Construction and assessment of risk models in medicine

by

Constantinos Kallis

A thesis submitted for the degree of Doctor of Philosophy

Department of Statistics

University of Warwick

Coventry

CV4 7AL

Acknowledgements

I would like to thank my supervisor Dr. Ewart Shaw for his help, guidance and support throughout my work. Moreover, I would also like to help my second supervisor Dr. Gill Grimshaw for giving to me the abdominal aortic aneurysm dataset that has been analysed in the first case study and her useful comments.

I wish to thank Dr. Sailesh Sankaranarayanan for providing the diabetic retinopathy data and his valuable co-operation and comments for the statistical and medical aspects of this study. Additionally, I would like to thank Dr. Vinod Patel for his useful suggestions for the diabetic retinopathy study.

I would like to thank the members of staff and fellow students at the Department of Statistics, University of Warwick for their comments about aspects of this research and their contribution to the improvement of my statistical knowledge and experience.

Special thanks to Dr. Fiona Steele and Dr. Denise Hawkes for useful comments and suggestions.

Also I would like to thank both examiners for several suggestions for improvement and pointing out technical and other issues related to re-presentation of the thesis.

Finally, I wish to thank my parents for supporting me financially and giving me encouragement and valuable advice throughout my life.

Declaration

I declare that this thesis is my own work, and has not been submitted for a degree at another university.

Abstract

This thesis investigates the application of classical and contemporary statistical methods in medical research attempting to bridge the gap between statistics and clinical medicine. The importance of using simple and advanced statistical methods in constructing and assessing risk models in medicine will be demonstrated by empirical studies related to vascular complications: namely abdominal aortic aneurysm and diabetic retinopathy.

First, data preprocessing and preliminary statistical analysis are examined and their application is investigated using data on abdominal aortic aneurysm. We illustrate that when dealing with missing data, the co-operation between statisticians and clinicians is necessary. Also, we show advantages and disadvantages of exploratory analysis.

Second, we describe and compare classification models for AAA selective screening. Two logistic regression models are proposed. We also show that it is important to assess the performance of classifiers by cross-validation and bootstrapping. We also examine models that include other definitions of abnormality, weighted classification and multiple class models.

Third, we consider the application of graphical models. We look at different types of graphical models that can be used for classification and for identifying the underlying data structure. The use of Naive Bayes classifier (NBC) is shown and subsequently we illustrate the Occam's window model selection in a statistical package for Mixed Interactions Modelling (MIM). The EM-algorithm and multiple imputation method are used to deal with inconsistent entries in the dataset. Finally modelling mixture of Normal components is investigated by graphical modelling and compared with an alternative minimisation procedure.

Finally, we examine risk factors of diabetic sight threatening retinopathy (STR). We show the complexity of data preparation and preliminary analysis as well as the importance of using the clinicians' opinion on selecting appropriate variables. Blood pressure measurements have been examined as predictors of STR. The fundamental role of imputation and its influence on the conclusions of the study are demonstrated.

From this study, we conclude that the application of statistics in medicine is an optimisation procedure where both the statistical and the clinical validity need to be taken into account. Also, the combination of simple and advanced methods should be used as it provides additional information. Data, software and time limitations should be considered before and during statistical analysis and appropriate modifications might be implemented to avoid compromising the quality of the study. Finally, medical research should be regarded for both statisticians and clinicians as part of a learning process.

Work reported elsewhere

Some of the work in this thesis has appeared in the following:

C. Kallis (2001) "Selective screening for abdominal aortic aneurysm". Technical Report, Department of Statistics, University of Warwick, UK.

C. Kallis, J. E. H. Shaw and G. Grimshaw (2002) "Exploratory analysis using graphical models". Poster and oral presentation at the RSS 2002 Conference, Plymouth, UK.

S. Sailesh, C. Kallis, S. Philip, M. Sein, S. Aldington, H. Randeva, E. Shaw, E. M. Kohner, E. Hillhouse and V. Patel (2003) "Clinical and hemodynamic risk factors for the presence of sight threatening retinopathy (STR) at presentation to a diabetic retinopathy clinic". Poster presentation at Diabetes UK 2003, Glasgow, UK.

S. Sailesh, C. Kallis, H. Randeva, S. Philip, M. Sein, S. Aldington, E. J. Shaw, E. Kohner, E. Hillhouse and V. Patel (2003) "Clinical Risk Factors predicting the presence of Sight Threatening Retinopathy at presentation to a Diabetic Retinopathy clinic". Poster presentation at 2003 ADA conference, New Orleans, USA.

V. Patel, S. Sailesh, C. Kallis, S. Philip, M. Sein, H. Randeva, E. Shaw, E. Hillhouse and E. Kohner (2003) "Clinical risk factors predicting the presence of Sight Threatening Retinopathy in patients with diabetes at presentation to a Diabetic Retinopathy clinic: A cohort study". Oral presentation at the 18th International Diabetes Federation conference, Paris, France.

and in other papers being written or co-written by C. Kallis.

Contents

1 Introduction **13**

1.1 Applications of statistics in medicine 13

1.2 Structure of this thesis 15

2 Preliminary analysis **16**

2.1 Abstract 16

2.2 Abdominal aortic aneurysm and screening 16

2.2.1 Medical screening 16

2.2.2 Case selection 17

2.2.3 Medical background 18

2.3 CASP data 19

2.4 Data cleaning 20

2.4.1 Missing data 21

2.4.2 Cross-checking 23

2.5 Data coding 25

2.5.1 Extracting data from text 25

2.5.2 Importance of sensible coding 26

2.6 Data description 27

2.7 Univariate analysis 28

2.8 Bivariate analysis 29

2.9 Multivariate exploratory analysis 31

2.10 Results 32

2.10.1 Missing data and cross-checking 32

2.10.2 Data coding 35

2.11 Data description 36

2.12 Univariate analysis 44
2.13 Bivariate analysis 49
2.14 Multivariate exploratory analysis 51
2.15 Conclusions 55

3 Regression analysis **56**

3.1 Abstract 56
3.2 Variable and model selection: risk factors and co-morbidities 56
3.3 Regression models 58
3.4 Classification models 59
3.4.1 Statistical methods 60
3.4.2 Machine learning approach 61
3.4.3 Neural networks 63
3.5 Model selection and validation 66
3.5.1 Regression models 67
3.5.2 Classification models 68
3.6 Statistical methods for selective screening 70
3.6.1 Weighted classification in medical screening 71
3.6.2 Misclassification costs 72
3.6.3 Optimum classification 73
3.7 Further topics 73
3.8 Use of clinical data sets 77
3.8.1 Bayes risk 77
3.9 Flexible selective screening 78
3.10 Statistical analysis and results 79
3.10.1 The area under the ROC curve as criterion 79
3.10.2 Proposed selective screening model 85
3.10.3 Age-adjusted abnormality thresholds 88
3.10.4 High risk abnormality threshold 90
3.10.5 Weighted classification results 91
3.10.6 Estimated Bayes risk 93
3.11 Multiple class models 94
3.11.1 Separating moderate from high risk cases 94
3.11.2 Separating low from moderate risk cases 95

3.11.3 Multi-categorical models 95
3.12 Measurement error and classification 96
3.13 Conclusions 97

4 Graphical modelling **99**
4.1 Abstract 99
4.2 Graphical modelling 99
4.2.1 Definitions 101
4.2.2 Loglinear models 104
4.2.3 Graphical Gaussian models 105
4.2.4 Mixed graphical models 106
4.2.5 Selection procedures in MIM 107
4.2.6 Chain graphs 109
4.2.7 Graphical model classifiers 110
4.3 Results from CASP data 112
4.3.1 Alternative coding 113
4.3.2 Bayesian Network Classifiers 115
4.4 Chain graphs 119
4.5 Occam's window model selection 125
4.6 EM algorithm imputation 130
4.7 Aortic diameter as mixture of components 141
4.8 Aortic diameter growth models for mixture of components 148
4.9 Conclusions 152

5 Case study II: Diabetic retinopathy **153**
5.1 Abstract 153
5.2 Medical background 153
5.2.1 Diabetes mellitus 153
5.2.2 Diabetic eye disease 155
5.3 EKSAGE data 159
5.3.1 Data cleaning and cross-checking 160
5.4 Missing data 162
5.5 Data coding 165
5.6 Data description 167

5.7 Univariate analysis 172
5.8 Bivariate analysis 174
5.9 Multivariate exploratory analysis 179
5.10 Odds ratio for risk factors 180
5.11 Hot-deck imputation 184
5.12 Graphical modelling 188
5.13 Conclusions 190

6 Conclusions and suggestions for further research 192

6.1 Conclusions 192
6.2 Suggestions for further research 194

A Standard statistical methods 197

A.1 Exploratory analysis 197
A.2 Univariate tests 198
A.2.1 Univariate analysis for continuous data 198
A.2.2 Categorical data analysis 200
A.3 Bivariate analysis 202
A.3.1 Graphical methods 202
A.3.2 Correlation 203
A.4 Multivariate exploratory analysis 204
A.4.1 Principal components analysis 204
A.4.2 Factor analysis 206
A.4.3 Multidimensional scaling 206
A.4.4 Cluster analysis 207
A.5 Regression models 208
A.5.1 Linear regression 208
A.5.2 Logistic regression 210

B Definitions for risk factor analysis 212

C Definitions of terms used for screening programmes 214

D EKSAGE project diabetic retinopathy protocol 216

List of Figures

2.1 Abdominal aortic aneurysm 18
2.2 Pattern of missingness for blood pressure variables 34
2.3 Histogram for aortic diameter 36
2.4 Q-Q plot for aortic diameter 37
2.5 Age distribution 38
2.6 Diastolic blood pressure 39
2.7 Systolic blood pressure 39
2.8 Logarithm of systolic blood pressure 40
2.9 Number of cigarettes per day 41
2.10 Years since last smoked 42
2.11 Alcohol consumption 42
2.12 Reciprocal of shifted alcohol consumption 43
2.13 AAA prevalence for each practice 48
2.14 Scatter plot matrix 49
2.15 Age and bps conditional on aac 52
2.16 Age and bps conditional on aac and com 52
2.17 Age and bps conditional on year of scan 53
2.18 Age and bps conditional on year of scan and AAA indicator 54

3.1 Area under the ROC curve for first proposed model 87
3.2 Area under the ROC curve for second proposed model 88
3.3 Proportion of cases above threshold against age 89

4.1 co-morbidities $\not\perp\!\!\!\perp$ bps 113
4.2 co-morbidities $\perp\!\!\!\perp$ bps|bpd 113
4.3 Naive Bayes Classifier 116

4.4 Augmented Naive Bayes Network 118
4.5 Block structure for chain graph 121
4.6 Forward stepwise AIC selection chain graph 122
4.7 Preliminary undirected model for CG-regression 124
4.8 Homogenous saturated model for EM-algorithm 131
4.9 Convergence diagnostics for DA augmentation 137
4.10 Mixture of components for aortic diameter 146
4.11 Misclassification error rate for each aortic diameter threshold 151

5.1 Representation of the normal eye 156
5.2 Standard scheme for diabetic retinopathy grading 157
5.3 Diabetic retinopathy complications 157
5.4 Q-Q plot of age at entry 169
5.5 Q-Q plot of body mass index 170
5.6 Q-Q plot of the logarithm of body mass index 171
5.7 Scatter plot matrix for complete dataset 175
5.8 Scatter plot matrix for diabetes type 1 patients 176
5.9 Scatter plot matrix for diabetes type 2 patients 177
5.10 Scatter plot matrix for complete dataset 178
5.11 Systolic blood pressure and diabetes duration conditional on diabetes eye disease condition 179
5.12 Systolic blood pressure and diabetes duration conditional on diabetes eye disease condition 180
5.13 Systolic blood pressure and diabetes duration conditional on diabetes eye disease condition 181
5.14 Association and influence structure for EKSAGE diabetes type 1 patients 190
5.15 Association and influence structure for EKSAGE diabetes type 2 patients 191

List of Tables

2.1 Pattern of missingness for removal criteria 33

2.3 Smoking level variable (smol) 40

2.4 Risk factor analysis results 46

2.5 Risk factor analysis results 46

2.6 Risk factor analysis results 47

3.2 First proposed logistic model 83

3.4 Second proposed logistic model 85

3.5 Selective screening results 86

3.6 Selective screening results 87

3.7 Age-related abnormality thresholds and number of cases below and above threshold 89

3.8 Proportion needed to be screened to achieve a given proportion of total benefit . . 93

4.2 Recoded smoking level variable (smol2) 114

4.4 Cardiac and vascular impairment indicators 114

4.6 Variables used in Naive Bayes Classifier 115

4.8 MIM output for Occam's window selection procedure 127

4.10 Results for Occam's window 129

4.12 smol2 distribution before and after EM imputation 132

4.14 Area under ROC curve for each imputed dataset 136

4.16 MI inference estimates 139

4.18 Aortic diameter of two components in MIM 143

4.20 Aortic diameter with three components in MIM 143

4.22 Aortic diameter with two components using heterogenous model 144

4.24 Aortic diameter with three components using heterogenous model 145

4.26 Aortic diameter with four components using heterogenous model 145

4.28 Three unequal variances normal components for aortic diameter 147

5.1 Pattern of missingness for 516 patients 164
5.3 Number and proportion of patients for each type of diabetes 168
5.5 Mean and standard deviation for each DM type 170
5.7 Odds ratio for blood pressure measurements 183
5.8 Matching criteria for average hot-deck imputation 185
5.10 Odds ratio for blood pressure measurements 186

B.1 General 2-way contingency table 212

C.1 2-way contingency table for selective screening 214

Chapter 1

Introduction

1.1 Applications of statistics in medicine

One of the important targets of statistical applications in medicine is to establish the relationship between the existence of an event of medical importance and the factors that are believed to be related to this event. Quantification of the relationships between factors in the data is one of the key aspects of any bio-statistical project. Furthermore, to decide about accepting or rejecting claims about links of specific phenomena to diseases and their associated symptoms, we need to examine the evidence that is available to us in the appropriate way.

Differences between individuals in the study in terms of their characteristics and their current medical state can often be explained to a certain degree by factors that are well established in the literature. For example, age is one of the attributes for an individual in a medical study that is usually included in several types of statistical analysis. The ageing process is believed to be a cause for many medical complications throughout our lives.

On the other hand, there are factors whose role in causing a number of diseases is a matter of controversy. A good example for this is smoking, which is claimed to be responsible for a number of diseases including lung cancer, myocardial infarction and other related medical complications. The link between smoking level and the degree of its damaging effect to the health of individuals is still a matter of debate, even though the majority of publications suggest that smoking is responsible for a large number of deaths that could have been avoided.

The motivation for this study is the application of statistical methods into the investigation of the association between medical factors in a study and the examination of the influence that the characteristics of an individual have on the presence of a specific clinical event. Nowadays,

the availability of computing resources and user-friendly statistical packages, has made statistical analysis in many ways a task that is easily performed by non-statistical medical analysts who subsequently present their results to the general public.

The dangers of misusing and abusing statistical analysis methods and their corresponding outcomes is widely evident in our time, where information about all sorts of events is easily collected and recorded to databases. In Altman (1999, pages 1-2), examples are given of results where medical research that are published in the media can often lead to conclusions that are not supported by the evidence provided.

In addition to that, Venables and Ripley (1997, page 1) state that one "of the effects of the information-technology era has been to make it much easier to collect extensive datasets with minimal human intervention. Fortunately the same technological advances allow the users of statistics to much more powerful 'calculators' to manipulate and display data".

There is a sense among some researchers outside the statistical community of including statistical methods whenever these methods are easily implemented and their outputs can be understood by an audience with little time to pay attention to the details. Especially in the medical area, it is sometimes more important to produce evidence as fast as possible rather than investigating the data thoroughly. Also, the pressure on data analysts to produce results that follow mainstream articles' requirements leaves little space for including statistical methods that have recently been made available because of the associated risk of being rejected as the statistical method used is "too difficult to understand".

On the other hand, there are a few members of the statistical community who might prefer to apply a statistical method that has been proven to give less biased results but is not readily available to the majority of statisticians, rather than compromising the quality of the results for the sake of speedy implementation. Clearly, only in an ideal world would we know which method is the best for the task we have to perform before we start analysing the data, and the best method is not always available to the data analyst.

Moreover, time limitations and lack of resources can result in the use of an inferior statistical method, even though another type of analysis would have been better had it been possible to be implemented. Essentially, the construction and assessment of risk models in medicine is an optimisation procedure, where the constraints do not always depend on the statistician or the data.

1.2 Structure of this thesis

In this thesis, we investigate the application of statistical methods that can be used in medical projects concerned with risk assessment. We demonstrate the implementation of simple and advanced methods and the associated difficulties and potential solutions when dealing with messy and complex data. Two case studies are presented to show the practical aspects of using contemporary statistical packages for investigating specific medical subjects.

Chapter 2 looks at data preparation as well as preliminary analysis of data from an abdominal aortic (AAA) aneurysm screening project. Missing data is shown to be a task where the statistical and clinical validity need to be considered. Exploratory analysis for identifying risk factors of AAA is also given.

In chapter 3, we compare classification models for AAA selective screening and we identify two logistic regression models as candidates for implementation. The use of cross-validation and bootstrapping for classification is also shown. Additionally, alternative definitions of AAA abnormality that have been proposed in literature are used to construct selective screening models.

In chapter 4 we demonstrate the application of graphical models. This type of models are considered for classification. The Naive Bayes classifier (NBC) as selective screening model for AAA is examined. Occam's window model selection is explored as an alternative of using a single model for classification. After that, the use of EM-algorithm and multiple imputation method for inconsistent data are demonstrated. We investigate whether the aortic diameter might be a mixture of normal distributions and use the findings to look into aortic diameter growth models and their relevance to screening programs for AAA.

Chapter 5 looks into finding risk factors related to sight threatening retinopathy (STR) for diabetic patients. We find that a number of variables are not statistically significant as predictors of STR but have been included as control variables because of the their clinical importance. Further exploratory analysis show the complex association structure of the variables in the data. The presence of STR is shown to be predicted by a number of blood pressure measurements and their significance is shown to depend on the way we deal with missing data in the sample.

In chapter 6, the conclusions are given and possible ways of improving and extending the methods used in this thesis are proposed.

The statistical analyses have been carried out mainly in S-Plus and R. The statistical package MIM, which has been obtained by the Department of Statistics, University of Warwick has been used for the implementation of graphical modelling techniques. Also, Stata, Microsoft Word and Excel have been used to record and manipulate the datasets included in this study.

Chapter 2

Preliminary analysis

2.1 Abstract

Data preparation and exploratory analysis are investigated and their application is examined using an empirical study on abdominal aortic aneurysm. We show the importance of taking into account both clinical and statistical considerations when dealing with missing data. Also, we show the usefulness as well as the limitations of exploratory analysis for identifying predictors of the presence of AAA. We find that co-morbidities and family history indicator, smoking level, age and diastolic blood pressure are significant predictors of abnormal aortic diameter.

2.2 Abdominal aortic aneurysm and screening

2.2.1 Medical screening

One of the important approaches to modern preventive medicine is *population screening.* This is defined in the NSC (National Screening Committee) first report (Health Departments of the United Kingdom 1998) as: "The systematic application of a test or inquiry, to identify individuals at sufficient risk of a specific disorder to warrant further investigation or direct preventive action".

In the NSC second report (Health Departments of the United Kingdom 2000), "a new definition is proposed to take into account the importance of informed choice and risk reduction and this is set out below".

"[Population screening is] A public health service in which members of a defined population, who do not necessarily perceive they are at risk of, or are already affected by a disease or its complications, are asked a question or offered a test, to identify those individuals who are more

likely to be helped than harmed by further tests or treatment to reduce the risk of a disease or its complications".

A subsequent report from the UK National Screening Committee asks "Healthcare staff involved in screening services to make a fundamental shift in their approach to their work. Instead of over-selling services, as they might have sometimes done in the past, staff should explain the limitations and risks associated with screening, as well as its benefits" (Kmietowicz 2000). In addition to that, "the public should be given a realistic view of the merits and flaws of screening so that they can make an informed choice and decline an invitation if they wish".

From the description above, it is obvious that medical screening may be extremely useful for preventing a disease from occurring or for early detection of a disease therefore maximising treatment effectiveness, but should be used with caution and only in those cases that the benefits outweigh the harm of using a screening test or detecting the disease before treatment is necessary.

In some cases, screening is the only possible option to prevent a life threatening event. An example is abdominal aortic aneurysm which is generally symptomless (Grimshaw et al. 1997). Furthermore, the individuals should know the benefits and the risks related to selecting and rejecting screening in order to take their decisions; in other words, the National Health system should be implementing elective screening schemes.

2.2.2 Case selection

One way of making a screening programme more efficient in terms of the resources needed is to select the part of the population that is most at risk and offer that population a discriminatory test. However, if the selection tool used does not have high specificity and relatively high sensitivity, then it will lead to the inclusion of a large number of cases that do not require the screening test and many cases will be at risk because they have been excluded from the test. In other words, both false negative and false positive are harmful in the population screened.

A possible approach to satisfying the requirements mentioned above is to include in the model predictors for which the "risk of the disease in the high-risk group relative to that in the low-risk group (relative risk)" is high (Hakama et al. 1979). These authors also attempt to improve the power of case selection by combining the risk indicators with the advantage of achieving better results in terms of discrimination.

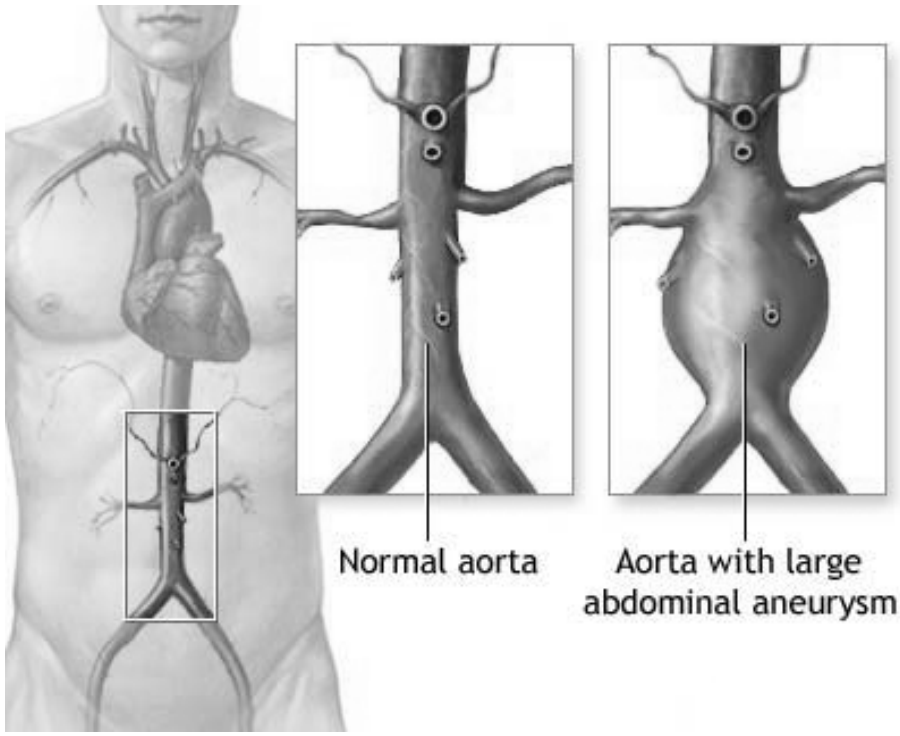
To be more specific, "given the size of the high-risk group (p% of the total population) and the ratio of the risks in the high- to low-risk group (relative risk, RR), the yield (the percentage of

cases detected at screening of all cases, y)" can be estimated from

$$y = \frac{\mathrm{RR}}{\mathrm{RR} + (100 - p)/p} \times 100\%$$

2.2.3 Medical background

An aortic aneurysm means dilation of the aorta and screening is arguably the most effective way of finding cases of abdominal aortic aneurysm (AAA) as few symptoms are evident before life threatening rupture of the aorta occurs.



Abdominal aortic aneurysm is detected by measuring aortic diameter by ultrasonographic scan.

Figure 2.1: Abdominal aortic aneurysm

The definition of presence of aortic aneurysm is related to an arbitrary threshold defining an abnormal aortic diameter. The most commonly used transverse dimension is 3 cm (Vardulaki et al. 1998, Grimshaw et al. 1994). A matter of controversy between researchers is the use of age related thresholds for aortic aneurysm screening (Grimshaw et al. 1992, Grimshaw et al. 1997). The prevalence varies between 1.3% and 12.7% depending on the age group screened and the criteria used for the definition of AAA (Wilmink et al. 1998 and Semmens et al. 2000). There is a trend for increasing prevalence both because of better diagnosis and increased incidence (Wilmink et al. 1998).

The mortality rate lies between 80% and 90% within a small period of time following rupture (Harris 1992, Scott et al. 1999 and Bergqvist et al. 1999). Repair of the aorta with a replacement arterial graft is the most common intervention for aortic diameter larger than 5.5 cm or greater in order to prevent rupture. It is obvious that the ability to find which cases have high risk of rupture could save a significant number of lives and reduce the cost of surgery and subsequent treatment, which is substantially higher for cases with ruptured aorta.

It is also important when considering elective surgery to estimate risks related to operation or any other intervention and compare them with the risk of rupture (Wilmink et al. 1998, Scott et al. 1999). An aspect in this case that has not been investigated thoroughly is the patient's and doctor's preferences and their effect on deciding about further action for large aortic aneurysm.

Because of the high mortality rate for ruptured aortas and the effectiveness of surgical intervention when aneurysm is diagnosed early, it has been suggested that patients should be periodically screened in order to identify people at risk of rupture. For those with aneurysms not yet ready for surgery, screening intervals currently vary from 3 months to 2 years and depend in an ad hoc way on the initial diameter and the estimated growth rate of the aorta.

2.3 CASP data

The data analysed for the abdominal aortic aneurysm case study were collected for the Community Aortic Screening Project (CASP) (Grimshaw 1993) between 1989 and 1993. It is stated that "the aim of the study was to explore the establishment of a mass screening programme in the Greater Birmingham area. The four objectives were:

- to study the feasibility of screening for abdominal aortic aneurysm in an urban environment
- to develop a scanning protocol
- an examination of definitions of abnormality and the prevalence of abnormal aortas
- to study the prevalence of associated medical conditions".

Grimshaw (1993) also says that "practices participating in the first phase of scanning were those practices with representatives on an ad hoc Community screening research committee and additional practices suggested by members of this group. These practices and clinicians from Queen Elizabeth Hospital have co-operated on many clinical projects. The practices are self selected only by the interest of a partner in innovative medicine. Current General Practice issues, the

demography of the practices, socioeconomic indices of the practice populations very widely. No attempt was made to differentiate patients on social or ethnic class".

The selection of the patients was based on findings from previous studies. The peak incidence of rupture is in the male population between 70 and 74 years old. The incidence below 65 years is known to be small; other screening programmes have started screening at age 65. The peak incidence of rupture begins to decline after 74 years old and the risks of surgery rise substantially beyond this age threshold because of related medical conditions.

Patients were invited to attend the local surgery for an ultrasound scan of the abdominal aorta by a single letter of invitation addressed to them by their own GP. All male patients between ages 65 and 75 were invited with the exception, in five practices, of the terminally ill. The practices with larger residential homes sent invitations to the patients only when it was clear that the patient would accept, or be suitable for, repair of the aorta, if this was necessary.

The number of letters of invitation sent in this first phase of the study was 3500. Further funding allowed the study to continue beyond the period of the first study and many additional patients were seen in the subsequent two years. A further 9500 invitations to screen were issued to a previously unscreened population (Grimshaw 1993).

The potential problems with using this group is that the results of the analysis might not applicable to any other age groups in the general population or for a screening programme for women of the same age group. A general problem with using observational data is the influence of unmeasured confounders that affect the selection into the screening programme. The sample we have in our study is by definition biased by patient self-selection and by their GPs wanting to buy the service.

The first step in constructing a screening model is preparing the data for analysis and subsequently exploring the data to identify important aspects of the data. Statistical methods for the objectives described above are presented in detail in the next sections.

2.4 Data cleaning

One of the most important and often disregarded steps of any statistical project is to ensure that the quality of the data is at a satisfactory level before proceeding to further analysis. The quality control might include the detection of missing values, logically or virtually impossible values, possibly arising from data coding errors, and in general apparently anomalous features (Cox and Wermuth 1996). All these aspects can be described by the terms *data checking* or *data cleaning* (Altman 1999).

Mistakes when data have been recorded or entered into the database can be a source of unexpected and misleading results. When discovered, they can be corrected and usually the person who is responsible for collecting the data can be helpful for this situation.

If information is available about the possible sources of mistakes or about the reasons for having low quality data in general, then there are other ways of dealing with this as we are going to see in the following paragraphs. It is crucial that before starting to analyse the data, we should be certain that the results produced are not influenced by any sources of unwanted bias. The phrase "garbage in garbage out" describes quite well the situation of producing wrong results from wrong data input (Altman 1999).

2.4.1 Missing data

It is not unusual in large studies why there are many reasons that some values have not been recorded. Especially in the medical area, this problem is present when attempting to analyse data retrospectively and it is not possible to collect further data after the end of the study. In some cases, even when the data are still collected (prospective studies), it might not be possible to have complete datasets because of limited resources, prohibitive cost or unavailability of the subject in the study.

A specific code is devised for indicating missing values and this is ideally done when entering the data into the database that is going to be used subsequently for statistical analysis. The person who is collecting the data or the database manager should make clear that the indicator of missing data would not be a source of confusion.

By that, we mean that the particular code should be a number that does not belong to the possible values for a specific variable. For example, if we are referring to age, then clearly a negative value or a very large number (e.g. 999) should be sufficient for indicating missing values. Another possible situation is to have different codes for missing values in order to indicate different reasons or different sources of data unavailability.

The use of blank spaces in statistical packages is often seen as convenient when entering large amounts of numbers in the database or when for a number of variables, the majority of the entries is actually missing. This might be a cause of problems for some statistical packages that do not recognise blank entries as valid; this can be dealt with through additional data coding as we would see later.

In addition to that, we should always be aware of the fact that different statistical packages handle missing data in different ways and sometimes it is not clear how the statistical analysis

has been done with regards to the presence of missing values. In most cases, missing values are omitted when constructing a model but this is not always desirable. There are situations where a large number of variables are included in the model and every subject with at least one missing value is excluded.

This might have as a result ignoring a great proportion of the sample and the results being severely biased. The only situation where this is not a problem is when a small number of subjects are removed from the study and they can be regarded as a "simple random subsample of all the data values" (missing completely at random) (Schafer 1997).

This situation is quite rare in practice; what is quite often assumed is that missing values are *missing at random (MAR)*. By this, we mean that "the probability that an observation is missing depends on the observed data but not on the missing data" (Schafer 1997, page 10). For example, if the value of the weight is missing for a number of individuals and missingness depends on their height, then the weight can be said to be missing at random.

On the other hand, it is possible, using the example above, that the weight is missing for the individuals that are overweight or obese. Then, the missing values can not be said to be missing at random and the missingness is non-ignorable. It is very difficult to be certain about the exact nature of the missing data unless we have additional information from other sources, e.g. the person who collected the data.

A possible way to deal with missing data when it is assumed to be missing at random is to impute the missing values. In Schafer (1997, page 1), *imputation* is described as "a generic term for filling in missing data with plausible values". There are many ways to do that; this includes simple methods such as replacing the missing entries by the mean or the median and more complex methods such as finding matching cases and get a random draw from the observed ones to replace the missing values (hot deck imputation) (Twisk et al. 2002).

Another way to deal with missing data is to use regression or other models to predict the missing values using the observed values and predictors that seem to be related to the variable for which we want to impute some entries. In addition to that, we can implement *multiple imputation* in order to derive several possible values for the missing entries. This is done to "reflect uncertainty about the true values of the missing data" (Schafer 1997, page 5).

Multiple imputation involves generating m complete datasets. We then analyse separately each of the imputed versions of the data as we would have done had we had complete datasets. Finally, we combine the results from each imputed dataset using the appropriate formulae to "produce a single inferential statement (e.g. a confidence interval or a p-value) that includes uncertainty due

to missing data" (Schafer 1997, page 5).

Moreover, there are situations that the missing values can be included in the study by including a missing category. This could be done when we are dealing with categorical values. It is not a good practice to replace missing values by the "majority class rule" (put the value that most of the subjects have for the particular categorical variable) as this might distort the results.

In all situations, before we analyse the data, we should clarify the ways we dealt with the missing data to the reader and if it is possible, to impute the data in different plausible ways and then compare the results. Sensitivity analysis is important in case of having non-ignorable missingness; in this situation, we impute the data using a number of models based on the information we have about the missingness mechanism.

2.4.2 Cross-checking

As mentioned above, there are situations that pairs of values should be investigated together in order to assess whether they should be regarded as outliers. We should extend our checks to all variables to see whether their values can be said to be reasonable depending on the values of other variables (Altman 1999, page 132).

The example of calculating body mass index (BMI) using weight and height of an individual is a good example of a variable where cross-checking is necessary. It is possible to have values of BMI that do not seem to be reasonable even though both height and weight lie into the range of reasonable values. Of course the main problem is the definition of reasonable values for height, weight as well as BMI.

In situations like the one mentioned above, there is a variety of possible solutions to deal with this problem. First, we can get information from the person that designed or collected the data that we are going to analyse. In this case, well defined range for each of the variables are often available hence it is easy for the data analyst to decide whether the values recorded or derived are reasonable or acceptable.

Alternatively, the literature related to the study might be a possible source of information for deciding about the acceptance or rejection of particular values. Nevertheless, it is again possible to include values that do not belong to the pre-defined range of acceptability and we may decide to investigate their influence on the results by including them in study.

Furthermore, consider the case of investigating the effect of previous pregnancies on kidney transplant. This information is "only relevant for women, and so for men should be set to missing or to a different code indicating not applicable" (Altman 1999, page 124).

In addition to that, it is possible to create, similarly to the indicators for outlying values, indicators for unacceptable values. In this way, we can include them in a regression model to assess the level of influence that they might have on the results.

Another way of overcoming the problem of unreasonable values detected by cross-checking the data is to derive new variable(s) from the original ones and include the derived variables in further analysis instead of the original. It is advisable to co-operate with the data collector or study designer in order to create replacements that will contain as much information as the original variables and at the same time avoiding the problem of illogical values.

For example, if pairs of systolic and diastolic blood pressure values do not seem to be reasonable, a possible way to overcome this problem is to create a categorical variable with a level indicating "doubtful combinations of measurements of blood pressure". Of course, that would mean that we have loss of information because of converting two continuous variables to a categorical; this will be further analysed in the section related to data coding.

Variables related to time events can also be a source of finding unacceptable combinations of values in the dataset. For example, if we need to calculate the age of the subjects in the study, it is possible to find negative values where there are possible mistakes or missing values in one or more variables used to compute age. In this case, great care should be taken of correcting the mistakes using additional information possibly extracted by other parts of the dataset.

Moreover, particular statistical programs and other software used for managing the data can be a source of errors. To be more specific, if the year for a time event is indicated by the last two digits (i.e. 90), some programs interpret that as 1890 or 1990 depending on the initial configuration of the software. Obviously, this type of error can be found by checking the range of the values; in other cases these are identified only after deriving a value with unreasonable value.

Additionally, the format of time events in one program does not mean that is acceptable in another program or even when it is recognised as valid, the value is actually the correct one. It is advisable in this case to try to include in the database time events using a simple format that is usually acceptable in the statistical packages. Separating the day, the month and the year of the event is possibly useful if for example we need to calculate the age using only the year of the event.

Also, it is preferable to use four digits instead of two for the year of the event to avoid possible misinterpretation from any software. Furthermore, we might need to cross-check that after transferring the data from one database to another, all the values have been copied successfully. There are situations that a number of entries are correct and because of some possible unrecognised format (even something like a blank space), the rest of the values are entered with errors or not at

all. Double checking that the difference between subsequent events has the correct sign is a first step of identifying possible problems.

Finally, for variables derived by a complicated procedure involving a number a functions, it is advisable to check the results of each step if possible. This might protect from unwanted mistakes possibly due to errors in the functions applied. Cross-checking in this case is much more difficult as reasonable but wrong values might be derived and with large datasets it is almost impossible to identify them by simple inspection. A good practice could be to run a few examples (e.g. using a random subsample that contains 1% of the data) to investigate quickly possible unexpected or illogical results.

The methods described above are by no means panacea for all possible problems that somebody might face and of course the proposed solutions are not always the optimal ones. But it is better to be prepared to spend sometimes a substantial amount of time examining the data and trying to identify all possible sources of problems before analysing the data. The amount of time is well spent if you bear in mind that the alternative might be running the analysis several times because of identifying these problems, possibly a few at a time.

2.5 Data coding

Equally important to data checking is the process of putting the dataset into a format that will be used for statistical analysis. Especially when the data recording has been done using paper notes, it is also important to transfer the information in the database by converting it by using coding.

It must be remembered that the results could be easily distorted by misinterpretation of the data values by the statistical software. For example, suppose that an entry in the database is a phrase that contains two words separated by a blank space. Some statistical packages will regard that as one entry but some others regard the blank space as "end of entry" character and will enter the second word as the value for the next variable.

Hence data coding is related to preparation of the data and it should be considered carefully before analysing the data to avoid unnecessary work and lost of valuable time.

2.5.1 Extracting data from text

In medical studies, it is often necessary to extract data from paper notes, sometimes in the form of comments. It is important to keep in the database these notes in text format in case we need to retrieve any part of this information in the future. It is common practice to search through these

comments by using text searching tools for particular keywords that provide sometimes crucial evidence about the medical condition of an individual.

Discussion with subject experts will provide the important keywords or phrases in the comments and the possible links between the text and other types of data available. For example, comments about related diseases might provide useful ways of building a medical profile for a patient and a possible way of identifying individuals that we need to treat.

Background information and existing literature related to our study are also useful in extracting and coding information from comments. In general, "one important role of a model is to provide a link with underlying substantive knowledge" and also "it will be often desirable to provide a direct link with previous published work in the field" (Cox and Wermuth 1996, page 18). Hence, data coding should take into account not only the objectives of the study but also ways of comparing the results of the study with corresponding results from previous and possibly well established and widely accepted publications.

Finally, we should remind the reader that entering data formatted as text is not always a trivial task. It is possible that additional work might be required to put this text into a form that will be acceptable by the database. In addition to that, there are statistical packages that do not provide text searching facilities for extracting useful data. Thus, the data analyst would probably need to transfer this information into another program and then insert the useful parts back into the software that will be used for statistical analysis.

2.5.2 Importance of sensible coding

There are situations that coding needs to be revised even though it is not clear how this should be done. For example, having a small group of subjects for some levels of a categorical data might prove an obstacle when trying to obtain parameter estimates related to those levels.

For instance, it is not unusual to have a study in which a small number of patients having a rare disease or belonging to a rare ethnic group are present. Particular models require a minimum number of subjects in order to compute specific parameters of the model, otherwise they do not return any results.

In order to avoid inestimability of parameters (Schafer 1997, page 206), we might merge a number of levels with small number of cases into one level that will represent a different group in the results of the analysis. Sensible coding is required as it is important to have estimates of the parameters but also models that will be useful and applicable.

It is tempting to recode in order to merge all small groups in the data into bigger ones to avoid

problems with inestimability all together. Consider though a study that we need to find useful variables to predict a rare disease; the small number of individuals having this disease is the most important part of the data.

Thus, recoding must take into account several factors that will affect the results of the statistical analysis. Information from subject experts as well as previous publications could be a useful aid to avoid possible pitfalls of data recoding.

2.6 Data description

Assuming that the quality of data is at an acceptable level and we have dealt with all possible types of data entry errors, we need to take a first look at the data using simple statistical tools that we have at our disposal. This is an important step for the purpose "of understanding the general characteristics of any dataset" (Everitt et al. 2001, page 37). For this reason, summary statistics and graphs of the variables are possible ways of obtaining information about key features of the data.

A frequencies table, which indicates the number of subjects we have for each of the possible values, is probably the first tool that we apply to summarise the variables in the dataset. Furthermore, the majority of the statistical packages include some option of obtaining the mean and the median, as well as the minimum and the maximum values when dealing with continuous variables. For categorical variables, a frequencies table should be usually sufficient to get a first picture of this type of data.

In addition to that, we can obtain more information about the distribution of the variables by graphical methods that are standard in any statistical software. For continuous variables, a histogram can be used; see appendix for further details.

Other possible graphical methods for continuous variables that are usually available are stem-and-leaf plots and box plots; see appendix for details. It is also important to identify outlying values that might be caused by typing or other types of error. These graphs could show a group of values of this kind not previously found during data checking stage. Hence, data description analysis might be repeated a number of times for a number of variables to ensure that quality of the data is appropriate for more sophisticated analysis.

Depending on the assumptions required for implementing specific statistical tests and constructing models, we might need to check whether these assumptions are fulfilled by the data we have. Using one of the types of graphs mentioned above, it is possible to assess whether a continuous variable follows a distribution of interest.

The Normal distribution is the most common of these distribution as a number of hypothesis tests and model procedures are applied on Normality assumptions. In this way, it is sometimes evident that we need to transform the data by applying the appropriate formula to the values of a variable.

It is also possible to test formally the distribution of a variable and to examine how successful a transformation is by applying parametric or non-parametric tests. Nevertheless, visual inspection of the data is sometimes a faster way of identifying departures from assumptions based on which the statistical modelling is going to be investigated.

In order to assess how close the distribution of a specific variable is to the distribution of interest is, quantile-quantile plots (Q-Q plots) are used. As normal distribution is for various reasons the most common distribution for comparison, the corresponding Q-Q plot is also known as normal probability plot or normal plot (Altman 1999, page 133) and it is found in a large proportion of publications related to data analysis; see appendix for further details.

Moreover, it is possible to use non-parametric estimation for the density function of a variable if we do not want to specify a parametric form for it. For example, by applying kernel estimators for estimating the density functions, we can plot the result in order to investigate visually the distribution of interest. By changing the bandwidth of the kernel function, we can obtain a "smooth" estimate (Everitt et al. 2001, page 45).

2.7 Univariate analysis

Graphical methods such as histograms and Q-Q plots have the main advantage of being simple methods of describing the data but they also have drawbacks. The interpretation of a graph depends on subjective criteria and the experience of the person who draws conclusions from the graph.

For example, it is difficult to know exactly what distribution a variable in our dataset follows just by looking at a histogram. Changing the number of groups of variable values or the breakpoints between these groups might have a dramatic effect on the shape of the histogram leading us to different.

Using the Q-Q plot involves assessing the closeness of the distributions that we would like to compare by the proximity of the points to a straight line. Of course the shape is also important: a plot with "U" shape is an indication of skewness and an "S" shape an indication that one distribution has heavier tails than the other. Again, the problem is that subjective judgement comes in and different analysts might have conflicting opinions concerning the conclusions derived by looking at

a graph.

Moreover, even if the distribution of the variable we examine seems close enough to the distribution of interest, we would like to measure how close it is. Another question is, using significance test, are they significantly different or not? Hence, based on the initial inspection of graphs related to the data we analyse, we would want to test whether are conclusions are valid and measure the uncertainty of these conclusions being right. For details related to univariate tests, see the corresponding part of the appendix.

2.8 Bivariate analysis

Equally important to the univariate analysis, where the variables in the data are examined separately is the analysis of these variables in pairs. In Altman (1999, page 277), three main purposes of such analyses are identified: possible association between two variables, predictability of one variable from the other and the level of agreement between the variables in the pair we examine.

There are several methods we could implement to investigate the bivariate structure of the dataset. Initially we examine pairwise relationships between variables by using graphical methods such as the scatter plot or the contour plot. An advantage of plotting bivariate data for the purpose of finding links between the members of the pair of variables we have is that it is often a simple way to find these links, similarly to drawing conclusions from any type of graph.

To put it in other words, "it is often convenient to present data pictorially. Information can be conveyed much more quickly by a diagram than by a table of numbers. It can also help a reader get the salient points of a table of numbers" (Bland 1996).

In addition to that, another method that is widely used in practice to assess the bivariate relationships in the dataset is estimation of the correlation between two variables. Furthermore the pattern of possible changes of a variable over the range of another variable could be examined. In statistical terms, this is equivalent to estimating the conditional distribution of one variable conditional on the values of a second variable. The appendix contains additional information for standard bivariate statistical methods.

Pairwise associations between variables using correlation coefficients can be found by chance when having a large number of variables and consequently a large number of significance tests. An additional aspect of correlation computation, when having more than two variables in the data is the influence of other variables on the association between two variables. Hence, adjustment might be considered and this can be done by calculating the partial correlation coefficient.

Sometimes, the influence of a third variable on the correlation of two other variables in the

dataset can be spotted by using the correlation matrix, where are all pairwise correlation coefficients are presented or by inspecting a scatterplot of the correlated variables. In other cases, multiple regression or other multivariate methods might be applied before discovering complex associations between variables. Additionally, external information such as the opinion of subject experts could be included by testing the degree of influence of other variables on correlation coefficient estimations.

Finally, we should interpret the results of correlation analysis with precaution in order to avoid being mislead by these results. For example, if Pearson's correlation is estimated to be small, this does not mean the variables are not associated but that the relationship between them is not linear; this might be due to "a peculiar (non-linear) relationship such as a cyclic pattern that exists between the average midday temperature and calendar month" (Altman 1999, page 296).

Furthermore, the importance of an association between variables in a study is sometimes unrelated to the statistical significance of the correlation coefficient. This is related to the fact that in the medical area, the results from correlation analysis can be judged as clinically irrelevant (Altman 1999, page 297). The square of the correlation coefficient is proposed as more realistic measure of association than the actual coefficient as it represents the proportion of variability of the data "explained" by the association between two variables.

A mistake in the interpretation of correlation between variables is to assume significant association means also causal relationship without additional evidence. It might be possible to explain the observed association by other means than inferring that one variable is causing the other variable. For example, there might be a group of variables that has direct influence on the variables that are highly correlated; thus, the association between the correlated variables depends on this group of variables.

In Hill and Hill (1991), it is mentioned that significant association should assumed as causation when a number of conditions are met. These are the strength of association, consistency of the association among different samples, specificity of the association, temporal relationship between cause and effect, the biological gradient or dose-response curve, biological plausibility, the coherence of the evidence, the experimental evidence and reasoning by analogy.

Hence, correlation analysis should be regarded as "an exploratory method for investigating interrelationships between among many variables" (Altman 1999, page 298). It is also very important, to avoid multiple hypothesis tests when having a large number of variables in order to avoid finding significant correlations which are found by chance and are not actual associations.

Thus, we need to apply multivariate methods that take into account the relationships between

more than two variables at a time. Nevertheless, pairwise associations investigations, such as correlation coefficient calculations are valuable and the results should be included into further statistical analysis.

2.9 Multivariate exploratory analysis

So far, we have presented methods for exploring the data either by investigating one variable at a time or by assessing pairwise relationships between the variables. The results from these steps are important to understand the nature of the data and also the importance of some variables in relation to the purpose of the study. From univariate and bivariate analysis, we could decide which variables should be considered for further analysis if we need, for example, to have only a small number of them in order to keep our models simple.

It is also crucial to understand the limitations of these methods and that they might be misleading if there are important effects from other variables in the data. Hence, we need to explore the multivariate relationships in the data and attempt to examine possible interactions between variables and also possible ways of summarising the data into a small number of features that explain variability between the observations.

Principal components analysis, factor analysis and multidimensional scaling are methods applied in order to reduce the number of variables needed by constructing combinations that describe well the differences between the observations. Additionally, cluster analysis is another method that can applied "to investigate multivariate data to determine whether the data consist of relatively distinct groups of similar individuals" (Everitt et al 2001, page 401). Details for multivariate methods usually found in statistical packages are given in an appendix.

From all the multivariate analysis methods mentioned above and described in the corresponding appendix, it is obvious that they are designed for continuous variables (e.g. principle components or factor analysis) or they depend on subjective criteria and they are more related to finding similarities between observations (e.g. cluster analysis). When dealing with categorical variables only or a mixture of categorical and continuous variables, we need to apply methods that take into account the exact nature of this data and use this information for assessing relationships between variables in the dataset.

In addition to that, we might need to include external information derived from the literature or subject experts. In some cases, some of the methods described previously can be applied to reduce the number of variables into components that are capable of explaining the variability between the observations.

Hence, we do not reject well established methods such as principle components analysis, factor analysis or clustering methods. We propose that these methods should be handled with precaution and be regarded as an exploratory tool and as means to discover data structure that otherwise might be difficult to reveal.

2.10 Results

The dataset made available for our statistical analysis included 3375 people that participated in the Birmingham Community Aneurysm Screening Project (CASP) under the same screening protocol. Data on past history for specific co-morbidities were included (myocardial infarction, peripheral vascular disease, chronic obstructive airway disease and other diseases of the circulatory system) within the dataset.

2.10.1 Missing data and cross-checking

From the initial sample, 374 cases with missing or irregular values were removed from further analysis; thus the size of the sample used is 3001. To be specific, these 374 cases have been removed according to at least one of the following criteria:

- missing date of birth (`dob`): 52 cases.
- missing date of scan (`dscan`): 28 cases.
- missing `age` (5 cases).
- the difference between age recorded and the age deduced by using date of birth and date of scan is more than a year (`bsa`): 67 cases.
- missing (1 case) or female gender (9 cases) (variable `gend`).
- missing diastolic blood pressure (`bpd`): 197 cases.
- missing systolic blood pressure (`bps`): 198 cases.
- apparent errors in recording for number of cigarettes per day (`smod`): 2 cases.
- missing practice (`prac`): 2 cases.
- recorded to belong in practices with only one case: 3 cases.

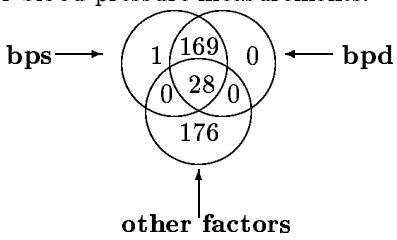
	prac	dob	dscan	amax	age	bsa	gend	bps	bpd	smod
3001	1	1	1	1	1	1	1	1	1	1
3	0	1	1	1	1	1	1	1	1	1
15	1	1	0	1	1	1	1	1	1	1
91	1	1	1	0	1	1	1	1	1	1
2	1	1	0	0	1	1	1	1	1	1
14	1	1	1	1	1	0	1	1	1	1
44	1	0	1	1	1	0	1	1	1	1
1	1	0	1	1	0	0	1	1	1	1
3	1	1	1	1	1	1	0	1	1	1
1	1	0	1	1	1	0	0	1	1	1
1	1	1	1	1	1	1	1	0	1	1
169	1	1	1	1	1	1	1	0	0	1
4	1	1	0	1	1	1	1	0	0	1
6	1	1	1	0	1	1	1	0	0	1
4	1	1	0	0	1	1	1	0	0	1
1	1	0	0	0	1	1	1	0	0	1
1	1	1	0	0	0	1	1	0	0	1
1	1	1	1	1	1	0	1	0	0	1
3	1	0	1	1	1	0	1	0	0	1
1	1	1	1	1	0	0	1	0	0	1
1	1	0	1	1	0	0	1	0	0	1
4	1	1	1	1	1	1	0	0	0	1
1	0	1	1	1	1	1	0	0	0	1
1	0	0	0	0	0	0	0	0	0	1
2	1	1	1	1	1	1	1	1	1	0

1: observed, 0: missing or irregular value

Table 2.1: Pattern of missingness for removal criteria

The pattern of missingness for removal criteria is shown in table 2.1 on page 33. In this case, 0 is the value denoting missing or irregular entry in the dataset.

As blood pressure measurements have approximately 6% of their entries missing, it is important to investigate further their pattern of missingness. The Venn diagram in figure 2.2 shows the distribution of missing values for blood pressure measurements:



At least one missing entry for all other factors

Figure 2.2: Pattern of missingness for blood pressure variables

In the remaining 3001 cases, there were 34 individuals with the doubtful recording of age. For these people, age was deduced by using the date of birth and date of scan. Additionally, 556 entries that appeared in the data file as blank entries have been replaced by zero. The replacement of the values has been applied with the agreement of the original data-collector Dr. Gill Grimshaw.

The above-mentioned 374 cases were removed to reduce bias in the subsequent analysis. When values are not entered in the database or the recorded values are treated as missing, it is assumed that they are missing at random. For example, unavailability of equipment for taking measurements is a possible cause of missingness.

Thus, on the one hand, missing blood pressure measurements can be attributed to missing equipment. On the other hand, blank entries for alcohol consumption can be replaced by zero as this value was regarded by the person entering the data as “common" and not worthy of writing or typing each time it occurs.

Finally, it should be remembered that there are cases where the missing at random assumption is known not to hold. If for example, blood pressure was not recorded for those with extremely low or high measurements, then this part of the data is missing not at random (MNAR).

The main problem in our dataset is missingness of age and blood pressure entries, both of which could be assumed MAR. To our knowledge, no missing values in this study can be regarded as MNAR. Thus, we could remove 374 cases assuming that this removal does not lead to biased results.

2.10.2 Data coding

For smoking history, two variables were recorded: the number of cigarettes per day and years since last smoked. A categorical variable for smoking level (never, moderate smoking, heavy smoking, inaccurate information and poor or inconsistent information) was derived from the recorded variables for smoking mentioned above.

