UNIVERSITY OF SYDNEY

Challenges associated with clinical studies and the integration of gene expression data

by

Anna Elizabeth Campain

A thesis submitted in fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Science
Department of Mathematics and Statistics

October 2012

'What's one and one and one and one and one and one and one and one and one and one and one and one?'
'I don't know', said Alice, 'I lost count.'
'She can't do addition!' said the Red Queen.

Lewis Carroll, *Through the Looking Glass* (1872)

Dedication

This thesis is dedicated to my family;

My mother for raising me to believe anything is possible, and all things will work out in the end. My father for teaching me that the *reason* that all things are possible is through planning and preparation. My sisters, who showed me that being the youngest, dimmest, clumsiest and ugliest is not vexing when you are dearly loved.

I dedicate this thesis also to my beloved husband, Stephen. This journey was not a burden, as I was loved and cared for throughout by you, with your belief in me, your gentleness, patience and prayers. Thank you for encouraging me to continue with my studies and for all our plans that you put on hold while I was working.

And finally this thesis is dedicated to our little darling, who through God's own timing, we still eagerly await. Your beautiful kicking, proding, hiccups and fidgets have kept me company these last eight and a half months.

Acknowledgements

I would like to thank and acknowledge my wonderful, inspiring and generous PhD supervisor, Dr Jean Yang. I am completely indebted to her for everything that she has done for me during this post-graduate marathon. I have 'highs' to celebrate because of her, and have pushed through the 'lows' believing in her wisdom. At the beginning of my PhD I was excited to work with such a fantastic statistician and mirror at least some of her skill. Now I see her as a brilliantly gifted friend.

I have also had a generous and proficient associate supervisor, Dr Samuel Müller. He has graciously shared his wisdom with me, especially regarding the clinical data side of my work. His wonderful, timely and accurate comments have helped shaped this work into something special, which I could not have done on my own. He has spent copious amounts of time shaping me into a better and more careful statistician.

This journey has been made more special with the work and dedication of Professor Terry Speed. Working with such a distinguished statistician was a great privilege. It has been a wonderful and fun experience collaborating closely with him. He spent so much time and energy on me, encouraging me to become a better statistician and bioinformatician.

I have had the great privilege of completing my PhD alongside many talented students in a range of scientific disciplines. To this end I would like to acknowledge and thank Ms Francine Marques and Ms Sarah-Jane Schramm. Francine has humbled me with her dedication to research, as she has delved and explored the genetic recesses of hypertension. As her PhD comes to a close too, I wish her enthusiasm and passion in her post-doc research. I would also like to thank Francine's supervisor, Professor Brian Morris, whose passion for biology infuses his laboratory. The beautiful Sarah-Jane has been a pleasure to work with, she motivates and encourages me with her scientific integrity and thoroughness as well as her deep understanding of what is truly important in life.

Professor Graham Mann and Clinical Professor Richard Scolyer have been incredible. To watch both straddling the clinical world and research arena with such passion has been amazing. It has been a magnificent opportunity to collaborate with these two doctors, while working with the melanoma data.

I would like to thank Associate Professor George Condous and Dr Jennifer Oats for their collaboration on the Early Pregnancy data, this large clinical data set has allowed me to explore many statistical challenges. Their collaboration has allowed me to grow extensively throughout my candidature.

These past few years has been a terrific joy, filled with fun and laughter along with hard work. My fellow statisticians, undergraduates, post-graduates and lectures have helped to make this time very special. In particular I would like to acknowledge Ellis Patrick, we have become great friends as we have competed for Jean's attention and affection. He is a gifted statistician who will be a great asset to the research community and I look forward to working with him in the future. I would also like to thank Ellis and John Ormerod for their time and kind words while reading, edit and helping me in the final stages of this mammoth project.

Abstract

Doctor of Philosophy

by Anna Elizabeth Campain

The progress of high-throughput biotechnologies has generated a myriad of large and complex data sets in many areas of medical research. The development of such data has had a terrific impact in statistics, producing new challenges and encouraging methodological development. In recent years, medical research has been able to generate both clinical and genomic level data for the same patients. The integration of such data may enable scientists to glean a greater understanding of complex diseases. Before such data types can be combined effectively, there are still many statistical problems associated with the analysis of these two data types separately. This thesis addresses some of these challenges and contributes to the advancement of their solutions.

The very common challenge of missing observations within clinical data is addressed using multiple imputation. However, multiple imputation algorithms are not routinely compared. To this end, a novel framework for the comparison of multiple imputation methods is developed within this thesis. Three popular multiple imputation methods are compared through a simulation study, highlighting strengths and weaknesses within them all. Model stability is a statistical problem that is especially prevalent when regression classes are unbalanced or observed clinical variables are rare. An original solution to model instability in a regression context is provided with stability gained through stratified bootstrap sampling.

Prior to the integration of clinical and gene expression data, further development of statistical methods for the integration of multiple expression studies is required. This thesis examines approaches used to combine expression data sets. Current expression integration approaches are compared with a newly developed method, 'meta Differential Expression via Distance Synthesis'. This thesis also highlights the two main ways data can be combined. The first is the integration of statistics obtained from individual expression data analyses. The second is the integration of the unanalysed expression data, allowing for a united data set in downstream analysis. Both paradigms are explored and advocated in different contexts.

Finally, a melanoma case study is used to highlight the importance of careful analysis of the individual data types, clinical and expression, prior to data integration. The solutions to several of the problems addressed are implemented within this study in combination with some rudimentary integration methods for combining the two types of data. This case study highlights the potential benefits of careful individual analyses of clinical and genomic data as well as the integration of this information when making survival time predictions.

Publications

Publications

- **A.E. Campain** and Y.H. Yang (2010) Comparison study of microarray meta-analysis methods, **11**:408 *BMC Bioinformatics.*
- **A.E. Campain**, F.Z. Marques, Y.H. Yang and B.J. Morris (2010) Meta-analysis of genome-wide gene expression differences in onset and maintenance phase of genetic hypertension, **56**:319–324 *Hypertension.*
- F.Z. Marques, **A.E Campain** P.J Davern, Y.H Yang, G.A Head and B.J. Morris (2011) Genes influencing circadian differences in blood pressure in hypertensive mice, **6**:4 e19203 *PLoS One.*
- F.Z. Marques, **A.E. Campain**, P.J Davern, Y.H Yang, G.A Head and B.J. Morris (2011) Global identification of the genes and pathways differentially expressed in hypothalamus in early and established neurogenic hypertension, **43**:766–771 *Physiological Genomics.*
- F.Z. Marques, **A.E. Campain**, E. Zukowska-Szczechowska, M. Tomaszewski, Y.H Yang, F.J. Charchar and B.J Morris (2011) Gene expression profiling reveals renin mRNA overexpression in human hypertensive kidneys and a role for microRNAs, **58**:1093–8, *Hypertension.*
- S.-J. Schramm, **A.E. Campain**, R. Scolyer, Y.H. Yang and G.J. Mann (2011) Review and cross-validation of gene expression signatures and melanoma prognosis. **132**:274–283 *Journal of Investigative Dermatology.*

Manuscripts under review and in preparation

- **A.E. Campain**, S. Müller, G. Condous and Y.H. Yang (2011) Stable logistic regression models in the presence of missing values and class imbalances. Under review, *Biostatistics.*
- G.J. Mann, G.M. Pupo, **A.E. Campain**, C.A. Carter, S.-J. Schramm, A. Pianova, S. Gerega, C. De Silva, K. Lai, J. Wilmott, M. Synott, P. Hersey, R.F. Kefford, J.F. Thompson, Y.H. Yang and R.A. Scolyer (2011) BRAF mutation, NRAS mutation and absence of an immune-related expressed gene profile predict poor outcome in stage III melanoma. Under review, *Journal of Clinical Oncology.*
- Y.H. Yang, **A.E. Campain** and T.P. Speed (2011) Finding differentially expressed genes in microarray data. Preprint.
- J. Riemke, **A.E. Campain**, T. Bignardi, I. Casikar, D. Alhamdan, D. Fauchon, R. Benzie, S. Müller, Y.H. Yang, M. Mongelli and G. Condous (2011) Development of a new model to predict viability at the end of the 1st trimester after a single visit to an Early Pregnancy Unit. Preprint.

Contents

List of Figures

List of Tables

Abbreviations

AE	ArrayExpress - `http://www.ebi.ac.uk/arrayexpress/`
AIC	Akaike Information Criterion
AUC	Area Under (the ROC) Curves
B-MI	Bootstrap samples and Multiple Imputation
B-MSE	Bootstrapped-Mean Squared Error
BP	Biological Process
BIC	Bayesian Information Criterion
Bss/Wss	Between Sum of Squares over Within Sum of Squares
CC	Cellular Component
cDNA	complementary DNA
CRL	Crown (to) Rump Length
DE	Differentially Expressed
DLDA	Diagonal Linear Discriminate Analysis
DNA	Deoxyribonucleic Acid
FC	Fold Change
FDR	False Discovery Rate
FHR	Foetal Heart Rate
GEO	Gene Expression Omnibus - `http://www.ncbi.nlm.nih.gov/geo/`
GLM	Generalised Linear Model
GO	Gene Ontology
GS	Gestational Sack
GST	Gene Set Tests
HT	HyperTensive
IC	Integrative Correlation
IUP	Inner-Unitary Pregnancy

KEGG	**K**yoto **E**ncyclopaedia of **G**enes and **G**enomes
LH	**L**yon **H**ypertensive
LMP	**L**ast **M**onthly **P**eriod
LOOCV	**L**eave **O**ne **O**ut **C**ross **V**alidation
MA-plot	A scatter plot of expression data displaying the $\log_2$(FC) on the y-axis and the average intensity on the x-axis
MF	**M**olecular **F**unction
MIA	**M**elanoma **I**nstitute **A**ustralia
mDEDS	**m**eta **D**ifferential **E**xpression via **D**istance **S**ynthesis
mRNA	**m**essenger **R**ibonucleic **A**cid
NM	**N**odal **M**etastasises
NT	**N**ormo**T**ensive
NUSE	**N**ormalised **U**nscaled **S**tandard **E**rror
PCR	**P**olymerase **C**hain **R**eaction
PLM	**P**robe **L**evel **M**odel
POE	**P**robability **O**f **E**xpression
QC	**Q**uality **C**ontrol
RLE	**R**elative **L**og **E**xpression
RMA	**R**obust-**M**ulti**A**rray analysis
ROC	**R**eceiver **O**perating **C**haracteristic
SHR	**S**pontaneously **H**ypertensive **R**at
SVM	**S**upport **V**ector **M**achine
WKY	**W**istart **KY**oto (normotenive rat)
VAS	**V**isual **A**nalogue **S**cale - A pain rating scale between 0 and 10.
YS	**Y**oke **S**ack

Abbreviations for data sets

EPU data	Riemke et al. (2011) clinical early pregnancy data
Cerutti data	Cerutti et al. (2006) hypertensive v normotensive rat data set
Clemitson data	Clemitson et al. (2007) hypertensive v normotensive rat data set
Grayson data	Grayson et al. (2007) hypertensive v normotensive rat data set
Rysä data	Rysä et al. (2005) hypertensive v normotensive rat data set
Bogunovic data	Bogunovic et al. (2009) melanoma data set
John data	Thomas John et al. (2008) melanoma data set
Jönsson data	Göran Jönsson et al. (2010) melanoma data set
Mann data	Mann et al. (2011) melanoma data set
Winnepenninchx data	Winnepenninckx et al. (2006) melanoma data set

Symbols

A_{PV}	The pre-validated vector
B	The total number of bootstraps
f	Number of unknown factor in the RUV-2 mega-analysis method
g	An index for the statistic measures in mDEDS
G	The total number of statistic measures used in mDEDS
h	An index for a batch within an expression data set
H	Total number of batches within an expression data set
i	An index representing an individual gene
I	The total number of genes
j	An index representing an individual sample
k	An index representing an individual data set
K	Total number of data sets within an analysis
m	The total number of imputations
n	The total number of samples
q	An index for a clinical variable or parameter
Q	Total number of clinical variables or parameters
r	An imputation
X	Data matrix, either of clinical or expression data when context is clear
x	Datum within a data matrix, often written with subscripts representing the row and column
$X^{(\mathrm{C})}$	Clinical data matrix $(n \times Q)$, when the context involves both clinical and expression data
$x^{(\mathrm{C})}$	Datum within the clinical data matrix
$X^{(\mathrm{E})}$	Expression data matrix $(I \times n)$, when the context involves both clinical and expression data

$x^{(\mathrm{E})}$	Datum within the expression data matrix
y	The response vector
α	Imputation coefficient estimate
$\boldsymbol{\beta}$	Vector of regression coefficients, with $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_Q)$
μ	Mean
π_j	Probability of the j^{th} event occuring
$\boldsymbol{\pi}$	The vector of the n event probabilities
σ^2	Variance
τ_{B}	Bootstrap inclusion threshold
τ_{MI}	Multiple imputation inclusion threshold

Chapter 1

Introduction

Statistical research in a medical context is an important component to both medical research and statistical method development. In the afterglow of the 'Human Genome Project' large amounts of research money and development time is being spent on different types of studies and diseases (both complex and simple). This thesis has been motivated by data sets containing typical statistical issues pertaining to work in bioinformatics and medical statistics. Within such research, numerous teams attempt to understand the genetic relationships, as well as potential treatments, for many devastating diseases. To date, this important work has had some success. Alizadeh et al. (2000) is a famous example of a study where microarrays were used in a clustering context to interrogate Diffuse Large B-Cell Lymphoma (DLBCL) samples. Their study's aim was to investigate why patients with the same disease appear to have a heterogeneous survival pattern. Alizadeh et al. (2000) proposed that there were several sub-classes of DLBCL, each with slightly different generic signatures and survival patterns. In another example, US Food and Drug Administration in 2007 authorised the very first microarray based diagnostic tool, the MammaPrint, which is used as a prognostic test for particular types of breast cancer (Cardoso et al., 2007; Mook et al., 2007), based on the genetic signature found in van't Veer et al. (2002).

The genomic era, with many massive data sets generated from high-throughput technology, has not diminished the importance of clinical data. Biologists, doctors and other researchers are becoming increasingly interested in combining clinical and genomic data. Some integrative methods to this effect have recently been developed (Boulesteix et al., 2008; Dettling and Bühlmann, 2004; Gevaert et al., 2006; Lê Cao et al., 2010; Tibshirani and Efron, 2002). However these methods are still in their infancy. The development of these methods are hindered due to many statistical challenges still to be addressed separately in the analysis of clinical and gene expression data. This thesis looks at

some of these statistical dilemmas including missing observations in clinical data, model instability as well as the integration of microarray data with other microarray studies. Only after such important problems have been adequately addressed can we begin to examine the questions involving the integration of clinical and microarray data.

Each of the sections in this introduction outlines parts of the thesis. The chapter continues by discussing clinical data and touching on some of the issues associated with this type of analysis (Section 1.1). An introduction to gene expression data and microarray analysis is presented in Section 1.2, discussing some of the developments that have taken place in recent years which have lead to a desire for cross study data integration. Finally, Section 1.3 highlights some approaches to the integration of clinical and gene expression data as a precursor to the melanoma case study in Chapter 5.

1.1 Introduction to clinical data

Clinical data is a traditional data type encompassing a wide range of observed data variables in a clinical or medical context. Such data can range from discrete and categorical to continuous values. These variables can cover a large range of observations from gender, blood type and heart rate to pathology results and genetic mutation information. This thesis makes use of a clinical data set obtained from the Early Pregnancy Unit at the Nepean Hospital Sydney Australia. This study relates to women who presented at the clinic with a complication related to their currently viable pregnancy in the first trimester (for more detail see Section 2.1). It is through this motivating example that the issues associated with clinical data are examined in Chapter 3.

1.1.1 Missing data

Missing data is a reality for a large proportion of studies and is apparent in diverse analysis situations, including data generated through clinical studies, questionnaires and censuses. Most statistical analysis methods assume a rectangular data set and much of the history and foundations of statistical development require complete data, that is where none of the observations are missing. There is a vast array of literature discussing missing data, for example Little and Rubin (1987), Rubin (1987) and Schafer (1999). These core texts elaborate on how the missing data can be observed within the data set, that is if the data is missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR). Different statistical problems that arise depend, amongst others, on how the missing data is distributed. A highly developed method of overcoming the problems of working with incomplete data sets is the use of imputation

and multiple imputation. Imputation allows the construction of a complete data set and hence downstream data analysis can proceed via well known complete data methods. This thesis focuses on multiple imputation. However, there are many other ways to handle missing data. Horton and Laird (2000), Ibrahim et al. (2005) and Little (1992) highlight such approaches including case deletion, maximum likelihood, fully Bayesian and weighted estimating equations. In the context of drug development, last observation carried forward and best observation carried forward are commonly employed methods (Barnes et al., 2008).

Despite missing data and data imputation being well studied statistical constructs, less well considered is how the *amount* of missing data affects later statistical analysis and to what extent does the amount of missing data need to be considered when dealing with incomplete data sets. Imputation methodology literature is littered with examples of incomplete data sets: in a study observing blood pressure measurements (van Buuren et al., 1999) a missingness of 12.5% is observed, Horton and Kleinman (2007) observe a similar amount ranging from 4–16% on several data sets. Larger proportions of missing data are reported, for example 43% in Horton and Lipsitz (2001) and nearly 100% in Stuart et al. (2009). There are conflicting views regarding how much missingness in a data set is detrimental. Acuna and Rodriguez (2004) suggest that '[...] 1–5% [missingness is] manageable. However 5–15% require sophisticated methods to handle, and more than 15% may severely impact any kind of interpretation'. Nishisato and Ahn (1995), when considering correspondence analysis expressed grave concerns for analysis interpretations when missingness was greater than their suggested 11%. Yet in contrast Rubin (1996) states '[T]he fraction of missing information is modest, e.g. $< 30\%$' suggesting that missingness up to 30% is still manageable for analysis methods involving imputation.

Most imputation algorithms will produce results regardless of the extent of the proportion of missing data and currently there exists a large variety of multiple imputation approaches, ranging from *ad hoc* to highly sophisticated statistical modelling. Chapter 3 develops a framework for the evaluation of different multiple imputation algorithms. This approach is used to evaluate and compare three popular multiple imputation methods, Amelia II (King et al., 2001), Mi (Rubin, 1987) and MICE (Buuren et al., 1999), and can be extended to consider other available methods and also multiple imputation methods developed in the future.

The three main questions addressed while considering missing data are:

1. How does the amount of missingness affect results obtained from the imputation procedure in a logistic regression or prediction context?

2. Can a critical point be identified where missingness is too great and the statistical validity of the downstream results and interpretation consequently needs to be questioned?

3. Through establishing an evaluation framework is it possible to consider the appropriateness of a particular method and to compare various available imputation algorithms?

The basis of the simulation study in Section 3.2 is to empirically investigate how regression coefficients depend on induced missingness and which imputation method handles this missingness most effectively. To this end, the Early Pregnancy Unit data is used, of interest is not a particular obtained final model but the effect of the missingness on the regression coefficients in weighted logistic regression models.

1.1.2 Model building and stability

To understand the relationship between a response variable and the explanatory variables (for example, between the condition of interest and the clinical variables observed), regression models can be developed. The generalised linear model (GLM) is a flexible family of statistical models which incorporates many aspects of statistical modelling including ordinary linear regression modelling and models for continuous, discrete or dichotomous responses. A GLM is made up of three elements: (i) the random component; (ii) the systematic component; and (iii) the link function.

There are two random components, the response variable $\boldsymbol{y}$ which in a GLM framework can be continuous or discrete and the error. The systematic component is a linear combination of explanatory variables. The explanatory variables can be the observed variables (for example the clinical values) or multiple combinations of (or powers of) the observed variables. These variables form the additive model $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_Q x_Q$ for Q variables, where the β's are parameters by which the explanatory variables are multiplied, and β_0 represents the intercept. The link function is the function that connects the expected value of the response variable (the random component) to the linear combination of explanatory variables (the systematic component). This is dependent on the type of data that makes up the random component. If the link function is $g(\cdot)$ and the expectation of $\boldsymbol{y}$ is μ then the link function is such that $g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_Q x_Q$. There are a multitude of different link functions depending on the data situation for example the identity link ($g(\mu) = \mu$) and the inverse Gaussian ($g(\mu) = \mu^2$) to mention two. For more information on GLM models introductory statistical model building texts include Agresti (2007); Lindsey (1997); McCullagh and Nelder (1998);

McNeil et al. (1996). In general, a GLM model can be written as,

$$g(\mu_j) = \boldsymbol{x}_j^T \boldsymbol{\beta} \tag{1.1}$$

where $j = 1, ..., n$, $\boldsymbol{x}_j$ is a vector of Q independent variables, and $\boldsymbol{\beta}$ is a vector of Q parameters.

For this thesis, responses are binary and a logistic regression model is used. The logistic regression model is a GLM with a *logit* link function, where $g(\mu) = \ln(\frac{\mu}{1-\mu})$, more detail is shown in Chapter 3.

The eventual aim of the GLM and other models is to describe the data. The variables in the model are reduced so that only the 'important' and/or 'informative' variables remain. Statistical models can be used for predicting and/or data explanation. The exact purpose for a model depends on the investigation and reasons undermining the data. Such wrestling stem from a philosophical root and the exact musing will differ depending on the statistician. Throughout this thesis an ideal model is parsimonious, and hence should both predict and explain as simply as possible the data and the relationships between the models. In some case (Section 3.5 and Chapter 5) the former has been sacrificed, to a small degree, for the latter. In some models, small perturbation in the original data set will have large effects on the final model selected (Breiman, 1996; Steyerberg et al., 2000). Such perturbations can be introduced in the form of re-sampling or sub-sampling. There are many reasons for this instability (or uncertainty), including the interrelationship and correlation between variables, and multicollinearity (Altman and Andersen, 1989). Large changes and erroneous conclusions can be made when models are unstable, and inaccuracies such as omitting variables by mistake can have large downstream ramifications (Chatfield, 1995).

Chapter 3 of this thesis outlines an approach to model instability which uses the bootstrap (Efron, 1979; Efron and Tibshirani, 1986) in an attempt to obtain stable models in the context of missing data and class imbalance.

1.2 Introduction to expression data

Gene expression is the process where the inherent information within Deoxyribonucleic Acid (DNA) are used in the synthesis of functional products developed by the gene, these outcomes are physical or biological. With the exception of red blood cells, every cell in the human body contains the same DNA. DNA is made up of molecules called nucleotides. These nucleotides are assembled head to tail to form a chain. There are four different bases that can be found as nucleotides, these are adenine (A), thymine

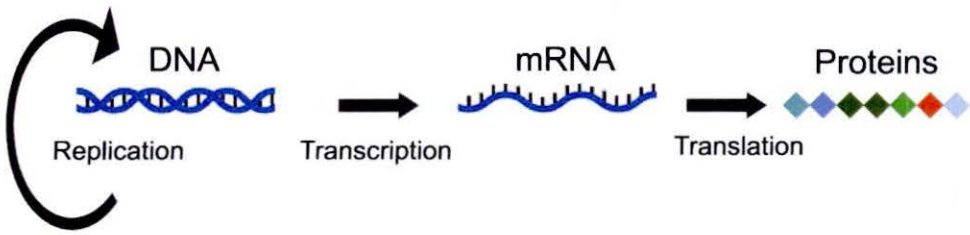


FIGURE 1.1: Central dogma of molecular biology, outlining the transcription and translation of DNA which leads to proteins. This simplified schema represents the flow of information within a biological system from DNA to mRNA and then to protein, with arrows representing the directions proposed for information transfer. DNA is self-replicating, and mRNA is constructed via transcription. This process of transcription is directed by the DNA template. In a very similar way the construction of the proteins via translation is directed by the mRNA template. In almost all cases, the direction of information flow is uni-directional, and the dogma, developed in 1958 by Francis Crick (Crick, 1970), has remained relatively unchanged.

(T), cytosine (C) and guanine (G). The nucleotides that make up individuals are 99.9% identical differing in arguably known locations along the human genome (Watson, 2008).

A DNA molecule consists of two nucleotide strands in a double helix shape. These chains of nucleotides are arranged in an 'anti-parallel' fashion and abide by Chargaff's rule (Chargaff, 1951). Chargaff's rule is that a hydrogen bond will be formed between complementary base pairs once these nucleotide strands are united. That is, adenine will bond with thymine (A to T) and cytosine will bond with guanine (C to G), creating a complementary double helix structure.

1.2.1 mRNA production and hybridisation

A direct result of Chargaff's rule is that if one strand of the DNA is known, the other strand can be inferred. The process of a DNA strand obtaining a correct complementary strand is known as *hybridisation*. When a gene is being expressed, information from the DNA is exposed and the mRNA (messenger ribonucleic acid) is created via a process known as transcription. Once this has taken place the mRNA may (or may not) travel through the cell to create amino acids or a protein via translation. These sequential stages are known as the 'Central Dogma of Molecular Biology' (Figure 1.1) and occur within a cell when a particular behaviour or expression is required. A good introduction can be found in Watson (2008), *Molecular Biology of the Gene*, which gives more detail on the biology behind hybridisation and gene expression.

In a highly simplified explanation, microarrays make use of the naturally occuring behaviour of hybridisation and attempt to observe which particular mRNAs are present in the sample. The detected mRNAs are used to infer the expressed genes. This is achieved in two parts. The first is by providing known and predetermined complementary strands

in the form of probes on the microarray slide. The second part is to measure which genes are expressing themselves, and to what degree in a sample. A fundamental assumption is that the measure of 'fluorescence' represents the level of intensity of hybridised probe and is *proportional* to the number of transcripts in the sample. In the human body, it is assumed that disruptions to the gene expression within cell groups can be linked to many diseases and such assumptions have led to a vast and growing area of research within bioinformatics.

1.2.2 Different microarray platforms

Since the initial spotted array (Schena et al., 1995) microarray technology has developed and different types of microarrays have been designed. Microarray companies including Agilent, Affymetrix and Illumina amongst others, along side some in-house (laboratory made) designs, have each developed their own microarrays with particular strengths and weaknesses. The three main approaches to microarray construction are spotted arrays, 'on-chip' DNA arrays as well as random bead arrays (Sreenivasulu et al., 2010).

In this thesis, a microarray platform will refer to *both* the type of array, for example Agilent, Affymetrix or Illumina as well as the version of array together. Obviously an Agilent and an Affymetrix array are considered different platforms, but also an Affymetrix GeneChip Rat Genome 230 array and an Affymetrix GeneChip Rat Genome U34 array are considered different platforms. Such a definition of a platform is applied because although it is slightly more straightforward to integrate two arrays from the same company (Affymetrix to Affymetrix for example) certain integrative steps are still required to integrate between versions.

Spotted arrays and 'on-chip' arrays are considered 'planar' arrays because the exact location of complementary nucleotide sequences are predetermined during the design phase for each slide type. Bead arrays are random in regards to the location and volume of the complementary nucleotide probe sequences, and are re-generated as each chip is constructed.

Spotted arrays Spotted arrays are created through a robotic mechanism that prints probes of complementary DNA (cDNA) or long oligonucleotides onto the microarray slide. Methods of attaching probes and the number of probes printed varies between slides. However, all such mechanisms are subject to between batch variability (as the solutions of RNA used to create spots are renewed) (Diehl et al., 2001), precision bias (accuracy of the robotic arm in placing the spots) and a printing/print-tip bias (for example, a damaged print-tip will case an error for all probes printed by that tip) and

many others. Recent developments attempt to remove some of these problems (Dufva, 2009; Okamoto et al., 2000).

On-chip 'On-chip' development is a technology that allows the probes to be built directly onto the surface of the slide. Affymetrix was one of the first companies to manufacture such slides. Their method makes use of photolithography to guide nucleotides onto the probe site one at a time to eventuate in a oligonucleotide probe (Auer et al., 2009). The method is an adaptation to the process of computer chip fabrication. To ensure the appropriate nucleotide is being added, a series of masks are developed (between 80–100), increasing the cost of the slides and a delay in the development of new or custom slides (Dufva, 2005). Nimblegene/FEBIT Technology/Roche diagnostics uses a similar light sensitive method to develop their slides but is able to increase flexibility by using mirrors, and hence is known as 'maskless' DNA synthesis (Nuwaysir et al., 2002; Singh-Gasson et al., 1999). Agilent also produces 'on-chip' developed arrays by using a series of phosphoramides, building oligonucleotides in a column in a similar fashion to PCR development (Hughes et al., 2001; Lausted et al., 2004).

Bead arrays Illumina's bead array makes use of bead type technology. Here the complementary oligonucleotide sequences are attached to microscopic beads (about 20–30 beads have the same probe sequence). All the different probe beads are combined and then spread over the slide which contains wells which are the size of one bead (Fan et al., 2006). This makes the location of each probe random on the chip. A code, integrated into the probe sequence is then scanned prior to hybridisation, so that the location of each oligonucleotide sequence is then known prior to hybridisation, this allows results to be processed (Dufva, 2009; Yeakley et al., 2002).

Despite the differences in platforms, there are two main overarching types of arrays, these are two-colour or one-colour arrays. Two-colour arrays include the spotted arrays, in-house arrays, some 'on-chip' arrays as well as the two colour Agilent array. Here two differently labeled samples (for example treatment and control) are hybridised together on a single array (Schena et al., 1995). Gene expression is measured relative to the other condition, as a ratio, eliminating the between array comparisons and biases. For the one-colour arrays, only one sample is hybridised to each array yielding gene intensities.

1.2.3 Individual microarray experiments

For individual microarray experiments, the techniques involved in the analysis have begun to stabilise. Such steps include image analysis, quality control and assessment,

determining differentially expressed genes and functional interpretation. Figure 1.2 is a flow-diagram highlighting these steps and some of the graphical tools used to aid in the process. There are many reviews regarding the methods of data analysis for a single data set. Such reviews include detailed outlines (Michiels et al., 2007) and papers detailing the process of single microarray analysis (Allison et al., 2006; Fan and Ren, 2006; Kreil and Russell, 2005; Roberts, 2008; Speed, 2003; Zhang et al., 2009). Owzar et al. (2008) considered some of the statistical challenges that arise and Sims (2009) addressed the interdisciplinary problems associated with bioinformatics as a whole.

Quality control (QC) is an important step in microarray analysis. The aim of such assessment is to determine if particular samples or arrays are of sufficient quality to be included in further analysis (Hartmann, 2005). QC measures are further detailed in Section 2.2.2. Preprocessing of microarray experiments typically involve three stages, background correction, probe-level normalisation and probe set summary (Smyth and Speed, 2003; Smyth et al., 2003). There are many different ways to perform preprocessing and techniques vary depending on the type of array used. After such analysis, the data from the individual samples within a study have been collated into a single between-sample normalised matrix.

Differentially expressed (DE) genes are genes that are biologically different between groups being considered in the analysis (for example genes different in their level of activities for particular conditions). Often it is the purpose of microarray experiments to estimate, or identify, statistically which genes are DE. Identification of DE genes can be considered as a two step process, (i) genes are ranked and, (ii) are tested for significance using a critical threshold. Typically either a number of highly ranked genes or the significant genes are used in downstream analysis. Ranking becomes increasingly important when due to lack of power there is a limited number of significant DE genes.

The ranking of genes can be achieved in numerous ways. The most common and simplest measure is fold change (FC) which represents the magnitude of mean expression differences between two classes. Another choice could be the t-statistic which takes variation of measurement into account. For this process, data is modelled using a linear model, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{y}$ represents the gene expression, $\boldsymbol{X}$ the design matrix, $\boldsymbol{\beta}$ is a vector of estimates of interest, for example the effect size and $\boldsymbol{\epsilon}$ is the error term.

To determine significance a critical threshold or value is determined. For the ranked statistics, values above the threshold are considered significant (or DE). There are many different thresholds. In the early days of microarray analysis, genes were considered significant if their FC was 2 or more. A different criterion is based on p-values, where genes are considered DE if they have a p-value less than 0.05. Multiple testing needs to be employed when using such a criterion (Dudoit et al., 2004) (for example Bonferroni

correction) because of the vast number of comparisons being made. DE genes can also be identified by controlling for 5% false discovery rate (FDR). Different rationales for the choice of cut-off are used with different statistics, but the most commonly used methods are FC alone, FDR alone or FC and FDR used together (Yang et al., 2011).

Gene set testing and functional interpretation are an important component to microarray analysis. Gene Set Tests (GST) consider the complete ranked gene lists and test to see if a set of pre-defined genes are clustered toward the top or bottom of the list by applying a Wilcoxson rank sum test. Such an approach asks the question *'is a set of pre-defined genes DE?'*. Functional interpretation facilitates biological interpretation regarding the obtained list of DE genes. A hypergeometric test is applied to the DE and non-DE genes, where genes are either included or excluded from an ontology (or gene function) in question. This approach addresses the question *'which ontologies are over (or under) represented within the list of DE genes?'*. Such tests are important to understand the biological links obtained from the analysis and relate the DE genes and significant ontologies back to the initial purpose of the investigation.

1.2.4 Combining different microarray platforms and results

Entrez IDs

Each probe on a chip is of a known nucleotide string, and has a unique label or ID, for example an AffyID in the Affymetrix environment. When comparing results between studies or platforms or performing downstream analysis on the selected probes, the probe ID itself is of limited use. There is a need to link this probe ID to its related gene or biological component of interest (for example chromosomal region). From the probe ID it is possible to map to an Entrez ID or GeneID (Maglott et al., 2011), which is a unique identifier for genes and other loci for a subset of model organisms[1]. Mapping to a gene-based identifier is an approach that allows comparisons across platforms. When using the Entrez ID as an identifier across platforms there are potential issues, for example, often multiple probes from a single array are mapped to a single Entrez ID. These probes may be different in chemical make-up but because they are much shorter than the regions mapped by an Entrez ID there is a many-to-one mapping. Also probes that mapped to the same EntrezID from multiple platform are likely to be different from one another, that is they are designed from different regions of the genes and hence the expression may or may not be comparable. The mapping of probes to IDs is not static, and care must be taken.

[1]There are many ways to achieve this mapping, (for example Ensembl or Entrez mappings), in this thesis Entrez IDs are used.

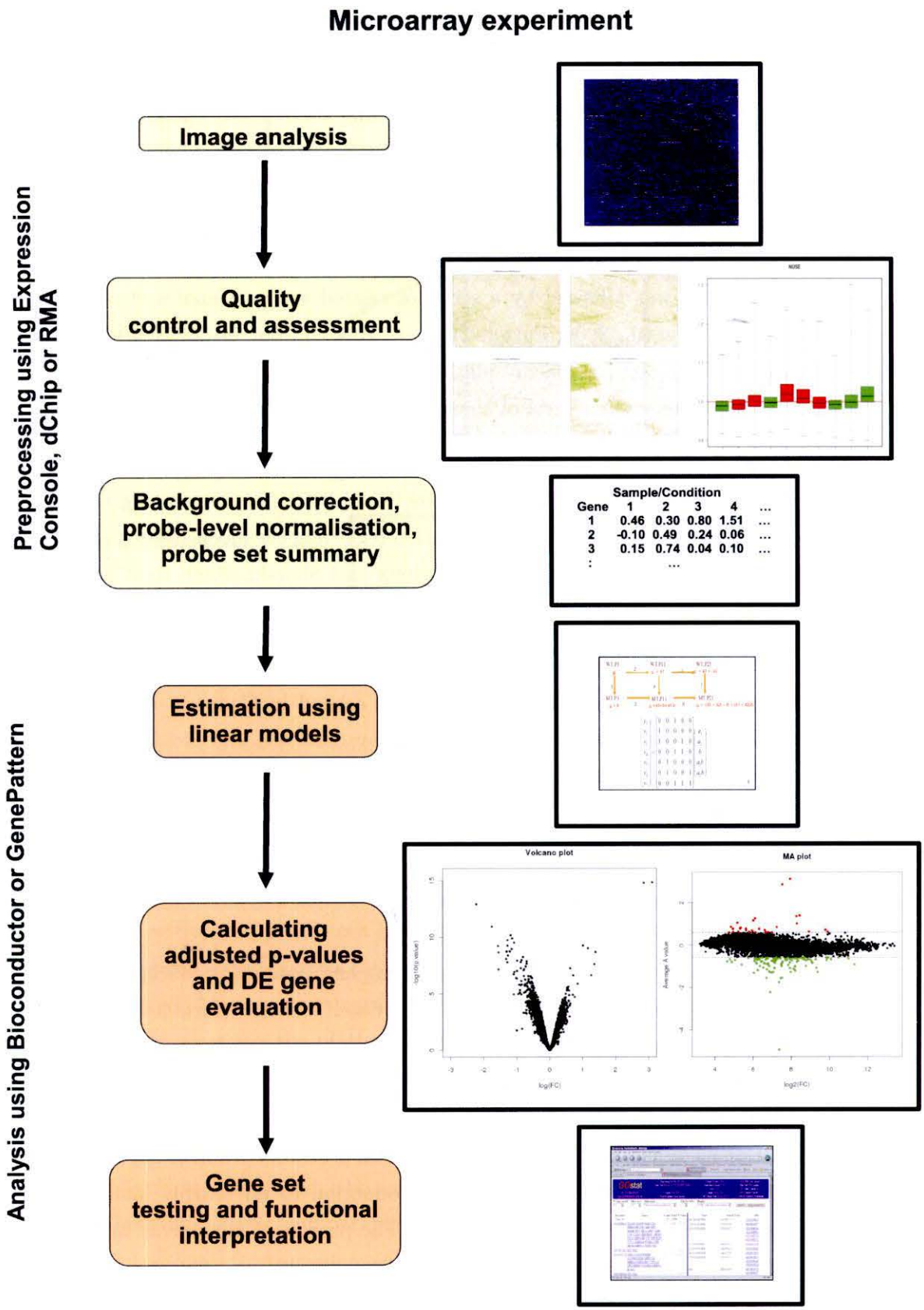


FIGURE 1.2: A flow-diagram representing the major steps involved in the analysis of an individual microarray experiment.

