of what is observed may be due to the different population sizes at each year, this can be equalized by subsampling from the cases and non-cases at each time point to be equal to the smalles number of cases and non-cases. The number of cases at year 5 was to small so this analysis only took place between years 6 and 12. So 101 cases were randomly sampled from each year and 2740 non-cases. The same analysis was then performed, identifying the values of $\hat{\delta}$ at each timepoint for all edges, and the same graph as before produced (figure 8.4). Fewer edges (217) were identified at any time point than previously due to reduced power. Looking at the three interesting edges from the initial analysis they seem to follow a similar pattern as before, although the edge between XXL VLDL TG and L VLDL FC is much reduced at year 12.

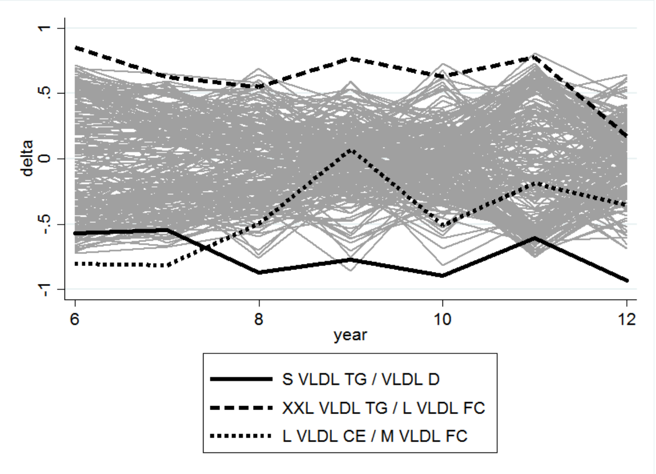


Figure 8.4: Each line on the plot represents an edge included in the network in at least 1 timepoint from the subsampled data analysis, the x-axis represents time and the y-axis the estimated $\hat{\delta}$.

Figure 8.5 shows a series of scatter plots, one for each year analysed, where each point on the plot represents each of the 3003 potential edges in the networks, with the position on the x-axis representing the correlation between the pair of variables of that particular edge in the non-cases group, with the y-axis representing the same but in the cases group. The small blue circles are edges where the p-value has not reached the threshold of 0.01 and are therefore not selected in that year's network and the larger red circles are those edges that are included in the network. The edges selected in year 5 are only those where the Pearson partial correlation is of an opposite sign in the cases and the non-cases, as the years progress and the sample size (in the cases group) increases, more subtle differences are picked up. But in general it appears that the edges identified are ones where there is a partial correlation with a large magnitude in the cases group and a correlation close to 0 in the non-cases group.

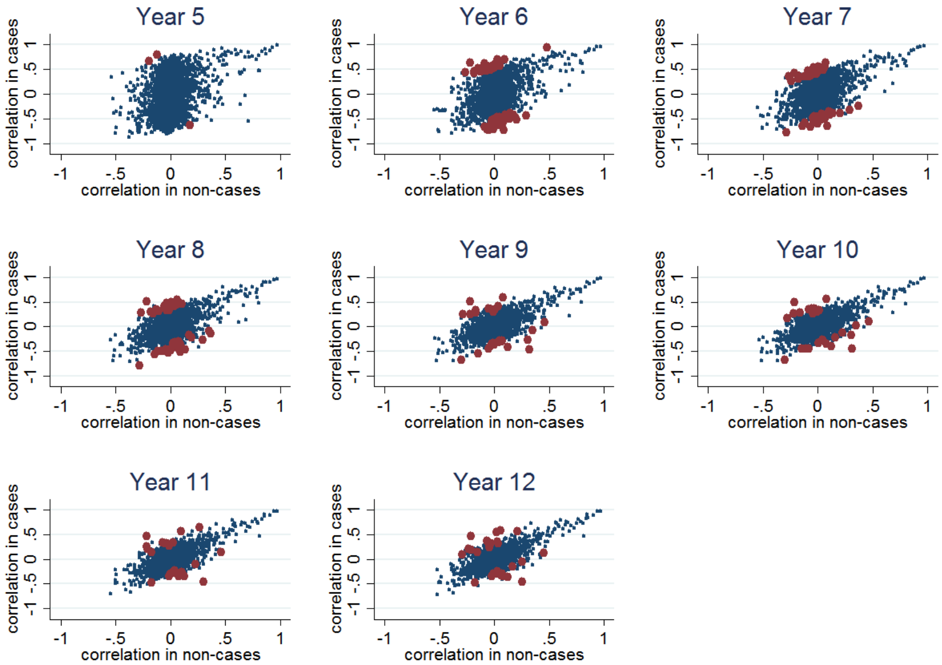


Figure 8.5: Each plot is a scatter of the 3003 potential network edges, with the x-axis representing the correlation in the non-cases group and the y-axis in the cases group. The small blue circles are the potential edges that do not have a p <0.01 so are not included in the year-specific differential network, whereas the larger red circles are the edges included in each differential network.

When considering degree centrality, the nodes identified as more central in the analysis restricted to the final time point (12 years) were described in the previous chapter, but there are 2 additional nodes that appear interesting in these year specific analyses. Small LDL phospholipids had a very high degree centrality in the differential networks generated at years 7 and 8, with a centrality of 8 and 11 in each of these networks. Similarly Glutamine had 8 and 6 edges in years 7 and 8 respectively, but for both these nodes, the degree centrality dropped off to 0 by the 12 year follow up. It may be that these metabolites are associated with an increased risk of CHD in the shorter term, but as time progresses the association becomes weaker.

8.2.2 Lasso regression

In chapter 4, the BWHHS data was analysed using a Cox regression model with lasso, to identify the metabolites most strongly associated with time to CHD event, as well as identify any non-zero interaction terms. The results from this Cox regression are contained in table 4.8

In the lasso regression analysis, coefficients are estimated for both the main association of the disease to each metabolite and all 2-way interactions between them. Lasso logistic regression analyses were carried out separately on the same year-specific datasets used in the previous section, with the λ parameters for each year allowed to be selected by the data at each wave. The summary of the numbers of non-zero coefficients in the analyses is shown in table 8.4. There is no overlap between these results and the results from the differential network analysis above, in fact not one of the edges selected by the differential network matches and of the non-zero interactions identified by the Cox regression with lasso. However, the time to event method used in the differential network analysis is quite different to that used in a Cox regression, and a method potentially more comparable to it would be to perform a logistic regression at each of the same points in time. Table 8.4 denotes the number of non-zero terms (main and interaction) identified at each yearly timepoint.

Table 8.4: Results from the 8 lasso regressions performed.

Year	Main effects identified	Interactions identified
5	2	0
6	3	16
7	5	33
8	3	17
9	3	18
10	5	33
11	5	37
12	4	33

As with the differential network analysis, at the 5 year time point there were very few coefficients (either for the main metabolites of any of their interactions) identified as being associated with risk of CHD. There were 3 metabolites identified as associated with risk of CHD at each of the time point from 6 years onwards. These were:

- IDL triglycerides, which were associated with an increased risk of CHD by each of the years 6-12
- Total cholesterol in HDL2, which was associated with a decreased risk of CHD by each of the years 6-12

- Monounsaturated fatty acids; 16:1, 18:1, which were associated with an increased risk of CHD by each of the years 6-12

There were also 8 interactions that were identified at each time point (apart from at year 5), these were interactions between:

- XXL VLDL cholesterol esters and small LDL free cholesterol
- XXL VLDL cholesterol esters and Citrate
- XL VLDL free cholesterol and Tyrosine
- XL HDL phospholipids and estimated degree of unsaturation
- XL HDL cholesterol esters and XL HDL triglycerides
- XL HDL cholesterol esters and Alanine
- Glucose and Citrate
- Glycine and Isoleucine

These are very similar to the set of interactions identified by the Cox regression, and as before they do not match up with any of the edges detected in the differential network.

8.3 Discussion

We have attempted to address a limitation of differential networks by extending it to incorporate time to event information, and in doing so have investigated whether associations between metabolites and CHD vary with time. The differential network analysis performed illustrated some of the difficulties associated with the method. In the early years there were too few observations to obtain satisfactory estimates for the edges in the network. Although this analysis has been framed as "time to event" as it includes information about the time until an individual has had a CHD event, but practically it treats time to diagnosis as a categorical variable and performs a series of differential networks at each time point, so we are considering whether the networks are stable across time, or whether they vary. An alternative to differential networks that uses a continuous outcome has been proposed by Valcarcel [105], and could have been a different approach taken in this chapter.