Initially, as can be seen above, we have separated individuals with ambiguous information about smoking level to two groups, namely "inaccurate" and "poor". The reason for the separation mentioned above is our intention to investigate whether there are substantial differences between the two groups. It must be pointed out that the data related to smoking should be analysed with caution as self-reporting was used on the data collection method.

To be specific, we have the following groups for smoking (the appropriate interpretation of blank entries in the data file was agreed with the original data-collector, Dr. Gill Grimshaw):

- both number of cigarettes per day and years since last smoked equal to zero (383 individuals): inaccurate information.
- smoked between one and ten cigarettes per day (820): moderate smokers.
- smoked more than 10 cigarettes per day (1480): heavy smokers (Smith et al. 1993 for further details).
- Blank entries in the data file for both covariates (263): never smoked.
- all other combinations (55): poor information.

In addition to that, an indicator for individuals with at least one of the co-morbidities or family history of abdominal aortic aneurysm has been included in further analysis. This can be justified by the fact that these co-morbidities are all considered as manifestations of degradation of the circulatory system and can be seen as symptoms of the same root cause.

Initially, AAA was deemed present if the maximum aortic diameter was 29 mm or more. At the end of the study, it was proposed that the standard threshold could be replaced by age-related thresholds for abnormality (Grimshaw et al. 1997). In addition to that, several sources in the literature mention that for aortic diameters in excess of 40 mm "a real rupture risk has been demonstrated" (Collins et al. 1988). Hence, the cut-off point of 40 mm could be used as the threshold of identifying and selecting high risk cases.

2.11 Data description

The variables contained in our analysis can be separated into two groups: categorical and continuous. It is possible that a variable that has been recorded on continuous scale, but is then converted to categorical, for several reasons. For example, aortic diameter was measured in millimetres; thus it is a continuous variable. As one of the objectives of this study is to predict whether an individual has aortic diameter above a certain level, we need to convert this to a categorical variable.

In addition to that, a number of categorical variables in this study have been combined to create a new categorical variable. When using exploratory statistical analysis, it is important to explore the data on both the original and the transformed form. In this way, it might be possible to change our coding or identify better ways of dealing with the original data.

As mentioned above, the key variable in the dataset we analyse is aortic diameter. The distribution of variable amax is show in figure 2.3 on page 36.

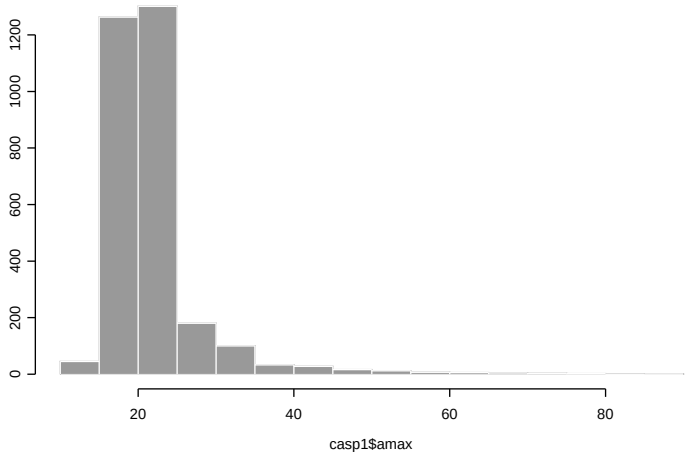


Figure 2.3: Histogram for aortic diameter

The majority of cases (2752 individuals, 91.7% of the sample) have diameters below 29 mm which is the threshold for aortic aneurysm. In order to assess whether amax follows Normal distribution, we use a Q-Q plot shown in figure 2.4 on page 37.

It is clear from the Q-Q plot that aortic diameter (`amax`) does not follow a Normal distribution. When applying smooth transformations such as logarithmic or square root, it is not possible to achieve normalisation of this variable (plots not shown).

From the Q-Q plot for aortic diameter, it might be concluded that the majority of the cases

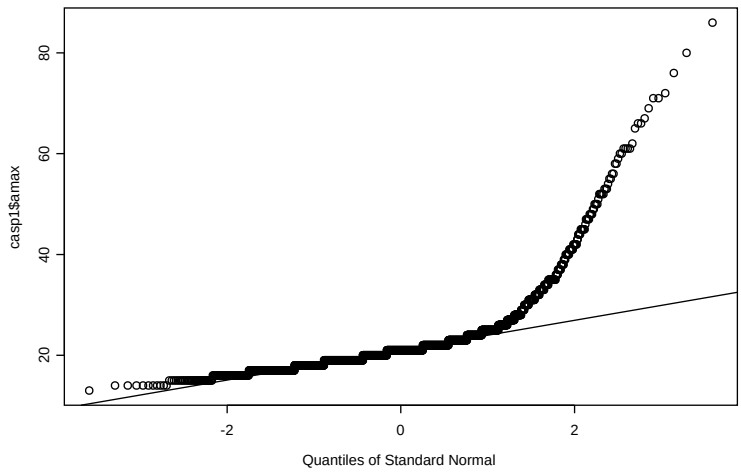


Figure 2.4: Q-Q plot for aortic diameter

follow one distribution and a small number another distribution. This is due to the fact that the distribution follows the straight line and there is point where it becomes a separate line segment.

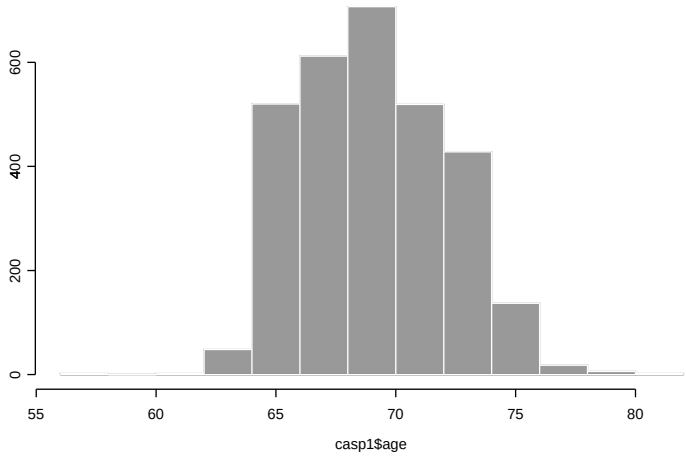
One of the key questions of this study is whether aortic diameter can be described as a simple mixture of normal distributions. This can be investigated to some extent by simple visual inspection of the corresponding Q-Q plot. We will attempt to answer this question at a later stage after exploring further the data and identifying the most significant variables associated with aortic diameter.

There are 249 people in this study with aortic diameter at least 29 mm wide. These are labelled as "abnormal" diameter group and will be the main focus of this study. Subsequently, we will attempt to study the individuals in this study by applying the threshold of 40 mm for the aortic diameter.

In this case, we try to examine separately the high risk of aortic rupture group (79 cases). Finally, we analyse the data by using a factor with three levels for risk of rupture: low, moderate and high. Classification is achieved by using both thresholds mentioned above (29 and 40 mm).

Next we investigate the age distribution of this sample. From the corresponding histogram (figure 2.5, page 38), age is roughly normally distributed with mean approximately 70 years and standard deviation of 3 years.

As we attempt to construct a case selection tool that can be easily implemented in clinical environments, we attempt to convert age (continuous scale) to meaningful categorical variables. Using the thresholds for defining clinically important age groups in Vardulaki et al. (2000), we define



Small number of individuals below 65 and above 75 years due to screening protocol.

Figure 2.5: Age distribution

a binary variable for indicating people in the sample with age at least 70 years. Subsequently, we construct another factor using as thresholds 65, 70 and 75 years (4 levels).

After that, we investigate the distributions of diastolic and systolic blood pressure (`bpd` and `bps` respectively). This has been done by plotting separately the histogram of these variables; see figure 2.6 on page 39 for diastolic and figure 2.7 on page 39 for systolic blood pressure.

It can be said that both variables are approximately Normal, even though systolic blood pressure might need transformation to remove a small amount of skewness. The histogram of the logarithm of systolic blood pressure (figure 2.8 on page 40) does not provide clear evidence about the necessity of transformation. Another possible way of blood pressure measurements normalisation may be achieved by applying Box-Cox transformations (Edwards 2000).

An additional difficulty is the interpretation of results on the transformed scale once the transformed variable is included in a risk model. In the same way as with age, we also investigate the usefulness of binary indicators for diastolic blood pressure at least 90 mmHg and systolic at least 160 mmHg. Both thresholds are clinical thresholds for hypertension (Vardulaki et. al 2000).

In addition to that, it is possible to combine diastolic and systolic blood pressures by a function of these variables. Mean arterial pressure and pulse pressure, which are respectively a weighted average ($\frac{2}{3}$`bpd`$+\frac{1}{3}$`bps`) and the difference (`bps-bpd`) are common in the medical literature. Furthermore, it might be useful to define the form of the combination that optimises a specific objective function.

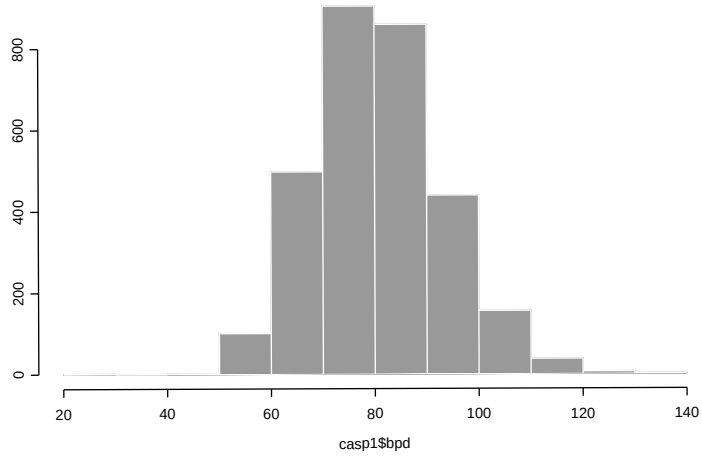


Figure 2.6: Diastolic blood pressure

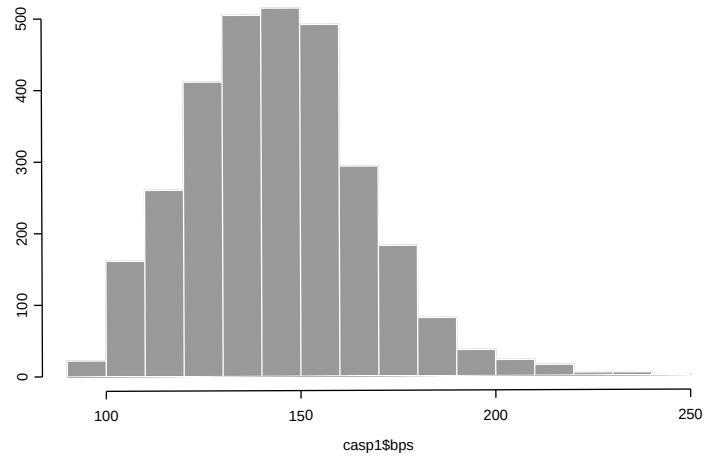


Figure 2.7: Systolic blood pressure

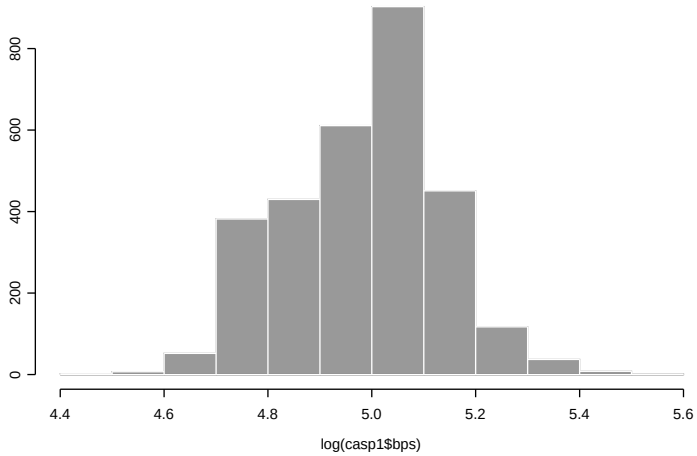


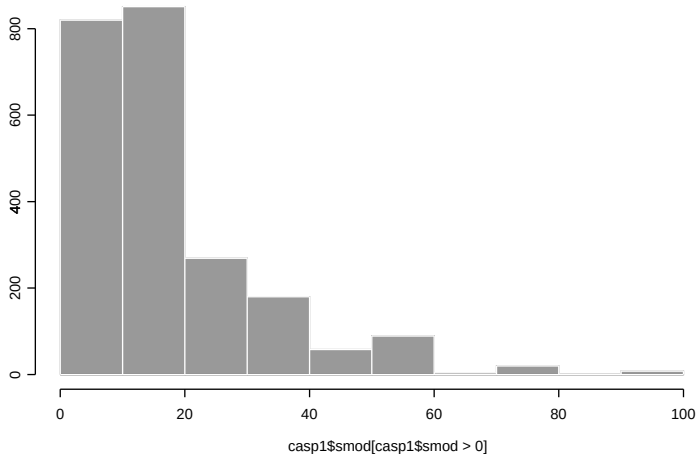
Figure 2.8: Logarithm of systolic blood pressure

For example, if we want to minimise misclassification error of a statistical model, it is possible to implement optimisation procedures that explore several possible combinations of diastolic and systolic blood pressure. The drawback of this could be that the result is not clinically meaningful and it is only applicable for a specific optimisation target.

Past smoking history has been recorded using two variables. These are number of cigarettes per day (`smod`) and years since last smoked (`lasm`). As mentioned previously, both variables have been combined to a categorical variable with five levels (`smol`). These indicate level of smoking (no, moderate or heavy) and inaccurate or poor (inconsistent) information. Useful information can be derived from the frequencies table for smoking level (table 2.3, page 40) and the histograms for smod (figure 2.9, page 41) and lasm (figure 2.10, page 42) variables .

0	1	2	3	4
263	820	1480	383	55

Table 2.3: Smoking level variable (smol)



312 missing values.

Figure 2.9: Number of cigarettes per day

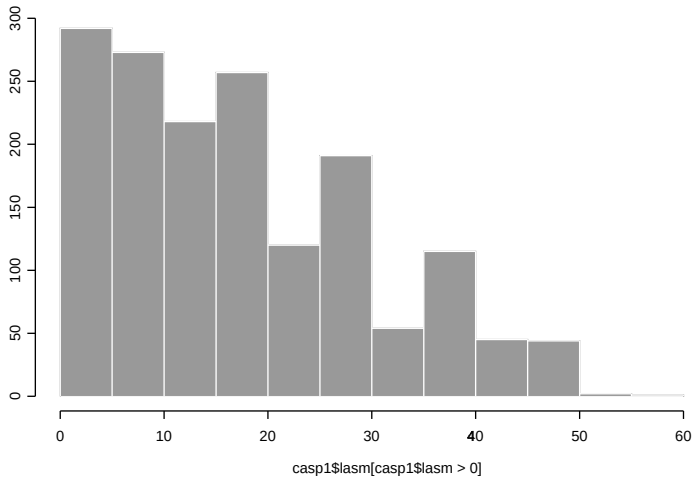
Further investigation is necessary for the reasons behind poor and inaccurate information for smoking. As this is based on self-reporting evidence, it is difficult to have a good explanation for each of the inconsistent entries. It might be possible though to replace the inconsistent data based on the other parts of the data or reasonable assumptions for missingness.

Another variable included in the analysis is an indicator for diabetes. Specifically, for 172 individuals, it is known that they are diabetic patients without any additional information about this disease (e.g. diabetes type 1 or 2, medication). It has been suggested that diabetes might play an important role in the pathogenesis of aortic aneurysm (Blanchard et al. 2000); hence we need to investigate its significance in our study.

Moreover, information has been recorded for alcohol consumption per week. The Q-Q plot for alcohol consumption (figure 2.11, page 42) indicates clearly that this continuous variable does not follow a Normal distribution. Using standard transformations (logarithmic or square root) had no effect on normalising alcohol consumption, mainly because of the large number of individuals with no alcohol consumption (results not shown).

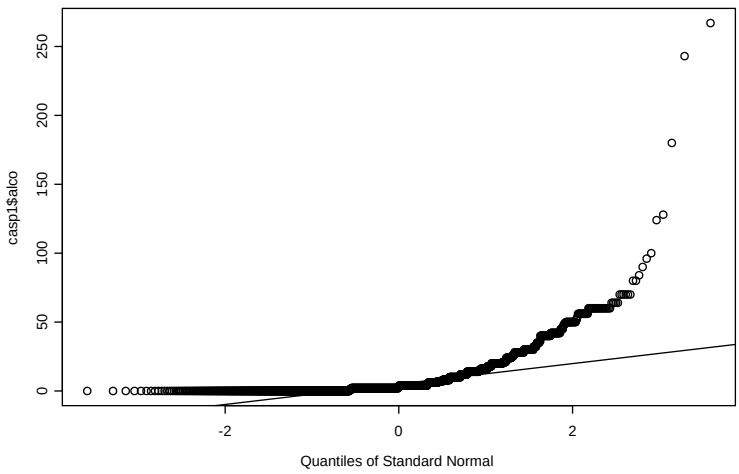
Another possible transformation is the reciprocal of shifted alcohol consumption which would represent time for consuming one unit of alcohol. From the Q-Q plot of $1/(1+\text{alco})$, we can see that this transformation does not achieve normalisation of alcohol consumption variable (figure 2.12, page 43).

Based on the Q-Q plots of alcohol consumption on the original and on transformed scales, it



267 missing entries.

Figure 2.10: Years since last smoked



Measured in units of alcohol per week

Figure 2.11: Alcohol consumption

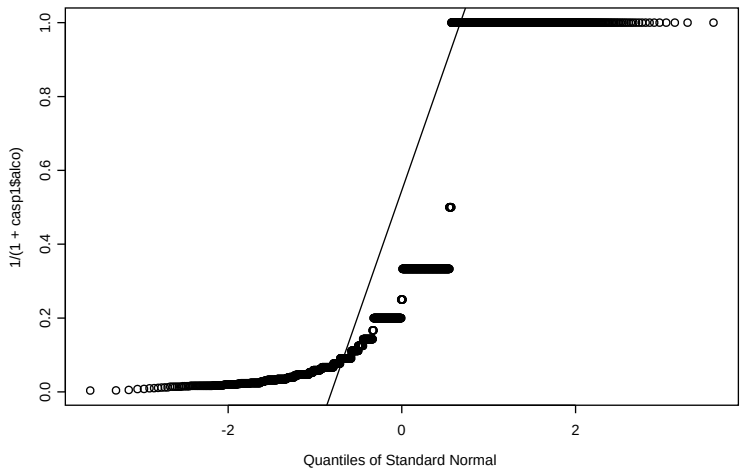


Figure 2.12: Reciprocal of shifted alcohol consumption

might be advisable to create a categorical variable. One possible way is to use the threshold of 20 units per week that is currently regarded by clinicians as the upper limit of moderate drinking. Alternatively, we could have a factor with 3 levels: 0 units (non-drinkers), 0-19 units (moderate drinkers) and at least 20 units (heavy drinkers).

Finally, the presence of several co-morbidities have been recorded as comments for each individual. These are the following:

- `mi`: myocardial infarction
- `cva`: cerebrovascular accident (stroke)
- `cvd`: cardiovascular disease
- `pvd`: peripheral vascular disease
- `coad`: chronic obstructive airway disease
- `cad`: coronary artery disease

These co-morbidities, family history and hypertension treatment can be regarded as complications of the circulatory system. Hence a binary variable can be used to indicate the presence of at least one of these medical conditions. Subsequently, it might be valuable to analyse the data by including these variables separately or combine them in a different way that is still clinically sensible.

Additional information about the individuals in the study might be derived from the practice they have been scanned for abdominal aortic aneurysm. As mentioned previously, two cases with missing practice and three cases representing the only people screened in their practice have been removed from further analysis. In addition to that, the "practice" variable would be useful in a predictive model if a screening program takes place in the same practices in another period of time.

Furthermore, date of ultrasound scan for measuring the diameter of the aorta might be included in the study. For 28 cases, date of scan was missing and they have been removed from further statistical analysis. The inclusion of date of scan related information for further analysis is not straightforward. If we want to apply a prediction formula for another sample, then day and month can be included. On the other hand, to investigate the possibility of temporal bias, year of scan could be equally important.

There are some temporal patterns that might be explained by additional information related to UK culture. For example, there no cases scanned during month 8 (August) and only 9 cases in September (month 9), probably due to summer holidays. Also, the small number of individuals tested in December can be attributed to Christmas' holidays. On the other hand, the reduction in scanned cases on day 3 and 27 could be explained by the fact that these particular days are close the beginning or end of each month.

2.12 Univariate analysis

The results of our statistical analysis in the following section are based on the use of a 2-way contingency table for comparisons between those with and those without the disease and also between those with and those without the presence of a risk factor. See in appendix C the definitions of important terms related to the results of the analysis mentioned below.

It is very important for constructing a case selection model to identify the variables that are more likely to optimally separate normal from abnormal cases. In a similar way to the analysis of risk factors undertaken by Smith et al. (1993) and using the same dataset, we investigated several predictors to find those that might be included in our model. It must be remembered that some variables may not be easily obtained, i.e. some variables may give better results but may be difficult, costly or ethically prohibitive to obtain.

The sample used in Smith et al. (1993) is a subset of the data used for this analysis that has been investigated earlier in the data collection process. It includes 2597 individuals and the cases with aortic diameter 29 mm have been regarded as normal. Hence, there are some differences in the results, and in our analysis more variables were included.

Tables 2.4 on page 46, 2.5 on page 46 and 2.6 on page 47 summarise the results of our risk factor analysis. These tables include the estimated prevalence for normal and abnormal cases, the results of χ^2-test for comparing prevalences and also the corresponding odds ratio and relative risk for each variable. Moreover, the attributable risk, defined as the fraction of the risk of having a disease that can be uniquely attributed to the presence of a specific risk factor, was estimated. In all cases, the 95% confidence interval and the p-value are also given. Bear in mind that the definition of abnormal cases is that the aortic diameter is at least 29 mm.

The following notation has been used:

- `mi`: myocardial infarction
- `ihd`: ischaemic heart disease
- `cvd`: cardiovascular disease
- `pvd`: peripheral vascular disease
- `coad`: chronic obstructive airway disease
- `bpd`: diastolic blood pressure at least 90 mmHg
- `bps`: systolic blood pressure at least 160 mmHg
- `smol`: smoking level
  - 0: never smoked
  - 1: moderate smokers
  - 2: heavy smokers
  - 3: inaccurate information
  - 4: poor information
- `diab`: diabetes present
- `ht`: hypertension treatment
- `alco`: alcohol consumption (at least a unit per week)
- `cva`: cerebrovascular accident (stroke)
- `co-mor`: presence of at least one of the co-morbidities or family history of AAA
- `age`: age at least 70 years

Risk factor	mi	ihd	cvd	pvd	coad
Prevalence of AAA for predictor absent	0.074	0.074	0.075	0.078	0.078
Prevalence of AAA for predictor present	0.180	0.175	0.140	0.141	0.143
Prevalence of predictor for normal cases	0.076	0.080	0.116	0.077	0.074
Prevalence of predictor for abnormal cases	0.185	0.189	0.209	0.141	0.137
χ^2 for prevalence	34.41	33.02	18.2	12.02	12.18
p-value for χ^2 value	<0.001	<0.001	<0.001	<0.001	<0.001
lower limit (2.5%) for relative risk	1.808	1.768	1.403	1.299	1.308
estimated relative risk of AAA	2.430	2.373	1.871	1.816	1.836
upper limit (97.5%) for relative risk	3.266	3.184	2.496	2.538	2.576
p-value for relative risk	<0.001	<0.001	<0.001	<0.001	<0.001
lower limit (2.5%) for odds ratio	1.938	1.89	1.455	1.331	1.341
estimated odds ratio of AAA	2.743	2.665	2.013	1.95	1.975
upper limit (97.5%) for odds ratio	3.883	3.758	2.785	2.856	2.908
p-value for odds ratio	<0.001	<0.001	<0.001	<0.001	<0.001
lower limit (2.5%) for attributable risk	0.058	0.058	0.041	0.018	0.017
estimated attributable risk of AAA	0.109	0.109	0.097	0.063	0.062
upper limit (97.5%) for attributable risk	0.157	0.158	0.150	0.107	0.105
p-value for attributable risk	<0.001	<0.001	<0.001	0.007	0.007

Table 2.4: Risk factor analysis results

Risk factor	bpd	bps	smol: 0	smol: 2	diab
Prevalence of AAA for predictor absent	0.069	0.074	0.087	0.059	0.084
Prevalence of AAA for predictor present	0.108	0.101	0.046	0.108	0.058
Prevalence of predictor for normal cases	0.355	0.323	0.091	0.480	0.059
Prevalence of predictor for abnormal cases	0.474	0.402	0.048	0.643	0.040
χ^2 for prevalence	13.83	6.32	5.28	24.25	1.48
p-value for χ^2 value	<0.001	0.010	0.022	<0.001	0.224
lower limit (2.5%) for relative risk	1.230	1.066	0.299	1.433	0.372
estimated relative risk of AAA	1.566	1.363	0.527	1.848	0.688
upper limit (97.5%) for relative risk	1.994	1.743	0.930	2.382	1.273
p-value for relative risk	<0.001	0.010	0.027	<0.001	0.233
lower limit (2.5%) for odds ratio	1.264	1.081	0.279	1.497	0.349
estimated odds ratio of AAA	1.634	1.404	0.505	1.950	0.669
upper limit (97.5%) for odds ratio	2.112	1.824	0.913	2.541	1.283
p-value for odds ratio	<0.001	0.010	<0.001	0.024	0.226
lower limit (2.5%) for attributable risk	0.072	0.016	-0.073	0.172	-0.044
estimated attributable risk of AAA	0.171	0.107	-0.043	0.295	-0.018
upper limit (97.5%) for attributable risk	0.260	0.190	-0.015	0.399	0.007
p-value for attributable risk	0.001	0.020	<0.001	0.003	0.154

Table 2.5: Risk factor analysis results

Risk factor	ht	alco	cva	co-mor	age
Prevalence of AAA for predictor absent	0.081	0.079	0.080	0.058	0.069
Prevalence of AAA for predictor present	0.096	0.085	0.250	0.140	0.098
Prevalence of predictor for normal cases	0.155	0.714	0.012	0.284	0.48
Prevalence of predictor for abnormal cases	0.181	0.731	0.044	0.510	0.574
χ^2 for prevalence	1.16	0.31	14.22	55.18	8.19
p-value for χ^2 value	0.281	0.580	<0.001	<0.001	0.004
lower limit (2.5%) for relative risk	0.869	0.818	1.833	1.885	1.109
estimated relative risk of AAA	1.185	1.079	3.106	2.396	1.418
upper limit (97.5%) for relative risk	1.615	1.423	5.263	3.045	1.814
p-value for relative risk	0.283	0.592	<0.001	<0.001	0.005
lower limit (2.5%) for odds ratio	0.860	0.818	1.902	2.026	1.132
estimated odds ratio of AAA	1.204	1.086	3.808	2.622	1.464
upper limit (97.5%) for odds ratio	1.686	1.442	7.624	3.395	1.892
p-value for odds ratio	0.279	0.568	<0.001	<0.001	0.004
lower limit (2.5%) for attributable risk	-0.027	-0.152	0.006	0.206	0.046
estimated attributable risk of AAA	0.028	0.053	0.030	0.297	0.169
upper limit (97.5%) for attributable risk	0.081	0.222	0.054	0.378	0.277
p-value for attributable risk	0.313	0.584	0.016	<0.001	0.009

Table 2.6: Risk factor analysis results

Based on the results in previous tables, the variables that are indicated to be used in the models in the following sections are co-morbidities and family history indicator, smoking level, age and diastolic blood pressure.

From the clinical point of view, as we have mentioned previously, the use of a single indicator can be justified by the fact that these co-morbidities are all considered as manifestations of degradation of the circulatory system and can be seen as symptoms of the same root cause. A related medical reason is the correct identification of each the co-morbidities and the accuracy of classification of highly correlated diseases. Previous studies have suggested that there is evidence of misclassification of between coronary heart disease and myocardial infarction (De Henauw et al. 1998 and Lenfant et al. 1998).

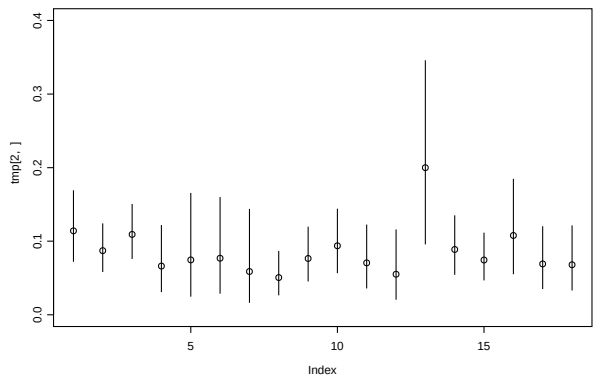
From the statistical point view, the inclusion of highly correlated variables with small number of cases for each combination of the variable levels might lead to problems associated with collinearity and sparse data. Trying to keep the predicting model as simple as possible has been another reason for reducing the number of predictors as far as possible. Additionally, the combined indicator for the co-morbidities has been found as stronger predictor than any of the co-morbidities separately.

Hence, for both clinical and statistical reasons, it has been decided to use the singe indicator. This is not without limitations, for example the research question involves identifying which

particular co-morbidities are associated with other risk factors in the study such as smoking.

For continuous variables, we compared the mean value of the predictor for the individuals with AAA with the corresponding mean value for those without AAA (using 5% significance level). In this way, the mean age of those with AAA is significantly different from the mean age of those without AAA. The same results are obtained for diastolic and systolic blood pressure and for the number of cigarettes per day. On the other hand, the mean number of years since last smoked and alcohol consumption for cases with AAA are not significantly different from those without AAA.

Furthermore, we investigate the influence of the practice at which each individual has been scanned for aortic aneurysm detection. The plot of the AAA prevalence and the corresponding 95% confidence for each practice can be seen in figure 2.13 on page 48.



95% confidence interval indicated by lines.

Figure 2.13: AAA prevalence for each practice

The corresponding χ^2-test indicates that practice is not significant even though this might be due to the small number of individuals with AAA in some practices.

Similarly to practice, we test the significance of day, month and year of scan results. In this case, we investigate whether abdominal aortic aneurysm prevalence is influenced by "hidden time effects" (Altman 1999, page 133).

In particular, the author mentioned above states that in a study "in which subjects are recruited over some months or years it is possible that there may be changes in the characteristics of the subjects or in the measurements made on them". No time related variable has been found as significant.

2.13 Bivariate analysis

In order to identify important associations between variables in CASP dataset, we analyse these variables in pairs. A scatter plot for each pair might provide useful insight about the type of relationship between these variables, for example if it is linear. Furthermore, the degree of agreement of the variables can be an indication of collinearity when both these variables are included as predictors in a model.

Instead of having separate scatter plots for each pair, it is possible to have a matrix of scatter plots. In this case, all possible pairs can be visually inspected at once and it might be possible to identify bivariate association patterns that are difficult to identify otherwise. Examples of these could be a non-linear relationship for a number of variable pairs which could be an indication for the necessity of transformations.

For the data related to abdominal aortic aneurysm we analyse, it would be useful to plot the response variable and other variables that are either found significant by univariate tests or they might be clinically important. The result is shown in figure 2.14 on page 49.

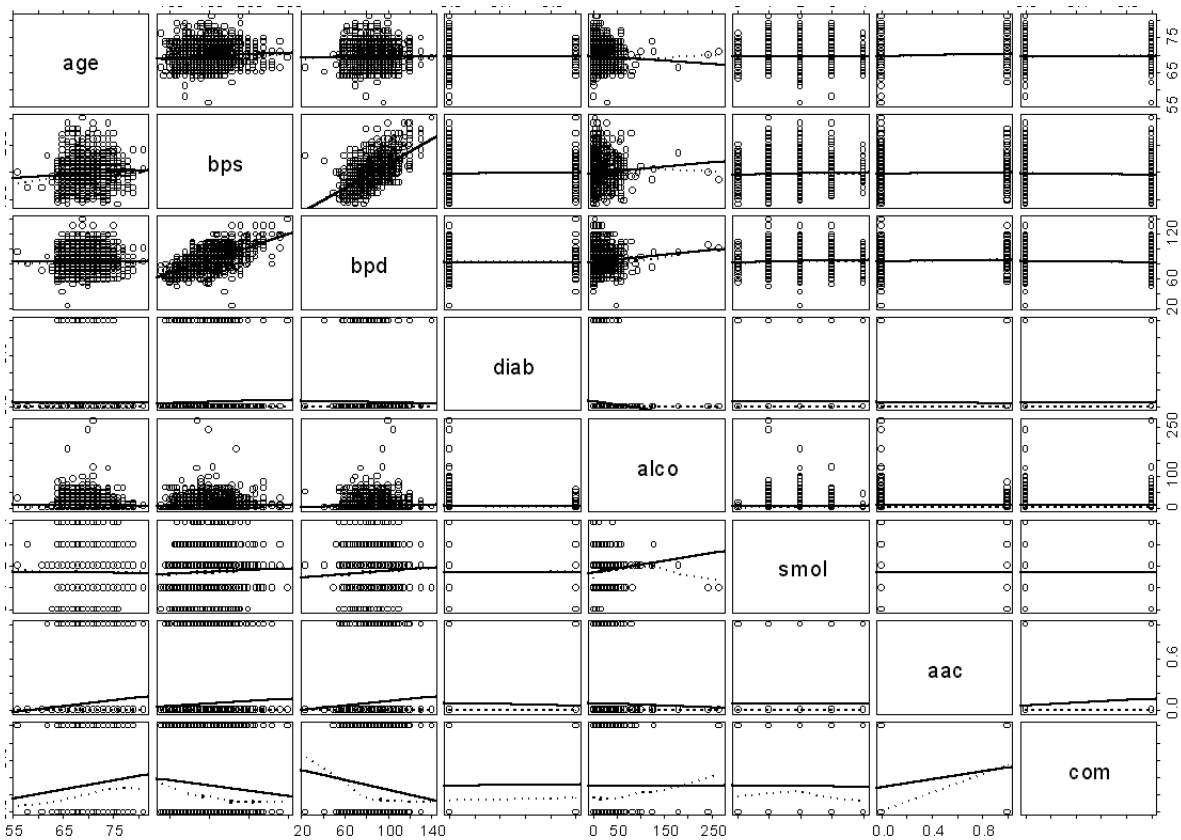


Figure 2.14: Scatter plot matrix

In each cell, we add the fitted linear and locally weighted regressions; these might be useful to reveal non-linear relationships between pairs of variables. Diastolic blood pressure (`bpd`) is highly correlated with systolic blood pressure (`bps`) and the relationship appears to be linear. Additionally, both blood pressure variables for abnormal AAA cases seem to be slightly increased on average when comparing with the corresponding normal cases.

On the other hand, the correlation between most of the other variables can be said to be small. Exceptions are the association between aortic aneurysm indicator (variable `aac`) and co-morbidities and family history of AAA indicator (variable `com`) and between AAA indicator and age.

From the scatter plot matrix, the relationship between aac and smol (smoking level) is not clear. This is due to the fact that both variables are categorical and a scatter plot might not be useful for such a pair of variables. Based on the results from univariate and bivariate analysis, we can conclude that com, age and blood pressure measurements might be useful predictors of AAA.

In some cases, it is also useful to quantify the degree of correlation between variables, especially when the type of bivariate relationship is clear from the scatter plot. In our case, we use Kendall's method to investigate the correlation between AAA and smoking level indicators. From the results (not shown), we can conclude that smoking level indicator is most probably not a useful predictor for AAA indicator. On the contrary, from χ^2-tests, we have seen that smoking level variable is associated with AAA response variable, a fact that leads to contradictory results. This might be due to the small number of cases for some levels of smoking variable which makes correlation and χ^2-tests unreliable.

If we remove the individuals with poor or inaccurate values of smoking level, Kendall's method shows that variable `smol` is significantly associated with variable `aac`. In other words, given two random individuals, the one with the higher `smol` also tends to have higher `aac`.

As we have contradicting results from the association significance tests, we need to look into the literature in order to confirm or reject the inclusion of variable `smol`. Because smoking has been suggested as one of the main causes of abdominal aortic aneurysm (see for example Vardulaki et al. 2000 and Blanchard et al. 2000), we will include the smoking level variable into further analysis.

Also, we might pursue alternative ways of analysing the data related to smoking. For example, we could implement different coding for this variable by joining levels with small number of individuals. In this case, we must take into account what each level indicates and avoid mixing groups of individuals that are unlikely to have similarities. In other words, we should try not to introduce a possible source of bias for the sake of numerical stability of statistical tests.

Finally, we could decide that the quality of data for some groups of individuals in our study is

not trustworthy and should be treated as missing. For smoking level variable, poor and inaccurate information levels can be regarded as indicators of unreliable data and might be imputed. Bear in mind that smoking level variable is based on self-reported information hence any value should be treated with caution.

2.14 Multivariate exploratory analysis

Similarly to bivariate relationships, we investigate associations of more than two variables at a time. In some cases, two variables might be marginally correlated but not correlated at all when this correlation is conditional on other variables. This phenomenon is known in the statistical literature as Simpson's paradox; see Edwards 2000 for examples.

Standard methods of multivariate exploratory analysis include, as we have previously seen, principal components and factor analysis, multidimensional scaling and cluster analysis. Before implementing these methods, it might be useful to investigate the conditional distributions of variables by graphical methods.

An example of a graphical multivariate method is the *conditioning plot or coplot*, which is a plot that "displays the bivariate relationship between two variables while holding constant (or 'conditioning upon') the values of one or more other variables" (Everitt et al. 2001).

For example, we can use the coplot to investigate the relationship between age and systolic blood pressure for each abdominal aortic aneurysm indicator. The result is shown in figure 2.15 on page 52.

From this conditioning plot, where we also included linear and locally weighted regressions, it appears that the relationship between age and bps for normal and abnormal AAA is similar but not identical. For normal AAA cases (left half of the figure), bps on average is increasing up to the age of 70 years and then it stays the same. On the other hand, for abnormal AAA individuals (right half of the plot), bps is the same up to 70 years on average and then it increases steadily.

If we include co-morbidities and family indicator (com) as additional conditioning variable, we obtain figure 2.16 on page 52. Comparing the lower half of figure 2.16 on page 52 with figure 2.15 on page 52, we can conclude that relationships between bps and age for all cases are the same as the corresponding relationships all individuals without co-morbidities or family history of AAA.

On the other hand, by examining the upper half of figure 2.16 on page 52, we might state that systolic blood pressure does not seem to increase with age when co-morbidities are present. This can be regarded as an illustration of the importance of having additional conditioning covariates when examining bivariate relationships.

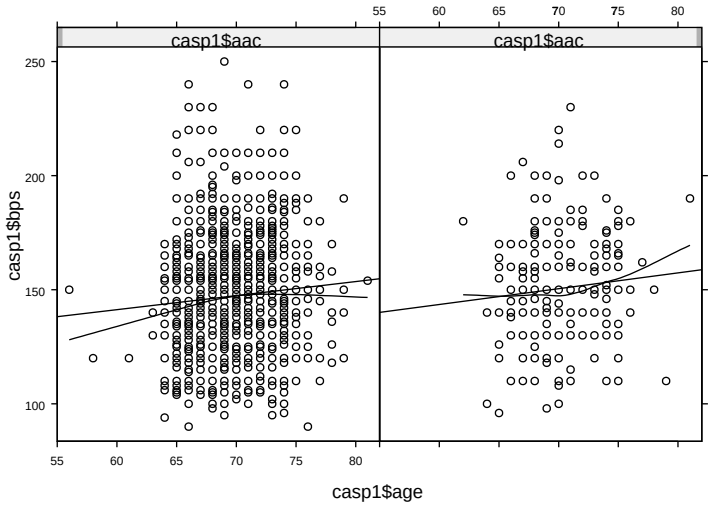
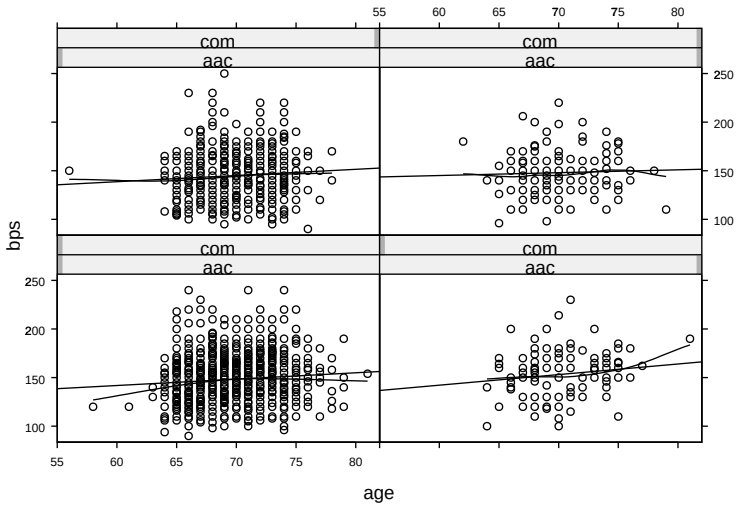


Figure 2.15: Age and bps conditional on aac



Right column:`aac`=1
Left column:`aac`=0
Top:`com`=1
Bottom:`com`=0

Figure 2.16: Age and bps conditional on aac and com

In order to verify the conclusions drawn from figure 2.16 on page 52 using association tests, we implement χ^2-tests for categorical versions of systolic blood pressure (bps) and age. Specifically, we apply 130 mmHg and 160 mmHg as thresholds for bps and 70 years as threshold for age and the corresponding factors are denoted as bpsgr and agegr2.

The cut-off points that have been applied (130 mmHg and 160 mmHg) correspond to the 25% and to 75% quantiles of the diastolic blood pressure distribution. In this way, the data has been divided into three parts: the lower quartile, the inter-quartile range and the higher quartile. This is often used for exploratory purposes to identify linear trend and we agree that regression is the appropriate method of identifying the relationship between the variables used. We used this example to illustrate the usefulness as well as the limitations of multivariate exploratory analysis.

We can conclude that the χ^2-test for the individuals with AAA (aac=1) and without co-morbidities (com=0) indicates no significant association, a fact that might be seen as contradiction to the conclusion drawn from the lower right cell of the conditioning plot in figure 2.16 on page 52. This might be attributed to the small number of cases that belong this group (127 individuals). The other χ^2-tests verify the conclusions derived by the corresponding conditioning plot.

Another possible question that can be answered by the conditioning plot is whether systolic blood pressure is increasing by age for individuals that have been scanned in different years. Again the results (figure 2.17 on page 53) shows that for different subsamples, the association between two covariates can be substantially different.

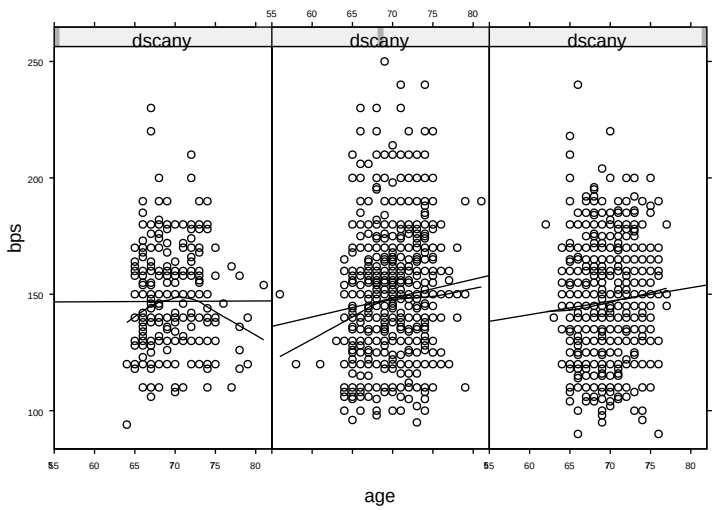


Figure 2.17: Age and bps conditional on year of scan

If we include AAA indicator (aac) as an additional variable to condition the relationship between

systolic blood pressure and age on, we have figure 2.18 on page 54. From this conditioning plot, we see that the association is different for different subgroups, even though we need to remember that the upper half of figure 2.18 on page 54 contains small number of individuals. In this case, χ^2-tests might not be reliable for verification of the conclusions drawn from the conditioning plot.

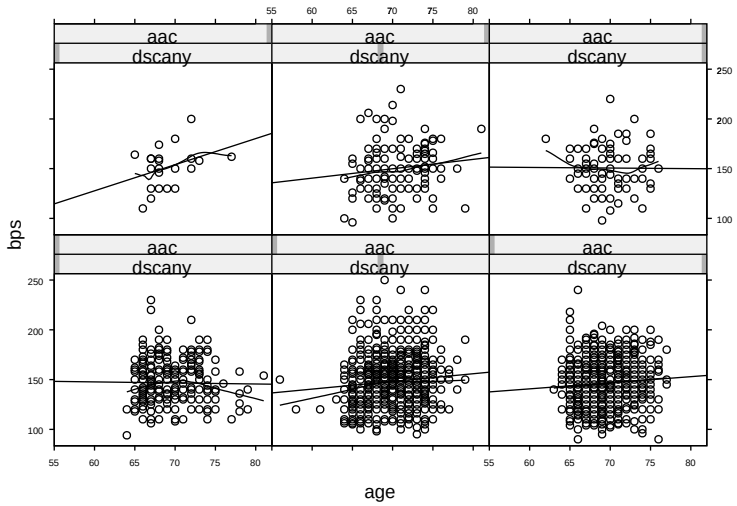


Figure 2.18: Age and bps conditional on year of scan and AAA indicator

Other methods of multivariate exploratory analysis such as principal components, factor analysis, multidimensional scaling and cluster analysis have been tried. Probably due to the fact that most of the covariates are not correlated between them, these methods confirmed that the variables in the data cannot be summarised into linear or other types of combination of covariates (results not shown). In other words, it is not possible to explain a large proportion of the variability between cases by using the multivariate exploratory methods mentioned above.

Hence, we can conclude that most interactions between predictors of AAA should not be included unless this is suggested by previous publications or medical experts. On the other hand, from conditioning plots, it can be suggested that there are indications of possible significant interactions for subsamples of the dataset.

We must be cautious about the significance of conclusions for parts of the data, as these might be spurious relationships that are falsely found as important. A possible reason for that is the fact that inspecting a large number of conditioning plots for different subsamples is similar to the implementation of a large number of hypothesis tests.

Hence, we will initially attempt to construct regression and classification models without interactions. Later on, we will investigate the significance of interactions for predictions. Finally,

we construct graphical models for exploratory and predictive purposes; this type of model allows searching for complex associations between variables that can be useful for different objectives.

2.15 Conclusions

Data preparation and exploratory analysis have been presented and their application has been examined using an empirical study on abdominal aortic aneurysm (AAA) screening. We showed that dealing with missing data and subsequent recoding of specific variables should take into account both clinical and statistical aspects of the analysis. For AAA screening, based on information collected about smoking habits, we have created an indicator with coding that reflects both the smoking level and the quality of the data.

Descriptive statistics commonly used in medical research have showed that even for a study with a small number of variables, the structure of the data might be very complex. Examples of the AAA data complexity have been demonstrated by examining the distributions of the aortic diameter and also the corresponding distributions of smoking and alcohol consumption.

Univariate analysis has been implemented to examine the relationship between the response variable and each of the predictors and has been shown to be very useful for many purposes. For the AAA study, co-morbidities and family history indicator, smoking level, age and diastolic blood pressure have been found as significant predictors of AAA presence.

Using bivariate analysis, we found, as we have expected, that diastolic blood pressure is highly correlated to systolic blood pressure. Smoking was found not to be associated with the AAA presence, a result that is contradicting the significance of smoking by univariate analysis.

Multivariate exploratory analysis showed the complex relationship between age and systolic blood pressure. In addition, we showed the limitations of exploratory analysis by comparing the results of analysis when adjusting for different control factors.

Chapter 3

Regression analysis

3.1 Abstract

In this chapter, we compare different classification models for AAA selective screening using the area under the ROC curve as criterion. We find that two logistic regression models can be used for selecting patients for screening. The first includes the co-morbidities and family history of AAA indicator, smoking level, and age group indicator as predictors of the presence of AAA and the second has in addition to the predictors of the first model raised diastolic blood pressure. The importance of assessing the performance of a screening classification model by cross-validation and bootstrapping is also shown. Other definitions of abnormality, weighted classification and multiple class models are also examined.

3.2 Variable and model selection: risk factors and co-morbidities

There are many epidemiological studies that identify specific risk factors for abdominal aortic aneurysm and related co-morbidities. Vardulaki et al. (2000) analysed their data related to abdominal aortic aneurysm by fitting logistic regression models. In this case, diastolic blood pressure, hypertension treatment, male gender and cigarette consumption level have been found as risk factors that increase significantly the risk of developing AAA.

Smoking has been identified as the main cause of AAA and some of the widely accepted risk factors such as hypertension and diabetes might not increase the risk of this disease (Smith et al. 1993). Furthermore, it has been suggested that AAA is not related to atherosclerosis and that smoking is associated with enzymes such as the metalloproteinases (Pyo et al. 2000). These enzymes are known to be active in other conditions such as chronic obstructive pulmonary disease.

Other studies related to AAA illustrate that the number of predictors of the presence of AAA is quite large and diverse. Singh et al. (2001) find that the prevalence of AAA increases with age, smoking and cholesterol is more prevalent amongst individuals taking antihypertensive medication. Sonesson et al. (1994) identify weight, height, body surface area and age as predictors of abdominal aortic diameter. Wilmink et al. (1999) show that the duration of smoking rather than smoking level predicts the AAA presence.

Rodin et al. (2003) state that the occurrence of AAA is related to male gender, age, cholesterol, smoking, height and higher systolic or diastolic blood pressure. Finally, in a study for testing the potential of a selective screening strategy for AAA, they find that age, place of birth, having a sister with AAA, smoking, history of myocardial infarction or coronary artery bypass, treatment of hypertension and diet for cholesterol are risk factors of AAA.

For the construction of AAA selective screening tests, based on the data collected for the CASP project, there are significant risk factors included such as smoking consumption level and duration of smoking, diastolic and systolic blood pressure, hypertension treatment and of course age (Smith et al. 1993). On the other hand, no information has been collected related to cholesterol levels and in some cases more reliable markers could have been used; for example serum levels of cotinine instead of using the number of cigarettes per day and the number of years since last smoked.

The fact that the information quantity and quality might not be at a desirable level can be partly attributed to the constraints we might have for data collection. These constraints are related to the information and the technology available at the time that the study is designed, and to ethical considerations. For example, genetic screening for several diseases might provide the ultimate way of identifying high risk cases but this is not likely to be used in a broad scale within the foreseeable future.

Another way to use this dataset more efficiently is to use several statistical methods to improve the performance of a potential classification method. For example, the mean arterial pressure

$$\texttt{map} = \frac{1}{3}\texttt{bps} + \frac{2}{3}\texttt{bpd}$$

might be used, although it was not found to be significantly correlated with AAA (see Vardulaki et al. 2000). Additionally, other variables such alcohol consumption level, diabetes and other co-morbidities might be found to improve the accuracy of a classification model.

Finally, we would like to emphasise that for a case selection classification method, some variables might not be used either because they are not easily obtainable or very expensive to obtain. In Breiman et al. (1984), it is mentioned that even though more successful classification can be made

by using invasive measurements, they are not used as they present some risk to the patient.

In the case of AAA, the variables obtained might not give the same sensitivity, specificity and positive predictive value as ultrasound screening but case selection classifiers might be useful as cost-efficient tools that can combined with ultrasound screening.

Hence, we believe that it is possible to achieve constructing a case selection or selective screening test based on information collected by a questionnaire that will identify high risk cases. The objective in this case might also be to prioritise high risk individuals for screening in a practice with a large number of patients to be screened.

3.3 Regression models

In the previous chapters, we described methods to explore the data in order to acquire useful information about the nature of the dataset. For instance, this could be the distributions of the variables and the relationships between these variables.

Furthermore, it is possible that the purpose of our study is to compare different groups of subjects by comparing, for example, characteristics of these groups such as the mean of one or more variables. In this case, the exploratory analysis is sufficient for giving answers to the questions set by the researcher.

On the other hand, "we might wish to describe the relation between (a number of variables), and thus be able to predict the value of one variable when we only know the other variable(s). Clearly the correlation coefficient does not perform these functions; it just indicates the strength of the association as a single number. We want a way of describing the relation between the values of (these) variables, and for this general problem we need the technique called *regression*" (Altman 1999).

Edwards (2000, page 7) states that "statistical techniques have two main purposes: explanation and prediction". Regression analysis tends to be applied mostly for prediction as the models are constructed mainly to estimate the values of one or more variables called *responses* by using other variables called *predictors*.

The same model can be seen as a way to describe the relationships between variables and specifically the conditional distribution of the response variables given the predictors. The relationships between the variables and the predictions from the regression model can be regarded as part of the information required for *decision making*.

In both cases, the objective is to construct a model that combines different parts of the dataset; the information from exploratory analysis should be included into the process of modelling. To be

more specific, consider the situation of trying to understand a particular situation described by a number of variables and at the same time attempting to forecast the values of some variables given the others.