Inconsistency of microarray results

As results from microarray experiments became more and more common a concern was raised; there was a general unsettledness regarding the reliability of microarray results as there appeared to be low levels of concurrence or repeatability between experiments (Kuo et al., 2002; Lee et al., 2004; Michiels et al., 2005; Miklos and Maleszka, 2004; Moreau et al., 2003; Tan et al., 2003; Yuen et al., 2002). To address this concern, the MicroArray Quality Control (MAQC) project was established, published first in Shi et al. (2006) and was followed by a series of related works (Canales et al., 2006; Fan et al., 2006; Guo et al., 2006; Patterson et al., 2006; Shippy et al., 2006). The MAQC study considered seven microarray platforms along with three alternative technologies and found that there *was* a level of consistency between the data sets studied.

The MAQC findings resulted in a flourish of response papers disputing their findings on a platform specific basis (Ha et al., 2009; Kerr, 2007), or discounting their findings completely (Chen et al., 2007). The contention regarding the inconsistency of microarray results is still open (for example see Boulesteix and Slawski, 2009; Russ and Futschik, 2010; Zhang et al., 2008).

Rhodes et al. (2004) suggested that forms of meta-analysis can reduce the impact of inconsistency of microarray results, which was echoed in Campain and Yang (2010). In Chapter 4, how the integration of microarray data sets can aid in the problem of inconsistent results is considered as well as how different integration methods affect the downstream interpretation of these difficult genes.

1.2.5 Integration of microarray data

Integration of microarray results is a challenging and important concept which offers a wide range of benefits (see Chapter 4 for more detail). Large efforts have been made to make publicly available data more accessible and usable (through auxiliary information). However, huge hurdles still exist regarding microarray integration. Platforms differ in probe content, including length and gene region considered, design, technology, relative hybridisation as well as labelling and experimental protocols (Yauk et al., 2004). Moreau et al. (2003) provided an overview and introduction to microarray analysis and the integration of multiple microarray data sets. The process of microarray integration has three major steps, as outlined in Figure 1.3:

1. Individual preprocessing,
2. Microarray integration and,

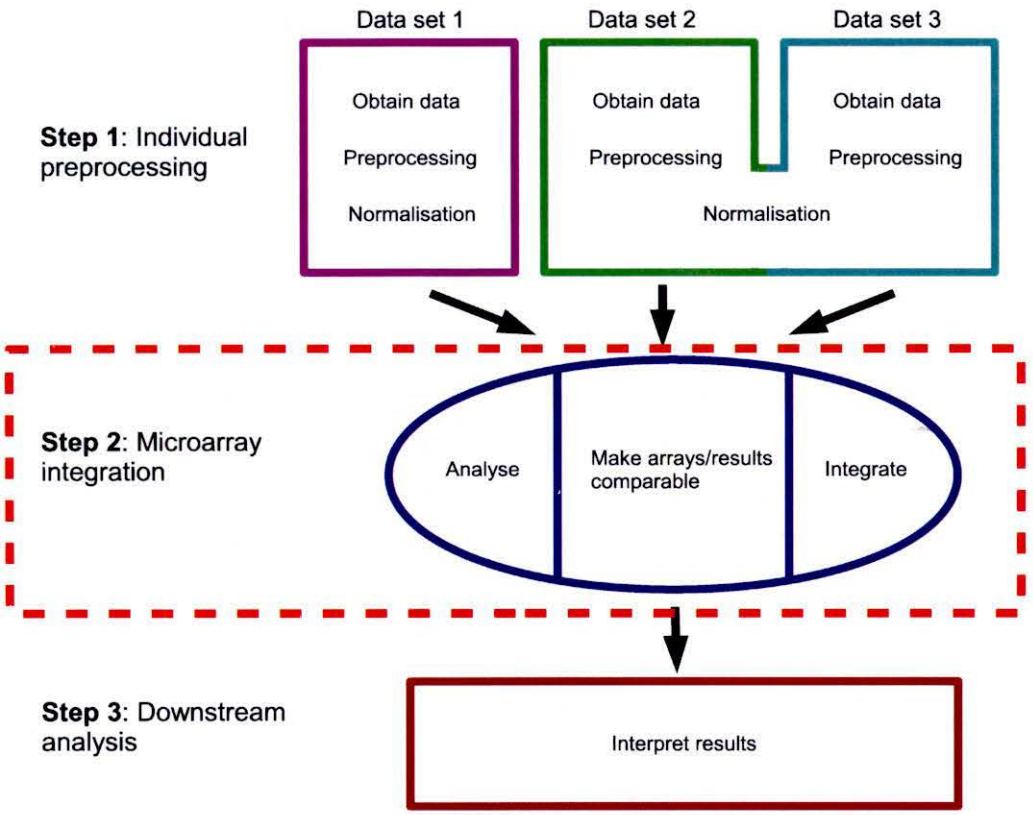


FIGURE 1.3: Graphical representation of steps involved in microarray data integration.

3. Downstream analysis and evaluation.

Individual preprocessing is the initial stage of any microarray experiment (integrative or not). Such tasks include acquiring the data from the experimental source or repository, performing quality control as well as preliminary normalisation. The only difference in an integration setting compared to a completely independent setting is that when two data sets have been developed on the same platform it is possible to normalise them together if desired. Microarray integration is an active area of research. The main aim of this step is to overcome the platform differences mentioned in Section 1.2.2 and to ensure that the largest differences in the analysis are because of biological differences of interest (for example case versus control), and not because of platform differences. Making annotation compatible is only one of the issues, other difficulties include the removal of platform biases and laboratory effects. Moreover, microarray integration needs to include an analysis, comparability and an integration phase but the order these are performed is open to discussion and alters results. Chapter 4 deals directly with this concept and offers several solutions and possibilities regarding array integration. Once microarray integration has been performed and results of the integration obtained, downstream analysis can continue. Such analysis is open ended and depends directly on

the research question including, for example, gene-set tests and ontology evaluations as well as polymerase chain reaction (PCR) validations.

1.3 Integration of clinical and gene expression data

Phenotype data, that is clinical data, is a cornerstone of medical research and microarray data offers a wealth of information at the molecular (expression) level. There is a desire to merge these two data types together. The integration is biologically driven as it is well observed that patients with the same clinical information may have extremely different disease courses and outcomes. It is hoped that by combining clinical and gene expression data the reason behind these differences may be exposed in a productive and diagnostic setting. Microarray data is homogeneous, being continuous readings from different probes or genes. Clinical variables are heterogeneous in nature with a wide range of variables observed including measurements, grades and pathology readings, most of which are discretised by physicians (Lê Cao et al., 2010). There are large advantages of integration. First there is potential for complementary information between the clinical and genomic data. Second, if the clinical information can reduce the number of genes needing to be observed, the cost of the genomic data could be reduced, as only a small sample of genes would be examined, reducing the overall cost of the diagnostic tools compared to a completely genomic approach.

Integrating microarray data with other forms of data, for example epidemiological and clinical data, although still in its infancy, has been used to help unlock some very complex diseases faced in the community. Whistler et al. (2003) discussed how such studies have allowed for a more in-depth study of chronic fatigue, highlighting that this could be a heterogeneous condition labeled as a single syndrome. Schwarz et al. (2009) used a graph theoretical approach to integrate the vast amounts of currently disparate information available for schizophrenia patients to attempt to aid therapeutic direction.

As with the initial problem with publicly available microarray data, Park et al. (2005) and Malin et al. (2010) noted that there are difficulties associated with integrating clinical data. Not only is security and patient privacy a major issue in medicine (this is not overly problematic for omics data) but clinical data also faces comparability, syntax and semantics problems which in coming years will need to be addressed.

1.3.1 Methods of integration

To integrate clinical and microarray data (and in some cases other types of omics data) several methods and frameworks have been postulated in recent years. Such approaches

depend on the data at hand as well as the biological question of interest. Tibshirani and Efron (2002) developed the popular notion of a pre-validated vector (see Section 5.4), summarising the microarray data as a form of cross-validated prediction variable within a regression setting. This method was built on by Lê Cao et al. (2010) who used a two phase approach to the pre-validated vector, making use of partial least squares and random forests in the aim to construct an optimal hybrid classification rule. Dettling and Bühlmann (2004) proposed a penalised logistic regression approach integrating the clinical variables and Gevaert et al. (2006) used a Bayesian networks concept to perform feature selection. Boulesteix et al. (2008) used mixture of experts models to combine the data in a non-linear way to obtain a combined signature.

Boulesteix et al. (2008) rightly remarked that there is a difficulty in evaluating the advantage of including (the expensive) microarray data with the clinical data. Heuristically the validation of integration becomes complicated because of the interplay between gene expression and clinical variables:

> Microarray data and clinical data may be redundant because the gene expression influences clinical variables or vice versa, or because both clinical and microarray variables are influenced by common latent unobserved mechanisms (Boulesteix et al., 2008),

and points out that additional biological knowledge is needed to address this question. Hence when integration takes place care needs to be taken. Tibshirani and Efron (2002) developed their method after concerns that other integrative methods were relying too heavily on the gene expression data, biasing results through methodology. Simply combining clinical and expression data may swamp the clinical data due to the large amount of expression results, but analysing them separately and joining them together in a model toward the end may be less than optimal when there is high correlation between results. Truntzer et al. (2008) evaluated if clinical or expression variables were being under or overestimated, and found that in general in integrative studies there is a tendency to overestimate the effects of genes due to the selection process and underestimate clinical variables due to the omission of relevant genes.

In this thesis, several difficulties associated with the analysis of clinical and gene expression data are discussed, some of these methods, concepts, analyses and results have already been published (or are currently under review) by the author in statistical and collaborative research. Chapter 2 discusses the data sets used to motivate the areas of concern addressed within this thesis. It includes description of the data sets and preliminary analyses including quality control and initial data assessments. Chapter 3 is concerned with some common problems encountered while analysing clinical data

including missing data, class imbalance and instability of models. It covers work from Campain et al. (2011). Chapter 4 develops the timely concept of gene expression integration and some of the difficulties encountered in this type of analysis and involves work published in Campain and Yang (2010), Campain et al. (2010) and Yang et al. (2011). The thesis concludes with a case study (Chapter 5) highlighting work by Mann et al. (2011) and Sarah-Jane Schramm et al. (2011). It concentrates on the clinical predictors of good survival prognosis coupled with a molecular signature. The motivation of the clinicians who initiated this study was to utilise the molecular and phenotype information to develop a predictor of Stage III melanoma. The predictive capabilities are assessed individually and as an integrated model. For validation purposes the molecular signature is also compared to publicly available Stage III melanoma data sets (Section 2.3.2) through a form of meta-analysis.

Chapter 2

Data sets

This chapter outlines the data sets used in this thesis. There are three data themes used to explore and motivate the statistical concepts presented. Section 2.1 outlines an early pregnancy data set, containing only clinical data. This data set is used throughout Chapter 3 and contains a proportion of missing data. Section 2.2 is an outline of four hypersensitive gene expression data sets. Results of quality control for each data set as well as individual preliminary analysis are also presented. This collection of data sets is used in Chapter 4 to examining meta- and mega-analysis. A collection of melanoma data sets is outlined in Section 2.3. These data sets include four publicly available gene expression data sets and one data set provided by collaborators, which includes both clinical and gene expression data. Melanoma data is used in Chapter 4 and explored extensively in Chapter 5.

2.1 EPU data

This data set has been provided by the Early Pregnancy Unit at the Nepean Hospital, NSW, Australia (Riemke et al., 2011). The data is from an observational study that was carried out at the Nepean Hospital between November 2006 and February 2009. The study considers first trimester pregnant women presenting at the unit with a viable intra-uterine pregnancy (IUP) at their first examination. To present at such a clinic, symptoms often include pain, bleeding, history of miscarriage and other reasons. The eventual diagnosis for each woman was established at the end of the first trimester as either viable or non-viable (miscarried). The purpose of the study was to build a model that has the ability to predict the outcome of the first trimester for the IUP at the initial consultation. This data set will be refered to as the 'EPU data' within this thesis.

Number of variables not observed	0	1	2	3	4	5	6	7	8	9
Number of subjects	63	143	99	68	25	13	3	1	0	1

TABLE 2.1: Number of variables not observed per sample for the EPU data.

A multitude of variables for each woman are observed including 21 historical, clinical and ultrasound variables for analysis. Historical variables include: age, ethnic group, parity, number of previous natural deliveries and caesarean sections, number of previous miscarriages and terminations, number of previous ectopic pregnancies and molar pregnancies, mode of conception, date of last monthly period (LMP), certainty of dates, gestational age in days by menstrual dates, and smoking. Clinical variables include: maternal height and weight, indication for ultrasound, bleeding, clots, presence of pain, gestational age in days by ultrasound, consistency with menstrual dates. The ultrasound measurements documented were gestational sac (GS) and yoke sac (YS) in three planes with means calculated, crown-to-rump length (CRL), foetal heart rate (FHR), ovarian pathology as well as presence and dimensions of a subchorionic haematoma.

The total number of samples for this data set is 416 of which 33 samples pertain to an eventual non-viable pregnancy at the end of the first trimester and 383 are viable pregnancies. Only 63 (15%) patients had a complete set of covariates, this included only 8 miscarriage cases (24% of the 33 miscarriage samples). A complete case analysis is not advisable with such a small percentage of the final data set (15%) as these samples are not necessarily indicative for the majority of samples. Table 2.1 shows the number of missing variables per sample and Table 2.2 shows the percentage of missing data per variable. There is an average overall missingness of 8.51%. Subchorionic bleed contains a missingness of over 50% and was included in the analysis after discussions with clinicians despite this large percentage.

Results for the EPU study were originally presented in Riemke et al. (2011), which serves as a reference for a full description and analysis. Riemke et al. (2011) initially considered 21 variables which, after model selection, was reduced to a final weighted logistic regression model containing eight clinical variables. Table 2.3 shows the odds ratios for the final weighted logistic regression model with odds ratios greater than 1 indicating a higher risk of miscarriage. A different analysis of this data set is included in Section 3.5, resulting in a slightly different model than the one in Riemke et al. (2011).

Variable	Number of missing data	Percentage
Age (of mother)	0	0.00
Number of natural deliveries	4	0.96
Number of previous caesareans	4	0.96
Number of previous miscarriages	2	0.48
Certain of dates (Y/N)	57	13.70
Gestational age (in days)	50	12.02
Bleeding (Y/N)	4	0.96
Clots (Y/N)	9	2.16
Smoker (Y/N)	21	5.05
VAS 0–10	65	15.62
CRL	1	0.24
FHR	15	3.61
Ultrasound gestational age (in days)	25	6.01
Consistent with menstrual dates (Y/N)	73	17.55
Both ovaries seen (Y/N)	60	14.42
Subchorionic bleed (Y/N)	231	55.53
Reason for attending clinic - Bleeding (Y/N)	8	1.92
Reason for attending clinic - Abdominal pain (Y/N)	8	1.92
Pain (Y/N)	9	2.16
GS mean (in mm)	21	5.05
YS mean (in mm)	77	18.51

TABLE 2.2: Number of missing data per variable for the EPU data.

Variable	Odds Ratio
Number of previous cesarians	0.44
Gestational age	1.05
Bleeding	1.93
Clots	6.12
Ultra sound gestational age	0.91
Consistent with menstrual dates	0.50
GS mean	0.88
YS mean	1.54

TABLE 2.3: Odds ratios from a weighted logistic regression for miscarriages and viable pregnancies, Riemke et al. (2011). Odds ratios larger than 1 indicate a higher risk of miscarriage.

2.2 Hypertensive versus normotensive rats

2.2.1 Data sets and samples

This is a collection of four data sets, all examining hypertensive and normotensive rats. All four data sets are publicly available as raw data via the Gene Expression Omnibus[1] (GEO) (Barrett et al., 2005) and ArrayExpress[2] (AE) (Parkinson et al., 2005) websites. The rats studied come from three different species, the spontaneously hypertensive rat (SHR), Lyon hypertensive (LH) and the Wistar Kytot (WKY) rat - a normotensive animal. All four studies were developed in independent laboratories, under their respective Affymetrix protocols, for independent purposes.

Cerutti et al. (2006), known henceforth as the 'Cerutti data' contains 15 samples, 5 SHR, 5 LH (that is 10 hypertensive samples) and 5 WKY samples. Clemitson et al. (2007), to be known as the 'Clemitson data' contains 10 samples, 5 SHR and 5 WKY samples. Grayson et al. (2007), the 'Grayson data', is the smallest study in the collection, containing 6 samples, 3 SHR and 3 WKY samples. All three of these data sets are generated on the Affymetrix GeneChip Rat Genome 230. Rysä et al. (2005) to be refered to as the 'Rysä data', is the largest and oldest of the studies, hybridised on the Affymetrix GeneChip Rat Genome U34 Array set A. It includes 23 samples, 12 SHR and 11 WKY. Table 2.4 contains a summary of the four different data sets including their rat species, source of the sample and published public database ID.

2.2.2 Quality control and preprocessing of arrays

Raw data for the Affymetrix expression microarrays consists of individual CEL files, which contains measured intensities and locations of probes on an array. Preprocessing includes background correction, probe-level normalisation and probe set summary. For all four data sets these steps were performed using the `affy` package (Gautier et al., 2004), with the probe-set normalisation performed via robust-multi-array averaging (RMA) (Irizarry et al., 2003). The quality control (QC) of each array is assessed by considering the probe level model (PLM) of the data. Different QC measures exist that can be used to assess the quality of arrays within individual microarray experiments as well as in integrative microarray experiments. These include, but are not limited to:

1. **Spatial image plots:** This is a pseudo image of the expression array, large area imperfections or poor hybridisations can be visualised. Such imperfections can be

[1] `http://www.ncbi.nlm.nih.gov/geo/`
[2] `http://www.ebi.ac.uk/arrayexpress/`

Data set	ID	Tissue	Samples	Platform	Number of unique genes
Cerutti et al. (2006)	AE: E-MEXP-357	Heart	10 HT (5 SHR, 5LH) 5 NT (WKY)		
Clemitson et al. (2007)	AE: E-TABM-45	Kidney	5 HT (SHR) 5 NT (WKY)	Affymetrix RAE230	10207
Grayson et al. (2007)	GEO: GSE 8051	Saphenous artery	3 HT (SHR) 3 NT (WKY)		
Rysä et al. (2005)	AE: E-GEOD-2116	Heart	12 HT (SHR) 11 NT (WKY)	Affymetrix rgU34a	4941

TABLE 2.4: Details of the four hypertension data sets used within this thesis, all data sets are publicly available and were analysed as a meta-analysis study in Chapter 4.

seen as large blemishes on the images of particular arrays or entire arrays that appear different to the cohort.

2. **Relative Log Expression (RLE) boxplots:** These are a series of boxplots, one per sample, where the log expression values for each sample are plotted. Good QC yields RLE plots that are centred around 0 with a small spread, with issues arising when boxes (representing spread) are larger than the majority or vary away from 0 (Gregory et al., 2007).

3. **Normalised Unscaled Standard Error (NUSE) boxplots:** A series of boxplots, one per sample, of the standard errors of the normalised samples. The NUSE values are centred around 1, low quality arrays might have boxplots that vary significantly from the remaining cohort, including having a larger spread to the boxplot or an elevated mean (Gregory et al., 2007).

4. **Hierarchical clustering:** Hierarchical clustering plots for each of the data sets were obtained and plotted including dendrograms. In this thesis, for all included clustering plots, the 500 most variable genes are used[3], judged on overall expression variability. The Euclidean distance was used as the distance metric with the complete agglomeration method. In an agglomerative approach (used in this thesis) samples are considered originally as singular and then combined into pair-clusters, which are then sequentially combined to other clusters until one large group is created. Different metrics are used to combine the clusters with the application of a linkage function. Samples are displayed in their clusters, with coloured bars below the dendrogram used to represent the different observed factors, biological and other, to determine the largest source of the variability.

All QC plots for the four data sets are shown in Appendix A. These combined QC measures suggest that two samples from the Rysä data set could be removed from downstream analysis, but have not because their quality is not overly poor. All other samples, from all other data sets, passed QC measures convincingly. The clustering algorithm recognises the differences in the classes (the top most coloured strip, subplot (d) Figures A.1 – A.4) as the main differences within the data. Visually considering the dendrogram is important, as it depicts how the samples cluster. The colours bars represent observed factors such as class, platform and study. This gives a visual suggestion as to the strongest factor relating the samples. In an ideal situation class should be the strongest relating factor with limit noise from other sources. The RMA algorithm implemented

[3] 500 of the most variable genes were selected as this was considered a manageable amount of genes, but not overly excessive.

in the `affy` (Gautier et al., 2004) package has been used to perform background correction, quantile normalisation and probe specific summarisation via the median polish algorithm.

When mapping between platforms is required for analysis, the individual probes for each platform are mapped to their corresponding Entrez IDs (Maglott et al., 2011) using the databases 'rae230a.db' and 'rgu34a.db' (version 2.4.1) from `Bioconductor` (Gentleman et al., 2004). When more than one probe pertained to the same Entrez ID, the mean value was taken as the expression level. Only overlapping genes were retained for further analysis after QC. A total of 4,678 unique genes are common to both platforms.

Each of the studies have been analysed individually. From the QC plots the boxplots indicate that expression levels across most arrays (within a data set) are comparable. There are however some arrays of concern, for example the grouping present in the Cerutti data (Figure A1 b). This grouping however is confounded with animal and date effects. Such confounding issues are common and need to be considered carefully when interpreting data from a biological perspective. Splotches are present in some of the spatial images, for example Figure A.2 (a), these blemishes are at an acceptiable level when compared to images from the PLM gallery[4] and do not cause the arrays to fail QC analysis.

2.2.3 Individual analysis

For each of the four data sets, individual analysis was performed. Several graphical tools as well as the number of DE genes are considered in each analysis. The genes that are considered DE and the number of these genes that are also positive control genes is recorded.

Graphical tools There are several graphical tools often used to aid in the interpretation of the analysis. Common plots produced may include:

1. **MA-plots:** MA-plots are scatter plots displaying the $\log_2$(FC) on the y-axis and the average expression value on the x-axis. In these plots, within this thesis, red indicates genes that have a positive FC > 1.5 (that is genes that are more highly expressed in the hypertensive samples compared to the normotensive samples) and the green indicates genes that have a negative FC < 1.5 (genes that are more highly expressed in the normotensive samples than the hypertensive samples). MA-plots are useful in identifying intensity dependent patterns in the ratios (Dudoit et al.,

[4]http://plmimagegallery.bmbolstad.com/

2000), and recognising the high abundant (high average expression value) genes with large FC values. Overall, these scatter plots help to visually gauge the signal to noise ratios within an experiment. See for example Figure A.1 (e).

2. **Volcano plot:** Volcano plots are a way to visualise results of the analysis. The volcano plot ($-\log_{10}$(p) on the y-axis and $\log_2$(FC) on the x-axis) quickly visualises the number of genes that are significant in both plotted conditions. Moreover, such plots allow the visualisation of the amount of *discordance* in the genes that are significant based on FC compared to those that are significant based on p-values. See for example Figure A.1 (f).

For each of the considered data sets, these interpretive tools are in Appendix A, plotted with their QC results. These plots suggest that there is only a small set of genes that are DE based on FC, and this further reduces when the dual conditions of FC and FDR are used as a combined selection criteria.

DE genes For each of the four data sets DE analysis was performed using least-squares regression (implemented through `limma`, Smyth, 2004). A set of DE genes can be selected in many ways, for example as the genes with an absolute FC greater than 1.5 **or** a FDR less than 0.05 **or** genes satisfying both criteria. Table 2.5 shows the number of genes recorded as DE for each data set. Table 2.6 shows the number of DE genes which are also in the positive control list (Section 2.2.4).

Sample size of the initial data sets has a large impact on the number of genes selected as DE. This is especially true when considering the Clemitson data ($n = 15$) and the Rysä data ($n = 23$). These data sets have a very large number of DE genes selected via FDR. For these two cases it appears that the FDR condition is no longer acting as a selection criteria when the dual conditions (FDR <0.05 and $|\text{FC}| \geq 1.5$) are used, because so many genes (at times as many as 25% percent) pass the FDR threshold. Interestingly the number of positive control DE genes selected by each of the single DE conditions was extremely low for each data set. This highlights the point that replication of DE genes across studies, as positive control genes are study verified genes, is very rare (Ein-Dor et al., 2005).

2.2.4 Important gene lists in microarray and integrative analysis

The hypertensive/normotensive data sets highlight some important gene lists in microarray and integrative analysis including positive control genes, house-keeping genes as well as consistent results between independent analyses. Each of these types of genes has a particular place and impact on the analysis along with the conclusions drawn.

Data set	$\|FC\| \geq 1.5$ (up,down)	FDR < 0.05	$\|FC\| \geq 1.5$ (up, down) **and** FDR < 0.05
Cerutti data	72 (53, 19)	396	58 (45, 13)
Clemitson data	113 (17, 96)	1178	113 (17, 96)
Grayson data	84 (53, 31)	35	30 (17, 13)
Rysä data	108 (41, 67)	972	108 (41, 67)

TABLE 2.5: Number of DE genes for each data set.

Data set	$\|FC\| \geq 1.5$ (up,down)	FDR < 0.05
Cerutti data	2 (0, 2)	3
Clemitson data	2 (0, 2)	7
Grayson data	3 (2, 1)	0
Rysä data	3 (2, 1)	11

TABLE 2.6: The number of DE genes for each data set that are also considered positive controls for hypertension.

Positive control genes Positive control genes are genes that have a documented relationship to the condition of interest. Hypertension is currently a relatively under-studied condition, affecting humans (Marques et al., 2011c), mice (Marques et al., 2011b,a) and rats (Campain et al., 2010) amongst other mammals. However, within current literature it is possible to establish a list of genes that have highlighted links to hypertension, on the individual study basis (Table 2.7, where genes are shown as gene symbols). It is common in microarray literature that genes reported in one study cannot be replicated in another (Ein-Dor et al., 2005; Zhang et al., 2009; Boulesteix and Slawski, 2009). A positive control gene list, containing 28 of these considered DE genes is monitored when the hypertensive data sets are analysed within this thesis. This monitoring is performed to establish whether or not positive control genes are repeatable in integrative studies.

House-keeping genes House-keeping genes are genes that are considered to have **no** differential expression across the condition being studied (Butte et al., 2001; Hsiao et al., 2001; Thellin et al., 1999; Warrington et al., 2000). Although this is conceptually true, this is *not* always the case for house-keeping genes (Ke et al., 2000; Suzuki et al., 2000). Moreover, such genes are well studied in humans and less well considered in rats. To obtain the list of house-keeping genes used in this thesis, the list of human house-keeping genes was obtained and mapped to the rat homologs, eventuating in a list of known rat genes, which are assumed to be house-keeping under the hypertensive versus normotensive conditions. With such a mapping 264 genes were retained as house-keeping genes and are listed in Table A.1.

Positive control gene symbols	Reference
Ace	Guo et al. (2005); Massie (1998)
Agt	Guo et al. (2005); Lévesque et al. (2004); Rogus et al. (1998)
Agtr1a	Le et al. (2003, 2004)
Agtr1b	Bertram and Hanson (2002); Langley-Evans (1997, 2000)
Agtr2	Min et al. (2008); Touyz et al. (1999)
Atp2a2	Cerutti et al. (2006); Kiec-Wilk et al. (2007)
ApoE	De Leeuw et al. (2004); Wenquan Niu et al. (2009)
Ccl2	Sanchez et al. (2007)
Cd36	Pravenec et al. (2001, 2003)
Dbh	Chen et al. (2010); Cubeddu et al. (1981); Greco et al. (1978)
Fstl1	Cerutti et al. (2006)
Gstm1	McBride et al. (2005)
Gstp1	Ohta et al. (2003)
Gstt1	Marinho et al. (2007)
Myh6	Campain et al. (2010)
Myh7	García-Castro et al. (2003)
Nos1	Paliege et al. (2006); Tambascia et al. (2001); Weichert et al. (2001)
Nos3	Guo et al. (2005); Kimura et al. (2003); Malhotra et al. (2004)
Nox1	Dikalova et al. (2005); Matsuno et al. (2005)
Nox4	Lu et al. (2010); Paravicini et al. (2004)
Nppa	Cerutti et al. (2006); Guo et al. (2005); Newton-Cheh et al. (2009)
Nppb	Newton-Cheh et al. (2009)
Pgm1	Cerutti et al. (2006)
Ren	Barrett and Mullins (1992); Hartner et al. (2006); Mullins et al. (1990)
Siat7A	Cerutti et al. (2006)
Sod1	Carlström et al. (2009)
Sod2	Archer et al. (2010); Rodriguez-Iturbe et al. (2007)
Uts2r	Watanabe et al. (2006, 2009)

TABLE 2.7: Positive control genes and their references.

Consistent DE genes Ideally DE genes should be consistently selected over multiple studies, considering the same conditions, regardless of the laboratory and array technology used (Campain et al., 2010). A gene that is DE across multiple studies is called a *consistent DE gene*. For the hypertensive/normotensive data sets the majority of DE genes should be DE across the multiple studies as they are comparing similar conditions. The number of consistent DE genes depends on the DE selection criterion, and the number of studies in which it is common. There is only a limited number of DE genes common to two or more of the considered hypertension studies (Table 2.8). There are 78 DE genes common to at least one other study over the complete range of DE criteria. Two consistent genes are also positive control genes, ApoE and Cd36, with Cd36 being DE in all studies under the FC criteria. This small overlap of DE genes confirms the findings of Ein-Dor et al. (2006).

Number of common studies		Selection criterion		
		$\|FC\| \geq 1.5$	FDR p-value <0.001	$\|FC\| \geq 1.5$ **and** FDR p-value <0.05
2 studies	DE genes	25	42	18
	Positive controls	ApoE	Cd36	Cd36
3 studies	DE genes	6	4	1
	Positive controls	-	-	-
4 studies	DE genes	1	0	0
	Positive controls	Cd36	-	-

TABLE 2.8: Number of consistent DE genes for the hypertensive/normotensive data sets.

2.2.5 Inconsistent DE genes

When the DE genes selected across each of the individual studies are compared, inconsistencies in the DE gene lists become evident. A gene is considered to be *inconsistent* if the gene is selected as DE in two (or more) of the studies, but the FC directions (up or down) do **not** agree. This inconsistency implies that one study found the gene in question to be more highly expressed in the condition (hypertension in this case) and another found it to be less highly expressed in the condition (that is more highly expressed in the normotensive animals). Such inconsistencies are a problem and something that needs to be carefully considered. An assumption in the integration of results from multiple studies is that results contain at least some level of consistency. If the **direction** of the DE genes are conflicting, difficulties may arise when integrating results where only non-directional statistics (such as p-values) are considered. In this thesis, to ensure the robustness of the inconsistent gene list, only genes that are DE based on **both** FDR and FC are considered.

When comparing the four hypertensive studies, 11 inconsistent genes (shown as gene symbols) are observed (Table 2.9), one of which is a positive control gene (Nppa). These inconsistencies could be due to a number of factors including small sample sizes, heterogeneity in studied samples or lab effects (Lai et al., 2009; Yang et al., 2011), all of which are shortcomings that mega-analysis attempts to overcome (to be discussed in Section 4.2). Figure 2.1 is an example of the scatter-plots available for each of the inconsistent genes. Plotted is the expression values for each sample within a data set, the classes are represented as different colours (red = hypertensive, black = normotensive). These plots show at the expression level what happens when inconsistent genes are being observed. Figure 4.8(a) is a heatmap of the inconsistent genes, the average FC for each gene is displayed for each of the four data sets. This plot provides a graphical representation of the inconsistencies for each gene. Green values in this heatmap imply negative average

Gene Symbol	Postive FC	Negative FC
Ak2	Clemitson	Rysä
Bet1	Cerutti	Clemitson
Clcn	Cerutti	Clemitson and Rysä
Dars	Cerutti	Clemitson
C0s2	Cerutti	Rysä
Ivns1abp	Cerutti	Clemitson
Nppa*	Rysä	Cerutti
Pfkp	Cerutti	Clemitson and Rysä
RGD1308772	Cerutti	Clemitson
Rrad	Cerutti	Rysä
Slc6a6	Rysä	Clemitson
* indicates an inconsistent gene that is also a positive control gene		

TABLE 2.9: 11 inconsistent genes with differing FC directions for two (or more) studies selecting them as DE genes.

FC values across a give data sets (expression levels are lower in hypertensive samples) and red values imply a positive FC level.

2.3 Melanoma data

2.3.1 Mann data set

The Mann et al. (2011) data set, refered to as the 'Mann data', studied in Chapter 4 and Chapter 5 contains two types of data: clinical (n=83) and gene expression data (n=79). This data is paired, but contains an additional four clinical samples. Samples are from Stage III melanoma patients, with biopsies taken from their metastasised sites. A more complete data and analysis description can be found in Chapter 5 and in the clinical paper Mann et al. (2011). A total of 79 tumour samples were obtained from the Melanoma Institute Australia (MIA) Biospecimen Bank, as a collection of fresh-frozen tumours, obtained with appropriate approval and informed consent. These samples were collected since 1996 through MIA, formerly the Sydney Melanoma Unit.

Clinical data

A range of clinical variables were obtained both from the sample itself as well as the patient. Samples were reviewed by a pathologist and the variables were assessed. The variables include: percentage of non-tumour cells, percentage of necrosis, degree of pigmentation, predominant cell shape and cell size of most cellular portion of tumour were

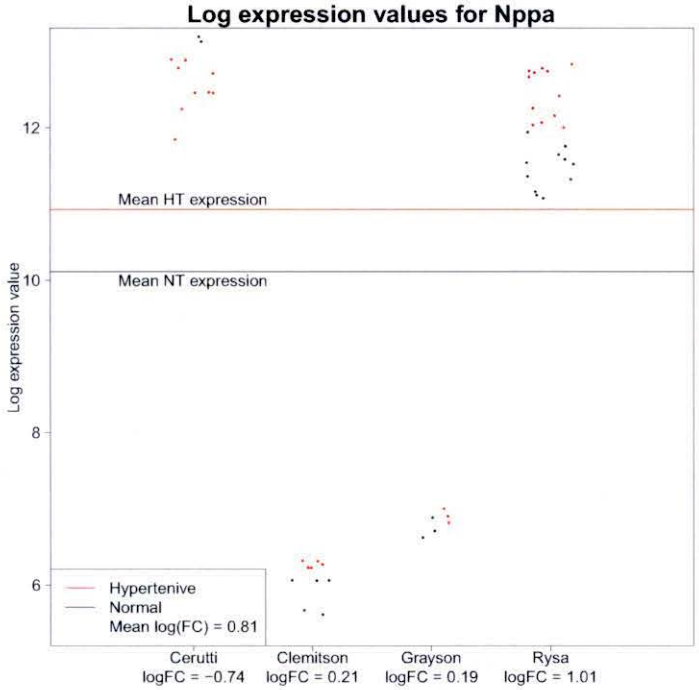


FIGURE 2.1: Nppa expression plot, this plot highlights the inconsistencies in results especially between expression levels in the Cerutti and Rysä data sets where FC is in opposite directions.

assessed. Pathologic variables were obtained for the samples. These include: number of nodes involved, largest nodal metastasis size and presence of extranodal spread. Other clinical variables include: age, gender, stage at diagnosis, body site (classified by pattern of sun exposure: continuous, intermittent, rarely exposed), presence of an associated nervous, degree of solar elastosis in the peritumoral skin, Breslow thickness (mm), Clark level, histologic melanoma subtype, and presence of regression, ulceration, vascular or lymphatic invasion.

Gene expression data

From the clinical samples, 79 paired expression arrays were also obtained. These were assayed using fresh frozen samples on Illumina Human Beadarrays.v3 arrays with 48,802 probes, from the metastasised tumour. The number of probes considered in the analysis was reduced to 26,085 after unexpressed probes were removed. A probe was considered unexpressed if the detection p-value was less than 0.01 (Du et al., 2008). The Mann et al. (2011) study took the 79 samples and reduced this down to the extremes of the survival groups:

1. **Class 1** *poor prognosis*, survival <1 year after surgical resection and died due to melanoma ($n = 25$), and;
2. **Class 2** *good prognosis*, survival > 4 years after surgery, with no sign of relapse ($n = 23$).

The clinical, pathologic and molecular parameters (including results of somatic mutation profiling) were analysed using multiple imputation and logistic regression for determinants of outcome. Detail of this complete analysis presented in Chapter 5.

2.3.2 Public melanoma data

In this thesis, a collection of four public melanoma gene expression data sets are studied. These studies examine the gene expressions of melanoma patients and develop gene signatures that distinguish between good prognosis (long term survival) and poor prognosis (short term survival). Definitions of classes and platforms used differ between studies and are summarised in Table 2.10. The examination of these data sets occurs in this thesis in two sections. The first is the analysis of the raw, or minimally preprocessed data for inhouse feature selection, classification and evaluation, in the meta-analysis melanoma case study (Section 4.4.2). The second makes use of their published gene lists, and the evaluation of these lists through classification as a method of validation through meta-analysis (Section 5.3.5).

Bogunovic data

Bogunovic et al. (2009), known as the 'Bogunovic data', is an examination of 33 metastatic melanoma lesions, processed on the Affymetrix Human Genome U 133 Plus 2.0 chip, in the Rockefeller University Gemonics Care laboratory according to the Affymetrix protocol. For the purpose of classification, two groups were defined based on survival time, those with prolonged survival (greater than 1.5 years) and those with 'shorter survival' (less than 1.5 years). Bogunovic et al. (2009) identified 266 genes/gene elements associated with post-recurrence survival. This signature was assessed for its predictive capacity against the 2001 Tumour-Node-Metastasis staging system, the presence of TIL, T-cell CD3 positive, and mitotic index.