In the earlier years (5/6/7) the partial Pearson correlation is very imprecisely estimated due to small numbers, so although in those years there are a higher number of edges detected these may be an artefact of this imprecision. However the higher numbers of edges detected here could also be that the metabolite profile of an individual is more closely associated with CHD events that occur within the first 7 years, and as time progresses the association becomes weaker.

There is one edge, that between large VLDL cholesterol esters and medium VLDL free cholesterol, which exists in the earlier networks but not in the later ones. This could just be a chance finding due to the imprecisely estimated correlation coefficients in the cases group or it could be that this edge is associated with CHD risk in the earlier years of follow up, but less so as time progresses.

Small LDL phospholipids and glutamine were identified as potentially "important" nodes in the 6/7-year risk of CHD networks, but not in the later years. It may be that these nodes are associated with CHD risk in the shorter term, but are not associated with the risk of CHD in the full 12 years.

There is no overlap between the findings of the differential network analyses and the lasso regression. This may be due to the fact that they are picking up different features in the data, or it may be due to the fact that there is little evidence of interactions between any of the investigated metabolites.

Chapter 9

Discussion

The main aim of this thesis was to evaluate statistical approaches for the exploration a high dimensional set of metabolic measurements, to identify whether any are involved in the aetiology of incident CHD events. The results would inform future studies of the mechanisms suggested by this exploration. We have used metabolomic data available within BWHHS as both a motivation and for illustration.
This aim led to the following two objectives:

1. Understanding the main features, and investigating the reliability, of the BWHHS metabolomic data.

2. Comparing statistical approaches for dealing with high dimensional data in terms of their suitability for aetiological investigations, focusing in particular onto the newly proposed differential networks (DN) approach. This also involved examining whether and how extensions of this approach to deal with time to event outcomes are suitable.

The first of these objectives was addressed in chapter 3 and the second addressed in chapters 6, 7 and 8, with chapters 4 and 5 providing some background.

9.1 Metabolomic data

Main features
The descriptive analysis of the BWHHS metabolomic data identified four main issues that must be addressed when including them in any statistical analysis on these data. These are:

1. High dimensionality

2. Correlation structure

3. Skewness

4. Zero values

The first issue, high dimensionality, is what first motivated the work of this thesis. When the aim of the analysis is to investigate the role of several metabolites on a particular outcome and the number of potential variables is high (relative to the sample size), an approach that relies on standard regression modelling may not be not be suitable. This is because the strategy of fitting say a linear or logistic regression model that includes all these potential explanatory variables would either not be estimable, if the number of regression coefficients exceeds the number of observations, or have extremely imprecise estimates because of lack of information. Some commonly applied methods for dealing with this are described and implemented in chapter 4.

The second issue, strong correlation structure, can lead to substantial problems when performing certain statistical analyses. This would occur when fitting any type of regression model that includes highly collinear variables. When this happens, the regression coefficients of all included variables are not, or are poorly, identifiable. There were a number of metabolites in the BWHHS which were defined as combinations of other metabolites that were also included in the data. Inclusion of all these metabolites in the same regression model would introduce perfect collinearity among some of the explanatory variables. For this reason it was decided to exclude those metabolites which were functions of other metabolites in the dataset. However, a very strong correlation structure among the metabolites remained even after these exclusions. This could be dealt with by implementing dimension reduction techniques, such as principal component regression performed in chapter 4. However when the aim is variable selection as opposed to dimension reduction, the results can be unstable when multicollinearity is present. Methods designed to deal with high dimensional data, such as Lasso regression, can mitigate this instability to a certain extent.

The third issue we encountered is skewness. Many of the metabolites in the BWHHS have distributions that are highly skewed, leading to considerations of transformations for analyses that require the data to be approximately normally distributed, as described in section 3.3.1. Different metabolites have different distributions and therefore certain transformations are more or less appropriate for different metabolites. In the applications it was decided to choose the transform most suitable for each metabolite separately, in order to obtain distributions for each that were closest to normality.

Finally, the fourth potential issue we identified was that a number of metabolites had many observations equal to exactly zero. These may be true zeroes, or they may be metabolite concentrations below the minimum threshold for detection. This results in a non-symmetric distribution of observations, which may cause problems when performing statistical analyses that require an assumption of a normal, or just a symmetric distribution. This would also cause a specific problem when attempting a log transformation for a selected metabolite if it included zero values, so a value half the size of the minimum observed value was used to replace zero in any metabolite that was log-transformed. The importance of this issue was highlighted in section 7.1.2, when zero values appeared to influence the results of a differential network analysis when defined using partial Pearson correlation coefficients.

Reliability of NMR metabolomic data
The short-term reliability of the NMR biomarkers available in the BWHHS was assessed using the data available on 37 women who provided a second blood sample a week after providing their first. Short-term reliability in this context refers to the consistency of biomarkers between two different blood samples within a single individual taken 1 week apart. The variation in the samples could be due to biological changes, differences in the method the sample was taken and measurement error. Of the 228 NMR biomarkers measured, 25(11.0%) were classed as having "poor" reliability, 135(59.2%) as having "good" reliability and 68(29.8%) as having "excellent" reliability. This is of interest because if the concentration of a metabolite is not well correlated with its concentration 1 week later then it is unlikely to be a stable indicator of long-term underlying health problems. These estimates of reliability were not as good as the estimated reliability of the 37 biomarkers obtained using standard methods from which 1(2.7%) was classed as "poor", 21(56.8%) were classed as "good" and 15(40.5%) were classed as "excellent". A key limitation of the generalizability of these conclusions was the long storage times for the samples used to obtain the NMR biomarkers - the blood samples were frozen for 11-13 years before ^{1}H-NMR spectroscopy was performed. Which could have contributed to the greater variation seen within the NMR biomarker concentrations compared with the standard biomarkers. Also the sample size was small (37) so the confidence intervals of the estimated intra-class correlation coefficients, used as the measure of reliability, were quite wide.

Finally, for those metabolites that had been previously measured using standard techniques, the agreement between the concentrations obtained using standard methods and the concentrations obtained using ^{1}H-NMR spectroscopy was assessed. There were some systematic differences in the absolute concentrations for three of the five metabolites measured using both techniques. However, in all five, both measurement methods were strongly correlated with one another, so for the purposes of assessing whether a metabolite is associated with an outcome of interest, the differences in concentration identified by the two methods should not be a cause of concern because the association between baseline LDL cholesterol concentration and CHD incidence can be identified using either method.

The results above were used to inform the analysis as reported in the remainder of the thesis, in chapter 4 where methods used for analysing data with a large number of exposures were applied to the BWHHS metabolomic dataset, and in chapter 6 onwards, where differential networks were explored in depth.

9.2 Differential Networks

In chapter 4 the methods of dimension reduction (using principal component regression) and variable selection (lasso regression) were described and applied to the BWHHS data, using both time to CHD event and CHD event in the follow up period as the outcomes. These two methods were selected as illustrative of two general classes of high dimensional data methods. They are useful for identifying individual metabolites, or groups of metabolites that are associated with the outcome. However, an alternative approach, the emerging method of differential networks, was explored in chapters 6 to 8, and used to identify if the association between pairs of metabolites differed in two groups.This was done first through simulations, then by implementing the method on the metabolomic dataset from the BWHHS and finally by exploring the possibility of its use with time to event data.

The literature review performed (described in chapter 6) identified that there was no one definitive differential networks method, with different researchers proposing various ways of defining a differential network, however most were based on an edge being defined as the difference in a measure of association between a pair of nodes in two different groups.