Then, the results from exploratory analysis are indicators of the complexity of the relationships between the variables, hence they can be seen as guidance for choosing the type of model and the variables to be included. Also, the nature of the relationships between predictors and responses e.g. strength of association and its form indicate the variables that are most likely to give accurate forecasts of the responses.

Moreover, the list of possible models constructed to describe the data and to predict response variables is endless; thus, it is more time efficient to explore the data initially and combine these results with external information in order to build a model that covers both explanation and prediction objectives by avoiding unnecessary statistical analyses.

Two of these regression models that appear in the medical literature are described below; the choice of type of model depends on the type of response variable. Specifically, if the response is on continuous scale, then linear regression is used. On the other hand, if the response is categorical and especially binary, logistic regression is implemented. Details about linear and logistic regression models are given in the appendix.

3.4 Classification models

Michie et al. (1994, page 1) state that the term classification covers "at its broadest ... any context in which some decision or forecast is made on the basis of currently available information, and a *classification procedure* is then formal method for repeatedly making such judgements in new situations".

Furthermore, classification is, according to authors mentioned above, "the construction of a procedure that will be applied to a continuing sequence of cases, in which each new case must be assigned to one of a set of pre-defined classes on the basis of observed *attributes* or *features*".

Three main groups of classification or *pattern recognition* methods according to Michie et al. (1994, page 2) are statistical approaches, machine learning methods and neural networks. Statistical methods are related mostly to Fisher's work on linear discrimination and to density estimation methods. Machine learning approaches include decision-trees and rule-based methods. Finally, neural networks are concerned with producing non-linear functions of the attributes by using layers of interconnected nodes.

Graphical models are another possible way of discriminating subjects into different classes by

their attributes. These models can be applied for classification even though they are not built primarily for this purpose; it is possible to modify the existing models in order to improve their performance. Naive Bayes is the most common classifier of this type found in the literature.

In all cases, classification models or *classifiers* as they are also called, have two main objectives: accurate predictions and uncovering the predictive structure of the problem (Breiman et al. 1984, page 6). Hence, the question of finding the best classifier is not always answered by a single model that has the best predictive performance e.g. by reducing the error rate. In other words, classification depends on what "the researchers are really interested in, and what it is they want to optimise" (Hand 1997, page 3).

3.4.1 Statistical methods

Logistic regression is applied not only to define the relationship between the response variable and the predictors, expressed by the corresponding regression coefficients but also to "predict the probability of a particular outcome in relation to several prognostic variables" (Altman 1999, page 355).

It is obvious that in this way, we are able to classify subjects in the study into two groups, depending on their probability of having a disease or in general being related to an event of interest. This type of analysis is called *discriminant analysis*.

Another statistical method of classifying individuals into distinct groups is Fisher's linear discriminant analysis. Specifically, "a hyperplane in the attribute p-dimensional attribute space is chosen to separate the known classes as well as possible. Points are classified according to the side of the hyperplane that they fall on" (Michie et al. 1994, page 18).

This method is reliable under the assumption that all the variables have a normal distribution with the same standard deviation within each group. A possible solution to the problem of unequal covariance matrices is quadratic discriminant analysis; a possible drawback of this method is the large of parameters that we need to estimate. See Michie et al. (1994) for further details about these methods.

The difference between logistic regression classification and Fisher's discriminant analysis is the criterion used for good separation between classes. A quadratic function is optimised by Fisher's method where logistic method maximises conditional likelihood. Michie et al. (1994) state that these methods give similar results when the attributes have normal distributions with equal covariances and different results when dealing with non-normal distributions and dissimilar covariances.

Hence, when for example, binary or nominal predictors are included, it is better to apply logistic regression classification as the assumptions for using linear or quadratic discriminant analysis are not met. On the other hand, when the assumptions for Fisher's method hold and we need a classifier for more than two groups, then linear and quadratic discriminant analysis seem more appropriate.

Density estimation methods, which are distribution-free (nonparametric) classification procedures are based on the estimation of the densities of the attributes for each class. Subsequently, using the densities and the prior probabilities for each class, the subjects in the study are classified. Kernel functions are the basic elements of the estimating the densities mentioned above (Michie et al. 1994, page 30).

K-nearest neighbour ($k - nn$) method is closely related to density estimation methods. In this case, a new observation is classified using the k nearest observations of this observation. The majority rule is applied to choose the class to which the new observation belongs to; in other words, if most of the k nearest belong to a particular class, then it is assumed that the observation under examination belong to this class. Michie et al. (1994, page 35) mention that "the nearest neighbour method is equivalent to the kernel density as the smoothing parameter tends to zero, when the normal kernel function is used".

For nearest neighbour method, the choice of metric for defining the distance between observations and the number k of neighbours is important. For example, if discrete variables are included in the data, then data are transformed so that "all observations lie in the unit hypercube" (Michie et al. 1994, page 35). Furthermore, the choice k might be done by cross-validation; the problem of choosing a rule when ties occur is another difficulty when using this method.

3.4.2 Machine learning approach

Recursive partitioning methods such as decision trees and rule-based approaches are procedures related to machine learning, that can be described as "automatic computing procedures based on logical or binary operations, that learn a task from a series of examples" (Michie et al. 1994, page 2). Furthermore, they state that the aim of these methods is to "generate classifying expressions simple enough to be understood easily by the human".

From what is mentioned above, it is clear that both accurate classification and *mental fit*, the ability to make the data structure understandable to humans is equally important. In medicine, these methods have been implemented for diagnosis of a particular disease from related symptoms as they are able to support their decisions with insight into causes; statistically derived systems

do not (Michie et al. 1994, page 52).

Machine learning methods have been seen as procedures that might have an advantage when categorical rather than numerical attributes are used and there are "strong and pervasive conditional dependencies among attributes" (Michie et al. 1994, page 53). By the term *conditional dependencies*, it is implied that there are important interactions between attributes and the response variable.

There are two broad inductive inference strategies implemented from classification trees and rule-based methods: specific-to-general and general-to-specific. On the one hand, specific-to-general procedures start with "a maximally specific rule for assigning cases to a given class" (Michie et al. 1994, page 54). Subsequently, the rule derived is becomes more general by dropping attributes one at a time. After that, specific rules can be explored and combined or separated further until all rules consistent to the data are found. A prespecified percentage of correctly classified cases defines the acceptability of a rule.

On the other hand, general-to-specific strategy starts by assigning all subjects to a single class, preferably the class where the majority of cases belong to. After that, each attribute in the data is used to split the sample into subsets; the split that minimises *impurity* is retained. An impure node is defined as a node where there are correctly classified cases and also incorrectly classified ones.

An impurity index, is "a measure for a given node, of the differences between the probabilities of belonging to each class" (Hand 1997, page 66); the same author also mentions several possible candidate functions used as impurity indices. Furthermore, "goodness of a split into subsets is the weighted mean decrease in impurity and the weights are proportional to subset sizes" (Michie et al. 1994, page 58).

For tree models, stopping rules for growth are applied to avoid overfitting the training data used to construct this model. To be more specific, for continuous data, it is possible to construct a tree such that all leaf (terminal) nodes contain only one case. This way of growing a tree will lead to "overlarge trees with many "twigs" which merely model random variation in the design set arising from the sampling procedure, rather than modelling any real underlying structure of the populations" (Hand 1997, page 71).

A possible rule is to stop when impurity is lower than a specific level; another is to prune the tree after finishing the growth procedure. Different criteria for pruning are used depending on the purpose of constructing such a model; for example, the number of nodes which is an indicator of the tree size can be applied. Growing and pruning trees are procedures that can be compared with

forwards and backwards stepwise regression; see Hand (1997) for further details.

Another approach for finding the optimum tree model is *tree averaging*. In this case, "one constructs several, even many, trees and averages their class predictions to produce a final classification" (Hand 1997, page 73). Of course, there are many possible ways of combining several models for classification; one of them is described in Hand (1997, page 73-74).

In addition to that, another partitioning criterion that can be implemented is combining predictors instead of splitting the data into subsets using one attribute at a time. The possible ways of such combinations is very large, especially when having a large number of variables to choose from. Another drawback of having multivariate splits is that "a sequence of splits on (such) splits can be complex to explain" (Hand 1997, page 74). A possible way to overcome this problem is including only a small number of variables and implement recursive partitioning by suitable linear combinations of these variables.

It should be stressed that we can convert tree models into an equivalent set of rules and vice versa. Hence, it is possible to derive from data a classification tree by implementing that is available to us and convert the results into a set of rules that "seem to lend themselves more into user-friendliness" (Michie et al. 1994, page 79).

Support Vector Machines (SVM) are another machine learning classification method. The input vectors (predictors) are mapped to a very high dimension feature space. SVM finds a linear decision surface that tries to maximise the margin of confidence of the classification of the data. Cortes et al. (1995) state that "special properties of the decision surface ensures high generalisation ability of the learning machine". This is attributed to the fact that the optimal margin classifier executes "Occam-Razor principle" by reducing the training data error and prevents over-fitting.

Finally, external information, either from preliminary analysis or from related literature and subject experts can be of great value on constructing classification trees. Furthermore, interaction between software and user can lead to better results; replacing some of the original attributes with transformations or combinations of them based on the results from the original attributes is an example of this interaction.

3.4.3 Neural networks

Research in the field of neural networks originates from "an interest in understanding and modelling the way organic brains function" (Hand, 1997, page 44). Firstly, a model similar to linear discriminants has been proposed, called a *neuron*, which "consists of a weighted sum of its inputs, followed by a non-linear function called the em activation function, originally a threshold function"

(Michie et al. 1994, page 84).

Subsequently, the same authors state that "it was established that the functionality of neural networks was determined by the strengths of the connections between neurons; ... if the network responds in a desirable way to a given input, then the weights should be adjusted to increase the probability of a similar response to the future. Conversely, if the network responds undesirably to an input, the weights should be adjusted to decrease the probability of a similar response".

The nodes are divided into input and output nodes; in a *multi layer perceptron structure*, a middle layer of nodes also exists and it is also called a hidden layer as it is not "visible to neither the inputs nor the outputs" (Michie et al. 1994, page 87). Additionally, they state that "unlike the input and output layers, (the hidden layer's) size is not fixed. The hidden layer is generally used to make a *bottleneck*, forcing the network to make a simple model of the system generating the data, with the ability to generalise to previously unseen patterns".

Moreover, feed-forward neural networks can be seen as a flexible way to generalise linear regression functions. For a neural network with one hidden layer, there are two steps in the estimation procedure. Firstly, the weighted sum of the attributes plus a constant, which is called the *bias* are transformed using an "activation function"; most of the time, this function is the logistic function.

In the second step, the sum of the outputs of the activation functions plus another bias term is transformed by an output function and that is the result used for discriminating the subjects in the sample into classes using their attributes. We can also mention here that the use of logistic function as activation function is one of the reasons why classification neural networks are regarded as generalisation of multiple logistic regression (Venables and Ripley 1997, page 490).

For classification, the objective is to define the weights that will optimise a criterion of goodness of fit. This can be achieved by minimising the sum of the squared differences between the actual and predicted response values. Another fitting criterion could be the weights that lead to maximisation of the likelihood of the data; see Michie et al. (1994, page 89) for further details.

Radial basis function networks are another type of neural network that can be implemented for classification. They consist of a layer of units with functions of the attributes and a layer of output functions that are connected to the attribute functions ; the form of the output is the same as the target vectors. They have a number of advantages over multi layer perceptron neural networks in some situations; see Michie et al. (1994, pages 93-96) for further details.

A particular advantage of neural networks is that they can approximate any continuous mapping closely if unlimited number of hidden layers can be used; this is why a two-way multi layer perceptron can be seen as a universal function approximator (Michie et al. 1994, page 88).

The same authors mention that the activations of the hidden nodes tend to form an orthogonal set of variables and it is possible to transform them into a principal component representation of the attribute space. In addition to that, this representation can act as a filter of the "lower-variance noise signal, provided the signal to noise ratio of the data is sufficiently high" (Michie et al. 1994, page 88). They also state that the hidden layers can be seen as "detectors of abstract features of the attribute space" and modelling involves only "the important underlying structure of the generating system".

Particular problems of neural networks are the choice of number of hidden layers since a small number will not achieve learning of the structure and a large number will lead to overfitting. Another problem is the removal of unnecessary nodes and addition of others that will improve the performance of the neural networks. Moreover, the stopping criterion when constructing this type of model is essential, especially when dealing with missing or noisy data and when the distributions of the attributes for each of the classes overlap (Michie et al. 1994, page 96).

Finally, neural networks are regarded as "black box" methods and their application is often rejected because of the fact that the way of deriving the results is not transparent as it is for example with logistic regression or classification tree models. This is related to the choice of objective for constructing the model.

Specifically, if "mental fit", the understanding of the structural relationships is important, then neural networks may not be the best candidate. On the other hand, if the performance of the neural network is superior to other methods and it will implemented with the single purpose of accurate predictions, then we should consider them as the best choice.

In medicine, understanding the decision process and fast deriving of the result are sometimes of greater importance than accuracy of classification and flexibility of modelling procedure. Hence, we believe that a neural network could be implemented as alternative solution to a medical classification problem and used only when its advantages outweigh its drawbacks.

External information and the opinion of clinicians about the characteristics of the decision support tool can provide useful ways of simplifying a neural network without compromising severely its predictive performance. Unfamiliarity of the majority of clinicians with this type of classifier makes the application of this type of model problematic; this does not necessarily mean that neural networks should be rejected without further consideration.

3.5 Model selection and validation

Choosing the best model for a specific study is not an easy task as there are objectives, sometimes conflicting, for which we need to find the optimum solution. For example, it is possible to attempt to find the data structure that explains the variability between observations and also separate individuals into groups with the smallest number of errors. In this case, a particular model might be good for classification but not acceptable for capturing variable interrelationships. Hence the problem is finding a way to compare different models and decide which is most suitable for our purpose.

As we have previously seen, different types of classification models are based on different fitting algorithms that optimise different functions. For example, logistic regression attempts to find maximum likelihood for a given dataset whereas classification tree models are based on discovering the best way to separate subjects into classes by their attributes. Thus, we need to measure the performance of candidate models in terms of our selection criterion; obviously, the model that is built using a fitting algorithm closely related to this criterion is more likely to be the best model.

An aspect that is sometimes disregarded is the way of choosing the appropriate selection criterion for finding the best model. In some cases, it is not clear how this criterion should be chosen as different factors affect this choice. Clearly defined objectives could guide us to the description of the optimum model and how different model selection methods can find the models that fit this description.

On the other hand, it is difficult to find available implementations of the "optimum selection method" in all cases. Specifically, if the several objectives in our study that can be roughly summarised by a scoring function, then the ideal model selection method would be to the one that finds the model with the best possible score. Some statistical packages do not provide the user the ability to implement optimisation procedures for user-defined functions not available in the program.

It might be said that the solution to the problem of not having a unique model that is the best for all objectives is to have different models that could be implemented depending on the aspect of the problem. In this case, the cost of constructing and implementing different models can be a possible obstacle. In addition to that, if different variables are needed for each model, this might imply additional cost of collecting more data that will not be used in every single model.

Hence, model selection should be expressed, whenever it is possible, in terms of existing criteria that can implemented by readily available methods or by methods that can be constructed by the user whenever the programming environment allows that. Moreover, we should remember

that additional external information can be useful for including or excluding specific models, thus reducing the number of models examined when attempting to find the best model.

For instance, in Altman (1999, page 340), the best model is related to "the ability of the model to predict the dependent variable or, equivalently to explain variation in that variable". Also it is mentioned that "some subjective assessment may be necessary, especially when different approaches yield different answers".

The variables included are usually chosen on their statistical significance but there are other reasons for keeping them in the model. To be more specific, "sometimes it is desirable to keep a variable in a model because past experience shows that it is important" (Altman 1999, page 340). On the other hand, the same author mentions that "if the aim is to identify important predictor variables then it makes sense to omit variables that do not contribute much to the model".

3.5.1 Regression models

Stepwise method is a model selection strategy that is available in almost all statistical packages. It is based on finding the best subset of variables by examining a group of candidate variables and subsequently adding or removing one of these according to a goodness of fit criterion. In this way, the number of models tested is reduced substantially; this could be crucial when the number of variables and observations is large.

For example, in regression model analysis there are usually three variations of stepwise selection. The first, forward selection starts with the predictor that has the strongest association with the response. After that, from the remaining covariates, the variable that explains the largest amount of the remaining variability is entered. This procedure stops until the remaining variables are not statistically significant when entered in the model.

The second variation is backward selection where initially all variables are included and the non-significant ones are removed from the model until only the predictors that explain significant amount of variability are included. The third variation is a combination of forward and backward selection; see Venables and Ripley (1997, page 220) for further details.

Another selection strategy is examining all possible models. As additional variables means also additional complexity of the model, it is not possible to compare different models by using only goodness of fit as criterion. Hence, we need a selection function that is optimised when the maximum explained variability is achieved by the model with the least number of parameters. Such criteria are Mallow's C_p statistic and also information related criteria such as AIC and BIC.

A possible drawback of stepwise strategy is the large number of tests conducted at each step

when a large number of candidate variables are included. This is related to finding significant results by chance; thus the number of variables tested should be restricted to the ones that are absolutely necessary. In situations where there is no relevant information about choosing the important variables, exploratory analysis can be very useful for this purpose.

Further details about poor predictive ability are given in Steyerberg et al. (2000) and Wang et al. (2004). Steyerberg et al. (2000) found that using 5% significance level leads to poor model performance when independent data are used to evaluate the model. Wang et al. (2004) show that stepwise regression ignores the influence of variables not selected by the procedure and the uncertainty due to the variable selection process.

On the other hand, stepwise methods are useful for the situations where are highly correlated predictors are included in the data. In this case, one of the correlated variables is removed as it is not likely to have significant contribution to the model given that another variable plays that role. In other words, it can be said that stepwise strategy removes the redundant variables from the model.

3.5.2 Classification models

For classification models, which are the main focus of this study, goodness of fit is desirable but more important is the ability of separating the objects in the study into classes by using their attributes. Several criteria can be used to evaluate the performance of classifiers; the most common of these criteria is misclassification error rate.

In Ripley (1996, page 18), a loss function is described for assigning costs to each misclassification of a classifier. According to the same author, if every misclassification is equally serious, then we can define this loss function to have value equal to zero for each correct decision and equal to 1 for each case of wrong decision. In this case, the total loss will be equal to the number of wrong decision; misclassification error rate is simply the number of wrong decisions divided by the total number of decisions taken by using the classification model.

A generalisation of misclassification error rate is derived by assigning different costs to each misclassification. For example, consider a medical discriminating rule applied to separate carriers of disease from non-carriers can be evaluated by counting the number of wrong decision. Usually it is more important to identify disease carriers, hence the cost of identifying an individual with the disease as normal (false negative) can be k times more than assigning a non-carrier as carrier (false positive).

In this case, the total cost is equal to the sum of the false positive decisions plus k times the

number of false negative classifications. In the same way, it is possible to estimate the expected loss by using the prior probabilities for each class and the costs of each misclassification. The optimum classifier is the one that has the minimum expected total cost.

There are two decision rules that are important when assessing the performance of a classification model. The first is *no-data* or *default rule*: we classify all objects in the dataset as members of the class with the highest prior probability; in this case, the attributes of the objects are ignored (Michie et al. 1994, page 13).

The second is *Bayes rule*, where decisions about classification of objects are based on the principle of minimising the expected cost of misclassification given the knowledge of the attributes. In the special case that misclassification costs are equal, Bayes rule is equivalent to allocating objects into the class with the greatest posterior probability.

Defining misclassification costs is usually based on information from subject experts or related sources in the literature. In many cases, there are disagreements between experts about misclassification costs; in Breiman et al. (1984, page 176), an example is presented where the ratio of false negative misclassification cost divided by the corresponding false positive has been modified after consulting with physicians.

Furthermore, Michie et al. (1994, page 14) mention that "even in situations where it is clear that there are very great inequalities in the sizes of the possible penalties or rewards for making the wrong or right decision, it is often very difficult to quantify them. Typically they may vary from individual to individual".

In the case of having a sample where a class has prior probability much higher than the corresponding probability of the remaining classing, classification models based on this sample have usually small error rate for the dominant class and large error for the other classes. Breiman et al. (1984, page 112) propose adjusting the priors to alleviate this problem; for example we can take equal priors or put different weights on each individual in the sample according to misclassification costs.

Another way to assess the performance of a classification model with two classes is to use sensitivity (defined as true positive rate for an event) and specificity (defined as true negative rate for the same event). If a cut-off point for allocating individuals into classes given their attributes is not clearly defined, it is possible to examine the performance of a classifier by the area under the receiver operating characteristic (ROC) curve.

To be more specific, we "plot sensitivity versus 1-specificity for each possible for each possible cut-off, and ...join the points" (Altman 1999, page 418). The area varies from 0.5 for a useless

discrimination model to 1 for an ideal model. The same author mentions that if the cost of false positive and false negative are equal, then the best cut-off point is the one that maximises the sum of the sensitivity and specificity.

Other methods take into account how well calibrated is a model according to the predicted probabilities; in this case, we compare the fraction of events p with predicted probabilities p. Brier and logarithmic score can be applied to measure exactly how well calibrated the classifiers are. These scores are called well-calibrated as they are maximised when the proportion is equal to p for all events assigned to occur with probability p. The Brier and logarithmic score are proper scoring rules as the expected penalty is minimised by stating one's true beliefs.

As classification is implemented by the attributes that are available in the sample, it is possible that there is an overlap between the two classes. In other words, there is a limit of separability as the information about class membership is constrained by the relevance of the attributes to the class variable.

The minimum possible error rate can be found by choosing the class with highest posterior probability; this way of classification is the optimal rule and it is referred in the literature as *Bayes rule*. The risk associated to this rule is called the Bayes risk and it can be used as a benchmark for other procedures. A possible way of estimating the lower bound for the Bayes risk is described in Ripley (1996, page 197).

3.6 Statistical methods for selective screening

Several models have previously been used to separate high probability from low probability cases in an optimal way. The most common type of model in the literature is the logistic regression model (Cole et al. 1996).

Other useful approaches include linear discriminants, classification and regression trees, neural networks, genetic algorithms, graphical models and combinations of different models (Michie et al. 1994). For all these methods, there are several aspects that should be investigated before deciding which model is the best for our purpose.

Ripley 1996 (page 58) points out that "it is quite common in medical diagnosis for the abundance of the classes in the training set not to reflect their importance in the problem". In this case a suggested way to deal with the associated biases is to use weighted classification.

Furthermore, the misclassification costs are closely related to the importance of identifying high risk cases and should be defined according the information given by subject specialists. In the medical area these specialists might be clinicians related to the disease under investigation or

health managers that are responsible for the implementation of the screening program.

Finally, the assessment of the classifiers constructed should be based, when it is possible, on the objective that they have been built for. In some cases, different methods favour different classifiers. For example, a model that has been constructed by maximum likelihood methods is not always the best for classification. Details about this phenomenon are given in Friedman et al. (1997).

3.6.1 Weighted classification in medical screening

A critical aspect of classification methods in medicine is the ability to identify important cases that are rare in the population from which the sample under consideration has been extracted. Ripley (1996) states that "Often, when the training data are a random sample from the population, the vast majority of cases are 'normals' yet the cost of misclassifying a diseased case as normal is l times higher than that of a false positive. In screening problems l can be ten or more".

Cost of misclassification is usually defined by the person responsible for conducting the study. It is also possible that when the data is analysed retrospectively, other types of cost might be found to be important. For example, cost can be initially defined in economic terms and later or the social or psychological cost of misclassification need to be taken into consideration.

To be more specific, for a data set with two classes, 'diseased' d and 'normal' n, we have to estimate the corresponding posterior probabilities $p(\mathsf{d}|x)$ and $p(\mathsf{n}|x)$ from the training data. The decision rule declares a case 'diseased' if $p(\mathsf{d}|x;\widehat{\theta}) > c$ for $c = 1/(1+l)$ less than 0.5, often much less. Since there are many 'normal' cases, the estimated posterior probabilities will be biased. There are two ways to overcome these biases.

The first is to subsample the 'normal' group in a random way. We denote the number of cases in class k by n_{k} and the proportions in the population by π_{k}. By using the fitted probabilities $p(\mathsf{k}|x;\widehat{\theta})$, we estimate quantities proportional to $p(\mathsf{k}|x)n_{\mathsf{k}}/\pi_{\mathsf{k}}$, the posterior probabilities under biased sampling. In this way, a case is declared 'diseased' if

$$\frac{p(\mathsf{d}|x;\widehat{\theta})\pi_{\mathsf{d}}/n_{\mathsf{d}}}{p(\mathsf{n}|x;\widehat{\theta})\pi_{\mathsf{n}}/n_{\mathsf{n}}} > \frac{1}{l} \Rightarrow \frac{p(\mathsf{d}|x;\widehat{\theta})}{p(\mathsf{n}|x;\widehat{\theta})} > \frac{1}{l\pi_{\mathsf{d}}n_{\mathsf{n}}/\pi_{\mathsf{n}}n_{\mathsf{d}}}$$

As $p(\mathsf{n}|x;\widehat{\theta}) = 1 - p(\mathsf{d}|x;\widehat{\theta})$ we have:

$$\frac{p(\mathsf{d}|x;\widehat{\theta})}{1-p(\mathsf{d}|x;\widehat{\theta})} > \frac{1}{l\pi_{\mathsf{d}}n_{\mathsf{n}}/\pi_{\mathsf{n}}n_{\mathsf{d}}} \Rightarrow \ldots \Rightarrow p(\mathsf{d}|x;\widehat{\theta}) > 1/(1+l\pi_{\mathsf{d}}n_{\mathsf{n}}/\pi_{\mathsf{n}}n_{\mathsf{d}})$$

Thus, by under-sampling the 'normal' group by a factor of l, this declares 'diseased' if the odds exceed one.

Another way of overcoming the problem of the biased training data set is by using weighting of cases. If 'normal' cases are weighted by a factor ω, where $\omega \approx 1/l$, the estimation biases will be minimised as the decision rule declares 'diseased' if $p(\mathsf{d}|x;\widehat{\theta}) > 1/(1+l\omega)$. The way that will be used depends on the sample size; sampling is preferred when the sample size is considerably large as the reduction in the size of the training set can have considerable computational benefits.

3.6.2 Misclassification costs

From the methods described above, it is obvious that the choice of l, the ratio of false negative misclassification cost divided by the corresponding false positive, is very important. In some cases, experts disagree about the estimation of this ratio (Breiman et al. 1984).

In problems related to the medical area, two ways of deriving misclassification costs are used. The first way is consultation with physicians or clinicians by using utility elicitation methods. The second way is the estimation the average misclassification costs by using prior information and the training data set.

On the one hand, the medical experts are asked to assess the severity of misclassifying a 'diseased' case as 'normal' and also to estimate the cost of the resources that would be needed to treat this case. In the same way, the cost of misclassifying a 'normal' case as 'diseased' should also be estimated.

It must be remembered that there are several factors affecting an expert's opinion about misclassification costs. A possible way of eliciting less biased assessments is combining opinions from different experts related to the specific problem.

On the other hand, it is possible by using prior information from previous studies to estimate the risk for a misclassified case and the cost related to it. If risks and costs depend on the predictors, then the average misclassification cost should be calculated. To be more precise, by using the notation in Ripley 1996 page 18, we have the misclassification probabilities:

$$\text{pmc}(\mathsf{k}) = Pr\{\widehat{c}(X) \neq \mathsf{k}, \widehat{c}(X) \in \{1,...,K\} | C = \mathsf{k}\}$$

where $\widehat{c} : X \to \{1,...,K\}$ is the classifier. The total risk is the total expected loss, viewing both the class C and the vector X as random:

$$R(\widehat{c}) = \sum_{\mathsf{k}=1}^{K} \pi_{\mathsf{k}} \text{pmc}(\mathsf{k})$$

3.6.3 Optimum classification

As we have previously seen, there are different misclassification costs for false positive and false negative cases and the overall misclassification cost can be used as a criterion for comparing different methods.

Breiman et al. 1984 (page 178) point out that by using the training set for this comparison, the estimates are somewhat optimistic and that the cross-validated misclassification cost should be used. The same principle must be applied for other criteria used and in some cases there is enormous difference between cross-validation and resubstitution (using the same data set for building a classifier and for assessing its performance).

In Hand (1997), the use of sensitivity Se (true positive rate) and specificity Sp (true negative rate) is analysed as a common strategy in epidemiological studies. Furthermore, in order to assess the performance of a classifier for a range of possible thresholds, the receiver operating characteristic (ROC) curve can be used.

ROC curves are a plot of sensitivity against (1-specificity) and in this case the overall performance is measured by the area under the ROC curve. If the cost of misclassification of class 0 is c_0 and the corresponding for class 1 is c_1, then the total cost is $\pi_0 c_0(1-\mathsf{Se}) + \pi_1 c_1(1-\mathsf{Sp})$. At the optimal threshold, where the cost is minimum, the slope of the ROC curve must be $s = c_1\pi_1/c_0\pi_0$. This notion can be used to identify 'good' threshold values.

Another way of comparing classifiers is the odds ratio defined as the ratio of the odds of being classified into class 0 given that the subject actually comes from class 0 and class 1 respectively and is equal to $\mathsf{Se} \times \mathsf{Sp}/(1-\mathsf{Se})(1-\mathsf{Sp})$. If class 0 has a very small prior, then the odds ratio is approximately equal to (positive predicted value)/(1-negative predicted value) (Hand 1997).

In comparing methods, there are classification methods that seem to be 'good' by using one criterion and 'bad' by another criterion. For these situations, the reason for building a classifier will determine the criterion that should be used.

For medical screening, low prevalence (low prior probability of a disease in the population) can produce a low positive predicted value (denoted by pp) even if the sensitivity and specificity are high. If, for example, π_0=0.01, Se=Sp=0.9, then pp=0.09.

3.7 Further topics

One particular problem of deriving estimates of classification accuracy from the same sample that has being used for constructing the classifier is that the results tend to be biased. In Michie et

al. (1994, page 107), the authors give an example of growing a very complex decision tree with 100% accuracy on the training dataset. They state that "in practice, complex structures do not always perform well when tested on unseen data, and this is one case of the general phenomenon of over-fitting data". Three possible ways of correcting the bias are using an independent test set, cross-validation and bootstrap.

The first option, which is using an independent test set for accuracy estimation, is the ideal option if the sample size is large and we can train a classifier without a part of the sample. Michie et al. (1994, page 108) adopted this strategy for samples of size 1000 or more. A particular aspect of this method is the possible loss of efficiency as the training set is reduced. Additionally, the results of the test set (usually 20-30% of the original data selected at random) may vary according to the way that the sample was split into training and test set.

Cross-validation gives an answer to the problem of testing the classifier only on a part of the initial dataset by dividing the sample into m subsamples. Each of these is used as test set for the classifier that is built on the remaining $m-1$ subsamples; this estimated classification performance is the average performance of these parts of the data. The results in this case are unbiased; in order to reducing the variability of these averages because of choosing a particular random split, we repeat the procedure a large number of times and we use as estimate the average over different splits.

In Ripley (1996, page 71) it is mentioned that cross-validation can be also used for model selection. Specifically, it can be used for estimating classification performance or goodness of fit. In addition to that, cross-validation procedure can be used to estimate parameters of the model. The same author states that model choice by NIC (Network Information Criterion) and leave-one-out cross-validation, in which a sample of size m is split into m subsamples, are asymptotically equivalent.

Another way to obtain accuracy estimates such as the misclassification error rate with less variability than cross-validation is to use bootstrap method. In this case, instead of using the original sample, we make a new sample of the same size as the original by resampling with replacement. In this case, approximately $1/e = 37\%$ is omitted from each sample and this part can be used as the test set for assessing classification performance.

Michie et al. (1994, page 108) state that "the penalty paid is that the estimated error-rate are optimistic (i.e. are biased downwards). The trade-off between bias and random error means that, as a general rule, the bootstrap method is preferred when the sample size is small, and cross-validation when the sample size is large". Furthermore, the same authors mention that "the

bootstrap and cross-validation estimates are generally close for large sample sizes".

Instead of using only the classification model with the best performance, we might combine different classifiers in order to improve this performance. In Ripley (1996, page 65), it is mentioned that "this would amount to combining the posterior probabilities (either plug-in or predictive) from a series of M models". Moreover, this can be done by using a weighted average of the predicted probabilities and these weights can be chosen by cross-validation.

In addition to that, in the complementary part of Ripley (1996, page 1), *bagging* (boostrap aggregating) is proposed as a possible way of combining classifiers. In this case, the idea is "to take an unweighted average of the predictions of, say 100, classifiers trained on training sets formed by resampling with replacement from the original training set". The same author mentions that motivation for this procedure comes from *unstable* classifiers "such as classification trees where a small change in the training set can lead to a large change in the classifier".

Boosting is a method that is based on the idea of combining classifiers by training them on sequentially formed learning datasets. To be more specific, the weights "of the examples which were classified incorrectly is increased relative to those which were classified correctly. For classifiers that do not accept weights on the examples, resampling with probabilities proportional to the weights can be used" (Ripley, 1996, complements page 2).

Boosting has been found to have better performance than bagging but it sometimes leads to worse performance than the original classifier. In addition to that, boosting has been proposed to minimise training error rate and not generalisation error (error for an independent sample). Hence, problems related to overlapping classes cannot be solved in general by boosting performance "to fit the training set perfectly" (Ripley, 1996, complements page 2).

When dealing with missing data for classification problems, there are models such as classification trees that assign observations into classes even when some of the attributes are not available. In Venables and Ripley (1997, page 419) and Breiman et al. (1984, page 142) possible ways of constructing classification trees using missing data and subsequently discriminating cases into groups from some missing attributes are given.

It is not always possible to construct a model with missing data; we might have in our disposal models that can only deal with complete observations. In addition to that, Breiman et al. (1984, page 146) warn that if the missingness does not occur at random, the results from algorithms that can be constructed and used for classification with missing data can be misleading.

Great care should be taken when removing cases with incomplete data from the study; sometimes this is done automatically by the software we are using without any warning about it. In

this case, it is possible that "the incompletely observed cases differ systematically from the completely observed cases. The completely observed cases that remain will be unrepresentative of the population for which the inference is usually intended" (Schafer 1997, pages 1-2).

Imputation, which can be described as the procedure of replacing missing data with plausible values, can be implemented when several types of statistical analysis are needed. The results of imputation will depend on the assumptions made by the imputation model should be sensible otherwise the imputed values will introduce bias to the results.

Specifically, "imputing averages on a variable-by-variable basis preserves the observed sample means, but it distorts the covariance structure, biasing estimated variances and covariances toward zero. Imputing predicted values from regression models, on the other hand, tends to inflate observed correlations, biasing them toward zero" (Schafer 1997, page 2).

The author mentioned above proposes multiple imputation by using the EM algorithm. This method attempts to replace each missing value by $m > 1$ simulated values. These values reflect uncertainty about the true values of the data and the result is to have m complete datasets that can be analysed by standard methods separately. Subsequently, the results are combined to give the result of interest (Schafer 1997, page 5).

Most of the statistical literature about multiple imputation methods is related to construction of procedures about different types of data and different types of inferences. Furthermore, it is usually assumed that a specific model for multiple imputation can be implemented when data are missing at random; the imputation model is usually defined by exploratory analysis of the data and it is also similar to the one that will be used in further statistical analysis.

In some cases, it might be necessary to explore the multiply imputed complete dataset in order to acquire further knowledge about their characteristics and decide about further analysis based on these results. In addition to that, graphical models or other methods of exploring the association or influence structure of each of the imputed datasets could reveal important relationships between variables in each of these datasets.

Furthermore, model selection using different criteria can be expanded into imputed datasets; after that, combining the results from different datasets will not be straightforward procedure as the selected models might be different. In addition to that, sensitivity analysis can be implemented if different imputation models seem sensible.

As multiple imputation is a method not widely used by experts in the medical area, the results from multiply imputed datasets should be presented in a way that these experts are familiar with and also the advantages of multiple imputation over other ad hoc methods of imputation (if they

exist) should be explained.

After all, many of the statistical analysis are intended for informing clinicians and other people that might not have the statistical background and knowledge of methods such as multiple imputation. It is not unusual that many statistical analyses seem unnecessarily complicated and are not used in practice. Hence, the data analyst should apply statistical methods that are appropriate for both the study and the individuals that will subsequently use it.

3.8 Use of clinical data sets

The data we use to construct the selective screening tool might not help us to discriminate successfully between low and high risk cases because it does not describe a significant proportion of the variability between these cases. In other words, we may have significant overlap between the two classes we wish to separate; thus, any classifier we apply will not be effective for the purpose for which it has been constructed.

In the situation described above, it is very useful to know the limits of separability, set by the information available. In case this information does not satisfy our objectives, e.g. a certain level of misclassification error rate, we should try to use additional information to improve the performance of the classification method we use.

In this section, a possible way of estimating the limits of separability by calculating the lower bound of the Bayes risk is explored. Methods of enhancing the use of the available information and introducing additional information, using as an example the data collected for an AAA screening program will also be described.

3.8.1 Bayes risk

The total risk, which is the total expected loss viewing both the class C and the vector X as random is defined as:

$$R(\widehat{c}) = \sum_{k=1}^{K} \pi_k \mathrm{pmc}(k)$$

The posterior probability of class k given $X = x$ is $p(k|x)$ where

$$p(k|x) = Pr\{C = k|X = x\} = \frac{\pi_k p_k(x)}{\sum_{l=1}^{K} \pi_l p_l(x)}$$

The optimal rule is to choose the class with highest $\pi_k p_k(x)$; this optimal classifier is also referred to as the Bayes rule. The value $R(c)$ of the total risk is called the Bayes risk; this value is the best

one can achieve if the π_k's and p_k's are known and provides a benchmark for all other procedures.

A possible way of estimating a lower bound for the Bayes risk is described by Ripley (1996). By running a 3-nn classifier on the training set and reporting $1/3$ the number of occasions on which the neighbours are two of one class and one of another, we have an approximation of the Bayes risk. The predictors should be scaled to have approximately equal ranges.

The justification of the use of the method describe above is given in Ripley (1996). To be more specific, the author states and proves the following (p.196-197):

"Proposition 6.3.: In the large-sample theory the means of the risk-averaged (3,2)-nn rule and the error rate of the (2,2)-nn rule are equal and provide a lower bound for the Bayes risk. The risk-averaged estimator has smaller variance".

In page 197, he states that based on proposition 6.3, this suggests estimating a lower bound for the Bayes risk in the way described in page 99 of the thesis. An example of this method is shown in Ripley (1996) p. 201 using data related to Pima Indians diabetes.

Before attempting to estimate the lower bound for the Bayes risk for our data, we need to identify which factors have been identified as important in previous studies. Furthermore, we need to identify the factors in our study that we likely to have under different data collection schemes.

Also, we should point out that different studies for the same disease may have different protocols for collecting data and implementing screening methods. Hence, the results of one study may not be applicable to another study.

3.9 Flexible selective screening

An additional aspect of great importance for selective screening is the application of the classification tool constructed to populations with different characteristics from the one whose sample was used for the construction of the classifier. The individuals in the sample might come from a specific area that has high prevalence of the targeted disease and high proportion of these cases might have other related co-morbidities present.

In this case, by applying the selective screening tool based on a specific sample to a target population with significantly lower prevalence of the disease and rare presence of related co-morbidities, many relatively high risk cases in this target population could be missed. An example of this is diabetes in different ethnic groups. For Caucasian population, the prevalence of diabetes is about 2% whereas for Asians and Afro-caribbeans the corresponding prevalence is about 15%.

It is apparent that when assessing the performance of the discrimination method, a test set that has not been used for the construction of this method is useful to obtain unbiased estimates of

the classifier's effectiveness. As mentioned previously, cross-validation and bootstrapping might be helpful as they lead to less biased results. Another potential problem of assessing a classification method's performance is confounding. In this case, there might be factors not recorded in the data that affect the occurrence of the event of interest.

For AAA selective screening as well as other screening programmes, genetic screening might provide additional information for the identification of possible causes of the disease. Thus, if this information is not available, then the selective screening tool may prove successful in one country but fail in another simply because of the different genetic structure of the populations screened. Another example of incomplete data that we might have to deal with arises from collecting and recording self-reported information for events such as smoking.

Hence, the need for flexibility when using selective screening becomes obvious. For the time being, genetic screening is not an option for massive screening programmes; this might take a decade or more to become common practice even in sophisticated health systems around the world. On the other hand, current reliable information should be used when constructing or modifying the selective screening tool.

Another aspect that shows the need for flexibility is the introduction of elective screening where the individuals decide whether they are screened or not after they are informed about the benefits and risks of the screening test. In this case, the classifier that will be used for selecting depends on the attitude of each case when giving information about themselves that will be subsequently used for screening. To put it in another way, each member of the population should be seen as a separate unit for which we may or may not have the information required for screening.

This does not mean that we need to have 'personal selective screening tools' for each member of the population; it means that the classifiers used should be flexible to deal with incomplete data and allow the possibility of inconclusive results. Elective screening could be seen as the classifier which can be labelled "patient choice screening" and might be a matter for further research.

3.10 Statistical analysis and results

3.10.1 The area under the ROC curve as criterion

There are several criteria that can be used to compare different classification models. In order to decide which is suitable, we must be guided by the objective for which the classifiers have been constructed. In our case, the primary objective is to identify the cases with the highest probability of abdominal aortic aneurysm. These individuals might subsequently be selected for a screening

program.

Several parameters of the problem are not strictly defined; for example there are no specific figures for sensitivity, specificity and proportion of cases to be screened. Thus, we need a criterion that will compare the models under consideration over a wide range of values for the parameters of interest; the area under the ROC curve is the most familiar criterion that covers these requirements.

Different model types by using various combinations of predictors have been compared by using the criterion mentioned above. To be more specific, logistic regression models, classification trees, neural networks, k-NN classifiers and boosting models have been assessed. For each model, the 95% confidence interval of the area under the ROC curve has been estimated by using a set of functions for the statistical package S-Plus, provided by Elizabeth J. Atkinson of the Mayo Clinic (Janisse, 1997).

Additionally, by using the same set of functions, it is possible to test the statistical significance of the difference between "two or more empirical curves that are constructed based on tests performed on the same individuals" (see DeLong et al. 1988). Further, it is argued that in this case "statistical analysis on differences between curves must take into account the correlated nature of the data". In this case, we use the complete dataset to build and test the models, thus the results are biased as they are overestimating the area under the ROC curve.

In order to get unbiased results about the estimated areas under the ROC curve, we used 3-fold cross-validation; in this case, the dataset is split at random into the learning set that contains two thirds of the original dataset and the remaining cases are used as the test set. The classifier is built on the learning set and the area under the ROC curve is estimated by using the predicted risk scores for the test set.

This procedure is repeated 1000 times for each model to assess the variance of the estimated area under the ROC curve. Hence, for each type of model and combination of predictors, we obtain a 95% confidence interval for the area under the ROC curve. In a similar way, we used bootstrapping combined with 3-fold cross-validation for the same purpose.

For the logistic regression model when the age groups categorical variable (less than 65, 65-69, 70-74 and at least 75 years), smoking level and co-morbidities and family history indicator are used, the mean area for the complete dataset is 0.682 and the standard deviation is 0.017, hence the 95% confidence interval is (0.649, 0.716). On the other hand, by using 3-fold cross-validation, the corresponding confidence interval is (0.651, 0.676). By applying bootstrapping 3-fold cross-validation, we obtain the interval (0.637, 0.676).

In the same way, we estimate the 95% interval confidence interval for the area under the ROC

curve using a classification tree model including age groups, smoking level and co-morbidities and family indicator. The result for the complete dataset is (0.666, 0.730) and not significantly different from the corresponding logistic regression model (compared by using the method in DeLong et al. 1988). On the other hand, the 95% confidence interval for the same model estimated by using is (0.629, 0.668) and is significantly different from the 3-fold cross-validation estimate for the logistic model (by using paired t-test).

Using the complete dataset, a neural network model has been used with the same predictors as the classification tree. The results are virtually the same as the classification tree model. Specifically the 95% confidence for the area under the ROC curve is (0.667, 0.731). Similar results were obtained for k-NN classifiers. The 95% intervals of the area under the ROC curve for 1-NN and 3-NN classifiers are respectively (0.667, 0.731) and (0.666, 0.730).

Ensembles of different models are combined by calculating the weighted sums of the predicted scores from each model. The weights are computed by using maximisation procedure for the area under the ROC curve. For the complete dataset, the area under the ROC curve for the combination of predicted scores from the logistic and the classification tree is 0.696. The corresponding area when the neural network and the 1-NN scores are combined with logistic and tree model scores is 0.698. Finally, the S-Plus library MART that implements boosting models gives an area under the ROC curve equal to 0.653. A possible reason for obtaining these results is the overlap between classes when the predictors mentioned above are included.

When diastolic blood pressure measurements are included in the model, the results are slightly improved but again they are not significantly different. For example, by flagging diastolic blood pressure above 90 mmHg with 95% confidence interval for the estimated odds ratio between 1.264 and 2.112, the estimated interval for the area under the ROC curve is (0.662, 0.728) for the complete dataset and (0.664, 0.689) for the 3-fold cross-validation method.

The results, compared by paired t-test are significantly different from the corresponding 3-fold cross-validation results without bpd. It also possible to have intervals for diastolic blood pressure defined by sensible criteria instead of a single split (Blanchard et al. 2000).

Diastolic blood pressure has been indicated as an important risk factor for abdominal aortic aneurysm but it may not be available, we therefore propose the use of two models: one with and one without diastolic blood pressure measurement. In this way, an important predictor will be included when it is available but it is not necessary when selecting the cases for screening. This is related to the need for flexible modelling according to the information available.

We now present the two proposed logistic regression models for screening in detail. The logit

of the logistic model has the following general form:

$g(x) = logit(p) = log(p/(1-p)) = b_0 + b_1 \times x_1 + + b_n \times x_n$,

where p is the probability of event occurrence, b_0 is the constant, b_i, i=1,..n the coefficients of the predictors and x_i, i=1,..,n the predictors.

To obtain the probability of the event, we use the formula:

$p = \exp(g(x))/(1 + \exp(g(x))$

The first logistic model proposed for aortic aneurysm screening contains as predictors indicators for the age groups, smoking level and co-morbidities and family history indicator. The list of indicators is shown below:

- age_0: at most 64 years old (reference category)
- age_1: between 65 and 69 years old
- age_2: between 70 and 74 years old
- age_3: at least 75 years old
- $smol_0$: never smoked (reference category)
- $smol_1$: moderate smoker
- $smol_2$: heavy smoker
- $smol_3$: inaccurate information
- $smol_4$: poor information
- com_0: no co-morbidities (reference category)
- com_1: at least one of the co-morbidities

The logit of the model with the indicators described above has the following form:

$logit(p) = -3.384 - 0.205agegr_1 + 0.057agegr_2 + 0.777agegr_3 + 0.556smol_1 + 0.931smol_2 - 0.487smol_3 - 0.015smol_4 + 0.897com_1$

The risk estimates for having abdominal aortic aneurysm obtained by the logistic model with the age groups, smoking level and co-morbidities and family history indicator are shown in table 3.2 on page 83:

Risk factors	Odds ratio (95% Conf. Inter.)	p-value
Age		
<=64 (refer.)	1.000	
65-69	0.815 (0.284, 2.336)	0.703
70-74	1.058 (0.370, 3.030)	0.916
>=75	2.175 (0.710, 6.664)	0.174
Smoking		
Never (refer.)	1.000	
Moderate	1.743 (0.921, 3.300)	0.088
Heavy	2.536 (1.381, 4.655)	0.003
Inaccurate	0.615 (0.260, 1.452)	0.267
Poor inform.	0.985 (0.266, 3.647)	0.982
Co-morbidities		
No (refer.)	1.000	
At least one	2.451 (1.880, 3.200)	<0.001

Table 3.2: First proposed logistic model

From table 3.2 on page 83 of the risk estimates, we can see that adjusting for smoking level and co-morbidities, the risk of having AAA is not significantly different if we compare the reference group (at most 64 years old) and the other categories at 5% significance level. Heavy smoking is significantly associated with the presence of AAA and the risk of a heavy smoker compare to a person that never smoked is increased by approximately 150%. Moderate smoking habit does not seem to have a significant effect on the occurrence of AAA compared to never being a smoker.

Finally, the presence of at least one of the co-morbidities increases the risk of having AAA by 145% even after controlling for age and smoking level. From the results shown above, it is clear that heavy smoking and the presence of related diseases or family history of AAA are both strong predictors of the presence of AAA.

The second model proposed for aortic aneurysm screening includes the age groups, smoking level and co-morbidities and family history indicator as before with the addition of a dummy variable for diastolic blood pressure at least 90 mmHg. The logit of the logistic model in this case is:

$logit(p) = -3.542 - 0.200agegr_1 + 0.051agegr_2 + 0.763agegr_3 + 0.487smol_1 + 0.857smol_2 - 0.540smol_3 - 0.095smol_4 + 0.934com_1 + 0.507bpdgr_1,$

where

- age_0: at most 64 years old (reference category)
- age_1: between 65 and 69 years old

- age_2: between 70 and 74 years old
- age_3: at least 75 years old
- $smol_0$: never smoked (reference category)
- $smol_1$: moderate smoker
- $smol_2$: heavy smoker
- $smol_3$: inaccurate information
- $smol_4$: poor information
- com_0: no co-morbidities (reference category)
- com_1: at least one of the co-morbidities
- $bpdgr_0$: diastolic blood pressure less than 90 mmHg (reference category)
- $bpdgr_1$: diastolic blood pressure at least 90 mmHg

The risk estimates for the occurrence of abdominal aortic aneurysm for the age groups, smoking level, co-morbidities and family history indicator and the "diastolic blood pressure at least 90 mmHg" variable are shown in table 3.4 on page 85:

As with the previous logistic model without the diastolic blood pressure indicator (table 3.2 on page 83), we can see in table 3.4 on page 85 that heavy smoking and the presence of co-morbidities are significant predictors of abdominal aortic aneurysm occurrence at the 5% level. The binary indicator for diastolic blood pressure above 90 mmHg is also significant at 5% level and the risk of having AAA with raised diastolic blood pressure (at least 90 mmHg) is 66% when compared to the lower level of diastolic blood pressure.

Finally, when age and diastolic blood pressure are included as continuous variables, the 95% confidence interval for the area under the ROC curve is (0.665, 0.731) for the logistic and (0.837, 0.876) for the tree model (calculated for the complete dataset) and are significantly different. On the other hand, the 3-fold cross-validation intervals are (0.671, 0.693) (logistic) and (0.543, 0.609) (classification tree) and are again different at the 5% significance level.

The results above show that classification tree models are over-fitting the data; hence if continuous variables are included for age and diastolic blood pressure instead of the corresponding categorical, the logistic model should be used to select the cases to be screened.

Risk factors	Odds ratio (95% Conf. Int.)	p-value
Age		
<=64 (refer.)	1.000	
65-69	0.818 (0.284, 2.361)	0.711
70-74	1.053 (0.365, 3.033)	0.924
>=75	2.144 (0.695, 6.613)	0.184
Smoking		
Never (refer.)	1.000	
Moderate	1.627 (0.858, 3.087)	0.136
Heavy	2.357 (1.280, 4.336)	0.006
Inaccurate	0.582 (0.246, 1.379)	0.219
Poor inform.	0.909 (0.244, 3.384)	0.887
Co-morbidities		
No (refer.)	1.000	
At least one	2.544 (1.947, 3.324)	<0.001
Bpd>=90 mmHg		
No (reference)	1.000	
Yes	1.660 (1.271, 2.170)	<0.001

Table 3.4: Second proposed logistic model

3.10.2 Proposed selective screening model

The main purpose of constructing a selective screening model is to select the part of the population that is most at risk and in this way improve the cost-effectiveness of the screening program. The information required should be easily obtained and also the cost of collecting it should not be high. One of the best ways of achieving this is via self-administered questionnaires.

It must be remembered that these will be filled by the members of the target population. Thus the questionnaires should be simple to use and should also provide the information that is necessary for deciding about further action in terms of the screening scheme for each individual.

Based on the results in the previous sections, the simplest model that fulfils the requirements analysed above should include three predictors: age groups (less than 65, 65-69, 70-74 and at least 75 years), smoking level (never smoked, moderate and heavy smoker) and an indicator of at least one of the co-morbidities or family history of AAA.

Different types of models achieve the same results in terms of the area under the ROC curve; in this case we selected the logistic regression model that is widely used for similar purposes. The more complex models, such as tree models, neural networks and boosting models did not achieve better results for this dataset in terms of maximising the area under the ROC curve.

As the level of performance of the selective screening in terms of sensitivity, specificity and proportion screened is not strictly defined, we should use several cut-off points for the risk score and then decide which is the optimum given the requirements of the program. In table 3.5 on page 86, an example is given when the predictors mentioned above are used.

Several parameters are estimated for each cut-off point of the risk score; see appendix B for the definitions of these parameters. Specifically, the 95% confidence interval for sensitivity, specificity and proportion screened are calculated; also positive and negative predictive value and positive and negative likelihood ratio are estimated.

In table 3.6 on page 87, the results for the logistic model with age and bpd as continuous variables are given. It must be remembered that the results in tables table 3.5 on page 86 and table 3.6 on page 87 are biased as the complete dataset is used to construct the classifier and estimate the parameters; cross-validation estimates might be used to overcome this problem.