John data

In Thomas John et al. (2008), refered to as the 'John data', researchers used inhouse oligonucleotide arrays (30,888 probes) to examine lymph node sections from 29 patients

with Stage IIIB and IIIC melanoma. Selecting a cut-off for good prognosis of 24 months survival, there were 16 poor-prognosis and 13 good-prognosis patients. The mean time to progression for the good-prognosis group was 40 months compared with four months in the poor-prognosis group. Multivariate analysis showed no statistically significant differences in age, gender, staging, the use of adjutant interferon therapy, or the presence of tumour-infiltrating lymphocytes between the two patient groups. From the list of 2,140 significantly DE genes, two outcome-related tests were developed, a 21-gene element predictor and a 5-gene qPCR predictor. In an independent sample of 10 tumours, the qPCR predictor correctly classified nine of them with respect to good or bad prognosis.

Jönsson data

Göran Jönsson et al. (2010), known henceforth as the 'Jönsson data', examined 57 subcutaneous and lymph node metastases, on Illumina Human Beadarrays.v2 (HumanWG-6 v2 Expression Beadchip) and the Illumina system according to the manufacturer's protocol. The raw data is not publicly available, however the data was normalised with Beadstudio v3 Illumina software using a cubic spline normalisation method. Data was also log-transformed and mean-centred across arrays. This level of processed data is in the public repository. For further classification analysis class groups are defined as individuals alive or dead, according to the clinical information provided.

Winnepenninckx data

Winnepenninckx et al. (2006), referred to here as the 'Winnepenninckx data', is the analysis of 58 primary cutaneous samples, on Agilent oligonucleotide whole-human-genome 44K dual colour microarrays. Winnepenninckx et al. (2006) identified a 254-gene expression signature. The two groups were classified such that the poor prognosis group had a distant metastasis within 4-years and the good prognosis group had no distant metastasis within 4-years. The original study also compared different prognosis groups. Patients were manually collated into a bad prognosis group if the primary melanoma was more that 4mm thick, or patients had an ulcerated melanoma that was more than 2 mm thick. The gene expression differences between these two groups were also considered for classification purposes.

Data set	Platform	Purpose of original study	Class cut-off
Mann et al. (2011)	Illumina Human Beadarrays.v3	DE analysis and classification	Good - Survival 4+ years Bad - Survival less than 1 year
Bogunovic et al. (2009)	Affymetrix Human Genome U 133 Plus 2.0	DE analysis and classification	Good - Survival 1.5+ years Bad - Survival less than 1.5 years
Thomas John et al. (2008)	Inhouse oligonucleotide arrays (30,888 probes)	DE analysis	Good - Survival 24+ months Bad - Survival less than 24 months
Göran Jönsson et al. (2010)	Illumina Human Beadarrays.v2	DE analysis and class discovery	Good - Alive Bad - Dead
Winnepenninckx et al. (2006)	Agilent 44K dual colour microarrays	DE analysis and classification	Good - no dist. met. within 4 years Bad - dist. met. within 4 years

TABLE 2.10: Summary of melanoma data set considered in this thesis.

Chapter 3

Clinical data

Statistics in clinical research faces many challenges including the presence of missing data and model instability. Clinical data often has a wide variety of variables to be considered, including discrete, continuous and ordinal, coming from diverse sources including measurements, counts, technological readings as well as subjective judgements in a clinical or medical context.

Although the elements considered in this chapter are common to many clinical data sets, this chapter in particular was motivated by early pregnancy clinical data provided by the Early Pregnancy Unit, part of the Nepean Hospital, Sydney, Australia, the 'EPU data'. Full description of this data was provided in Section 2.1. To summarise some of the characteristics, the EPU data is highly imbalanced (8% minority class), contains approximately 8.5% missing data overall and has only 63 complete out of a total of 416 cases. The data contains a fully observed binary response, viable or miscarriage pregnancy, and covariates range from maternal history to ultrasound variables, and are either continuous or categorical. This data set is studied at the end of this chapter, providing an example of the use of the presented solutions to missing data and unstable models. Different and altered versions of the EPU data are also used for the two simulation studies that explore elements of missing data and unstable regression models.

Although binary responses are not the only response within a clinical setting they are common, representing for example treated versus non-treated, alive or dead, presence or absence of a particular variable/condition. In a binary class setting one may construct a predictive model using weighted logistic regression. The binary response vector $\mathbf{y}$ has length n with classes coded 0 and 1. Let $\boldsymbol{X}$ be an $n \times Q$ matrix for a sample of size n and Q explanatory variables. The probability of an 'event' occuring is π_j, $(\pi_j = P(Y_j = 1))$.

Let π denote the vector of the n event probabilities. We write

$$\ln(\frac{\pi}{1-\pi}) = \boldsymbol{V}\boldsymbol{\beta}, \tag{3.1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_Q)$ and $\boldsymbol{V}$ is a $n \times (Q+1)$ data matrix, with the first column containing all 1's and the remaining $n \times Q$ matrix corresponding to $\boldsymbol{X}$. The parameter vector $\boldsymbol{\beta}$ can be estimated via classical methods such as 'maximum likelihood methods', 'least-squares' or 'weighted least-squares' (Venables and Ripley, 2002). Observations can be weighted according to a weight vector $\boldsymbol{w}$ such that its components sum to n, the sample size. In such a case the regression parameters are estimated based on maximising the weighted log-likelihood.

When applying logistic regression there are a range of problems that are commonly encountered. Two major challenges are:

1. Missing data in the explanatory variables, the response variables, or both.
2. An unstable final model, where small changes in the data produce large changes in the final model.

Currently there is a large variety of multiple imputation approaches available to address the problem of **missing data**, ranging from *ad hoc* to highly sophisticated statistical modelling. However, very little is known regarding how these various methods compare. Very recently, Abrahantes et al. (2011) compared a range of multiple imputation methods, but how these methods depend on the amount of missing data was not investigated. In this chapter an empirical study is used to investigate how regression coefficients depend on induced missingness and which imputation method handles missingness most effectively.

At times model building may lead to the construction of an **unstable model**. Unstable models are models with non-reproducible coefficients. For such models small perturbations in the data used to construct the model result in large model changes. Moreover they often have very poor predictive qualities. This is especially prevalent when the data contains a highly imbalanced class distribution. Solutions to class imbalances have been proposed, for example to either over or under sample class samples to avoid imbalance (Weiss and Provost, 2001). But when sample sizes are relatively low this is difficult. Furthermore, questions have been raised about the statistical validity of such approaches. Imbalanced class distributions are a common problem in clinical data, and can exacerbate the issue of unstable models. Solutions to handling these unbalanced analysis questions are offered, changing the performance measures used and applying weights within regression models to compensate for the vastly different class sizes.

As model uncertainty and unstable regression models are a real and important concern (Chatfield, 1995) a model stabilising approach based on the bootstrap is considered in this chapter, originally developed in Campain et al. (2011). This method, the 'B-MI approach' is examined in a simulation study, and compared to other currently implemented methods in the context of missing data with imbalanced class distributions.

This chapter is presented in two parts. The first part explores missing data and the effects of imputation and the second part considers model instability and some of the issues associated with this challenge. Part one is made up of Sections 3.1 and 3.2. Section 3.1 considers missing data in detail, giving an overview of current solutions to the problem of missingness found within data sets. Section 3.2 is a comparison of multiple imputation methods, offering a unique framework for the comparison of multiple imputation algorithms. Part two consists of Sections 3.3 and 3.4. Section 3.3 considers some of the elements of unstable models and provides a novel stabilizing method for producing consistent regression models through the use of bootstraps. Section 3.4 is the development and comparison of the B-MI approach in a missing data context with class imbalance. Section 3.5 is a case study of the EPU data set, incorporating all the elements considered within the chapter and demonstrating their use and success in a clinical environment.

3.1 Missing data

Data sets, where some of the observations are missing, are common in statistics. However, most of the methods developed for statistical analysis require complete data sets. Imputation is the process of obtaining estimates for missing data within a data set. Various methods both ad hoc and statistically sophisticated exist for this procedure. In many non-statistical fields it is still common practice to delete all cases with missing data, due to the general feeling that exploring methods of imputation in data sets is 'making data up' (Stuart et al., 2009). However, over the last 30 years advances in imputation have been made and imputation is considered statistically far superior to complete case analysis.

There are three special data missing structures, missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). The structure behind these different cases and the effects they have on downstream analysis have been well studied in Little and Rubin (1987), Rubin (1987) and Schafer (1997). Methods for imputation are varied, a list of such resources can be found in Harel and Zhou (2007), Horton and Kleinman (2007) or Bramer et al. (1997). It is well known that both MCAR and MAR are not too much of a concern when analysing missing data. In the context

of the EPU data, MCAR would result from missingness being independent from both observable and unobserved factors,for example the loosing of data, and MAR would result from missingness not attributed to the data itself, for example an accidental missing of a question. However, MNAR patterns pertain to non-ignorable missingness. Such missingness occurs when the reason for missing data is attributed to an unobserved factor. For the EPU data, MNAR could occur when data is intentionally missed, for example the deliberate missing of a question. Knowing the correct structure for the missing data aids in the accuracy of the imputation procedure. Unfortunately it is not always possible to establish the missingness structure in a data set. Despite this, analysis can often proceed by using an imputation method that is based on an observed data likelihood and incorporating as much covariate information as possible. If this is performed, even if the missing structure is MNAR one can often impute using a MAR approach (Schafer, 1997).

3.1.1 Complete case

Complete case analysis is performed when only data samples with a complete set of observations are used in the analysis. Such a stringent criteria is often used in practice, but is generally not statistically sound. Only considering complete cases can be problematic in several ways. For example, the size of the data set is often greatly reduced in turn reducing statistical power, and exacerbating variable selection and model stability issues. Furthermore, there may be an underlying reason for the missing data and by removing all non-complete samples such structure could be eradicated from the analysis.

3.1.2 Single imputation

Performing the imputation process once, to obtain a completed data set is known as single imputation. Single imputation is a well known approach, Rubin [1987, p11.] comments:

> 'Single imputation, that is, filling in a value for each missing value, is probably the most common method for handling item non-response in current survey practice. There are two major attractive features supporting this practice. First, standard complete-data methods of analysis can be used on the filled-in data set. Second, [the effort of imputation] need be carried out only once by the data producer.'

Single imputation's main point of contention is that as each estimated data point is treated as an observed data point and therefore the variability of estimators, inherent due to imputation, is not taken into account.

3.1.3 Multiple imputation

Multiple imputation is repeatedly performing single imputation, say m times, only achievable for imputation processes with a stochastic component. For each missing data point multiple estimates are calculated for its value instead of one. This essentially means that m complete data sets are developed for further analysis. Hence, multiple imputation retains the advantage of single imputation, namely complete data sets for downstream analysis, but is superior to single imputation as the variability of the unknown values are taken into account. To employ multiple imputation each completed data set undergoes statistical analysis, independent from the other sets. Once the parameters of interest are obtained, for example regression coefficients, the results are then aggregated and their between and within imputation variabilities are combined. 'Rubin's rule' (Graham et al., 2007; Rubin, 1987) stipulate that an overall estimate can be calculated via averaging these estimates, $\bar{\alpha}_q = \frac{\sum_{r=1}^{m} \hat{\alpha}_{rq}}{m}$, where $\hat{\alpha}_{rq}$ is the point estimate for the rth imputation for the qth estimated parameter and $\bar{\alpha}_q$ is the average parameter estimate.

Let T_q be the combined within imputation variance and between imputation variance for the qth parameter in question, such that

$$T_q = \frac{\sum_{r=1}^{m} SE_{rq}^2}{m} + [1 + \frac{1}{m}] \frac{\sum_{r=1}^{m} (\hat{\alpha}_{rq} - \bar{\alpha}_q)^2}{(m-1)}, \tag{3.2}$$

the first component is the average of the squared standard errors (SE_{rq}) of the particular estimated parameter and the second component is the scaled variance over the m imputed data sets. It follows that the standard error for $\bar{\alpha}_q$ after multiple imputation is estimated by $SE_q = \sqrt{T_q}$.

In a model selection context, aggregating the m coefficient estimates for each variable can involve an inclusion frequency. An inclusion frequency is the measure of how prevalent a variable is in the numerous models constructed. The inclusion threshold stipulates how high the inclusion frequency must be for a variable to be considered in the final model (Austin and Tu, 2004; Heymans et al., 2007). For example, if a variable is present in seven out of 10 multiple imputations, for this variable to be included in the final model the inclusion threshold must be below 70%. The multiple imputation inclusion frequency is denoted as τ_{MI}.

Wood et al. (2008) considered a range of methods and conditions for combining parameter estimates from multiple imputation models. Very high as well as very low inclusion thresholds lead to inappropriate selection methods. One of the appropriate methods is produced when predictors are selected based on a frequency threshold (e.g. 50%). Let τ_{MI} be the inclusion threshold and $\hat{\rho}_q$ be the estimated inclusion frequency from the data as a percentage for the qth estimated parameter. The average of m multiple imputation estimates is

$$\bar{\alpha}_q = \frac{\sum_{r=1}^{m} \hat{\alpha}_{rq}}{m} \cdot I(\hat{\rho}_r \geq \tau_{\text{MI}}). \tag{3.3}$$

In this chapter, the inclusion threshold with respect to the multiple imputation models, τ_{MI}, will always be set at 50%. The selection of τ_{MI} is interesting in its own right, but the tuning of such a parameter is not addressed within this thesis. The value of 50% was considered to be a compromise between having an inclusion threshold that was too high, and hence an overly sparse model, or too low, and hence having an average coefficient estimate that incorporates a large number of zero estimates.

3.2 Multiple imputation algorithms comparison

Most imputation algorithms are designed to impute data regardless of the proportion of missingness evident in the data set. This section is a unique framework designed to compare different imputation methods to assess the validity of the methods and their appropriateness as the amount of missing data within a data set increases. The validity of methods is addressed based on models developed after imputation, regression coefficient estimation and downstream classification. Designed as a simulation study, for each imputation method two questions are addressed:

1. How the *amount* of missingness affects multiple imputation and the downstream results and;
2. If a critical proportion of missingness can be reached, such that beyond this point statistical interpretation needs to be questioned.

Missing data in classification has been studied in recent years. Grzymala-Busse and Hu (2001) as well as Farhangfar et al. (2008) have noted that on average imputation improves classification, compared to not imputing. Markey et al. (2006) considered how missing data in the test set impacts the classifiers performance when a complete data set was used to construct a classifier using neural networks.

In this study, resubstitution error rates are used to measure how much information is lost due to missing data and subsequent imputation as well as which imputation method is most appropriate in this context. Data is classified using both weighted logistic regression and random forests (Breiman, 2001).

This section continues by considering a range of different imputation methods in detail (Section 3.2.1), then the unique comparison framework is outlined (Section 3.2.2) and the evaluation criteria used to assess the different imputation approaches (Section 3.2.3). The simulation study and results are shown in Section 3.2.4.

3.2.1 Imputation algorithms

As different multiple imputation algorithms are tailored for different situations, the amount of missingness present in the data may affect these algorithms to varying degrees. Therefore three popular algorithms will be compared in this study, Amelia II (King et al., 2001), MICE (Buuren et al., 1999) and the Mi (Rubin, 1987) imputation method. These methods are used to explore the extent to which different imputation method(s) can reproduce adequate amounts of information. These three procedures have been considered because of their ease of use, availability and prevalence in literature (especially MICE). For example, all three methods are available as pre-written software applications, either as stand alone packages or packages in `R` (R Development Core Team, 2005). Imputations are drawn from some form of predictive distribution of the incomplete values. To do this one must model the predictive distribution for these imputations based on the observed data (Little and Rubin, 1987). However, at times this can be difficult, particularly when the total number of parameters, Q, increases (Dempster et al., 1977). Another reason for this selection of algorithms is that these different imputation methods all make use of alternative approaches and approximations to overcome the problems of constructing a predictive model and overcoming the arduous nature of this problem as the number of parameters increases.

Before describing the algorithms in more detail some notation is introduced. Let $\boldsymbol{X}$ be an $n \times Q$ matrix where the columns represent Q covariates and the rows n samples and $\boldsymbol{y}$ be a vector of response variables. Let $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{y})$ represent the data set in question, $\boldsymbol{X}_q$ represents the qth covariate ($q = 1, ..., Q$), with $\boldsymbol{X}_q^{\text{obs}}$ being the observed values and $\boldsymbol{X}_q^{\text{mis}}$ the unobserved or missing values of $\boldsymbol{X}_q$, similarly x_{jq} represents the jth sample ($j = 1, ..., n$) of the qth covariate, this data point is either observed or unobserved. The matrix $\boldsymbol{X}_{\text{nominal}}$ is the component of the $\boldsymbol{X}$ matrix that contains nominal variables and $\boldsymbol{X}_{\text{ordinal}}$ is the component of the $\boldsymbol{X}$ matrix that contains ordinal variables. In this thesis

the vector $\boldsymbol{y}$ is assumed to be completely observed and as the focus is on categorical variables, each level of category or group is termed a class.

Amelia II

Amelia II (King et al., 2001) imputes data by combining a bootstrap and an expectation maximisation (EM) approach under the assumptions that the missingness structure in the data is MAR and variables are jointly multivariate normal. Because of the jointly multivariate normal assumptions, Amelia II can be used for categorical analysis. Parameters required for data modelling include a vector of means for the Q covariates, $\boldsymbol{\mu}$, and a covariance matrix, for the Q covariates, $\boldsymbol{\Sigma}$. The algorithm draws m bootstrapped samples, the number of required multiple imputations, of size n', from the data set $\boldsymbol{Z}$. If not otherwise specified n' is chosen to be n. The EM algorithm is used to produce from the bootstrap estimates, $\hat{\boldsymbol{\mu}}_r^*$ and $\hat{\boldsymbol{\Sigma}}_r^*$ for $\boldsymbol{\mu}$ and a $\boldsymbol{\Sigma}$, $r = 1, ..., m$.

Consider the case where x_{jq} is missing and needs to be estimated. Let $\boldsymbol{X}_{j,-q}$ be the jth sample with the qth covariate removed. To impute the missing value Amelia II makes use of linear regression and the pair of bootstrap estimates $\hat{\boldsymbol{\mu}}_r^*$ and $\hat{\boldsymbol{\Sigma}}_r^*$ and uses these to calculate $\hat{\boldsymbol{\beta}}_r^*$. To estimate $\tilde{x}_{jq}^r$ the rth imputed value for x_{jq}, regress $\boldsymbol{Z}_{j,-q}^{\text{obs}}$ the observed data for the jth sample not including the qth covariate, but including the response variables on x_{jq} letting,

$$\tilde{x}_{j,q}^r = \boldsymbol{Z}_{j,-q}^{\text{obs}} \hat{\boldsymbol{\beta}}_r^*$$

at the point of data imputation. Variability is obtained for this estimate through the bootstrapping component of the imputation process and the differing $\hat{\beta}^*$ estimates. For this study, Amelia II is implemented from the contributed `R` package `Amelia` (Honaker et al., 2008), and the package's suggested parameters were applied for continuous and nominal variables, i.e.

```
amelia(x=X, m=5, noms=colnames(X_nominal), ords=colnames(X_ordinal))
```

is an example of the `R` code implemented.

MICE

Multiple imputation by chained equations (MICE) (Buuren et al., 1999) makes use of the Gibbs sampler, allowing for the generation of random variables from a marginal distribution directly without having to calculate the density (Casella and George, 1992). The Gibbs sampler is considered a multivariate extension to chained data augmentation

(Tanner, 1991) and hence is a series of chained equations. To begin, missing values are initialised in some way, for example by random draws from the observed marginal distribution. Then the Gibbs sampler is applied, with t as an iteration counter, and $\boldsymbol{X}_q^{\text{mis}}$ the missing values from the qth covariate:

$$\text{For imputations required for } \boldsymbol{X}_1^{\text{mis}} \text{ draw } \boldsymbol{X}_1^{t+1} \text{from } P(\boldsymbol{X}_1|\boldsymbol{X}_2^t, \boldsymbol{X}_3^t, ..., \boldsymbol{X}_Q^t),$$

$$\text{for imputations required for } \boldsymbol{X}_2^{mis} \text{ draw } \boldsymbol{X}_2^{t+1} \text{ from } P(\boldsymbol{X}_2|\boldsymbol{X}_1^{t+1}, \boldsymbol{X}_3^t, ..., \boldsymbol{X}_Q^t),$$

$$\vdots$$

$$\text{for imputations required for } \boldsymbol{X}_Q^{\text{mis}} \text{ draw } \boldsymbol{X}_Q^{t+1} \text{ from } P(\boldsymbol{X}_Q|\boldsymbol{X}_1^{t+1}, \boldsymbol{X}_2^{t+1}, ..., \boldsymbol{X}_{Q-1}^{t+1}),$$

so that all covariates have been imputed. Iteration stops when a convergence criterion is reached or a maximum number of iterations have passed. Care needs to be taken when using MICE, especially when data is binary or ordinal as chained equations have a tendency to produce separable data, that is a perfect prediction of an outcome by a predictive variable or combination of predictive variables (Tanner, 1991). This is particularly prevalent when there is a highly imbalanced class distribution in the outcome variable or when missingness is large (Su et al., 2011). Buuren et al. (1999) highlight that convergence can not be guaranteed except in special cases such as the multivariate normal and suggest to be careful when using MICE if missingness is large. For MICE the `R` package `mice` (Buuren and Oudshoorn, 2007) using the default parameters is used, i.e.

```
mice(data=X, m=5).
```

For more details on using MICE see Azur et al. (2011).

Mi

The Mi procedure makes extensive use of the predictive mean matching method developed by Rubin (1987). A variable of interest is imputed using other variables as predictors and the posterior mean is calculated given these predictors and the posterior predictive distribution. Imputation is performed by finding an observed value having the closest predictive mean. This observed value is used as the imputed value. The advantage of such a method is that only realistic values, or pre-observed values, are imputed. Mi overcomes the problem of separation, a potential problem in algorithms such as MICE, by transforming difficult to model variables. Examples of this are adding

dummy variables in the modelling of strictly positive values and the addition of noise (multiple types) for collinear values (Su et al., 2011). Mi is implemented through the `R` package `mi` (Gelman et al., 2009), again using the default parameters, i.e.

```
mi(object=X, n.imp=5).
```

3.2.2 Comparison methodology

Within this simulation study two separate approaches are used to compare multiple imputation methods. Imputation methods are considered powerful if they can reconstruct the information lost due to missing data, and if this reconstruction can be performed even when the proportion of missingness is high. The percentage of missingness relates to the proportion of missing covariate values.

Weighted logistic regression is used extensively in the first method. Through bootstrapping, the distributions of the logistic regression coefficients are examined. How these estimated distributions change as the percentage of missing data increases is compared across the multiple imputation methods. The bootstrapped mean squared error from these bootstrapped coefficients is also used to aid comparison of multiple imputation methods, where consistently low values indicate a resistance to change because of an increase in missingness.

The second comparison method considers prediction accuracy in a classification framework. Data is classified using two classification methods, weighted logistic regression and random forests (Breiman, 2001). The consistency of the prediction accuracy across multiple imputation methods as the amount of missingness increases is indicative of an imputation method's ability to reconstruct discriminant information for partially complete data sets.

To examine how the amount of missingness affects the imputation results, an iterated bootstrapped logistic regression procedure is employed. Figure 3.1 contains a graphical representation of the process, which consists of the following six step algorithm:

Algorithm 1: Coefficient distribution from logistic regression

1. Obtain a data set with no missing data.
2. Draw stratified bootstrap samples. For example, the bootstrap samples could be stratified to preserve the original class distribution, with $n' > n$, where n' is the size of the stratified sample. Under sampling is also possible depending on the data set in question.

3. Using the complete bootstrapped data set from step 2, create several data sets, for example 10. The first such data set is the complete data set and subsequent data sets contain an increasing amount of induced missingness (this could be a series of 5%, 10%,...,45% missingness). Missingness is induced MCAR with the response variable left complete. In this study nine data sets have been used.

4. Impute each of the incomplete data sets m times.

5. For each level of missingness, apply a weighted logistic regression analysis to the data sets, resulting in coefficient estimates for each of the covariates.

6. Repeat steps 2.–5. B times, to obtain bootstrapped distributions for the coefficient estimates for each covariate, as the amount of missingness increases.

Note: Bootstrap sampling is included in the process firstly to increase the number of samples in the datasets and secondly to increase the amount of randomness in the process. By bootstrapping hundreds of times the estimates of the distributions of the coefficients can be observed. These distributions can then be compared to the distributions when there is no missingness in the data set. Hence, the ability of the imputation method to regain information lost due to missingness can be explored.

3.2.3 Evaluation criteria

The bootstrapped distributions for the data sets containing missingness can be compared against the bootstrapped distribution of the complete data. Hence, the effect of the amount of missingness and how well particular imputation methods regenerate the original information can be examined. Here, two selection criteria have been considered (i) bootstrapped mean squared error; and (ii) prediction accuracy.

Bootstrapped mean squared error

To compare the distributions of the coefficient estimates for the data sets with induced missingness against the estimated coefficient distributions for the complete data the bootstrapped mean squared error (B-MSE) is calculated. Let $\hat{\beta}^b_{q,0}$ be the coefficient estimate for the complete data set for the bth bootstrapped sample, where $b = 1, ..., B$, for the qth variable where $q = 1, ..., Q$ and $\hat{\beta}^b_{q,k}$ be the coefficient estimate for the bth bootstrapped sample with k% missing, where $k = 5,\ 10,\ 15, ...,\ 45$, then the B-MSE is

$$\text{B-MSE}_{q,k} = \frac{\Sigma_{b=1}^{B}(\hat{\beta}^b_{q,0} - \hat{\beta}^b_{q,k})^2}{B}.$$

The B-MSE is obtained for all variables in the data set and all available imputation algorithms, here Amelia II, MICE and Mi.

Prediction accuracy

To examine how the amount of missing data affects data classification, the above data, induced with a range of proportions of missingness, is classified using random forests (Breiman, 2001) and logistic regression. Classification is considered under two paradigms. The first is to consider classification in a traditional sense using an imputed data set to construct and test a classification rule. The second is to evaluate how information is lost through imputation. To achieve this, consideration is taken to how error rates increase when classifying data that has been imputed when the classification rule was constructed on the complete data set.

These two aims are performed in separate stages:

1. To observe how missingness affects classification when the classifier is constructed and tested on imputed data, repetitively impute data sets for a given level of missingness and classify using v-fold cross validation. (i) For random forests the final class is the majority vote of the m classifiers, (ii) for logistic regression the m models are combined prior to evaluation.

2. To examine how information is lost through the presence of missing data, the error rates are considered when a classifier is built on the complete data set and tested on an imputed data set (with differing levels of missingness). The complete data set is used to construct a classification rule. The resubstitution error rate is obtained as the baseline. Using the *same* classification rule, imputed data sets are also classified, obtaining a modified resubstitution error rate. Comparing these error rates to the baseline resubstitution error rate, the decrease in accuracy can be attributed to the missing data and subsequent imputation.

The first stage in the classification process is typical classification of multiply imputed data sets with v-fold cross validation. Although this stage can be informative as to if the induced missingness affects classification, it struggles to highlight how the amount of missingness and imputation affects classification. The reason for this is that the same level of missingness is being used to build and to test the classifier. If it is assumed that in a data set with $k\%$ missingness there is a loss of information, after imputation this loss of information is still evident throughout the whole data set. Constructing and testing a classifier on this data set, even using cross validation to reduce bias, does not allow the evaluation of how much information is lost due to the missing data.

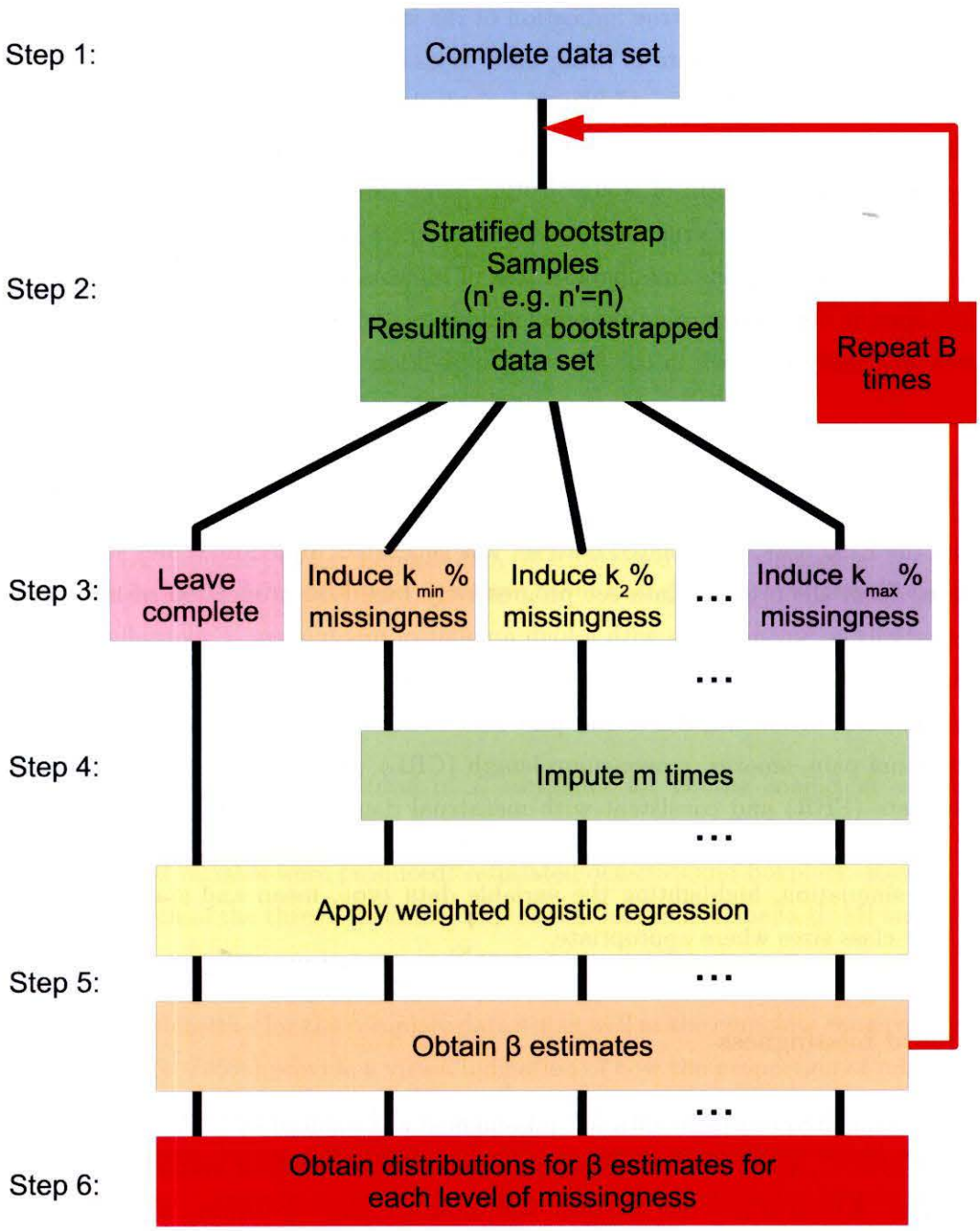


FIGURE 3.1: Methodology for the comparison of coefficient distributions as the amount of missingness increases.

To examine how classification accuracy is affected by the loss of information from the amount of missing data and imputation, the second stage of the classification method is performed. This stage makes use of resubstitution error rates. Although in most classification contexts the resubstitution error rate is known to be biased downward and can not be used to gain a true indication of the independent error rate, in this context the resubstitution error rate is being used to examine how information is lost due to missingness within a classifier. A classification rule is constructed based on the complete data set. Subsequently data sets with a range of imputed amounts of missingness are classified (creating a form of resubstitution error rate). An increase in this error rate when compared to the original resubstitution error rate obtained from the complete data set is caused by the missingness, loss of information, and subsequent imputation. More specific details regarding the cross validation and resubstitution error rates for this particular study is shown in the following subsection.

3.2.4 Simulation study

From the EPU data, a simulated data set was generated. Individuals and variables were deleted from the original data set progressively based on proportion of missing data, sequentially removing items with a high amount of missingness. Eventually obtaining a complete data set with 270 cases (coincidently with 8% miscarriages) and 12 variables (age, number of natural deliveries, past-miscarriages, gestational age, bleeding, clots, abdominal pain, smoker, crown-rump length (CRL), gestational sac (GS) mean, foetal heart rate (FHR) and consistent with menstrual dates), five of which are common to Riemke's final model (Table 2.3). Table 3.1 shows a summary of the data, which is used in this simulation, highlighting the variable data type, mean and standard deviation (SD) or class sizes where appropriate.

Induced missingness

For the simulated complete data set, missingness was induced by MCAR at proportions of 5%, 10%, ..., 45%, with the response variable (miscarriage or viable pregnancy) left complete. A total of $B = 999$ stratified bootstrap samples were drawn, with stratification used to preserve the original class distribution. Data sets were multiply imputed five times and a 5-fold cross validation was used for classification. Weights used in the logistic regression were chosen to ensure an even class distribution (Section 3.3.1).

Variable	Data Type	Mean	SD	Class Sizes
Outcome	binary	-	-	248 v 22
Age	integer	27.69	5.14	-
Number of natural deliveries	integer	0.71	0.98	-
Past-miscarriages	integer	0.61	1.03	-
Gestational age	integer	59.73	13.98	-
Bleeding	binary	-	-	159(yes) v 111(no)
Clots	binary	-	-	31(yes) v 239(no)
Abdominal pain	binary	-	-	157(yes) v 113(no)
Smoker	binary	-	-	50(yes) v 220(no)
CRL	continuous	19.56	18.09	-
GS mean	continuous	31.82	18.43	-
FHR	integer	146.75	26.19	-
Consistent with menstrual dates	binary	-	-	206(yes) v 64(no)

TABLE 3.1: Summary of the complete data ($n = 270$) used in this simulation, highlighting the variable data type, mean and SD or class sizes where appropriate.

Coefficient distribution

After running 'Algorithm 1' a total of B estimates for $\boldsymbol{\beta}$, the coefficient vector for the covariates, were obtained. From these results two types of plots for each level of missingness and variable were produced: estimated densities and boxplots. Results were obtained for each of the three considered imputation methods, Amelia II, Mi and MICE. Examples of these plots can be seen in Figures 3.2 and 3.3.

The estimated densities for the complete data set as well as the complete range of missing proportions (5% – 45%) provide a visual indication of how the proportion of missingness affects the distributions. Here, the density of the complete data set is considered as the 'gold standard' because data imputed with no loss of information would reproduce this distribution. Closeness of the other bootstrapped distributions to the gold standard implies that the imputation method was able to re-capture, to some degree, the lost information for that particular proportion of missingness.

Boxplots of the log-ratios of the imputed data set coefficients over the complete data set coefficients are included. In these plots, estimates are paired by bootstrap draw. A log-ratio greater than zero implies that the coefficient with induced missingness is greater than the gold standard. An increased width in the log-ratio boxplots is indicative of

Past-miscarriages	Amelia II		Mi		MICE	
Amount Missing	Med	Var	Med	Var	Med	Var
Complete	0.06	0.13	0.06	0.13	0.06	0.13
5%	0.04	0.12	0.04	0.14	0.06	0.15
10%	0.02	0.11	0.04	0.16	0.06	0.16
15%	0.02	0.10	0.04	0.18	0.05	0.19
20%	0.00	0.10	0.03	0.19	0.04	0.19
25%	0.00	0.10	0.05	0.20	0.04	0.20
30%	0.01	0.09	0.03	0.30	0.04	0.27
35%	0.00	0.09	-0.02	0.35	0.10	0.25
40%	0.00	0.07	0.03	0.31	0.06	0.26
45%	0.00	0.07	0.00	0.45	0.06	0.31

TABLE 3.2: Summary of the median and variance for the bootstrap distributions for the coefficient 'past-miscarriages' as the amount of induced missingness increases. Data has been imputed using Amelia II, Mi and MICE.

the increase in variability of the ratios, and as the median shifts away from zero, this suggests that bias occurs.

The summary of the bootstrapped coefficients (Table 3.2 and 3.3) and plots are provided for only two variables, 'past-miscarriages' and 'clots'. Results for the other ten variables are shown in Appendix B Section B.1 and are omitted here because results for variables with large coefficients, relative to their standard errors, are similar to 'clots' and results with coefficients close to zero are similar to 'past-miscarriages'.

Figure 3.2 contains the bootstrapped distributions and boxplots for 'past-miscarriages', a nominal variable that often results in small regression coefficients for the bootstrapped samples. This variable was not in the final model of Riemke et al. (2011) and is representative for redundant variables. The results displayed in Figure 3.2 are typical of all the small effect variables studied. Amelia II produces a series of reliable results. Here reliable is that the medians are close to the original complete data set median, and the variance of the distributions are close to that of the original distribution, even as the amount of missingness increases to 45%. The medians do exhibit some small shrinking toward zero, together with a decreasing variability (Table 3.2). As the amount of missingness increases, Mi and MICE increase the distributional variability for the coefficient 'past-miscarriages' (Table 3.2). Both these imputation methods exhibit a minor

Clots	Amelia II		Mi		MICE	
Amount Missing	Med	Var	Med	Var	Med	Var
Complete	-2.36	1.00	-2.36	1.00	-2.36	1.00
5%	-2.14	0.87	-2.28	1.02	-2.37	1.11
10%	-1.99	0.74	-2.27	1.42	-2.35	1.29
15%	-1.82	0.69	-2.19	1.10	-2.41	1.58
20%	-1.63	0.92	-2.14	1.18	-2.45	1.60
25%	-1.49	0.53	-2.06	1.30	-2.43	1.81
30%	-1.37	0.46	-2.06	2.28	-2.39	2.41
35%	-1.20	0.48	-1.97	2.70	-2.43	2.39
40%	-1.06	0.45	-2.06	2.17	-2.35	2.20
45%	-0.93	0.73	-2.11	4.10	-2.32	2.76

TABLE 3.3: Summary of the median and variance for the bootstrap distributions for the coefficient 'clots' as the amount of induced missingness increases. Data has been imputed using Amelia II, Mi and MICE.

bias trend as the distributions tend to become more symmetrical as the variance inflates. The Mi imputation appears to inflate the median to a greater extent than MICE. These results are made more evident by a visual inspection of the boxplots, showing the log-ratio of imputed coefficient estimates to complete coefficient estimates, paired by bootstrap draw.