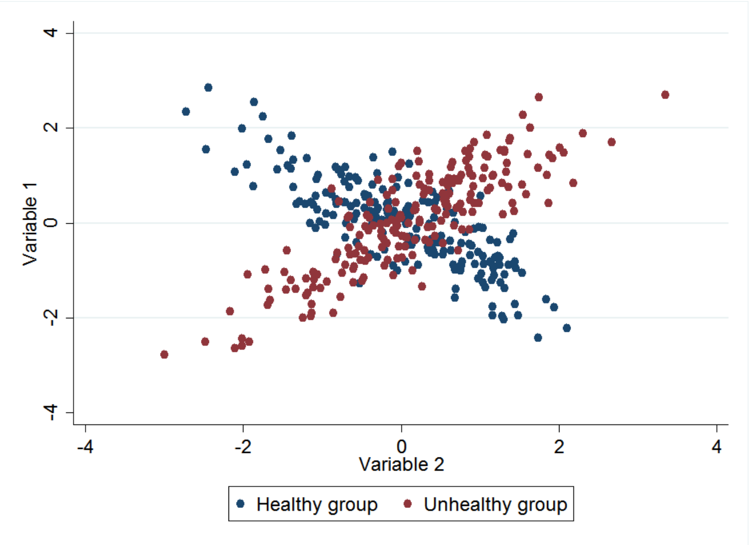


Figure 9.1: Scatter plot of simulated observations from a pair of variables, with data points coloured by groups - blue: healthy individuals, red: unhealthy

One topic not addressed in the literature was the interpretation of what an edge in a differential network means. In the typical association networks described in chapter 5 an edge between a pair of nodes represents an association (often a correlation) between those nodes. In a differential network an edge is a more complicated concept, many papers using a simple diagram as shown in figure 9.1 to show what the distribution of a pair of variables that have different correlations in two groups looks like. However this led to a question - what situations could lead to such a distribution of variables? The main applications considered in this thesis concern the investigation of the aetiology of a particular disease. Hence, if a pair of variables are identified that have a distribution as shown above, how does this inform us about the mechanism leading to the outcome of interest? To investigate this three potential data generating models were hypothesized and a series of simulations performed to observe what differential network would result from each generating model.

9.2.1 Simulation study

There were three main purposes for carrying out the three sets of simulations:

1. to relate the data generating structure to the resulting differential networks
2. to relate the differential network results to those obtained from standard regression models
3. to assess the performance of differential networks under some realistic scenarios

In the literature we reviewed, only two papers [72, 88] had performed simulation studies in order to assess the performance of the method. Both papers opted for simulations where the cases and non-cases were generated separately, with a different marginal correlation defined for each group. Such simulated data did indeed lead to a different correlation structure in the two groups, but their generation (and then examination) did not consider alternative scenarios in which such data could arise.

Alternative data generating models

In order to keep our analysis manageable, our simulations were limited to estimating a single edge in the differential network, so only investigating the relationship between one pair of variables, while allowing for the presence of at least another, all encompassing variable. As the edge between the pair of variables is the fundamental building block of a differential network, understanding which situations give rise to the identification of an edge should help in interpreting the results from a more complex differential network analysis. Three scenarios were hypothesized and used to define the data generating models for the simulations, these were:

(A) A pair of variables are joint causes of disease

(B) Disease modifies the joint distribution of a pair of variables

(C) There is a common cause of the disease and of the joint distribution of a pair of variables

We considered these three scenarios exhaustive of the mechanisms that would give rise to an edge between metabolites. Scenario A reflects the temporal order of the variables collected during a cohort study (with the metabolites measured at entry among disease free individuals and disease occurring at a later time).

Scenario B would most likely arise in a cross-sectional study where disease status is measured at the same time as the metabolites. However it could also occur in the context of a cohort study if disease is latent at the time of recruitment. Scenario C could occur in either cohort or cross sectional studies and possibly represents a very likely scenario when the variables considered are not on the causal path to disease.
Diagrams with proposed hypothesized causal models were produced for each of these scenarios, illustrated in figures 9.2, 9.3 and 9.4

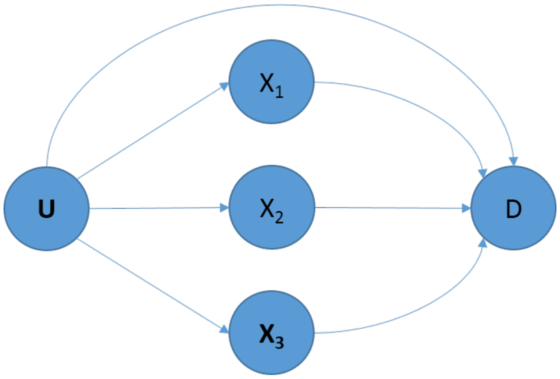


Figure 9.2: Hypothesized scenario A, where the three metabolites X_1, X_2, and X_3 have a common cause U and all three together with U are causes of D

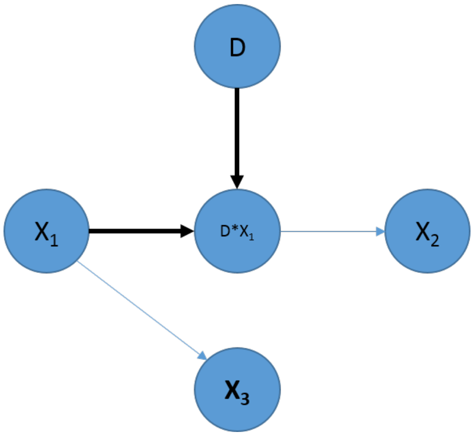


Figure 9.3: Hypothesized scenario B where X_1 and X_3 are always correlated but X_1 and X_2 are only correlated when disease is present. The D*X_1 node represents the product of D and X_1, so when D=0 there is no association between X_1 and X_2 and when D=1 there is an association between X_1 and X_2.

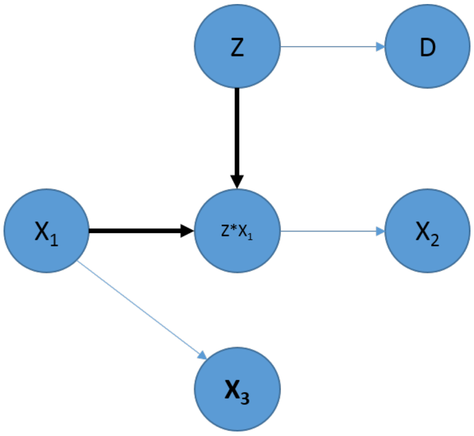


Figure 9.4: Hypothesized scenario C where X_1 and X_3 are always correlated but X_1 and X_2 are correlated only when Z is present. Since Z is a cause of D this implies that X_1 and X_2 are more strongly correlated when D is present - The Z*X_1 node represents the product of Z and X_1, so when Z=0 there is no association between X_1 and X_2 and when Z=1 there is an association between X_1 and X_2

In the data generating model from scenario A the disease status was caused by the simulated exposure variables. In scenario B the data generating model was based on the disease modifying the association between a pair of variables, and in C an unmeasured variable modifies the association between a pair of variables and is also associated with the disease.

Data generating structures When investigating data generating model A, it was found that if a variable modified the effect of another in terms of its association with the outcome, then an edge in a differential network would be induced. Also, if either both of the variables are strong independent causes of the outcome, or only one is and the variables are correlated then again an edge would be induced, however for this to be true the variable would have to be extremely strongly associated with the outcome. So it is fair to say, practically, that if the effect of that variable on the outcome is modified by another variable then an edge should be identified between that pair of variables in the differential network (if the effect is strong enough relative to the sample size).

It was also identified, by analysing data generating model B, that if the disease modifies the relationship between a pair of variables then this can also induce an edge in a differential network. So here the disease status is defined prior to the concentration of X_1 and X_2 (or in our simulations, at least prior to X_2). So if this data generating model is true, a differential network will not provide information about the causes of disease, but could provide information regarding potential markers of disease.

This is contrasted with data generating model A, where the disease status is influenced by the pair of variables (metabolites X_1 and X_2, say), the disease risk is (in part) defined by an individual's concentration of X_1 and X_2. In scenario B, the situation is reversed, an individual's concentration of X_1 and X_2 is (in part) defined by their disease status (In the simulated model disease only defines the concentration of X_2).

In scenario A the development of disease is thought of as occurring later than the measured concentration of the metabolites, so if the data generating model is assumed to be true, then a differential network would give an indication of the aetiology of disease, highlighting possible pairs of metabolites that modify the effect of each other on the disease. This could be thought of as similar to identifying a set of significant interactions in a multivariable logistic regression.

In scenario C, like scenario B, the concentrations of X_1 and X_2 have no influence on the risk of disease. However, unlike scenario B, the disease status is not necessarily defined prior to the measurement of the metabolite concentrations. Therefore if this data generating model were true then a differential model may be able to provide information regarding early indicators of disease, which although not necessarily providing information regarding the causes of disease, could aid early diagnosis or highlight a predisposition towards a disease, so could be of use in a cohort study analysis. These observations however depend on the simplicity of the structures considered. As regards scenario A in particular they depend on the absence of an arrow from U to D. This will be explored further in section 9.2.3

Power, relative to data generating model In scenario A, if the interaction between the two variables was equivalent to an odds ratio of 1.1, it was estimated that an overall sample size of 7000 was required in order to detect an edge using a threshold of $p<0.01$ with 95% power. To detect an edge with this sample size the prevalence of the outcome must be close to 50%. If the prevalence is less than 50% then a larger overall sample size would be required, however if the number in the smaller group (between diseased and not diseased) falls below 2000 then no matter how large the overall sample size there will not be 95% power to detect an edge due to an interaction of this strength.