Cut-off point (probabilities)	0.03	0.05	0.07	0.09
lower limit (2.5%) for sensitivity	0.961	0.901	0.654	0.500
estimated sensitivity	0.968	0.912	0.671	0.518
upper limit (97.5%) for sensitivity	0.974	0.922	0.688	0.536
lower limit (2.5%) for specificity	0.125	0.277	0.569	0.731
estimated specificity	0.138	0.293	0.587	0.746
upper limit (97.5%) for specificity	0.150	0.310	0.605	0.762
lower limit (2.5%) for proportion screened	0.859	0.708	0.416	0.259
estimated proportion screened	0.871	0.724	0.434	0.276
upper limit (97.5%) for proportion screened	0.883	0.740	0.452	0.292
positive predictive value	0.092	0.105	0.128	0.156
negative predictive value	0.979	0.973	0.952	0.945
positive likelihood ratio	1.123	1.290	1.625	2.039
negative likelihood ratio	0.232	0.300	0.560	0.646

Table 3.5: Selective screening results

For the logistic model that includes the categorical variable for the age groups, smoking level and co-morbidities and family history indicator (table 3.2 on page 83), figure 3.1 on page 87 is showing the plot of sensitivity versus 1-specificity for the whole range of possible cut-off points.

In Hosmer and Lemeshow (2000), it is stated that the discrimination provided by the model implementation is acceptable if the the area under the ROC curve is between 0.7 and 0.8, excellent if the corresponding area is between 0.8 and 0.9 and outstanding for an area larger than 0.9. Using this criterion, the proposed classifier has discriminative performance which is likely to be acceptable with minor improvement. With the inclusion of "raised diastolic blood pressure" indicator (at

Cut-off point (probabilities)	0.03	0.05	0.07	0.09
lower limit (2.5%) for sensitivity	0.961	0.880	0.727	0.577
estimated sensitivity	0.968	0.892	0.743	0.594
upper limit (97.5%) for sensitivity	0.974	0.903	0.759	0.612
lower limit (2.5%) for specificity	0.133	0.304	0.519	0.668
estimated specificity	0.146	0.321	0.537	0.685
upper limit (97.5%) for specificity	0.159	0.338	0.555	0.701
lower limit (2.5%) for proportion screened	0.851	0.680	0.468	0.321
estimated proportion screened	0.864	0.697	0.486	0.339
upper limit (97.5%) for proportion screened	0.876	0.713	0.504	0.356
positive predictive value	0.093	0.106	0.127	0.146
negative predictive value	0.980	0.970	0.958	0.949
positive likelihood ratio	1.133	1.314	1.605	1.886
negative likelihood ratio	0.219	0.336	0.479	0.593

Table 3.6: Selective screening results

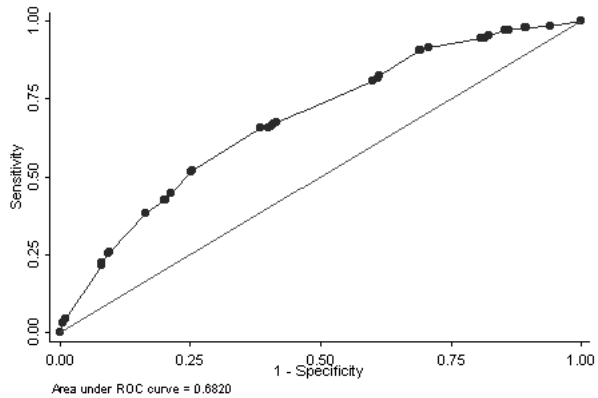


Figure 3.1: Area under the ROC curve for first proposed model

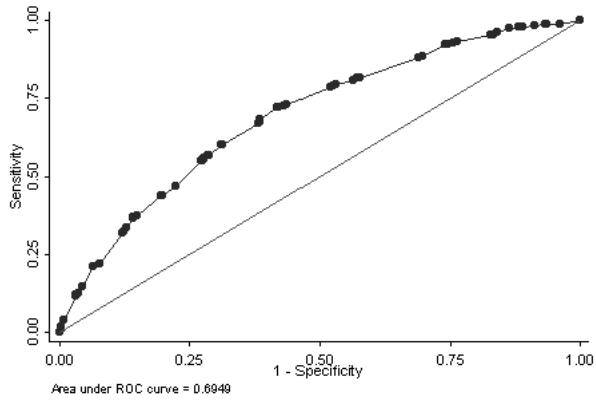


Figure 3.2: Area under the ROC curve for second proposed model

least 90 mmHg) (model shown in table 3.2 on page 83), the discrimination is slightly improved; see figure 3.2 on page 88.

Both graphs are useful when clinicians would choose a specific cut-off point for separating cases to be screened for abdominal aortic aneurysm based on the information available on the predictors. We might conclude that both the classifiers mentioned above have the potential of being used for screening but there is also need for further improvement by including additional predictors.

3.10.3 Age-adjusted abnormality thresholds

An alternative definition for the presence of abdominal aortic aneurysm is based on the fact that "the aorta expands throughout adult life" (Grimshaw et al. 1995). In this paper and also in Grimshaw et al. (1997), age-adjusted abnormality thresholds are calculated using Letter Value Analysis (Hoaglin et al. 1983).

To be specific, the upper letter value D, representing 6% of the population is used to define the abnormal cases; for example, for 60 years old male individuals, the threshold is 24 mm and for individuals at age 70, the corresponding threshold is 37 mm (see table 2 in Grimshaw et al. 1997).

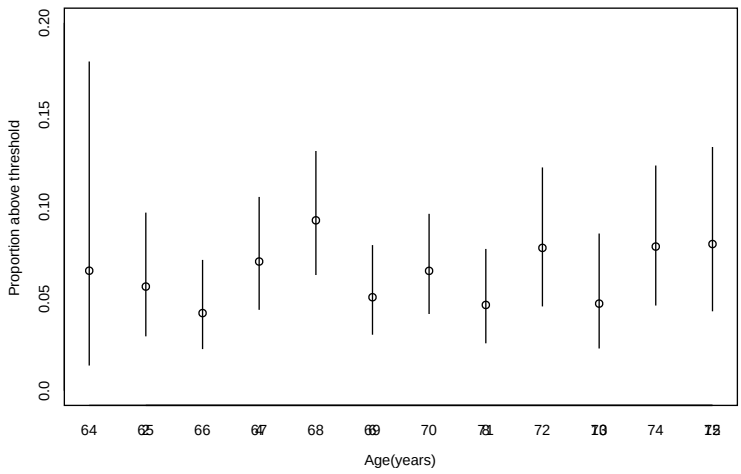
In the same paper, table 3 includes the age-adjusted abnormality thresholds that have been calculated "using robust regression from data points given for letter D in table 2 and illustrated by the line in figure 3"; see table 3.7 on page 89 and figure 3.3 on page 89. By using this criterion, there are 191 cases defined as abnormal.

For our data, χ^2-tests indicate that the most important predictors are age, diastolic blood pressure, smoking level and the presence of at least one of the co-morbidities or family history of AAA. Furthermore, odds ratio, relative risk and attributable risk are similar for each of the

Age (years)	Threshold (mm)	Number below threshold	Number above threshold
64	27.5	43	3
65	28.0	200	12
66	29.0	295	13
67	30.0	278	21
68	30.5	284	29
69	31.0	336	18
70	32.0	330	23
71	32.5	286	14
72	33.0	202	17
73	34	201	10
74	35	200	17
≥75	35.5	150	13

Few cases with age below 64 years not shown.

Table 3.7: Age-related abnormality thresholds and number of cases below and above threshold



Vertical lines indicate 95% confidence intervals for proportions.

Figure 3.3: Proportion of cases above threshold against age

important predictors.

In addition, the area under the ROC curve has been estimated for two models of selective screening: a logistic regression model including age groups, smoking level and co-morbidities and family history indicator, and the same model including age and diastolic blood pressure. The 95% confidence interval for the area under the ROC curve for the complete dataset model with age groups is (0.634, 0.711).

The corresponding interval for the model including age and diastolic blood pressure is (0.652, 0.730). The intervals are similar, suggesting that in this case, both models can be used for selective screening modelling of abdominal aortic aneurysm. As before, the use of cross-validated estimates of the area under the ROC curve will give less optimistic results.

The use of age-related abnormality thresholds is, as mentioned previously, a matter of controversy between researchers. Until more information is available about the importance of differentiating the thresholds of AAA abnormality according to age and well designed clinical trials assessing the results in terms of cost-effectiveness and reducing mortality are conducted, we suggest that it may prove difficult to adopt them for selecting individuals for screening.

On the other hand, there are clinicians (e.g. radiologists) that advocate the implementation of age-related thresholds based on their experience. Hence, it might be possible to assess the importance of these thresholds initially by collecting evidence from places that use them. A possible improvement of this method might also be to apply risk-related thresholds or benefit-cost ratio thresholds.

3.10.4 High risk abnormality threshold

Collins et al. 1988 mention that a real rupture risk has been demonstrated for aortic diameters in excess of 40 mm. Based on that, we define high risk of aortic rupture indicator to be a binary variable that is equal to 1 for all individuals in our study (79 cases) with aortic diameters at least 41 mm wide and equal to 0 otherwise.

By using risk factor analysis for the covariates included in our data, we find that the predictors that are significant are age groups indicator, smoking level, diastolic blood pressure (at least 90 mmHg) and co-morbidities and family history indicator. The only difference between this analysis and the others in previous sections is that in co-morbidities and family history indicator, chronic obstructive airway disease variable should be replaced by hypertension treatment indicator.

To be more specific, the co-morbidities categorical variable is equal to 1 if at least one of the following medical conditions or risk indicators is present:

- Myocardial infarction (mi)
- Cerebrovascular accident (stroke) (cva)
- Cardiovascular disease (cvd)
- Peripheral vascular disease (pvd)
- Coronary artery disease (cad)
- Family history of abdominal aortic aneurysm
- Hypertension treatment

Moreover, for high risk selective screening models, the 95% confidence interval for the area under the ROC curve is (0.657, 0.782) with age groups and (0.674, 0.802) with age and bpd (complete dataset estimates); the AUC's are not significantly different. Both models are useful because the individuals filling the questionnaires might not know their diastolic blood pressure measurement.

Finally, it should be emphasised that using selective screening for high risk cases with aortic diameters in excess of 40 mm does not make the models for abnormal cases obsolete. Both types of models can be used simultaneously in order to optimise the results of screening; this can be achieved if selective screening models are also regarded as priority setting models.

In other words, the cases to be screened can be selected using the model for abnormal cases and after that, priority for screening the selected individuals associated with greater probability of aortic rupture can be set by the model for high risk cases. This might be more useful in practices with large number of cases for screening, where a long period of time would be required for testing all these cases.

3.10.5 Weighted classification results

Previously, the importance of weighted classification in medical screening was analysed and the use of misclassification costs investigated. It is apparent that the criterion used to compare different models that are assessed for selective screening will determine the effect of weighting.

Additionally, it should be stressed that weights and misclassifications costs are partly based on subjective assessments of the importance of identifying abnormal cases; moreover, the effect of weighted modelling will in general vary with the criterion of abnormality and the definition of the target group.

The formulae for obtaining the estimated misclassification cost ratio have been given previously. When we applied these to our data, using information from the literature, this ratio has been found to be approximately 10, a similar value to that used in other screening problems (Ripley 1996).

Furthermore, there are other ways for obtaining weights for a number of cases; for example other predictors can be used to improve the estimation of the misclassification cost ratio or the expectation of life for each individual. In our case, we used the expectation of life based only on the age of a subject; the figures used can be found in the 1997-1999 interim life tables for males in England produced by the UK government's actuary's department.

The weighted models have been compared using the estimated area under the ROC curve. When the misclassification cost ratio is used, the 95% confidence intervals for this area are identical to the ones obtained when weights are not included. On the other hand, when the expectation of life has been the criterion of weighting individuals, the interval for the area under the ROC curve is (0.650, 0.716) (for the model with age groups); the corresponding interval with age and bpd is (0.665, 0.731) (estimated by using the complete dataset).

It must be emphasised that by including weighting when choosing the individuals for screening, we modify the criteria for selection of the target group in a predetermined way. For example, we will tend to find younger abnormal cases if we use age alone to calculate life expectation.

As the expectation of life is lower for cases with co-morbidities (Smith et al. 1993 and Cole et al. 1996), weighting of these cases should be lower than average; a possible way of achieving this is to deduct the years of life lost because of the presence of the co-morbidities from the expected life span. In addition, the quality of life might also be included in the weights by using QALY (Quality Adjusted Life Years); see Cox et al. (1992) for details.

Weighting individuals can be described as applying importance weights to each of the individuals. The importance weights can be defined for each individuals separately or to groups of individuals with common characteristics or same values for some of the variables in the data. Weighting errors is a special case of weighting individuals as the importance weights are the same for all individuals that belong to the same class of the response variable.

Life expectancy allows different weighting according to the function that defines the importance of each individual. All individuals with the same life expectancy are weighted equally. As expected, weighting errors does not change the area under ROC, a fact that is confirmed by our results.

Finally, the method described in Cole et al. (1996) for comparing the estimated total benefit when different percentages of the population are screened has been used. In the words of Cole et al. 1996 "logistic regression analysis was used to derive the risk of AAA for an individual based

on his risk factors" and then "for each individual the potential benefit of screening was calculated by multiplying his risk of developing AAA by his expectation of life". Finally "the proportion of the total benefit that would be obtained by restricting screening to a given proportion of the population with the greatest potential benefit was calculated by summation".

The results derived from our data and the corresponding results in Cole et al. (1996) paper (last column in table 2, where all risk factors are used) are shown in table 3.8 on page 93 below:

Estimated total benefit(%)	50	55	60	65	70	75	80	85	90
Proportion screened (%)	27	32	37	43	48	54	60	67	75
Cole et al. 1996 (%)	5	–	8	–	11	–	17	–	29

Table 3.8: Proportion needed to be screened to achieve a given proportion of total benefit

The large differences between our results and the results in Cole et al. (1996) paper can be attributed to the data used for constructing the models. On the one hand, our study was a cohort study including 3001 individuals from the general population, and the predictors used can be easily obtained by using self-administered questionnaires.

On the other hand, Cole et al. (1996) study used a case-control study, which was small (78 cases and 99 controls), hospital-based, and included variables such as diastolic and systolic blood pressure and serum high density lipoprotein (HDL) that might not be known to the individuals. Finally, in Cole et al. 1996, it is mentioned that "the controls were not screened and some of them might have had AAA".

3.10.6 Estimated Bayes risk

It is valuable for any data set used for classification to estimate the maximum potential for discrimination between the classes; the method we applied in our case was the calculation of the lower bound of the Bayes risk for different sets of predictors. Full details about this method can be found in Ripley (1996).

In the previous sections, we proposed three models for selective screening for abdominal aortic aneurysm. For the first model, where the predictors are age groups, smoking level and comorbidities and family history indicator, the 95% confidence interval for the lower bound of Bayes risk is (0.069, 0.079). The default error rate for our data is equal to the proportion of abnormal cases, which is 0.083.

For the second model, which includes also diastolic blood pressure at least 90 mmHg indicator, the corresponding interval is (0.067, 0.077), showing that this predictor does not improve signif-

icantly the separability of the classes. In both cases, this interval has been calculated by 1000 estimates of the Bayes risk and the differences can be attributed to the random selection of k nearest neighbours in the S-Plus function used.

On the other hand, by using age and diastolic blood pressure as continuous variables, the 95% confidence interval for the lower bound of Bayes risk is (0.052, 0.061); this indicates that the overlap between classes (normal and abnormal) is less by using the predictors mentioned above.

Thus, the simple models (e.g logistic regression) perform nearly optimal and the advanced models (for example boosting models) can hardly outperform them. It must be remembered that it is possible to improve the discriminative power of the data and reduce the lower bound for the Bayes risk by including additional variables.

3.11 Multiple class models

Previously, we described the application of models where the individuals in our dataset are classified as abnormal if their aortic diameter is at least 29 mm. Additionally, we investigate the possibility of separating cases above 40 mm as a distinct group for whom "a real rupture risk has been demonstrated (Collins et al. 1988). In this section, explore the implementation of multiple class models by integrating the two models mentioned above into a single classifier.

To be more precise, we label cases as low, moderate or high risk of aortic rupture according to their aortic diameter. A person is described as high risk if his diameter is above 40 mm and low risk if the corresponding diameter is less than 29 mm; the others are said to have moderate risk of rupture. In this way, we have 2752 low risk, 170 moderate risk and 79 high risk cases.

Moreover, we investigate models for separating moderate from high risk of rupture based on information collected by self-administered questionnaires. In the same way, we construct and test classifiers for discriminating low from moderate risk individuals. Finally, we implement the use of multi-categorical models; specifically, proportional odds logistic models and classification tree models are applied.

3.11.1 Separating moderate from high risk cases

We can separate moderate from high risk of aortic rupture individuals by comparing the prevalence of high risk of rupture for those with and those without a specific risk factor. Further details about the application of this method and description of the risk factors have been given previously.

The results indicate that the only risk factor whose prevalence for high risk cases is significantly different at the 5% level from the corresponding prevalence for moderate risk individuals is hyper-

tensive treatment. For continuous variables, we compare the mean value of the predictor for the two groups mentioned above. Only diastolic blood pressure is found to be significantly different (results not shown).

If we compare these results with the corresponding ones for discriminating low risk cases from the others, we can see that many predictors are not useful for separating the two groups under investigation. As expected, when using logistic model and classification tree models for case selection, the area under the ROC curve (AUC) for both models indicates that it is not possible to identify with acceptable accuracy high risk cases.

To be more precise, the 95% confidence interval for the logistic model AUC is (0.396, 0.548) and for the classification tree model (0.459, 0.577). Specific description of the method applied to obtain the confidence intervals for the AUC have been given previously.

3.11.2 Separating low from moderate risk cases

In a similar way of constructing and testing classification models for moderate and high risk cases, we investigate the possibility of separating low from moderate risk of aortic rupture cases. From prevalence tests for categorical and continuous variables included in our dataset, myocardial infarction, chronic obstructive airway disease, cardiovascular disease, diastolic blood pressure above 90 mmHg, heavy smoking and age are significant for building the case selection models.

The 95% confidence interval for the area under the ROC curve is (0.636, 0.675) for the logistic model and (0.611, 0.656) for the classification tree model. Based on the results for the the area under the ROC curve, we might conclude that these models can be used for case selection. An additional model at this stage could be a classifier that separates low risk individuals from high risk individuals.

3.11.3 Multi-categorical models

In previous sections, the constructed models should be combined with another classifier in order to separate the individuals included for case selection into three distinct groups. For example, the model for separating moderate from high risk cases should be combined with a discrimination procedure that identifies low risk cases. In this case, we apply two-stage case selection, where the information collected by self-administered questionnaires can be used in both stages.

Another way of separating the three risk groups is to build multi-categorical classifiers (models where the response variable is a k-class factor, $k > 2$) and test them according to suitable criteria. Two possible candidates when constructing this type of models in S-Plus are proportional odds

logistic regression and classification tree model.

The difference between having a multi-categorical classifier and two-stage case selection is that we need to build and test one model instead of a pair of two classifiers. Also, that means additional difficulties when comparing the single multi-categorical classifier and the pair of two-class classifiers.

To be more specific, the criterion used to compare performance of the two-class models, the area under the ROC curve (AUC), is not directly applicable to problems involving more than two classes. In Hand et al. 2001, a simple generalisation of the AUC for multiple class classification problems is proposed. The overall performance of the classification rule in separating the c classes is the average of the pairwise areas under the ROC curves:

$$M = \frac{2}{c(c-1)} \sum_{i<j} \widehat{A}(i,j)$$

where $\widehat{A}(i,j)$ is the AUC for classes i and j. In the same paper, it is mentioned that the standard deviation of M can be derived by using bootstrap methods.

In our case, for the proportional odds logistic model, the overall performance indicator M is equal to 0.632 when the complete dataset is used; the corresponding indicator for the classification tree is 0.695. It must also be mentioned that there are significant differences between the pairwise AUC included for calculating M.

For both models, the pairwise AUC for moderate and high risk of aortic rupture cases is less than the corresponding values of the AUC for the other two pairs. Especially in the case of the logistic model, the AUC for moderate and high risk cases is 0.529 indicating that for this task, the classification model is almost useless.

3.12 Measurement error and classification

In Grimshaw et al. 1992, it is stated that "the accuracy with which abnormalities can be detected is of crucial importance to a screening programme". For the Birmingham Community Aneurysm Project (CASP) "the aorta was examined with a Pie Data 150s scanner using a 3.5 MHz mechanical sector transducer".

For twenty patients, this measurement was compared to computer tomography (CT) measurement that is regarded as a "gold standard". The differences between the two measurements are assumed to be normally distributed and the estimated parameters for this distribution are: mean 0.1 mm and standard deviation 1.8 mm.

Clearly, given any threshold for defining abnormal aortic diameter, some cases might be mis-

classified as their actual diameter is different from that estimated. In order to assess the impact of this, we calculate the estimated area under the ROC curve (AUC) for the logistic and the classification tree model after adding random noise to the measurement of the diameter of the aorta. The noise added is normally distributed with the same mean and standard deviation as the differences between the ultrasound scanner and the CT measurement.

The 95% confidence intervals for the AUC are (0.650, 0.674) for the logistic model and (0.626, 0.663) for the classification tree model. The corresponding measurements without random noise are (0.651, 0.676) for the logistic model and (0.629, 0.668) for the classification tree model. Clearly, in terms of the estimated AUC, the differences are very small.

It is maybe more important to use this to calculate the number of misclassified cases for a specific threshold and the corresponding misclassification cost because of the ultrasound measurement inaccuracy. As the misclassification cost is not defined accurately for any given aortic diameter, we have not attempted to calculate the cost of having biased estimates for the diameter.

Other variables such diastolic and systolic blood pressure are measured with error. Hence, it is possible to study the effect of measurement error of predictors on classification accuracy. For categorical variables in the study, such as co-morbidities and family history indicator, we might assume that some of the patients have not been assigned to their actual factor level.

Nevertheless, due to the lack of information about measurement error in variables other than aortic diameter, we will not attempt to estimate misclassification accuracy for other possible types of error. Related literature and expert's opinion might be a possible source for measurement errors whenever that type of information is deemed reliable and relevant to our study.

3.13 Conclusions

In this chapter, we have compared different modelling approaches for AAA selective screening. Using the area under the ROC curve as criterion, two logistic regression models are chosen. The first includes as predictors of AAA presence the co-morbidities and family history of AAA indicator, smoking level, and age group indicator and the second has in addition to the predictors of the first model raised diastolic blood pressure. The performance of other types of models is similar to the logistic model.

The importance of assessing the discriminative power of a classifier by cross-validation and bootstrapping has been shown. With the inclusion of age and diastolic blood pressure on continuous scale, the classification tree outperforms the logistic model when the corresponding areas under the ROC curve are compared on the full data. The situation is reversed when the two models

mentioned above are compared using cross-validation and bootstrapping estimations of the area under the ROC curve.

Using different abnormality definitions, we showed that it is possible to use the same predictors as for the standard definition of abnormality (aortic diameter at least 29 mm) for screening. For the age-adjusted thresholds, age groups, smoking level and co-morbidities can be used as well as age and diastolic blood pressure on continuous scale. For the high risk abnormality threshold (diameter above 40 mm), the same predictors can be included in the selective screening classifier.

Weighted classification models have been implemented. Weighting misclassification errors as expected have no effect on the area under the ROC curve estimate. Other criteria for weights used are expectation of life and total benefit of screening. We showed that 50% of the total benefit of screening can be achieved by screening 27% of the target population and 90% benefit by screening 75%.

The Bayes risk estimation has been found close to the default error rate. This indicates that there might significant overlap between classes and the advanced models should not be expected to improve discrimination performance when compare to simpler models such as the logistic model.

Multiple class models showed that it is not possible to separate moderate from high risk cases. In addition to that, multiple class models tested show that it is possible to use proportional odds logistic regression and classification tree models for this type of selective screening.

Finally, adding measurement error to the aortic diameter has no influence on the discriminative power of the proposed logistic models.

Chapter 4

Graphical modelling

4.1 Abstract

In this chapter we present the implementation of graphical models using an empirical study with data from an aortic aneurysm screening project. We consider different types of graphical models for classification as well as for identifying the influence and association structure of the variables. The Naive Bayes classifier (NBC) is examined as selective screening model and found as possible candidate for the AAA study. Occam's window model selection is implemented and the EM-algorithm as well as multiple imputation method are explored. Graphical modelling is used to investigate whether the aortic diameter can be described as a mixture of normal distributions. The mixture with three components and unequal variances both using the EM-algorithm in MIM and minimisation procedure in S-Plus confirm the current clinical thresholds for aortic diameter. Finally, the growth model proposed for aortic diameter identified 38mm as the diameter threshold that gives minimum misclassification error rate.

4.2 Graphical modelling

The main motivation behind using graphical models is the fact that reliance on pairwise marginal associations may be very misleading (Edwards 2000, page 8). An example of this is a phenomenon of having results when analysing the complete sample that are reversed when analysing each of the subsamples separately; it is known in the literature as Simpson's paradox.

Independence and conditional independence are two terms that are necessary to describe Simpson's paradox. Two variables are independent if their joint distribution is equal to the product of their marginal distributions. On the other hand, two variables are conditionally independent given

a third variable if they are independent for each possible value of this third variable.

Using terms of marginal and conditional distribution, "Simpson's paradox refers to a reversal in the direction of association between the marginal and conditional distributions" (Edwards 2000, page 9). In other words, two variables can have marginal association, which means they are not independent and also they can have no conditional association given another variable, which means they are conditionally independent.

A graph is a structure of vertices, representing variables in the dataset, and edges, representing associations between variables. Undirected graphs are the graphs where the edges are undirected, meaning that by the edges we represent association and not influence or causal relationship between variables. In this instance, a missing edge between two vertices implies that the variables corresponding to these vertices are conditionally independent given all the other variables in the model.

In addition to that, directed graphical modelling can be implemented where there is information about the structure of the model. For example, information about the temporal structure of the data in longitudinal studies, can be used when constructing a directed graph so that the direction of the edges is compatible with the temporal structure of the data. Furthermore, information about the influence of one variable onto another can be included in a similar way to temporal relationships.

At the stage of exploratory analysis, we apply graphical model analysis to find the association structure of variables in our data; in others words using undirected graphical models. In addition to that, it is possible to explore the data further by including influence structure information into association exploratory analysis; that means use directed graphs for finding influence relationships between variables.

The same models applied for exploratory analysis can also be used for other reasons. To be more specific, "statistical techniques have two main purposes: explanation or prediction. For example, studies may be set up with the broad goal to increase insight or understanding about a problem, without a more specific practical purpose. These stand in contrast to, for example, meteorological studies whose objective is to predict future weather, or economic studies for which the goal is to predict next year's gross national product" (Edwards 2000, page 7).

Initially, we would try to construct graphical models for the purpose of explanation and to understand the association between variables in our data. Hence, we will implement simple graphical models and try to integrate knowledge gained by other types of exploratory analysis described previously into constructing these models. Later on, when constructing models for prediction of

specific events, we would include the information gained by graphical models. Also, we could apply graphical models for prediction, probably with specific modifications that will give to these models better predictive performance; the exploratory stage is also important for this purpose.

Three different types of undirected graphical models could be implemented for exploratory analysis depending on the variables include in our data. For discrete (categorical) variables, loglinear models are implemented; for continuous variables, a possible class of corresponding models are Gaussian graphical models. Finally, mixed models need to be applied when having both discrete and continuous variables.

If it is necessary to explore influence structure of the data, directed acyclic graphs are a possible answer. Furthermore, chain graphs, which are a mixture of directed and undirected edges for both categorical and continuous variables can be used for exploring the data. This type of graphical models should be implemented "when good prior knowledge of direction of influence is available" (Edwards 2000, page 12).

Further details about graphical modelling can be found in many publications in this area. For the interested reader, we recommend Cowell et al. (1999) for a general overview of the graphical modelling approach and its applications. We should also mention Lauritzen and Spiegelhalter (1988) that is very frequently cited in publications related to this topic.

4.2.1 Definitions

For the statistical analysis of our data, we used the statistical package MIM (Mixed Interaction Models). It is important to give some definitions related to graphical models that are useful for understanding the results derived by statistical tools that can implemented in MIM. These definitions can be found in Edwards (2000).

Homogeneity: If I is a fixed grouping factor, and Y is a response, then if the densities $f_{Y|I}(y|i)$ are constant over i, then we call this homogeneity.

Independence: Two random variables X and Y are independent if

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

or equivalently if the conditional density of one variable is not a function of the other, for example

$$f_{Y|X}(y|x) = f_Y(y)$$

Conditional independence: This is very important in graphical modelling. For three random

variables X,Y and Z, X and Y are conditionally independent given Z if for each value z, X and Y are independent in the conditional distribution given Z=z. In this case, the notation used is

$$X \perp\!\!\!\perp Y|Z$$

Pairwise Markov property: For undirected graphs, the edge between all pairs of variables X and Y is omitted if $X \perp\!\!\!\perp Y|$(the rest); for all other pairs, an edge is drawn between them. Thus, if two variables are not adjacent in the graph, then they are conditionally independent given the rest. This is known as the pairwise Markov property for undirected graphs.

Global Markov property: For undirected graphs, if two sets of variables u and v are separated by a third set of variables w, then $u \perp\!\!\!\perp v|w$.

Deviance: Under multinomial sampling with N observations, the likelihood of a table $\{n_{\mathbf{k}}\}$ of arbitrary dimension $\mathbf{k}$ such that $N = \sum_{\mathbf{k}} n_{\mathbf{k}}$ is

$$L(\{p_{\mathbf{k}}\}|\{n_{\mathbf{k}}\}) = \frac{N!}{\prod_{\mathbf{k}} n_{\mathbf{k}}!} \prod_{\mathbf{k}} p_{\mathbf{k}}^{n_{\mathbf{k}}}$$

The values of $p_{\mathbf{k}}$ that maximise this expression for a given model are called the maximum likelihood estimates (MLEs) and are written $\widehat{p}_{\mathbf{k}}$. The logarithmic function is monotonic, hence the MLEs also maximise the log likelihood

$$l(\{p_{\mathbf{k}}\}|\{n_{\mathbf{k}}\}) = ln(\frac{N!}{\prod_{\mathbf{k}} n_{\mathbf{k}}!}) + \sum_{\mathbf{k}} n_{\mathbf{k}} ln p_{\mathbf{k}}$$

The deviance of a model M_0 is the likelihood ratio test of M_0 versus the unrestricted (saturated) model M_f, i.e.,

$$G^2 = 2(\widehat{l}_f - \widehat{l}_0)$$

where $\widehat{l}_f$ and $\widehat{l}_0$ are the maximised log likelihoods under M_f and M_0 respectively.

Model formulae (for graphical models in MIM): These have the form

$$d_1, d_2, ..., d_r / l_1, l_2, ..., l_s / q_1, q_2, ..., q_t$$

where

$$d_1, d_2, ..., d_r$$

are **discrete** generators (discrete variables only),

$$l_1, l_2, ..., l_s$$

are **linear** generators (each containing exactly **one continuous** variable),

$$q_1, q_2, ..., q_t$$

are **quadratic** generators (contain both discrete and continuous variables)

The generators are closely related to the probability density function

$$f(i,y) = p_i|2\pi\Sigma_i|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(y-\mu_i)'\Sigma_i^{-1}(y-\mu_i)\}$$

written in canonical form:

$$f(i,y) = \exp\{\alpha_i + \beta_i' y - \frac{1}{2} y'\Omega_i y\} \text{ or}$$

$$f(i,y) = \exp\{\alpha_i + \sum_{\gamma\in\Gamma} \beta_i^{\gamma} y_\gamma - \frac{1}{2}\sum_{\gamma\in\Gamma}\sum_{\eta\in\Gamma} \omega_i^{\gamma\eta} y_\gamma y_\eta\}$$

where Γ is the set of continuous variables

The generators specify how these parameters may vary over i in the following way. Each parameter, i.e. $\alpha(i)$, each element of the q-vector $\beta(i)$, and each element of the symmetric $q \times q$ matrix $\Omega(i)$, is assumed to be expressible as a factorial combination of terms depending on various subsets of the discrete variables.

For example, if there are two variables J and K, then under some models we may restrict the way the discrete parameter varies over j and k, by constraining it to satisfy $\alpha(j,k) = u(j) + w(k)$. The discrete generators specify this expansion, by specifying the maximal terms not set to zero. So here the generators would be J and K. In the same way, the linear generators containing a continuous variable x determine the factorial expansion for the element of $\beta(i)$ corresponding to x.

Similarly for two continuous variables, w and x, say, the factorial expansion for the (w,x) element of $\Omega(i)$ is given through the quadratic generators that contain w and x. There are two syntax rules for model formulae.

First, for every linear generator, there must be a discrete generator that is at least as large. Subsequently, for every continuous variable in a quadratic generator, there must be a linear generator that is at least as large. It can be shown that these follow from the requirement that the models are invariant to location and scale transformations of the continuous variables (Edwards

2000).

Using the notation for the canonical form of the probability density function

$$f(i,y) = \exp\{\alpha_i + \sum_{\gamma\in\Gamma} \beta_i^{\gamma} y_{\gamma} - \frac{1}{2}\sum_{\gamma\in\Gamma}\sum_{\eta\in\Gamma} \omega_i^{\gamma\eta} y_{\gamma} y_{\eta}\}$$

we can summarise the rules for graphical model generators in the following list:

- A, B discrete, $A \perp\!\!\!\perp B|$(the rest) $\Rightarrow \forall\gamma, \eta \in \Gamma$ set α_i, $\beta_i^{\gamma}, \omega_i^{\gamma\eta}$ that contain an interaction involving AB equal to zero. No discrete, linear or quadratic generator contains AB.
- A discrete, X continuous, $A \perp\!\!\!\perp X|$(the rest) $\Rightarrow \forall\eta \in \Gamma$ set $\beta_i^{X}, \omega_i^{X\eta}$ that contain an interaction involving A equal to zero. No linear or quadratic generator contains AX.
- X, Y continuous, $X \perp\!\!\!\perp Y|$(the rest) $\Rightarrow$ set ω_i^{XY} equal to zero. No quadratic generator contains XY.

4.2.2 Loglinear models

For categorical data, we need to implement models that link variables with number of subjects for each combination of levels of these variables. We can summarise cross-classification of variables in a contingency table; for each cell of this table we have to estimate the probability that an observation falls in the specific cell. Loglinear models "are so called because they apply linear models to the logarithms of the expected cell counts" (Edwards 2000, page 14).

These logarithms are linked to the variables via parameters that we estimate; these parameters are known as interaction terms. Depending on the model under consideration, some of the interaction terms in model formula can be omitted, or equivalently set to zero, depending on the interactions that are significant.

The likelihood of contingency table assuming multinomial sampling can be used to judge how well a model under consideration fits the data. To be specific, we compare maximised log likelihoods of the candidate model and the saturated (unrestricted) model via the likelihood ratio test of these likelihoods; in the literature, this test is known as the deviance of the candidate model (Edwards 2000, page 16).

We can assess whether a model under examination is significantly different from the saturated model because of the fact that the deviance is "asymptotically $\chi^2_{(k)}$ distributed, where k (the degrees of freedom) is given as the difference in number of free parameters between the candidate and the unrestricted model" (Edwards 2000, page 16). The candidate model is rejected if the deviance is larger than $\chi^2_{(k)}$.

Hierarchical loglinear models have the characteristic of setting higher-order interactions to zero when a lower order interaction that these high-order interactions contain has been omitted. Graphical models are a subclass of the hierarchical models and are defined by setting specific two-factor and related higher-order interactions equal to zero.

This fact means that "higher-order interactions included in the model are defined by the two-factor interaction terms" (Edwards 2000, page 17). Additionally, the link between graphical models and conditional independence relationships between variables is derived from the fact that omitting a two-way interaction corresponds to having the corresponding factors conditionally independent given all the other variables.

The expected cell counts estimated by applying a loglinear model can be found by iterative methods; the most popular is the iterative proportional scaling (IPS) algorithm. This estimation is based on a formula that involves the marginal table of observed counts and the corresponding table of fitted counts. These tables are directly linked with the generators in the model, which are the subsets of interaction terms included in this models and are are regarded as the building blocks of a graphical model. For more details see Edwards (2000).

4.2.3 Graphical Gaussian models

For continuous variables, we construct graphical models on the assumption that the joint distribution of these variables is multivariate normal. In the literature they are also known as covariance selection models; this name come from the fact that the covariance matrix, or to be precise, the inverse of the matrix, called the precision matrix is of fundamental importance in the formulation of this type of models.

Specifically, for a pair of these variables, their conditional distribution is bivariate normal. From their covariance matrix, it is possible to calculate their partial correlation coefficient, which is their correlation taken all the other variables into account. Then, this partial correlation is equal to zero if the corresponding term in the precision matrix is zero. Hence, "two variables are independent given the remaining variables if and only if the corresponding element of the inverse covariance is zero" (Edwards 2000, page 36).

Similarly to loglinear model analysis, we compare a candidate with the saturated model by using the deviance that follows χ^2 distribution with degrees of freedom equal to the number of free parameters; in this case this is equal to the number of edges missing from the candidate model. We should mention that a missing edge between two variables in a graphical model is equivalent to assuming the these variables are conditionally independent given the rest.

4.2.4 Mixed graphical models

If both discrete and continuous variables are included in the data, a combination of loglinear and graphical Gaussian models can be implemented to assess the relationships between variables. The general principal behind mixed graphical models is the factorisation of density function into two parts, the discrete part assuming multinomial distribution and the continuous part assuming multivariate normal distribution for each combination of levels of the categorical variables.

In other words, for each possible configuration of values of the discrete variables, the mean and the covariance of the multivariate normal depend on this configuration. That is why the multivariate normal distribution is also called conditional Gaussian (CG) distribution (Edwards 2000, page 59). If this distribution is the same for all possible combinations of levels, then we have homogenous model; the general case model, where the CG-distributions are different is called heterogenous model.

There are specific rules for constructing a mixed graphical model; these are related to the generators of the model which correspond to the three parts of this model: discrete, linear and quadratic. For example, for each linear generator there must be a corresponding discrete generator such that the intersection of the linear generator and the set of discrete variables belongs to the set of discrete generators in the model. Also, for each quadratic generator and each continuous variable, we should have a corresponding linear generator (Edwards 2000, page 63).

Moreover, the graph of the model and in particular the removal of edges corresponds to setting specific generators equal to zero. For instance, the edge between two discrete variables is omitted if no discrete generator contains the interaction between these two variables. This is equivalent to the statement that these two discrete variables are conditionally independent given the rest of the variables.

In the same way, removing an edge between a discrete and a continuous variable is equivalent to the omission of all quadratic and linear generators that contain the corresponding interaction. A similar rule is applied for the case of conditional independence between two continuous variables.

As for loglinear and graphical Gaussian models, the deviance between a model under consideration and the saturated model has asymptotic χ^2 distribution "with degrees of freedom given as the difference in the number of free parameters between the two models" (Edwards 2000, page 68). The same test for deviance significance can be implemented for comparison of two alternative models in order to find the one that fits better the data.

When having a large number of variables, it might be possible to break down the graphical model into simpler models without loss of information. Decomposable models are graphical models "for

which explicit maximum likelihood estimates for the parameters can be derived" (Edwards 2000); they are regarded as one of the best candidates for this purpose.

For example, parameter estimation is computationally more efficient as there are closed form for the maximum likelihood estimates. Also, it is easier to interpret decomposable models as "they are equivalent to a sequence of univariate conditional models (regressions)" (Edwards 2000, page 93).

Deviance criterion is mentioned above as a possible way to decide about the best candidate model. At this stage, we are looking into the application of graphical models for exploratory analysis of data. Hence, it is important to have a simple way of assessing the relationships between variables in the dataset and also the ability to compare them using other types of model criteria that take into account other characteristics of the model.

For example, we might need to balance goodness of fit with parsimony in order to avoid overfitting and unnecessary complexity. Information criteria such as Akaike's information criterion (AIC), which chooses the model that maximises likelihood and the smallest number of parameters might be the appropriate criterion especially when having a large number of variables. Similar to this criterion is the Bayesian information criterion (BIC).

Furthermore, we need to have the ability at the exploratory stage, to include external information from the literature or from subject experts. In terms of graphical models, this should be translated into removing or adding specific variables, transforming or combining variables and also adding or removing edges in the graph. An additional virtue of the software implemented is the ability to present the results in a way that it will be easy to explain to medical experts that are not familiar with this type of modelling.

One possible software candidate that fulfils the requirements described above is a statistical program called MIM (Mixed Interactions Models); see Edwards (2000) for further details. It is necessary at this stage to describe the selection procedures available in this computer program and explain how the results can be included in the exploratory analysis of data and also in further statistical analysis such as construction of models.

4.2.5 Selection procedures in MIM

The approaches to model selection implemented in MIM are based on the assumption that "little prior knowledge or relevant theory is available, and so model choice becomes an entirely empirical, exploratory process" (Edwards 2000, page 157). Stepwise selection is the most popular of the approaches available in MIM; at each step, edges are added or removed according to some predefined

criterion.

Another way to select the appropriate model is to apply EH-procedure, which "seeks the simplest models consistent with the data" (Edwards 2000, page 167). Finally, information criteria such as AIC and BIC are used for choosing the optimum model.

The standard implementation of stepwise method is backward selection. In this case, edges are removed if the deviance difference between the candidate and the currently accepted model indicates that the edge removed is not significant using χ^2 tests. The procedure stops when all edges that are present in the model are significant or they cannot be removed; there are several reasons why an edge is not eligible for removal.

Firstly, specific edges can be fixed in the model; this is a possible way to include external information in the modelling process even though in the exploratory analysis it might be preferable to select a graphical model based only on the data. Another reason is the principle of coherence; in this case, an edge that has been found significant at one step, cannot be removed. Finally, decomposability of the graph might be another reason for the ineligibility of an edge in the graphical model.

Other options available in MIM for model stepwise selection are forward selection, which gives quite similar results to backward selection and also headlong selection, where edges are included or removed after being chosen at random order. It is also possible to ignore coherence in selection and also to allow non decomposable graphical models to be accepted as valid.

The EH-procedure, named after the initials of its authors (Edwards and Havranek), has the aim of finding the simplest model consistent with the data. It uses overall goodness of fit tests; in this case a model is accepted or rejected based on the χ^2 deviance test. Models are classified as accepted or rejected based on the coherence principle; the procedure stops when all candidate models have been classified.

Finally, AIC and BIC are used to select the optimum model. Both criteria are a compromise between goodness of fit (by assessing maximum likelihood) and complexity of the model (using the number of free parameters). The current version compares all possible models, thus it might be time consuming when a large number of variables are present in the candidate models.

There are particular advantages and disadvantages in the selection methods described above; choosing which one to implement depends on a number of reasons. For example, a possible disadvantage of EH-procedure compared with stepwise selection is the use of tests for overall goodness of fit instead of deviance tests. Edwards (2000, page 173) state that "with high dimensional contingency tables, such tests are known to be less reliable than tests based on the deviance differences".

When it is feasible, information criteria selection should be implemented as it is a straightforward procedure and it takes into account the complexity as well as the goodness of fit of the model. The number of variables might make this type of selection computationally expensive; nevertheless, we believe that for exploratory analysis we need to find a model that indicates the simplest association structure.

Also model selection based on information criteria avoids the problem of having a large number of significance tests. To be more specific, in the case of stepwise and EH-procedure, "the overall error properties are not related in any clear way to the error levels of the individual tests" (Edwards 2000).

Finally, for all selection procedures, we should remember that when it is possible, we can include prior knowledge either from other related studies or based on expert's opinion about the association structure. We can either add or remove edges and test whether the new model is acceptable by the corresponding criterion.

In this case, the model derived and the information extracted from it will be closer to current practice and can be more acceptable from subject specialists. This does not mean that newly discover associations found by graphical modelling should be rejected immediately. We need to examine careful the conclusion derived from our exploratory analysis since these are going to be integrated into model construction and other types of statistical analysis.

4.2.6 Chain graphs

When prior knowledge about study subject is available, especially when this is related to possible influence of some variables onto others, then it is possible to explore the data by implementing directed graphs. In this case, the directed edge should indicate direction of influence; similarly, it is possible to include the temporal structure of the data in the models. This is particularly important when dealing with longitudinal data, where several measurements are taken over time for an individual.

For example, if biological characteristics such as race and gender are included, it might be better to replace undirected edges between these variables and others such as food preference with directed edges. In other words, as it does not make sense to test the influence of food preference on race or gender, we should prefer to test whether the influence of race and gender on food preference is significant.

In this instance, "it may be legitimate to call gender and race determinants of the response" (Edwards 2000, page 190). It it also stated that gender and race cannot be regarded as causes as

they are no ways of having an intervention or treatment to change them.

One of the important differences between undirected and directed graphs is the interpretation of conditional independence in case of a missing edge between two variables. On the one hand, an edge is omitted from an undirected graph if the variables are conditionally independent given all the other variables in the graph. On the other hand, a missing edge in a directed graph means conditional independence of the variables given all prior variables (Edwards 2000, page 193).

Chain graphs are a mixture of directed and undirected edges. When having variables that we do not have prior information about direction of influence between them, then we connect them using an undirected edge between them and test its significance.

In MIM, in case of exploring the data using chain graphs, we need to define the ordering of the variables; those variables that we want to put on equal footing are included in the same block and are linked using undirected edges. Otherwise, directed edges connect all the other variables; the ordering of the variables is also related to the test ordering.

To be more specific, as conditional independence of variables is related to prior variables only, the edges of prior variables are tested first for significance and subsequently the edges for next group of variables are assessed based on the results from the previous group of variables; see Edwards (2000) for further details.

4.2.7 Graphical model classifiers

As we have seen in the previous sections, there are many different types of classification models that can be applied to separate subjects in the study into groups by their attributes. The best solution is rarely known prior to the implementation of different models as there are several parameters that define the optimum discrimination rule. These parameters could be minimum misclassification error rate, user friendliness of the model or optimisation of some utility function that is specifically used for a specific problem.

The StatLog project is a comparative study for different types of classifications models that attempts to define criteria or indications by which a user will select the type of model under specific circumstances. Three types of measures of datasets are applied for this comparison: simple, statistically based and information theoretic measures (Michie et al. 1994).

Other types of models, either of similar type to the ones compared in StatLog or others that have not been included in this study might be implemented. In addition to that, improvements to classification algorithms lead to us to the conclusion that there is no guarantee that the results are applicable to the classifiers that are available today.

Another aspect of the problem of finding the optimum classifier that is quite often ignored is the availability of a model to the constructor of the classification rule and the end user. There is no point in constructing a "state of the art" classification model that needs computing facilities for implementation and it costs a large of money when the end user cannot afford the means necessary for this implementation.

Also, there are situations where a practical rule in the form of a flow chart or guidance table is preferable to an accurate but slow program as speed might be more important than minimisation of misclassification cost. An example of this is the use of decision support systems in intensive care units where clinicians have to make decisions for life threatening situations in a matter of seconds.

This does not mean that complicated models are not acceptable into solving problems of classification in the medical area. Graphical models are particularly useful in understanding complex relationships between variables in the data. Their implementation is sometimes problematic, especially when the end users are not familiar with this type of models and they are not willing to spend time and money into implementing something that they might believe is unnecessarily complicated.

In this case, simplification of the way graphical models are applied might be crucial in making results derived by such models acceptable. It is also possible to find a way of implementing the results by constructing a decision rule based on these results that is suitable for the environment where classification is going to be needed.

The graphical models that are used for exploratory analysis of the data are usually the optimum in terms of maximum likelihood or some other criteria such AIC and BIC that penalise complexity. In other words, using this type of model selection criteria, belonging to the class of *minimum description length* (MDL) scores, we obtain the best model that achieves minimisation of "the error of learned (graphical model) over all the variables in the domain" (Friedman et al. 1997).

The same authors state that "minimising this error, however, does not necessarily minimise the local error in predicting the class variable given the attributes". This can be seen from the fact that the log likelihood function of this model can be factorised to sum of the log likelihood of the conditional distribution of the class variable given the attributes and the likelihood of all the attributes. When there are many attributes, "using MDL (or other nonspecialised scoring functions) for learning (graphical models) may result in a poor classifier".

Naive Bayes classifier is a directed graphical model where it is assumed that all the attributes are independent given the class variable. Michie et al. (1994, page 40) state that "the assumption of independence makes it much easier to estimate (the posterior probabilities of each class given the

attributes) since each attribute can be treated separately". This type of model has been successful in many classification problems; Friedman et al. (1997) mention that "the performance of Naive Bayes is somewhat surprising, since the above assumption is clearly unrealistic".

From the exploratory analysis, it is possible to identify which variables are not related to the class variable and remove them from the classification graphical model. To be more specific, for a directed acyclic graphical model, also called *Bayesian network*, we can define relevant attributes "on the notion of a *Markov blanket* of a variable X, which consists of X's parents, X's children, and the parents of X's children in a given network structure G" (Friedman et al. 1997).

Other available information, such as strong correlation between variables can be included in the model; in the case of Naive Bayes model edges are added between variables when the association between them is significant. Friedman et al. (1997) call these models augmented Naive Bayesian networks and the edges added augmenting edges.

The same authors propose a procedure of adding edges and their direction and also the possibility of using separate models of association structures for each class; these are known in the literature as *Bayesian multinets*. External information can be included in these models by adding or removing edges according to indications from subject experts or other related publications.

It is obvious from the description of different types of models that the optimisation criterion is the touchstone by which the best classifier will be found. Hence, as we mentioned previously, we need to take into account model selection criteria and performance assessment procedures in order to clarify in which cases a particular criterion is appropriate. This will also be useful for deciding what actions need to be taken in order to improve the performance of a particular model based on the results of model selection and comparison.

4.3 Results from CASP data

In Edwards 2000, "the inadequacy of studying only pairwise associations between variables" is illustrated by giving three examples related to Simpson's paradox. This paradox "refers to a reversal in the direction of association between the marginal and the conditional distributions".

For example, using variables included in the data we analyse, we have the apparent association in marginal distribution shown in figure 4.1 on page 113. This result indicates that co-morbidities and family history of AAA indicator (com) is not independent from systolic blood pressure (bps).

On the other hand, if we include diastolic blood pressure (bpd), the result is shown in figure 4.2 on page 113. In this case, we can see that systolic blood pressure is independent from co-morbidities and family history of AAA indicator given diastolic blood pressure. It is obvious that the apparent

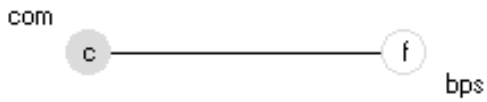


Figure 4.1: co-morbidities $\not\perp\!\!\!\perp$ bps

association in marginal distribution in figure 4.1 on page 113 is different from that in conditional distribution (figure 4.2 on page 113).

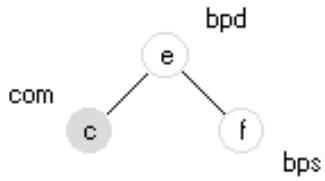


Figure 4.2: co-morbidities $\perp\!\!\!\perp$ bps|bpd

Furthermore, in Edwards (2000) it is mentioned that "a cause can have a positive effect in two subpopulations, but a negative effect when these are combined" and "the apparent paradox is due to this mistaken causal interpretation".

It is obvious from what is mentioned above that "reliance on pairwise marginal associations may be very misleading. It is necessary to take a multivariate approach by including all relevant variables in the analysis so as to study the conditional as well as the marginal associations. This is the basis of graphical modelling" (Edwards 2000).

4.3.1 Alternative coding

We have previously seen that information about smoking has been recorded by two variables: number of cigarettes per day and years since last smoked. Later, because of the possible lack of reliability of self-reported values for smoking, we have combined the initial variables into a factor with levels corresponding to level of smoking or type of inconsistent information (table 2.3 on page 40).

To avoid creating problems with statistical inference in graphical modelling because of data sparseness, we can combine levels 3 (inaccurate information) and level 4 (poor information) of smoking level variable coded as smol. The frequencies table of the recoded smol variable, labelled as smol2 is shown in table 4.2 on page 114.

0	1	2	3
263	820	1480	438

Table 4.2: Recoded smoking level variable (smol2)

It might seem very odd to have such a large proportion of heavy smokers in the sample. Dr. Gill Grimshaw, who was responsible for data collection has provided additional information about the reasons behind the high proportion of heavy smokers. The individuals in this cohort were all of an age to be either fighting or reservists in the Second World War. During Second World War, free cigarettes were provided to those being in the army; therefore it should not be a surprise that a large proportion of them became heavy smokers during that period of time. This is reflected by the fact that the majority reported YES to the question "Have you ever smoked?"

On the other hand, it is possible to recode co-morbidities and family history indicator (variable com) into two binary variables. These variables, labelled as car and vas are shown in table 4.4 on page 114. We use this alternative coding to investigate which type of impairment is more useful for predicting AAA and how these impairment are linked to other factors in the data.

Level 1 indicates no-presence of the specific impairment and 2 presence of the impairment indicated by the variable. The specific coding (1 and 2) has been used to make the data compatible with MIM requirements for data analysis. To be more precise, factor variables accepted with letters, negative numbers or zero as indicators of factor level.