Figure 3.3 depicts results from the 'clots' variable which frequently produced the largest coefficients in the bootstrapped analysis of this data set and is also the variable with the largest coefficient in Riemke's final model. 'Clots' is a binary variable and results are typical for all large coefficient variables within this data set including those that are continuous or categorical. The shrinkage bias exhibited by Amelia II toward zero is very pronounced for this variable, as the median for these distributions is large, although variability decreases from the original distribution with the introduction of missingness (Table 3.3). For the Mi imputation method the variability increases as the amount of missingness increases, this variability tends to drastically increase as the amount of missingness reaches about 35–45%. The inflation bias of the median is still evident for the Mi imputation method however for large coefficients this bias appears to not be present when MICE is used.

For the two variables the B-MSE has been plotted for each imputation method (Figure

3.4). A small B-MSE implies that the two distributions (the complete and the incomplete) are similar. The B-MSE can be used to examine the relationship between bias and variability. Figure 3.4(a) illustrates the B-MSE when coefficients are small. From Figure 3.2 it is clear Amelia II induces only a small amount of shrinkage, as a result this imputation method produces a competitive B-MSE particularly for large proportions of missingness, Both Mi and MICE have an increasing variance regarding the proportion of missingness and thus for percentages of missingness greater than 15%; Amelia II is better than MICE, which in comparison is always slightly better than Mi. Figure 3.4(b) encapsulates the B-MSE behaviour when the coefficients are large. Amelia II suffers from an increase in bias and this contributes to a large B-MSE, this induced bias is greater than the increase in variability produced by Mi and MICE, implying that Amelia II performs poorly, comparatively. MICE and Mi are comparable in this case with MICE often producing a smaller B-MSE than Mi. Based on this study, taking into consideration bias tendencies and B-MSE, overall MICE appears to be the most competitive imputation method, albeit Amelia II can be very successful when missingness is extreme and variables are redundant.

Prediction accuracy

Classification error rates were obtained after applying the two stages of classification described in Section 3.2.3. The first stage produced a series of error rates obtained for a 5-fold cross validation and the second stage used resubstitution-type errors to observe the loss of information due to missing data and imputation. The classification error rates for both the random forests and the logistic regression models are plotted in Figures 3.5 and 3.6 for the two stages of classification, respectively. As the amount of missingness increases the classification error rate is expected to increase if the imputation methods are losing information critical to classification. Also provided in Figure 3.6 are the split error rates relating to the classification accuracy of the two individual classes, viable and miscarriage. These split error rates are obtained for the three imputation methods and provide insight into how classification deteriorates as missingness increases.

Stage 1: V-fold classification Figure 3.5 displays the 5-fold cross validation classification error rates for both the random forest classifier and the logistic regression models as the amount of missingness increases. Imputed data was used to train and test the classifier. The classification error rate was inflated when missingness was present compared with the classification error rate of the complete data set. However, once missingness is introduced into the data set, the inflated error rates appear near constant despite the increase in the amount of missingness. It is possible that the reason for this

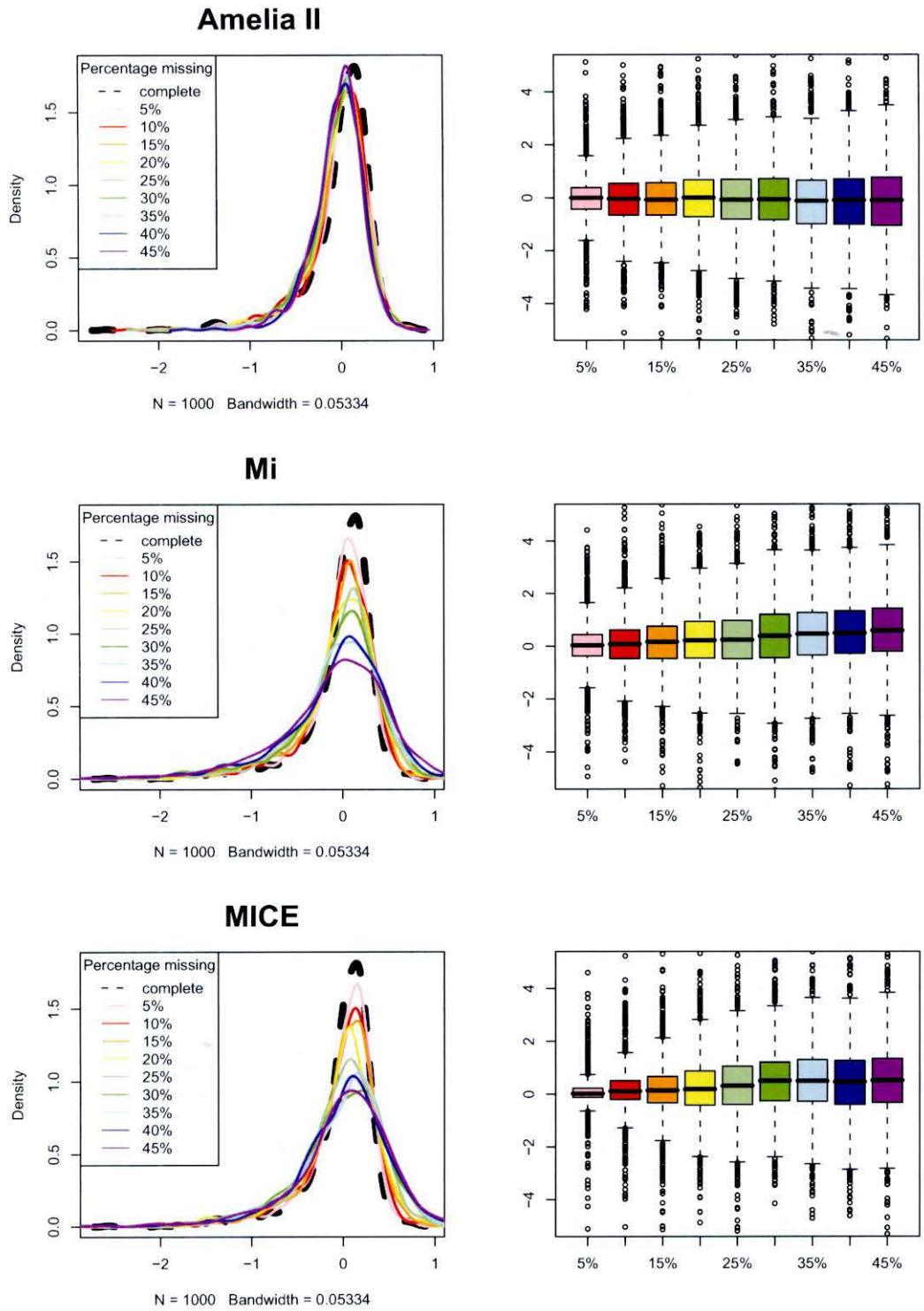


FIGURE 3.2: **Redundant variables:** Bootstrapped distributions and boxplots for estimated coefficients for the 'past-miscarriages' variable as the amount of missingness increases. Boxplots depict the log-ratios for imputed data set coefficients to complete data set coefficients, paired by bootstrap draw. Imputation methods used include Amelia II, Mi and MICE.

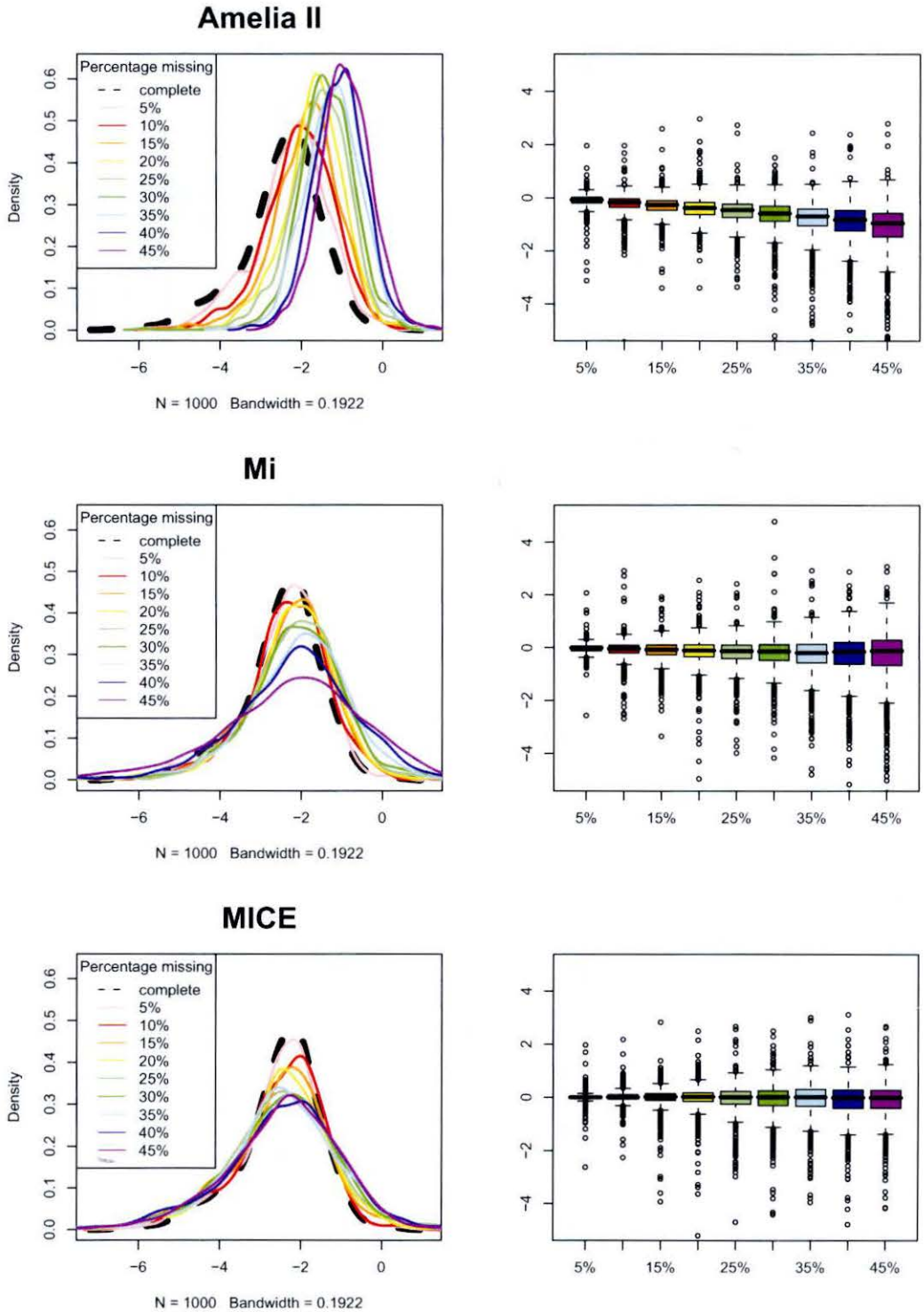


FIGURE 3.3: **Important variables:** Bootstrapped distributions and boxplots for estimated coefficients for the 'clots' variable as the amount of missingness increases. Boxplots depict the log-ratios for imputed data set coefficients to complete data set coefficients, paired by bootstrap draw. Imputation methods used include Amelia II, Mi and MICE.

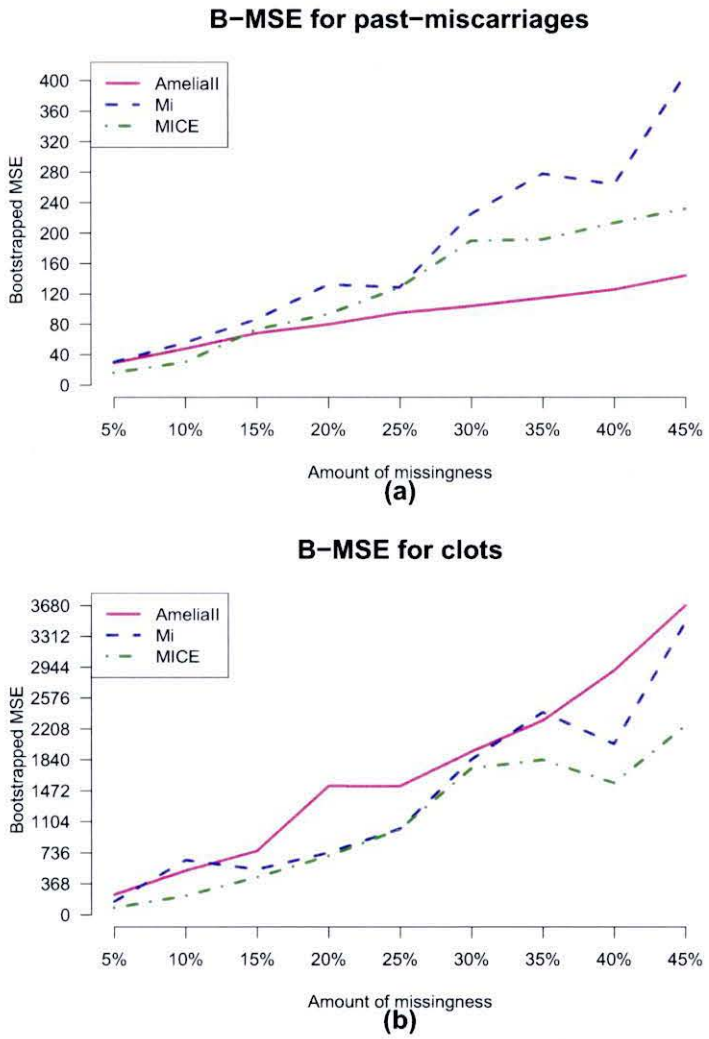


FIGURE 3.4: B-MSE for Amelia II, Mi and MICE relating to the estimated coefficients for the variable (a) 'past-miscarriages' and (b) 'clots'. A smaller B-MSE implies the bootstrapped distribution for the imputed data was similar to the bootstrapped distribution for the complete data.

stems from the methodology (Section 3.3.2). Here the imputed data is used to build **and** test the classifier. Some information is lost through the process of data deletion and imputation, hence the classifier is not built as accurately and it follows that the classification error rates are inflated. Interestingly no imputation method appears to stand out when imputed data is used during the construction of a classifier. These results agree with the findings of Grzymala-Busse and Hu (2001) as well as Farhangfar et al. (2008). The high error rates, especially for the weighted logistic regression, may relate to the data set containing a highly imbalanced class distribution or also due to logistic regression not being flexible enough. For these reasons it is possible that the classifiers

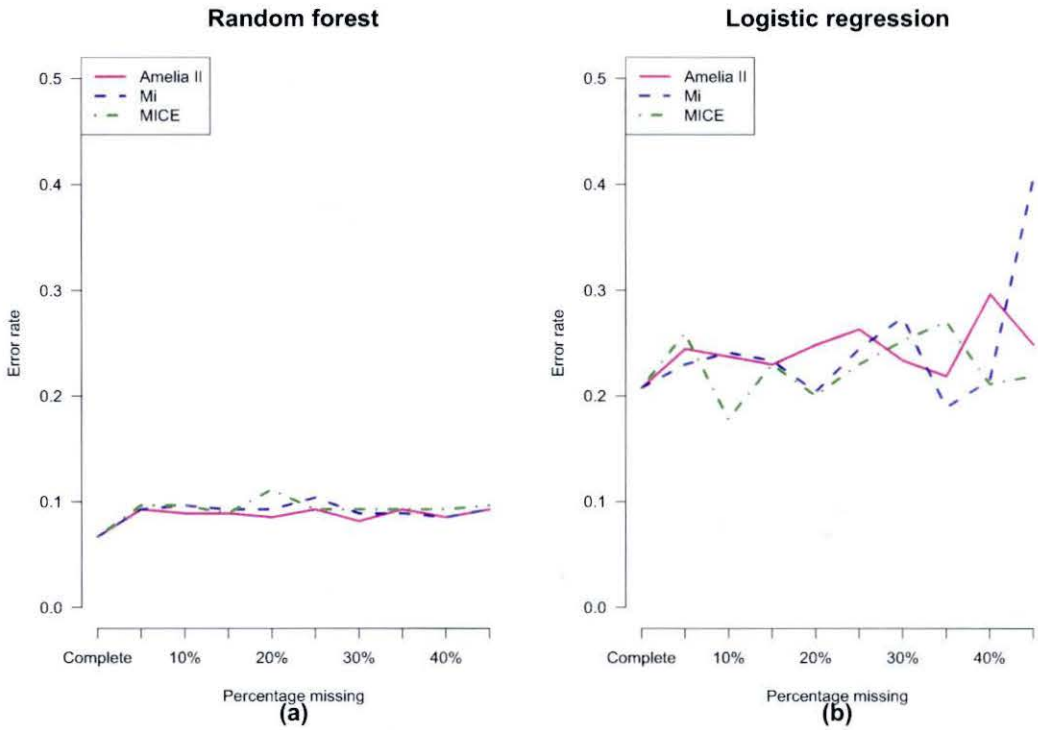


FIGURE 3.5: Classification error rates for the (a) random forest classifier and (b) logistic regression in dependence of missingness, using $m = 5$ imputations and 5-fold CV.

are struggling to correctly separate the data set (complete or imputed). Hence, classification is poor regardless of induced missingness. It is important to note that these error rates show that, for this data set, any amount of missingness yields error rates greater than that of the complete data set (Figure 3.5). It is evident that missingness affects the ability to classify data as the distinction between the two classes is reduced when MCAR missingness with subsequent imputation is present.

Stage 2: Loss of information through missingness Figure 3.6 (a) and (b) contain the second stage of classification where resubstitution type errors are used to examine information loss for both the random forests and logistic regression classifiers. To evaluate how the amount of missingness is affecting the classification of the data set, a careful evaluation of the resubstitution error rates was applied. The random forest classifier and the logistic regression classifier have markedly different error rates. Error rates for the random forests after an initial spike as soon as missingness is introduced hover around 10%, logistic regression produces error rates around 20%. Regarding random forests, considering Figure 3.6 (c), (e) and (g) it becomes clear that as the amount of missingness increases there is a rapid increase in the amount of samples classified as viable pregnancies (the majority class). Hence, the apparent low error rates, and the drop in

error rates cannot be attributed to the accuracy of the classifier, but instead to the lack of sensitivity in the classifier. Error rates are heading toward 8%, the error rate for the naïve classifier, where all samples are simply classified as the majority class. This loss of sensitivity in the classifier is caused by the loss of information due to missingness.

With the increase in missingness the weighted logistic regression classifier's ability to accurately distinguish between classes reduced slightly. Although this classifier has a larger overall prediction error rate than random forests it does not suffer from the majority class bias (that is classifying all samples to the majority class). The inclusion of weights in the logistic regression classification compensates for the minority class, and hence typically avoids the loss of sensitivity which occurred when using the random forest classifier (Figure 3.6 (d), (f) and (h)). With the deterioration of both classifier's ability to discriminate between classes as the amount of missingness increases, there is no strong indication for the more reliable classifier when using imputed data. Schmid et al. (2001) compared neural networks, logistic regression and classification trees and found them almost equally robust for moderately sized data sets ($n = 500$).

Figure 3.5 and 3.6 highlight that there is a large discrepancy in performance between logistic regression and random forest classification. This interesting difference exists regardless of testing type (cross validation or resubstitution) and the proportion of missing data. Comparisons of classification methods goes beyond the scope of this thesis.

Other settings and data

The results of this simulation study are generalisable to data sets with more observations than covariates. For real data, if time permits, it is recommend to apply 'Algorithm 1' to a simulated data set having a similar structure (amount of missingness, moments, correlations, etc) as the real data set. This allows the estimation of the amount of shrinkage towards zero and the increase in variability of the parameter estimates, both due to the amount of missingness and the imputation procedure used.

3.2.5 Multiple imputation algorithm comparison conclusion

This study provides a framework for assessing different imputation methods and examines how the amount of missing data affects these multiple imputation approaches. Comparison of methods has been achieved through examining how the multiple imputation methods and the amount of missing data alters logistic regression coefficients as well as how these elements impact upon prediction accuracy using logistic regression and random forests. This novel framework was used to compare Amelia II, Mi and MICE but

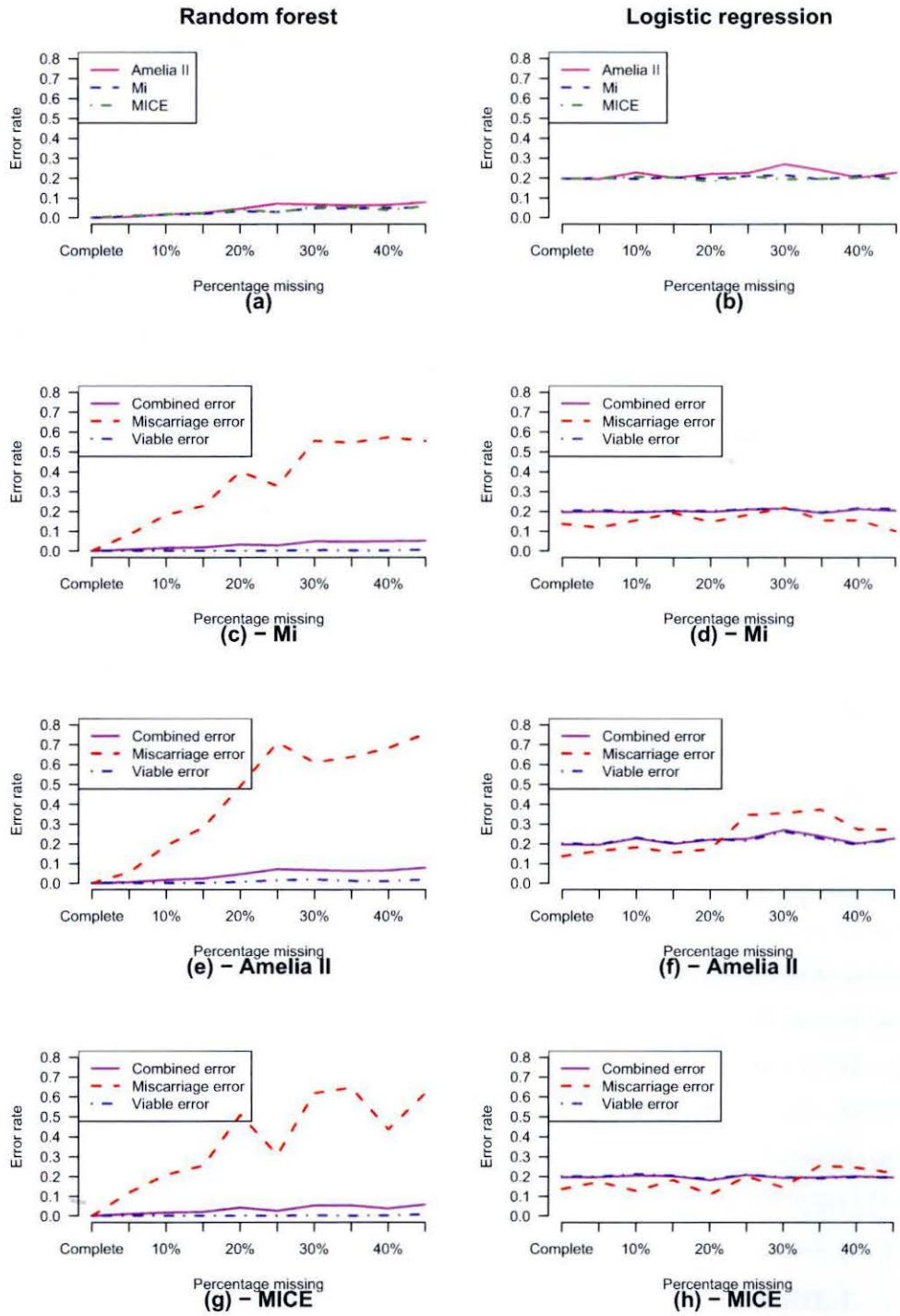


FIGURE 3.6: Resubstitution error rates for (a) random forests and (b) logistic regression, as the amount of missingness increases. Average error rates across all $m = 5$ multiple imputations are plotted, (c,e,g) resubstitution split error rates for Mi, Amelia II and MICE from random forest classification, (d,f,h) resubstitution split error rates for Mi, Amelia II and MICE from logistic regression classification. 'Combined error' is the overall error rate, 'Miscarriage error' is the error associated with the miscarriage samples and 'Viable error' is the error associated with the viable samples.

its application is not limited to these imputation methods, and can easily be extended to compare any other current method or multiple imputation method developed in the future.

Inducing missingness ranging from 5% to 45%, in data from the EPU study gives insight into how the proportion of missing data affects the statistical interpretation of the analysis performed. It was found that, increasing the amount of missingness (i) alters the coefficient distributions from the original coefficients' distributions developed under the complete data set such that depending on the imputation method used this may increase the variably or shrink the mean toward zero, and (ii) decreases prediction accuracy under downstream analysis. The downstream effect is very extreme when missingness is large (40–45%) but is manageable through multiple imputation. It has been shown that conclusions from data sets with excessive amounts of missingness can be misleading, particularly when imputation algorithms produce biased estimates, and therefore more care is needed in such situations.

Different imputation methods have different effects on the data. For Amelia II, there is evidence of shrinkage of the mean toward zero and this shrinkage increases proportionally to the amount of missing data. This is in line with the conclusions in Abrahantes et al. (2011). For the imputation methods Mi and MICE there is a tendency for variability to increase, however, the median produced by these imputation methods is consistent with that of the complete data set. When coefficients are large and notably most likely to be retained in the final model, the increase in variability is less pronounced especially when considering the MICE algorithm. Under the B-MSE criterion, MICE maintains a closer fit to the original distribution as missingness increases compared to the other imputation methods. Within the classification paradigm, all imputation methods applied to the data show degraded prediction accuracy in a similar way.

In summary, it has been shown empirically that different imputation methods affect differently the downstream analysis. Based on various criteria the recommendation is to use MICE over Amelia II and Mi. However, it should be noted that these investigations are based on the EPU study, and hence results are to some extent a function of that data. Readers are encouraged to apply the investigation to their own data if time permits using a range of available imputation methods.

3.3 Instability in model selection - logistic regression

Unstable models are considered to have occured when a small perturbation in the data set produces large changes in the final model constructed. This can be caused by a

number of reasons:

1. The feature selection method, especially if a computational method is employed without statistical thought (Murtaugh, 1998) and (Austin and Tu, 2004);
2. Highly correlated variables within the data set (Cessie and Houwelingen, 1992);
3. When parameter estimates are unstable;
4. Imbalanced class distributions.

This section will mainly focus on the last two points. Unstable models are more probable in the case of small samples or a heavily imbalanced class distribution. Detection of unstable models is important as non-reproducible models struggle to accurately predict independent data. To examine how predictive models handle future outcomes, the stability of a model can be examined by random splits in the data, producing training and validation sets (Beyene et al., 2009). There are several contributing factors affecting the occurrence of unstable models. These include variable correlation and multicollinearity within a model as well as high proportions of class imbalance.

It is important to detect an unstable model throughout the analysis process. To highlight an unstable model one may hold out a proportion of data, and construct a model with the remaining data. By observing the retained models after a number of trials, one can examine if the models being constructed are similar enough to be considered stable. Moreover, the held out portion of the data can also be used to evaluate the model, and thus obtain an estimation as to the predictive capabilities of the model on independent data. This is graphically represented in Figure 3.7, which illustrates a loop that could be employed.

Presently the theoretical justification of the Akaike or Bayesian information criterion (AIC and BIC, respectively) is based on the assumption that a complete data set is available. Logistic regression model selection becomes more difficult in the presence of missing covariates. Only recently first results appeared. Yang et al. (2005a) addressed model selection in presence of missing data in a Bayesian framework, Schomaker et al. (2007) examined different model selection criteria including selection after imputation and down-weighting incomplete cases, and Consentino and Claeskens (2011) studied alterations to the original AIC. The model selection method in this section makes use of the BIC together with the bootstrap and is performed after imputation. This approach is developed to address simultaneously the difficulties involved with missing covariates and model instability.

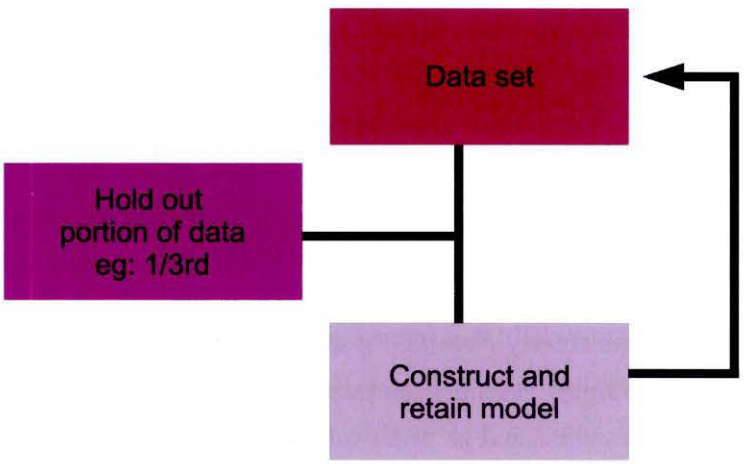


FIGURE 3.7: Graphical representation of a method to establish if a model is stable. By repeating this process multiple times and comparing the retained models one can make an informed decision if stability measures are required in downstream analysis.

3.3.1 Class imbalance distributions

An imbalance in class distribution is said to occur when one class is highly represented compared to the other class(es) present. High class imbalances occur in many prediction settings and frequently a high level of prediction accuracy is required for the rare but important events. A famous example of class imbalance is a study involving oil spills (Matwin et al., 1998), but this statistical issue is especially poignant in medical studies. For example, rare disease diagnosis or as here miscarriages exhibit class imbalance. Class imbalance not only affects the accuracy of the classifier but also the model construction and performance evaluation methods. Class imbalance becomes increasingly challenging when the data contains other categorical covariates. For example, in a regression context one must ensure that there are sufficient events in each cell of the contingency table so that the data is not considered overly sparse. Sparseness is more common when classifications have several categories or class imbalance is acute (Agresti, 2007). Without sufficient observations statistical interpretation becomes questionable with large standard errors and increased potential to overfit the data. Therefore, the difficulty increases when missingness is highly prevalent. Two elements to address the problem of class imbalance are presented: (i) change the performance measure; and (ii) alter the class distribution for increased prediction accuracy.

Change performance measures

The confusion matrix is used to define a performance criterion, for known and predicted 'events'/'non-events' (Table 3.4). The results are tabulated such that a is the number of true positives (where an event is correctly considered positive) and d is the number of

	Event	Non-event
Predicted event	a	b
Predicted non-event	c	d

TABLE 3.4: An example confusion matrix for prediction.

true negative samples (correctly classified non-event samples), b is the number of false positives (non-event samples classified as event samples) and c is the number of false negatives (event samples classified as non-event samples).

In standard prediction paradigms, performance is a measure of accuracy such that

$$\text{accuracy} = \frac{a+d}{n}, \tag{3.4}$$

where $n = a+b+c+d$. However, when there is a large imbalance in class distribution, poor classification performance, that may be completely misclassifying one group, may incorrectly suggest a high level of performance accuracy. For example if a rare disease is present in 1% of a particular sample, complete misclassification will yield a classifier with 99% accuracy.

The receiver operating characteristic (ROC) curve was developed in the 1960's to summarise data from signal detection experiments in psychophysics (Hanley, 2005). It visually represents the relationship in the trade off between true positive and false positive classification. At development the area under the ROC curve (AUC) was suggested as a measure of performance accuracy. Provost et al. (1998) continued with the benefits of the ROC curve suggesting that it is a good indicator of a classifier's performance in a wide range of prediction settings.

Altering class distribution for increased prediction accuracy

There are many ways of altering the class distribution of a data set. Examples include both under and over sampling. Under sampling involves reducing the samples present in the majority class so that a desired distribution between majority and minority classes is reached. Over sampling draws minority class samples with replacement to artificially inflate the number of samples within this class (Breiman et al., 1984). Both approaches have drawbacks, under sampling potentially discards information and over sampling due to repeating identical samples may lead to over-fitting the classifier in question. A more comprehensive review and more complex sampling procedures have been developed by Chan and Stolfo (1988) and Chawla et al. (2002).

Weiss and Provost (2001) analysed the effect of class distribution on learning in general. In an empirical study of 24 data sets they varied the class distributions of data sets undergoing classification and observed how these error rates were altered as the distributions changed. Weiss and Provost suggested that the naturally occuring class distribution is not often optimal in producing the best performing classifier and that a more appropriate distribution is that of a 50% minority class balance.

Instead of under or over sampling, apparent class distributions can be obtained by weighting the observations and applying a weighted least squares or maximum likelihood approach. The weight vector $\boldsymbol{w}$ can be obtained in a range of ways. For example an even class distribution yields a weight vector with w_1 and w_2 pertaining to samples from the appropriate classes with

$$w_1 = \frac{n/2}{n_1} \qquad \text{and} \qquad w_2 = \frac{n/2}{n_2}, \tag{3.5}$$

where $n = n_1 + n_2$ and n_i is the number of samples in class i.

3.3.2 Stabilising methodology: The B-MI approach

Bootstrap samples and **m**ultiple **i**mputation are combined in this method, and it is therefore called the *B-MI approach* (Campain et al., 2011). This is a novel method that produces a stable logistic regression model with good predictive properties. The method has been developed under the assumption that there are at most a manageable proportion of missing data in $\boldsymbol{X}$. An overall proportion of missing data less than 30% is considered manageable by Rubin (1996) and this was confirmed in the simulation study in Section 3.2. A graphical representation of the B-MI approach is shown in Figure 3.8 and in the following each of the six stages is detailed.

Stage 1: Multiple imputation The data set undergoes m multiple imputations, producing $\mathbf{Z}^1, ..., \mathbf{Z}^m$ complete data sets.

Stage 2: Bootstrapping the $\mathbf{Z}_r$'s The jth paired bootstrap sample is drawn, $\boldsymbol{b}_j^*$ denotes the sample's index vector which is fixed over $r = 1, ..., m$. Let $\mathbf{Z}_{\boldsymbol{b}_j^*}^r$ denote the bootstrapped imputed data for the jth bootstrap sample. Note that the m imputed and bootstrapped data sets are paired, that is they share the same observed values but have different imputed values. Stratified bootstrap samples are drawn, which were shown to be crucial in preserving robustness qualities in the model building process (Müller and Welsh, 2005, 2009). Here, stratification was such that a consistent class distribution was

ensured and the proportion between originally complete samples and samples completed via imputation was maintained.

Stage 3: Model fitting and selection A weighted logistic regression is fit to each of the m data sets (which have been imputed and bootstrapped) which is reduced by an automated variable selection procedure. A BIC like criterion is used with penalty multiplier of $\frac{1}{2}\ln(n)$ instead of $\ln(n)$, but other criteria can be used instead such as the AIC or BIC. This gives, for the m data sets, estimated bootstrap parameters $\hat{\boldsymbol{\beta}}^r_{\boldsymbol{b}^*_j} = (\hat{\beta}^r_{0,\boldsymbol{b}^*_j}, ..., \hat{\beta}^r_{Q,\boldsymbol{b}^*_j})$ where those components not retained by the variable selection procedure have zero estimates.

Stage 4: Aggregation over MI Each variable q (where $q = 0, ..., Q$) is retained or not in each of the m models. The bootstrap inclusion frequency, $\hat{\rho}_{q,\boldsymbol{b}^*_j}$, is the proportion of times variable q is retained for the jth bootstrap sample i.e

$$\hat{\rho}_{q,\boldsymbol{b}^*_j} = \frac{1}{m}\Sigma^m_{r=1} 1\{\hat{\beta}^r_{q,\boldsymbol{b}^*_j} \neq 0\}. \tag{3.6}$$

To aggregate across the m models, only coefficients that satisfy $1 \geq \hat{\rho}_{q,\boldsymbol{b}^*_j} \geq \tau_{\text{MI}} \geq 0$ are non-zero. For the qth covariate the aggregated coefficients are therefore

$$\overline{\beta}_{q,\boldsymbol{b}^*_j} = \frac{\Sigma^m_{r=1}\hat{\beta}^r_{q,\boldsymbol{b}^*_j}}{m} \cdot 1_{\{\hat{\rho}_{q,\boldsymbol{b}^*_j} \geq \tau_{\text{MI}}\}}. \tag{3.7}$$

This averaging method yields $\overline{\boldsymbol{\beta}}_{\boldsymbol{b}^*_j}$, a vector of length $Q+1$ which are the coefficient estimates for the jth bootstrap sample.

Stage 5: Repeat For $j = 1, ..., B$ repeat stages 2–4 and let $\overline{\boldsymbol{\beta}}^*$ be the $B \times (Q+1)$ matrix containing all the $\overline{\boldsymbol{\beta}}_{\boldsymbol{b}^*_j}$'s.

Stage 6: Stable model construction Stable variables need to be selected and included in the 'stable variable set'. Let $\hat{\boldsymbol{\rho}}_{\bullet,\overline{\boldsymbol{\beta}}^*} = (\hat{\rho}_{0,\overline{\boldsymbol{\beta}}^*}, ..., \hat{\rho}_{Q,\overline{\boldsymbol{\beta}}^*})$ be an inclusion frequency vector such that

$$\hat{\rho}_{q,\overline{\boldsymbol{\beta}}^*} = \frac{1}{B}\Sigma^B_{j=1} 1\{\hat{\rho}_{q,\boldsymbol{b}^*j} =\neq 0\} \tag{3.8}$$

with $0 \leq \hat{\rho}_{q,\overline{\boldsymbol{\beta}}^*} \leq 1$. Variables $q = 1, ..., Q$ are considered stable if $\hat{\rho}_{q,\overline{\boldsymbol{\beta}}^*} \geq \tau_{\text{B}}$, where τ_{B} is the 'bootstrap inclusion threshold' ($0 \leq \tau_{\text{B}} \leq 1$). Let the set of the s stable variables be denoted as $\alpha = \{q_1, ..., q_s\}$.