In scenario B, the overall sample size had a negligible effect compared to the effect of the smaller group size. So to achieve 95% power to detect a difference in the partial correlation between the cases and non-cases of 0.2, there must be at least 500 individuals in each group. A difference of 0.16 can be detected at 80% power at this sample size. If the sample size drops to 100 in the smaller group, a difference of 0.4 is required for 95% power to detect an edge (0.32 for 80% power). A detailed analysis of power was not performed using data generating model C because of its similarity to model B.

Comparison to standard regression models It was of interest to identify how similar the value of δ_{12} in a differential network model (i.e. the quantification of an edge) was to the interaction term (γ_{12}) in a standard logistic regression model, where disease status is the outcome, and a pair of metabolites are exposures, with an interaction between the two metabolites included.

There was a linear relationship between the γ from the logistic regression and the value of δ in the simulated differential networks. In fact in scenario A, when the metabolites X_1 and X_2 were uncorrelated the value of δ and γ are equal in the range -0.2 to 0.2. Figure 9.5 illustrates this relationship.

In scenario A, the logistic regression was more efficient at identifying a significant $\hat{\gamma}_{12}$ between X_1 and X_2 than the differential network method was at identifying a significant $\hat{\delta}_{12}$, whereas in scenarios B and C the differential network method was more efficient. This may be self explanatory, as in scenario A we generated the data using a logistic model i.e. we defined an interaction term γ_{12} which induced a non-zero δ_{12} in the differential network, and in scenarios B and C we defined a difference in the correlations δ_{12} which in turn induced a non-zero γ_{12} in the logistic regression model. Of note with regards to these simulations is that their interpretation applies to both conditional and marginal correlations, because the simulated numbers could be generated either marginally or condi-

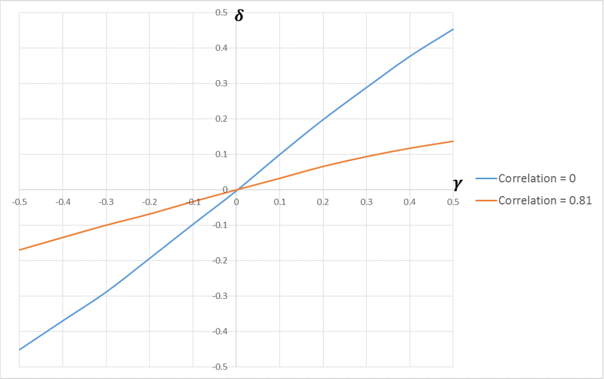


Figure 9.5: Chart showing the estimated value of δ_{12} against the value of γ_{12} for the situation where X_1 and X_2 are uncorrelated (blue) and when the partial correlation of X_1 and X_2 in the overall dataset is 0.81 (red) (Sample size N=10000)

tionally (on other variables).

9.2.2 Marginal vs. partial correlations

In chapter 7 the relative merits of the use of marginal and partial correlations in differential networks was discussed. Using partial correlations provides a “truer” picture of the differences in the case and non-case networks, as the estimated edges are adjusted for all the other nodes in the network to avoid confounding. However, by performing a differential network analysis using marginal correlations could also uncover some features of the data not immediately apparent, so could also be a useful exercise when carrying out a data exploration exercise on some new data. In particular we showed that partial correlations that do not account for all data generating variables (the Us in the model) or not all variables downstream from the Us, may lead to spurious edges. The possibility of adding one metabolite at a time to move from the marginal model to the partial model was also explored (figure 7.22), which can provide an insight into the differences between the marginal and the partial results.

9.2.3 Differential networks using the BWHHS data

Differential network using marginal correlations The differential network based on marginal Pearson correlation coefficients, adjusted for age, is displayed in figure 9.6.

In this differential network, 38 edges were included involving 33 nodes (all other nodes were excluded from the diagram). There were three metabolites identified

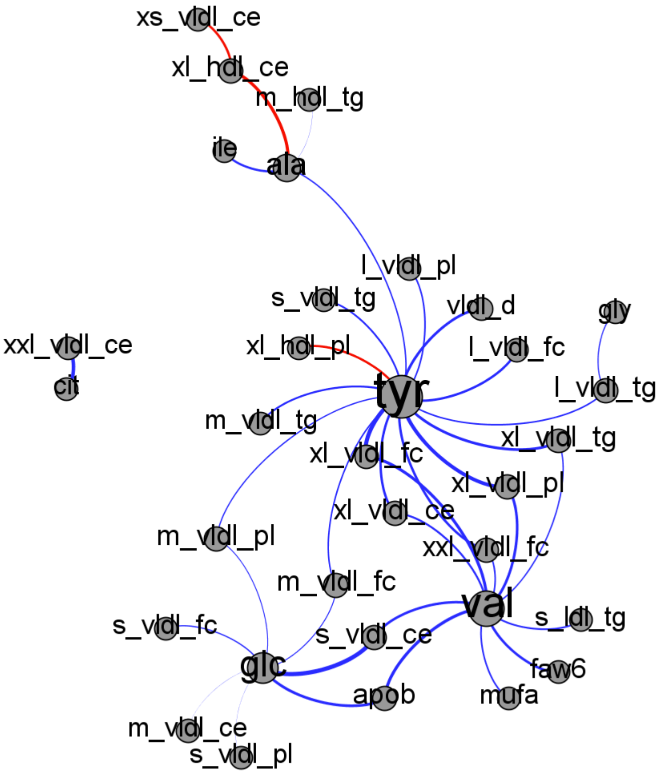


Figure 9.6: Differential network of the BWHHS data formed using marginal correlations - blue = negative estimated δ, red = positive estimated δ

as being potentially important because they were involved in several of these edges (i.e. they were highly central nodes). - Tyrosine, Valine and Glucose, none of which were identified in the univariable analysis as being associated with CHD. The strongest individual edge identified was that between glucose and small VLDL cholesterol esters (The top 5 edges are shown in table 7.1 in chapter 7).

This is a network based on marginal Pearson correlation coefficients, so each edge in the network is estimated only using the data from the pair of metabolites that the edge connects, unadjusted for any of the other metabolites. If we consider the 3 hypothesized data generating models from chapter 6, and we were to believe scenario A was true we may interpret it as these highly central nodes are strongly associated with the outcome (which can be verified by performing a univariable logistic regression with the selected metabolite as an exposure and CHD as the outcome). This was not the case for these 3 metabolites, so a second interpretation could be that the results are indicating that the highly central nodes are modifiers of the association between other metabolites and disease. When the results from the pairwise logistic regressions performed in section 7.3.2.2 were checked, it was observed that there were a large number of significant interactions between these three metabolites and the other metabolites within the dataset. This may lend weight to the hypothesis that these nodes are modifiers of other metabolites association with disease.

Given the assumption in scenario B that the disease is the cause of the difference in correlations observed in a differential network, and that we have excluded all those with the disease at baseline, model C seems a more plausible data generating model than B. In this instance we would consider the unmeasured variable Z from the model to be a pre-disease status or some other cause of CHD. If we were to believe scenario C was true we could interpret the highly central nodes by suggesting that it was this unmeasured cause of disease (or the early stages of the disease itself) that was modifying the association between Tyrosine (for example) and a number of other metabolites.

Many of the nodes that the central nodes are connected to via an edge are they themselves very highly correlated with one another (e.g. the edges from Tyrosine are mostly connected to VLDL metabolite concentrations). The differential network obtained using marginal Pearson correlation coefficients is undadjusted, so if there is a "significant" edge detected between a node (let us call it X) and one node from a highly correlated set, it is likely that there will also be edges from X to each of the nodes in the correlated set. For example there is an edge between Alanine and Isoleucine in the differential network, these are "standalone" amino acids not very strongly correlated with another set of metabolites. But if for example instead of 1 measure of Isoleucine, it was broken down into 9 sub-divisions all very strongly associated with one another, then an edge would probably exist from Alanine to each one of these sub-divisions, giving Alanine a high degree centrality. It is possible that this is contributing to the high degree centrality observed in Tyrosine.