Variable	Absent(=1)	Present(=2)
Cardiac (car)	2476	525
Vascular (vas)	2283	718

Result of recoding co-morbidities and family history indicator (com)

Table 4.4: Cardiac and vascular impairment indicators

Cardiac impairment is present when at least one of the following co-morbidities is present:

- `cad` (coronary artery disease)
- `mi` (myocardial infarction)
- `cvd` (cardiovascular disease)

Vascular impairment is present when at least one of the following co-morbidities is present:

- `pvd` (peripheral vascular disease)

- `ht` (hypertensive treatment)
- `cva` (cerebrovascular accident or stroke)

4.3.2 Bayesian Network Classifiers

Using the alternative coding for the data presented previously, we initially attempt to construct graphical models that will be implemented for classification. One of the commonly used Bayesian Network classifiers is the Naive Bayes Classifier (denoted as NBC) which has been described previously.

The variables used to construct the Naive Bayes Classifier for our data can be displayed in MIM as shown in table 4.6 on page 115. The graphical representation of this type of classifier is shown in figure 4.3 on page 116.

Variable	Label	Type	Levels (discrete)
g	age	Continuous	-
h	bps	Continuous	-
i	bpd	Continuous	-
a	aaa	Discrete	2
b	smol2	Discrete	4
c	car	Discrete	2
d	vas	Discrete	2
e	alc	Discrete	2
f	diab	Discrete	2

Type of variable and number of levels for factors also shown in table.

Table 4.6: Variables used in Naive Bayes Classifier

The way to construct the Naive Bayes classifier in MIM is as follows. Initially, we define the block recursive structure of the model (Edwards 2000). In this case, the first block contains only abdominal aortic aneurysm indicator (labelled aaa).

In the second block, the rest of the variables are included and in this way we indicate that we investigate the distribution of the risk factors in the model conditional on the disease indicator. At this stage, we can choose which variables to exclude from the second block to make the model simpler or to remove factors that might not be easily available.

After that, a possible way to construct Naive Bayes classifier is to define a *main effects* model (complete independence between variables) and then add the directed edges from aaa to all the other variables. It is clear that the risk factors (variables b to i in figure 4.3 on page 116) are conditionally independent given the class variable (variable a in the graph).

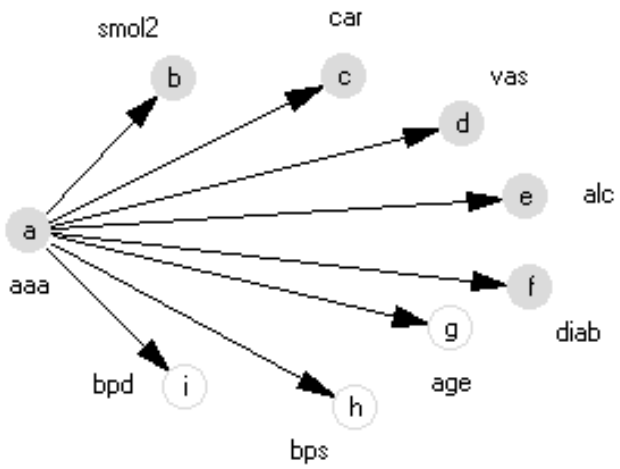


Figure 4.3: Naive Bayes Classifier

Another way to construct this model is to define the saturated model (heterogenous or homogenous) and then remove all the edges between risk factors. Obviously, we might decide to exclude some variables from the model by removing the edge that links aaa and the variables we want to exclude.

In Edwards (2000), it is stated that in MIM "the command `Classify` can be used to compute predicted classifications using the maximum likelihood discriminant analysis method. Each observation is assigned to the level g with largest estimated density $\widehat{f}(g,j,y)$", where "

$$\widehat{f}(g,j,y) = p_i|2\pi\Sigma_i|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(y-\mu_i)'\Sigma_i^{-1}(y-\mu_i)\}$$

with $i = (g,j)$ where g is a level of the grouping factor G, j is a $(p-1)$-tuple of discrete measurement variables [and] the q-vector y contains the continuous measurement variables". "The density estimate can either use all available observations or use the leave-out-one method; that is, the density for each observation is estimated using all available observations except the one in question".

At the present time, the MIM classification option (command classify) is not available for directed graphs. To overcome this problem, it is possible to use the *moral graph* of the directed

graphical model. This is obtained by joining all the parents of each node and subsequently replacing all arrows in the graph with undirected edges (lines) (Edwards 2000).

In Cowell et al. (1999), Lemma 5.9 states that if a probability distribution P "admits a recursive factorisation according to the directed acyclic graph D, it factorises according to the moral graph D^m and therefore obeys the global Markov property relative to D^m".

In our case, classification by using the directed version of Naive Bayes classifier is equivalent to the corresponding classification of the moral graph of this classifier. As there is only one parent node in the graph, corresponding to AAA indicator, the moral graph can be obtained simply by replacing all directed edges by the corresponding undirected ones. In MIM, this can be done by "switching off" the block mode option and define the moral graph of the Naive Bayes Classifier using undirected graphical model definition.

Classification can be implemented by using the complete dataset to construct and derive the predicted class and log-densities for each level of the predicted factor (variable aaa). To obtain less biased results, we might use the option of leave-one-out (n-fold cross-validation) classification. In this case, one observation is removed from the learning dataset and it is used for prediction.

Leave-one-out is repeated for all observations and can be computationally demanding with large datasets. Other possible options to obtain less biased results are to use bootstrapping or cross-validation with a smaller number of groups. Also, we might implement a combination of bootstrapping and classification that has been previously used for logistic and classification tree models for this dataset.

After obtaining the predicted class and the corresponding log-densities from complete and leave-one-out classification, we need to estimate the area under the ROC curve for each option. As this is not currently available in MIM, we extract the results from MIM and import them in S-Plus.

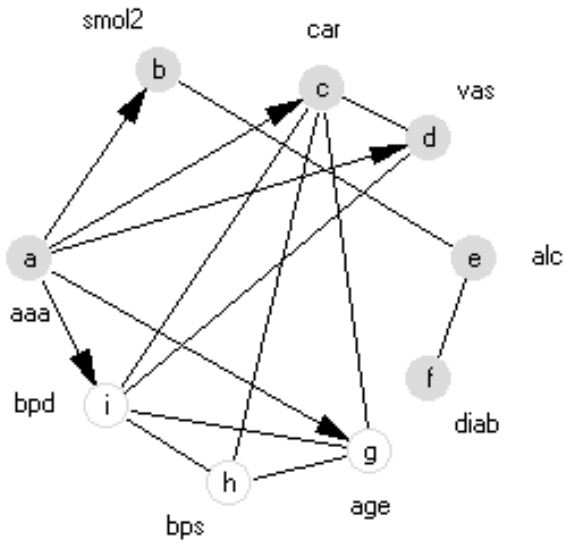
The log-densities for variable AAA level 1 (normal cases) is always larger than the corresponding log-densities for level 2 (abnormal cases). This has as result that all cases are predicted by the Naive Bayes Classifier as normal and the estimated area under the ROC curve is 0.502. The corresponding 95% confidence interval is (0.500, 0.503). Similar results are obtained if we use the complete data and leave-one-out estimates of the log-densities for level 1 and 2 of the variable (results not shown).

The results are unexpected as fitting 8 risk factors, even if unrelated, should give better area under the ROC curve alone. After checking MIM with other combinations of variables, there seems to be a problem with the classification command in MIM. Naive Bayes classifiers can be fitted in R using library e1071 that includes fitting of Naive Bayes models with both categorical

and continuous variables. The 95% confidence interval of the area under the ROC curve for the Naive Bayes model using the complete data is (0.590, 0.652).

In Friedman et al. 1997, the use of augmented Naive Bayesian networks is proposed where strong correlation between variables can be included in the model. It might be possible to include or exclude edges from the graph by using related literature or expert's opinion (Hojsgaard 1996).

A possible way to construct an augmented Naive Bayesian network is to start from the complete independence directed graphical model. Subsequently, we add directed edges between the AAA indicator and the other risk factors and also undirected edges between the risk factors (block 2 in the recursive block structure). In our case, we used forward selection using AIC criterion for deciding about the addition of an edge, either directed or undirected. The result is shown in figure 4.4 on page 118.



Forward selection using AIC criterion.

Figure 4.4: Augmented Naive Bayes Network

In the same way as described for Naive Bayes classifier, we obtained the results for the augmented Naive Bayes Network in terms of predicted AAA class, log-densities for each AAA level. Again, all cases are predicted by the graphical model to belong to level 1 (normal cases) and the estimated area under the ROC curve is similar to the corresponding one for the Naive Bayes Classifier. As explained earlier, the problems with the classification command in MIM for these

models does not allow us to evaluate the performance of augmented Naive Bayes Networks in this statistical package.

4.4 Chain graphs

Edwards 2000 (page 203) states that "although problems with complete causal orderings [of the variables] seem to be fairly unusual in applications, partial orderings are often available. For example, an epidemiological study might involve the following characteristics of sample of individuals:

1. Familial characteristics, such as parental genotype.
2. Genetic characteristics, such as individual genotype.
3. Demographic characteristics, such as sex and ethnic group.
4. Social and economic factors, such as occupation, socioeconomic status, and educational background.
5. Lifestyle characteristics, such as tobacco use, diet, and physical exercise.
6. Biological characteristics, such as elevated cholesterol and body mass index.

Clearly, the familial characteristics are antecedent to the demographic characteristics, which themselves are antecedent to the lifestyle and biological characteristics. It may be reasonable to assume that the socioeconomic factors are antecedent to the lifestyle and biological characteristics, although this is clearly a nontrivial assumption". A good example of using chain graphs in epidemiology and an illustration of the advantages of the chain graph when compared with logistic regression can be found in Didelez et al. (2002).

Chain graphs, which have been described previously in detail, can be used to analyse data where possible partial orderings are available. Block recursive graphs, an alternative term in the literature for chain graphs are a combination of undirected graphs and directed acyclic graphs (DAGs). For the specifics about the dependence chain and Markov properties of chain graphs, the reader can find extensive description in Cowell et al. (1999) and Edwards (2000).

The variables in CASP data that have been included in Bayesian Network classifier previously (table 4.6 on page 115), can be also used to construct chain graphs to investigate possible influence and association structures of the variables. The objective of this part of the study is to estimate the joint distribution of the factors in the study.

It is also possible to implement the chain graphs mentioned above as classification models. In Michie et al. (1994), it is mentioned that "since the process of building the network does not take

into account the fact that we are only interested in classifying, we should expect as a classifier a poorer performance than other classification oriented methods. However, the built networks are able to display insights into the classification problem that other methods lack".

We will not attempt to label these chain graphs models as causal or intervention models as causal interpretation and the possible effect of intervention require additional modelling assumptions and possibly different formulation. Details about the use of chain graphs as causal and intervention models can be found in Lauritzen et al. (2002).

Modelling with chain graphs in MIM can be implemented by the undirected CG-distribution models (Edwards 2000). An exception of this rule is when there are discrete responses and continuous covariates, where we need to include all interactions between the covariates in all models under consideration. "When there are discrete response and continuous covariates, the CG-regression models ... can be used, but maximum likelihood estimation for these models is computationally more difficult" (Edwards 2000, page 209).

Initially we use age as discrete variable by using appropriate thresholds. A possible configuration can be derived by the age thresholds that have been implemented previously for logistic models for abdominal aortic aneurysm (Vardulaki et al. 2000). In this way, we will be able to implement chain graph modelling using maximum likelihood estimation without the need of using CG-regression models.

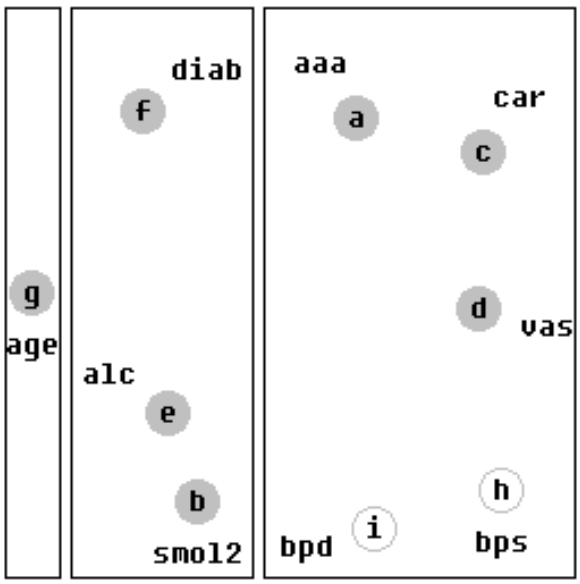
From the variables in table 4.6 on page 115, age can be regarded as possible determinant of all the other factors. At the same time, we assume that age can not be modified by any other factor available in the data. Hence, age might be seen as the parent node for all the variables in the chain graph model.

Moreover, smoking level (variable b denoted as smol2) is possibly affecting variables in the study such diastolic (variable i, bpd) and systolic (h, bps) blood pressure, the indicators of cardiac (c, car) and vascular (d, vas) complications and the presence of abdominal aortic aneurysm (a, aaa). Also, smol2 might be assumed to be on equal footing with alcohol consumption level indicator (variable e, alc) and diabetes indicator (diab, variable f).

For the group of variables bpd, bps, car, vas and aaa, it is not clear from the literature which variable can be regarded as cause or determinant of another variable in this group. With cross-sectional data as the one we use, we can also investigate the possible association between these variables. Thus, we put all the variables in this group on equal footing.

The block structure derived by the assumptions described above can be seen in figure 4.5 on page 121. Using this block structure, we start from a model assuming complete independence for

the variables within each block and for the variables in different blocks. In other words, there are no undirected edges between variables in the same block and no directed edges between candidate parent and children nodes in the chain graph.



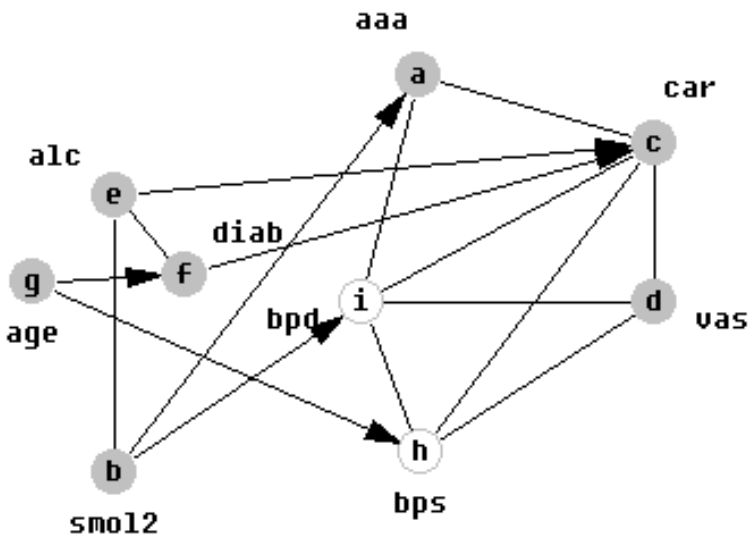
No association between variables (main effects model).

Figure 4.5: Block structure for chain graph

Subsequently, we use forward stepwise procedure to test significance of directed and undirected edges in the chain graph. AIC was used as fitting criterion and also unrestricted search was implemented to obtain a parsimonious model without being restricted to decomposable models. The result is shown in figure 4.6 on page 122.

From figure 4.6 on page 122, we can see that age has influence on the presence of diabetes and also systolic blood pressure. Furthermore, smoking level (smol2) is associated with alcohol consumption but not age and affects diastolic blood pressure and aortic aneurysm but not systolic blood pressure or cardiac or vascular complication indicators.

AAA indicator (variable a in the graph), which is the key variable of this study, is associated with cardiac but not vascular complications. Also, AAA is linked with diastolic but not systolic blood pressure, a result that confirms the level of significance of each blood pressure measurement for predicting AAA in logistic regression models.



Directed edges indicate influence and undirected edges state association between variables.

Figure 4.6: Forward stepwise AIC selection chain graph

Finally, smoking level is the only risk factor from those tested for influence on AAA indicator that is significant on the 5% significance level. Diabetes and alcohol consumption level are not significant determinants of AAA but are important for cardiac complications; this indicates the information that can be derived from the implementation of chain graphs is not restricted to a particular variable of interest.

Age, that has been used as a factor with 4 levels, is indicated as influential for the presence of diabetes and systolic blood pressure. In order to investigate the role of age on when included in a chain graph on continuous scale, it is necessary to apply CG-regression models that have been described previously.

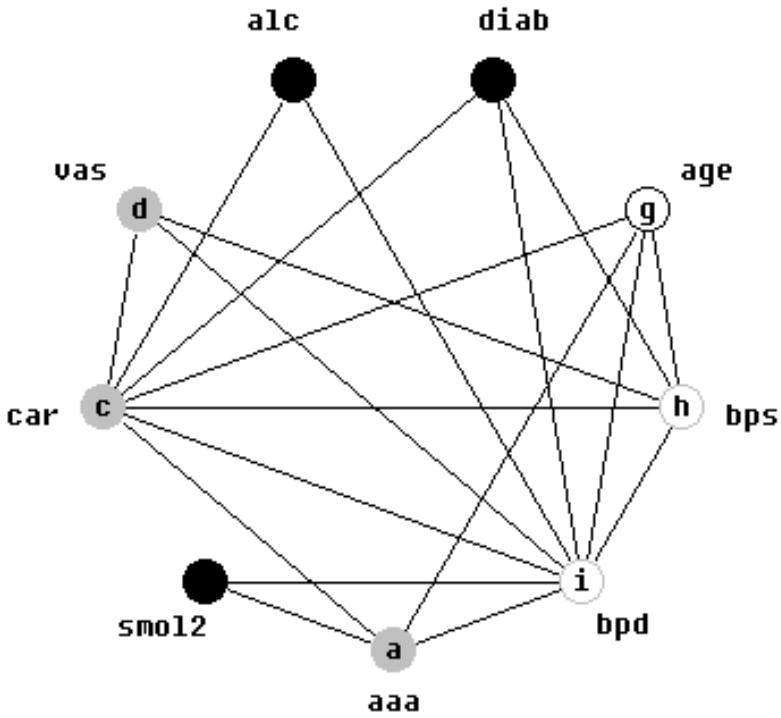
The ME-algorithm, the fitting procedure that CG-regression modelling is based on, "resembles the EM-algorithm closely but maximises the conditional rather than the marginal likelihood. In that and other senses it can be regarded as dual to the EM-algorithm" (Edwards 2000, page 315).

Additionally, the same author states that "for several reasons, when specifying a CG-regression model, it is best to use a full marginal model, that is to say, include all interactions between the covariates". In our data, it is possible to regard age, smoking level, diabetes and alcohol consumption as covariates and the rest of the variables as responses. All the edges between covariates are not eligible for removal and that is implemented in MIM by using the command *fix* (Edwards 2000).

Because of the large number of variables in the CG-regression model, convergence of the algorithm can be slow. A way to overcome this problem might be first to "select a preliminary undirected model, and then, as it were, fine-tune the analysis with CG-regression models, using the undirected model as a point of departure" (Edwards 2000). For this model, age, smoking level (smol2), diabetes and alcohol consumption are covariates and all the edges between them are "fixed" and not eligible for removal. AAA variable is associated with smoking level, cardiac impairment indicator, age and diastolic blood pressure.

CG-regression convergence occurs after 5 iterations when using the model selected by undirected stepwise regression (figure 4.7 on page 124). The CG-regression selected model is the same as the one chosen by the undirected selection procedure.

It is clear that using age on continuous scale and CG-regression (figure 4.7 on page 124) leads to different conclusion when compared to figure 4.6 on page 122 (maximum likelihood estimate using age as discrete variable). Thus, we can see that when estimating the joint distribution of the same variables, we need to be careful about the assumptions that we base on our conclusions as they might lead to completely different and sometimes contradicting results.



Age, smol2, diab and alc variables fixed as covariates.

Figure 4.7: Preliminary undirected model for CG-regression

Further information from the literature and from subject experts can indicate other possible block structures and subsequently different chain graphical models. These models can be compared by using likelihood ratio tests or information criteria differences. In addition to that, it is possible to estimate the predictive performance of the graph for a variable of interest such as the AAA variable which is the key variable in our study.

4.5 Occam's window model selection

In Madigan et al. (1994), specific drawbacks of stepwise selection are given. Specifically, the use of multiple tests and the comparison of non-nested models are two of the particular difficulties associated with stepwise selection. Furthermore, the use of p-values is described as controversial, even when there are only two models to be compared, because of the "conflict between p-values and evidence".

In addition to that, "perhaps most fundamentally, conditioning on a single selection model ignores model uncertainty and so leads to underestimation of the uncertainty about the quantities of interest" (Madigan et al. 1994). The approach used by these authors is to discard the models that "predict the data far less well than the best model in the class [of models of interest]".

To be specific, the models that are eligible for selection belong to the set

$$A' = \left\{ M_k : \frac{\max_l p(M_l|D)}{p(M_k|D)} \leq c \right\}$$

In the examples in Madigan et al. (1994), c=20 is used "by analogy with the popular .05 cutoff for p-values".

In Raftery (1995),

$$2\log\left(\frac{\max_l p(M_l|D)}{p(M_k|D)}\right)$$

is approximated by

$$BIC_k - BIC_l$$

where BIC_l is the minimum BIC corresponding to the maximum posterior probability. BIC stands for Bayesian Information Criterion and is also called Jeffreys-Schwarz criterion. It is equal to

$$-2 \times maximised\ Log - likelihood - 0.5 \times p \times log(N)$$

where p is the dimension of the model (number of free parameters) and N is the number of

observations.

Hence

$$\frac{\max_l p(M_l|D)}{p(M_k|D)} \leq c \Rightarrow \exp\left[-\frac{1}{2}(BIC_l - BIC_k)\right] \leq c \Rightarrow$$

$$-\frac{1}{2}(BIC_l - BIC_k) \leq \log c \Rightarrow BIC_l - BIC_k \geq -2\log c \Rightarrow$$

$$BIC_k - BIC_l \leq 2\log c$$

Subsequently, Madigan et al. (1994) use as criterion for reduction of the eligible models mentioned above "one of the most widely accepted norms of scientific investigation, Occam's razor. Let E represent the evidence and let $pr(H|E)$ represent the probability of a specified hypothesis H given the evidence E. Occam's razor states that if

$$pr(H_1|E) = pr(H_2|E) = ... = pr(H_k|E)$$

for hypotheses $H_1, ..., H_k$, then the simplest among $H_1, ..., H_k$ is to be preferred".

In this way, the models that should be excluded according to Occam's razor from set A' defined above, belong to set B where

$$B = \left\{ M_k : \exists M_l \in A', M_l \subset M_k, \frac{pr(M_l|D)}{pr(M_k|D)} > 1 \right\}$$

Thus, the set of acceptable models is A, where

$$A = A' \backslash B$$

Occam's window model selection is not currently available in MIM and we need to adapt the algorithm proposed by Madigan et al. (1994) according to the programming environment available. We will also need to combine MIM with a standard statistical package (in our case S-Plus version 4.5) to estimate quantities that can not be directly computed in MIM. Because there is a problem with the classification command in MIM, we need to use library e1071 in R for the Naive Bayes classifiers selected in MIM.

An example of this combination has been used previously to estimate the predictive performance of the Naive Bayes Classifier (NBC) and the augmented Bayes network by the area under the ROC curve. For Occam's window, we will demonstrate the implementation procedure for the Naive Bayes Classifier in figure 4.3 on page 116. In this case, we will attempt to find all the models that are acceptable by Occam's window procedure by removing edges between AAA and the rest of the

variables.

The first step is to remove an edge between AAA and one of the variables in block 2 and compute the deviance from the base model, the log-likelihood and the corresponding AIC (denoted as IC) and BIC by the exclusion of the specific edge. The results in MIM are shown in table 4.8 on page 127.

MIM classification model	Deleted edge	DF	AIC	BIC
af,ae,ad,ac,ab/ai,ah,ag/ai,ah,ag	-	1252	84479.803	84647.991
af,ae,ad,ac,b/ai,ah,ag/ai,ah,ag	ab	1255	84512.349	84662.516
af,ae,ad,c,ab/ai,ah,ag/ai,ah,ag	ac	1253	84505.638	84667.819
af,ae,d,ac,ab/ai,ah,ag/ai,ah,ag	ad	1253	84487.245	84649.426
af,e,ad,ac,ab/ai,ah,ag/ai,ah,ag	ae	1253	84477.978	84640.159
f,ae,ad,ac,ab/ai,ah,ag/ai,ah,ag	af	1253	84479.431	84641.612
af,ae,ad,ac,ab/ai,ah,g/ai,ah,g	ag	1254	84495.200	84651.374
af,ae,ad,ac,ab/ai,h,ag/ai,h,ag	ah	1254	84482.356	84638.530
af,ae,ad,ac,ab/i,ah,ag/i,ah,ag	ai	1254	84488.082	84644.256

Naive Bayes classifier (af,ae,ad,ac,ab/ai,ah,ag/ai,ah,ag) used as base model.

Table 4.8: MIM output for Occam's window selection procedure

We can see from the results that the model with minimum BIC after removing only one edge is af,ae,ad,ac,ab/ai,h,ag/h,ai,ag (after removing edge ah from the NBC). After that, we compare models by the formula

$$BIC_k - BIC_l \leq 2 \log c$$

and then set $c = 20$ (corresponding to 5% significance level). That means

$$BIC_k - BIC_l \leq 6$$

Thus, we compare the minimum BIC value (84638.530 for model af,ae,ad,ac,ab/ai,h,ag/h,ai,a) with the corresponding value for the other models. The Naive Bayes Classifier

af,ae,ad,ac,ab/ai,ah,ag/ai,ah,ag

with all the edges between AAA and the other variables is not an acceptable model as the difference in BIC is 9.461. In the same way, the models that are acceptable are shown below:

- af,e,ad,ac,ab/ai,ah,ag/ai,ah,ag BIC: 84640.159
- f,ae,ad,ac,ab/ai,ah,ag/ai,ah,ag BIC: 84641.612
- af,ae,ad,ac,ab/ai,h,ag/h,ai,ag BIC: 84638.530
- af,ae,ad,ac,ab/i,ah,ag/i,ah,ag BIC: 84644.256

The same procedure should be repeated for each of the four models shown previously until we find all the acceptable models. For simplicity, we assume that we only allow one edge to be removed for the NBC model, hence these four graphical models will be used to demonstrate the implementation of Occam's window in MIM.

First, we need to estimate the posterior probability of each model. In Raftery (1995), as we have previously seen,

$$2\log\left(\frac{\max_l p(M_l|D)}{p(M_k|D)}\right)$$

is approximated by $BIC_l - BIC_k$ where BIC_l is the minimum BIC corresponding to the maximum posterior probability. From this approximation, we can deduce that $p(M_k|D)$ is approximately equal to

$$\frac{\exp\left[-\frac{1}{2}(BIC_k - BIC_l)\right]}{\sum_k \exp\left[-\frac{1}{2}(BIC_k - BIC_l)\right]}$$

Equivalently, in Raftery (1995), $p(M_k|D) \propto \exp(-\frac{1}{2}BIC_k)$, thus

$$p(M_k|D) \approx \frac{\exp\left[-\frac{1}{2}(BIC_k)\right]}{\sum_{l=1}^{K} \exp\left[-\frac{1}{2}(BIC_l)\right]}$$

Using one of the two formulae mentioned above, the estimated posterior probabilities $p(M_k|D)$ for each selected model, denoted as p are:

- af,e,ad,ac,ab/ai,ah,ag/ai,ah,ag p: 0.258
- f,ae,ad,ac,ab/ai,ah,ag/ai,ah,ag p: 0.125
- af,ae,ad,ac,ab/ai,h,ag/h,ai,ag p: 0.584
- af,ae,ad,ac,ab/i,ah,ag/i,ah,ag p: 0.033

Subsequently, we estimate the probabilities for AAA variable by using each of the four selected models shown above. In this way, it is possible to have estimates from each model separately and also from the weighted average of the probabilities of each observation.

The formula applied to compute the weighted average of the log-densities can be found in Madigan et al. (1994). Specifically, the authors state that if "Δ is the quantity of interest, such as a parameter, a future observation, or the utility of a course of action, then its posterior distribution given data D, is

$$pr(\Delta|D) = \sum_{k=1}^{K} pr(\Delta|M_k, D)pr(M_k|D)$$

This is an average of the posterior distributions under each of the model, weighted by their posterior model probabilities".

For CASP data, we will estimate the area under the ROC curve using library e1071 in R by each model separately and their weighted average. The results are shown in table 4.10 on page 129.

MIM classification model	Omitted Edge	Weight	Mean AUC	AUC 95% C.I.
af,e,ad,ac,ab/ai,ah,ag/ai,ah,ag	ae	0.258	0.617	0.581, 0.653
f,ae,ad,ac,ab/ai,ah,ag/ai,ah,ag	af	0.125	0.617	0.582, 0.652
af,ae,ad,ac,ab/ai,h,ag/ai,h,ag	ah	0.584	0.617	0.581,0.653
af,ae,ad,ac,ab/i,ah,ag/ai,ah,ag	ai	0.033	0.611	0.575, 0.647
Weighted average	-	-	0.615	0.579, 0.650

Four submodels of Naive Bayes classifier and weighted average used.

Table 4.10: Results for Occam's window

From the results in table 4.10 on page 129, we conclude that the area under the ROC curve (AUC) for each of the four submodels of the Naive Bayes classifier are almost equal. The same can be said about the weighted average estimate.

In Cowell et al. (1999) say that the average of the posterior distributions should "give better predictions (as evaluated by means of the logarithmic scoring rule) than expected from the use of the a single model". Given this statement, the fact that the area under the ROC curve by the weighted average is less than all but one of the single models could be seen as unexpected. To the best of our knowledge, it is not clear whether statement above about the expected improvement of predictive performance applies when the area under the ROC curve is used instead of the logarithmic scoring.

When it is possible to investigate all possible models by Occam's window's rules (Madigan et al. 1994), the implementation in MIM can be performed by the command *select*. In Edwards (2000) it is stated that this command "uses a brute force approach, viz., each model in the class specified is fitted and the criterion calculated. This is only feasible for problems of small to moderate dimension".

For example, there are 256 submodels of the Naive Bayes Classifier model we described previously, hence it is possible to compute the BIC for each of the 256 models. Subsequently, we have two options described in Raftery (1995): strict or symmetric version of Occam's window. In the first case, we apply both rules whereas in the second case we apply one the first rule. Using the symmetric version of Occam's window, we find 6 acceptable models shown below:

- model:ab,ac,ad,e,f/ag,h,i/i,h,ag BIC: 84620.584 p: 0.502
- model: ab,ac,d,e,f/ag,h,i/i,h,ag BIC: 84622.020 p: 0.245
- model: ab,ac,ad,e,f/g,h,i/i,h,g BIC: 84623.968 p: 0.092

- model: ab,ac,ad,e,f/ag,ai,h/h,ai,ag BIC: 84624.319 p: 0.078
- model: ab,ac,d,e,f/g,h,i/i,h,g BIC: 84625.404 p: 0.045
- model: ab,ac,d,e,f/ag,ai,h/h,ai,ag BIC: 84625.755 p: 0.038

In a similar way, it is possible to a set of acceptable models by using the model selection strategy of Edwards and Havranek known as EH-search (Edwards 2000 and Madigan et al. 1994).

4.6 EM algorithm imputation

In previous sections, we have seen that it was necessary to recode information about smoking. Specifically, the data collector used variables smod (number of cigarettes per day) and lasm (years since last smoked) to record smoking related information. Because the data recorded has been regarded in some cases as unreliable, it has been recoded. The new variable, indicating level of smoking or type of reporting inconsistency is labelled smol (smoking level) and its distribution is shown in table 2.3, page 40.

After that, the two levels of unreliable or inconsistent smoking level have been merged into a single level to avoid inference problems when estimating graphical model parameters. The frequencies table for the new variable for smoking, labelled smol2, can be seen in table 4.2 on page 114. There are 438 individuals out of 3001 (14.6% of the sample) that have reported smoking related information that might be seen as inconsistent.

Another possible way to deal with the inaccurate smoking data entries is to regard them as missing and attempt to impute them by using other, probably more reliably recorded, variables. In other words, we discard the data that might be a source of bias for our results and replace it with the assumption that it is "inconsistent at random". This term, similarly to "missing at random" defined previously, implies that the data inaccuracy does not depend on the actual value that is recorded. The *EM-algorithm* is a possible method that can be applied to impute the inconsistent values for smoking by regarding them as missing at random.

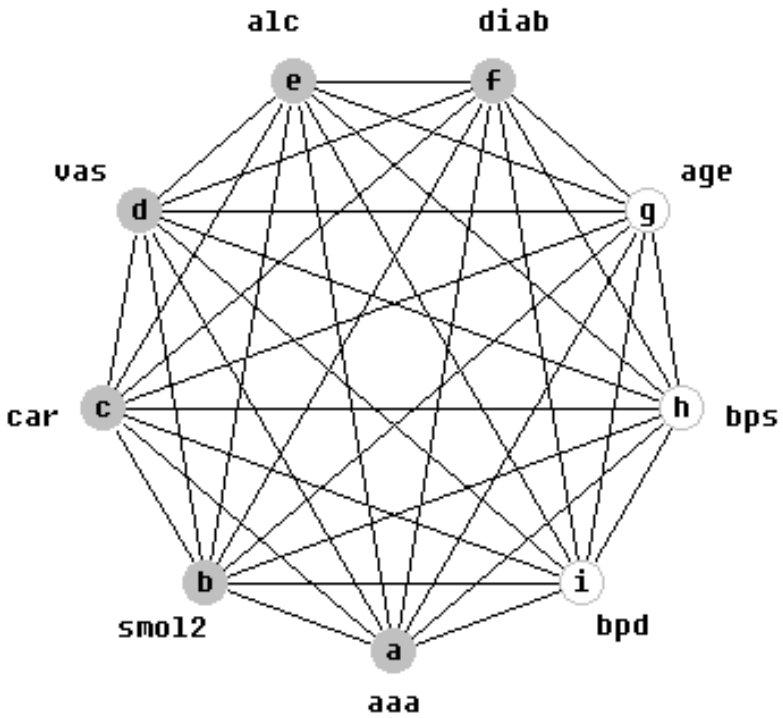
Schafer (1997) states that "EM [algorithm] capitalises on the interdependence between missing data Y_{mis} and parameters θ. The fact that Y_{mis} contains information relevant to estimating θ, and θ in turn helps us to find likely values of Y_{mis}, suggests the following scheme for estimating θ in the presence of Y_{obs} alone: 'Fill in' the missing data Y_{mis} based on an initial estimate of θ, re-estimate θ based on Y_{obs} and the filled-in Y_{mis} and iterate until the estimates converge".

Additionally, Edwards (2000) mentions that "each cycle in the [EM] algorithm consists of two steps: an E (expectation) step and an M (maximisation) step. In the E-step, expected values of the

sufficient statistics given the current parameters estimates and the observed data are calculated. In the M-step, new parameter estimates are calculated on the basis of the expected sufficient statistics using the ordinary algorithms". Further details about properties of the EM-algorithm, its convergence and the related topic of *data augmentation* can be found in Schafer (1997).

The implementation of EM-algorithm in MIM for is executed by using the command *EMFit* (Edwards 2000 page 261). Specifically, the current model is used and the expected likelihood and the corresponding change of likelihood at each step are printed. The procedure stops when the change is less that the convergence criterion. Missing values need to be replaced by the symbol "*" before entering the data into MIM.

For simplicity, we demonstrate this application of the EM-algorithm in MIM for the Naive Bayes classifier (NBC) described previously. Initially, we use the homogenous saturated model shown in figure 4.8, page 131 for imputation of the 438 inaccurate values for smoking level variable. Subsequently, we assess the performance of the NBC model with the inclusion of the imputed values for smoking as part of the learning and testing datasets.



All two-way interactions included.

Figure 4.8: Homogenous saturated model for EM-algorithm

The EM-algorithm, when using the homogenous saturated model, converges after 27 steps and the reported deviance (-2*Log likelihood in MIM) is equal to 79901.48 (MIM output not

shown). Table 4.12 on page 132 contains the distribution of smoking level variable (b, labelled smol2). From this table, we observe that the majority of missing values have been imputed at level 3 (heavy smokers). This might be an indication that the imputation model is not sensible for replacing the inconsistent entries in the data and the results based on this imputation should be used with caution as they might be misleading.

Smoking level code	Before imputation	After imputation
1	263	264
2	820	835
3	1480	1902
Total	2563	3001

438 missing observations in smol2 not shown in MIM output.

Table 4.12: smol2 distribution before and after EM imputation

The next step is to classify each individual in terms of belonging to level 1 (normal AAA) or 2 of variable a (AAA indicator). Using the NBC model in R library e1071, we compute the predicted probabilities for each individual in the study. The 95% confidence interval for the area under the ROC curve for the model described above is (0.558, 0.632).

Apparently, it is possible to use the same data we imputed above for other types of statistical analysis. These might be model selection using stepwise procedures, Occam's window method and EH-search. Additionally, we might decide that a model less complicated than the homogenous saturated graphical model is sufficient for imputation. Since we intent to assess the performance of Naive Bayes classifier (moralised undirected version), it might be sensible to implement the NBC as imputation model.

Convergence of the EM-algorithm in MIM occurs after 6 steps and the deviance at the final step is 81953.2346. In this case, all the missing values for smoking level variable b (smol2) have been imputed at level 3 (heavy smokers). The 95% confidence interval for the area under the ROC curve (0.566, 0.641) and it is similar to the results derived by homogenous model EM imputation.

Another way that we might implement to impute the missing or inconsistent information is multiple imputation. This method that has been described previously will be applied to replace 438 entries of smoking level variable (smol2). In this occasion, we are assuming that we have "inaccuracy at random", hence we might include AAA indicator and all the other factors in the study to impute the inconsistent entries. Also, we are treating the inaccurate values as missing and we do attempt to study specific patterns of misreporting of data and possible associations of inaccuracy with data structures.

For simplicity, we will implement multiple imputation (MI) under the assumption that the joint distribution of the variables in our sample is multivariate normal. Schafer (1997, page 147) states that "even if some of the incompletely observed variables are clearly non-normal, it may still be reasonable to use the normal model as a convenient device for creating multiple imputations".

Schafer (1997) also says that "for example, it may be quite reasonable to use a normal model to impute a variable that is ordinal (consisting of a small number of ordered categories), provided that the amount of missing data is not extensive and the marginal distribution is not too far for being unimodal and symmetric ...real data often do not conform to normality, and it is important to know whether the multiple-imputation procedures advocated in this book are robust to departures from the modelling assumptions".

Furthermore, Schafer (1997, page 148) says that when "using the normal model to impute categorical data, however, the continuous imputes should be rounded off to the nearest category to preserve the distributional properties as fully as possible and to make them intelligible to the analyst. We have found that the normal model, when used in this fashion, can be an effective tool for imputing ordinal and even binary data in instances where constructing a more elaborate categorical-data model would be impractical".

Currently, there are some statistical packages such as SAS and S-Plus that can be used to implement multiple imputation and the subsequent analysis by combining the results from each imputed dataset. Further details about inference used in the statistical packages mentioned above for a scalar quantity and multidimensional estimands of a statistical model based on multiply imputed datasets can be found in Schafer (1997).

In our case, we will implement MI analysis in Norm for Windows (Version 2.03, November 2000) by J. L. Schafer as we believe that it is likely to be used by statisticians and medical experts as an "easy to use" resource for deriving multiply imputed datasets. We need to emphasise that for the imputation of the inaccurate values of smoking level variable, there might be other more appropriate approaches.

In this instance, a categorical procedure (based on saturated multinomial and loglinear models) or mixed continuous and categorical method (based on the general location model) (Schafer 1997, p.399) might be preferable for imputation of smoking level variable. However, we will use the normal approximation as an illustration of a more general approach to the problem of multiple imputation and the possible advantages and disadvantages of imputing categorical data under normality assumptions.

In addition to that, there are occasions where the data analyst might not have access to specific

commercial statistical packages where multiple imputation procedures are available. Norm for Windows is a free package with complete documentation including brief explanation about each step of the procedure with examples. Another non-commercial package for implementing multiple imputation is R, which is "an Open Source statistics project" and "a conscious attempt to provide a high-quality environment for leading-edge statistics which is available to everyone" (Ripley 2002).

For illustration, we will analyse the results from multiple imputation for the Naive Bayes Classifier (NBC) described previously (figure 4.3 on page 116). From the graphical representation of the NBC, we conclude that smol2$\perp\!\!\!\perp$other predictors|AAA. Hence, the imputation model that is appropriate for the NBC graphical model might include only smoking level variable (smol2) and AAA (abdominal aortic aneurysm indicator).

On the other hand, it is possible that the statistical analysis after multiple imputation will include other types of graphical models such the Augmented Naive Bayes Network shown in figure 4.4 on page 118. In this case, variable (smol2) is associated with other predictors, hence it might be necessary to include other variables in the model. Also, in the case of applying forward stepwise procedures to test the significance of adding edges between variables in the graph, we might need to have all the possible factors in the imputation model.

Currently, in contrast to the imputation based on the EM-algorithm that has been shown previously, multiple imputation is not currently available in MIM (Edwards 2000). As with the implementation of Occam's window method (Madigan et al. 1994), we will combine procedures from different statistical packages to obtain the results that are required.

It is also important to say that for this case study, it is also possible to use multiple imputation of data that will be analysed by logistic regression and classification tree models for predicting abdominal aortic aneurysm indicator (AAA). The procedure is this case will be similar to the analysis using the NBC graphical model as in both cases, the area under the ROC curve will be the parameter estimated by multiple imputes.

First, we assume that multiply imputed datasets will be used only to derive result from the NBC model. In this occasion, we will use AAA indicator that is fully observed to replace the missing values of smoking level variable (smol2). In S-Plus, we construct a matrix with AAA indicator and smol2 and we replace entries at level 3 of smol2 by -3, which we will subsequently define as missing value code in NORM.

Furthermore, we need to extract the data, which in this case is a 3001×2 matrix, into a word processor, where we change the format of the entries in the file exported from S-Plus. As trivial as it might sound, reformatting the data so that it can be entered into another program is sometimes

the reason behind entry errors. This type of error might not be noticed by the data analyst and could lead into mistakes and derivation of wrong conclusions.

After that, we can obtain results for the missingness pattern and the variable distributions by the command *Summarise*. In this case, 15% of smoking level entries are regarded as missing and also the means and standard deviations of the variables are calculated with the assumption that they follow a Normal distribution. Subsequently, we estimate the means, variances and covariances of the data using the EM algorithm. Convergence is achieved after 8 iterations.

The next step depends on the number of imputations that are required for subsequent analysis. At this stage, it is possible that we need to implement exploratory analysis to choose the appropriate model or checking the possibility of transforming variables in the data. In Norm help file, it is suggested that "it can be quite useful to create one special imputation for exploratory purposes. First, run the EM algorithm to obtain maximum likelihood (ML) or posterior mode estimates of the model parameters. Then use the Impute from parameters procedure to generate one imputation under these estimates.

When a single imputation is generated in this manner, any quantities estimated from this dataset will tend to be close to estimates obtained from a set of m proper imputations. Exploratory or diagnostic plots produced from this dataset will be typical of plots created from any one of the imputed datasets. Statistical significance of effects will be somewhat overstated, however, because this single imputation will not properly reflect parameter uncertainty.

Model selection procedures applied to this data set will tend to detect all significant or important effects, and perhaps also some unimportant ones. If you use this one imputed dataset to select a model, you should then discard the results, refit the model to a set of m proper multiple imputations, and obtain proper estimates and standard errors using Rubin's rules for scalar estimands as implemented in NORMŠs MI inference for scalar estimands procedure".

Before imputing data, we need to implement Data Augmentation (DA), which is described in NORM as "an iterative simulation technique, a special kind of Markov chain Monte Carlo (MCMC). In DA there are three types of quantities: observed data, missing data, and parameters. The missing data and parameters are unknown. DA alternately performs the following steps:

I-step: Impute the missing data by drawing them from their conditional distribution given the observed data and assumed values for the parameters

P-step: Simulate new values for the parameters by drawing them from a Bayesian posterior distribution given the observed data and the most recently imputed values for the missing data.

Alternating between these two steps sets up a Markov chain that converges to a stationary

distribution, the joint distribution of the missing data and parameters given the observed data. DA bears a strong resemblance to the EM algorithm, and may be regarded as a stochastic version of EM. It is useful for multiple imputation of missing data.

By running DA for a large number of cycles, and storing the results of a few I-steps along the way (with enough cycles in between to ensure independence), one obtains proper multiple imputations of the missing data".

In our case, the diagnostics plots to investigate DA convergence can be seen in figure 4.9 on page 137. From the NORM outputs, we can deduce that convergence has been achieved, thus we can use the parameters from DA to impute data for exploratory analysis. Since we have decided that the model for subsequent analysis will be the NBC graphical model, it is not necessary to apply exploratory analysis to this dataset. More details about data augmentation and the corresponding convergence diagnostics can be found in Schafer (1997).

After that, we implement multiple imputation by data augmentation using a single chain of 5000 iterations. After every 1000 iterations, one data imputation is performed. From the results and the corresponding diagnostics (not shown), we can conclude that the chain has converged and the five imputed datasets are appropriate for subsequent analysis.

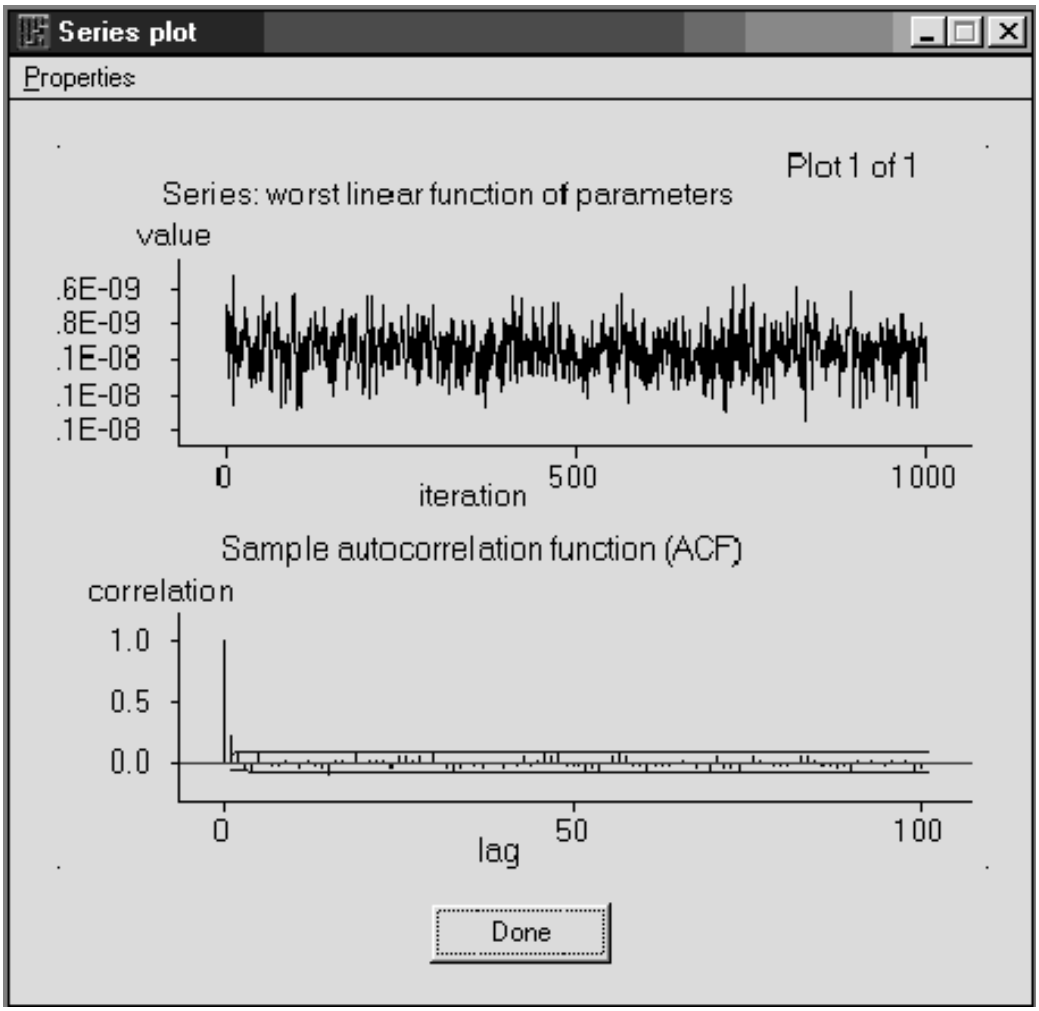
The following step after multiple imputation is to estimate the area under the ROC curve (AUC) for each imputed dataset. For this purpose, we need to import each dataset from NORM into S-Plus and then subsequently combine MIM and S-Plus so that we will estimated the AUC for the Naive Bayes Classifier (NBC). At this stage, we remind the reader that it is possible to have other graphical models for classification and we use the NBC for simplicity.

The estimates for the AUC by the NBC for each imputed dataset are shown in table 4.14, page 136. We can see that the area under AUC estimates for different models are not substantially different, hence we expect the combined AUC should be similar to any of the area under the AUC estimates shown in the table mentioned above.

Imputation model	Mean area under AUC	95% conf. interval
1	0.556	(0.519, 0.592)
2	0.556	(0.519, 0.593)
3	0.550	(0.514, 0.587)
4	0.558	(0.521, 0.595)
5	0.555	(0.518, 0.592)

Table 4.14: Area under ROC curve for each imputed dataset

In Schafer (1999), the formulae for obtaining combined estimates from multiply imputed datasets



Parameters used for exploratory analysis imputation.

Figure 4.9: Convergence diagnostics for DA augmentation

are given. Specifically, these formulae are related to the repeated-imputation inference for a quantity Q that denotes "a generic scalar quantity to be estimated, such as a mean, correlation, regression coefficient, or odds ratio. Let Y denote the intended data, part of which is observed (Y_{obs}) and part of which is missing (Y_{mis}).

Let $\hat{Q} = \hat{Q}(Y_{obs}, Y_{mis})$ denote the statistic that would be used to estimate Q if complete data were available, and let $U = U(Y_{obs}, Y_{mis})$ be its squared standard error. We must assume that with complete data, tests and intervals based on the normal approximation

$$\frac{\hat{Q} - Q}{\sqrt{U}} \sim N(0,1)$$

would be appropriate".

For the area under the ROC curve (AUC), it is possible to have normal approximation by subtracting 0.5 from the mean estimate. In other words, if the scalar quantity Q is the mean AUC and U the AUC squared standard error, then

$$\frac{\hat{Q} - 0.5}{\sqrt{U}} \sim N(0,1)$$

In addition to that, Schafer (1999) says that in "the absence of Y_{mis}, suppose that we have $m > 1$ independent simulated versions or imputations $Y_{mis}^{(1)} \ldots Y_{mis}^{(m)}$. From these we calculate the imputed-data estimates $\hat{Q}^{(l)} = \hat{Q}(Y_{obs}, Y_{mis}^{(l)})$ along with their estimated variances $U^{(l)} = U(Y_{obs}, Y_{mis}^{(l)}), l = 1, ..., m$. The overall estimate of Q is simply the average

$$\overline{Q} = m^{-1} \sum \hat{Q}^{(l)}$$

To obtain a standard error for $\overline{Q}$, one must calculate the between-imputation variance

$$B = (m-1)^{-1} \sum (\hat{Q}^{(l)} - \overline{Q})^2$$

and the within-imputation variance

$$\overline{U} = m^{-1} \sum U^{(l)}$$

The estimated total variance is

$$T = (1 + m^{-1})B + \overline{U}$$

and tests and confidence intervals are based on Student's t-approximation

$$\frac{\overline{Q} - Q}{\sqrt{T}} \sim t_\nu$$

with degrees of freedom

$$\nu = (m-1)\left[1 + \frac{\overline{U}}{(1+m^{-1})B)}\right]^2$$

Notice that if Y_{mis} carried no information about Q, then the imputed-data estimates $\hat{Q}^{(l)}$ would be identical and T would reduce to $\overline{U}$. Therefore

$$r = \frac{(1+m^{-1})}{\overline{U}}$$

measures the relative increase in variance due to missing data, and the rate of missing information in the system is approximately

$$\lambda = \frac{r}{1+r}$$

A more refined estimate of this fraction ... is

$$\lambda = \frac{r + 2/(\nu+3)}{1+r}"$$

In our case, we combine the estimates table 4.14, page 136, first by subtracting 0.5 from the mean estimates and the compute the combined AUC in NORM. The output is shown in table 4.16, page 139. We need to remind the reader that before combining the results using *MI inference:Scalar* in NORM, we have subtracted 0.5 from all mean estimates. Hence, before reporting the results as shown in table 4.16 on page 139, we have added 0.5 to all the mean estimate and to lower and upper limit of the 95% confidence interval (95% L.L. and 95% U.L. respectively).

Mean	St. Er.	T-ratio	DF	p-value	95% L.L	95% U.L.	% mis
0.555	0.0191	2.89	4868	0.0039	0.518	0.592	2.9

Naive Bayes Classifier (NBC) used for imputation and analysis

Table 4.16: MI inference estimates

The results in table 4.16, page 139 indicate that the estimated area under AUC are similar, as expected, to the corresponding from each imputed dataset. Furthermore, from the p-value in the table mentioned above, it can be deduced that the estimated area under the AUC is significantly different from 0.5, hence the NBC model is not completely worthless for classification.