To obtain the final model, weighted logistic regression and model aggregation (using τ_{B}) is applied to the m imputed data sets for the stable variables $q \in \alpha$. (Note that

traditional multiple imputation theory suggests that imputation should be completed prior to any variable selection, Schafer, 1997). In this way stable variables have been selected and a measure of the intrinsic variability due to missing data and imputation has been retained.

There are some important aspects of the B-MI approach that need further attention. These include:

1. The choice of the inclusion frequency τ_{B},
2. The evaluation of the final model,
3. Assessing stability of variables and,
4. The evaluation of confidence intervals of the regression parameters for the final model.

Inclusion frequency τ_{B} The use of inclusion frequencies for model building and variable selection is an emerging concept. Here the inclusion frequency thresholds are used to select important variables, which are then used (based on the m complete imputed data sets , $\boldsymbol{Z}^1, ..., \boldsymbol{Z}^m$) to create a final model. Both, Müller and Welsh (2010) in a $Q < n$ context, and Meinshausen and Bühlmann (2010) in a $Q \gg n$ paradigm, used a range of inclusion frequencies to tune parameters of model selection procedures. An inclusion threshold has been shown to be conducive to bootstrap methods (Austin and Tu, 2004). To select an optimal τ_{B}, a range of values can be considered. By examining the bootstrap inclusion frequencies one is able to obtain an understanding of the instability of variables within models. In unsimulated cases where neither the true models, nor the full extent of the data complexity is known, selecting τ_{B} can be challenging. One could consider the trade-off between the number of variables desired in the model, that is whether it is parsimonious, and the required stability of the selected model and the expense of including a false positive variable. This could be achieved by considering a range for τ_{B} (as performed in the simulation study) and selecting an appropriate threshold. It is interesting in its own right to investigate the statistics of selected variables in determining τ_{B}. When such investigations are not possible one could suggest to call the variable q stable if $\hat{\rho}_{q,\overline{\boldsymbol{\beta}}^*} \geq \frac{1}{2\tau_{\mathrm{B}}}$. This rule of thumb is based on results reported in the following simulation study. The threshold τ_{B} can vary depending on the required stability for a particular model and context.

Evaluation of the final model It is desirable for a method to produce models that have a high prediction accuracy. To evaluate the predictive properties a validation set,

if available, can be used, which is independent from the main data used to build the model, but representative of the data as a whole. If such an independent validation set is not available, cross validation can be used instead, that is v random training and validation data splits are made. After the B-MI approach is applied and a model is constructed the AUC for the cross-validation set and the resubstitution AUC for the training set are retained. The validation AUC of each model is obtained using the held out data and is used to assess the predictive capabilities of the model in question. The B-MI approach is repeated for the v data splits. The AUC results are used to get a realistic understanding of the prediction because of the class imbalance.

Assessing stability of variables By examining the inclusion frequencies of a variable over the v simulated models, one is able to obtain a better understanding of the stability of variables within models. Some methods, for example, consistently yield models that select the same variables and the inclusion frequencies for these variables would be high with non-selected variables consistently low. Other methods, with a greater degree of instability, result in models with a range of variables with no particular variable set being more frequent, resulting in a large number of mid-range inclusion frequencies (around 50%). A stable model is one that consistently selects similar variables even when different perturbation of the data (here obtained through random splits) are observed. Hence, for this particular subset of variables, they will have high inclusion frequencies (around 70–90%).

Bootstrapped confidence intervals Bootstrapped confidence intervals can be constructed for the parameter estimates. There are multiple methods, ranging from the basic percentile intervals to the more complex and potentially more accurate bias-corrected and accelerated (BC_a) interval or the approximate bootstrap confidence (ABC) interval (Efron and Tibshirani, 1993).

3.4 Examining the B-MI approach

This simulation study investigates two elements. The first section considers the optimal parameters for the B-MI approach, addressing important components such as the τ_B tuning parameter. Second the B-MI approach with selected τ_B is compared with three other methods currently used in data analysis with missing values, including complete case analysis, where only samples with a complete set of variables are used, single imputation and multiple imputation. As class imbalance is a contributing factor to the unstable nature of the models being produced the simulation is applied both with and

without weights to compare how weights can affect the models being produced and AUC values are used to indicate prediction performance.

3.4.1 Simulated data

In this simulation study, data was constructed in two parts, the first component was the regression data, which was obtained from a subset of the EPU data and the second component was the response vector, simulated from a constructed model.

The design matrix contains 18 variables from part of the EPU study, these variables were selected to be used in the simulation study so that the simulated data would reflect real clinical data with similar properties including range, balance and distribution. Only a subset was selected so that the data would be relativity independent of other variables in the model. This criteria was chosen because multicollinearity and high levels of interdependencies within data sets increases the instability of models and a detailed study examining this issue is beyond the scope of this thesis. Table 3.5 contains a summary of the variables including their range, mean and data type.

The responses were generated from a logistic regression model with known parameter vector, $\boldsymbol{\beta}_{\text{true}}$, so that there was a high class imbalance (similar to the original EPU study) with 70 events and 346 non-events. To model the simulated responses, the vector pertaining to the probability of success for each sample, $\boldsymbol{\pi}_{\text{sim}}$, was given by

$$\pi_{j,\text{sim}} = \frac{1}{1 + e^{\mathbf{v_j^T}\boldsymbol{\beta}_{\text{true}}}} \tag{3.9}$$

with $j = 1, ..., n$. The variables are labeled A through U with the $\boldsymbol{\beta}_{\text{true}}$ values in Column 2 of Table 3.5. The responses were pseudo-random draws selected from a binomial distribution with probability of success $\boldsymbol{\pi}_{\text{sim}}$. The resulting data set has an overall missingness of 9%.

To examine the original stability of the simulated model, data was imputed once using MICE, which according to Section 3.2 is the best method as it has the smallest bias, and a logistic regression model was applied. The third column of Table 3.5 '90% CI for $\hat{\boldsymbol{\beta}}_{\text{true}}$' shows the 90% confidence intervals obtained from this first imputation of the data. The intervals in bold do not contain zero. From these results it is clear that model instability is present even when considering the complete data set, and not just a subset, as some variables that have a zero coefficient are producing confidence intervals not including zero, indicative of their inclusion in the model.

Variable	β_{true}	90% CI for $\hat{\beta}_{\text{true}}$	Range	Mean/Frequency	Type
–	2	**(3.00, 7.77)**	–	–	Intercept
A	0	(-0.048, 0.06)	15–42	27.27	Integer
B	0	(-0.41, 0.17)	0–6	0.73	Integer
C*	-1.5	**(-2.11, -0.82)**	0–3	0.23	Integer
D	0	**(-0.68, -0.09)**	0–6	0.63	Integer
E	0	**(-2.00, -0.65)**	0–1	310 / 106	Binary
F	0	(0.00, 0.06)	2–126	60.22	Continuous
G*	2	**(0.70, 1.75)**	0–1	179 / 237	Binary
H*	2.6	**(1.67, 3.39)**	0–1	364 / 52	Binary
J	0	(-0.88, 0.37)	0–1	325 / 91	Binary
K	0	(-0.05, 0.21)	0–10	1.06	Integer
L	0	(-0.03, 0.00)	67–193	147.22	Continuous
M*	-1.5	**(-1.96, -0.85)**	0–1	141 / 275	Binary
N	0	(-0.63, 0.43)	0–1	104 / 312	Binary
P*	-2	**(-2.19, -1.08)**	0–1	284 / 132	Binary
Q	0	(-0.62, 0.38)	0–1	276 / 140	Binary
S	0	(-0.85, 0.31)	0–1	227 / 189	Binary
T*	-0.2	**(-0.14, -0.06)**	1–79	29.14	Continuous
U	0	(-0.55, 0.02)	1–10	4.54	Continuous

* Indicates a variable in the true model

TABLE 3.5: Regression coefficients for the simulation. Such β_{true} values result in a highly imbalanced class distribution, for this simulated data set. Bold confidence intervals do not include zero.

3.4.2 Evaluation criteria

To evaluate the B-MI method, and its respective parameters, as well as to compare the abilities of the different methods, two evaluation criteria will be used: (i) stability for the variables and (ii) the prediction accuracy of the final model. To achieve these two results $v = 250$ simulation runs were generated. In each run the data was split into two sets; a validation set and a training set, in a 1/3 – 2/3 ratio. The splitting of the validation and training set was obtained through stratified pseudo-random draws so that the originăl class distribution within each set was maintained.

3.4.3 B-MI method - tuning τ_{B} and validating the use of weights

Within the B-MI approach there are many parameters to consider, perhaps the most important is τ_{B}, the inclusion threshold. To investigate how the bootstrap inclusion threshold alters the model being constructed, τ_{B} varied from 60% to 100%, that is a variable was only present in the final model if it was present in 60% to 100% of all bootstraped samples. Missingness within the data set was handled by applying MICE with $m = 25$, and $B = 500$ bootstraps. The aggregation of multiple imputation models

was restricted using $\tau_{MI} = 0.50$. To examine the importance of weights, the B-MI approach was applied using weighted logistic regression and standard logistic regression. When not clarified, the assumption is that the B-MI approach with weighted logistic regression was used.

Results of tuning and evaluating the B-MI method

The results from the B-MI analysis of the simulated data sets are summarised in Table 3.6 and Figure 3.9. Table 3.6 reports *inclusion frequencies*, that is the proportion of times a particular variable was in the final model out of the 250 simulation runs, as a percentage. The redundant variable 'E' is present in a large number of the logistic regression models while τ_B is low, once τ_B reaches 0.75 this false positive variable begins to drop out of the models. Table 3.6 highlights the division that exists between the variables in the true model and those not in the true model. The highlighted variables are those considered stable. A variable is considered stable if its inclusion frequency is above the stability threshold, where **the stability threshold** $= \frac{1}{2\tau_B}$, this split in the inclusion frequencies indicates that the final models are stable and consistently select only the true β variables, once τ_B is large enough. Figure 3.9 demonstrates that there is a consistency in the predictive capabilities between the resubstitution AUC values and the validation set AUC values implied by boxplots with a similar range. The average validation set AUC value is 83.15% and resubstitution AUC of 87.16%.

Figure 3.10 shows the number of variables at each τ_B value that are at or above a particular inclusion frequency. The plot shows that as the inclusion threshold τ_B increases the number of variables in the constructed models decreases.

3.4.4 Comparison of methods

To evaluate the B-MI approach, the simulated data set is used to compare the B-MI method with three methods currently used when data is missing:

1. Complete case (CC) analysis, always performed on samples with no missing variables (applied with and without weights);
2. Single imputation (SI) (MICE imputation algorithm with $m = 1$, with and without weights);
3. Multiple imputation (MI) (MICE imputation algorithm with $m = 25$[1] and $\tau_{MI} = 0.50$, with and without weights) and;

[1] '$m = 25$' was selected as this is a large number of imputations which still allows for practical computation times for this data set.

B-MI – τ_B	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.98	1.00
H*	100.00	100.00	100.00	100.00	99.20	98.00	94.80	87.60	67.20	0.00
T*	99.60	99.60	99.60	99.60	99.20	98.00	92.40	78.00	51.60	0.00
M*	100.00	99.20	97.20	97.20	94.80	87.60	78.00	59.20	33.20	0.00
C*	98.80	98.00	97.60	94.80	88.40	80.00	71.60	49.60	22.80	0.00
P*	93.60	89.20	80.40	73.20	62.80	48.40	37.20	17.20	4.80	0.00
G*	89.60	84.80	78.80	72.00	63.60	50.40	36.00	14.40	5.20	0.00
E	92.40	87.20	78.40	64.80	56.40	46.00	36.80	21.60	10.80	0.00
D	76.80	69.20	60.80	52.00	39.60	28.80	19.20	9.20	4.80	0.00
U	66.00	60.80	50.00	37.60	27.60	17.20	6.80	2.40	0.80	0.00
B	20.00	14.80	8.80	6.00	4.40	2.80	1.20	0.00	0.00	0.00
L	22.00	14.80	9.20	4.80	3.60	1.20	0.80	0.00	0.00	0.00
F	17.60	12.00	6.80	3.60	1.60	0.80	0.80	0.40	0.00	0.00
Q	8.00	5.60	4.00	2.40	1.20	0.80	0.00	0.00	0.00	0.00
J	6.80	4.40	2.80	2.80	1.60	0.40	0.00	0.00	0.00	0.00
S	1.60	1.20	1.20	0.40	0.40	0.40	0.40	0.00	0.00	0.00
N	2.80	1.60	0.80	0.40	0.40	0.40	0.00	0.00	0.00	0.00
K	5.20	2.40	0.80	0.40	0.40	0.00	0.00	0.00	0.00	0.00
A	1.20	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

TABLE 3.6: **Simulated data:** Inclusion frequencies for varying inclusion threshold τ_B. Variables are ranked in order of stability across all considered τ_B values. For each value of τ_B the highlighted variables are considered stable, as their inclusion frequency is above $1/2\tau_B$.

4. B-MI approach (B-MI) (MICE imputation algorithm with $m = 25$, $\tau_{MI} = 0.50$ and $\tau_B = 0.75$, with and without weights).

Results of comparison study

Tables 3.7 and 3.8 along with the boxplots in Figure 3.11 depict the results from the other methods used to analysis the simulated data set. Methods are compared based on their AUC values as well as their frequency stability. In all cases the complete data set was used (that is all 416 samples) except in the CC analysis 64 of the 416 (15%) of the cases were complete in which 21 (33%) were events.

The inclusion frequencies in Table 3.7 give insight into the stability of the features selected by each method. Both the SI and MI approaches include the true model variables in the majority of the 250 constructed models. The MI approaches are more stable than the SI methods as they yield larger inclusion frequencies ranging from 82% for variable 'C' in the weighted case, to 100%, where as the inclusion frequencies for the SI analysis were as low as 68% for variable 'P' in the unweighted case. This was not the case for the CC analysis, with some variables in the true model being picked up only a limited

number of times, for example variable 'H' selected in only 34% and 42% of the models with and without weights, respectively.

Not only is the consistency in selecting the true variables important but so too is the exclusion of the redundant variables. The selection of redundant variables is a problem for all the compared alternative methods, with inclusion frequencies for redundant variables as large as 96% (variable 'L', MI_W method). When these methods are compared against the B-MI approach, the superiority of the B-MI with weights approach becomes clear. With $\tau_B = 0.75$, all variables included in the true model have an inclusion frequency above 72% and the highest inclusion frequency of a true redundant variable is less than 65% (variable 'E'), highlighting the separation between included and excluded variables. Such separation is not the case for SI nor MI methods with all four (weighted and unweighted) approaches having variables with inclusion frequencies larger for some redundant than included variables, in particular consider variables 'E', 'D' and 'L' for these cases.

The variables flagged for concern at the very initial set up of the simulation, (namely variables 'D', 'E' and 'Q'), tend to have a larger inclusion frequency (especially 'D' and 'E') than other redundant variables (Table 3.7). For the SI and MI approaches these inclusion frequencies range from 82–92%, highlighting the severe problem of variable selection. The impact of these problematic variables is reduced when the B-MI approach is used especially as τ_B increases.

The AUC values indicate the predictive capabilities of the different methods summarised in Table 3.8. Figure 3.11 contains the boxplots for the resubstitution AUC which is a biased estimate, as well as the more informative validation AUC. For the validation set AUC, 0.74 was obtained for CC (with and without weights), 0.84 for SI with weights and 0.835 without and 0.85 for MI with weights and 0.84 for MI without weights. The AUC values for the B-MI approach are comparable, albeit slightly lower, to these existing methods, with a resubstitution AUC of 0.87 and a validation AUC of 0.83 for the B-MI with weight.

All methods considered are those that are used in current practice, yet some of these methods contain important problems that affect downstream interpretation. For example the use of CC analysis can mean that analysis is being applied on data that does not represent the original data or the population, and the use of SI underestimates the variably within the analysis. In this data set, CC analysis drastically reduced the data set and resulting error rates so that they were not representative of the original sample. For the original 416 samples only 15% of all cases are complete. This changes the distributions of some of the variables, for example the relative frequency of variable 'E' changes from 74.5% in the entire data set, to 90.6% in the complete case only data set

(p=0.007, from χ^2-test). For CC analysis, the reduced number of samples makes model evaluation difficult and over-fitting is evident.

	CC	CC_W	SI	SI_W	MI	MI_W	B-MI ($\tau = 0.75$)	B-MI_W ($\tau = 0.75$)
H*	**42.00**	**34.40**	**100.00**	**100.00**	**100.00**	**100.00**	**99.20**	**100.00**
T*	**46.40**	**41.60**	**98.80**	**94.80**	**100.00**	**96.00**	**97.60**	**99.60**
M*	**93.60**	**95.60**	**99.60**	**99.60**	**99.60**	**100.00**	**98.80**	**97.20**
C*	**48.00**	**47.60**	**91.20**	**78.00**	**92.80**	**82.00**	**77.20**	**94.80**
P*	**86.40**	**81.60**	**68.40**	**74.40**	**85.20**	**86.80**	**53.60**	**73.20**
G*	**86.40**	**86.80**	**92.80**	**81.20**	**96.40**	**84.80**	**74.80**	**72.00**
E	52.40	45.60	83.20	86.40	91.60	92.00	54.40	64.80
D	39.60	42.00	47.20	86.00	50.00	92.00	10.80	52.00
U	41.20	38.40	40.40	18.00	29.20	11.60	4.00	37.60
B	62.40	58.40	34.00	16.80	36.00	16.00	6.40	6.00
L	40.00	46.40	40.40	92.40	46.40	96.40	7.20	4.80
F	32.40	38.80	10.80	10.80	6.80	6.40	0.00	3.60
Q	41.60	37.20	5.20	12.00	2.80	7.20	0.00	2.40
J	53.20	49.20	6.40	8.00	4.00	2.40	0.00	2.80
S	38.40	44.00	3.20	25.20	0.80	22.40	0.40	0.40
N	40.40	39.60	2.40	3.20	0.00	0.40	0.00	0.40
K	32.80	35.20	5.20	30.80	2.80	22.40	0.00	0.40
A	42.80	37.60	1.60	4.80	0.00	1.60	0.00	0.00

TABLE 3.7: **Simulated data:** Inclusion frequencies for the 18 variables in the simulated data set. Comparisons for different analysis methods including Complete Case (CC), Single Imputation (SI), Multiple Imputation (MI) and Bootstrapped Multiple Imputation (B-MI) for $\tau = 0.75$ both using and not using a weighted logistic regression model. In bold are the variables in the simulated model and highlighted are the variables that have a larger inclusion frequency than these simulated model variables.

3.4.5 Simulation conclusions

The inclusion frequencies are used to gauge the stability of the models, with the B-MI (with weights included) approach resulted in the most stable models. For the B-MI approach, as τ_B increased the separation between variables in the true model and redundant variables increases, and the stable variables became evident. This was not the case with the other methods, the MI approaches frequently selected variables not in the true model for example variable 'E' and 'L', and the CC approach frequently missed non-redundant variables, for example variable 'H'.

Regarding predictive capabilities, all the compared methods had comparable AUC values. In the CC case, it was clear that the data was being over-fit, however in all other cases the resubstitution AUC was around 0.87 and the validation AUC was about 0.84.

Method	Mean validation AUC	Mean resub AUC
Complete Case (CC)	74.38%	99.81%
Complete Case with weights (CC_W)	74.24%	99.87%
Single Imputation (SI)	84.24%	88.41%
Single Imputation with Weights (SI_W)	83.49%	87.74%
Multiple Imputation (MI)	84.84%	87.94%
Multiple Imputation with Weights (MI_W)	84.21%	87.47%
Bootstrap Multiple Imputation (B-MI) without weights ($\tau_B = 0.75$)	82.93%	86.28%
Bootstrap Multiple Imputation (B-MI_W) with weights ($\tau_B = 0.75$)	83.15%	87.16%

TABLE 3.8: Mean validation and mean resubsitution AUC values for different analysis methods.

Such a high validation set AUC is indicative of good predictive capabilities of the final models. For the B-MI approach the choice of τ_B affected the AUC values because when τ_B was very high, (about 0.90) the number of variables in the final model was reduced and the AUC values suffered resultantly. This reduction in predictive capabilities was illustrated by the downward trend in AUC boxplots as τ_B increases (Figure 3.9).

The decision to add weights or not to logistic regression models within the B-MI approach is very data set dependent. Analysis for this simulation was performed both with and without weights (non-weighted data results can be found in Appendix B, Section B.2). Regarding the inclusion of weights in the B-MI approach, Table 3.8 shows that including weights increases the AUC prediction values and Table 3.7 highlights the increased stability enjoyed by the B-MI approach with weights.

Comparing weighted and unweighted cases for the other compared methods (Table 3.7) the increased frequency of the selection of non-redundant variables, coupled with the results from Weiss and Provost (2001) leads us to suggest that weights should be considered when highly unbalanced data sets are used.

These results indicate that a data dependent, optimal inclusion threshold exists. The optimal inclusion threshold allows for variables with non-zero coefficients to consistently be included in the model, but variables with extremely small or a null coefficient to be left out of the model. In the simulation study an inclusion threshold of 0.75–0.90 produces favourable results. When τ_B reaches 0.75 the variables consistently selected to be included in the model reflect the simulated data. In real applications the true model

is typically unknown. Therefore, inclusion and other tuning parameters are selected by taking into account model complexity, for example counting the number of variables in the model, and parsimony, that is striving to provide a small and useful final model.'

In addition to the simulation study described, a range of simulations over a range of class imbalances were carried out. When class imbalances was more extreme, as expected stable models were more difficult to obtain, with a higher amount of variability in the inclusion frequencies. With the higher class imbalance and the greater variability, AUC values also decreased. The difference between the size and accuracy of final predictive models in the weighted and unweighted cases was more extreme, with a larger final model providing a better AUC estimate.

3.5 Case study: EPU data

3.5.1 Data description

The motivating data set underpinning the methodology was provided by the Early Pregnancy Unit located at the Nepean Hospital, Sydney, Australia, (EPU data). Results for this study were originally presented in Riemke et al. (2011), which serves as a reference for a full description and analysis. A complete data description can be found in Chapter 2 Section 2.1.

3.5.2 B-MI Model

The B-MI approach with weighted logistic regression was performed on the EPU data set. Table 3.9 contains the recommended parameters obtained from the simulation, these values have been obtained from the previous simulation study.

B-MI Parameters		
Multiple Imputation	Algorithm	Mice
	m	25
Bootstraps	B	500
Model reduction	Criteria	BIC
	τ_{MI}	50%
	τ_B	75%

TABLE 3.9: Parameters used in the EPU study, selected from results obtained from the simulation study Section 3.4

For the bootstrapping, stratification was employed so that (i) the class distribution remained the same and (ii) the distribution of missingness within the samples was maintained, that is the distribution of near complete samples (less than 3 out of 21 variables missing) and samples with high levels of missingness (more than 3 out of 21 variables missing).

Variable	Mean regress coef	Odds ratio	Inc. freq
GS mean	-0.15	0.86	0.972
Clots	2.11	8.23	0.934
Number of previous caesareans	-0.90	0.41	0.884
Subchorionic bleed	-0.28	0.75	0.876
VAS 0–10	-0.04	0.97	0.758

TABLE 3.10: The final model, found using bootstrapping, multiple imputation and weights presented as regression coefficients.

Table 3.10 shows the final model. The coefficients, and subsequent odds ratios were found by constructing a model on the five selected variables and the multiple data sets. To evaluate the stability and predictive accuracy of the model, 250 training and validation splits were randomly produced from the data. The 'Inc. freq' column in Table 3.10 indicates the stability of the included variables and Figure 3.12 shows the validation and resubstitution AUCs. This resulted in an estimated validation AUC of 75.78%.

The coefficients obtained from this model (Table 3.10) are considered the final β estimates for the logistic regression model. The variables retained in the final model are in agreement with known factors associated with miscarriage, in particular 'clots' and 'GS mean' (Choong et al., 2003; Tower et al., 2000). The effect associated with 'number of previous caesareans' is also in agreement with results reflecting modern caesarean surgeries, namely that such a procedure has no effect on miscarriage rates (Nielsen and Hökegaard, 1984; Siddiqi et al., 1988).

3.5.3 Note regarding information criterion feature selection

Modern model selection procedures, applied in this thesis, are based on information criteria rather than extracting p-values from repeated hypothesis testing. The major concern is to consider feature selection as a group of interacting variables as opposed to unique, independent variables. An approach which selects a model having smallest AIC value has not necessarily all partial p-values smaller than a nominal level of 5%, i.e. confidence intervals can contain zero. Claeskens and Hjort (2008) advocate strongly an

information theoretic approach to model selection. As a result of this selection paradigm confidence intervals become redundant and at times uninformative, and have hence not been induced in this final model, or models appearing in subsequent chapters. It is of course possible to construct confidence intervals is desired using the BC_a or other such intervals.

3.6 Conclusion

Missing data, imbalance of class distribution and unstable models are common hurdles often needing to be overcome in a clinical data context. The final model should be one with stable variables, consistently selected when perturbations in the data set are made. This model should also have consistent and good predictive capabilities.

Missing data is a challenge in many data sets, and although there are many ways to overcome this problem, multiple imputation is currently the most statistically robust. Section 3.2 provides a framework to compare different imputation methods. Included in this chapter is the comparison of Amelia II, Mi and MICE. The comparison framework's capacity is not limited to these three algorithms and can easily be extended to compare any other current methods, or multiple imputation methods developed in the future. Comparisons take into consideration how the amount of missing data changes downstream analysis including logistic regression coefficients and prediction accuracy. Through a simulation study it was found that as the proportion of missing data increases in a data set the distribution of regression coefficients increase in variability or suffer a shrinkage of the mean toward zero. Downstream analysis is severely affected when missingness is extreme (40–45%).

Such conclusions were achieved through examining how the multiple imputation methods and the amount of missing data alters the logistic regression coefficients as well as how these elements impact upon prediction accuracy using logistic regression and random forests. Although different imputation methods are affected in different ways, MICE maintains the closest fit to the original distributions as the amount of missing data increases and is therefore recommended as the method of choice. A proportion of missing data around 30% is still manageable through multiple imputation, concuring with Rubin (1987).

If a final model does not have stable variables nor good predictive capabilities it is of little use in describing the data or being used to interpret independent data. The B-MI method presented in Section 3.4 makes use of multiple imputation to overcome missing data and bootstrapping to result in a final stable model coupled with regression weights

and appropriate model evaluation methods to handle class imbalance. Throughout the study MICE was used as the multiple imputation algorithm. Tuning the bootstrap inclusion threshold, τ_B, is important for selecting stable models and through a simulation study it was extrapolated that as a guide $\tau_B = 0.75$ should be used to produce stable models when the multiple imputation inclusion threshold is set such that $\tau_{MI} = 0.50$. When compared to other currently available modelling methods that handle missing data, the B-MI method with weights produced superior results. For this comparison superiority was based on stability and the construction of a model that correctly reflects the model used to simulated the data.

3.7 Manuscripts under review

This chapter includes the work under review in Campain et al. (2011) and Riemke et al. (2011). The EPU analyses were conducted by the author with Associate Professor George Condous's groups at the Acute Gynaecology, Early Pregnancy and Advanced Endosurgery Unit at the Nepean Centre for Perinatal Case, University of Sydney, Nepean Hospital, working in particular with Dr Jennifer Oates (Riemke).

- **A.E. Campain**, S. Müller, G. Condous and Y.H. Yang (2011) Stable logistic regression models in the presence of missing values and class imbalances. Under review, *Biostatistics*.
- J. Riemke, **A.E. Campain**, T. Bignardi, I. Casikar, D. Alhamdan, D. Fauchon, R. Benzie, S. Müller, T.H. Yang, M. Mongelli and G. Condous (2011) Development of a new model to predict viability at the end of the 1st trimester after a single visit to an Early Pregnancy Unit. Preprint.

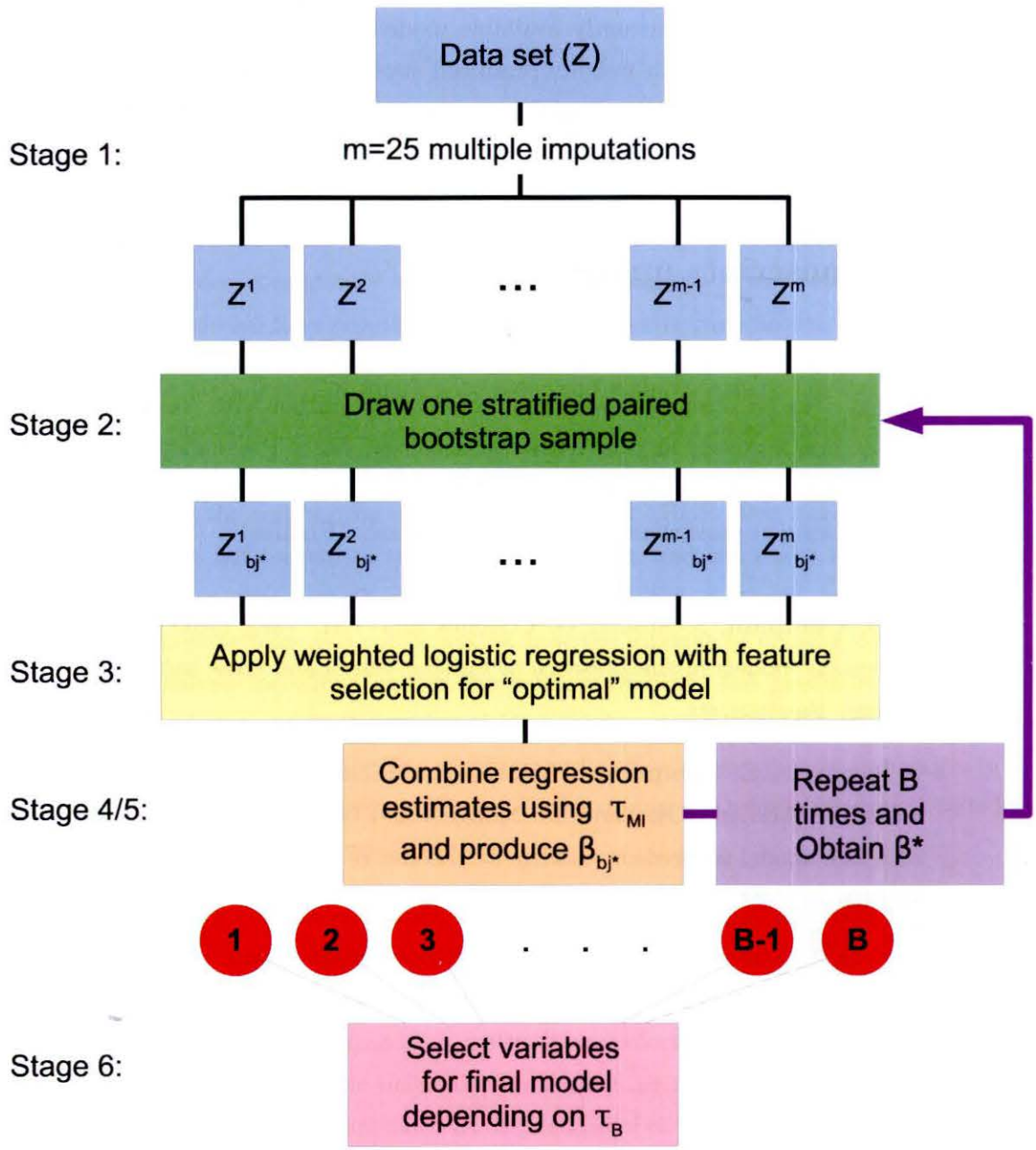


FIGURE 3.8: Graphical representation of the proposed method for variable selection within a logistic regression model.

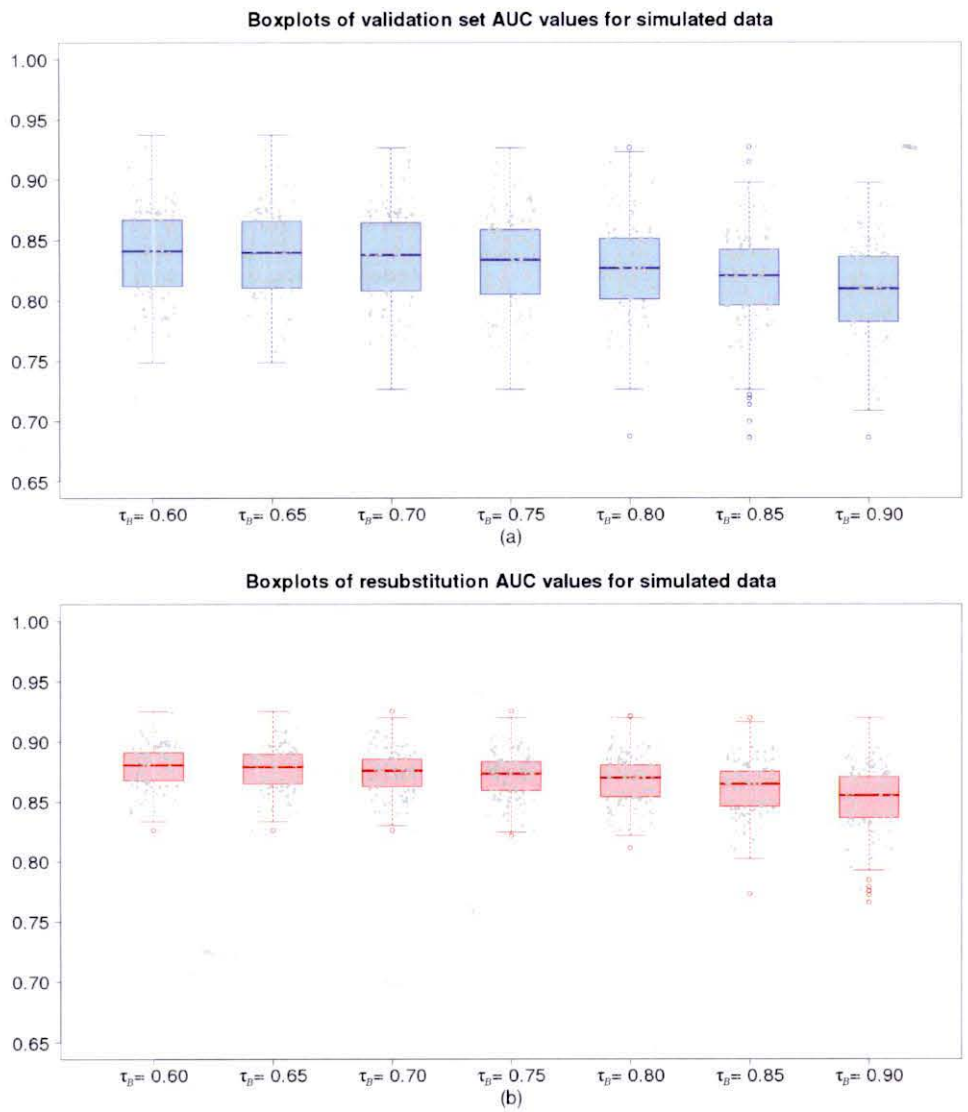


FIGURE 3.9: Boxplots of 250 **(a)** validation and **(b)** resubstitution AUC values. As τ_{B} increases the resubstitution AUC and validation AUC values decrease. Here the reduction in the number of selected variables in the model reduces predictive capabilities, for $\tau_{\mathrm{B}} \geq 0.75$.

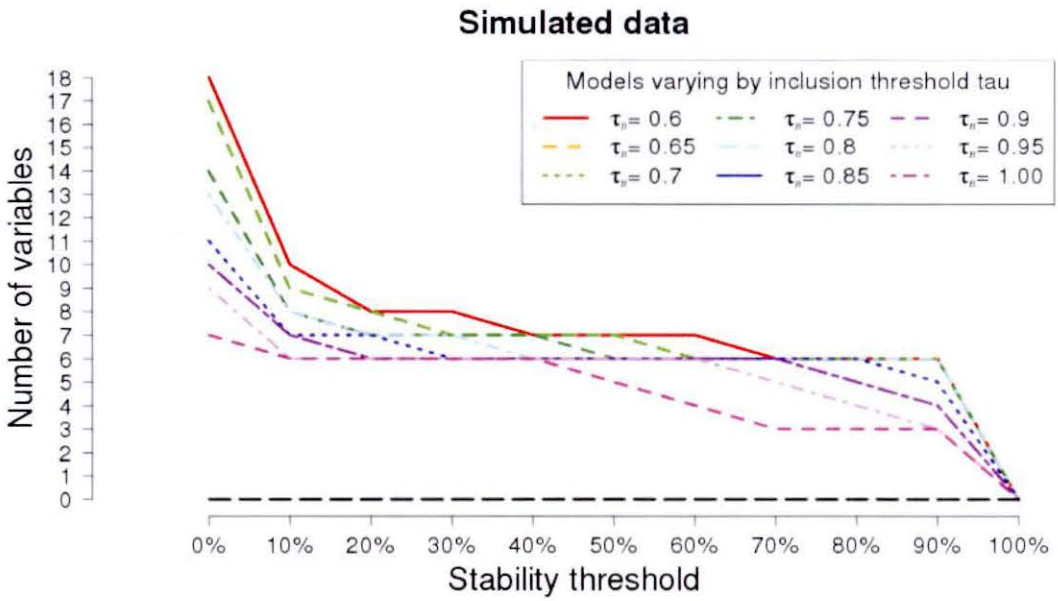


FIGURE 3.10: **Simulated data:** The stability of the variables in the models built using B-MI, as the inclusion frequency τ_B changes. This plot shows the number of stable variables for each of the varying inclusion frequencies. The stability threshold varies from 0% to 100%, and plotted is the number of variables that are considered 'stable' as the stability threshold increases. The selection of stability threshold depends on the required stability of the variables in the final model, and hence on the context of the model being developed.