It is also important to bear in mind that the criteria for inclusion in this network is a p-value of less than 0.01, which given there are 3003 tests taking place is a very liberal threshold for inclusion (a Benjamini-Hochberg adjusted threshold would be p <0.000017, and would yield no edges in the network). So the ability of a differential network based on this sample size yielding strong evidence for any particular edge is very limited. In chapter 6 we identified a sample size of 2000 required in the smaller group (in this case the diseased group) to give high power of detecting edges due to a moderate interaction effect if scenario A was true, whereas in the BWHHS there were only 182 cases. However, it should be remembered that this is an exploratory analysis to aid identifying candidate metabolites/relationships for further study, so as a result tyrosine, valine and glucose are the strongest candidate metabolites for further examination and the edges listed in table 7.1 are specific pairs of metabolites whose relationship with each other could be investigated further with respect to their relationship with

CHD. Finally, as these are marginal correlations, the estimates for any edges may be confounded, as discussed in section 7.1.3.1 this can lead to too many edges in the network, whereas using partial correlations can provide more specific results.

There were many more negative edges in the differential network than positive edges, due to the fact that the correlations within both the cases and non-cases tended to be postivie, with a higher correlation typically found in the non-cases. In this thesis no investigation was performed into whether there was any difference in the nature of negative and positive edges.

Differential network using partial correlations The differential network based on partial Pearson correlation coefficients is displayed in figure 9.7.

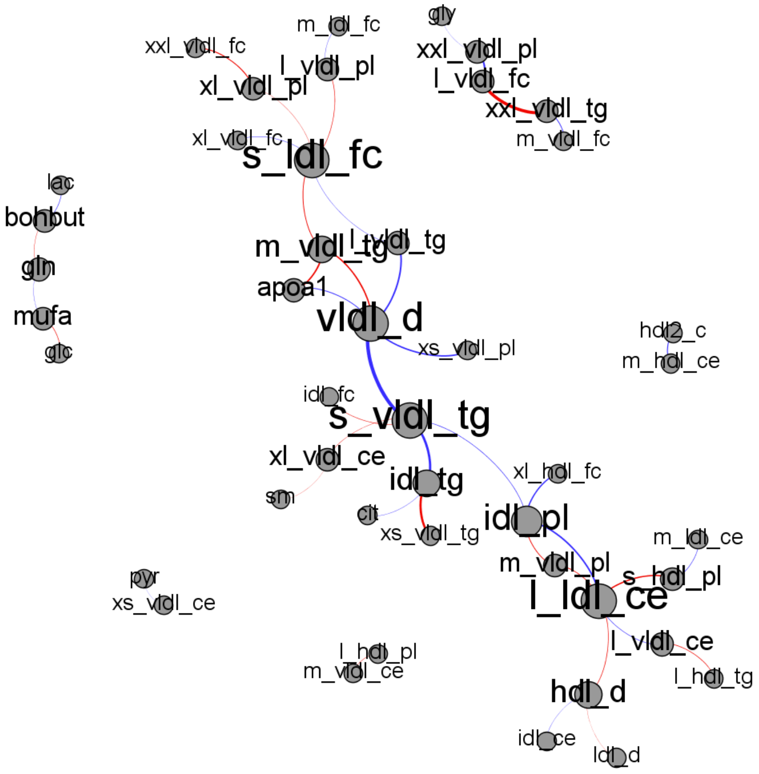


Figure 9.7: Differential network of the BWHHS data formed using partial correlations - blue = negative estimated δ, red = positive estimated δ

The differential network formed using partial correlations has 42 edges involving 45 different nodes. There were 4 nodes which had the joint highest degree (5): Small VLDL triglycerides, large LDL cholesterol esters, small LDL free cholesterol and VLDL diameter. The 5 edges relating to the 5 largest differences in correlation ($\hat{\delta}$) are displayed in table 7.4.

The differential network using partial Pearson coefficients was denser than the differential network using marginal correlations, although there were fewer obvious "hub" nodes, so was a more dispersed network. More of the VLDL metabolites are involved in this network than in the marginal differential network, suggesting that, after adjustment for all other metabolites, these VLDL variables may be modifying the effect on disease of other variables in the network. Or they could be strong, independent risk factors for disease and the estimated δs are echoes of this strong effect (as in scenario A). It is also possible that the association between these pairs of variables is modified by the propensity that an individual has to go on to develop CHD in the subsequent 12 years (as in scenario C). The interpretation of this network is similar to that of the marginal correlation differential network, except for the fact that the edges estimated now are adjusted for all other metabolites.

However, given the highly correlated nature of the BWHHS data it may be that spurious effects are being observed as described in section 7.1.2.1. Because there are a number of extremely highly correlated groups of metabolites, adjusting for all of them can in fact cause a spurious edge to appear in the network between a pair of nodes that are only distantly associated with the association that truly differs between the disease states.

Again when interpreting the findings of this differential network the role of chance is the most plausible explanation for any extreme results observed. There were 3003 potential edges in the network and a p-value of <0.01 was the criteria for inclusion. This allows exploration of the data in the network form that wouldn't be possible with a more conservative threshold, but it does mean that the exploratory analysis performed should be interpreted through a sceptical lens - it may just be that, for a differential network analysis, there is no signal within this data to uncover.

There is no overlap between the findings of the differential network analyses and the lasso regression. None of the interactions identified by a lasso regression were included as edges in the differential network. Nor were any of the highly central nodes in the network highlighted as important main effects in the lasso regression. This may be due to the fact that they are picking up different features of the data, but it also may just be that there is no signal (or the signal is too weak) to pick out, and the edges selected in the network and the interactions selected by the regression are just found by chance.

9.2.3.1 Other considerations

Beyond the limitations of the model itself described above, there are other decisions made throughout this analysis that would have influenced the results a great deal. The metabolite values used in the analysis were transformed from the raw concentrations to more closely comply with the assumptions required of the Pearson correlations performed. The transformations selected for each metabolite were the "best" transformation available from 3 potential options (best being the transformation that resulted in the lowest skewness). By transforming the metabolites using different transformations for each, potentially true linear associations were lost by transforming at least one of a pair of metabolites onto a different scale.

A different, and potentially more effective strategy might have been to choose a single transformation for all metabolites that provided the lowest overall skewness and thus meet the assumption of multivariate normality. This would hopefully make any true associations that exist in the data more likely to be identified in the final analysis. Another solution to this problem could be to use Spearman correlations instead of Pearson correlations as the measure of association, as it does not require the same assumptions of joint normality required for Pearson correlation. This would have the additional benefit of reducing the impact of data where there a high number of zero values.

More fundamentally there are problems with both of the two methods of differential networks proposed (using marginal or partial correlations). When using marginal correlations it is possible the the edges observed are explained via other nodes in the network, so the edge identified may relate to a distant association via other variables rather than a direct association between the two nodes. At the other extreme we use partial correlations to measure the association between a pair of nodes, adjusting for every other node in the network. When we have the situation where there are a high number of nodes in a network and a large proportion are strongly correlated with one another we have little power to identify an edge. Two potential strategies for dealing with this could be:

1. More stringent variable selection criteria
2. Limit the number of variables adjusted for

The first of these two methods is the simpler approach, it is simply to use the partial correlation as our measure of association but to exclude variables that introduce high collinearity into the model. Previously only variables that introduced perfect collinearity were excluded (i.e. when one variable was a sum of two or more others) but we could set a pragmatic threshold to exclude variables where there are a number that are strongly correlated with one another.

The second is to perform an exploration of the data between the two extremes described. Rather than adjusting for none (as in marginal), or all (as in partial),

of the other nodes in the network, we could adjust for a subset of representative nodes. One potential way to select this set of representative nodes could be to first generate a network using all our data.
Then we find the network modules defined by this overall network (as per the method described in section 5.2.6). We could then select one node from within each network module to represent that module, and using this set of nodes as a representative set of nodes from the network we can adjust the associations accordingly.

9.2.4 Results from the time to event analysis

In chapter 8 we attempted to include information on time to event to investigate whether that uncovered edges in the network that were not identified in the network where CHD event in the follow up period was used as the outcome of interest. This was done by, at the end of each year of follow-up, comparing the group of individuals who had suffered a CHD event by that time with those who had survived until that point. Up until the 5th year there were too few events to estimate the differential networks successfully, so this was performed from years 5 to 12.

There was one additional edge identified using this method (when compared to the static differential network (using partial correlations) in chapter 7, between large VLDL cholesterol esters and medium VLDL free cholesterol, which was potentially identified as important in the earlier years but had disappeared by the final year. There were also 2 additional nodes, glutamine and small LDL phospholipids that were identified as important in the 6th and 7th years of follow up, but by the final 12 year follow-up they were no longer found to be as central to the network.