After that, we have imputed smol2 inconsistent entries by including all other factors in NBC graphical model. The distribution of variable smol2 of each imputed dataset indicated that the smol2 imputed values are similar to the ones imputed from the multiple imputation with AAA indicator only (results not shown). Thus, we might be able to assume that the AUC derived by the NBC model would be almost identical under the assumption the both imputation models are sensible options for this task they have been implemented.

Schafer (1999) states that multiple imputation (MI) "is not the only principled method for handling missing values, nor is it necessarily the best for any given problem. In some cases, good estimates can be obtained through a weighted estimation procedure ...In fully parametric models, maximum-likelihood estimates can often be calculated directly from the incomplete data by specialised numerical methods, such as the EM algorithm. The estimates obtained through such procedures may be somewhat more efficient than those from MI, because they involve no simulation.

Given sufficient time and resources, one could perhaps derive a better statistical procedure than MI for any particular problem. In real-life applications, however, where missing data are a nuisance rather than a major focus of scientific enquiry, a readily available, approximate solution with good properties can be preferable to one that is more efficient but problem-specific and complicated to implement. MI is not the only tool available, but it is a handy one and a valuable addition to any data analyst's toolkit".

If the purpose of multiple imputation described above is to estimate the parameters of homogenous and heterogenous graphical models (Edwards 2000), then MI under the assumption that the variables joint distribution can be a multivariate Normal is highly unlikely to be appropriate. For the homogenous model, it is possible to impute the data in S-Plus or any other statistical package by the general location model (Schafer 1997).

Finally, in the occasion of having chain graphical models, when using MI, it is essential to take into account the temporal and causal structure of the model. Hence, it might be beneficial for the person that will impute the data, to have a model that will be related as far as possible to the analyst's most likely models.

Schafer (1999) states that when "the imputer's model is more general (i.e. makes fewer assumptions) than the analyst's, then MI leads to valid inferences with perhaps some loss of power, because the additional generality may add extra variation among the imputes $Y_{mis}^{(1)}, ..., Y_{mis}^{(m)}$. When the imputer makes more assumptions than the analyst - and the extra assumptions are plausible - then the MI estimate $\overline{Q}$ may become more precise than any estimate derived from the observed

data and analyst's model alone, a property that [is called] 'super-efficiency'. In such cases, MI intervals tend to be narrower than intervals derived purely from the analyst's model, and they also tend to be conservative with higher-than-nominal coverage probability.

The only serious danger of inconsistency arises when the imputer makes more assumptions than the analyst and these additional assumptions are unwarranted. For example, consider a situation where a variable is imputed under a no-interactions regression model and the analyst subsequently looks for evidence of interactions; if interactions are present, then the MI estimates of them will be biased toward null values.

In practice, this means that an imputation model should reasonably preserve those distributional features (e.g. associations) that will be the subject of future analyses. Above all, the processes of imputation and analysis should be guided by common-sense".

4.7 Aortic diameter as mixture of components

In previous section, we have used the aortic diameter, which is the key variable of this study as categorical variable. We have mainly applied the threshold of 29 mm to separate the individuals in the sample into normal and abnormal AAA cases, thus creating a binary variable labelled AAA indicator.

Furthermore, we have included the cut-off point of 40 mm and the result is a factor with three levels indicating low, moderate and high risk of aortic rupture. Another way of dividing the sample into clinically important groups is the application of age-related thresholds for the aortic diameter.

As we have previously mentioned, one of the important questions of this case study is whether aortic diameter can be described as a simple mixture of normal distributions. The first step was to investigate the distribution of aortic diameter by simple visual inspection of the corresponding Q-Q plot (figure 2.4 on page 37).

The conclusion from the investigation mentioned above is that the aortic diameter does not follow a Normal distribution. In addition to that, smooth transformations such the logarithmic or square root do not achieve normalisation of the aortic diameter distribution. The fact that in the Q-Q plot we have a straight line segment for the majority of cases and then a separate line segment for the other individuals might be an indication that the aortic diameter distribution is a mixture of at least two components.

In Edwards (2000, page 109) an example of separating a distribution into two components is given. In this example, it is assumed that we have two normal distributions with common variance. Specifically, it is assumed that the observed values "have been drawn from a density of the form

$$f(y) = p_1(2\pi(\sigma^2))^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y-\mu_1)'(\sigma^2)^{-1}(y-\mu_1)\right\} +$$

$$(1-p_1)(2\pi(\sigma^2))^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y-\mu_2)'(\sigma^2)^{-1}(y-\mu_2)\right\}$$

where p_1 is the proportion of the first component, and the components are normal densities with means μ_1 and μ_2 and common variance σ^2. This is known as the two-component mixture model. We wish to estimate these parameters and, if possible, to test whether $\mu_1 = \mu_2$, i.e., whether the data are adequately described with one component" (Edwards 2000).

Furthermore, a discrete latent variable is included in the graphical model by declaring in MIM a factor with two levels and all its values missing. The EM-algorithm is applied to impute the missing values of the latent variable and after the algorithm has converged, the maximum likelihood estimates of the parameters are derived. Moreover, "the probability for each observation of deriving from the first or second component can be obtained by using the *Summary* command [in MIM]" (Edwards 2000, page 112).

Finally, it is also possible in MIM to test the adequacy of the mixture model and to perform a likelihood ratio between the model with and the model without an edge between the latent and the observed variable mentioned above. For the example described in Edwards (2000, pages 109-112), it is pointed out that the inference from MIM results in this case is wrong.

Specifically, two reasons are given for having the wrong inference. "First, there are two fewer parameters in the simpler model, not one: the parameter p_1 is inestimable, but this is not detected by MIM, so the degrees of freedom [for the likelihood ratio test] should be two, not one. Secondly, the likelihood ratio does not have an asymptotic χ^2 distribution" (Edwards 2000, page 112).

For the aortic diameter in our sample, we adopt a similar strategy to identify the mixture components of the distribution. Initially, we examine whether the aortic diameter distribution can be separated into two components. Each of the components is assumed to follow a Normal distribution and the two distributions have the same variance.

The EM-algorithm in MIM, when started with random values in MIM by the command *emfit*, is possible to show no change of the log likelihood after converging in a small number of steps. In Edwards (2000, page 110), it is mentioned that in a similar situation that also occurred in the example that we might "suspect that [the log likelihood] was very flat at the starting point. We try some other starting points by repeating the command [*emfit*] a few times".

Convergence occurs for the model with two normal components for the aortic diameter after 38 steps. After that, we examine the output of the EM-algorithm to derive the estimated parameters

for the mixture of normal components. Moreover, we use a likelihood ratio to examine whether one component is adequate. The results from MIM are in table 4.18 on page 143.

Component	Mean	Variance	Count
1	22.413	42.203	1484.777
2	22.042	42.203	1516.223

Mixture of two normal components with equal variance assumed.

Table 4.18: Aortic diameter of two components in MIM

From MIM output in table 4.18 on page 143, we can see that the two components have almost equal means and almost the same number of individuals in each component. Furthermore, from the likelihood ratio test, the null hypothesis that one component is adequate cannot be rejected. Hence, the aortic diameter cannot be separated into two normal components with common variance.

The next step is to assume that there are three normal components with common variance. In a similar way to the model with two components shown previously, we fit a graphical model with a latent factor with three levels and all its values missing. The EM-algorithm converges after 57 steps and the results for the estimated components can be seen in table 4.20 on page 143.

Component	Mean	Variance	Count
1	49.141	17.194	100.272
2	21.295	17.194	1946.501
3	21.295	17.194	954.226

Mixture of three normal components with equal variance assumed.

Table 4.20: Aortic diameter with three components in MIM

The estimated components parameters in table 4.20 on page 143 indicate that two of the three normal components have identical means. In this occasion, we might conclude that two components are actually identified. This can be confirmed by imputing the component variable (b variable) level for each individual in the sample. In addition to that, the likelihood ratio test indicates that one component hypothesis is rejected, thus the aortic diameter can be separated into three components.

From the results shown above, it is not clear why two of the three components are identical and that after imputation, one of two identical components contains no individuals. A possible explanation can be that the assumption of common variance is not valid and restricts the model search to homogenous graphical models. From the histograms of the components separated by imputation (not shown), it is clear that none of the two normal components follow a Normal

distribution.

After that, we assume that a mixture of two normal components with unequal variances might be appropriate. Specifically, we make use of a heterogenous graphical model with a two level latent factor with all its value missing and subsequently estimated by the EM-algorithm. The density function of the mixture has the form

$$f(y) = p_1(2\pi(\sigma^2)_1)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y-\mu_1)'(\sigma^2)_1^{-1}(y-\mu_1)\right\} +$$

$$(1-p_1)(2\pi(\sigma^2)_2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y-\mu_2)'(\sigma^2)_2^{-1}(y-\mu_2)\right\}$$

where p_1 is the proportion of the first component, and the components are normal densities with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 respectively.

Convergence occurs after 26 steps and the deviance (-2* log likelihood in MIM output) is equal to 16772.663. Furthermore, we can see that in table 4.22 on page 144 that the variance of the first component of the mixture (138.237) is much larger than the corresponding variance of the second component. Also, by imputation of component indicator variable b, we derive that 287 cases belong to the first component.

Component	Mean	Variance	Count
1	34.277	138.237	358.656
2	20.590	6.818	2642.344

Mixture of two normal components with unequal variance assumed.

Table 4.22: Aortic diameter with two components using heterogenous model

In addition to that, the individuals imputed into the first component of the mixture shown above have aortic diameters at least 28 mm and the cases in the second component corresponding diameters at most 27 mm. This might be regarded as indication of separation of the sample into two distinct subsamples and the cut-off point is 28 mm. The threshold for abnormal aorta in this study is 29 mm, which shows that the two normal components found for the aortic diameter distribution can be said to be in agreement with current clinical practice.

Similarly to the model with two normal components with unequal variances, we repeat the analysis to investigate whether more components are required. For the three components, convergence occurs after 131 steps and the deviance is 16635.886. Compared to the model with two components, the deviance between these models (16772.6630-16635.886=136.777) is found to be significant by using a likelihood ratio test with 3 d.f. (p-value<0.001).

The estimated parameters for a mixture of three components are shown in table 4.24 on page 145. In addition to that, by implementing imputation for the component indicator (variable b), the individuals in the sample are separated by their aortic diameter into three distinct groups. The cut-off points in this occasion are 27 mm and 39 mm, another indication that the results can be said to be in agreement with the clinical thresholds for separating individuals with low from moderate risk of aortic rupture (29 mm) and moderate from high risk (41 mm).

Component	Mean	Variance	Count
1	45.148	175.511	117.801
2	20.378	5.977	2528.157
3	27.779	26.713	355.042

Mixture of three normal components with unequal variance assumed.

Table 4.24: Aortic diameter with three components using heterogenous model

A heterogenous graphical model has been implemented for examining the possibility of having four components. The EM-algorithm converged after 1487 steps and the deviance is 16612.0674. Compared to a model with three normal components, it is significantly different, hence we might consider the possibility of having an additional subgroup in the study, not currently used in clinical practice. The estimated parameters are shown in table 4.26 on page 145. The thresholds derived by imputation for discriminating individuals according to their aortic diameter are 23 mm, 28 mm and 41 mm.

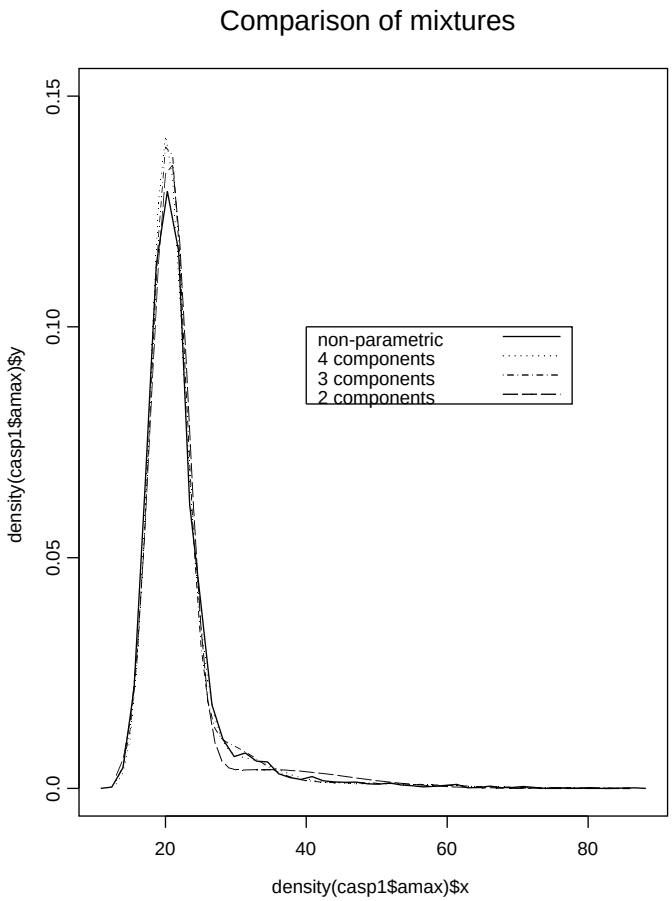
Component	Mean	Variance	Count
1	19.718	4.534	1903.577
2	22.700	6.580	757.033
3	30.316	27.837	240.197
4	46.881	176.592	100.193

Mixture of four normal components with unequal variance assumed.

Table 4.26: Aortic diameter with four components using heterogenous model

A possible way to assess the differences between the mixture models with unequal variances is to compare their density functions with the non-parametric density of aortic diameter (Everitt et al. 2001). The plot, shown in figure 4.10 on page 146, indicates that the mixture models density functions are similar to each other and to the non-parametric density function.

Hence, by the plot and the likelihood ratio tests mentioned above for the aortic diameter, we might conclude that the assumption that there two distinct populations in the sample seems likely to be correct. Also, it is more likely that there are three or even four subgroups that can be



Comparison of mixture models with non-parametric density function.

Figure 4.10: Mixture of components for aortic diameter

identified and the current clinical thresholds seem to be in agreement with the findings of the analysis.

Finally, we should mentioned that the results derived by graphical models with latent variables can be confirmed by a minimisation procedure available in S-Plus library MASS. In Venables et al. (1998, page 287) an example of fitting a mixture model with two components with unequal variances is given.

For our dataset, we applied a similar procedure for a model with three components and having as starting values of the minimisation procedure, the estimated means and variances for each risk of aortic rupture group. The results, shown in table 4.28 on page 147 are almost identical to the results found in MIM (table 4.24 on page 145).

Component	Mean	St. Deviation	Proportion
1	20.3785	2.44562	0.843
2	27.7998	5.16765	0.118
3	45.166	13.2472	0.039

Minimisation S-Plus procedure output.

Table 4.28: Three unequal variances normal components for aortic diameter

In Edwards 2000 (page 103), it is mentioned that "a wide variety of latent variable problems, including mixture models, latent class models, factor analysis-type models, and other novel models can be handled in this manner". For example, for social sciences, it is possible to examine whether "the manifest variables are conditionally independent given one or more latent variables" (Edwards 2000, page 108) by fitting an appropriate graphical model.

Additionally, latent variable graphical models can be implemented for measurement error and misclassification problems. In the case of aortic diameter and blood pressure in our data, the inaccuracy of the devices used for measurement indicate a potential investigation using latent variable models.

Also, doubts about smoking level or the presence of co-morbidities can be handled by the models mentioned above. The assumption in this case is that there is information available to estimate misclassification error for different patterns of inaccuracy.

Hence, it is important to apply latent variable graphical models for problems such as identifying the components of aortic diameter. Other options that require further investigation is the possibility of having a mixture of other distributions for the aortic diameter such as Poisson or Gamma distributions.

The results should be regarded with caution and should always be examined for statistical and

clinical validity. In the analysis shown above, it might be concluded that the proposed mixture models seem to give similar separation into aortic diameter subgroups as the clinical thresholds that have been used in other types of analysis applied in this case study.

4.8 Aortic diameter growth models for mixture of components

From the results for the possible mixtures of normal components of aortic diameter described above, we might conclude that there distinct groups of patients in the population. It is not clear though from all the statistical methods shown previously how such groups might be formed. In other words, we need to identify a causal mechanism for the mixture of components, which will be linked to aortic diameter growth and the probability of rupture.

In the data we have analysed, aortic measurements and other related factors are only included for the first visit. Hence, we are not able to construct growth models using longitudinal data by implementing one of the statistical methods currently available. Nevertheless, we might be able to combine results from the literature to define possible ways of aortic diameter growth for each of the mixture components found previously.

Vardulaki et al. (1998) estimated growth rates by initial diameter, age and sex by a random effects model. Furthermore, the same authors say that "the logarithms of individual growth rates in the population were assumed to be normally distributed with an unknown mean μ and unknown variance σ^{2}". Additionally, estimates of growth rate were derived for each health centre in the study for different initial aortic diameter bands or age bands.

It might be possible to construct aortic diameter growth models for our dataset by assuming that the growth parameters estimated in Vardulaki et al. (1998) would be similar. Moreover, we might need to compute our parameter estimates for different groups than in Vardulaki et al. (1998), thus it is not a straightforward implementation.

To be specific, for the mixture of three normal components shown previously, we might assume that there different growth patterns for each of these components. We denote aortic diameter of each individual in the study by Y and the time that elapsed since initial diameter measurement by t. Each of the components Y_i, i=1,2,3, follows a Normal distribution $N(\mu_i, \sigma_i^2)$ with probability p_i. The estimates for each of the three components of the mixture mentioned above are shown in table 4.24 on page 145 and in table 4.28 on page 147.

At this stage, we need to make some assumptions about each of the components of the mixture

described above. In particular, we might say that for the majority of cases, the mean and the variance of the distribution will not be changed as time after initial aortic measurement t increases. Thus, for component Y_1, the parameter estimates $p_1 = 0.85$, $\mu_1 = 20$ and $\sigma_1 = 2.5$ are fixed $\forall t$.

For the second component Y_2, the parameters will be allowed to change with time t and we assume that aortic diameter growth occurs similarly to the group of cases at Chichester with initial diameters 30-39 mm (Vardulaki et al. 1998). Using the estimated growth percentage for the Chichester group mentioned above, we propose that for component Y_2 with $p_2 = 0.1$, the corresponding mean and standard deviation are

$$\mu_2 = \alpha_{20} + \alpha_{21} t = 28 + 0.55t$$

and

$$\sigma_2 = \beta_{20} + \beta_{21} t = 5 + 0.15t$$

In the same way as for component Y_2, the third component Y_3 with associated probability $p_3 = 0.05$ can be said to be affected by increasing time after initial diameter t. Using the average growth percentage of Chichester groups with initial diameters 40-49 mm and $\geq$50 mm, we have

$$\mu_3 = \alpha_{30} + \alpha_{31} t = 45 + 2.25t$$

and

$$\sigma_3 = \beta_{30} + \beta_{31} t = 13 + 0.5t$$

From the results shown above, it is clear that the majority of individuals under investigation, we assume that the aortic diameter growth in negligible and the risk of aortic rupture is small. Thus, the individuals in this group will not be included in further screening tests for abdominal aortic aneurysm. On the other hand, the other two groups are expected to have aortic growth that depends on the time that has passed since initial diameter.

Moreover, it is possible that some of the individuals will shift from component Y_2 to component Y_3 when the diameter has grown to such an extent that the individuals have entered the final stage of aortic diameter growth. This is in agreement with Vardulaki et al. (1998), where it is mentioned that "the growth rate of AAAs less than 50 mm in diameter is substantially lower than the rate observed in those greater than 50 mm".

Clinical factors such as age and smoking level might be possible modifiers of the aortic growth pattern. In Vardulaki et al. (1998), the individuals that belong to the age band 70-74 years have

approximately 1% additional growth of their aorta when compared with the cases with ages 65-69. Based on the potential influence of age described above, we define an indicator I that takes value 1 when age is at least 70 years and 0 otherwise and we include that in the proposed growth models.

For component Y_1, we assume that the corresponding mean and standard deviation μ_1 and σ_1 are not affected substantially by time increase after initial aorta screening test t or aging (expressed by indicator I described above). For components Y_2 and Y_3, the relationships between growth parameters and factors t and I, assuming that I affects only the mean of the components are:

$$\mu_2 = \alpha_{20} + \alpha_{21}t + \alpha_{22}I = 28 + 0.55t + 0.3I$$

$$\sigma_2 = \beta_{20} + \beta_{21}t = 5 + 0.15t$$

$$\mu_3 = \alpha_{30} + \alpha_{31}t + \alpha_{32}I = 45 + 2.25t + 0.45I$$

$$\sigma_3 = \beta_{30} + \beta_{31}t = 13 + 0.5t$$

In the same way, it is possible to include smoking level effects into the growth pattern as well as effects related to other variables such as blood pressure and diseases indicating cardiac and vascular complications. A study that includes information about risk factors over time might be able to help into understanding the effect of each of the factors on the aortic growth pattern.

Moreover, the probability of rupture and the potential interventions for reducing this probability might be also said to be affected by the way that the aorta grows over time. Hence, the use of graphical models and more specifically chain graphs might be proven a suitable way of modelling change of the aorta over time and the identifications of the reasons for which an individual with normal aorta is shifted to the abnormal aortic diameter group.

In order to investigate the benefit of abdominal aortic aneurysm screening tests that are implemented to identify individuals with high risk of aortic rupture, we need to find the threshold that is optimum in terms of benefit. For simplicity, we assume that the misclassification error rate should be minimised and also that growth depends only on the initial diameter. In addition to that, we suppose that the risk of rupture depends only the aortic diameter value (Vardulaki et al. 1998).

For example, five years after the initial screening tests, the majority of cases that belong to the component Y_1 are not affect by aortic growth. On the other hand, using the equation for growth given above, component Y_2 parameters will change from $\mu_2 = 28$ and $\sigma_2 = 5$ to $\mu_2 = 30.75$ and

$\sigma_2 = 5.75$. In the same way, Y_3 will shift from $N(45, 13)$ to $N(56.25, 15.5)$.

Suppose that at this stage, we need to identify the individuals with diameter at least 50 mm five years after the initial screening. Then, from the Normal distributions shown above, the probability that an individual that belongs to either component Y_2 or Y_3 has diameter at least 50 mm can be easily computed. In this instance, for $Y_2 \sim N(30.75, 5.75)$, $p(Y_2 \geq 50) = 0.0004$, whereas for $Y_3 \sim N(56.25, 15.5)$, $p(Y_3 \geq 50) = 0.6566$.

Assuming that Y_3 is the part of the population we would like to identify, a screening test becomes a problem of separating individuals in Y_3 from the other cases at the initial screening stage. For the initial screening test, $p_2 = 0.10$, $Y_2 \sim N(28, 5)$, $p_3 = 0.05$ and $Y_3 \sim N(45, 13)$. The misclassification error rate for threshold t will be

$$mer = p_2 \times p(Y_2 > t) + p_3 \times p(Y_3 \leq t)$$

The plot of aortic diameter thresholds against the estimated misclassification error rate is shown in figure 4.11 on page 151. In this case, the minimum misclassification error rate is achieved when the aortic diameter threshold of 38 mm is used. Similarly, it is possible to find the optimum screening threshold for other criteria, for example the area under the ROC curve or a utility function that incorporates different misclassifications costs.

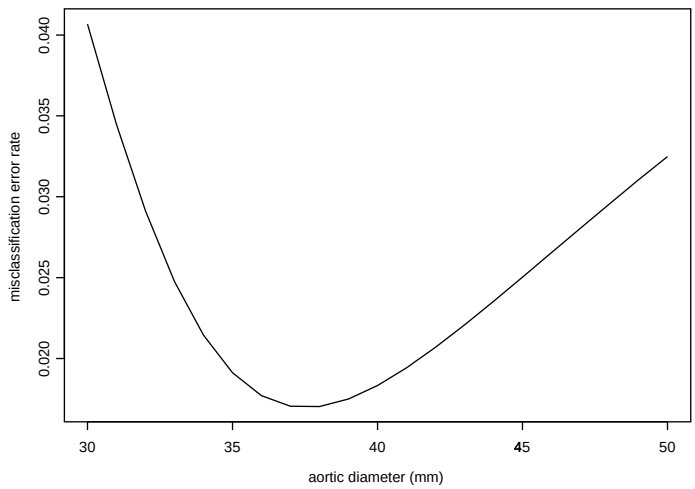


Figure 4.11: Misclassification error rate for each aortic diameter threshold

4.9 Conclusions

In this chapter we presented the application of graphical models using data from an aortic aneurysm screening project. We considered different types of graphical models for classification as well as for identifying the influence and association structure of the variables. Different relationships between variables when age groups have been replaced by age on continuous scale and chain graphs are replaced by CG-regression.

The Naive Bayes classifier (NBC) has been tested as selective screening model for our AAA study. The NBC models can be used for the objective they have been tested even though the logistic models in the previous chapter achieve larger area under the ROC curve. The classification command in the software MIM we have used gives unexpectedly wrong results and we use library e1071 in package R instead.

Occam's window model selection has been implemented where several models are considered simultaneously. The EM-algorithm and multiple imputation method for graphical model classification have been explored to deal with missing entries in the dataset. In both cases, it is possible that the imputation model is inappropriate for replacing inaccurate measurements of smoking level.

Finally graphical modelling has been implemented to investigate whether the aortic diameter can be described as a mixture of normal distributions. The mixture with three components and unequal variances both using the EM-algorithm in MIM and minimisation procedure in S-Plus confirm the current clinical thresholds for aortic diameter risk levels. Finally, the growth model for aortic diameter identified 38mm as the diameter threshold that gives minimum misclassification error rate.

Chapter 5

Case study II: Diabetic retinopathy

5.1 Abstract

In this chapter, we present a case study to illustrate the use of different statistical methods to identify risk factors related to sight threatening retinopathy (STR) at first visit of diabetic patients to the eye clinic. We find that gender, smoking level and body mass index are not statistically significant as risk factors of STR but nevertheless should be included as control variable as they are variables of "known clinical importance". Bivariate and multivariate exploratory analysis show the complexity of the association between systolic blood pressure and duration of diabetes. Different blood pressure measurements have been tested as determinants of the presence of STR. We show that the imputation of missing values is crucial by comparison of "complete observations only" analysis and "hot-deck imputation". Finally, we used graphical modelling to allow for complex relationships between the variables in the data.

5.2 Medical background

5.2.1 Diabetes mellitus

Diabetes mellitus is "a condition in which there is a chronically raised blood glucose concentration. It is caused by an absolute or relative lack of the hormone insulin, i.e. insulin not being produced from the pancreas or there is insufficient insulin or insulin action for the body's needs" (Williams et al. 1999, page 2).

A description of a disease similar to diabetes is given in Ebers papyrus from Egypt (1550 BC) and has been also noted by Indian physicians in the fifth and sixth centuries AD (Rudnicka et

al. 2000). "The word 'diabetes' comes from the Greek [word ΔIABAINEIN], meaning 'to pass through'. It was first used by Aretaeus of Cappadocia in the 2nd century AD. Aretaeus gave a clinical description of the disease, noting the increased urine flow, thirst and weight loss, features which are instantly recognisable today" (Williams et al. 1999, page 6).

Diabetes is diagnosed by testing individuals with chronic hyperglycaemia. Tests that are currently applied in clinical practice include fasting plasma glucose (FPG) test, random plasm glucose test and oral glucose tolerance test (OGTT). The criteria for diabetes diagnosis varied until late 1970s, which had as a result wide variations in the reported prevalence of diabetes.

In 1997, an "Expert Committee of the American Diabetes Association proposed modifying the diagnostic criteria for diabetes, by lowering the fasting plasma glucose at which diabetes can be diagnosed" (Williams et al. 1999, page 17). Moreover, Rudnicka et al. (1999, page 1) mention that "diabetes should be considered in people with one or more of the following:

- polyuria (excessive passage of urine)
- polydipsia (excessive or abnormal thirst)
- lethargy
- weight loss
- recurrent infections or inflammations
- blurring of vision
- ulcerated feet
- poor healing
- hypertension
- ischaemic heart disease
- obesity
- family history of diabetes
- peripheral vascular disease"

Classification of diabetes by the World Health Organisation (WHO) (1980, revised 1985) "included the two common types of diabetes which were identified and classified by a clinical description of the patients, i.e. having either insulin-dependent diabetes mellitus [*IDDM*] (type 1

diabetes), because they are judged to need insulin to survive, or non-insulin-dependent diabetes mellitus [*NIDDM*] (type 2 diabetes), where insulin deficiency is less severe and insulin replacement is not essential to preserve life" (Williams et al. 1999, page 20).

In 1997, the American Diabetes Association "abandons the terms IDDM and NIDDM, which may be confusing because they are based on treatment and not aetiology, and retains the terms type 1 and type 2 diabetes. Most of the cases of type 1 are due to autoimmune destruction of the islet B cells [of the pancreas]; type 2 diabetes is caused by insulin resistance with an insulin secretory defect" (Williams et al. 1999, page 21).

Long term complications of diabetes, described in detail by Rudnicka et al. (1999) and Williams et al. (2000) include:

- Hypertension, defined as a blood pressure (BP) greater than 140/90 mmHg, is present in 70% of the diabetic population.
- Coronary heart disease (CHD), which is the most common and important complication of diabetes, accounting for at 75% of deaths.
- Lipid disorders, where in poorly control diabetes are identified by increased triglycerides, increased LDL cholesterol and decreased HDL cholesterol.
- Peripheral vascular disease is related to foot ulceration and amputation.
- Diabetic nephropathy, the commonest cause of renal failure in the UK.
- Diabetic neuropathy, combined with peripheral vascular disease is the leading cause of diabetic foot problems.
- Diabetic eye disease, which will be described in detail in the next section.

It is clear from the complications given above that diabetes mellitus is a medical condition associated with other life-threatening and "quality of life" reducing diseases. Further information about epidemiology and aetiology of type 1 and type 2 diabetes, as well as description of diabetic treatment are given in Rudnicka et al. (1999) and Williams et al. (2000).

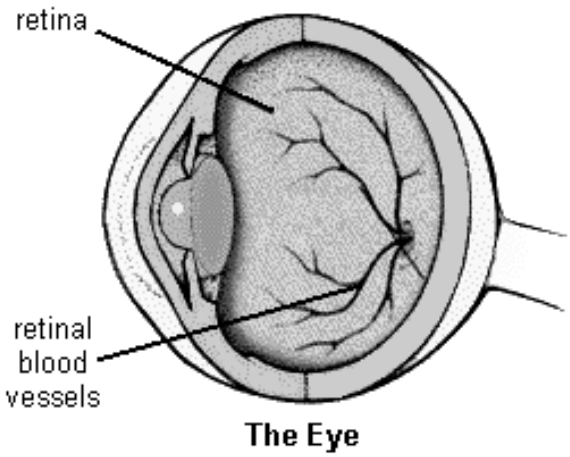
5.2.2 Diabetic eye disease

Diabetic retinopathy is a common complication of both type 1 and type 2 diabetes and is "the major cause of registerable blindness in the working population in Western countries" (Rudnicka et al. 1999, page 32). In Fong et al. (2003), it is mentioned that during "the first two decades of

disease, nearly all patients with type 1 diabetes and >60% of patients with type 2 diabetes have retinopathy.

In the Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR), 3.6% of younger-onset patients (type 1 diabetes) and 1.6% of older-onset patients (type 2 diabetes) were legally blind. In the younger-onset group, 86% of blindness was attributable to diabetic retinopathy. In the older-onset group, in which other eye diseases were common, one-third of the cases of legal blindness were due to diabetic retinopathy".

In figure 5.1 on page 156, the retina and its corresponding blood vessels are shown in a representation of a normal eye. In Williams et al. (2000, page 16), it is mentioned that "diabetic eye disease primarily affects the retinal blood vessels, but diabetes also accelerates cataract formation (lens opacities). The lesions of diabetic retinopathy can be grouped into five categories, according to the features seen on *ophthalmoscopy* - background, *preproliferative* and *proliferative* retinopathy, advanced diabetic eye disease and *maculopathy*".

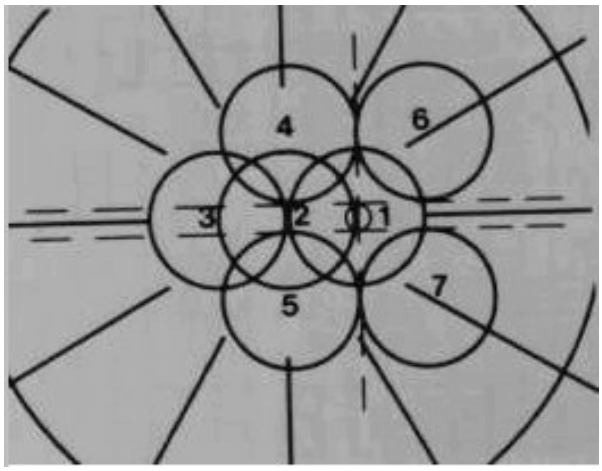


Retina the important part for this study.

Figure 5.1: Representation of the normal eye

The method for determining the presence and severity of diabetic retinopathy is using retinal photographs taken during eye examination (ophthalmoscopy). The schematic representation of the seven photographic fields that are used in stereo fundus photography and is regarded as the standard procedure for retinopathy grading is shown in figure 5.2 on page 157.

Specifically, each of the seven standard fields mentioned above covers 30 degrees. Field 1 is centred on the optic disc, field 2 is centred on the macula and field 3 is just temporal to the macula. Fields 4 - 7 are tangential to horizontal lines passing through the upper and lower poles of the disc and to a vertical line passing through its centre.

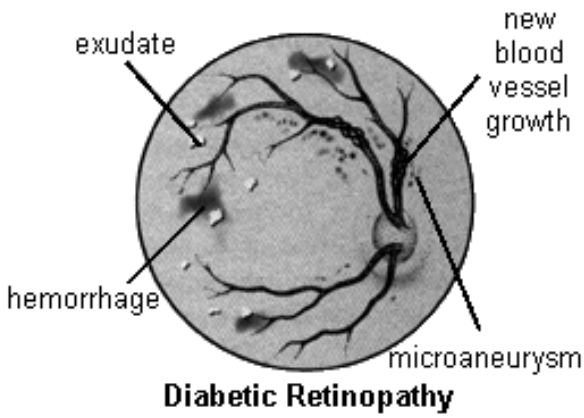


Seven photographic fields for coverage of different eye areas.

Figure 5.2: Standard scheme for diabetic retinopathy grading

The most common diabetic eye disease complications of are shown in figure 5.3 on page 157. The natural history of diabetic retinopathy, as described by Fong et al. (2003) involves progression "from mild nonproliferative abnormalities, characterised by increased vascular permeability, to moderate and severe nonproliferative diabetic retinopathy (NPDR), characterised by vascular closure, to proliferative diabetic retinopathy (PDR), characterised by the growth of new blood vessels on the retina and posterior surface of the vitreous.

Macular edema, characterised by retinal thickening from leaky blood vessels, can develop at all stages of retinopathy. Pregnancy, puberty, blood glucose control, hypertension, and cataract surgery can accelerate these changes".



Combination of complications indicate disease stage.

Figure 5.3: Diabetic retinopathy complications

Furthermore, Fong et al. (2003) point out that *sight-threatening retinopathy* "is rare in type 1 diabetic patients in the first 3-5 years of diabetes or before puberty. During the next two decades, nearly all type 1 diabetic patients develop retinopathy. Up to 21% of patients with type 2 diabetes have retinopathy at the time of first diagnosis of diabetes, and most develop some degree of retinopathy over time.

Vision loss due to diabetic retinopathy results from several mechanisms. Central vision may be impaired by macular edema or capillary nonperfusion. New blood vessels of PDR and contraction of the accompanying fibrous tissue can distort the retina and lead to tractional retinal detachment, producing severe and often irreversible vision loss. In addition, the new blood vessels may bleed, adding the further complication of preretinal or vitreous hemorrhage. Finally, neovascular glaucoma associated with PDR can be a cause of visual loss".

Treatment of diabetic retinopathy include glycemic control, blood pressure control and aspirin treatment (Fong et al. 2003). Moreover, the same authors mention that *laser photocoagulation* "is effective at slowing the progression of retinopathy and reducing visual loss, but the treatment usually does not restore lost vision.

Because these treatments are aimed at preventing vision loss and retinopathy can be asymptomatic, it is important to identify and treat patients early in the disease. To achieve this goal, patients with diabetes should be routinely evaluated to detect treatable disease".

The proposed ophthalmologic examination schedule for diabetic patients, according to Fong et al. 2003 is the following:

- Type 1 diabetes
  - First examination: Within 3-5 years after diagnosis of diabetes once patient is age 10 years or older
  - Minimum routine follow-up: Yearly
- Type 2 diabetes
  - First examination: At time of diagnosis of diabetes
  - Minimum routine follow-up: Yearly

Further details about diabetic eye disease laser treatment and methods for diabetic retinopathy screening can be found in Rudnicka et al. (1999).

5.3 EKSAGE data

The population included in our study, labelled *EKSAGE* (Eva Kohner Study At George Eliot), consisted of patients attending the diabetic retinopathy clinic from 1964 to 1994. It has been the policy of the diabetologist to refer patients with eye complications to the Hammersmith diabetic retinopathy clinic for treatment as well as those patients with long duration of diabetes.

All patients were receiving standard diabetes treatment available at the time. The sample composed of type 1 or the "younger onset" type of diabetes who had their diabetes diagnosed before the age of 30 years and type 2 or the "older onset" type, those who were at least 30 years old when diagnosed with diabetes.

Data on patients were prospectively collected onto diabetes trial card. Information with regards to hemodynamic factors (standing and supine blood pressure), biochemical factors have been annually recorded. Details on physical examination (peripheral pulses, vibration sense) were also recorded.

All patients had detailed eye examination with information recorded on intraocular pressure, refractive errors, and visual acuity with retinal photographs collected and graded during each visit as a part of the clinical care.

Over three years the data was entered into Microsoft Access database designed for the purpose of the study. Data was cross-checked periodically; we give further details for data cleaning in next section. Prior to 1979, data entries for a number of factors were recorded in mg and all these values were converted in to SI system at the time of entry for the purpose of analysis.

Advantages of the EKSAGE cohort for risk factors analyses are its large size, the inclusion of patients at clinically important stages and good documentation of retinopathy severity. In addition data on standing and lying blood pressure, detailed physical examination findings and biochemical parameters which were meticulously documented over time.

Limitations are that it is not population based and patients who had retinopathy less severe than non-proliferative diabetic retinopathy in either eye are not in the cohort as their diabetologist/primirary care physician managed them.

In the many studies available with regards to progression of diabetic retinopathy, there is a failure to identify these factors clearly because of the variability of the disease, and the numerous factors that may influence its course and outcome. Moreover, future studies of progression of diabetic retinopathy are marred by the fact that effective treatment is now established and the guidelines for when to commence treatment and the techniques of treatment are well documented.

Clinical factors attributed to retinopathy may be divided into systemic and non-systemic (ocu-

lar) factors. Despite good control of glycaemia and blood pressure, it is still not clear why certain patients develop and progress to sight threatening retinopathy (STR). Various mechanisms have been postulated in the pathogenesis and progression.

The pattern of STR in diabetes suggests that the process leading to the development of proliferative retinopathy consists of different stages and that progression through each stage may be governed by different factors (Krolewski et al. 1986). A multifactorial model for the development of sight threatening diabetic retinopathy is assumed but the mechanisms of action risks of the identified factors is not entirely clear (Rand et al. 1985).

Our study included diabetic patients referred to hospital for various forms of retinopathy. It is not addressing the natural history of development or progression of diabetes or related complications in a population. The aim is to identify among those patients that visit the eye clinic for the first time potential risk factors that separate the patients with sight threatening retinopathy (STR) from those with non-sight threatening retinopathy (NSTR).

The findings from this data might not be applicable to the whole population or to the general diabetic population as it is a selected sample of those diabetic patients referred to the eye clinic because of eye problems. The data contains patients with several visits at the hospital. All visits after the first have not been used as the objective at this stage was identifying risk factors at first visit before any treatment or any other medical intervention administered at the eye clinic.

5.3.1 Data cleaning and cross-checking

Our data includes 1445 diabetic patients with at least one visit at the eye clinic. Because of the large number of variables, different data frames have been constructed in S-Plus and Excel that represent different aspects of an individual's recorded information. Each data frame column represent a variable and each data frame row represent a visit for a patient. There are 9001 rows and about 20 columns in each of the six data frames.

The id number for each patient is denoted by variable `ekno` and the maximum `ekno` in the data we used for statistical analysis is 2582. Specifically, there are 1137 patients in EKSAGE for which there is no eye data available or it was not possible to collect it at present time. There are several reasons behind the unavailability of missing eye and the results from the data we had in our disposal should be treated with caution.

It would be desirable to obtain and analyse data for the patients that the eye data is available in hospital records and compare the results of the current dataset with the corresponding results from the "updated" data. It should be emphasised that there are no reasons to believe that the

patients are different from those patients that we use in our analysis. Hence, we might assume that the inclusion of the patients in the future will not alter our results substantially.

On the other hand, there are other patients for whom the eye data is missing and will not be possible to be retrieved in the future from medical or other records. For these patients, other data such personal information, biochemical markers and hemodynamic measurements are available. Hence, it is possible to include these patients in models where variables related to eye data will not be present.

A possible suggestion to avoid the removal of the patients mentioned above is to impute the missing eye data and include in subsequent analysis for diabetic retinopathy. Due to the large number of eye related factors needed to grade a patient in terms of diabetic eye disease and the substantial number of individuals with missing eye data, it has been decided to exclude these patients from further analysis. As mentioned previously, we have no reasons to believe that eye information missingness depends on the eye condition in any way, hence we can assume that no bias is included by removing these people from subsequent analysis.

From year of visit and `ekno` variable, we have created a new variable by merging `ekno` and `year` variables. This new variable, labelled `ekyid` can be used as an identification code for each patient-year and to verify that the correct visit number has been assigned to each visit of each patient. At this point, we have found some disparities between visit numbers that have been given by Microsoft Access and then extracted as Excel spreadsheet files for importing them in S-Plus.

The reason for some of the inconsistencies mentioned above was possibly due to the fact that some patients have additional visits without eye data. As there are 9001 patient-years, it was not possible to check manually each visit number and verify that each `ekyid` corresponds to the correct year and ekno. A possible way to deal with such a task is to create appropriate algorithms (functions in S-Plus) for cross-checking and identification of data entry errors and disparities.

In this way, we have constructed variable `vinu` (visit number). There are 1445 entries corresponding to visit 1 information, 1249 patients with at least two visits and a patient with information about 29 visits. This does not mean that a follow-up visit is always a year after the previous one for every patient.

Additionally, we need to identify the number of visits that each patient has in the dataset. The variable `novi` (number of visits) provided another way of cross-checking the year of visit and the visit number for each data entry. There are, for example, 196 entries with "exactly one visit" indicator (196 patients), 386 entries with "exactly two visits" (193 patients with exactly two visits each), one patient with exactly 27 visits and another patient with precisely 29 visits.

For some unknown reason, the last year for some patients has been identified at the data exportation by Access as the first year for these patients. Manual identification and correction of data inconsistencies as the ones mentioned above is time consuming and is accompanied with the risk of adding typing errors. By the application of relevant S-Plus code, cross-checking, data correction and additional verification has been achieved in relatively short time and in a reliable way.

In complicated studies such as the EKSAGE study, the data analyst should be prepared to spend a substantial amount of time verifying the data before any statistical analysis. Further verification and cross-checking was needed for other variables and in some cases it was necessary to check the medical records of the patients to correct data inconsistencies.

For example, the age of a patient at each visit is usually recorded in the medical records and has been entered into the database without verification. Subsequently, from their date of birth and year of visit, some disparities have been found and corrected. For some individuals, the date of birth has not been recorded, hence it was not possible to check whether the age entry made in the hospital is reliable.

Finally, there were a few entries where the computed age by the difference between year of visit and year of birth was below zero. The reason for this was that in Access, the year of visit has been entered as 1960 whenever it was missing and if the year of birth was after 1960, then the calculated age was a negative value. In such cases, the medical records or subsequent entries for the same patient have been used to correct the type of errors described above.

As there are more than a hundred variables in the study, we will initially present those variables that have been included in subsequent statistical analysis we have implemented. The decision for the inclusion of only a small number of clinical factors in the study has been based on previous studies and the clinician's opinions about the importance of each variable present in the data.

Moreover, it has been suggested by the medical experts that it is important to identify which parts of the recorded medical information is important for the presence of sight threatening retinopathy. Hence, we initially analyse the data only from the year of entry for each patient.

5.4 Missing data

There are 1445 diabetic patients with at least one visit. With further investigation, we identified individuals in the study that we might remove from subsequent analysis. Specifically, according to the clinicians' point of view about having a simple initial investigation, we have removed 61 patients from the study according to the following criteria:

- 1 patient with height (variable `ht`) less than 120 cm.
- 1 patient with right intra-ocular pressure (variable `rtiop1`) more than 40 mmHg.
- 3 patients with right intra-ocular pressure (variable `ltiop1`) more than 40 mmHg.
- 43 patients with missing duration of diabetes at visit 1(variable `dmdav`).
- 14 patients with sight-threatening retinopathy indicator (variable `stri`) missing.

Furthermore, we have removed from the study all those patients for who have at least one missing value for the factors present in the data. At this stage, it has been decided that creatinine (variable `creat`), triglycerides (variable `trig`) and HbA_{1c} (variable `hba1c`) will not be included in the analysis due to the large number of missing values in each of the these variables. Specifically, there are 677 patients with missing value of creatinine, 734 with no records of HbA_{1c} at entry and 1013 without triglycerides entries.

By applying the data removal criteria, we remove an additional group of 516 diabetic patients. The final number of individuals in the sample is 929, a fact that indicates that we have excluded almost 36% of the cases that have at least one visit at the eye clinic. This is by no means a small proportion that we have removed and we should be aware of this fact when we interpret our results.

The pattern of missingness for the 516 patients shown in table 5.1 on page 164. In this occasion, for each of the factors used as an exclusion criterion, we have the following number of missing values:

- lying systolic blood pressure (`sbply` in mmHg): 14
- lying diastolic blood pressure (`dbply` in mmHg): 14
- standing systolic blood pressure (`sbpst` in mmHg): 115
- weight (`wt` in kg): 55
- body mass index, the ratio of weight divided by height^2 (`bmi` in kg/m^2): 132
- uric acid (`urac` in mmol/l): 182
- cholesterol (`chol` in mmol/l): 116
- retinal perfusion pressure, a weighted function of `sbply`, `dbply` and mean intraocular pressure (`rpply` in mmHg): 218

The last proportion of patients removed raises the question of biased results of subsequent statistical analysis. Hence, we might examine whether the variables in the models we construct have the same distribution for the 929 individuals included and the 516 excluded from the study.

rpply	chol	urac	bmi	wt	sbpst	dbply	sbply	cases
0	0	0	0	0	0	0	0	929
0	0	0	0	0	1	0	0	36
0	0	0	1	0	0	0	0	50
0	0	0	1	0	0	1	0	3
0	0	0	1	1	0	0	0	15
0	0	0	1	1	1	0	0	4
0	0	1	0	0	0	0	0	37
0	0	1	0	0	1	0	0	5
0	0	1	1	0	0	0	0	3
0	0	1	1	0	1	0	0	1
0	0	1	1	1	0	0	0	2
0	0	1	1	1	1	0	0	1
0	1	0	0	0	0	0	0	34
0	1	0	0	0	1	0	0	2
0	1	0	1	0	0	0	0	2
0	1	0	1	1	0	0	0	2
0	1	1	0	0	0	0	0	31
0	1	1	0	0	1	0	0	3
0	1	1	1	0	0	0	0	5
0	1	1	1	1	1	0	0	1
1	0	0	0	0	0	0	0	89
1	0	0	0	0	0	1	1	1
1	0	0	0	0	1	0	0	8
1	0	0	0	0	1	1	1	3
1	0	0	1	0	0	0	0	7
1	0	0	1	0	1	0	0	2
1	0	0	1	0	1	1	1	1
1	0	0	1	1	0	0	0	3
1	0	0	1	1	1	0	0	3
1	0	0	1	1	1	1	1	3
1	0	1	0	0	0	0	0	34
1	0	1	0	0	1	0	0	16
1	0	1	1	0	0	0	0	2
1	0	1	1	0	1	0	0	1
1	0	1	1	1	0	0	0	7
1	0	1	1	1	1	0	0	2
1	1	0	0	0	0	0	0	4
1	1	0	0	0	1	0	0	1
1	1	1	0	0	0	0	0	7
1	1	1	0	0	1	0	0	9
1	1	1	0	0	1	1	1	3
1	1	1	1	1	0	0	0	5
1	1	1	1	1	1	0	0	4
1	1	1	1	1	1	1	1	3

Removal criteria according to clinicians' opinion.

Table 5.1: Pattern of missingness for 516 patients

For example, it is possible to implement a t-test to compare the mean of the age of the patients in the initial analysis and the mean of those patients removed by one of the criteria described above. From the results (p-value=0.100), it can be said that the difference of the mean age for the two groups of individuals mentioned above is not statistically significant at the 5% level. The 95% confidence interval of the difference between the means of the two samples defined above is (-0.261, 2.934).

In the same way, we might compare the mean of other variables contained in this part of the study and also investigate the joint distribution of specific groups of factors. As there is a large number of possible ways to examine differences between included and excluded patients, it is possible to find significant differences by applying multiple tests. Hence, we will initially examine the dataset with 929 patients and then compare the results derived with the corresponding results derived by imputed datasets.

5.5 Data coding

There are factors in our study that we need to use as categorical variables. In this case, the coding should be defined according to the number of individuals in each level of the factor. For example, if there are few patients that belong to a specific group with a particular characteristic, it might be possible to include these patients in a larger group with similar characteristics. Large heterogeneity between people in a certain level of a factor might lead to biased results.

In the EKSAGE study, one of the key factors in the analysis is diabetes type (variable `dmty`). There are 297 type 1 patients and 632 type 2 patients. As we have previously mentioned, the aetiology for each type of diabetes is different. Hence, we will analyse each diabetes type group of patients separately.

Other factors included at this stage of the analysis are the following:

- gender (`gend`)
  - type 1: 179 male (gend=1) and 118 female (gend=2)
  - type 2: 373 male and 259 female
- race (`race`)
  - type 1: 269 white (race=4), 13 indo-asian (race=2), 5 afro-caribbean (race=3) and 10 other race (race=1)
  - type 2: 474 white, 104 indo-asian, 28 afro-caribbean and 25 other race.

- diabetes treatment (`insul`)
  - type 1: 215 insulin (insul=1), 77 pills (insul=2) and 5 not known (insul=3)
  - type 2: 161 insulin, 234 pills and 237 not known
- smoking status (`smoker`)
  - type 1: 118 smokers (smoker=2), 54 ex-smokers (smoker=1), 117 non-smokers (smoker=4) and 8 not known (smoker=3)
  - type 2: 142 smokers, 143 ex-smokers, 336 non-smokers and 11 not known.

The grading protocol for diabetic retinopathy includes information about several factors related to eye complications. Sight-threatening retinopathy indicator (variable `stri`) is defined by comparing the condition of right eye and left eye and the corresponding complications for each eye. Subsequently, worse eye retinopathy and maculopathy gradings are used to define whether sight-threatening retinopathy is present.

To be specific, there are five categories for diabetic retinopathy grading, without taking into account macular complications:

- Grade 1 No retinopathy
- Grade 2 Background retinopathy
- Grade 3 Preproliferative retinopathy
- Grade 4 Proliferative retinopathy
- Grade 5 Advanced diabetic eye disease

Moreover, maculopathy is identified by one of the following rules:

- Any patients with macular edema:
  - mild
  - moderate
  - severe
  - cystic
- Any patients with at least two exudates in macular area
- Any patients with any exudates in macula but the type of exudates:

- Circinate
- Plaque
- Scattered and circinate
- Scattered and Plaque
- Scattered, circinate and plaque

Sight-threatening retinopathy is present if worse eye retinopathy is at least grade 4 or maculopathy is present in at least one of the eyes. At this stage, it was necessary to write specific S-Plus functions to identify the patients in the study with sight-threatening retinopathy at each visit. For more details about definition of each of the features included in grading, the complete grading protocol for the EKSAGE project can be seen in the appendix.

From the details about diabetic eye disease mentioned above, it is clear that the response variable in our models (variable denoted as `stri`), is a combination of a large number of clinical factors. Thus, it might be also possible to implement statistical analysis for each of the factors contained in the EKSAGE study protocol to examine the role of specific factors as predictors of sight-threatening retinopathy.

At the current stage of study, we will assume that the clinicians' opinion about diabetic retinopathy grading is sufficient for defining the response variable (`stri`). Also, according to the experience of the medical experts and the literature related to diabetic eye disease, we implement statistical modelling with the purpose of understanding the importance of blood pressure as predictor of diabetic retinopathy.

5.6 Data description

Before constructing statistical models for investigating the association between predicting variables in the dataset and the response variable (sight-threatening retinopathy indicator, `stri`), it is important to investigate the distribution of the variables mentioned above. We will examine the distribution of these variables for the complete dataset and for each diabetes type separately.

Before that, we present key categorical factors such as gender, race, diabetes treatment and smoking status and their corresponding coding and frequencies table. The most important categorical variable in our study is sight-threatening retinopathy indicator (`stri`), defined by the grading protocol given in the appendix. The distribution of stri for each diabetes type is shown in table 5.3 on page 168.