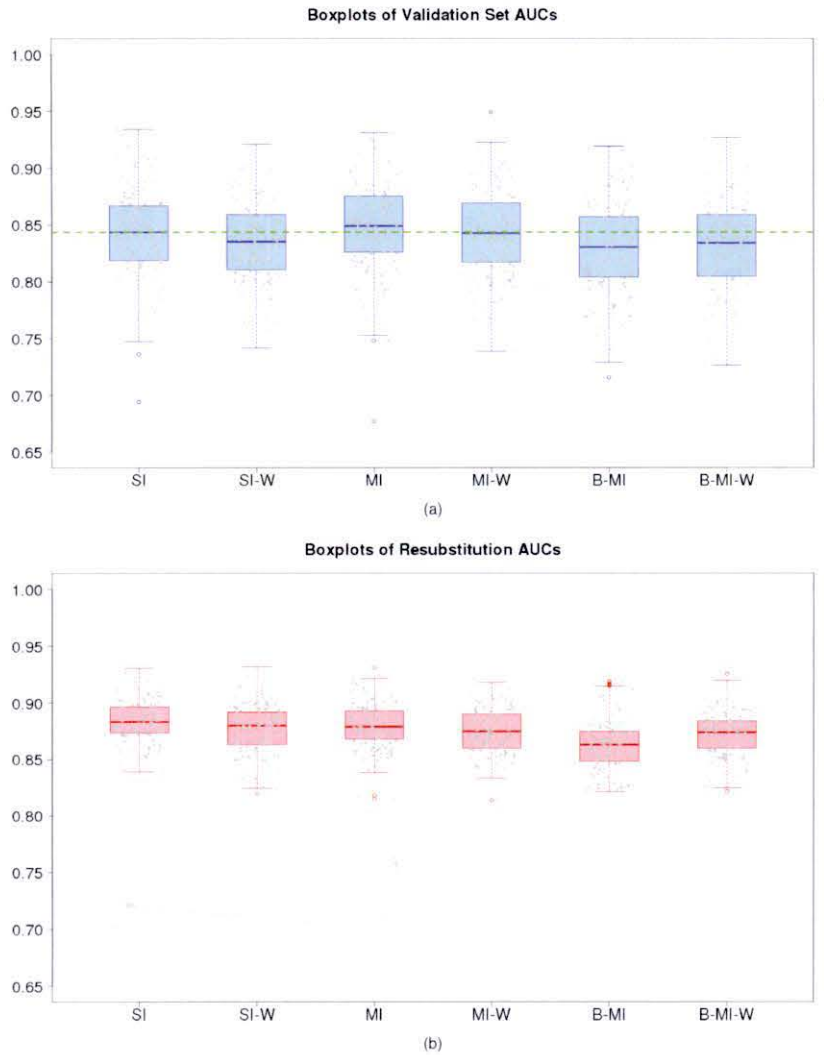


FIGURE 3.11: Boxplots of **(a)** the validation and **(b)** resubstitution AUC values for 250 random training set splits for multiple analysis methods. Validation AUCs can be used to estimate the predictive capabilities of the model construction methods.

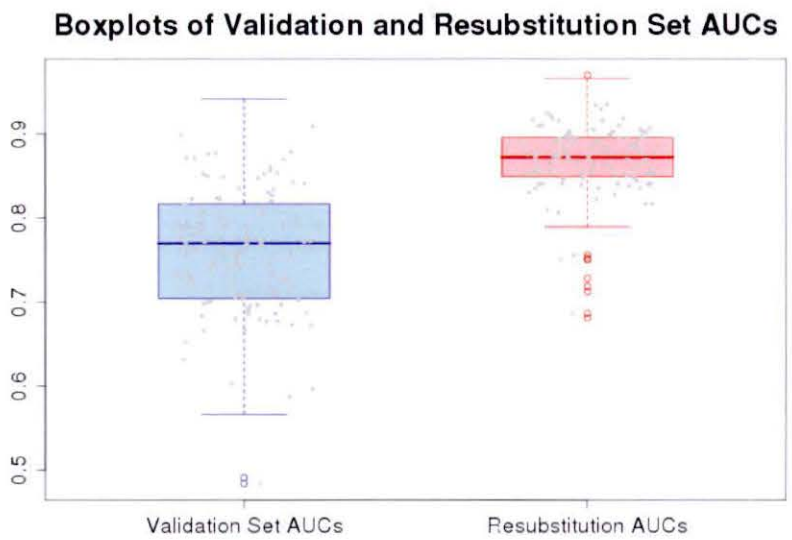


FIGURE 3.12: Boxplots of the AUC values for 250 random validation and training set splits. Validation and resubstitution AUC values are shown for $\tau_{\mathrm{B}} = 0.75$.

Chapter 4

Integration of gene expression data

Science has embraced microarray technology and due to extensive usage in recent years there has been an explosion in publicly available data sets. Examples of repositories available for the access of such data include Gene Expression Omnibus[1] (GEO) (Barrett et al., 2005), ArrayExpress[2] (AE) (Parkinson et al., 2005) and Stanford Microarray Database[3] (SMD) (Hubble et al., 2009), as well as individual researchers' and institutions' websites. The usefulness of these data sets has not been superannuated, when used wisely they may yield a depth of information. Demand has increased to effectively utilise these data sets in current research as additional data for analysis and verification.

The integration of data sets deals with analysis methods that traditionally incorporate the synthesis or at times review of results from data sets that are independent but related (Normand, 1999). Integrating data sets has a range of benefits. For example, power can be added to an analysis, obtained by the increase in sample size of the study. This aids the ability of the analysis to find effects that exist and is termed 'integration-driven discovery' (Choi et al., 2003). Integration can also be important when studies have conflicting conclusions as they may estimate an average effect or highlight an important subtle variation (Hong and Breitling, 2008; Normand, 1999).

There are a number of issues associated with integrating gene expression studies. These include problems common to traditional data set integration such as overcoming different aims, design and populations of interest. There are also concerns specific to gene

[1] `http://www.ncbi.nlm.nih.gov/geo/`
[2] `http://www.ebi.ac.uk/arrayexpress/`
[3] `http://genome-www5.stanford.edu/`

expression data including challenges with probes and probe sets, differing platforms being compared and laboratory effects (Campain and Yang, 2010). As different microarray platforms contain probes pertaining to different genes, platform comparisons are difficult when comparing these differing gene lists. Often the probes in the intersection are the only ones retained for further analysis. Moreover, when probes are mapped to their 'Entrez IDs' (Maglott et al., 2011) for cross platform comparisons often multiple probes pertain to the same gene. Due to reasons ranging from alternative splicing to probe location these probes may produce different expression results (Ramasamy et al., 2008). Ideal methods for aggregating these multiple probe results in a meaningful and robust way is currently the topic of much discussion, but not part of this thesis. Laboratory effects are important because array hybridisation is a sensitive procedure. Influences that may affect the array hybridisation include different experimental procedures and laboratory protocols (Irizarry et al., 2005), sample preparation and ozone level (Fare et al., 2003). For more details on the integration of microarrays in general as well as difficulties associated with such analysis refer to Cahan et al. (2007); Fierro et al. (2008); Grützmann et al. (2005); Larsson et al. (2006); Ramasamy et al. (2008); Rhodes et al. (2002).

In this chapter, data integration will be considered in two levels; 'meta-analysis' and 'mega-analysis':

- **Meta-analysis** is the integration of the *statistics* from different microarray studies. Such a method looks at how genes correlate to a phenotype within a data set *after* the analysis, and has been termed 'relative' by Larsson et al. (2006) and 'indirect' by Fierro et al. (2008). Statistics or p-values from multiple studies are compared or aggregated to obtain features which are commonly considered important. There are multiple meta-analysis methods currently available including Fisher's inverse chi-squared (Fisher, 1950), 'GeneMeta' (Choi et al., 2003; Gentleman et al., 2008), 'Probability of expression matrices' (Parmigiani et al., 2002) and 'RankProd' (Breitling et al., 2004). A new meta-analysis approach was produced by Campain and Yang (2010), 'Meta differential expression via distance synthesis' (mDEDS) which can be used to identify differentially expressed (DE) genes from multiple data sets. This new method makes use of multiple statistical measures across data sets to obtain a DE list.

- **Mega-analysis** refers to integrative methods that combine *data sets* prior to analysis, making one large or 'mega' data set. This method is also known as 'absolute' (Larsson et al., 2006) or 'direct' (Fierro et al., 2008) meta analysis. A central component to this approach is a normalisation method that can handle this type of data integration. Such methods include, preprocessing data together,

applying study and platform effects in downstream models, or sophisticated batch correction/normalisation approaches including ComBat (Johnson et al., 2007) and RUV-2 (Gagnon-Bartsch and Speed, 2011). Traditional microarray analysis tools are then used on this large data set post normalisation.

To explore meta- and mega-analysis, this chapter contains three main studies concerning intergrative analysis. These studies look to compare integrative methods in a novel way, highlighting different aspects and strengths of meta- and mega-analysis methods:

1. **Case study 1:** Section 4.4.1 is a simulation study used to compare meta-analysis methods when the list of differentially expressed genes is constructed and known. Performance is measured with receiver operating characteristic (ROC) curves as well as the area under these ROC curves (AUC) (see Section 3.3.1 for more details).

2. **Case study 2:** Section 4.4.2 is a case study which considers the meta-analysis of three different melanoma studies in a classification context. The data sets used in this study include the Bogunovic, Jönsson and Mann data sets (Sections 2.3.1 and 2.3.2). Performance in this regard is based on the ability of classifiers constructed after meta-analysis to predict results for an independent data set.

3. **Case study 3:** Section 4.4.3 is a case study considering the DE analysis of hypertensive versus normotensive rats. This case study compares four independent data sets, the Cerutti, Clemitson, Grayson and Rysä data sets (Section 2.2), and highlights the use and effectiveness of meta- and mega-analysis. Integrative approaches are used to examine conflicting conclusions obtained from the DE analysis of the four data sets separately and how the use of appropriate methods may suggest a compromise to these inconsistencies.

This chapter continues with a description of meta- and mega-analysis methods (Sections 4.1 and 4.2), including the description of mDEDS (Campain and Yang, 2010). Section 4.3 outlines the performance assessment methods that can be used to compare the effectiveness of different meta- and mega-analysis methods, and to evaluate their success for data sets in question. Section 4.4 contains three case studies where different aspects of the integrative methods are observed and compared. As described above, the first case study is a simulation comparison and the other two are concerned with real data. A discussion of meta- and mega-analysis methods follows in Section 4.5.

Notation while working in integrative expression data needs to be considered with care. Let $\boldsymbol{X}$ represent an expression matrix $(I \times n)$, with $i = 1, ..., I$ genes and $j = 1, ..., n$ samples. The rows represent genes and the columns represent samples, with each element

of the matrix denoted as $x_{i,j}$. Note that such a data matrix is the transpose of most standard data matrices within statistics, but this notation is commonly employed in the microarray setting. If there are $k = 1, ..., K$ data sets, n_k represents the number of samples in the kth data set. For simplicity, and without loss of generality, we focus on dichotomous response; i.e., two-group comparisons. We designate groups as treatment T and control C. For two-channel competitive hybridisation experiments, logarithms of the relevant fluorescent intensity measurements are obtained. We assume that the comparisons of log-ratios are all indirect; that is we have n_T arrays in which samples from group T are hybridised against a reference sample R, and we can obtain n_T log-ratios, $M_{T_j} = \log_2(T_j/R)$; $j = 1, ..., n_T$ from group T. In an identical manner n_C log-ratios are also calculated from group C. For single colour arrays such as Affymetrix oligonucleotide array experiments, we have n_T chips with gene expression measures from group T and n_C chips with gene expression measures from group C.

4.1 Meta-analysis

Within this section several methods will be presented. Such methods include the very classical Fisher's inverse chi-squared method as well as methods developed in the microarray context including GeneMeta, Probability of expression matrices, RankProd and mDEDS. Meta-analysis is mainly used in a DE context, to obtain a list of DE genes using results from multiple analyses. Such methods lend themselves to classification, for although feature selection is a slightly different concept, the strengths of an integrated DE list have the potential to aid in the development of an informative feature list.

These five methods are compared with the performance of the 'data set cross-validation' method. 'Data set cross-validation' is a naïve approach which assumes that results obtained from one study are directly applicable to another study. For example if a classification rule was constructed on data set A, it could be used to predict results from data set B. This method is considered here as a very simple meta-analysis approach, as information is being gathered from one study and then being applied to another. Although 'Data set cross-validation' is attractive for its simplicity and heuristic sense it is not recommended because in practice vast amounts of between study variability renders the classifier mute. This last approach is applied as a comparison tool throughout this study. Although ideally a classification rule should be directly applicable to independent and unique data addressing the same conditions, the poor performance of the 'Data set cross-validation' method highlights that this not the case for real expression data.

Fisher's inverse chi-squared

Fisher, in the 1930s developed a meta-analysis method that combines the p-values from independent data sets. One of a plethora of methods for combining the p-values (Fisher, 1950) is the Fisher summary statistic,

$$S_i = -2\sum_{k=1}^{K}\log(p_{ik}), \tag{4.1}$$

which tests the null hypothesis that for gene i, there is no differences in expression means between the two groups. The p-value p_{ik} is the p-value for the ith gene from the kth data set. In assessing S_i, the theoretical null distribution should be χ^2_{2K}. It is also possible to extend Fisher's method by producing weights for different data sets based on, for example, quality.

GeneMeta

One of the first methods that integrates multiple gene expression data sets was proposed by Choi et al. (2003) who described a t-statistic based approach for combining data sets with two groups. An implementation of this method is found in `GeneMeta` (Gentleman et al., 2008) an R package containing meta-analysis tools for microarray experiments.

Choi et al. (2003) described a meta-analysis method to combine estimated *effect-sizes* from the K data sets. In a two group comparisons, a natural effect size is the t-statistic. For a typical gene i, the effect size for the kth data set is defined as

$$d_k = \frac{\bar{T}_k - \bar{C}_k}{S_{\mathrm{p}k}}, \tag{4.2}$$

where $\bar{T}_k$ and $\bar{C}_k$ represent the means of the treatment and the control group respectively in the kth study. $S_{\mathrm{p}k}$ is the pooled standard deviation for the kth data set.

For K number of observed effect sizes, Choi et al. (2003) proposed a random effects model

$$d_k = \mu + \delta_k + \varepsilon_k \ , \ \varepsilon_k \sim N(0, s_k^2)$$

where μ is the parameter of interest, s_k denotes the within study variances and $\delta_k \sim N(0, \tau^2)$ represents the between study random effects with variance τ^2. Choi et al. (2003) further mentioned that when $\tau^2 = 0$, δ_k denotes the between study effect in a fixed effect model, which assumes that the difference of observed effect sizes are from sampling error alone. The random effects model is then estimated using a method

proposed by DerSimonian and Laird (1986) and a permutation test is used to assess the false discovery rate (FDR).

Probability of expression matrix

Given a data set, the probability of expression (POE) method (Parmigiani et al., 2002) transforms the expression matrix $\boldsymbol{X}$ to $\boldsymbol{E}$ which is an indicator matrix representing latent classes. Each matrix element, E_{ij}, is defined as the chance of multiple conditions present across n samples within gene i. The transformed matrix, $\boldsymbol{E}$, consists of three values $-1, 0, 1$ that represent the conditions 'under-expressed', 'not differentially expressed' and 'over-expressed' respectively. After the transformation into a POE matrix, genes of interest are established using 'integrative correlation' (IC) (Parmigiani et al., 2004). Notice that this integrative correlation method is not restricted to be used with a POE matrix. The method IC begins by calculating all possible pairwise Pearson correlations ($\hat{\rho}^k_{(i,i')}$, where $i \neq i'$) between genes i and i' across all samples within a data set k. Thus, a pairwise correlation matrix D is generated with $R = \binom{I}{2}$ rows representing the number of pairwise correlations and K columns representing the number of data sets.

For a selected pair of data sets k and k', let $\bar{\rho}^k$ and $\bar{\rho}^{k'}$ denote the means of the correlations per study. Gene-specific reproducibility for gene i is obtained by only considering comparisons that contain the ith gene. That is

$$I_i(kk') = \sum_{i'=1}^{I} (\hat{\rho}^k_{(i,i')} - \hat{\bar{\rho}}^k)(\hat{\rho}^{k'}_{(i,i')} - \hat{\bar{\rho}}^{k'}), \tag{4.3}$$

where $i \neq i'$. When more than two data sets are being compared, all integrative correlations for a particular gene are aggregated. This method provides a combined ranking for genes across K data sets.

This method is implemented in the R package `metaArray` which contains a number of meta-analysis methods. The main function in this package is a two steps procedure which transforms the data into a POE matrix and followed by a gene selection method based on IC.

RankProd

RankProd is a non-parametric meta-analysis method developed by Breitling et al. (2004). Fold change (FC) is used as a selection method to compare and rank the genes within each data set. These ranks are then aggregated to produce an overall score for the genes across data sets, obtaining a ranked gene list.

Within a given data set k, pairwise FC (pFC) is computed for each gene i as

$$T_1/C_1,\ T_1/C_2,...,\ T_{n_{T1}}/C_{n_{C1}} \tag{4.4}$$

producing $n_T \times n_C$ pFC_{lmk} values per gene with $l = 1, ..., n_T$ and $m = 1, ..., n_C$. The corresponding pFC ratios are ranked and we may denote this value as $\text{pFC}_{(irk)}$, where I is the number of genes ($i = 1, ..., I$) and R is the number of pairwise comparisons ($r = 1, ..., R$) between samples. Then the rank products for each gene i is defined as

$$RP_i = (\prod_{k=1}^{K} \prod_{r=1}^{R} pFC_{(irk)})^{\frac{1}{R}}. \tag{4.5}$$

Expression values are independently permuted B times within each data set relative to the genes, the above steps are repeated to produce $RP_i^{(b)}$, where $b = 1, ..., B$. A reference distribution is obtained from all the $RP_i^{(b)}$ values, and the adjusted p-value for each of the I genes is obtained. Genes that are considered significant are used in future analysis.

Meta differential expression via distance synthesis

There are many different ways of estimating DE genes and different results are obtained from different microarray platforms analysed as evident in the inconsistent results obtained from multiple studies (Boulesteix and Slawski, 2009; Russ and Futschik, 2010; Zhang et al., 2008). The assumption behind the novel mDEDS method (Campain and Yang, 2010), is the consistency of DE genes, in that genes that are truly DE will be estimated as DE regardless of the platform or the statistic used. 'Meta differential expression via distance synthesis' (mDEDS) makes use of multiple statistical measures from all the considered data sets, to obtain an integrated DE list. It is extended from 'Differential expression via distance synthesis', DEDS (Yang et al., 2005b), which is designed to obtain DE gene lists from different DE measures. Example DE measures include standard and moderated-t stat (Smyth and Wettenhall, 2003), FC, SAM (Tusher et al., 2001) and the B-statistic, amongst others.

The DE method DEDS works under the assumption that true DE genes should score highly within a set of non-dominated genes, over a range of statistical measures. Through permutations these highly scoring genes are calculated and ranked in order of overall significance when compared to the null results generated by the sample permutations. The mDEDS approach uses these non-dominated genes both within and between data sets from different platforms and still using a range of DE measures constructs a ranked list. Consistently high ranked genes are then considered DE via mDEDS. This method

endeavours to be robust against two elements. The first, when different measures produce significantly different ranked lists, and the second, 'platform specific bias', when particular platforms produce results that are more favourable to particular gene sets. Campain and Yang (2010) compare DEDS to mDEDS and find the ability for mDEDS to be robust against platform bias results in a more successful DE tool in a meta-analysis context.

The mDEDS process uses several key steps:

1. Let there be $k = 1, ..., K$ data sets and $g = 1, ..., G$ appropriate (DE measuring) statistics, hence there will be $K \times G$ statistics for each of the $i = 1, ..., I$ genes. Let t_{ikg} be the statistic for the ith gene, from the kth data set for the gth DE measure. Assuming large values indicate increased DE genes, let the observed coordinate-wise extreme point be

$$E_0 = (\max_i(t_{i11}), ..., \max_i(t_{i1G}), ..., \max_i(t_{iKG})). \tag{4.6}$$

2. Locate the overall (observed, permutation) extreme point E:

   (a) Each of the K data sets is permuted B times by randomly assigning n_T arrays to class 'T' and n_C arrays to class 'C', producing $b = 1, ..., B$ sets of K data sets. For each permuted data set the G number of DE statistics are recalculated yielding t^b_{ikg}. Obtain the corresponding coordinate-wise maximum:

$$\boldsymbol{E}_b = (\max_i(t^b_{i11}), ..., \max_i(t^b_{i1G}), ..., \max_i(t^b_{iKG})). \tag{4.7}$$

   which is a vector of length KG

   (b) Obtain the coordinate-wise permutation extreme point E_p by maximizing over the B permutations,

$$\boldsymbol{E}_p = (\max_b(E_{b11}), ..., \max_b(E_{b1G}), ..., \max_b(E_{bKG})). \tag{4.8}$$

   (c) Obtain E as the overall coordinate-wise maximum: $E = \max(\boldsymbol{E}_p, E_0)$.

3. Calculate a distance d from each gene to E. For example, one choice for a scaled distance is

$$d_i = \sum_{k=1}^{K} \sum_{g=1}^{G} \frac{(t_{ikg} - E_{kg})^2}{\text{MAD}(t_{ikg})^2}, \tag{4.9}$$

   where MAD is the median absolute deviation from the median. Order the distances, $d_{(1)} \leq d_{(2)} \leq ... \leq d_{(n)}$.

Hence a ranking of genes to be used in downstream analysis is obtained.

4.2 Mega-analysis

Mega-analysis is an integrative approach that combines K smaller data sets $(\boldsymbol{X}_1, ..., \boldsymbol{X}_K)$ into a larger data set $(\boldsymbol{X})$ often after some form of scaling and/or expression adjustment. The main component of mega-analysis is 'normalisation', hence each mega-analysis method can be considered a normalisation method. Four such normalisation methods are presented including Null correction, where no normalisation occurs, Quantile normalisation, ComBat and RUV-2.

Null correction

The 'Null correction' is a naïve method where several data sets are combined into one larger data set for downstream analysis, with no normalisation applied to the data. This method is rarely performed in practice. To perform Null correction Entrez IDs from the expression platforms (if the platforms are different) are matched and the data sets are simply placed together forming a larger matrix from the several smaller ones.

Quantile normalisation

Quantile normalisation is a method applied to the data obtained from the Null correction approach. However, the additional stage in the analysis forces the quantiles of the different data sets to be identical. Ideally this means that no batches can be determined by examining the distributions of the expression levels post integration. Quantile normalisation is a common method for correcting data found in literature. This method is typically not used in isolation, and often when a linear model is applied to evaluate DE genes, both a study and platform (if required) effects are included in the model.

ComBat

Johnson et al. (2007) proposed a batch correction method, designed for small sample sizes, known as 'ComBat' making use of a parametric (and non-parametric) empirical Bayes framework. Although ComBat was initially designed to correct for batch effects, it can also be applied in a mega-analysis context where each of the data sets are considered independent batches.

Initially the samples are standardised gene-wise, leaving genes with a similar overall mean and variance. Assuming the model

$$Y_{ijh} = \mu_i + \boldsymbol{Z}\boldsymbol{\beta}_i + \gamma_{ih} + \delta_{ih} + \varepsilon_{ijh}, \tag{4.10}$$

where Y_{ijh} is the log expression value for the ith gene, from the jth sample within the hth batch and $\boldsymbol{Z}$ is the design matrix for the sample conditions, with the error term distributed normally with zero mean and variance σ_i^2. The parameters μ_i, $\boldsymbol{\beta}_i$ and γ_{ih} are estimated as $\hat{\alpha}_i$, $\hat{\boldsymbol{\beta}}_i$ and $\hat{\gamma}_{ih}$ for $h = 1, ..., H$ and $i = 1, ..., I$ via the gene-wise ordinary least-squares approach constrains $\Sigma_h n_h \hat{\gamma}_{ih} = 0$ for $h = 1, ..., H$. With n being the total number of samples from all studies, $\hat{\sigma}_i^2$ is estimated such that $\hat{\sigma}_i^2 = \frac{1}{n}\Sigma_{ij}(Y_{ijh} - \hat{\alpha}_i - \boldsymbol{Z}\hat{\boldsymbol{\beta}}_i - \hat{\gamma}_{ih})^2$. The standardised data X_{ijh} is calculated by

$$X_{ijh} = \frac{Y_{ijh} - \hat{\alpha}_i - \boldsymbol{Z}\hat{\boldsymbol{\beta}}_i}{\hat{\sigma}_i}. \tag{4.11}$$

Assumptions are made regarding the distribution of X_{ijh}, namely $X_{ijh} \sim N(\gamma_{ih}, \delta_{ih}^2)$ and the prior distributions for the batch effect parameters take the form $\gamma_{ih} \sim N(Y_h, \tau_h^2)$ and $\delta_{ih}^2 \sim$ Inverse Gamma(λ_h, θ_h) with these hyperparameters being estimated via the methods of moments empirically from the data. For the batch effect parameters, γ_{ih}^* and δ_{ih}^{2*}, the Empirical Bayes estimates are given by the conditional posterior means and the Empirical Bayes adjusted data γ_{ijh}^* for all i, j and h, can be calculated using these batch effects such that

$$\gamma_{ijh}^* = \frac{\hat{\sigma}_i}{\hat{\delta}_{ih}^*}(X_{ijh} - \hat{\gamma}_{ih}^*) + \hat{\alpha}_i + \boldsymbol{Z}\hat{\boldsymbol{\beta}}_i. \tag{4.12}$$

RUV-2

Remove Unwanted Variation (RUV) in two steps (RUV-2) (Gagnon-Bartsch and Speed, 2011), is a normalisation method that makes use of control genes to attempt to remove the unwanted variation in the data set. Factor analysis is performed just on the control genes, and the resulting factors are modelled within a linear regression model. Factor analysis is the process of obtaining several components, although not directly observed, to capture the unwanted variation. This is achieved through some method such as singular factor decomposition or PCA evaluation, these elements are then modelled along with the other known confounders.

Let the linear model

$$\boldsymbol{X} = \boldsymbol{Y}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{F}\boldsymbol{\alpha} + \epsilon \tag{4.13}$$

be used to model the expression data with n arrays and I genes. $\boldsymbol{X}$ is the expression matrix, let $\boldsymbol{Y}$ be a matrix whose columns are the factors of interest of length Q, for example the class of the samples. The matrix $\boldsymbol{Z}$ has columns which are the observed covariates (study, lab, batch) and $\boldsymbol{F}$ is a matrix where the columns are all unobserved. There are f such unobserved variables, and it is important to select the value of f carefully after factor analysis. The RUV-2 method, and variants, are not the first methods to attempt to use factor analysis for decomposing and modelling variability, this can be seen in for example Leek and Storey (2007) and Listgarten et al. (2010). However, the unique concept in the Gagnon-Bartsch and Speed (2011) approach is the use of only control genes to achieve this purpose.

Approaching the problem using control genes, ensures that too many factors are not included within the modelling process. If too many factors are modelled it is possible to remove from the data set the factor of interest, especially if such an effect is strong and one of the first few factors obtained by factor analysis. Control genes that can be used for this purpose include, for example, housekeeping genes (see Section 2.2.4) or spike-in controls placed by the manufacturer on the microarray platform. The concept behind the use of such genes is that control genes are not affected by the condition of interest, so variation within these genes are sources of unwanted variation due to other effects. Equation 4.13 is restricted to only concern the control genes

$$\boldsymbol{X}_c = \boldsymbol{Y}\boldsymbol{\beta}_c + \boldsymbol{Z}\boldsymbol{\gamma}_c + \boldsymbol{F}\boldsymbol{\alpha}_c + \epsilon_c. \tag{4.14}$$

The assumption is that the control genes do not change throughout the analysis, therefore $\boldsymbol{\beta}_c$ would be a vector of zeros. Equation 4.14 can be reduced even further with the assumption that there are no variables to be modelled within $\boldsymbol{Z}$ (a simplifying assumption, see Gagnon-Bartsch and Speed (2011) for more details). Hence the resulting equation is

$$\boldsymbol{X}_c = \boldsymbol{F}\boldsymbol{\alpha}_c + \epsilon_c \tag{4.15}$$

which leads to an estimate of $\hat{\boldsymbol{F}}$ through factor analysis. To estimate $\hat{\boldsymbol{F}}$, many methods can be used, for example SVD or the EM algorithm. Depending on the data, the number of factors modelled, f is selected. In practice f can be selected based on how well the data cluster (and hence the unwanted variation is removed) and the p-value distribution, how well the model fits the theoretical assumptions.

4.3 Performance assessment

Assessing the performance of different meta- and mega-analysis methods is important to understand and compare methods, this is non-trivial, however. Ideally good mega-analysis normalisation methods should remove external artefacts from the combined data sets (such artefacts include batch, study and other non-biological effects) as well as increase the number of 'discovered' DE genes at particular FDR thresholds (Gagnon-Bartsch and Speed, 2011).

In this thesis, meta- and mega-analysis adjustment will be assessed in six ways including considering the genes selected as DE, ROC curves when data is simulated, error rates in a classification paradigm, hierarchical clustering, raw p-value distribution and observing the behaviours of the control genes.

1. **DE genes**: The aim of meta- and mega-analysis is to increase the distinction between the groups being compared. Observing the number of genes selected as DE after integration allows an assessment of this increased separation. The genes selected as DE are also important. Typically meta- and mega-analysis methods can be evaluated using pre-published gene lists and noticing the concordance of the obtained DE gene list and published material. Observing if the DE genes are expected to be DE (from the positive control list) or expected to be non-DE (from the house-keeping gene list) is indicative of the integrative method's success. This process, however, is subject to publication bias. As a result other methods should be used either as replacement evaluation tools or in tandem to such an approach.

2. **ROC Curves**: For simulated data where the 'true' DE gene list is known, an integrative method's performance can be measured via ROC curves. ROC curves are created by plotting the true positive rates versus the false positive rates for the obtained DE genes. Performance is indicated by how close the plots are to the upper left hand corner of the ROC space. The AUC is also used as a comparison tool, with AUC values close to one indicating an accurate DE list. Some further details on ROC curves and AUC were given in Section 3.3.1.

3. **Classification error rates**: As classification depends heavily on the feature list used to construct the discriminant rule, a classification framework can be used to assess the performance of DE lists, although strictly speaking feature selection and obtaining a DE list are not identical processes. In a classification rule, if all else is held the same (for example the samples and the classifier) the difference in classification error rates can be attributed to the feature list. Hence, if meta- and

mega-analysis methods are used as the feature selection method for a classifier, the discriminatory abilities of the DE lists can be assessed.

4. **Hierarchical clustering**: Sample clustering can be achieved by observing the clustering similarities between samples. Ideally samples will cluster together based on biological factors and not based on non-biological factors. Hierarchical clustering is produced by considering a subset of the genes (often a number of the most variable genes) and observing, based on these genes, which samples are the most similar to each other measured by cluster dissimilarity.

5. **Raw p-value distribution**: P-values should be uniformly distributed across the interval 0–1 under the null hypothesis. A common underlying assumption in most DE analyses of gene expression studies is that the majority of the genes satisfy the null hypothesis: that is no differential expression. Histograms of p-values should hence resemble a horizontal line, perhaps with an inflated number of genes with very small p-values, representing the biologically DE genes.

6. **Control genes**: Control genes, both positive controls and house-keeping genes, are genes with an expected response to the comparison (either to be linked with the condition of interest or not, Section 2.2.4). Considering if such control genes behave as one would expect after integration can help to understand if the adjustment was valid. Observing the location of the house-keeping genes on MA-plots is a visual indication regarding the behaviour of these genes. Genes can also be ranked in order of increasing FDR. A curve representing the increase in FDR can be plotted with the control genes noted. By ranking the genes it is possible to see where each control gene is located. It is an assumption that positive control genes, being linked to the condition of interest would cluster with small FDR values, and house-keeping genes would be located more evenly throughout the complete set of genes, perhaps more heavily represented at the non-DE end of the spectrum.

In the case studies to follow different performance assessment measures are used depending on the context. For case study 1, ROC curves and AUC values determine the success of the meta-analysis methods. For case study 2, a classification context is established and error rates are used to assess the different DE gene lists obtained after meta-analysis. For case study 3, the number of DE genes, clustering, distribution of the p-values and observing the control genes on FDR curves are used as tools for comparing the meta- and mega- analysis methods.

4.4 Application studies

In the following three case studies, different elements of meta- and mega-analysis methods are explored.

4.4.1 Case study 1: Simulation

This simulation study compares the performance of the different meta-analysis methods. These methods include, Fisher's inverse chi-squared method, known as 'Fisher', GeneMeta, RankProd, mDEDS and two different POE methods. For the POE methods, two gene selection methods are used, the IC as well as Bss/Wss, where Bss/Wss is the ratio of the between sum of squares to within sum of squares (Dudoit and Fridlyand, 2003). Distinction will be made between them using the terms POE_{IC} and $\text{POE}_{\text{Bss/Wss}}$ to indicate what type of analysis was performed after the construction of the POE matrix.

Data was simulated to represent three separate gene expression data sets. The simulation approach is adapted from Ritchie et al. (2006). A non-parametric bootstrap simulation is used to generate a matrix of non-differentially expressed genes. Samples are constructed with replacement from the original data, such that a balanced binary class distribution is established. Within the simulation it is assumed that the expression data is sampled from three different data sets, so three such matrices are generated. These matrices are used as an underlying 'background' data set with DE genes imposed on top of them. This background noise contains the latent characteristics of an actual microarray data with no biologically DE genes.

DE genes are simulated with a 2-fold increase in fold change. Two types of DE genes are simulated: (i) 'true' DE genes, and (ii) 'platform specific' DE genes. True DE genes are the same genes within each of the three generated data sets, representing biologically relevant DE genes. Platform specific DE genes simulate platform bias, apparent within DE genes from microarray experiments (Bosotti et al., 2007). These genes are randomly selected from all the genes in the data sets, with the exclusion of the true DE genes, and are generated independently for each data set. This simulation taps into the important property that a powerful meta-analysis tool has the ability to correctly distinguish a true DE gene which is DE across multiple platforms from a DE gene which is simply a platform phenomenon.

Nine data sets were simulated, with the percentage of DE genes changing, varying between 2.5%, 4% and 10% (with three data sets at each percentage level). For each data set, half of the DE genes were true DE genes (and hence the same for all the data sets at the same percentage level) and the other half of the DE genes were platform

Meta-analysis method	AUC values at different percentages of DE genes		
	2.5%	4%	10%
RankProd	0.999	0.998	0.995
mDEDS	0.998	0.998	0.994
Fisher	0.996	0.993	0.982
GeneMeta	0.861	0.866	0.876
POE_{IC}	0.483	0.492	0.491
$POE_{BSS/WSS}$	0.489	0.490	0.487

TABLE 4.1: AUC values for the different meta-analysis methods in the simulation comparison, ranked is order of AUC values for the 10% DE genes simulation.

specific (hence different for all the data sets at the same percentage level). The three data sets at each percentage level had a different number of simulated samples, 150, 100 and 80 samples respectively, each with 20,000 genes. The data sets at the same DE percentage level were analysed together as a simulated meta-analysis.

Results Figure 4.1 shows the ROC curves for the 10% DE gene level, (5% true and 5% platform specific DE genes). Only this study is presented here as these results are indicative of all considered DE percentages, the remaining results are shown in Appendix C. An ideal ROC curve is as close to the top left-hand corner as possible. This plot highlights that the two POE methods ($POE_{Bss/Wss}$ and POE_{IC}) are struggling to calculate the true DE gene list. Table 4.1 contains the AUC values for all three different DE gene percentage levels for the different meta-analysis methods. An AUC value of 0.5 corresponds to a completely random model. $POE_{Bss/Wss}$ and POE_{IC} appear to continue to produce low AUC values for all percentages. GeneMeta and RankProd perform adequately. Interestingly RankProd decreases in accuracy as the percentage of DE genes in the simulated data increases. Fisher and mDEDS perform competitively with accuracy for both decreasing slightly as the percentage of DE genes increases from 2.5% to 10%.