As with the original analysis, there was no overlap between the findings of the differential network analyses and the lasso regression. As this analysis was limited by the low number of observations in the earlier years of study, this led to poorly estimated correlations in the cases group, and as a result few edges detected. So these observed inconsistencies between the two methods are likely to be due to lack of power, even more than in the 12-year analysis (in chapter 7.

9.3 Concluding comments

In this thesis a metabolomic dataset from the BWHHS cohort study was described and the metabolites short-term reliability assessed, this identified the metabolites which were more stable, and therefore were better candidates for use in investigating the associations with CHD over a long follow-up period.

This study was used to motivate our investigation of the emerging method of differential networks. This was addressed by initially simulating individual network edges to gain an understanding of potential interpretations of differential network analysis, then by applying it to the metabolite data from the BWHHS to discover the differential network formed from comparing those who developed CHD in a 12-year follow-up period. Then finally, an extension of the analysis to incorporate time to CHD event was proposed and implemented, with some limitations encountered.

Differential networks may be a useful additional tool in performing exploratory analyses on high-dimensional datasets such as -omics data where causal/aetiological structure is still unknown. This thesis identified potential scenarios where a differential network could identify important interactions or be used to identify biomarkers for future disease risk in cohort studies, although whether it can provide useful information over and above that obtained via other high dimensional methods is uncertain.

Bibliography

[1] London School of Hygiene and Tropical Medicine : Department of Non-communicable Disease Epidemiology. British women's heart and health study, November 2008.

[2] Lawlor DA, Davey Smith G, Patel R, and Ebrahim S. Life-course socioeconomic position, area deprivation, and coronary heart disease: findings from the british women's heart and health study. *Am J Public Health*, 95(1):91–97, 2005.

[3] Watt HC, Carson C, Lawlor DA, Patel R, and Ebrahim S. The influence of life course socio-economic position on health behaviours in older women: findings from the british women's heart and health study. *Am J Public Health*, 99(2):320–327, 2009.

[4] Amuzu A, Carson C, Watt HC, Lawlor DA, and Ebrahim S. Influence of area and individual lifecourse deprivation on health behaviours: findings from the british women's heart and health study. *Eur J Cardiovasc Prev Rehabil*, 16(2):169–173, 2009.

[5] Kim LG, Carson C, Lawlor DA, and Ebrahim S. Geographical variation in cardiovascular incidence: results from the british women's heart and health study. *BMC Public Health*, 10(696), 2010.

[6] Kaptoge S, Di Angelantonio E, Lowe G, Pepys MB, Thompson SG, Collins R, and Danesh J. C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: an individual participant meta-analysis. *Lancet*, 375(9709):132–140, 2010.

[7] Sarwar N, Gao P, Seshasai SR, Gobin R, Kaptoge S, Di Angelantonio E, Ingelsson E, Lawlor DA, Selvin E, Stampfer M, Stehouwer CD, Lewington S, Pennells L, Thompson A, Sattar N, White IR, Ray KK, and Danesh J. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *Lancet*, 375(9733):2215–2222, 2010.

[8] Swerdlow DI and (IL6R MR) consortium. The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis. *Lancet*, 379(9822):1214–1224, 2012.

[9] Lawlor DA, Taylor M, Bedford C, and Ebrahim S. Is housework good for health? levels of physical activity and factors associated with activity in elderly women. results from the british womens heart and health study. *J Epidemiol Community Health*, 56:473–478, 2002.

[10] Choi M, Prieto-Merino D, Dale C, Nüesch E, Bowling A, Ebrahim S, and Casas JP. Effect of changes in moderate or vigorous physical activity on changes in health-related quality of life of elderly british women over seven years. *Qual Life Res.*, 2012.

[11] May M, Lawlor DA, Patel R, Rumley A, Lowe G, and Ebrahim S. High molecular weight adiponectin is not associated with incident coronary heart disease in older women: a nested prospective case control study. *Eur J Cardiovasc Prev Rehabil*, 14(5):839–869, 2007.

[12] Sattar N, Watt P, Cherry L, Ebrahim S, Smith GD, and Lawlor DA. High molecular weight adiponectin is not associated with incident coronary heart disease in older women: a nested prospective case control study. *J Clin Endocrinol Metab*, 93(5):1846–1849, 2008.

[13] Amuzu A. Cohort profile : British women's heart and health study. Unpublished manuscript profiling the BWHHS cohort held at LSHTM.

[14] Vehtari A, Mäkinen V-P, Soininen P, Ingman P, Mäkelä SM, Savolainen MJ, Hannuksela ML, Kaski K, and Ala-Korpela M. A novel bayesian approach to quantify clinical variables and to determine their spectroscopic counterparts in ^{1}H NMR metabonomic data. *BMC Bioinformatics*, 8, 2007.

[15] Idle JR and Gonzalez FJ. Metabolomics. *Cell Metabolism*, 6(5):348–351, 2007.

[16] Waterman CL, Kian-Kai C, and Griffin JL. Metabolomic strategies to study lipotoxicity in cardiovascular disease. *Biochimica et Biophysica Acta*, 1801(3):230–234, 2010.

[17] Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, and Kell DB. Metabolomics by numbers: acquiring and understanding global metabolite data. *TRENDS in Biotechnology*, 22(5):245–252, 2004.

[18] Lewis GD, Asnani A, and Gerszten RE. Application of metabolomics to cardiovascular biomarker and pathway discovery. *Journal of the American College of Cardiology*, 52(2):117–123, 2008.

[19] Serkova NJ, Standiford TJ, and Stringer KA. The emerging field of quantitative blood metabolomics for biomarker discovery in critical illnesses. *American Journal of Respiratory and Critical Care Medicine*, 184(6):647–655, 2011.

[20] Ala-Korpela M, Korhonen A, Keisala J, Hörkkö S, Korpi P, Ingman LP, Jokisaari J, Savolainen MJ, and Kesäniemi YA. ^{1}h nmr—based absolute quantitation of human lipoproteins and their lipid contents directly from plasma. *Journal of Lipid Research*, 35(6):2292–2304, 1994.

[21] Tukiainen T, Kettunen J, Kangas AJ, Lyytikäinen LP, Soininen P, Sarin AP, Tikkanen E, O'Reilly PF, Savolainen MJ, Kaski K, Pouta A, Jula A, Lehtimäki T, Kähönen M, Viikari J, Taskinen MR, Jauhiainen M, Eriksson JG, Raitakari O, Salomaa V, Järvelin MR, Perola M, Palotie A, Ala-Korpela M, and Ripatti S. Detailed metabolic and genetic characterization reveals new associations for 30 known lipid loci. *Human Molecular Genetics*, 21(6):1444–1455, 2012.

[22] Soininen P, Kangas AJ, Würtz P, Tukiainen T, Tynkkynen T, Laatikainen R, Järvelin MR, Kähönen M, Lehtimäki T, Viikari J, Raitakari OT, Savolainen MJ, and Ala-Korpela M. High-throughput serum nmr metabonomics for cost-effective holistic studies on systemic metabolism. *The Analyst*, 134(9):1781–1785, 2009.

[23] Dettmer K, Aronov PA, and Hammock BD. Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, 26(1):51–78, 2007.

[24] Nicholson JK, O'Flynn MP, Sadler PJ, Macleod AF, Juul SM, and Sönksen PH. Proton-nuclear-magnetic-resonance studies of serum, plasma and urine from fasting normal and diabetic subjects. *Biochemical Journal*, 217(2):365–375, 1984.

[25] Michigan State University. Nuclear magnetic resonance spectroscopy.

[26] The Royal Society of Chemistry. Introduction to nuclear magnetic resonance spectroscopy.

[27] Beckonert O, Keun HC, TMD Ebbels, Bundy J, Holmes E, Lindon JC, and Nicholson JK. Metabolic profiling, metabolomic and metabonomic procedures for nmr spectroscopy of urine, plasma, serum and tissue extracts. *Nature Protocols*, 2:2692–2703, 2007.

[28] http://what-when-how.com/acp-medicine/diagnosis-and-treatment-of-dyslipidemia-part 1/. Lipoprotein structure.

[29] Hewett P and Ganser GH. A comparison of several methods for analyzing censored data. *Annals of Occupational Hygiene*, 51(7):611–632, 2007.

[30] Brereton RG. *Chemometrics for Pattern Recognition.* Wiley, 2009.

[31] King G and Zeng Langche. Logistic regression in rare events data. *Political Analysis*, 9:137–163, 2001.