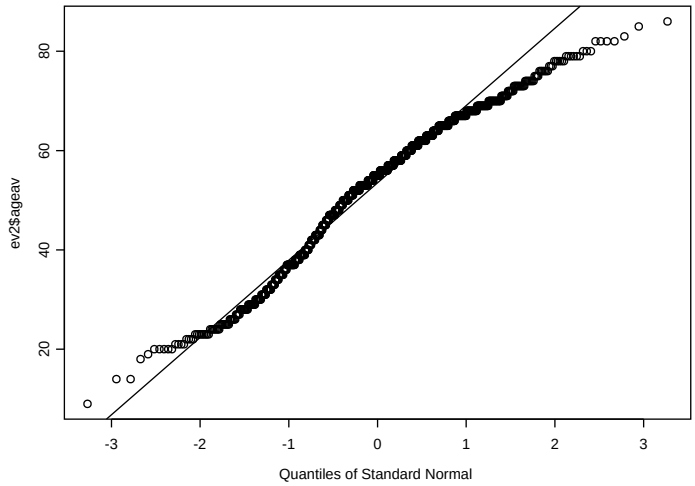
Variable	DM type 1 n (%)	DM type 2 n (%)
Sight-threatening retinopathy		
Yes	245 (82.5)	497 (78.6)
No	52 (17.5)	135 (21.4)
Race		
White	269 (90.6)	474 (75.0)
Indo-Asian	13 (4.4)	105 (16.6)
Afro-Caribbean	5 (1.7)	28 (4.4)
Others	10 (3.3)	25 (4.0)
Gender		
Male	179 (60.3)	373 (59.0)
Female	118 (39.7)	259 (41.0)
Smoking level		
Non-smokers	117 (39.4)	336 (53.2)
Ex-smokers	54 (18.2)	143 (22.6)
Smokers	118 (39.7)	142 (22.5)
Not known	8 (2.7)	11 (1.7)
Insulin treatment		
Insulin	215 (72.4)	161 (25.5)
Tablets	77 (25.9)	234 (37.0)
Not known	5 (1.7)	237 (37.5)

Categorical variables

Table 5.3: Number and proportion of patients for each type of diabetes

From table 5.3 on page 168, we can see that 82.5% of diabetes type 1 and 78.6% of diabetes type 2 have diabetic eye disease at the stage that sight is under threat. The high proportion of diabetic patients in this sample with eye disease in high-risk level is expected as they are the individuals referred to an eye clinic.

Age at visit 1 (at entry) distribution (variable `ageav`) for all patients is shown in figure 5.4 on page 169. It is apparent that the age for the complete dataset does not follow a Normal distribution. Type 1 diabetic patients are usually less than 30 years old when referred to an eye clinic whereas type 2 patients are at least 30 years old.



Distribution for 929 patients (complete dataset).

Figure 5.4: Q-Q plot of age at entry

Hence, age at entry for all patients is expected to be a mixture of two distributions that correspond to each diabetes type part of the sample used in the analysis. The Q-Q plots for the age at entry distribution for each diabetes type can be said to be close to a Normal distribution (results not shown).

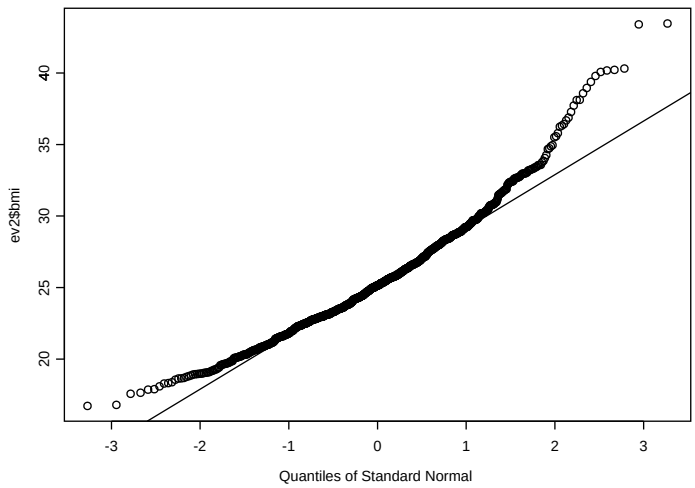
Diabetes duration at entry (variable `dmdav`) is another variable on continuous scale that might play an important role in explaining the presence of diabetic eye disease. Patients with type 1 diabetes are expected to have longer diabetes duration than the corresponding group with type 2 diabetes. The distribution of each diabetes type group can be seen in table 5.5 on page 170.

Body mass index (variable `bmi`) is defined as the ratio of weight in kilograms divided by the square of height (measured in metres). From the Q-Q plot of bmi (figure 5.5 on page 170) for the complete dataset, it can be said that this variable does not follow a Normal distribution.

Variable	DM type 1 Mean (s.d.)	DM type 2 Mean (s.d.)
Systolic lying blood pressure (mmHg)	142.7 (26.02)	160.7 (28.64)
Diastolic lying blood pressure (mmHg)	84.8 (12.56)	89.8 (14.03)
Mean arterial pressure (mmHg)	104.1 (15.67)	113.5 (17.28)
Retinal perfusion pressure (mmHg)	53.3 (10.59)	59.0 (11.28)
Age (years)	38.4 (11.44)	60.0 (9.87)
Body mass index (kg/m^2)	23.8 (3.31)	26.4 (4.09)
Diabetes duration at entry (years)	21.4 (8.15)	9.8 (7.63)
Uric acid (mmol/ml)	0.29 (0.09)	0.32 (0.09)
Cholesterol (mmol/ml)	6.20 (1.66)	6.14 (1.66)

Continuous variables

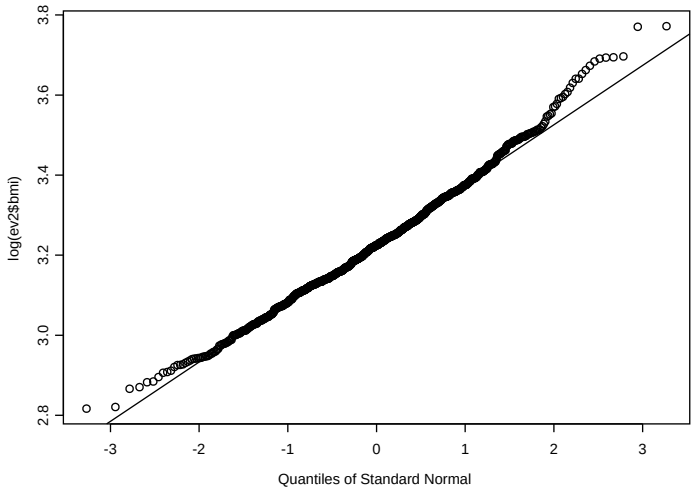
Table 5.5: Mean and standard deviation for each DM type



Distribution for 929 patients (complete dataset).

Figure 5.5: Q-Q plot of body mass index

If we apply the logarithmic transformation, by examining figure 5.6 on page 171, this transformation seems appropriate for achieving normalisation of bmi distribution. The same transformation gives similar normalisation for the distribution of bmi for each diabetes type (results not shown).



Distribution for 929 patients (complete dataset).

Figure 5.6: Q-Q plot of the logarithm of body mass index

Uric acid (`urac`) and cholesterol (`chol`) are two other factors measured on continuous scale and might be proven important for predicting the presence of sight-threatening retinopathy. Both variables mentioned above follow a Normal distribution when transformed by the logarithmic transformation. This transformation achieves normalisation to a great extent for both type 1 and type 2 diabetic patients (results not shown).

Blood pressure measurements are the main focus of the first part of the study. To be precise, we examine whether several variables related to systemic and ocular blood pressure might be useful predictors of sight-threatening retinopathy at entry. Furthermore, we investigate the extent of eye complications' risk reduction that can be achieved by administering blood pressure treatment.

There are three blood pressure variables that can be directly obtained by measuring the patients with the appropriate medical equipment. These are systolic (bps) and diastolic blood pressure (bpd), which are indicators of systemic blood flow and are the most common blood pressure factors in bio-statistical research projects. In addition to that, intra-ocular blood pressure (iop) for each eye can be measured by the clinicians and are useful as indicators of eye complications and abnormalities in ocular blood flow.

Other blood pressure related factors might be derived by the directly measured indicators mentioned above. These are mean arterial pressure (map) and pulse pressure (pp), which are respectively a weighted average ($\frac{2}{3}$bpd + $\frac{1}{3}$bps) and the difference (bps-bpd) and are also common in the medical literature. Additionally, postural systolic pressure, which is the difference between lying and standing systolic blood pressure might be another useful factor associated with diabetic eye disease complications.

For diabetic patients with eye complications, retinal perfusion pressure (rpp), which is equal to $\frac{2}{3}$map − iop, is another function of systemic and ocular blood pressure measurements. It has been proposed in the medical literature as a useful predictor of diabetic retinopathy and we will include this variable in our statistical analysis.

5.7 Univariate analysis

The next in our study will be to examine the association between the predictor variables presented previously and sight-threatening retinopathy indicator (`stri`), the binary variable which is the event of interest. For categorical covariates, we will implement χ^2-tests and for continuous factors t-tests or Mann-Whitney tests whenever the distribution of the variables could not be assumed to follow a Normal distribution.

The results for the association between `stri` and gender for the complete dataset (p-value=0.9486) and each diabetes type group of patients (p=05659 for type 1 and p=0.5773 for type 2) indicate the association between the response and the predictor variable mentioned above is not statistically significant using 5% significance level. Hence, it might be conclude that gender is not likely to be an important predictor of advanced diabetic eye complications.

On the other hand, this does not necessarily mean that gender should not be included in other types of statistical analysis. In Sonis (1998), it is mentioned that confounding cannot be assessed with a statistical test. To be more specific, assessment of confounding "by testing the statistical significance of baseline differences or the significance of a potential confounding factor in a multivariate model can produce underestimates or overestimates of the true association between an exposure and an outcome".

Furthermore, the same author states that we should not include all covariates in a multivariate model to control confounding. The reason for not including covariates in the model that are not confounders is that this "may produce underestimates or overestimates of the effect in question, as well as artificially widened confidence intervals" (Sonis 1998).

Finally, the author mentioned above says that in order to "prevent problems resulting from

these misunderstandings, researchers should consider drawing causal models prior to conducting the research and use the change-in-estimate criterion, rather than a statistical test, to detect confounding". Graphical models and specifically chain graphs related to association and influence data structures (Edwards 2000) might be a possible way of assessing possible causal models.

In the same way as mentioned previously for gender, we test the strength of the association between categorical variable `race` and `stri`. In this case, race is not significantly associated with advanced diabetic retinopathy, even though χ^2-test may not be appropriate due to the small number of individuals in most of the contingency table cells.

On the other hand, race might be an important predictor for type 2 diabetic patients' eye complications. The difference between type 1 and type 2 associations for a specific covariate (`race`) can be attributed to the different aetiology of each type of diabetes.

It might be possible that the causal mechanism for type 1 diabetic ocular disfunction is not different for each of the ethnic groups. On the other hand, variable `race` seems to play an important role into finding type 2 diabetic patients with diabetic eye disease and it is possible that the causal mechanism might not be the same for all races.

Moreover, smoking level is not found by association tests to be significantly correlated to sight-threatening retinopathy, even though there are small number of individuals in some of the contingency tables. Specifically, in table 5.3 on page 168, we can see that level 3 of smoking variable (not known level for variable `smoker`), there are few patients without sight-threatening retinopathy.

A possible way to overcome the problem of small number of cases in some cells of the contingency tables mentioned above is to remove the individuals from the test or merge factors levels with similar meaning. For `smoker` variable, we might assume that the removal of patients with unknown smoking status will not lead to biased χ^2-test results. For type 1, the p-value from the χ^2-test is 0.6097 and the corresponding p-value for type 2 is 0.9802. Hence, we can say that based on the tests mentioned above, smoking status factor is not significantly associated with `stri`.

Another categorical factor that could be an important predictor for `stri` is insulin treatment (`insul`). Similarly to the adjustments for smoking status factor described previously, we have removed unknown insulin treatment status for type 1 diabetes group of patients' tests. From the results (p=0.031 for type 1 and p=0.4381 for type 2), the conclusion extracted is the significant (at 5% level) association of insulin treatment only for type 1 diabetic patients.

For continuous variables, as many of them do not follow a Normal distribution, a t-test might not be the appropriate choice for comparison between the group of patients without and the

patients with sight-threatening retinopathy. An alternative non-parametric test in this case is the Mann-Whitney test, which is implemented in S-Plus by the equivalent Wilcoxon rank sum test for two sample data.

The results for body mass index indicate that there is no significant difference for bmi for either diabetes type group. In the same way, we have found that age at entry is significant for type 1 only and also that diabetes duration is not significant for diabetes type 2 only (results not shown). Uric acid is not significant for either diabetic group and cholesterol is significant only for type 1 patients.

Blood pressure measurements can also be tested for significant association with advanced diabetic eye disease using univariate tests mentioned above. In this way, we find that lying systolic and diastolic blood pressures (variables `sbply` and `dbply` respectively) are significant for both diabetes types (results not shown).

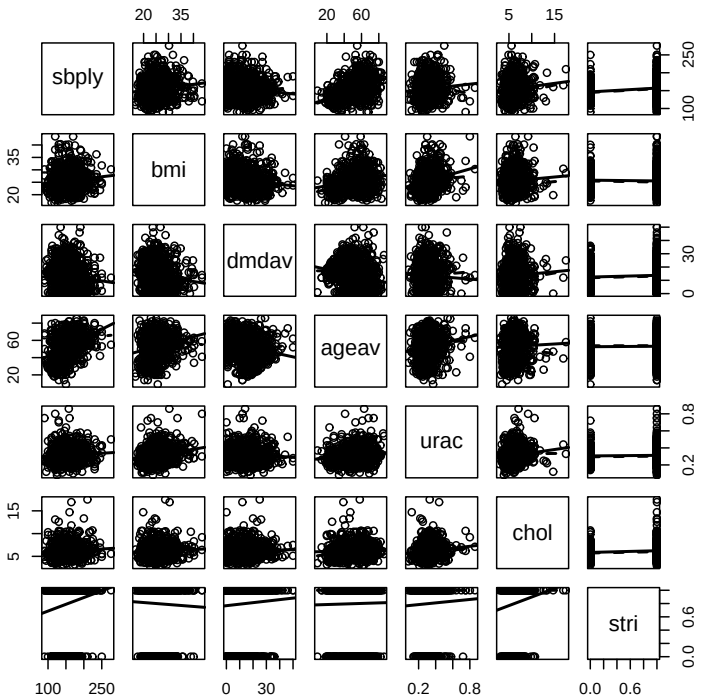
Moreover, functions of directly measured blood pressures can be tested in the way mentioned above. For example, mean arterial pressure and retinal perfusion pressure are both significantly associated with sight threatening retinopathy at 5% significance level for both types of diabetes. On the other hand, postural systolic pressure is significant only for type 1 diabetic patients.

From the results in this section, we have identified possible predictors for sight-threatening retinopathy indicator (`stri`). The univariate association tests might be regarded as a indication that needs to be confirmed by other more complicated or sophisticated statistical analyses that take into account additional variables or other types of information. Such analyses could be performed by methods and procedures appropriate for investigating bivariate relationships in the dataset.

5.8 Bivariate analysis

The scatterplot matrix is a possible way of examining associations between factors in our study. For the complete dataset, the results are shown in figure 5.7, page 175, where we have used the response variable (`stri`) and continuous predictors. In the scatterplot mentioned above, we can see that lying systolic blood pressure (`sbply`) is increasing when age at entry (`ageav`) is increasing and also that uric acid (`urac`) is proportional to body mass index (`bmi`).

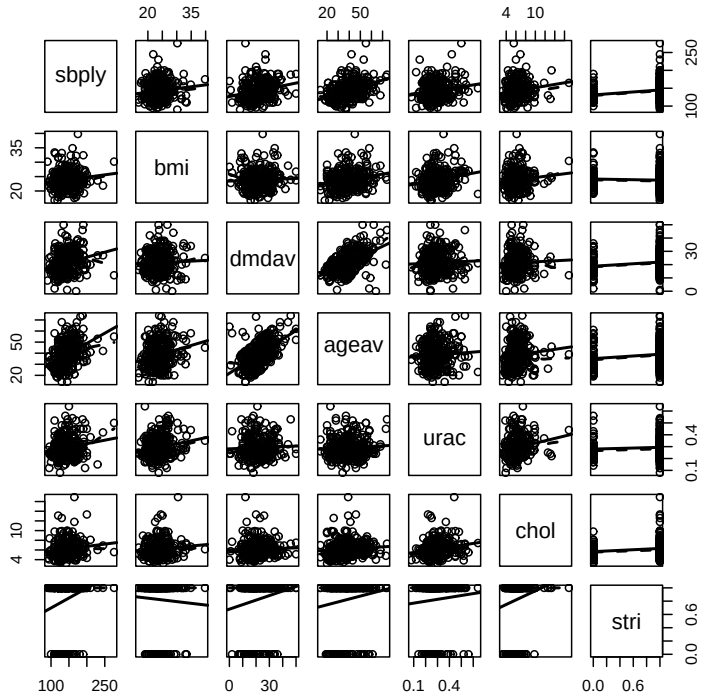
On the other hand, diabetes duration (`dmdav`) is decreasing when age at entry (`ageav`) is increasing. This result might be attributed to the fact that both types of diabetes are included in the sample used for deriving the scatter plot mentioned above. Hence, it might be helpful to examine the corresponding scatter plots for each diabetes type group separately. Diabetes type 1 output (using statistical software R to avoid problems with large size plot in S-Plus) is shown



Response variable and continuous predictors included.

Figure 5.7: Scatter plot matrix for complete dataset

in figure 5.8, page 176. The corresponding scatterplot for diabetes type 2 patients can be seen in figure 5.9, page 177.



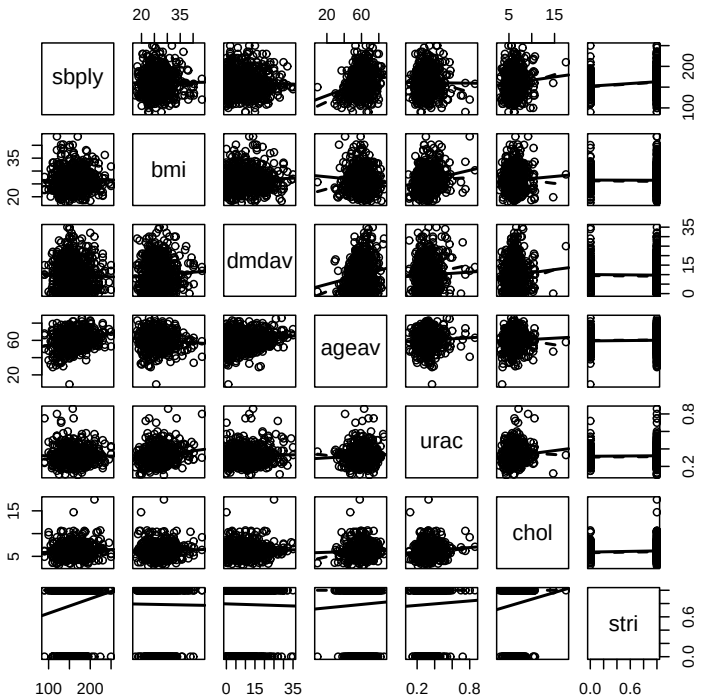
Response variable and continuous predictors included.

Figure 5.8: Scatter plot matrix for diabetes type 1 patients

In both figure 5.8, page 176 and figure 5.9, page 177, we can conclude that diabetes duration is increasing with increasing age at entry. Compared to the corresponding scatter plot matrix cells of `ageav` and `dmdav`, we can see that the association between the variables mentioned before is reversed for each subsample. This is another example of the phenomenon labelled as Simpson's paradox.

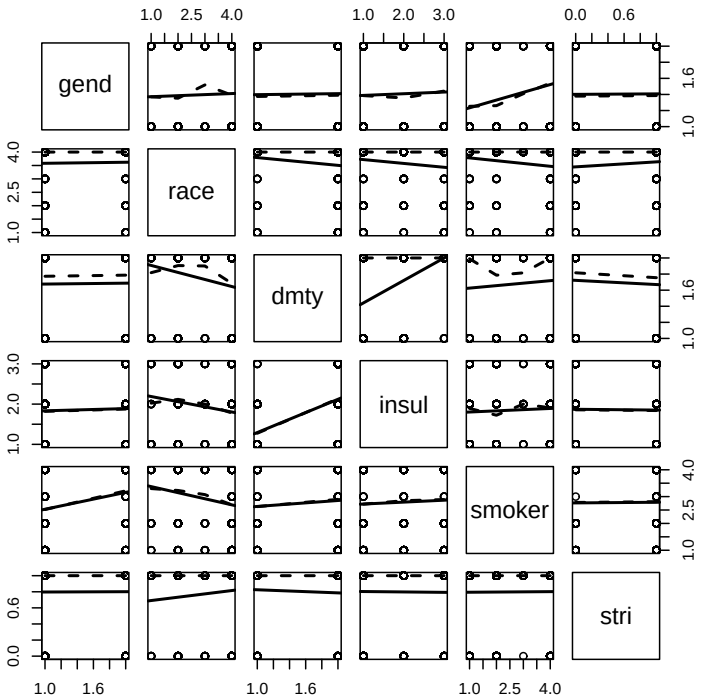
In the same way, we can have additional factors in the scatter plot matrix that are important in the EKSAGE study. Categorical variables might be inserted into to the type of bivariate analysis described previously and the proportion of patients for each combination of categorical factor levels will be shown. For example, in figure 5.10, page 178, we can see that the proportion of patients with `stri` present increases from Indo-Asian to White `race`, remains the same for different smoking levels and decreases with is larger for type 1 than type 1 diabetics.

Additional information might be gained by scatter plot matrices with combinations of categorical and continuous variables. The drawback of including a large number of factors in a scatter



Response variable and continuous predictors included.

Figure 5.9: Scatter plot matrix for diabetes type 2 patients



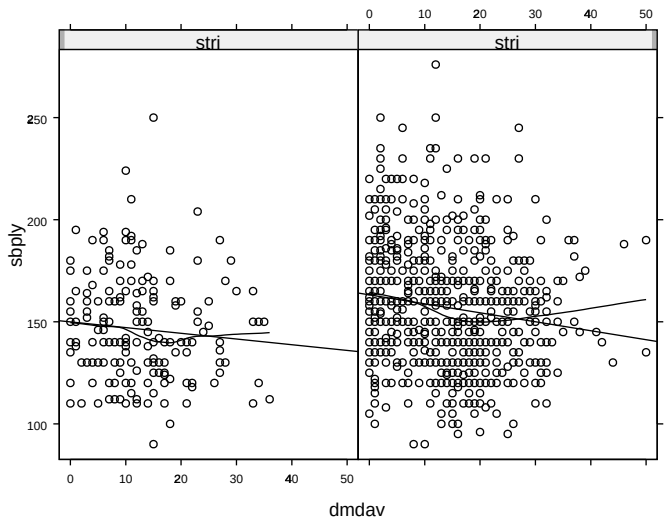
Response variable and categorical predictors included.

Figure 5.10: Scatter plot matrix for complete dataset

plot matrix is the difficulty of inspecting visually a large number of data patterns to identify associations. Furthermore, it might be necessary to quantify the correlation between covariates in the study and if possible, implement significance tests such as Kendall's method.

5.9 Multivariate exploratory analysis

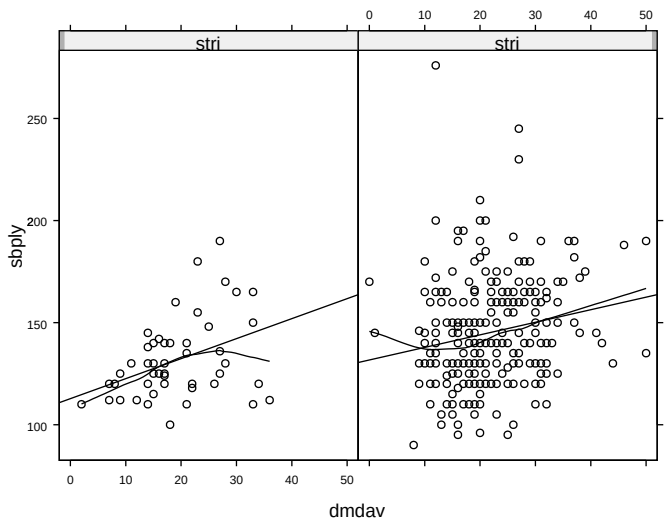
In addition to bivariate exploratory analysis, it might be useful to implement association graphical methods such as conditioning plots that have been described previously. For instance, in figure 5.11, page 179, we can see that for the complete dataset, systolic lying blood pressure (variable `sbply`) is decreasing when diabetes duration (variable `dmdav`) is increasing up to approximately 20 years. For more than 20 years duration, systolic blood pressure is increasing and this pattern is almost identical for both patients without (left half of the plot) and with advanced diabetic eye disease (right half of coplot).



Number of patients larger when `stri`=1 (disease present, right half of the plot)

Figure 5.11: Systolic blood pressure and diabetes duration conditional on diabetes eye disease condition

Moreover, for diabetes type 1 patients only, the conditioning plot in figure 5.12 on page 180, systolic lying blood pressure has a different pattern of change with diabetes duration conditional on sight-threatening retinopathy level. Specifically, for patients without advanced diabetic eye disease (left half of coplot), `sbply` increases steadily up to 25 years duration of diabetes and then it is decreasing. On the other, for individuals with `stri`=1, systolic blood pressure decreases up to 20 years duration and then it's increasing.



Diabetes type 1 patients number larger when `stri`=1 (disease present, right half of the plot)

Figure 5.12: Systolic blood pressure and diabetes duration conditional on diabetes eye disease condition

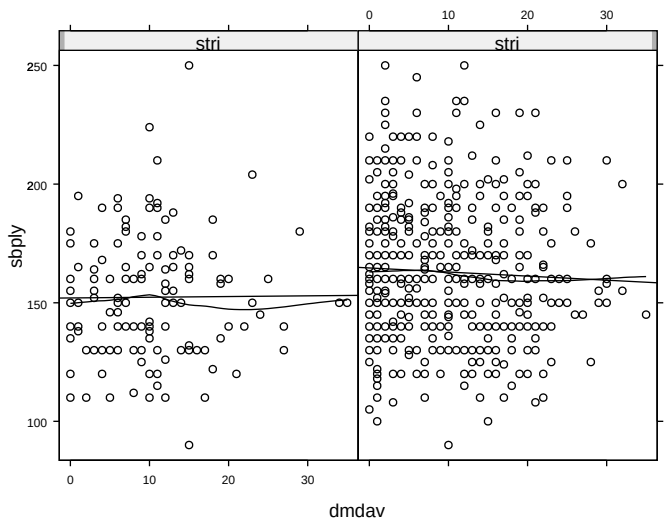
On the contrary, for patients that belong to type 2 diabetic group (figure 5.13 on page 181), systolic lying blood pressure has small fluctuations with diabetes duration for individuals without and with sight-threatening diabetic eye complications. It is clear from the conditioning plots mentioned above that the association structure between factors in our study is not likely to be identical for different subsamples in the diabetic population investigated.

Furthermore, we might need to quantify the strength of the association for each group examined by graphical multivariate exploratory methods. We must remind the reader that a large number of conditioning plots is similar to a large number of hypothesis tests. Also, "data dredging" even by graphical methods is accompanied with the possibility of finding apparently important associations which are not necessarily statistically or clinically significant.

5.10 Odds ratio for risk factors

At this stage, we are interested in investigating the association between different measurements of blood pressure with sight-threatening retinopathy. The importance of this task is to examine the extent of reducing risk of having advanced diabetic eye disease by administering blood pressure medication to diabetic patients.

According to the clinicians' opinion, there are other factors in the study that might have influence on the association between blood pressure level and diabetic eye disease stage. Hence,



Diabetes type 2 patients number larger when `stri`=1 (disease present, right half of the plot)

Figure 5.13: Systolic blood pressure and diabetes duration conditional on diabetes eye disease condition

we need to control for possible confounders of the association mentioned above and we should remember that these confounders are not necessarily statistically significant for the event of interest (presence of sight-threatening diabetic retinopathy).

For type 1 diabetic group, we control for the following factors:

- Gender
- Race
- Smoking level
- Body mass index
- Age at entry
- Diabetes duration at entry
- Uric acid
- Cholesterol

For type 2 diabetic patients, we control for the same confounders as for type 1 patients with the addition of insulin treatment.

The results for type 1 and type 2 groups of patients for different blood pressure measurements are shown in table 5.7, page 183. To calculate the odds-ratio for each blood pressure variable, we have implemented logistic regression models as implemented in Design S-Plus 4.5 library (2/5/2000 version). The logistic regression model has as response variable the indicator for sight-threatening retinopathy (`stri`) and as predictors the confounders for each diabetes type given previously and one of the blood pressure factors.

Subsequently, it was necessary to construct a data frame with all the variables mentioned above and subsequently implement procedure included in the Design library to change the format of the data and to use the command *lrm* instead of the standard logistic regression procedure in S-Plus. Furthermore, in order to obtain the odds ratio for a specific range for each blood pressure measurement, it is necessary to give two values for each variable so that their difference is the range of interest.

For example, for `sbply`, giving the values of 120mmHg and 130mmHg is one of the possible ways of having the odds ratio of 10mmHg change for this variable. It is assumed that the risk of sight-threatening retinopathy changes linearly with systolic lying blood pressure. If a transformed variable is used, it is not possible to use the standard procedure to estimate the odds ratio for the event of interest for a change on the original scale. For example, if we include a blood pressure measurement after applying logarithmic transformation, the odds ratio will depend on the proportion and not the difference between measurements.

For each variable, the first line indicates the lower and higher value and the difference for which the odds ratio is estimated and also the estimated effect (regression coefficient) and its associated standard error and 95% confidence interval limits. In the second line of each factor, the odds ratio for a given range and the 95% confidence interval are shown.

The results contained in table 5.7, page 183 are the following:

- Systolic lying blood pressure (`sbply`)
- Diastolic lying blood pressure (`dbply`)
- Retinal perfusion pressure (`pply`)
- Postural systolic pressure (`posys`)
- Mean arterial pressure (`map`)

If it is necessary to transform the risk factor for which we are interested in obtaining the odds ratio, then the procedure described above should be modified. For example, suppose that the

Risk factors	Odds ratio per 10 mmHg (95% C.I.)
DM type 1	
Systolic lying blood pressure	1.24 (1.05, 1.47)
Diastolic lying blood pressure	1.59 (1.16, 2.17)
Retinal perfusion pressure	1.75 (1.19, 2.56)
Pulse pressure	1.16 (0.94, 1.44)
Postular systolic pressure	1.31 (1.00, 1.71)
Mean arterial pressure	1.50 (1.14, 1.96)
DM type 2	
Systolic lying blood pressure	1.13 (1.05, 1.23)
Diastolic lying blood pressure	1.36 (1.17, 1.59)
Retinal perfusion pressure	1.49 (1.22, 1.80)
Pulse pressure	1.08 (0.98, 1.20)
Postular systolic pressure	1.08 (0.96, 1.21)
Mean arterial pressure	1.28 (1.13, 1.46)

Odds ratio adjusted for all confounders specified by clinicians.

Table 5.7: Odds ratio for blood pressure measurements

logarithmic transformation should be applied to predictor x_i. Then, we have for the log odds the following expression:

$$\log(\frac{\widehat{p}}{1-\widehat{p}}) = c_0 + c_1 x_1 + ... + c_i \log(x_i) + ... + c_n x_n$$

which can be written as

$$(\frac{\widehat{p}}{1-\widehat{p}}) = \exp(c_0) exp(c_1 x_1) ... (x_i)^{c_i} ... \exp(c_n x_n)$$

The odds ratio for two measurements m_1 and m_2 for the risk factor x_i is

$$\frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \left(\frac{m_1}{m_2}\right)^{c_i}$$

Hence, if it required to use x_i on the original scale measurement and not the transformed scale, the odds ratio can be estimated for a specific proportion $\frac{m_1}{m_2}$ rather than the difference $m_1 - m_2$.

For a $\frac{p}{100}\%$ increase in the risk factor x_i, we label proportion increase by R where $R = 1 + \frac{p}{100}$. Hence, as $R = \frac{m_1}{m_2}$, we have that the odds increase by R^{c_i}. For a small increase, where it can assumed that $R \simeq 1$ or $p \simeq 0$, we have the following expression:

$$R^{c_i} \simeq 1 + (R-1)c_i$$

Finally, a confidence interval for the odds ratio increase can be found using the coefficient $\widehat{c_i}$ and its associated standard error $\widehat{s}(c_i)$. For example, the 95% confidence interval for the increase in odds ratio can be approximated by the following interval:

$$(R^{\widehat{c_i}-1.96\widehat{s}(c_i)}, R^{\widehat{c_i}+1.96\widehat{s}(c_i)})$$

5.11 Hot-deck imputation

As we have seen previously, we have removed from the dataset we analysed 516 patients, according to the clinicians' criteria and when at least one of the values for the factors in data is missing. The large number of the individuals not in the statistical analysis might be a possible reason for having biased results. At this stage, instead of removing the cases with missing values, we apply the hot-deck imputation method so that the number of patients not in the study will be reduced. Hence, we will have 1384 subjects in our analysis.

Hot-deck imputation is described in Twisk et al. (2002) as an approach where "the average value of, or a random draw from a subset of comparable cases (e.g. cases with the same gender, age, etc.) is imputed for the missing values". For the EKSAGE study, we have constructed S-Plus functions for each of the factors in the study with matching criteria according to the clinicians' opinions.

For example, for imputing Body Mass Index (BMI) missing values, we implement the mean hot-deck method using age at entry, gender and diabetes type as matching criteria. Subsequently, for the cases where there no matching individuals exist for the combination of the variables mentioned above, we have included only a subset of the initial matching criteria. For BMI imputation, age at entry and gender have been included to identify comparable cases and after that only gender.

The list of the variables used as matching criteria is shown in table 5.8 on page 185. For factors such as age at entry and diabetes duration, the clinicians' opinion has been that it is not sensible to impute the missing values for the factors. In fact, 43 patients have been removed from further analysis because their diabetes duration entry was not present in the data.

On the other hand, it might said that mean hot-deck imputation has specific drawbacks which might affect the results of our statistical analysis. Since the imputed values are identical to the average of all comparable cases, it can be said the method of imputation describe above is appropriate for estimating the mean of a variable but it underestimates the corresponding variance.

In Schafer (1997), it is mentioned that for the estimates of methods such as hot-deck imputation, "standard errors, p-values and other measures of uncertainty calculated by standard complete-

- BMI
  - Age at entry, gender, diabetes type
  - Age at entry, gender
  - Gender
- uric acid
  - Age at entry, gender, diabetes type
  - Age at entry, gender
  - Gender
- cholesterol
  - Age at entry, gender, BMI, diabetes duration
  - Age at entry, gender, BMI
  - Age at entry, gender
  - Gender
- Systolic, diastolic and intraocular blood pressure measurements
  - Age at entry, gender, diabetes type, BMI
  - Age at entry, gender, diabetes type
  - Age at entry, gender
  - Gender

Table 5.8: Matching criteria for average hot-deck imputation

data methods could be misleading, because they fail to reflect any uncertainty due to missing data". Hence, when it is possible, other methods such imputation by the EM-algorithm or multiple imputation might be more appropriate methods to deal with missing entries.

Thus, we only use mean hot-deck imputation to compare the results for the odds-ratio from complete-case analysis and hot-deck imputation analysis. At this point, we would like to remind the reader that exploratory analysis might be also useful after imputation. Transformations of factors in the analysis and associations between covariates could be different when the removed patients are inserted in the analysis. It is not clear whether the results from complete-case and hot-deck imputation are applicable to the diabetic population under investigation.

In the same as with the results in table 5.7, page 183 (from complete-case analysis), the estimated odds-ratio results for blood pressure changes for mean hot-deck imputed data are shown in table 5.10, page 186.

Risk factors	**Odds ratio per 10 mmHg (95% C.I.)**
DM type 1	
Systolic lying blood pressure	1.22 (1.08, 1.38)
Diastolic lying blood pressure	1.41 (1.14, 1.75)
Retinal perfusion pressure	1.74 (1.30, 2.32)
Pulse pressure	1.19 (1.01, 1.40)
Postular systolic pressure	1.22 (1.01, 1.49)
Mean arterial pressure	1.39 (1.15, 1.67)
DM type 2	
Systolic lying blood pressure	1.12 (1.06, 1.19)
Diastolic lying blood pressure	1.32 (1.18, 1.49)
Retinal perfusion pressure	1.38 (1.18, 1.61)
Pulse pressure	1.09 (1.01, 1.17)
Postular systolic pressure	1.13 (1.03, 1.25)
Mean arterial pressure	1.26 (1.14, 1.39)

Hot-deck mean imputation applied to some factors

Table 5.10: Odds ratio for blood pressure measurements

A comparison of the odds ratio in table 5.7, page 183 and the corresponding results in table 5.10 on page 186 indicate that most of the estimates and their associated 95% confidence intervals are similar. There are though factors, for example diastolic lying blood pressure (`dbply`) for diabetes type 1 patients, where the mean estimate and 95% upper limit for the odds ratio might be regarded as markedly different.

In addition to that, postural systolic pressure (`posys`) for diabetes type 2 patients and lying pulse pressure (`pply`) for both diabetes type patients have small differences in the lower 95% limit of the odds ratio. Nevertheless, the differences for the factors mentioned above affect the conclusion

that can derived from this type of analysis.

For the complete cases only data, the 95% confidence interval for the three factors described above extents from values below 1 to values above 1. Thus, we can not conclude whether a change with the specified range shown in tables will significantly alter the risk of having sight-threatening retinopathy.

On the other hand, using the hot-deck imputation analysis results, the lower limits for the 95% odds ratio confidence intervals contain only values above 1. Hence, these results suggest that the `stri` risk reduction can be achieved by reducing these blood pressure measurements. Clearly, this contradiction of conclusions described above can be attributed to the difference between the samples that have been used for this type of statistical analysis.

Multiple imputation and imputation by implementing the EM-algorithm might be other possible methods of obtaining estimates of the odds ratio for the blood pressure measurements shown above. In both cases, the procedures need to applied with great care as the imputation model could have substantial influence on the results. Sensitivity analysis is a possible way of examining the influence of the imputation model on the results.

As we are dealing with a longitudinal study, where some patients have additional measurements to the ones at entry, it might be possible to impute the missing values at visit 1 by the subsequent entries. Longitudinal imputation methods that are briefly described in Twisk et al. (2002) include last value carried forward, linear interpolation and individual longitudinal regression. Also multiple imputation can be implemented, where measurements from other visits are also included in the imputation model.

The amount of data collected for the EKSAGE study is enormous and due to time limitations, it was not possible to include all the visits recorded at the current stage. In addition to that, the purpose of the analysis described previously assumed that only the data at entry is available and this situation is likely to arise in future applications of the results.

In other words, as it is possible that there might be missing or doubtful entries for some of the factors for several reasons, we need to have procedures that deal with missingness in a way that it will be possible to be implemented in clinical practice. It might not difficult to construct an imputation table, similar to a classification tree model, where the missing values for a patient might be replaced by sensible replacements. Nevertheless, whenever it is possible, statistically sound methods should be used to avoid having biased or even invalid results.

5.12 Graphical modelling

Classification and regression models for prediction of sight-threatening retinopathy indicator (variable `stri`) at entry could be the next step in our statistical analysis. According to the clinicians' opinion, the key issue is not to make successful forecasting about the eye condition at entry but to identify the factors that are associated with the presence of advanced diabetic eye disease complications.

In practice, the factors that could be used as predictors for response variable `stri` will be usually collected and recorded at the eye hospital where soon after tests will be performed to examine the eyes of the patient. Hence, even if we were able to predict diabetic retinopathy at entry, with the current clinical practice it will be useful only for a short period of time.

A potential application of a predictive model might be to construct a case selection tool for diabetic patients if it is not possible to check all patients in a meaningful period of time. To be specific, it might be possible that due to the large number of patients or because of limited health resources, patients will be at danger of loosing their eyesight as a result of delayed checking and identification of diabetic retinopathy. In this case, identifying the diabetic patients that are most at risk will be valuable.

At this stage, it is important to investigate the relationship between the risk factors in the EKSAGE study that are likely to be collected at the first eye hospital visit. The exploratory analysis described in previous paragraphs is an indication of the complicated association between the covariates and the response variable `stri`. Furthermore, it is possible that association and influence structure will be changing with time (Klein et al. 1995).

Graphical modelling has been used by de Fine Olivarius et al (2001) to investigate the relationship between diabetic retinopathy and other related risk factors. In this case, the authors have implemented a two-block chain graph (Edwards 2000) and subsequently examined the association structure between all the variables except diabetic retinopathy and the influence strength of all predictors on diabetic retinopathy indicator.

The chain graph described above has been constructed in such a way that it is similar to a multinomial logistic regression model where the distribution of the categorical response variable (diabetic retinopathy indicator) conditional on all the other variables is examined. At the same time, the association structure with age and sex connected by undirected edges with blood pressure measurements and other related hemodynamic factors might be seen as dubious.

As described in Edwards (2000), the chain graph structure should reflect the possible causal ordering of the variables included in the analysis. Hence, an undirected edge between age and

blood pressure described in de Fine Olivarius et al (2001) might not be sensible as it is possible to have an effect of age on blood pressure but it is unlikely that blood pressure will alter age.

Furthermore, the study by de Fine Olivarius et al (2001) was related to newly diagnosed type 2 diabetic patients whereas in the EKSAGE study both type type 1 and type 2 patients are included. Moreover, the individuals in our investigation have been referred to a diabetic eye clinic because of possible eye complications and the diabetes duration for a large number is not comparable to the corresponding in the paper by de Fine Olivarius et al (2001).

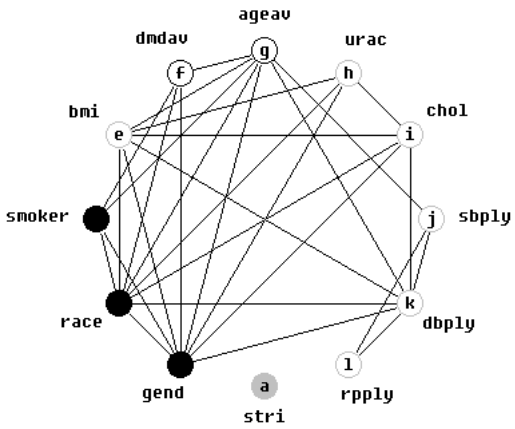
Finally, the authors mentioned above have implemented their chain graph construction method in the statistical software Digram. As this program allows only categorical variables to be included, continuous variables have been categorised, a fact that might lead to loss of information and the possibility of having biased results. In our case, we have implemented chain graph construction in the statistical software MIM, where it is possible to include both categorical and continuous variables (Edwards 2000).

The chain graph in our case will consist of two blocks that will contain the covariates and the response variables. Based on the information from the literature and the clinicians, the factors used for diabetes type 1 patients analysis are:

- e: bmi (BMI)
- f: dmdav (diabetes duration at entry)
- g: ageav (age at entry)
- h: urac (uric acid)
- i: chol (cholesterol)
- j: sbply (lying systolic blood pressure)
- k: dbply (lying diastolic blood pressure)
- l: rpply (retinal perfusion pressure)
- a: stri (sight-threatening retinopathy indicator)
- b: gend (gender)
- c: race (ethnicity)
- d: smoker (smoking level)

In this case, we assume that gender (variable b), age at visit (variable g), diabetes duration (variable f), race (variable c) and smoker status (variable d) are the covariates and the other variables are the responses. As there are continuous covariates and discrete responses, we implement CG-regression analysis as described in Edwards (2000).

The result of CG-regression is shown in figure 5.14 on page 190. From this graph, we can see that sight-threatening retinopathy indicator (variable `stri`) is not significantly associated with or influenced by other factors included in this type of analysis. Additional information can be extracted from figure 5.14 on page 190 for the importance and strength of the associations between factors and the influence of the covariates on the response variables.



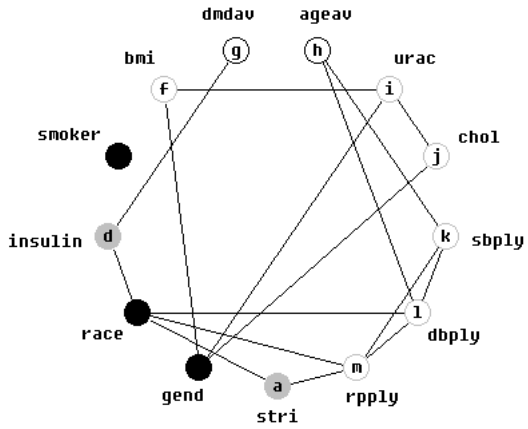
Stepwise and CG-regression applied in MIM

Figure 5.14: Association and influence structure for EKSAGE diabetes type 1 patients

The corresponding association and influence structure for diabetic type 2 patients, where insulin treatment indicator is included in the response variables, is shown in figure 5.15 on page 191. In this occasion, the sight-threatening indicator is associated only with race and `rpply`. CG-regression could not be fitted for some unknown reason and the result was obtained by stepwise backward procedure. The results from both figure 5.14 on page 190 and figure 5.15 on page 191 might be repeated with hot-deck imputed data or with other combinations of blood pressure measurements.

5.13 Conclusions

In this chapter, we presented a case study to illustrate the use of different statistical methods to identify risk factors related to sight threatening retinopathy (STR) at first visit of diabetic patients to the eye clinic. First, by univariate exploratory analysis we found that gender, smoking level and body mass index are not statistically significant as predictors of STR at 5% level but nevertheless



Insulin treatment added in response variables

Figure 5.15: Association and influence structure for EKSAGE diabetes type 2 patients

should be included as control variable as they are factors of "known clinical importance". Race is shown to be significant predictor of STR for diabetes type 2 patients.

Using a scatter plot matrix to implement bivariate exploratory analysis, we showed that the association between diabetes duration at entry and age at entry changes from negative to positive by controlling for diabetes type. Multivariate exploratory analysis using conditioning plots showed the complexity of the association between systolic blood pressure and duration of diabetes.

Different blood pressure measurements have been tested as determinants of the presence of STR adjusting for gender, race, smoking level, body mass index, age at entry, diabetes duration at entry, uric acid and cholesterol; for type 2 insulin treatment was added in the group of the control variables. We had demonstrated by comparison of "complete observations only" analysis and "hot-deck imputation", postular systolic pressure for diabetes type 2 patients and lying pulse pressure for both diabetes types changed from non-significant to significant predictors. This illustrated that the imputation of missing values is of great importance and should not be regarded as nuisance that should be dealt with in a "quick and dirty" way.

Finally, we used graphical modelling to allow for complex relationships between the variables in the data. For diabetes type 1 patients, STR is not associated with any other factors whereas for type 2 STR is linked with race and retinal perfusion pressure. This showed that apart from the statistical significance of an association, the clinical importance should be taken into account.

Chapter 6

Conclusions and suggestions for further research

6.1 Conclusions

From the contents of this thesis, it is apparent that the application of statistics in medicine is by no means a trivial task. Also, it can be said that there is no definitive or unique way of performing a specific type of statistical analysis, even at the stage of exploratory analysis. As we have seen throughout this study, at several stages of this project it was necessary to choose between different techniques and methods that seem suitable for solving a particular problem.

The complicated nature of the data with missing and sometimes unreliably reported or recorded entries is only the tip of the iceberg of the challenges that are encountered by medical statisticians. We have shown in many occasions the importance of data cleaning and coding in the two case studies included in this thesis. In addition to that, the medical expert's knowledge and guidance into understanding the data and subsequently the construction and assessment of the corresponding risk model has been proven valuable.

The exploratory analysis has been shown to be particularly important into finding the structure of the data and identifying important predictors for medical complications. At the same time, the limitations of preliminary statistical analysis are also evident, a fact that is often ignored by clinicians insisting on presenting only "simple" and "well-established" methods in medical publications.

For the AAA case study, we compared different classification models and we found that two logistic regression models are giving results that suggest that these models can be used for selective screening. The estimate of the Bayes risk suggests that more complex models would not make much

difference as there is evidence of significant overlap between classes. In addition to that, cross-validation and bootstrapping should be implemented as the performance of the classifier should be considered on the part of the data that has not been used for constructing the classification model.

Graphical modelling has been demonstrated as particularly useful for identifying complex data structures. For the AAA screening data, the area under the ROC curve for the Naive Bayes classifier (NBC) is less than the corresponding area of the logistic model. This might be attributed to the different way the logistic and NBC models classify cases into classes. Discriminative classification methods such as the logistic model are estimating directly the posterior probabilities for each class.

Generative classifiers such as the NBC estimate the corresponding probabilities model by first modelling the class-conditional probability for each class and then employ the Bayes' rule. Chan et al. (2002) state that "it is widely believed that discriminative classifiers are to be preferred since the discriminative criterion is more closely related to the classification error". The same might be true when Occam's window selection method has been implemented using NBC models.

The graphical modelling has been particularly successful when implemented for identifying the possible normal components of the aortic diameter. The fact that the results have been confirmed independently by a minimisation procedure and agree with the clinical definitions of abnormality suggest that this might be regarded as a sensible result. Further confirmation of these results need to be considered using independent samples from other studies.

For the second case study about sight threatening retinopathy (STR), we clearly demonstrated that relying only on statistical significance for identifying important risk factors is not appropriate for this dataset. Specifically, gender, smoking level and body mass index are not statistically significant as predictors of STR but they are included as control variables because clinicians have suggested that they have been shown as important in other major studies.

On the other hand, using "complete cases only" and "hot-deck imputation", we have demonstrated that these two methods that are common in the medical literature for dealing with missing data give contradicting results. In this case, the clinicians' opinion does not lead to robust results and we need to consider alternative methods for imputation.

The overall conclusion of this thesis is the fact that applying statistics in medicine is an optimisation process that requires a multidisciplinary approach that requires the involvement of both clinicians and statisticians. We showed that finding the optimal risk model requires combination of simple and advanced statistical methods. The limitations of medical studies with regards to the time and the data available and also the capabilities of the statistical software should be taken into account. Modifications to the methods used might be necessary to avoid damaging the quality of

the study. Finally, both statisticians and clinicians should be prepare to learn from each other and from the whole process of using statistical methods in medical research.

6.2 Suggestions for further research

There are aspects of this study that suggest further investigation and possible improvements. For example, instead of excluding patients with missing values or using hot-deck imputation to avoid complex imputation methods (diabetic retinopathy case study), it might be preferable to apply multiple imputation or maximum likelihood imputation with the EM-algorithm.

Another possibility is to use sensitivity analysis to examine the effect of using different imputation models and techniques on the odds-ratio estimates. For the abdominal aortic aneurysm, it might have been better to impute the doubtful entries of smoking level indicator by methods that take into account that the imputed values are levels of a categorical variable. Also, additional analysis might indicate the variables that predict missingness or inaccuracy in both cases studies.

As the statistical methods that have been included in this study are improved or extended, it might be possible to repeat the analysis with updated or alternative versions of these methods. For example, when this research started, there were few packages for applied graphical modelling that did not require advanced statistical knowledge or the use of prior information. Currently, the freeware statistical package R includes a variety of libraries for graphical models with manuals that include examples of implementing this type of analysis, thus making it easy for a non-statistician to understand the basic principles and the implementation procedure.

Additional data can be always useful into extending or revising the results of medical research. For example, longitudinal data about aortic diameter growth and other related predictors might be proven crucial into understanding the pathogenesis of abdominal aortic aneurysm. In addition to that, graphical modelling analysis can be extended in this way into the discovery of association and influence relationships of other variables over time.

Multilevel analysis, which is "a methodology for the analysis of data with complex patterns of variability, with a focus on nested sources of variability" (Snijders et al. 1999) could be a possible way to investigate the effect of clinical practice on prevalence of aortic aneurysm. In this case, we assume that clinical practices in the study are a random sample and we are interested in explaining the variability between individuals in the study taking into account the hierarchical structure of the data.

Furthermore, it is possible to implement multilevel models for investigating the aortic diameter growth patterns and compare the results to the corresponding ones from graphical models. Possible

ways of implementing this type of modelling are related functions and procedures available in packages such as R, S-Plus, SAS and Stata. For more sophisticated and flexible multilevel analysis, it might be worth considering packages such as MLwiN and WINBUGS; see Goldstein (2003) for further information about multilevel modelling software, resources and further developments.

We would also like to propose a few ideas for further research that are related to selective screening and its application to abdominal aortic aneurysm screening programs. For example, the selection process might be improved in the future by additional information for the individuals to be screened, such as genetic information.

Furthermore, the new strategy proposed by the UK National Screening Committee for the implementation of elective screening schemes should be investigated in detail. To be specific, more information is needed about the influence of choosing or declining an invitation for screening on the general health of the population and the cost-effectiveness of a program. It will also be valuable to investigate the factors affecting the individual's decision about participating in screening schemes.

Moreover, there is much knowledge that is available from previous trials which should be used when designing and applying screening programs. Moreover, the combination of previous and new information is more likely to lead to a successful strategy. Hence, co-operation with other researchers in the same field might be proven an important way of improving the effectiveness of the screening programme.