4.4.2 Case study 2: Melanoma study

Three melanoma data sets are used for the examination of the meta-analysis methods in a classification context. The Bogunovic, Jönsson and Mann data sets were selected, as introduced in Sections 2.3.1 and 2.3.2. Two of the selected data sets are generated from the Illumina Human Beadarrays platform v2 and v3 (Jönsson data and Mann data respectively) and the third is from an Affymetrix platform, HG 133 Plus 2.0 (Bogunovic data). All three methods have clear (albeit slightly different) class distinctions between

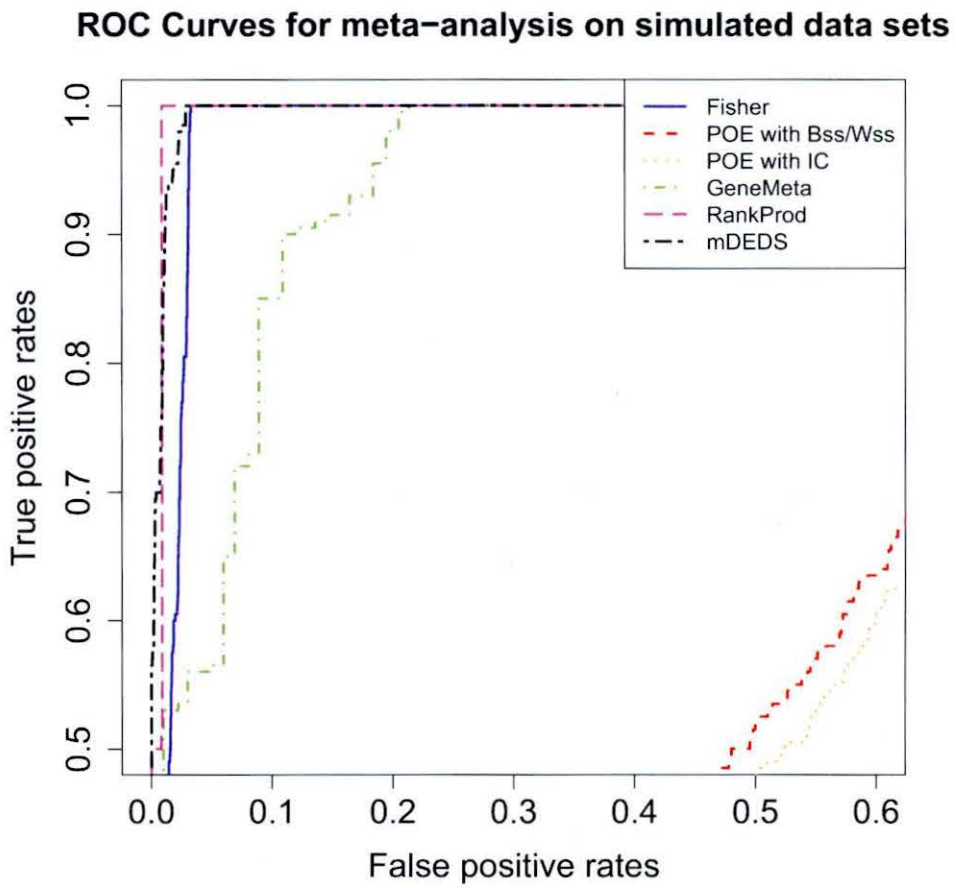


FIGURE 4.1: ROC plots for simulated data using different meta-analysis methods for the 10% DE gene level (5% true, 5% platform specific DE genes) simulation.

good and bad prognosis groups. The Mann data and the Jönsson data were integrated together using the meta-analysis methods: Cross-validation, Fisher, GeneMeta, $\text{POE}_{\text{Bss/Wss}}$, POE_{IC}, mDEDS, and RankProd. In the case of the cross-validation meta-analysis method genes were ranked based on Bss/Wss (all other meta-analysis methods produce a ranked gene list). A discriminant rule was constructed using support vector machines (SVM) (Hastie et al., 2009) on a ranked subset of these genes. The subset ranged from the top 10 ranked genes to 500 genes in increments of 10. This discriminant rule (built from the Jönsson and Mann data) was then used to classify the independent Bogunovic data set.

Results Figure 4.2 displays the error rates for the classification of the Bogunivic data set using SVM. In this analysis the number of genes used to build the classification rule varies from 10 to 500, as seen on the horizontal axis of the graph. The meta-analysis methods are split into two separate plots for readability. The mean and minimum error rates for each of the meta-analysis methods can be found in Table 4.2. The majority of

Meta-analysis method	Mean error rate	Min error rate
mDEDS	0.2655	0.1212
RankProd	0.3055	0.1515
Fisher	0.4806	0.2121
GeneMeta	0.3709	0.2727
$POE_{Bss/Wss}$	0.3327	0.2727
POE_{IC}	0.4842	0.2727
Cross-validation	0.3467	0.2424

TABLE 4.2: Mean and minimum error rates for SVM using LOOCV when the Bogunivic data set is classified using a gene list obtained via meta-analysis from the Jönsson and Mann data sets.

FIGURE 4.2: LOOCV error rates as the number of genes increases from 10 to 500, when the Bogunovic data set is classified using a gene list obtained via meta-analysis from the Jönsson and Mann data sets, with discriminant rule constructed via SVM.

the applied meta-analysis methods successfully capture discriminating DE genes across the two data sets (the Jönsson and Mann data), to distinguish between good and bad survival prognoses. The Fisher and POE_{IC} meta-analysis methods performed the most poorly with mean classification error just below 50% (48.06% and 48.42% respectively).

4.4.3 Case study 3: Hypertension study - DE analysis of hypertensive versus normotensive rat samples

The data used in this case study comes from four independent rat data sets, each observing hypertensive and normotensive animals. These data sets come from two different Affymetrix platforms, the Cerutti data (10 hypertensive, 5 normotensive), the Clemitson data (5 hypertensive, 5 normotensive) and the Grayson data (3 hypertensive and 3 normotensive) all come from GeneChip Rat Genome 230, the Rysä data set (12 hypertensive, 11 normotensive) on the other hand comes from Affymetrix GeneChip Rat Genome U34 Array set A. More detail regarding the individual data sets can be found in Section 2.2. Data has been preprocessed and reduced to the 4,678 intersecting genes identified via their Entrez IDs. Only using the intersecting genes allows results from individual analysis to be compared between the two platforms.

The purpose of this case study is to consider how meta- and mega-analysis can be used to integrate data addressing the same scientific question and how such integration methods can be informative over individual analysis especially when inconsistencies are evident between individual results. The individual analyses for the four data sets is presented in Section 2.2.3.

This case study makes use of two important types of genes: (i) positive control genes, these genes are known to have a relationship to the condition of interest; and (ii) house-keeping genes, genes selected because of the assumption that they do not vary with regard to the condition of interest. More details regarding each of these gene lists are shown in Section 2.2.4.

Mega-analysis results

Four currently available mega-analysis methods were used to combine the four public hypertensive/normotensive rat data sets. Once this normalisation was performed a large expression data set was obtained and downstream analysis proceeded as in the case of an individual study. In each case, data was modelled using least squares regression applied through `limma` (Smyth, 2004). The four mega-analysis methods included:

1. Null correction, where data was simply combined together and modelled using a class effect.
2. Quantile normalisation, where data was combined and then quantile corrected and modelling with a class and platform effect.
3. ComBat, with each data set considered as an independent batch, that is $h = k$.
4. RUV-2, with the EM approach used for factor analysis. House-keeping genes (Section 2.2.4) are the control genes used to perform the factor analysis within this case study. Although a range of different genes are possible, this particular set of genes were selected because of the different platforms being combined throughout the mega-analysis comparison. The number of unknown factors, f, modelled into the data was four, with one factor of interest, class.

Several methods (Section 4.3) were used to evaluate these mega-method normalisation approaches, first to evaluate if the adjustments were helpful to the analysis and second to compare methods to one another.

DE genes It is desirable that the number of genes 'discovered' as DE increases as the distinction between the two factors of interest (in this case class) increases. The mega-analysis normalisations are designed to increase this distinction, hence the number of DE genes should increase. Table 4.3 shows the number of DE genes from the complete gene lists and Table 4.4 shows the number of DE genes from the positive control list. Three different DE criteria are used. The first being genes that have an absolute FC greater than 1.5, the second that the FDR is less than 0.05 and the third being that both

Mega-analysis	$\|FC\| \geq 1.5$ (up, down)	FDR < 0.05	$\|FC\| \geq 1.5$ (up, down) **and** FDR < 0.05
Null correction	41 (35, 6)	5	3 (0, 3)
Quantile normalisation	26 (17, 9)	238	21 (13, 8)
ComBat	20 (12, 8)	841	20 (12,8)
RUV-2	22 (18, 4)	761	22 (18, 4)

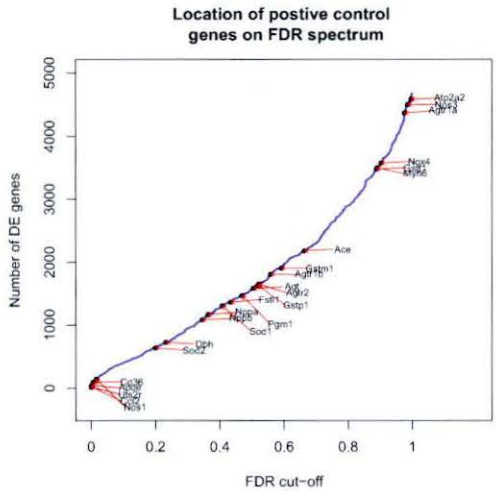
TABLE 4.3: Number of DE genes after the four data sets were integrated using the four mega-analysis methods.

Mega-analysis method	Positive control genes		Inconsistent genes	
	$\|FC\| \geq 1.5$ (up, down)	FDR < 0.05	$\|FC\| \geq 1.5$ (up, down)	FDR < 0.05
Null correction	2 (2,0)	0	1(1,0)	0
Quantile normalisation	2 (1, 1)	3	0	1
ComBat	1(0,1)	8	0	5
RUV-2	2(1,1)	8	0	3

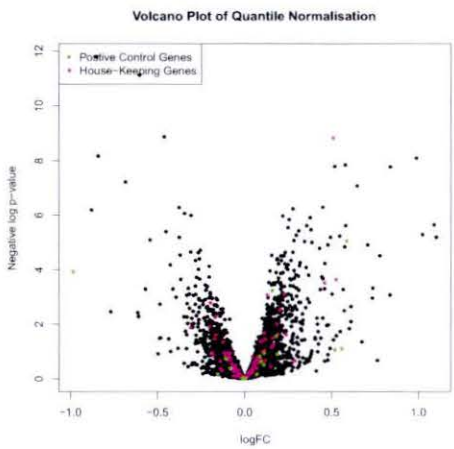
TABLE 4.4: The number of positive control genes and inconsistent genes, within the DE gene lists for each mega-analysis method.

these criteria are satisfied. The more sophisticated mega-analysis methods (ComBat and RUV-2) result in more DE genes, both overall and in the case of the positive control list. Figure 4.3 shows the number of genes selected as DE as the cut-off for the FDR value varies from 0 to 1. Highlighted in these plots is what the required FDR cut-off values need to be for the positive control genes to be DE. These plots attempt to address if a mega-analysis method is successful in consistently placing positive control genes, which have been externally validated, toward the more significant end of the FDR spectrum. ComBat and RUV-2 adjustments seem to produce small FDR values in the positive control genes. Due to the p-value violations made by the Null correction adjustment method, see Figure 4.4 (b), an FDR curve has not been plotted as it produces very limited information. Figure 4.3 also contains volcano plots for the three mega-analysis methods. These plots show a relationship between p-values and FC. Ideally the positive control genes (green) should be plotted higher in the graph and the pink house-keeping genes should be lower with smaller $-\log_{10}$(p-value) and FC scores. This is not the case for this analysis, suggesting that the positive control genes are not highly informative and there are potential DE changes in the house-keeping genes.

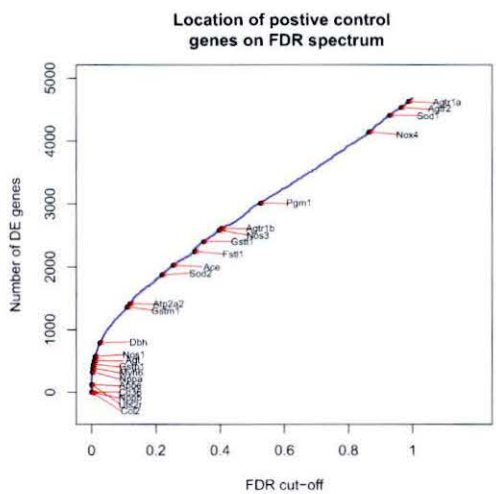
Hierarchical clustering Observing sample clustering allows the observation of the driving source of variation within a data set. Hierarchical clustering of each of the four methods is plotted in Figure 4.4 (for the Null and Quantile correction methods) and



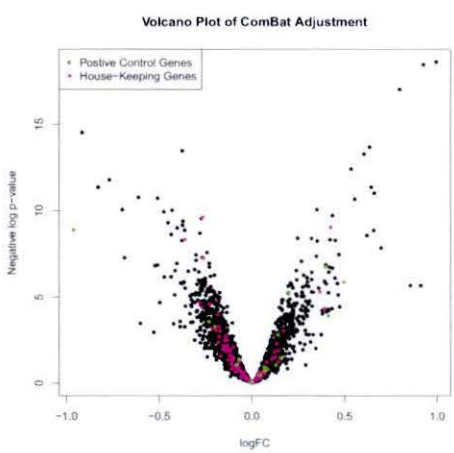
(a) Quantile normalisation varying FDR cut-off value



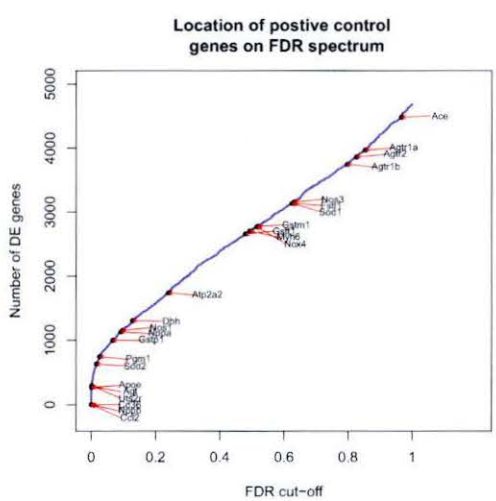
(b) Quantile normalisation Volcano plot



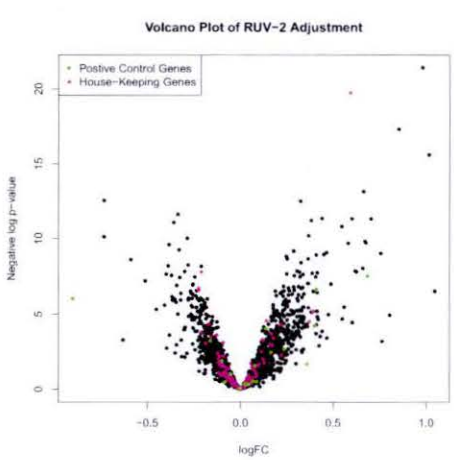
(c) ComBat adjustment varying FDR cut-off value



(d) ComBat adjustment Volcano plot



(e) RUV-2 adjustment varying FDR cut-off value



(f) RUV-2 adjustment Volcano plot

FIGURE 4.3: **(a, c, e)** Number of DE genes when the FDR p-value cut-off is varied, highlighted on each plot is when particular positive control genes become DE. **(b, d, f)** Volcano plots for mega-analysis methods representing the relationship between p-values and FC. Positive control genes are plotted in green and house-keeping genes are plotted in pink. These plot are constructed for the mega-analysis methods Quantile correction, ComBat and RUV-2 adjustment

Figure 4.5 (for the ComBat and RUV-2 methods). Cluster plots were drawn using the top 500 overall most variable genes using Euclidean distance and the complete agglomeration method. The coloured bars beneath each of the dendrograms highlight features of each of the samples. The first two bars are of special interest as these represent the sample class (blue = normotensive, red = hypertensive) and the platform (light blue = Affymetrix 230A chip, purple = Affymetrix U34A chip) respectively. For the Null correction and the Quantile normalisation all samples within the *platform* cluster together, this indicates that the largest source of variation within this analysis is the platform effect. Conversely, for ComBat and RUV-2, platform is no-longer the dominating source of variation but instead it is *class* (seen much more clearly using the RUV-2 approach). The purpose of the integration is to perform a DE analysis based on class (genes that are DE between hypertensive and normotensive rats) and as a result the largest source of desired variation after mega-normalisation should be the class factor.

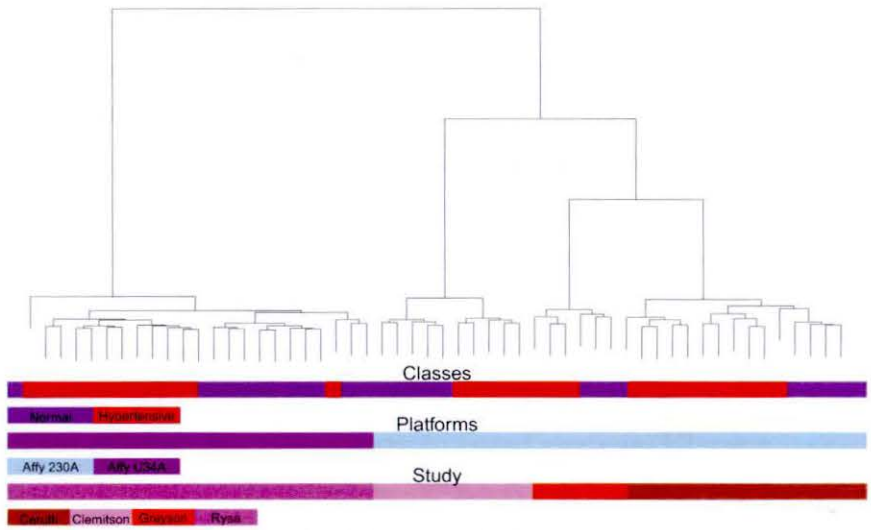
Raw p-value distribution The distribution of the raw p-values after mega-analysis indicates whether assumptions regarding independence and the null hypothesis of limited differential expression for the majority of genes has been violated. Figures 4.4 and 4.5 include histograms of the raw p-values of the combined data sets after each of the mega-analysis normalisation methods have been performed. Ideally the p-value distribution should be uniform, with a possible spike in frequency at the low end of the spectrum indicating the DE genes. Quantile normalisation and the RUV-2 adjustment are closest to the uniform distribution. The Quantile normalisation method yields overall fewer genes with extremely low p-values, implying that the uniform frequency across the entire distribution is higher than that of the RUV-2 method. P-values produced after Null correction do not conform with the assumptions required for a reliable DE analysis.

Meta-analysis results

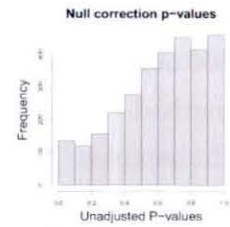
The four public data sets were analysed independently and four of the meta-analysis methods were used to integrate these results. The meta-analysis methods used include:

1. Fisher's Inverse Chi-squared method ('Fisher'),
2. GeneMeta,
3. mDEDS and
4. RankProd.

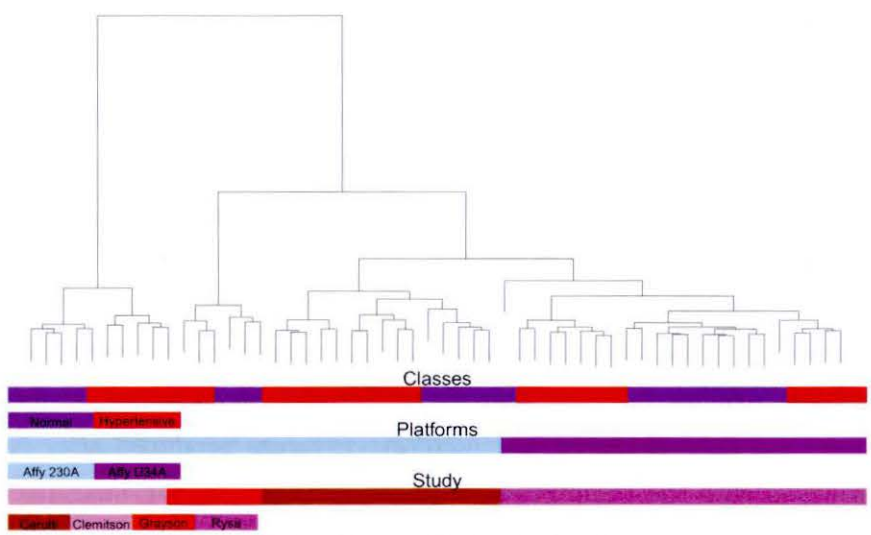
The number of genes found to be DE and the results from the positive control genes were observed during the analysis.



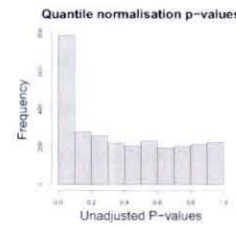
(a) Null correction cluster plot



(b) Null correction p-value distribution



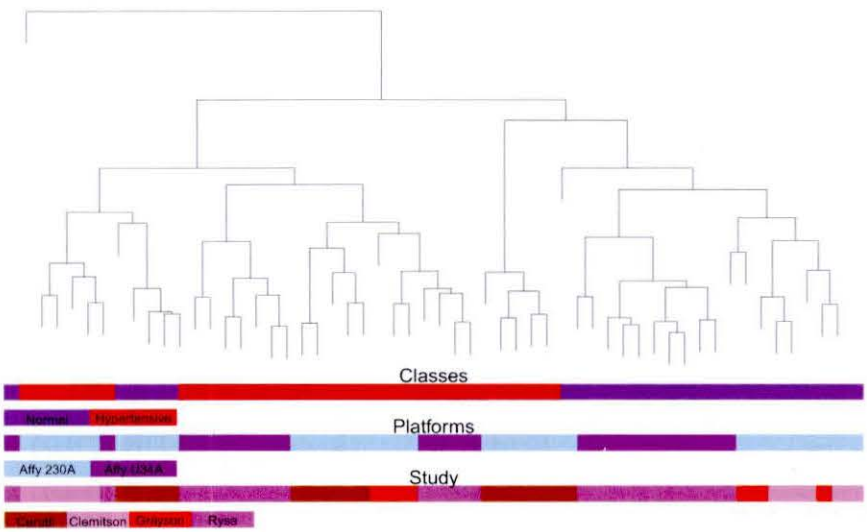
(c) Quantile normalisation cluster plot



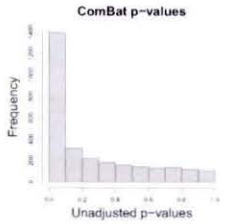
(d) Quantile normalisation p-value distribution

FIGURE 4.4: Hierarchical cluster plots and p-value histograms of Null correction and Quantile normalisation mega-analysis normalisation methods.

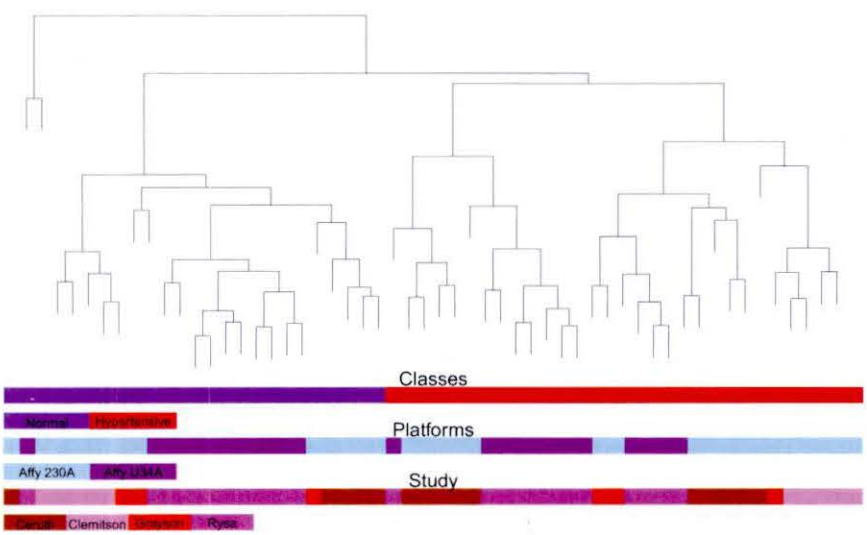
DE genes The number of DE genes found from the complete gene lists, for the different meta-analysis methods are reported in Table 4.5. Fisher and RankProd select a large number of DE genes when FDR is used as the selection criterion. RankProd is the only method able to produce FC results. When considering only the positive control or the inconsistent genes (Table 4.6), for the Fisher and RankProd methods there are a large proportion of these genes selected as DE. That is, for the Fisher method 7/28 (25%) of the positive control are identified as DE (using FDR < 0.05) and 8/11 (73%) of the inconsistent genes were selected as DE. For the RankProd method 8/28 (29%) of the positive controls are considered DE and 9/11 (82%) of the inconsistent genes



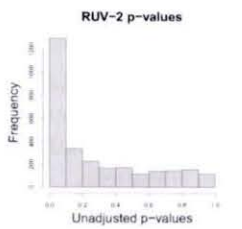
(a) ComBat adjustment cluster plot



(b) ComBat adjustment p-value distribution



(c) RUV-2 adjustment cluster plot



(d) RUV-2 adjustment p-value distribution

FIGURE 4.5: Hierarchical cluster plots and p-value histograms of ComBat adjustment and RUV-2 adjustment mega-analysis normalisation methods.

are considered DE. Both GeneMeta and mDEDS struggle to produce DE genes based on p-values. Both these methods are intended to be gene ranking methods, however permutations were produced to generate p-values for comparative purposes. By construction, GeneMeta and mDEDS produce very sparse low FDR estimates resulting in their inability to produce DE genes based on the FDR DE criterion.

Control genes Figure 4.6 contains plots of the number of genes to be considered DE as the FDR cut-off increases from 0 to 1. Figure 4.6 (a) and (b) highlight when the *positive control genes* become DE for the Fisher and RankProd methods. Due to the

Meta-analysis	$\|FC\| \geq 1.5$ (up, down)	FDR < 0.05	$\|FC\| \geq 1.5$ (up, down) **and** FDR < 0.05
Fisher	NA	605	NA
GeneMeta	NA	40	NA
RankProd	21(10, 11)	785	0
mDEDS	NA	54	NA

TABLE 4.5: Number of DE genes for each data set, using meta-analysis methods. Note: Fisher, GeneMeta and mDEDS do not produce FC values.

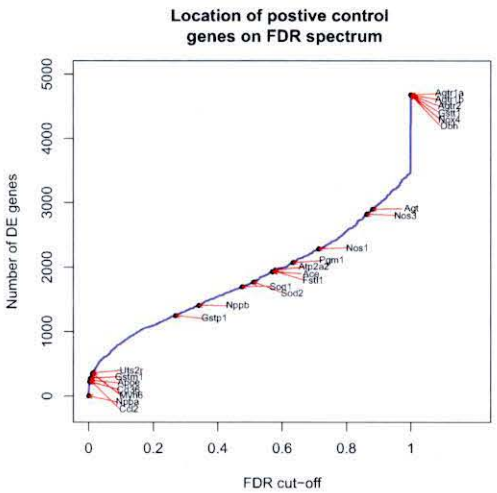
Meta-analysis method	Positive control genes		Inconsistent genes	
	$\|FC\| \geq 1.5$ (up, down)	FDR < 0.05	$\|FC\| \geq 1.5$ (up, down)	FDR < 0.05
Fisher	NA	7	NA	8
GeneMeta	NA	0	NA	0
RankProd	1(0, 1)	8	0	9
mDEDS	NA	0	NA	1

TABLE 4.6: The number of DE genes for each data set which is also considered a positive controls for hypertension, or an inconsistent gene when analysed using meta-analysis methods. Note: Fisher, GeneMeta and mDEDS do not produce FC values.

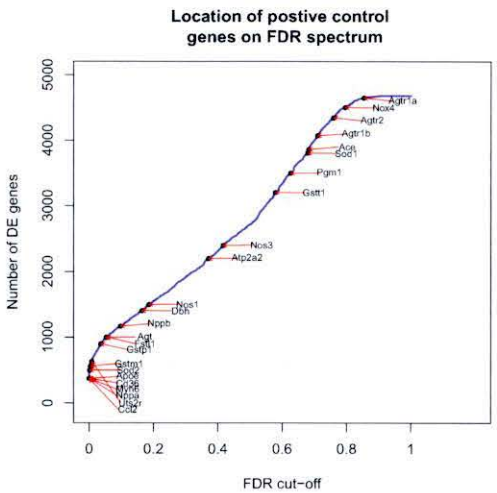
design of this gene list, ideally the positive control genes should become DE toward the lower end of the FDR spectrum. Although for the Fisher and RankProd methods there appears to be a small cluster of these genes with a low FDR value, both these methods overall rank the positive control genes throughout the entire distribution. Figure 4.6 (c) and (d) highlight when the *inconsistent genes* become DE, for the Fisher and RankProd methods, respectively. For both these methods the inconsistent genes appear in the first half of the spectrum perhaps reflecting that all these genes are DE in their independent studies, just with differing FC directions. It is also possible that the meta-analysis methods are inadequately approaching such discrepancies. This would be especially true in the case of the Fisher meta-analysis method which implements the aggregation of p-values, not the consideration of FC direction.

4.5 Discussion

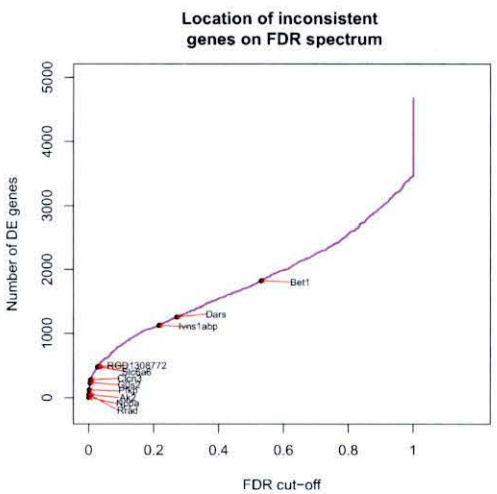
Meta- and mega-analysis methods have been explored in the three case studies considered in this chapter. Both levels of integration approaches have distinct advantages and disadvantages.



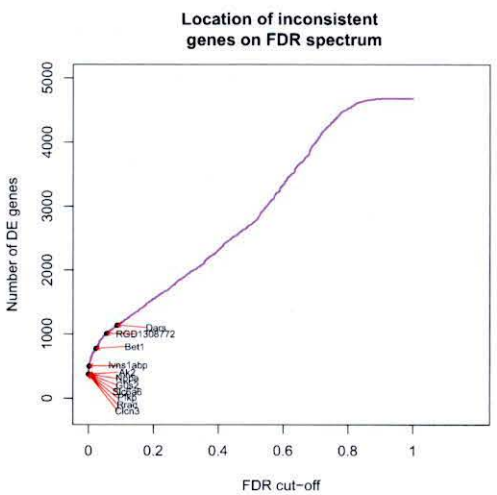
(a) Fisher with positive control genes highlighted



(b) RankProd with positive control genes highlighted



(c) Fisher with inconsistent genes highlighted



(d) RankProd with inconsistent genes highlighted

FIGURE 4.6: The control genes as the FDR p-value cut-off is varied. These plots highlight when each of the positive control genes or inconsistent genes become DE for the Fisher's inverse chi-squared and RankProd meta-analysis methods.

Meta-analysis methods analyse data sets separately and integrate the resulting statistics. A meta-analysis approach over a mega-analysis approach allows for several advantages. First, the raw results of a microarray experiment need not be known for some of the meta-analysis methods and, second, data sets can be vastly different in the technologies used to create microarrays which result in vastly different ranges and variances (Guerra et al., 2008). Moreover, when studies are extremely different in regards to purpose, careful meta-analysis techniques may allow a new, independent question to be considered (Campain et al., 2010). As data sets are treated separately, such analysis is susceptible to small data set issues such as minimal power for each data set to detect DE genes. Independent data set analysis is also susceptible to outliers, where effects are more extreme

when sample sizes are small. When considering mega-analysis, the major advantage of mega-analysis over meta-analysis is that a large data set is constructed, and from this downstream analysis is applied. In this way the combined large data set avoids small sample issues. But to apply mega-analysis one must have access to the raw (if possible) or processed expression data, not just resulting statistics.

Performance validation The simulation study (case study 1) coupled with the melanoma case study (case study 2) offers insight into the seven meta-analysis methods compared in this chapter. It is important to validate meta-analysis methods, although at times this is difficult to perform. Some meta-methods are simple variants of common classical statistical methods, others offer more sophisticated responses to specific issues faced in the microarray environment. A large proportion of meta-research deals with DE genes and the process of obtaining a DE list from multiple data sets. Unfortunately DE gene lists are elusive because the true biological DE gene lists are typically not known. Often for validation purposes DE lists are compared to other published DE lists with the level of congruency indicative of the success of the meta-method. This approach suffers from publication bias (Dudoit and Fridlyand, 2003) because continuously pre-published information is being published, with little validation to the variations that are occuring. An alternative assessment criteria, utilising the classification framework, offers an intuitive validation process with interpretable results. Classification performance relies heavily on the accuracy of the classifier's feature list, which is traditionally taken from the DE list.

In case study 2 meta-analysis validation was performed using SVM classification. SVM was chosen as it is an efficient and accurate classifier for microarray data and in recent years is gaining popularity. This study could have been conducted using any number of classifiers provided feature selection is not performed implicitly by the classifier. The varying DE list obtained from the meta-methods are the only varying component in the comparison. Therefore a reduction in classification error can be attributed to the meta-analysis method.

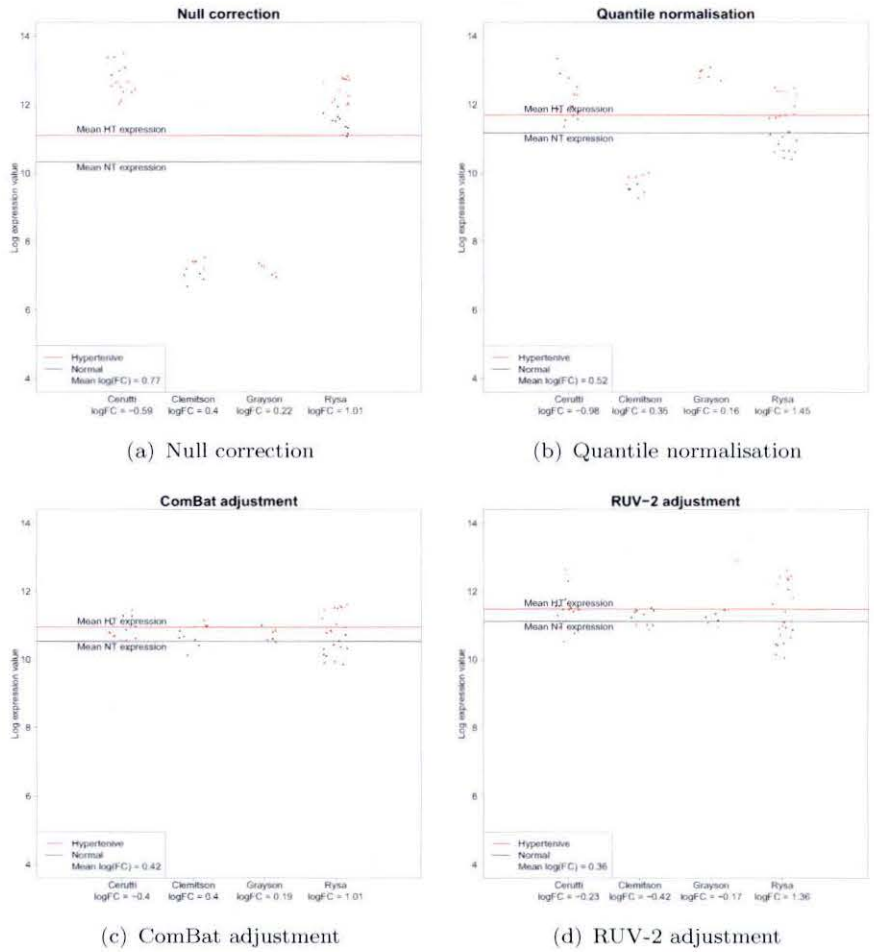
Within case study 3 mega-analysis methods have been compared via several methods. Hierarchical clustering was used to see if the mega-normalisations have allowed the factors of interest to be the most distinctly variable elements within the data set. The distributions of raw p-values have been observed to ensure that methods are not violating distributional and independence assumptions. The number of 'discovered' genes for each of the mega-analysis methods was also considered, under the assumption that as the true factor of interest becomes more clear the number of DE genes will increase. Alongside

these comparison methods results for control genes, both positive controls and housekeeping genes have been monitored. Genes with expected behaviours, be it that they are linked with the condition or they have no change, should still behave in the expected way regardless of the data-manipulation method applied.

The validation methods here are by no means faultless and as the number of meta- and mega-analysis methods increases it is paramount that validation methods are also developed. Heuristically, any form of integrative adjustment needs to increase the number of DE genes truly associated with the condition of interest, and decrease the false positives so that research into linked genes can take place. This process is made increasingly difficult in that the genes truly associated with the condition of interest are still unknown.

Expression unification Mega-analysis seeks to unify data sets, and for each gene to give an indication as to the expression differences between classes over all the combined data sets. Figure 4.7 is a scatter plot of the expression values of the inconsistent and positive control gene Nppa. In this plot it can be seen what mega-analysis normalisation achieves based on the expression levels of individual samples. Originally, each data set was reporting extremely different expression levels for this gene, these results were extreme enough to flag this gene as inconsistent. This difference is seen in particular between results from the Cerutti and Rysä analyses. As the mega-analysis normalisation methods become more sophisticated (for example in the case of ComBat and RUV-2) the differences in the data sets are removed yet the signal in the data is not removed completely. However, based on FC alone, after integration, this gene is no longer DE.

Comparison to individual analysis Figure 4.8 contains two heatmaps of average FC for the positive control genes and the inconsistent genes for (a) the individual analysis and (b) the FC values after mega-analysis normalisation. Genes that have a negative FC are plotted in green with genes with a positive FC are plotted in red, a black value represents no difference between the two conditions. In the inconsistent gene cases a gene is considered a DE gene but the FC direction changes between the different studies. Mega-analysis unifies these inconsistencies by giving an overall value. For some of the positive control genes the average FC signal is increased (for example the Cd36 case) after mega-analysis normalisation. This change highlights that after adjustment there is an increased distinction between classes. One might assume that mega-analysis adjustments are just weighted averages of the expression values for the individual data sets for the inconsistent genes. But the differences in expression values between the different mega-analysis methods for genes such as G0s2 and Rrad would suggest that this is perhaps not the case.



(a) Null correction

(b) Quantile normalisation

(c) ComBat adjustment

(d) RUV-2 adjustment

FIGURE 4.7: Expression plots after mega-analysis adjustment is applied, for gene Nppa. In the more sophisticated and successful mega-analysis normalisation methods, the inconsistencies within the data sets is no longer a dominating factor.