[32] Snijders T and Bosker R. *Multilevel Analysis.* Sage Publications, 1999.

[33] Fleiss JL. *The Design and Analysis of Clinical Experiments*. Wiley Classics Library, 1999.

[34] Rubin DB. Matching to remove bias in observational studies. *Biometrics*, 29:307–317, 1973.

[35] Altman DG and Bland JM. Measurement in medicine: the analysis of method comparison studies. *The Statistician*, 32:307–317, 1983.

[36] Ala-Korpela M. ^{1}h nmr spectroscopy of human blood plasma. *Progree in Nuclear Magnetic Resonance Spectroscopy*, 27:475–554, 1995.

[37] Navarro SL, Brasky TM, Schwarz Y, Song X, Wang CY, Kristal AR, Kratz M, White E, and Lampe JW. Reliability of serum biomarkers of inflammation from repeated measures in healthy individuals. *Cancer Epidemiology, Biomarkers & Prevention*, 21(7):1167–1170, 2012.

[38] Epstein MM, Breen EC, Magpantay L, Detels R, Lepone L, Penugonda S, Bream JH, Jacobson LP, Martnez-Maza O, and Birmann BM. Temporal stability of serum concentrations of cytokines and soluble receptors measured across two years in low-risk hiv-seronegative men. *Cancer Epidemiology, Biomarkers & Prevention*, 22(11):2009–2015, 2013.

[39] Breir M, Wahl S, Prehn C, Fugmann M, Ferrari U, Weise M, Banning F, Seissler J, Grallert H, Adamski J, and Lechner A. Targeted metabolomics identifies reliable and stable metabolites in human serum and plasma samples. *PLoS one*, 9(2), 2014.

[40] Zheng Y, Yu B, Alexander D, Cooper DJ, and Boerwinkle E. Medium-term variability of the human serum metabolome in the atherosclerosis risk in communities. *OMICS: A Journal of Integrative Biology*, 18(6):364–373, 2014.

[41] Floegel A, Drogan D, Wang-Sattler R, Prehn C, Illig T, Adamski J Joost H-G, Boeing H, and Pischon T. Reliability of serum metabolite concentrations over a 4-month period using a targeted metabolomic approach. *PLoS one*, 6(6), 2011.

[42] Townsend MK, Clish CB, Kraft P, Wu C, Souza AL, Deik AA, Tworoger SS, and Wolpin BM. Reproducibility of metabolomic profiles among men and women in 2 large cohort studies. *Clinical Chemistry*, 59(11):1657–1667, 2013.

[43] Chadeau-Hyam M, Campanella G, Jombart T, Bottolo L, Portengen L, Vineis P, Liquet B, and Vermeulen RCH. Deciphering the complex: Methodological overview of statistical models to derive omics-based biomarkers. *Environmental and Molecular Multigenesis*, 54:542–557, 2013.

[44] Witten DM and Tibshirani R. Survival analysis with high-dimensional covariates. *Statistical Methods in Medical Research*, 19:29–51, 2010.

[45] Dudoit A, Shaffer JP, and Boldrick JC. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103, 2003.

[46] Benjamini Y and Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1):289–300, 1995.

[47] Shaffer JP. Multiple hypothesis testing. *Annual Review of Psychology*, 46:561–584, 1995.

[48] Westfall PH and Young SS. *Resampling based multiple testing: Examples and methods for p-value adjustment*. Wiley, 1993.

[49] Meinshausen N, Maathuis MH, and Bühlmann P. Asymptotic optimality of the westfall-young permutation procedure for multiple testing under dependence. *The Annals of Statistics*, 39(6):3369–3391, 2011.

[50] Benjamini Y and Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

[51] Jolliffe IT. *Principal Component Analysis*. Springer-Verlag, 1986.

[52] Bartholomhew DJ, Steele F, Moustaki I, and Galbraith JI. *Analysis of multivariate social science data*. Chapman and Hall, 2008.

[53] Chaterjee S and Hadi AS. *Regression analysis by example (Fourth edition)*. Wiley, 2006.

[54] Hoerl E and Kennard RW. Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12(1):69–82, 1970.

[55] Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58:267–288, 1996.

[56] Hosmer DW and Lemeshow S. *Applied Logistic Regression, Third Edition*. Wiley, 2013.

[57] Kalbfleisch JD and Prentice RL. *The statistical analysis of failure time data, 2nd edition*. Wiley, 2002.

[58] Schoenfeld D. Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239–241, 1982.

[59] Laszlo Barabasi. Network science book.

[60] Lewis TG. *Network Science: Theory and Applications*. Wiley, 2009.

[61] Newman MEJ. The structure and function of complex networks. *Society for Industrial and Applied Mathematics Review*, 45(2):365–375, 2003.

[62] VanderWeele TJ and Robins JM. Signed directed acyclic graphs for causal inference. *Journal of the Royal Statistical Society*, 72:111–127, 2010.

[63] Zachary WW. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977.

[64] Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, and Barabasi A-L. The human disease network. *PNAS : Applied Physical Sciences*, 2007.

[65] Mäkinen V-P, Forsblom C, Thorn LM, Wadén J, Kaski K, Ala-Korpela M, and Groop P-H. Network of vascular diseases, death and biochemical characteristics in a set of 4,197 patients with type 1 diabetes (the finndiane study). *Cardiovascular Diabetology*, 8, 2009.

[66] Keeling MJ and Eames KT. Networks and epidemic models. *Journal of the Royal Society, Interface*, 2(4):295–307, 2005.

[67] Christley RM, Pinchbeck GL, Bowers RG, Clancy D, French NP, Bennett R, and Turner J. Infection in social networks: Using network analysis to identify high-risk individuals. *American Journal of Epidemiology*, 162(10), 2005.

[68] Welch D, Bansal S, and Hunter DR. Statistical inference to advance network models in epidemiology. *Epidemics*, 3(1), 2011.

[69] Zhang Y and Tao C. Network analysis of cancer-focused association network reveals distinct network association patterns. *Cancer Informatics*, 13(3):45–51, 2014.

[70] Talwar P, Silla Y, Grover S, Gupta M, Agarwal R, Kushwaha S, and Kukreti R. Genomic convergence and network analysis approach to identify candidate genes in alzheimer's disease. *BMC Genomics*, 15(199), 2014.

[71] Valcárcel B, Würtz P, Seich al Basatena N-K, Tukiainen T, and Kangas AJ et al. A differential network approach to exploring differences between biological states: An application to prediabetes. *PLoS ONE*, 6(9), 2011.

[72] Valcárcel B, Ebbels TM, Kangas AJ, Soinenen P, Ala-Korpela M, Järvelin MR, and de Iorio M. Genome metabolome integrated network analysis to uncover connections between genetic variants and complex traits: an application to obesity. *Journal of the Royal Society, Interface*, 11(94), 2014.

[73] Albert R and Barabasi A-L. Statistical mechanics of complex networks. *Review of Modern Physics*, 74:47–97, 2002.

[74] Ruohonen K. *Graph Theory.* Tampere University of Technology, 2008.

[75] Freeman LC. Centrality in social networks conceptual clarification. *Social Networks*, 1:215–239, 1978.

[76] Faust K. Centrality in affiliation networks. *Social Networks*, 19:157–191, 1997.

[77] Newman MEJ and Girvan M. Finding and evaluating community structure in networks. *Physical Review E*, 69, 2004.

[78] Fortunato S. Community detection in graphs. *Physics reports*, 486:75–174, 2010.

[79] Blondel V, Guillaume J, Lambiotte R, and Mech E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics*, 2008.

[80] Cox D and Wermuth N. *Multivariate Dependencies.* Chapman and Hall, 1996.

[81] Krumsiek J, Suhre K, Ilig T, Adamski J, and Theis FJ. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *B¡C Systems Biology*, 5(21), 2011.

[82] Schäfer J and Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 2005.

[83] Ledoit O and Wolf M. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10:603–621, 2003.

[84] Finn JD. *A General Model for Multivariate Analysis.* Holt, Rinehart and Winston, 1974.

[85] Bastian M, Heymann S, and Jacomy M. Gephi: an open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media*, 2009.

[86] Jacomy M, Venturini, Heymann S, and Bastian M. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLOSone*, 9(6), 2014.

[87] Fuller T, Ghazalpour A, Aten JE, Drake TA, Lusis AJ, and Horvath S. Weighted gene coexpression network analysis strategies applied to mouse weight. *Mammalian Genome*, 18, 2007.