For the second case study, time limitations have not permitted the analysis of the data for visits other than the first. Progression of diabetic retinopathy and the effectiveness of laser treatment might be understood further by analysing the complete series of measurement for as many patients as possible.

Multilevel event history models might be useful for the situation described above as we have to deal with several measurements for each individual and the possibility of censored observations. Additionally, as diabetic retinopathy progression and the corresponding treatment can be regarded as correlated event histories, it might be necessary to consider *multilevel multiprocess* event history models (Goldstein 2003).

Another issue is the additional inclusion of the biochemical marker HbA_{1c} that is the current indicator of the glucose level in the blood. The main obstacle in the inclusion of HbA_{1c} has been the unavailability of this measurement for all patients that entered the study before 1979.

Furthermore, for the patients that came to the eye hospital between 1979 and 1987, HbA_1 has been measured, the values of which can not be easily converted into HbA_{1c} units. Again, it is possible to use information from the experts or the literature to convert HbA_1 to HbA_{1c} using

different conversion formulae and subsequently conduct sensitivity analysis to estimate the extra uncertainty imposed by the lack of knowledge of HbA_{1c} for all patients.

We would like to state that this study is by no means exhaustive in terms of identifying the difficult issues that medical statisticians or other medical experts might face when constructing a risk model and afterwards assessing its validity. One of the objectives of this study has been to understand and describe the difference between knowing a particular statistical technique and its potential applications and implementing this technique and subsequently presenting the results to non-statisticians.

It can be said that there is a lot of work that needs to be done into clarifying to non-statisticians the possible disadvantages a method that they understand and prefer to apply and is inappropriate for the question they want answers for. Methods such as logistic regression, classification trees, neural networks and graphical models have become available in the last few years in many statistical packages and the medical experts are becoming familiar with these techniques and the benefits of using them in medical research.

Finally, it should be remembered that for different people, the best treatment for a medical problem is sometimes defined in different ways and it is not always possible to have a unique solution. Subjectivity plays a crucial role into choosing the criteria by which screening programmes are evaluated and risk models' performance is assessed. Thus, the construction and assessment of risk models in medicine should be regarded as an optimisation problem with constraints that need to be identified well before attempting to find a possible solution.

Appendix A

Standard statistical methods

A.1 Exploratory analysis

A histogram is a plot of the different values of a variable or groups of values if the number of the values is large against the number of cases or the frequency of them. The grouping of values in order to define the number of groups and the break points that separate groups are usually predefined by a set of formulae. Some statistical packages allow more flexibility into constructing a histogram by allowing the user to change the number of groups and the breakpoints if the initial options seem to be inappropriate or insufficient to summarise the information about a particular variable.

Stem-and-leaf can be regarded as "an enhanced histogram" (Venables and Ripley 1997, page 170). The values of the variable are divided into groups and instead of having the height, as in the case of a histogram, the second digit of the value is printed. It is not particularly successful when having wide range of possible values; in this case, the initial way of plotting should be changed in order to summarise values in a different way.

A boxplot is "a way to look at the overall shape of a set of data" (Venables and Ripley 1997, page 172). The median, as well the values of the quartiles are plotted. Values defined as extreme (e.g. outliers) are plotted in the edges of the graph; the cut-off point for extreme values is predefined by a related formula and usually is more or less than 1.5 times the interquartile range (the range between the 25th and the 75th percentile).

In particular, "the normal plot is based on two ideas. First, the cumulative frequency distribution gives a better idea of the shape of the data than does the frequency distribution. The cumulative frequency distribution for data that are Normally distributed has an S shape. It is

however, difficult to judge Normality from the cumulative frequency distribution, which is where the second idea comes in. Because all Normal distributions are precisely the same shape, we can stretch the vertical scale to make the cumulative distribution function a straight line if the data are Normal. Departures of the sample data from Normality are thus easily seen as departures from a straight line" (Altman 1999, page 133).

A.2 Univariate tests

A.2.1 Univariate analysis for continuous data

A simple test for continuous variables available in the majority of statistical packages is a t-test, usually applied for comparing one or two groups of observation. To be more specific, we consider this test when for instance "we wish to compare the mean of a single group of observations with a specific value" (Altman 1999, page 183). Furthermore, "the use of t-test is based on the assumption that the data for each group (with independent groups) ...have an approximately normal distribution" (Altman 1999, page 199).

The first alternative, when we do not have Normal distribution, is to transform the data. For example, we could use log transformation if the variable skewed; the results though would not be on the original scale and we would need to back-transform in order to have the results on the original scale. The same rule applies when trying to derive a confidence interval for the mean of distribution: transformation of the data almost always requires back-transformation of the results.

The other alternative, when the variable does not follow Normal distribution is to use non-parametric tests. These tests are also known as "distribution-free tests" and "they hold under relatively mild assumptions regarding the underlying populations from which the data are obtained" (Everitt et al. 2001, page 72). For a single sample two tests that are commonly used are the sign test and the Wilcoxon sign rank sum test (Altman 1999, page 186).

When comparing two groups of observations, there is a distinction between tests applied to independent groups and the corresponding tests for paired groups. For example, when comparing the mean age of a group of healthy with a group of ill individuals, we are referring to independent groups. In this case we might have different sample sizes for these groups. On the other hand, comparing the mean length of right and left arm for a number of people refers to paired groups and the assumption required are different from the ones for independent groups.

To be specific, for independent groups, "parametric methods require the observations within each group to have an approximately Normal distribution, and the standard deviations in each

group should be similar. ... For paired data ... there is no assumption that each set of observations should be normally distributed, but there is a different assumption of Normality" (Altman 1999, page 180). The corresponding assumption of Normality for paired differences is that the differences between the values of each pair are normally distributed.

The null hypothesis when applying a t-test for independent groups is that the means of these groups are equal. The significance level for rejecting the null hypothesis is usually $\alpha = 5\%$. It is also possible to derive the $100(1-\alpha)\%$ confidence interval for the difference between two means; for details about the formulae required see Everitt et al. 2001 (page 65).

The corresponding non-parametric test for comparison of two independent groups is the Mann-Whitney which is also called the Mann-Whitney-Wilcoxon test. The null hypothesis tested is that the two groups have the same distribution and the alternative hypothesis is that they have different location. Further details about this test can be found in Altman (1999, page 194).

For paired groups of observation, "we are interested in the average difference between the observations for each individual and the variability of the differences. By looking at these differences we effectively reduce the analysis to a one sample problem" (Altman 1999, page 189). The corresponding non-parametric method for paired groups is the Wilcoxon matched pairs signed rank sum test.

When comparing three or more independent groups of observations for a continuous variable, instead of comparing two groups at a time, it is better to conduct a single analysis called one way analysis of variance (ANOVA).

In this case, we partition the total variation into two components: variation between individuals within groups and difference between group means. "The test is based on a comparison of the observed variation between the means of the groups with that expected from the observed variability between subjects. The comparison takes the general form of an F test" (Altman 1999, page 206).

Subsequently, if we find that the groups are significantly different, it is possible to compare the groups pairwise in order to identify the groups that contribute mostly to the significant result of the ANOVA test. The danger in this case is that, as each test has 5% chance of rejecting the null hypothesis, when actually it is true, the probability of type I error is increased because of the large number of tests conducted.

A possible way to solve this problem is using an adjustment for the p-value called the Bonferroni method. The assumptions in one way analysis of variance are the normality of group observations, the equality of group variances and the independence of observations.

It also possible to apply the corresponding non-parametric way analysis of variance, which is the Kruskal-Wallis test. This test is "an obvious mathematical extension of the Mann-Whitney test" (Altman 1999, page 213).

A.2.2 Categorical data analysis

When allocating subjects in a study into one of the possible groups according to one or more of their characteristics, then we can study these groups by examining a frequency table indicating the proportion of subjects that belong to each of these groups. Depending on the number of groups which define the number of levels or categories for the variable we construct to describe the attribute that separates the subjects, we have to implement different types of statistical analysis.

To be more specific, "when there are only two categories for one of the variables, for example whether a patient has a particular symptom or not, the data can be summarised as the proportion of the total number of individuals in one of the categories" (Altman 1999, page 229). In addition to that, the results from analysing the data of this type as a frequency table should be the same as when analysing the same information using proportions.

For one group of individuals, it is possible to examine whether the proportion of cases with a particular characteristic is equal to a specific number and also estimate the confidence interval in which this proportion lies. When testing the null hypothesis that the proportion is equal to a particular quantity, it might be necessary to use a continuity correction, particularly if the sample size is small, as the method applied for this test "uses the continuous Normal distribution as an approximation to the discrete Binomial distribution".

If we are dealing with two independent groups and we want to compare them by using a categorical variable with two levels, one possible way to do that is to compare the proportion of subjects that belong to the category of interest. In this case, we need to estimate the difference between the proportion in each group and estimate if necessary the corresponding confidence interval for this difference.

Another option is using hypothesis testing to examine whether the difference between the proportions mentioned above is significantly different from a quantity of interest. Setting this quantity equal to zero is equivalent to testing if these proportions are significantly different. A continuity correction might be also necessary, depending on the sample sizes of the groups.

A different approach is needed when comparing two proportions on paired observations. The reason for this is that "the groups are individually matched, thus we should not treat the observations as independent" (Altman 1999, page 236). In this occasion, we split the sample into four

parts depending on the presence or absence of the characteristic of interest in each of pair members and subsequently we analyse the data by estimating the difference between the proportions. The formula applied for this difference and the associated standard error can be seen in Altman (1999).

An alternative approach for comparing different groups when using categorical variables is to analyse frequency tables constructed by cross-tabulation of two categorical variables. Specifically, we compare the number of cases expected in each of the frequency table cells with the corresponding number of observed cases. The formula applied is related to the square of the difference between observed and expected observations and X^2 statistic estimated by this formula follows χ^2 distribution.

A specific requirement for this type of analysis "attributed to the statistician W. G. Cochran is that 80% of the cells in the frequencies table should have expected frequencies greater than 5 and all cells should have expected frequencies greater than 1" (Altman 1999). For 2 by 2 tables only, when the sample sizes are small and the requirement mentioned above is not met, then Yates' continuity correction is used.

In Altman (1999, page 253) it is mentioned that for 2 by 2 tables, "there is an alternative approach for tables with very expected frequencies, known as Fisher's exact test". In this case, it is assumed that the row and column totals are fixed; see Altman (1999, pages 254-257) for further details. If the samples are paired, then it is not possible to use the standard χ^2 test. McNemar's test, known also as the test of paired proportions should be applied.

If the categories of the variable have an ordering, "we should make use of the ordering to increase the power of the statistical analysis. When the groups are ordered we usually expect any differences among the groups to be related to the ordering. Failure to take account of the ordering of the groups is a common statistical error" (Altman 1999, page 261).

The χ^2 test for trend is applied to examine if the variation among groups can be attributed "to a trend of proportions across the groups and the remainder. The method of evaluating a trend is effectively to fit a straight line to the proportions and see if the slope of the line is significantly different from zero" (Altman 1999, page 261). An alternative method in the case of ordered categorical variables is apply the Mann-Whitney statistic.

Finally, for 2 by 2 tables, it might be necessary to compare the risk of having a disease between two groups of patients using one of their characteristics. Relative risk and the odds ratio are possible ways of estimating the risk difference between these groups. For combining several 2 by 2 tables, it is possible to combine their results in order to obtain an overall assessment of the risk. Mantel-Haenszel method is applied to obtain the combined odds ratio.

All the tests mentioned above are usually implemented when it is required by the objectives of the study, e.g. for comparing two or more groups in the data. Additionally, the results from these tests can be used for further analysis such as model construction. Specifically, the significance level of a comparison test between groups, when one or more variables are used for separating subjects into these groups, can be regarded as indication of the importance of these variables.

For example, if the results of a t-test for independent groups indicate significant difference between the mean age of these groups, then it might be suggested that age should be included in a statistical model in order to control for age effects on the result. When having a large number of possible predictors, hypothesis testing can be a useful way of identifying the most significant ones and included initially only them in the model.

A.3 Bivariate analysis

A.3.1 Graphical methods

A scatter diagram or scatter plot as it is usually called is a graph where the values of a variable are plotted against the values of a second variable in the dataset. From this type of graph, we are able to investigate the relationship between two variables by the shape that the points in the scatter plot have. If for example, these points seem to lie on a straight line, then we can infer that the bivariate relationship in this case is approximately linear.

Furthermore, for two linearly related variables, if increase in the values of one variable are associated with increase in the other variable of the pair we investigate, then we can conclude that these variables are positively associated or positively correlated. On the other hand, increase of the one variable that results in decrease of the other means negative association. Finally if the points lie into a cloud where there is no distinct pattern, there the variables can be said to be unrelated.

Depending on the shape of the pattern of points, it is also possible to infer other types of association. For example, we can conclude that there is non linear relationship between two variables if the points seem to lie on a curve. In addition to that, we can use the scatter plot to examine whether a linear relationship between variables exists after transforming one or both variables.

According to Everitt et al. (2001, page 96), "an important parameter of a scatterplot that can greatly influence our ability to recognise patterns is the aspect ratio, the physical length of the vertical axis divided by that of the horizontal axis". In the example given, plotting the points in the scatter diagram after restricting the range of the variable to the time of particular interest, a

trend is clearly revealed that is was not apparent when the original scale has been used.

Finally, it must be remembered that a scatter plot is useful for describing the bivariate relationships in the data but should not be regarded as formal statistical analysis. The information derived from this type of plot should be combined with the results from other types of analysis, e.g. univariate tests, in order to draw firm conclusions about the nature of the data and suggest further statistical analysis that might be necessary.

It is possible to estimate density of two variables by enhancing a scatterplot. This is done by estimating bivariate density functions; for further details see Everitt et al. (2001, pages 102-103). Furthermore, smoothers can be included in the scatter diagram, especially when the bivariate analysis indicates that the association between two variables is complex. Local regression models can be fitted to the data by using polynomials of a specific degree; the result also depends on the smoothing parameter of the function.

A.3.2 Correlation

If we want to quantify the degree of association between two variables, one possible way to do that is to apply a method that estimates the correlation coefficient or as it is called for simplicity the correlation between the variables. The standard method involves the computation of the correlation coefficient related to Pearson's method and it "measures how close the points are to a straight line" (Bland 1996).

To be more specific, correlation coefficient takes values that lie in the interval [-1,1]. If this value is positive, this means that the variables are positively associated and if it is equal to 1, then the points in the scatter plot for these variables lie on a straight line. On the other hand, a negative correlation indicated negative association between the variables. A value close to zero shows no bivariate relationship.

Furthermore, it is possible to estimate the confidence interval for Pearson's correlation coefficient and also to test whether this coefficient is significantly different from a specific value. The associated hypothesis test is valid when "the two variables are observed on a random sample of individuals and the data for at least one of the variables have a Normal distribution in the population. For the calculation of a valid confidence interval, both variables should have a Normal distribution" (Altman 1999, page 279).

If the variables do not follow Normal distribution, then it might be possible to transform one or both the variables to achieve Normality or to use a non-parametric method to assess the degree of association between these variables. Another assumption that is made in order to have valid

estimate of Pearson's correlation coefficient is the independence of the observations involved. To put it in other words, "the analysis is not valid when there is more than one observation for some or all subjects" (Altman 1999, page 282).

An addition aspect related to the estimation of correlation that we need to take into account is the nature of the sample that is used. A few outliers might have substantial influence on the result; we should try to have a random sample from the population we want to study. Also, the correlation of a mixture of different samples would have unexpected and misleading results; it is better to apply this method to the different subgroups separately.

Moreover, when we are dealing with longitudinal data, where several measurements are taken from each individual for a number of variables, we should be very careful when assessing the correlation of pairs of variables that are measured over time. Specifically, this correlation might be due to time trends; for example blood pressure measurements are known to follow a daily cyclic pattern and apparently these daily measurements will be highly correlated. Hence, it might be necessary to remove the time trends (Altman 1999, page 283).

As mentioned previously, non-parametric correlation methods are applied in the case of variables that do not have Normal distribution. The two popular methods of this type, namely Spearman's and Kendall's correlation coefficients are based on ranking the variables under examination and then calculating the corresponding rank correlation by comparing the orderings. This rank correlation coefficient, using one of the two methods mentioned above, "has the advantage of not specifically assessing linear association but more general association" (Altman 1999, page 287).

Furthermore, it is possible to estimate the confidence interval for the rank correlation and it can be applied to categorical data as well as to continuous data. As rank correlation is regarded as less restrictive than Pearson's correlation, it should be preferred especially in situations that it is difficult to assume that the variables involved follow Normal distribution.

A.4 Multivariate exploratory analysis

A.4.1 Principal components analysis

The basic idea, when applying principal components analysis, is to find linear combinations of the variables to describe the variation of a set of multivariate data (Venables and Ripley 1997, page 382). Additionally, these linear combinations "are derived in decreasing order of importance so that the first principal component accounts for as much as possible of the variation of the original data, the second for as much of the remaining variation subject to being uncorrelated with the

first, and so on" (Everitt et al. 2001, page 381).

The objective of using this method is to find a few components in order to reduce the number of variables without substantial loss of information. Additional benefits of applying this method is the information we might get about the structure of the data and also the degree of association between the variables and whether it is possible to summarise them or if we should retain the original variables in further analysis.

The coefficients are the eigenvectors of the covariance or correlation matrix; the correlation method is used when the variables have different scales. Moreover, the eigenvalues of these matrices are the variances of the principal components. In the presence of outliers, we need to replace the correlation matrix by the minimum volume ellipsoid estimator, which is robust in terms of influence of maverick observations.

Choosing the number of components required depends on subjective criteria. A common method is to select the components that describe a specific proportion of the variance; usually this figure lies between 70% and 90%. Also, the results of principal components analysis are derived from the correlation method, we can choose the components that have eigenvalues which at least equal to one.

The coefficients of the linear combinations of the original variables are called loadings of the principal components. These provide important information about the structure of the data as well as the contribution of each of the original variables in each of the components. For example, if a group of variables have coefficients in the loadings with different signs, this means that the contrast between these variables is important for the purpose of finding ways of describing the variability between observations.

Finally, it is possible to select the number of components by plotting the eigenvalues in the screeplot, which is "a barplot of the variances of the principal components" (Venables and Ripley 1997, page 384). In this case, we are looking for "an "elbow" in the curve" (Everitt et al. 2001, page 383); in other words, we exclude components that do not seem to make any difference in describing variability in the data.

As linear combinations of the original variables are constructed, it is not possible to apply principal components analysis when categorical data are included in the dataset. Nevertheless, it is possible to combine a large of continuous variables into a small number of components, especially if there are a lot of highly correlated continuous variables. A possible drawback of this could be that we do not know if possible bias is introduced on the relationship between categorical and continuous variables or the associations between categorical ones when the effect of some of the

continuous variables is substantial.

A.4.2 Factor analysis

Similar to principal components analysis is another method called factor analysis. In this case, the observed variables are regarded as expressions of "underlying fundamental quantities called factors and factor analysis seeks linear combinations of the factors" (Venables and Ripley 1997, page 405).

Moreover, "the correlation between each pair of observed variables results from their mutual association with these factors" (Everitt et al. 2001, page 389). Hence, the observed variables are conditionally independent given the factors; it should be stressed that these factors are latent variables. Description of the underlying methodology of finding factors in this type of analysis is given in Venables and Ripley (1997) and also in Everitt et al. (2001).

The coefficient of the latent variables are known as factor loadings. In order to have factors that we can easily interpret, we might use rotation of the loadings matrix. Furthermore, it is possible to select the number of factors needed to describe the data by using an appropriate test; see Everitt et al. (2001, page 393) for further details.

Both principal components and factor analysis attempt to explain the observed correlations or covariances and in many cases will lead to similar results. In addition to that, both methods are ineffective when the observed variables are "almost uncorrelated" (Everitt et al. 2001, page 396).

On the other hand, there is an important difference between these two methods. Factor analysis is based on the hypothesis that "a set of latent variables exist and they are adequate to account for the covariances of the variates, although not for their full variances. Principal components analysis is merely a transformation of the data and they have no part corresponding to the specific variates of factor analysis" (Everitt et al. 2001, page 396).

A common problem for both methods is handling categorical data and especially variables with no ordering between different levels, namely nominal data. As mentioned above, it might be possible to apply factor or principal components to the continuous part of the data, but we should be cautious when including the results into further analysis.

A.4.3 Multidimensional scaling

Distance methods are "methods based on representing the cases in a low-dimensional Euclidean space so that their proximity reflects the similarity of their variables. To do so, we have to produce a measure of similarity" (Venables and Ripley 1997, page 385). The default distance in statistical packages is Euclidean; dissimilarities are distances used when categorical variables are included in

the data and are related more to cluster analysis.

Multidimensional scaling is a popular distance method "which seeks a configuration such that distances between points best match those of the distance matrix". Further details about this method can be found in Venables and Ripley (1997, page 385). It should be stressed that this method, when using Euclidean distance and the results are plotted, is equivalent to plotting the first principal components.

Again, the drawback of this method is that it is not suitable for categorical data. Hence, we need to apply a similar method, such as cluster analysis, especially when we are dealing with nominal data.

A.4.4 Cluster analysis

In medicine, "discovering that a large set of patients can be partitioned into a small number of groups or clusters, within which patients have very similar characteristics may have important implications both for treatment of the disease and for understanding its etiology" (Everitt et al. 2001, page 401).

Cluster analysis is related to discovering group structure amongst the cases in the dataset. This method is based on "a measure of similarity or dissimilarity between cases" (Venables and Ripley 1997, page 389). There are several possible measures of similarity and dissimilarity; these measures are labelled in the literature as coefficients. The authors mentioned above state that "for categorical variables, most dissimilarities are measures of agreement".

There are two broad categories into which cluster analysis fall into. The first is agglomerative hierarchical clustering and its main characteristic is that initially each subject in the study is defined as a separate cluster. Subsequently, similar clusters are merged into a single cluster and this process continues until all subjects belong to a cluster. Different measures of distances lead to different groups or clusters; the interpretation of the clustering and the assessment of success of the method is mostly subjective.

The second category of cluster analysis methods is k-means clustering and the algorithm applied in this case "chooses a pre-specified number of cluster centres to minimise the within-class sum of squares from those centres. As such, it is most appropriate to continuous variables, suitably scaled" (Venables and Ripley 1997, page 391). As it is not possible to examine every possible partitioning of the dataset into k clusters, the general method of finding the optimum clustering is to separate the individuals into k groups and then move each individual to one of the clusters until the clustering criterion has beeb optimised.

Another approach to clustering is to assume that "the population from which the observations arise consists of c subpopulations each corresponding to a cluster"; the objective is to choose the optimum clustering to maximise likelihood. This method is described in Everitt et al. (2001, page 409) under the label "classification maximum likelihood".

A.5 Regression models

A.5.1 Linear regression

First, we describe the type of model where the response variable is on continuous scale. To keep things simple, we describe this method when there is only one predictor used; then, we show how linear regression is applied with more that one predictors, which are also called explanatory variables.

The objective of linear regression analysis is fitting a straight line that gives the best possible prediction of the response variable using the explanatory variable. This means that this line would be as close as possible to the data points. *Residuals* can be defined as the difference between the distance between actual and predicted values of the response; they can be used to assess how well the model fits the data.

To be more specific, least squares linear regression, the standard method for this particular problem, minimises the sum of the squares of the residuals. It also minimises the variance of the residuals; this is "a measure of the "goodness-of-fit" of the line" (Altman 1999, page 302).

The assumptions for using linear regression analysis, are linear relationship between predictor and response and also that the response values for each predictor value are normally distributed and with equal variances (Altman 1999, page 303). Another assumption is that the responses are conditionally independent given the values of the predictors.

The residuals follow normal distribution with mean equal to zero if these assumptions are met; this can assessed by a Q-Q plot of these residuals. Also by plotting the residuals against the predictor values are a possible way of identifying departure from the assumptions mentioned above.

The equation of the regression line can be implemented to estimate the response variable from the predictor. Regression line is an estimate of the mean response for each value of the predictor; it is also possible to derive a confidence interval for this prediction to account for the uncertainty of this estimation.

Furthermore, it is possible to use hypothesis testing to assess the significance of the estimated

slope of the regression line and also the significance of the model in terms of explaining the variability of the response variable. It is important to remember that this variability might be due to unknown sources; this is called *random variation*.

Precautions should be taken to avoid unnecessary problems because of misuse of the regression model. Specifically, Altman (1999, page 317) mentions the importance of avoiding predictions outside the range of the observed data as the relationship between predictor and response might be different outside this range. Also, predictions of the predictor by the response should not be derived by a linear regression model that has been constructed the other way around.

Finally, pooling data from groups where there is big difference between the relationship of response and predictor is not recommended; it is better to analyse these groups separately. It is also important to remember that this difference between groups might be due to a strong effect of other variables on the regression between one predictor and the response. Hence, it is sometimes necessary to construct a linear regression model that takes into account the relationship of a group of predictors and the response variable.

Multiple linear regression models involve a response variable that is predicted by a group of explanatory variables (covariates or predictors). The equation describing the relationship between predictors and response includes a constant term and regression coefficients that correspond to each of the covariates in the model.

Everitt et al. (2001, page 181) describe the interpretation of each covariate coefficient as "the expected change in the response variable associated with a unit change in the corresponding explanatory variable, when the remaining explanatory variables are held constant". In Mosteller and Tukey (1977, page 319), it is mentioned that this interpretation is valid when the response variables "are not closely related, either functionally or statistically". It is preferable to interpret to describe each predictor coefficient as the difference between predicted values when they have the same values for all the other other predictors and a unit difference for the predictor of interest.

Altman (1999, page 337) refers to the possible situations that multiple regression is helpful. Firstly, when studying the relationship between two variables, the possible effect of other "nuisance" variables needs to be removed. Additionally, choosing between possible predictors for prognosis of an event or combining them into a single prognostic index is another situation where multiple regression could be implemented.

Similarly to linear regression with one predictor, we can use hypothesis testing to assess significance of each covariate in the model and also if the models explains a significant amount of response variability. Furthermore, plots of residuals can be important to diagnose possible mod-

ifications of a models such as transforming a group of variables or adding polynomial terms of predictors. Venables and Ripley (1997) describe a number of regression diagnostics available in many statistical packages.

Analysis of covariance is applied when we need to include categorical predictors in a linear regression model. For instance, the regression coefficient of a binary covariate "indicate the average difference in the dependent variable between the groups defined by the binary variable, adjusted for any differences between the groups with respect to the other variables in the model" (Altman 1999, page 339).

In addition to that, a categorical predictor with $k > 2$ levels can be converted to $k - 1$ binary variables. In this case, significance of a coefficient indicates the magnitude of difference between the groups indicated by the corresponding binary variable. In case of discrete predictors with ordered categories, Altman (1999, page 339) suggests that the variables could be included taking into account the ordering; the approach of converting this type of covariate into binary variables is not recommended.

Finally, we should mention that when the response variable is on a continuous scale and all the predictors are categorical, then there are situations where multi-way analysis of variance is applied. For example, when more than one measurement is taken for each person under different experimental conditions, then the analysis is regarded as generalisation of paired t-test; see Altman (1999, page 326) for further details.

A.5.2 Logistic regression

When the response variable in the model is not on continuous scale but discrete, then standard linear regression models need to be modified to take into account this fact. A binary response variable can be included in a prediction model by transforming it using the *logit* transformation. In this case, the appropriate model is called *logistic* regression model.

This model belongs to the general class of generalised linear models. The distribution of the response variable given its mean is a distribution from the exponential family. Possible options of distributions that belong to this family are, among others, the binomial distribution, when the response is binary and Poisson distribution, when the response is number of occurrences of a particular event.

Logistic regression is the most popular solution of predicting a dichotomous response variable. Using the logit link ensures that mean of the response, which represents the proportion of individuals with the specific characteristic, lies in the interval [0,1]. Furthermore, logit link is more

convenient than other links function such as the probit and the complementary log-log as logit transformation represents the log-odds and the exponential of the coefficients represent adjusted odds ratios (Everitt et al. 2001, page 209).

Similarly to linear regression models, hypothesis testing can be applied to estimate the significance of each predictor. Furthermore, goodness of fit tests are used to assess whether the maximised likelihood of a candidate model is significantly different for the corresponding likelihood of the saturated model, which is the maximum achievable likelihood. For further details, see Everitt et al. (2001, page 212).

Appendix B

Definitions for risk factor analysis

	Risk factor absent	Risk factor present
Disease absent	a	c
Disease present	b	d

Table B.1: General 2-way contingency table

- $\widehat{p}$: estimated probability
- D: disease present
- R: risk factor present
- $\overline{\mathsf{D}}$: disease absent
- $\overline{\mathsf{R}}$: risk factor absent

Prevalence of AAA for predictor absent$=\dfrac{b}{a+b} = \widehat{p}\,(\mathsf{D}|\overline{\mathsf{R}})$

Prevalence of AAA for predictor present$=\dfrac{d}{c+d} = \widehat{p}\,(\mathsf{D}|\mathsf{R})$

Prevalence of predictor for normal cases$=\dfrac{c}{a+c} = \widehat{p}\,(\mathsf{R}|\overline{\mathsf{D}})$

Prevalence of predictor for abnormal cases$=\dfrac{d}{b+d} = \widehat{p}\,(\mathsf{R}|\mathsf{D})$

Relative risk of AAA for predictor: $RR = \dfrac{d/(c+d)}{b/(a+b)} = \dfrac{\widehat{p}\,(\mathsf{D}|\mathsf{R})}{\widehat{p}\,(\mathsf{D}|\overline{\mathsf{R}})}$

The ratio of the risk in the group with the predictor present divided by the risk in the group with the predictor absent.

The standard deviation for the logarithm of the relative risk is $sd(\log(RR)) = \sqrt{\dfrac{1}{a} - \dfrac{1}{a+c} + \dfrac{1}{b} - \dfrac{1}{b+d}}$. The 95% confidence interval for relative risk is $\exp(\log(RR) - 1.96 \times sd(\log(RR))$ to $\exp(\log(RR) + 1.96 \times sd(\log(RR))$.

Odds ratio of AAA for predictor: $OR = \dfrac{d/c}{b/a} = \dfrac{\widehat{p}\,(\mathrm{D}|\mathrm{R})/\widehat{p}\,(\overline{\mathrm{D}}|\mathrm{R})}{\widehat{p}\,(\mathrm{D}|\overline{\mathrm{R}})/\widehat{p}\,(\overline{\mathrm{D}}|\overline{\mathrm{R}})}$

The ratio of the odds of having AAA in the group with the predictor present divided by the corresponding odds in the group with the predictor absent.

The standard deviation for the logarithm of the odds ratio is $sd(\log(OR)) = \sqrt{\dfrac{1}{a} + \dfrac{1}{b} + \dfrac{1}{c} + \dfrac{1}{d}}$. The 95% confidence interval for odds ratio is $\exp(\log(OR) - 1.96 \times sd(\log(OR))$ to $\exp(\log(OR) + 1.96 \times sd(\log(OR))$.

Risk of AAA attributable to predictor: $R_A = \dfrac{P(A) \times (R-1)}{1 + P(A) \times (R-1)}$, where R is the relative risk. Also $R_A = \dfrac{\widehat{p}(\mathrm{R}) \times (\dfrac{\widehat{p}\,(\mathrm{D}|\mathrm{R})}{\widehat{p}\,(\mathrm{D}|\overline{\mathrm{R}})} - 1)}{1 + \widehat{p}(\mathrm{R}) \times (\dfrac{\widehat{p}\,(\mathrm{D}|\mathrm{R})}{\widehat{p}\,(\mathrm{D}|\overline{\mathrm{R}})} - 1)}$

If A denotes the presence of predictor and B the presence of AAA, then the risk of AAA attributable to the predictor, denoted by R_A, is the fraction of P(B) that can be uniquely attributed to the presence of the predictor A.

Furthermore, by using the two way contingency table mentioned above, R_A is estimated by $\dfrac{a \times d - b \times c}{(a+b) \times (b+d)}$ and the standard deviation for logarithm of 1-R_A is $\sqrt{\dfrac{c + R_A \times (a+d)}{(a+b+c+d) \times b}}$. The 95% confidence interval for risk of AAA attributable to predictor is $1 - \exp(R_A + 1.96 \times sd(\log(1 - R_A))$ to $1 - \exp(R_A - 1.96 \times sd(\log(1 - R_A))$ (see Fleiss 1981 for details).

Appendix C

Definitions of terms used for screening programmes

	Below cut-off point	Above cut-off point
Disease absent	a	c
Disease present	b	d

Table C.1: 2-way contingency table for selective screening

- $\widehat{p}$: estimated probability
- D: disease present
- R: above cut-off point

Sensitivity= $\dfrac{d}{b+d} = \widehat{p}\,(\mathsf{R}|\mathsf{D})$

The probability that a test result will be positive when the disease is present (true positive rate, expressed as a percentage).

Specificity= $\dfrac{a}{a+c} = \widehat{p}\,(\overline{\mathsf{R}}|\overline{\mathsf{D}})$

The probability that a test result will be negative when the disease is not present (true negative rate, expressed as a percentage).

Positive predictive value= $\dfrac{d}{c+d} = \widehat{p}\,(\mathsf{D}|\mathsf{R})$

The probability that the disease is present when the test is positive (expressed as a percentage).

Negative predictive value= $\dfrac{a}{a+b} = \widehat{p}\,(\overline{\mathsf{D}}|\overline{\mathsf{R}})$

The probability that the disease is not present when the test is negative (expressed as a percentage).

Positive likelihood ratio= $\frac{\text{Sensitivity}}{\text{1-Specificity}} = \frac{\widehat{p}\,(\mathsf{R}|\mathsf{D})}{1-\widehat{p}\,(\overline{\mathsf{R}}|\overline{\mathsf{D}})}$

The ratio between the probability of a positive test result given the presence of the disease and the probability of a positive test result given the absence of the disease.

Negative likelihood ratio= $\frac{\text{1-Sensitivity}}{\text{Specificity}} = \frac{1-\widehat{p}\,(\mathsf{R}|\mathsf{D})}{\widehat{p}\,(\overline{\mathsf{R}}|\overline{\mathsf{D}})}$

The ratio between the probability of a negative test result given the presence of the disease and the probability of a negative test result given the absence of the disease.

Appendix D

EKSAGE project diabetic retinopathy protocol

Decisions agreed by Prof Eva Kohner, Dr Vinod Patel and Dr S Sailesh

Section 1:

Clinical Grading Allocation for the EKSAGE Study.

Five categories (this includes patient with or without macular exudates or macular edema at any stages)

Grade 1 No Retinopathy Grade 2 Background retinopathy Grade 3 Preproliferative Retinopathy Grade 4 Proliferative Grade 5 Advanced Diabetic Eye diseases

*Identifying patients with Maculopathy (Given below: Section 2)

Grade 1 Only Level 10 No Retinopathy

Grade 2 Level 14, 20, 30,35a, 35d, Background Retinopathy Level 43a, 47c

Grade 3 Level 35 b, 35f, 43b, 47a, 47b, 47d, Preproliferative retinopathy Level 53a, 53b, 53c, and 53d

Grade 4 Level 60, 61a, 61b, 61c, 61d, Proliferative Retinopathy Level 65a, 71b, 71c and 71d

Grade 5 Level 65b, 71a, 75, 81and 99 Advanced Diabetic Eye disease

Note:

1) For levels see attached sheet (Appendix1)

2) 35 a includes both venous occlusion and venous loops, however venous occlusion is an earlier stage and falls in to the background retinopathy and Venous loop falls in to Preproliferative. Patients with Venous loop can be separated from "35 a" level which is now included in grade 2, and included in the preproliferative group (grade 3)

3) Patients with RP only ie "61 a" should be in advanced category depending on number of RP (we decided if $\geq$ 3 RP in one field or if RP $<$ 2 but associated with Visual loss and macular traction/ detachment to categorise this as Advanced)

4) Vitreous Hemorrhage and Preretinal have been included in the Advanced Diabetic eye disease

Section 2: Maculopathy

Patients belonging to any of the above Grades can have Macular lesions. They have to be identified by

Any patients with Macular edema: mild Moderate Severe Cystic

Any patients with Exudates in Macular area $\geq$ 2

Any patients with any Exudates in Macula but the type of Exudates

Circinate Plaque Scattered + Circinate Scattered + Plaque Scattered+ Circinate + Plaque

Note:

Patients with NO Macular Edema and NO Macular Exudates will be considered as those without Maculopathy for the purpose of the study

Appendix 1

Modified UKPDS retinopathy allocation scale Level Severity Definition

10 DR absent Micro-aneurysms and other characteristics absent (No diabetic retinopathy ie No M & H, HE, CWS or IRMA)

14* DR questionable HE, CWS or IRMA definite; MA absent (CWS, HE or IRMA but no M & H)

20 MA only MA definite, other characteristics absent

30 Any exudates and CWS only (no MA or Hemorrhage)

35** Mild NPDR One or more of the following:

a. Venous loops = 2 in 1 or more field (Also includes venous occlusions and any number of loops as no mention about number of loops)

b. CWS or VB or IRMA = 1 in 1 or more field (Venous beading has been given a higher grading hence not to be included in this) c. Ignored *

d. HE = 2 in 1 or more field

(Also includes any HE ie > 1 but <3 + M & H) e . HE $\leq$ 3 in 1 or more field

f. CWS § 2 in 1 or more field

43 Moderate NPDR One (only) of the following:

a. HMA = 3 in 3-4 fields or HMA = 4 in 1 field

b. IRMA = 2 in 1-2 fields (See IRMA gradings)

47 Moderately severe One (only) of the following: NPDR

a. Both Level 43 definitions

b. IRMA = 2 in 3-4 fields (See IRMA gradings)

c. HMA 4 in 2-3 fields

d. VB = 2 in 1 field (This VB is to be ignored as any VB has been allocated a higher scale (53d) moreover it is not possible to extract information on the number of beading from the notes)

53 Severe NPDR One or more of the following:

a. At least 2 of the last 3 Level 47 definitions

b. HMA $\geq$ 4 in 4 fields

c. IRMA $\geq$ 3 in 1 or more field

d. VB = 2 in 2 or more fields

(Includes any venous beading and combinations)

60 Any NVE 1

61 Mild PDR

a. FPD or FPE present (with NVD and NVE absent) (Includes any RP)

b. NVE = 2 in 1 or more field

c. NVD 1 only

d. Any RP/FPD and any NV

65 Moderate PDR Either of the following:

a. NVE $\geq$ 3 in 1 or more field or NVD= 2 (With VH or PRH = 1 or absent) (VH or PRH = 1 and 2 has been tallied to Preretinal haemorrhage in the Vitreous haemorrhage combo box)

b. VH or PRH = 2 in 1 field (With NVE $<$ 3 in 1 field AND with NVD absent) (VH or PRH $>$ or equal to 3 has been tallied as Haze as it was considered as to cause haze the VH would be larger and occupy more than one field)

71 High Risk PDR (1) Any of the following:

a. VH or PRH $\geq$ 3 in 1 or more field (Haze)

b. NVE $\geq$ 3 in 1 or more field AND VH or PRH $\geq$ 2 in 1 or more field +(Preretinal)

c. NVD = 2 AND VH or PRH $\geq$ 2 in 1 or more field +(Preretinal)

d. NVD $\geq$3

75 High Risk PDR (2) NVD $\geq$3 AND VH or PRH $\geq$2 in 1 or more field (haze)

81 Advanced PDR Retina obscured due to VH or PRH (unable to visualise in the VH combo box)

99 Cannot grade

(A new box has been created- if retinopathy can be graded it has to be entered as can grade so the programme automatically computes the grade. If it can not be graded it has to be entered as to why (Cataract, RD, Corneal opacities, No eye data, No photograph etc - any information entered in this box other than "can grade" will automatically grade as 99, meaning: cannot grade)

IRMA grading for allocation scale:

1. IRMA $\geq$ 3 in 1 or more fields 4

2. IRMA = 2 in 3-4 fields 3

3. IRMA = 2 in 1-2 fields 2

4. IRMA = 1 in 1 or more fields 1

Appendix 2:

DR diabetic retinopathy

HE hard exudates

CWS cotton wool spots

IRMA intra-retinal microvascular abnormalities

MA micro-aneurysms

NPDR non-proliferative DR

VB venous beading

HMA haemorrhages/microaneurysms

PDR proliferative DR

NVE new vessels elsewhere (>1 DD from disc)

NVD new vessels disc (within 1 DD of disc margin)

FPE/RP fibrous proliferations elsewhere (>1 DD from disc)

FPD/RP fibrous proliferations disc (within 1 DD of disc margin)

VH vitreous haemorrhage

PRH preretinal haemorrhage

DD disc diameter.

RP Retinitis preproliferans

** NPDR Levels 35 and above require presence of micro-aneurysms.

Any HE has to be more than 35 or 20 (if only one exudates)

Any NV has to be more than or equal to 60

Any VH has to be more than 65

Appendix 3

PC Yes or No PC type PRP Pan retinal photocoagulation FP Focal or fill in Photocoagulation

No of sittings Number of sessions of PC before the next visit PC for see grading below

Photocoagulation this year

For

New vessel 1

Macular Exudates 2

Macular Edema 3

Both (exudates and New vessels) 4

Don't know 5

Rubeosis 6

Vitreous Haemorrhage 7

Retinal Detachment 8

Bibliography

ALTMAN, D. G. (1999). *Practicals statistics for medical research.* Chapman and Hall/CRC.

BERGQVIST, D. (1999). Management of small abdominal aortic aneurysms. *British Journal of Surgery* 86: 433–434.

BLANCHARD, J. F., H. K. ARMENIAN, AND P. P. FRIESEN (2000). Risk factors for abdominal aortic aneurysm: Results of a case-control study. *American Journal of Epidemiology* 151(6): 575–583.

BLAND, M. (1996). *An introduction to medical statistics.* Oxford University Press.

BREIMAN, L., J. H. FRIEDMAN, R. A. OLSHEN, AND C. J. STONE (1984). *Classification and regression trees.* Wadsworth International Group.

CHAN, K., T. LEE, P. S. AMD M. GOLDBAUM, R. WEINREB, AND T. SEJNOWSKI (2002). Comparison of machine learning and traditional classifiers in glaucoma diagnosis. *IEEE Transactions on Biomedical Engineering* 49(9): 963–974.

COLE, C. W., G. B. HILL, W. J. MILLAR, A. LAUPACIS, AND K. W. JOHNSTON (Spring 1996). Selective screening for abdominal aortic aneurysm. *Chronic Diseases in Canada* 17(2): 51–55.

COLLINS, J., L. ARAVJO, J. WALTON, AND D. LINDSELL (1988). Oxford screening programme for abdominal aortic aneurysm in men 65-75 years. *Lancet* 2: 613.

CORTES, C. AND V. VAPNIK (1995). Support-vector networks. *Machine Learning* 20: 273–297.

COWELL, R. G., A. P. DAWID, S. L. LAURITZEN, AND D. J. SPIEGELHALTER (1999). *Probabilistic Networks and Expert Systems.* Springer-Verlag.

COX, D. R., R. FITZPATRICK, A. E. FLETCHER, S. M. GORE, D. J. SPIEGELHALTER, AND D. R. JONES (1992). Quality-of-life assessment: Can we keep it simple? *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 155(3): 353–393.

COX, D. R. AND N. WERMUTH (1996). *Multivariate dependencies.* Chapman and Hall.

DE FINE OLIVARIUS, N., N. V. NIELSEN, AND A. H. ANDREASEN (2001). Diabetic retinopathy in newly diagnosed middle-aged and elderly diabetic patients. prevalence and interrelationship with microalbuminuria and triglycerides. *Graefe's Arch Clin Exp Opthalmol* 239: 664–672.

DE HENAUW, S., P. DE SMET, W. AELVOET, M. KORNITZER, AND G. D. BACKER (1998). Misclassification of coronary heart disease in mortality statistics. evidence from the who-monica ghent-charleroi study in belgium. *Journal of Epidemiology and Community Health* 52: 513–519.

DELONG, E. R., D. M. DELONG, AND D. L. CLARKE-PEARSON (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44: 837–845.

DIDELEZ, V., I. PIGEOT, K. DEAN, AND A. WISTER (2002). A comparative analysis of graphical interaction and logistic regression modelling: self-care and coping with a chronic illness in later life. *Biometrical Journal* 44: 410–432.

EDWARDS, D. (2000). *Introduction to Graphical Modelling.* Springer.

EVERITT, B. AND S. RABE-HESKETH (2001). *Analyzing medical data using S-Plus.* Springer.

FLEISS, J. L. (1981). *Statistical methods for rates and proportions.* Wiley.

FONG, D. S., L. AIELLO, T. W. GARDNER, G. L. KING, G. BLANKENSHIP, J. D. CAVALLERANO, F. L. F. III, AND R. KLEIN (2003). Diabetic Retinopathy. *Diabetes Care* 26(1): 226–229.

FRIEDMAN, N., D. GEIGER, AND M. GOLDSZMIDT (1997). Bayesian network classifiers. *Machine Learning* 29: 131–163.

GOLDSTEIN, H. (2003). *Multilevel Statistical Models.* Arnold.

GRIMSHAW, G. (1993). *The natural history of the abnormal aorta diagnosed by mass screening.* Ph. D. thesis, University of Birmingham.

GRIMSHAW, G. M. AND M. F. DOCKER (1992). Accurate screening for abdominal aortic aneurysm. *Clin. Phys. Physiol. Meas.* 13(2): 135–138.

GRIMSHAW, G. M. AND J. M. THOMPSON (1997). Changes in diameter of the abdominal aorta with age: an epidemiological study. *Journal of Clinical Ultrasound* 25: 7–13.

GRIMSHAW, G. M., J. M. THOMPSON, AND J. D. HAMMER (1994). A statistical analysis of the growth of small abdominal aortic aneurysms. *European Journal of Vascular Surgery* 8: 741–

746.

HAKAMA, M., E. PUKKALA, AND P. SAASTAMOINEN (1979). Selective screening: theory and practice based on high-risk groups of cervical cancer. *Journal of Epidemiology and Community Health* 33:257–261.

HAND, D. J. (1997). *Construction and assessment of Classification Rules.* John Wiley and Sons.

HAND, D. J. AND R. J. TILL (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning* 45:171–186.

HARRIS, P. L. (1992). Reducing the mortality from abdominal aortic aneurysms: need for a national screening programme. *British Medical Journal* 305:697–9.

HEALTH DEPARTMENTS OF THE UNITED KINGDOM (1998a). First report of the NSC.

HEALTH DEPARTMENTS OF THE UNITED KINGDOM (2000b). Second report of the NSC.

HILL, A. B. AND I. D. HILL (1991). *Bradford Hill's principles of medical statistics.* Edward Arnold.

HOAGLIN, D., F. MOSTELLER, AND J. W. TUKEY (1983). *Understanding Robust and Exploratory Data Analysis.* Wiley.

HOJSGAARD, S. (1996). Learning Structures from Data and Experts. *Mathematics and Computers in Simulation* 42.

HOSMER, D. AND S. LEMESHOW (2000). *Applied Logistic Regression.* John Wiley and Sons Inc.

JANISSE, J. (1997). ROC functions.
http://lib.stat.cmu.edu/s-news/Burst/7616.

KLEIN, J. P., N. KEIDING, AND S. KREINER (1995). Graphical models for panel studies, illustrated on data from the Framingham heart study. *Statistics in Medicine* 14:1265–1290.

KMIETOWICZ, Z. (2000). Warn the public of screening limitations, staff told. *British Medical Journal* 321:914.

KROLEWSKI, A. S., J. H. WARRAM, L. I. RAND, A. R. CHRISTLIEB, E. J. BUSICK, AND C. R. KAHN (1986). Risk of proliferative diabetic retinopathy in juvenile-onset type 1 diabetes: a 40-yr follow-up study. *Diabetes Care* 9(5):443–452.

LAURITZEN, S. L. AND T. S. RICHARDSON (2002). Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society. Series B* 64(3):321–361.

LAURITZEN, S. L. AND D. J. SPIEGELHALTER (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical*

Society Series B 50(2): 157–224.

LENFANT, C., L. F. L, AND T. THOM (1998). Fifty years of death certificates the framingham heart study. *Annals of Internal Medicine* 129: 1066–1067.

MADIGAN, D. AND A. E. RAFTERY (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* 89(428): 1535–1546.

MICHIE, D., D. J. SPIEGELHALTER, AND C. C. TAYLOR (1994). *Machine learning, neural and statistical classification.* Ellis Horwood.

MOSTELLER, F. AND J. W. TUKEY (1977). *Data analysis and regression.* Addison-Wesley Publishing Company.

PYO, R., J. K. LEE, J. M. SHIPLEY, J. A. CURCI, D. MAO, S. J. ZIPORIN, T. L. ENNIS, S. D. SHAPIRO, R. M. SENIOR, AND R. W. THOMPSON (2000). Targeted gene disruption of matrix metalloproteinase-9 (gelatinase b) suppresses development of experimental abdominal aortic aneurysms. *The Journal of Clinical Investigation* 105(11): 1641–1649.

RAFTERY, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology* 25: 111–163.

RAND, L. I., A. S. KROLEWSKI, L. M. AIELLO, J. H. WARRAM, R. S. BAKER, AND T. MAKI (1985). Multiple factors in the prediction of risk of proliferative diabetic retinopathy. *New England Journal of Medicine* 313(23): 1433–8.

RIPLEY, B. D. (1996). *Pattern Recognition and Neural Networks.* Cambridge University Press.

RIPLEY, B. D. (2002). Statistical Methods Need Software: A View of Statistical Computing. In *RSS 2002 Conference.*

RODIN, M. B., M. L. DAVIGLUS, G. C. WONG, K. LIU, D. B. GARSIDE, P. GREENLAND, AND J. STAMLER (2003). Middle age cardiovascular risk factors and abdominal aortic aneurysm in older age. *Hypertension* 42(1): 61–68.

RUDNICKA, A. R. AND J. BIRCH (2000). *Diabetic Eye Disease Identification and comanagement.* Butterworth Heinemann.

SCHAFER, J. L. (1997). *Analysis of incomplete multivariate data.* Chapman and Hall/CRC.

SCHAFER, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research* 8: 3–15.

Scott, R. A. P., H. A. Ashton, M. J. Lamparelli, G. Harris, and J. W. Stevens (1999). A 14-year experience with 6 cm as a criterion for surgical treatment of abdominal aortic aneurysm. *British Journal of Surgery* 86: 1317–1321.

Semmens, J. B., P. E. Norman, M. M. D. Lawrence-Brown, and C. D. A. J. Holman (2000). Influence of gender on outcome from ruptured abdominal aortic aneurysm. *British Journal of Surgery* 87: 191–194.

Singh, K., K. H. Bonaa, B. K. Jacobsen, L. Bjork, and S. Solberg (2001). Prevalence of and risk factors for abdominal aortic aneurysm in a population-based study. *American Journal of Epidemiology* 154(3): 236–244.

Smith, F. C. T., G. M. Grimshaw, I. S. Paterson, C. P. Shearman, and J. D. Hamer (1993). Ultrasonographic screening for abdominal aortic aneurysm in an urban community. *British Journal of Surgery* 89: 1406–1409.

Snijders, T. A. B. and R. J. Bosker (1999). *Multilevel Analysis: An introduction to basic and advanced multilevel modeling.* Sage Publications.

Sonesson, B., T. Lanne, F. Hansen, and T. Sandgren (1994). Infrarenal aortic diameter in the healthy person. *European Journal of Vascular Surgery* 8(1): 89–95.

Sonis, J. (1998). A closer look at confounding. *Family Medicine* 30(8): 584–588.

Spencer, C. A., K. Jamrozik, P. E. Norman, and M. M. D. Lawrence-Brown (2000). The potential for a selective screening strategy for abdominal aortic aneurysm. *Journal of Medical Screening* 7: 209–211.

Steyerberg, E. W., M. Eijkemans, F. Harrell, and J. Habbema (2000). Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small datasets. *Statistics in Medicine* 19: 1059–1079.

T, T. W., C. Quick, and N. Day (1999). The association between cigarette smoking and abdominal aortic aneurysms. *Journal of Vascular Surgery* 30: 1099–1105.

Twisk, J. and W. de Vente (2002). Attrition in longitudinal studies: How to deal with missing data. *Journal of Clinical Epidemiology* 55: 329–337.

Vardulaki, K. A., T. C. Prevost, N. M. Walker, N. E. Day, A. B. M. Wilmink, C. R. G. Quick, H. A. Ashton, and R. Scott (1998). Growth rates and risk of rupture of abdominal aortic aneurysms. *British Journal of Surgery* 85: 1674–1680.

Vardulaki, K. A., N. M. Walker, N. E. Day, S. W. Daffy, H. A. Ashton, and

R. SCOTT (2000). Quantifying the risks of hypertension, age, sex and smoking in patients with abdominal aortic aneurysm. *British Journal of Surgery* 87: 195–200.

VENABLES, W. N. AND B. D. RIPLEY (1997). *Modern applied statistics with S-Plus*. Springer.

WANG, D., Z. WENYANG, AND A. BAKHAI (2004). Comparison of Bayesian model averaging and stepwise methods for model selection in logistic regression. *Statistics in Medicine* 23: 3451–3467.

WILLIAMS, G. AND J. C. PICKUP (1999). *Handbook of Diabetes*. Blackwell Science.

WILMINK, A. B. M. AND C. R. G. QUICK (1998). Epidemiology and potential for prevention of abdominal aortic aneurysm. *British Journal of Surgery* 85: 155–162.