[88] Gill R, Datta S, and Datta S. A statistical framework for differential network analysis from microarray data. *BMC Bioinformatics*, 11, 2010.

[89] Kujala M, Nevalainen J, März W, Laaksonen, and Datta S. Differential network analysis with multiply imputed lipidomic data. *PLOS One*, 10(3), 2015.

[90] de la Fuente A. From differential expression to differential networking identification of dysfunctional regulatory networks in diseases. *Trends in Genetics*, 26(7), 2010.

[91] Ideker T and Krogan NJ. Differential network biology. *Molecular systems biology*, 8(565), 2012.

[92] Castro C, Krumsiek J, Lehrbach NJ, Murfitt SA, Miska EA, and Griffin JL. A study of caenorhabditis elegans daf-2 mutants by metabolomics and differential correlation networks. *Molecular Biosystems*, 9(7), 2013.

[93] Walley AJ, Jacobson P, Falchi M, Bottolo L, Andersson JC, Petretto E, Bonnefond A, Vaillant E, Lecoeur C, Vatin V, Jerna M, Balding D, Petteni M, Park YS, Aitman T, Richardson S, Sjostrom L, Carlsson LMS, and Froguel P. Differential co-expression analysis of obesity-associated networks in human subcutaneous adipose tissue. *International Journal of Obesity*, 36(1):137–147, 2012.

[94] Chu J, Lazarus R, Carey VJ, and Raby BA. Quantifying differential gene connectivity between disease states for objective identification of disease-relevant genes. *BMC Systems Biology*, 5(89), 2011.

[95] Odibat O and Reddy CK. Ranking differential hubs in gene co-expression networks. *Journal of Bioinformatics and Computational Biology*, 10, 2012.

[96] Bockmayr M, Klauschen F, Györffy B, Denkert C, and Budczies J. New network topology approaches reveal differential correlation patterns in breast cancer. *BMC Systems Biology*, 7, 2013.

[97] Gambardella G, Moretti MN, de Cegli R, Cardone L, Peron A, and di Bernardo D. Differential network analysis for the identification of condition-specific pathway activity and regulation. *Bioinformatics*, 29(14), 2013.

[98] Danaher P, Wang P, and Witten DM. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society*, 76(2):373–397, 2014.

[99] Zhao SD, Cai TT, and Li H. Direct estimation of differential networks. *Biometrika*, 101(2), 2015.

[100] Xia Y, Cai T, and Cai TT. Testing differential networks with applications to the detection of gene-gene interactions. *Biometrika*, 2015.

[101] Drton M and Perlman MD. Model selection for gaussian concentration graphs. *Biometrika*, 91(3):591–602, 2004.

[102] Drton M and Perlman MD. A sinful approach to gaussian graphical model selection. *Journal of Statistical Planning and Inference*, 138(4):1179–1200, 2008.

[103] Wermuth N and Cox DR. Distortion of effects caused by indirect confounding. *Biometrika*, 95(1):17–33, 2008.

[104] De Stavola BL, Daniel RM, Ploubidis GB, and Micali N. Mediation analysis with intermediate confounding: Structural equation modeling viewed through the causal inference lens. *American Journal of Epidemiology*, 181(1):64–80, 2015.

[105] Valcárcel B, Ebbels TMD, and De Iorio M. Variance and covariance heterogeneity analysis for detection of metabolites associated with cadmium exposure. *Statistical Applications in Genetics and Molecular Biology*, 13(2):191–201, 2014.

Appendix A

Results from an alternative version of the univariable analysis performed in chapter 4 are shown in the below table. Previously the linear and quadratic terms were assessed in a joint test, here we perform an analysis just of the linear associations. Some (13) additional weaker linear associations are identified here that were not picked up in the analysis when using the joint test. Mean diameter for VLDL particles is the strongest of these, where a one standard deviation increase in the mean diameter of VLDL is estimated to be associated with a 33% increase in the odds of disease. However, this analysis does not identify Creatinine as being associated with the odds of CHD, despite strong evidence of an association found in the original analysis in chapter 4, picked up there because there may be a non-linear association between Creatinine and the outcome. The table below shows the results for completeness.

Table 1: Metabolites associated with the odds of CHD in the follow up period, statistically significant at the Westfall-Young corrected threshold of p <0.00117, ordered by p-value. Odds ratios relate to a one standard deviation change in the metabolite concentration.

Variable	Odds Ratio (95% CI)	p-value
Total cholesterol in HDL2	0.72 (0.62, 0.83)	<0.0001
Triglycerides in very small VLDL	1.40 (1.20, 1.62)	<0.0002
Serum total triglycerides	1.39 (1.20, 1.61)	<0.0003
Monounsaturated fatty acids; 16:1, 18:1	1.40 (1.20, 1.63)	<0.0004
Total cholesterol in HDL	0.72 (0.61, 0.83)	<0.0005
Triglycerides in small VLDL	1.38 (1.19, 1.60)	<0.0006
Triglycerides in medium VLDL	1.38 (1.19, 1.60)	<0.0007
Triglycerides in IDL	1.36 (1.18, 1.58)	<0.0008
Triglycerides in VLDL	1.39 (1.19, 1.62)	<0.0009
Concentration of medium VLDL particles	1.37 (1.18, 1.59)	<0.0010
Total lipids in medium VLDL	1.36 (1.17, 1.58)	<0.0011
Phospholipids in very large HDL	0.70 (0.59, 0.83)	<0.0012
Free cholesterol in medium VLDL	1.37 (1.18, 1.59)	0.0001
Cholesterol esters in large HDL	0.70 (0.59, 0.83)	0.0001
Total cholesterol in large HDL	0.70 (0.59, 0.83)	0.0001
Triglycerides in large VLDL	1.40 (1.19, 1.65)	0.0001
Total lipids in large HDL	0.71 (0.60, 0.84)	0.0001
Phospholipids in medium VLDL	1.36 (1.17, 1.58)	0.0001
Free cholesterol in large HDL	0.70 (0.59, 0.84)	0.0001
Concentration of large HDL particles	0.72 (0.61, 0.84)	0.0001
Concentration of large VLDL particles	1.40 (1.18, 1.65)	0.0001
Total lipids in large VLDL	1.39 (1.18, 1.64)	0.0001
Mean diameter for HDL particl	0.73 (0.62, 0.85)	0.0001
Phospholipids in large HDL	0.73 (0.62, 0.85)	0.0001
Phospholipids in large VLDL	1.40 (1.18, 1.65)	0.0001
Free cholesterol in large VLDL	1.37 (1.17, 1.61)	0.0001
Glycoprotein acetyls, mainly a1-acid glycoprotein	1.32 (1.15, 1.52)	0.0001
Mean diameter for VLDL particles	1.32 (1.14, 1.52)	0.0002
Total cholesterol in large VLDL	1.36 (1.16, 1.61)	0.0002
Concentration of small VLDL particles	1.33 (1.14, 1.55)	0.0002
Triglycerides in small LDL	1.33 (1.14, 1.54)	0.0002
Free cholesterol in very large HDL	0.73 (0.62, 0.86)	0.0003
Triglycerides in small HDL	1.33 (1.14, 1.56)	0.0003
Estimated degree of unsaturation	0.76 (0.65, 0.88)	0.0003
Bonferroni threshold		
Concentration of very large HDL particles	0.74 (0.62, 0.87)	0.0004
Total lipids in very large HDL	0.74 (0.63, 0.87)	0.0004
Total cholesterol in medium VLDL	1.32 (1.13, 1.54)	0.0004
Total lipids in small VLDL	1.32 (1.13, 1.53)	0.0004
Triglycerides in very large VLDL	1.34 (1.14, 1.57)	0.0004
Cholesterol esters in large VLDL	1.34 (1.14, 1.58)	0.0005
Concentration of very large VLDL particles	1.33 (1.13, 1.56)	0.0005
Cholesterol esters in very large VLDL	1.32 (1.13, 1.54)	0.0005
Total lipids in very large VLDL	1.33 (1.13, 1.56)	0.0005
Triglycerides in LDL	1.29 (1.12, 1.50)	0.0007
Total cholesterol in medium HDL	0.77 (0.66, 0.90)	0.0009
Triglycerides in large LDL	1.29 (1.11, 1.49)	0.0009
Cholesterol esters in medium HDL	0.77 (0.66, 0.90)	0.0009
Total cholesterol in very large VLDL	1.31 (1.11, 1.53)	0.0010
Free cholesterol in chylomicrons and extremely large VLDL	1.30 (1.11, 1.53)	0.0011