Application of meta- and mega-analysis The integration of data sets offers a way to enhance the robustness of microarray technology. The 'data set cross-validation' meta-analysis approach observed in this chapter encapsulates a very real problem with microarrays; gene lists selected from one platform or study have a limited ability to be transfered. This is highlighted by their inability to be used to classify samples generated by another platform or data set, as demonstrated by the 34.6% error rate obtained via this method (Table 4.2). For the melanoma study, some meta-analysis approaches were able to increase the accuracy of cross platform classification when compared to this naïve method, at times the error reduced by near 10% as shown in Table 4.2. This indicates that the added power through more sophisticated meta-analysis methods produces more robust and reliable results, eventuating in a gene list that is not platform dependent but truly indicative of the disease.

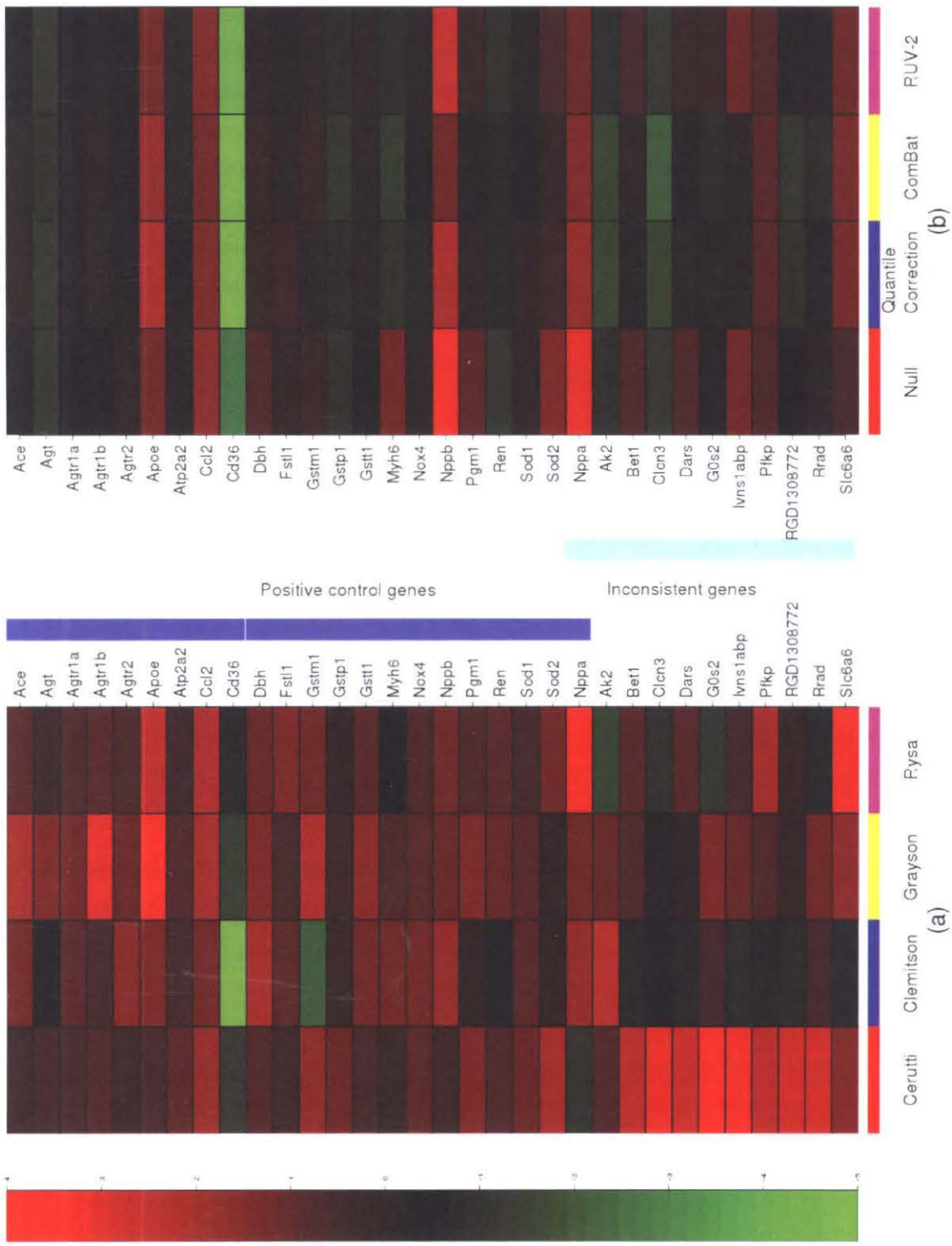


FIGURE 4.8: Heatmaps of FC values for the positive control genes and the inconsistent genes. Heatmaps are produced for (a) the individual studies and (b) values after mega-analysis normalisation. Red values indicate a positive FC (that is the gene in question is more highly expressed in the hypertensive samples then the normotensive samples).

Cross-platform meta-analysis Cross platform meta-analysis multiplies the level of complexity in this particular analysis paradigm, as observed in Campain and Yang (2010). The meta-analysis complexity is suggestive of the meta-method one should employ. In Campain and Yang (2010) two levels of meta-analysis complexity were considered: (i) when meta-analysis is performed across similar platforms, for example Affymetrix with Affymetrix; (ii) when meta-analysis is performed across disparate platforms, for example Affymetrix with oligo arrays. It was shown that the mDEDS method was able to behave competitively in both situations. The melanoma meta-method comparison has confirmed results, in that mDEDS performed well in a cross-platform classification context between two Illumina beadarray data sets and an Affymetrix data set.

Both long-oligo and beadarrays are compared. These platforms vary remarkably with differences ranging from probe length to construction. In this high complexity environment, POE_{IC}, GeneMeta and Fisher's inverse chi-squared method struggle to obtain a gene list robust enough for cross platform classification. Two different reasons could contribute to the decrease in accuracy of the meta-analysis methods as the level of complexity increases. The meta-analysis methods could be over-fitting the data, methods that model the data are particularly susceptible to this, for example GeneMeta. Conversely, some feature selection methods may not capture the complexity of the data, this is potentially occuring in the POE_{IC} case. The Fisher's inverse chi-squared meta approach does not take into consideration the actual intensities of each spot on the microarray, albeit at times this method is ideal, for example when individual intensities are unknown, or when the characteristics of the study vary greatly (Guerra et al., 2008).

Within such a complex environment mDEDS is able to perform DE analysis well, as this method makes use of the different data sets but does not try to fit a full parametric model to the data. The mDEDS method uses multiple statistical measures while developing its ordered gene list. Using multiple measures aids robustness as more of the variability can be encapsulated within the meta-analysis method. It is possible that the multiple platforms and multiple measures draw enough diversity to begin to transcend cross platform variability and produce a reliable gene list. The variation in some of the meta-analysis method's abilities within classification suggests that different tools are beneficial depending on the meta-analysis project.

Batch correction Batch effects are considered non-biological differences that make samples in different batches of microarray development not directly comparable. Such effects are inevitable when additional samples or replicates are added to array data sets or when multiple studies are being combined or integrated together, pooling across different labs, array types or platforms (Rhodes et al., 2004). In most cases batch

effects are inevitable as non-biological variations are observed simply through multiple, apparently identical, amplification and hybridisation. Over time the average size of microarray studies has increased as the cost of such experiments decreased and a greater appreciation of the variability within the microarray studies was gained. This implies that for a number of studies the time taken to produce the data has increased. As a result within studies temporal, spatial and other artefacts has increased. These are collectively known as batch effects (Yang et al., 2011). Batch effects can completely overshadow the effect of interest and confound the DE genes. As a result, powerful batch correction methods are vital for microarray research. Batches obtained separately with time delays, for example a year, can be considered as separate batches, which resemble individual data sets on similar platforms.

It is possible to speculate that mDEDS can be used in a batch correction context. By using mDEDS one can borrow strength from the multiple batches yet avoid particular batch bias. Mega-analysis can also be used as a method for normalising batch effects, where each batch is treated as a separate study.

Open questions There are still many open questions regarding the integration of expression data sets. For example, questions pertaining to mismatched probe sets across platforms and the handling of multiple probes for the same genes. More research within these areas would greatly aid the integration of microarrays and increase the ability to make use of the current plethora of information laying dormant in these public repositories. However, once more of these types of tools for integration have been developed, meta-analysis will save time, money and scientific resources.

There are several issues needing to be resolved each time such an analysis is applied other than that of method validation. Quality of the individual data sets is paramount. Each of the data sets studied here have undergone individual QC (Section 2.2 and Appendix A, Figures A.1-A.4). Ramasamy et al. (2008) gives an overview of the elements and conditions to consider when selecting data sets to integrate. When quality is compromised with this level of the analysis, informative integrated results after mega-analysis normalisation can not be expected. Publicly available data sets need to be coupled with factors of interest relating to the expression samples. Too often data sets are available with only limited information regarding the obvious factor of class, but also less direct factors such as batch and development sequence. Such information can aid in the normalisation process, as attempts are made to remove such effects through mega-analysis normalisation.

4.6 Conclusion

Individual gene expression microarray analysis has been around for almost 15 years. Methods of analysing these single experiments is well established. There is an increased need for methods of data integration and normalisation to ensure as much information is acquired from this data as possible. This need is driven by several factors including; (i) the increase in the amount of microarray data now available in public repositories and (ii) the increased size of individual experiments, introducing batch and time effects into single experiments.

This chapter examined some of the solutions to data integration in gene expression microarray analysis. Two main approaches were considered, a high level integration where data is combined at the statistics level, termed *meta-analysis*, and a lower level integration where data is combined at the rawest level available into a large data set termed *mega-analysis*. This chapter has begun to untangle some of the methods of integrating, both in a mega- or a meta- sense depending on the research context. Establishing the 'best' meta- or mega-analysis method is cumbersome, first because the methods studied in this thesis are not exhaustive and second because circumstances shape the meaning and usefulness of an optimal method. Mega-analysis offers the overarching advantage, as after mega-analysis normalisation data can be analysed as one would analyse an individual data set, hence leading to increased power, intuitive methods and conclusions. In contrast meta-analysis combines statistics such as p-values or test-statistics, and offers a convenient solution when the raw data is not available, or studies are not directly comparable.

New methods of data integration will continue to be developed. In particular cross-study normalisation and batch effect elimination are becoming readily accessible tools with different approaches being designed to address particular research questions. Further growth in this area is expected in coming years. To this end, development into the evaluation methods of the different meta- and mega-analysis methods needs to be an active area of research so that the most effective method is used appropriately. With appropriate and validated meta- and mega-analysis methods, along with adequate quality control of the individual data sets, appropriate statistical research can begin to make use of this wealth of information in the public domain.

4.7 Publications

This chapter includes work published or accepted in Campain and Yang (2010), Campain et al. (2010), Marques et al. (2011a), Marques et al. (2011b) and work under review in

Yang et al. (2011) and Marques et al. (2011c). Some of the meta- and mega-analysis work was conducted by the author with Professor Terry Speed, Department of Statistics University of California at Berkeley and the Walter and Eliza Hall Institute of Medical Research Melbourne, and the hypertension analysis was performed with Professor Brian Morris, Basic and Clinical Genomic Laboratory, School of Medical Sciences and Bosch Institute, University of Sydney, working in particular with Ms Francine Marques.

Publications

- **A.E. Campain** and Y.H. Yang (2010) Comparison study of microarray meta-analysis methods, **11**:408 *BMC Bioinformatics.*
- **A.E. Campain**, F.Z. Marques, Y.H. Yang and B.J. Morris (2010) Meta-analysis of genome-wide gene expression differences in onset and maintenance phase of genetic hypertension, **56**:319–324 *Hypertension.*
- F.Z. Marques, **A.E Campain** P.J Davern, Y.H Yang, G.A Head and B.J. Morris (2011) Genes influencing circadian differences in blood pressure in hypertensive mice, **6**:4 e19203 *PLoS One.*
- F.Z. Marques, **A.E. Campain**, P.J Davern, Y.H Yang, G.A Head and B.J. Morris (2011) Global identification of the genes and pathways differentially expressed in hypothalamus in early and established neurogenic hypertension, **43**:766–771 *Physiological Genomics.*
- F.Z. Marques, **A.E. Campain**, E. Zukowska-Szczechowska, M. Tomaszewski, Y.H Yang, F.J. Charchar and B.J Morris (2011) Gene expression profiling reveals renin mRNA overexpression in human hypertensive kidneys and a role for microRNAs, **58**:1093–8, *Hypertension.*

Manuscripts under review

- Y.H. Yang, **A.E. Campain** and T.P. Speed (2011) Finding differentially expressed genes in microarray data. Under review, Preprint.
- F.Z. Marques, **A.E Campain** P.J Davern, Y.H Yang, G.A Head and B.J. Morris (2011) Genes influencing circadian differences in blood pressure in hypertensive mice, **6**:4 e19203 *PLoS One.*

Chapter 5

Melanoma: An integrative case study

The incidence of melanoma, one of the most deadly forms of skin cancer, is on the rise (Fecher et al., 2007; Geller et al., 2002; Gray-Schopfer et al., 2007; MacKie et al., 2002; Thompson et al., 2005). It has become more prevalent in industrialised nations over the past 25 years, with Australia having the highest concentration of incidences in the world (Balch et al., 2001). Despite this increase, there has been minimal success regarding new treatment therapies since the late 1970s (Winnepenninckx et al., 2006). Melanomas developing distant metastases, labeled Stage III and IV, occur in about 15% of patients with primary melanomas (Tsao et al., 2004). For Stage III patients, about 30–40% die within one year and another 30–40% will survive beyond four years. There are novel, targeted and potentially aggressive systemic therapies being developed (Lorigan et al., 2008) for treating such a condition. However, this has been hindered as there is no way to identify patients who could potentially benefit from such therapies (Göran Jönsson et al., 2010; Mann et al., 2011).

For Stage III patients, the heterogeneity of survival outcomes (Ravo et al., 2008) matched to clinical variables, suggests that there are possibly molecular sub-clusters within Stage III melanoma patients (Hoshida et al., 2008; Rangel et al., 2008). Therefore, it would be advantageous to classify melanomas that have already undergone metastases into categories that may predict patient survival (Balch and Soong, 2008). To classify patients currently there are several methods, including Tumour-Node-Metastasis staging as well as well-defined clinical and pathologic variables such as Breslow thickness, ulceration and mitotic rate used to anticipate these groupings (Balch et al., 2001).

Mann et al. (2011) endeavoured to establish, for Stage III melanomas, if gene expression profiling can predict survival outcomes for patients. Moreover, we examined if

these molecular profiles can contribute further when used together with the clinical prognostic models. Such an integration aims to add another dimension to understanding the survival outcomes and predicting the survival experience for individual patients.

This chapter continues with the analysis of the Mann et al. (2011) data in three parts, each part consisting of a component of the analysis as illustrated in Figure 5.1. The first is the analysis of the clinical data, including data description, model construction and final prediction evaluations (Section 5.2). The second is the analysis of the expression data, which includes preprocessing, DE analysis and evaluation of the classification model and molecular signature (Section 5.3). The third component of this chapter is an outline of integrative methods applied to this study, the method finally selected and final model evaluation (Sections 5.4 and 5.5).

As this chapter contains the analysis of both clinical and gene expression data, there is a need to differentiate between different data matrices and types. Let $\boldsymbol{X}^{(\mathrm{C})}$ represent the clinical data matrix ($n \times Q$) and let $\boldsymbol{X}^{(\mathrm{E})}$ represent the gene expression matrix ($I \times n$). It is important to recall that the rows of $\boldsymbol{X}^{(\mathrm{C})}$ are the columns in $\boldsymbol{X}^{(\mathrm{E})}$. Each sample contains both clinical data (Q variables) and gene expression data with expression values for I genes.

5.1 Experiment aim and design

This experiment was based on the observed extremes in survival times of Stage III melanoma patients. Some individuals after treatment live for four or more years (and are considered 'treated' of the condition, a good prognosis), other individuals succumb to the condition within a year. In total, 83 Stage III melanoma patients were observed, there was clinical data for all 83 patients but expression data for only 79 of these individuals. Tumour samples were obtained from the Melanoma Institute Australia (MIA) Biospecimen Bank, with each tumour being from a patient from whom informed consent had been obtained with approval dating from 1996. Tumours were from patients where distant metastases were not known to be present at the time of tumour banking, and specimens were macro-dissected at banking and reviewed to meet minimum tumour cell content criteria. The motivation behind this analysis was to establish if one could predict the survival prognosis (good or bad) for individual patients. Presently there is very little known in clinical practice regarding this matter. The clinicians hoped that the introduction of expression variables would aid analysis. It is hoped that individuals with poor prognosis expectations may be willing to try experimental treatments to potentially aid their survival.

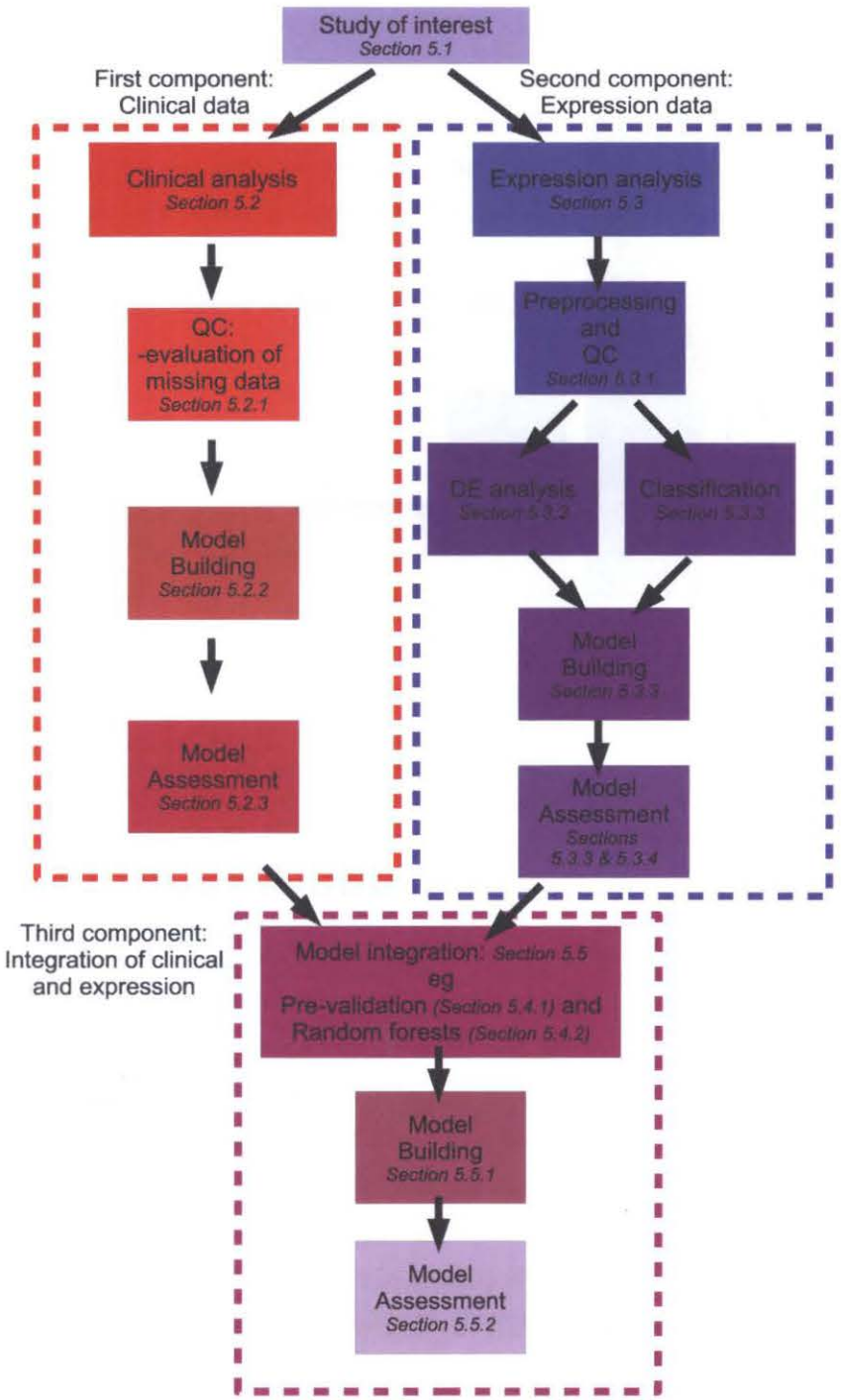


FIGURE 5.1: Flow-diagram of the steps involved in the analysis of the Mann et al. (2011) data, involving the integration of clinical and expression data.

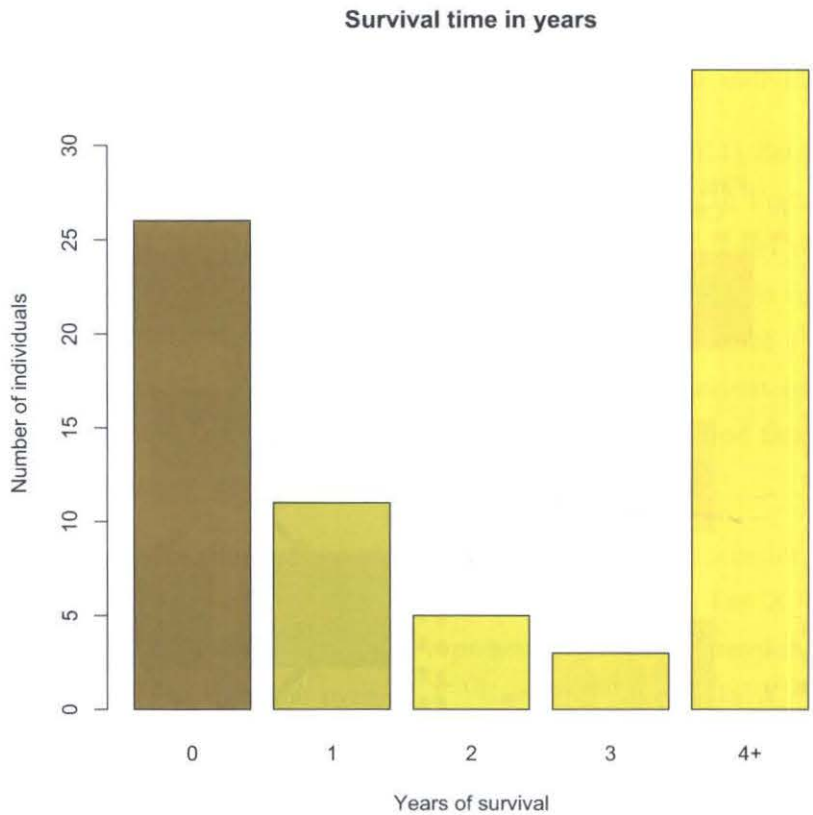


FIGURE 5.2: Histogram of the survival times of the 83 patients, *not conditioned on reason for death.* This plot highlights the extremes in the survival times of individuals with Stage III melanoma.

Figure 5.2 is a histogram of the survival times (not conditioned on reason for death) of the patients in the study, highlighting the two extreme survival groups within our data. Preliminary analysis comparing the differences between short and long term survival indicated that very little molecular signal was present in the data (both expression or clinical). Several definitions of good and bad prognosis were explored, examining the number of DE genes based on the different classes. Further detail of these class definitions and the number of DE are in Appendix D. For this analysis we focus on two survival extremes, a patient is considered within the good prognosis group if they survived more than four years with no sign of relapse ($n = 23$), and a bad prognosis group for those that survived less than one year and died due to melanoma ($n = 25$). This retained a total of 48 patients with matched expression and clinical data.

5.2 Clinical data

Pathologic, clinical and mutation information was obtained from each patient, together making up the 'clinical data' component of the analysis. Pathologic information includes percentage of non-tumour cells, percentage of necrosis, degree of pigmentation, predominant cell shape as well as cell size. Clinical information obtained includes age, gender, stage at diagnosis, location on body, presence of an associated nevus, Breslow thickness and Clark level amongst others. The mutation information observed includes BRAF, NRAS and PI3KCA mutations. More detail regarding the variables included in the clinical data set are shown in Mann et al. (2011).

5.2.1 Missing data

As advised in multiple imputation literature (Schafer, 1997, 1999), imputation was applied to the original data set, not just the selected 48 samples that were to be used in downstream analysis. After imputation, the completed data sets are reduced to the desired samples. The initial clinical data set consisted of 83 patients with Stage III melanomas and 33 variables. Clinical variables ranged from stage at diagnoses and survival status to tumour morphology and histology. Missing data was present in this data set at an overall average of 10.4%, such a level of missing data is relevant but not overly extreme (Campain et al., 2011; Rubin, 1996). Five variables contained no missing information and two variables contained missing data at levels greater than 25% (29% and 46%) (Figure 5.3). All variables were included in the analysis after discussions with clinicians. Only 36 (44%) patients contained a complete set of covariates (Figure 5.4). Case deletion is not considered a statistically appropriate method to overcome missing data with this proportion, 44%, of complete cases. Multiple imputation, making use of Amelia II [1] (King et al., 2001) with $m = 5$ [2], was employed to overcome missing data.

5.2.2 Model building and assessment

The dichotomous survival situation of Stage III melanoma patients was exploited to construct a regression model to predict the survival experience of future patients. A total of 48 patients were included in this stage of the analysis, 25 patients died within a year and 23 lived for more than four years with no sign of relapse. This data

[1] Amelia II was used here instead of MICE as this collaborative research was performed prior to the multiple imputation comparison study in Chapter 3.

[2] Between five and 10 multiply imputed data sets is recommended in literature (Rubin, 1987), although more can be used.

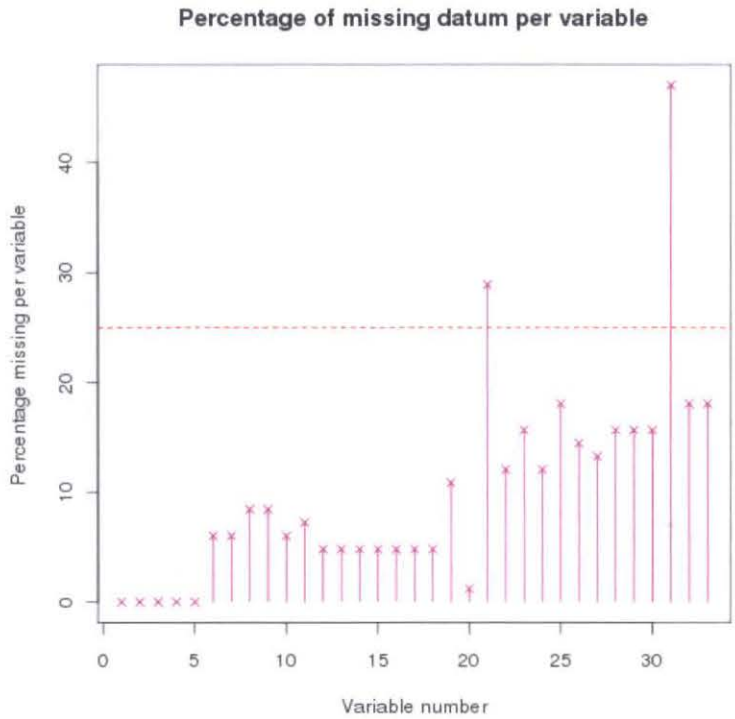


FIGURE 5.3: Graphical representation of the percentage of missing data, by variable. The average overall percentage of missingness is 10.4%, and there are two variables that contain a percentage of missingness over 25%. Although the amount of missingness is high for these two variables, they were retained in the analysis after discussions with the clinicians.

was analysed using a logistic regression model. A binary response variable of survival time was observed (1 = Survived greater than four years, with no sign of relapse, 0 = died within a year, due to melanoma). In total, 14 of the clinical variables were removed from the regression analysis because there was not adequate categorical sampling within the two groups to produce a stable regression model.

For the logistic regression, let $\boldsymbol{X}^{(\mathrm{C})}$ be the covariate matrix with dimensions (48 x 20), 48 samples and 20 variables, the first being a column of ones and the remaining 19 pertaining to the clinical variables in the data, $\boldsymbol{y}$ is a zero-one response vector of length 48 with 0 representing a death and 1 representing survival. The probability of an event occurring is π_j, ($\pi_j = P(Y_j = 1)$) and $\boldsymbol{\pi}$ denotes the vector of the 48 event probabilities. The matrix $\boldsymbol{X}^{(\mathrm{C})}$ contains missing covariates, let $\boldsymbol{X}^{(\mathrm{C})}{}_r$ be the rth imputation ($r = 1, ...m$). Let $\boldsymbol{\beta} = c(\beta_0, \beta_1, ..., \beta_{19})$, be a vector of parameters representing the coefficients of the 19 variables and the intercept term. For each imputation of the data set write,

$$\ln(\frac{\hat{\boldsymbol{\pi}}}{1-\hat{\boldsymbol{\pi}}}) = \boldsymbol{X}^{(\mathrm{C})T}_r \hat{\boldsymbol{\beta}}_r, \tag{5.1}$$

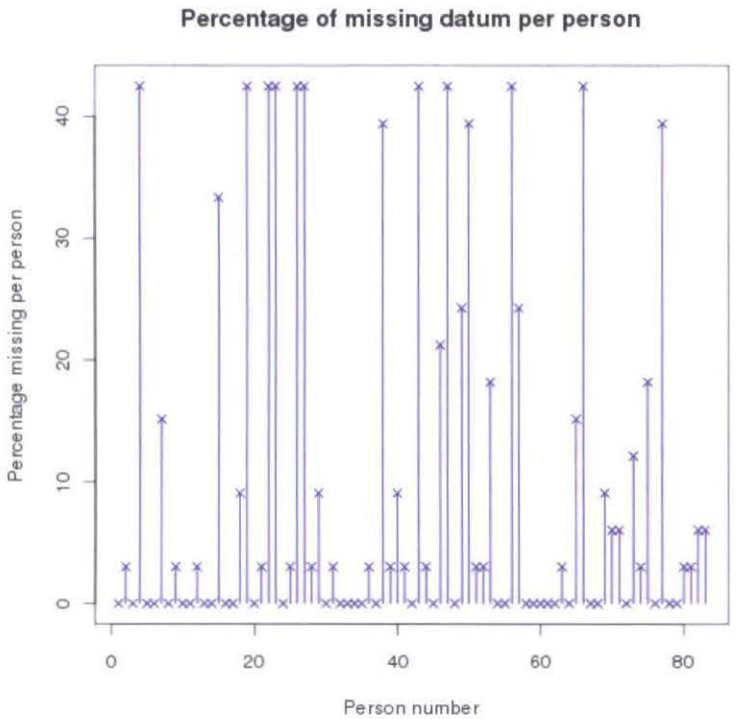


FIGURE 5.4: Graphical representation of the percentage of missing data, by sample. Only 44% of samples are complete and a complete case analysis is not advisable in such a situation.

where $\hat{\beta}_r$ are the estimated coefficients for the rth imputation. For these regression models, variables are reduced to form a parsimonious model using the BIC selection criteria. When logistic regression is applied to the multiply imputed data sets, m estimates for each covariate are produced. These regression models were aggregated with a multiple imputation inclusion frequency, τ_{MI}, of 0.5. More details on multiple imputation and the aggregation of regression models when it is applied is shown in Chapter 3.

5.2.3 Final model and prediction

The final logistic regression model with multiple imputation produces a leave-one-out cross validation (LOOCV) error rate of 27%. Table 5.1 consists of six clinical variables; BRAF mutation, cell size, nodal metastasises (NM), NRAS mutation, pigment and primary stage (which has three levels, 1, 2, and 3). The analysis showed that while better and worse prognosis tumours tended to differ by several features of their antecedent primary melanoma, only early stage at presentation (primary stage = 1) and the presence of a NM component were predictive of better survival once the Stage III disease was present. The association of a NM component with better prognosis might reflect a propensity for a localised rather than spreading growth pattern of the disease. The

Variable	Mean Coefficient
BRAF mutation (Yes/No)	-2.74
Cell Size	-1.80
NM	2.80
NRAS mutation (Yes/No)	-3.19
Pigment	-1.12
Primary Stage = 2	-3.90
Primary Stage = 3	-2.13

TABLE 5.1: Coefficients for the logistic regression model based on the clinical data, with $m = 5$ multiple imputations.

findings of an association of a NRAS and BRAS mutation with worse survival outcome agrees with other published work relating to Stage III melanomas (Göran Jönsson et al., 2010).

5.3 Expression data

This section illustrates some of the typical analysis methods used to understand the expression data. These include performing QC analysis, establishing a DE gene list, constructing a classification model and verification of the significance of some of the results.

5.3.1 Preprocessing

The expression data consists of 79 Illumina 'IlluminaHumanv3' arrays. These 79 arrays were matched to 79 clinical samples.

Data from the 79 Illumina beadarrays was preprocessed and analysed using R. Quality control was performed on the chips using the `lumi` package (Du et al., 2008) in R/Bioconductor (Gentleman et al., 2004). Based on the quality assessment all arrays were deemed suitable for further analysis.. Data normalisation was performed using a variance-stabilising transform (VST) (Lin et al., 2008) and quantile normalisation as implemented in the `lumi` package for R/Bioconductor. To reduce false positives, unexpressed genes (based on a detection p-value cut-off of 0.01, Lin et al., 2008) were removed from the data set. This reduced the number of probes being analysed from 48802 to 26085, 53% of the total number of probes on the Illumina human array. Throughout

the analysis, annotation of the Illumina arrays was performed on `illuminaHumanv3.db` with R version R-2.11.0. More details are shown in Chapter 2, Section 2.3.1.

5.3.2 DE analysis

A linear model was applied to identify DE genes between the two groups; died within a year due to melanoma and survived more than four years with no sign of relapse. The effects of interest were estimated using the log intensity values from the normalised arrays for both data sets. For a typical gene, the gene's log intensity value for the sample j is denoted as y_j where $j = 1, ..., 48$. Then, the gene expression for gene i can be linearly modelled by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon \tag{5.2}$$

where $\boldsymbol{y}$ is a vector of log intensity values, $\boldsymbol{X}$ is a design matrix, $\boldsymbol{\beta}$ is a vector of parameters and ϵ, is a normally distributed error term. The estimable parameters within $\boldsymbol{\beta}$ included the intercept (α) and the class effect (κ) at two levels, with 1 = indicating survival greater than four years with no sign of relapse. The values associated with the experiment were presented as

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ \vdots \\ y_{46} \\ y_{47} \\ y_{48} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \kappa \end{pmatrix} + \epsilon. \tag{5.3}$$

The robust linear parameter estimate for β_j was estimated using the functions implemented in the `limma` package (Smyth, 2004). Table D.1 case (f) contains the number of DE genes when different cut-off values were used, under this good and bad prognosis definition. Genes selected by controlling for 5% FDR were used for downstream analysis, including ontology results and further investigations by biologists.

Conclusions with practical and biological interpretations should examine the data from many aspects. To this end, DE analysis has been performed within this analysis to deepen our understanding, and gain as much information as possible, regarding the behaviour of genes within these two melanoma cases.

5.3.3 Classification modelling and prediction

Genes used to construct a classifier, known as the 'molecular signature', are not necessarily the same DE genes selected above. Genes were selected for the molecular signature via the 'median robust' method. Genes were ranked based on the difference of the two group medians, $(\tilde{x}_{\text{good}} - \tilde{x}_{\text{poor}})$, where $\tilde{x}$ represents the median of a group. Diagonal linear discriminant analysis (DLDA) (Hastie et al., 2009) was used as a classifier and the number of genes in the molecular signature varied from 10 to 500 in increments of 10 genes. LOOCV was employed to estimate the prediction accuracy for the optimal number of genes in the signature. LOOCV was used above other cross-validation methods because of the small number of samples in the study. such a cross-validation method allows for the largest number of samples to be used to construct the classifier, yet still retains an unbiased performance measure. Other forms of feature selection and classification methods were employed, but such approaches did not produce greater accuracy (Appendix D).

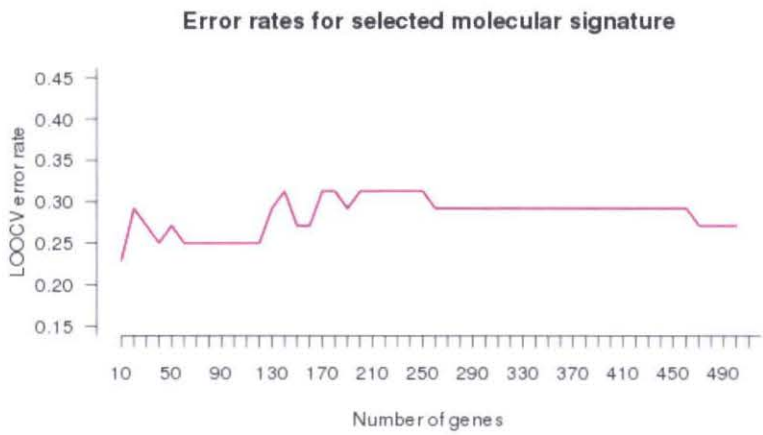


FIGURE 5.5: The varying LOOCV error rates as the number of genes used to construct a DLDA classifier changes from 10 to 500. The lowest error rates occur when 10 genes (22% LOOCV error) and 60 genes (25% LOOCV error) are used.

Figure 5.5 shows how the LOOCV error rate (y-axis) varies as the number of genes used to construct the classifier range from 10 to 500 (x-axis), genes were ranked based on the median difference between survival groups. Using LOOCV, a molecular signature of 10 genes produced an error rate of 22% and a molecular signature of 60 genes produced an error rate of 25%.

5.3.4 Functional evaluation of the DE genes

Ontology and KEGG analysis The ontologies and Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathways of the selected genes were analysed, using both gene set tests (GST) and gene ontology (GO) analysis. GST were performed on ranked ontologies, including biological process (BP), cellular component (CC), molecular function (MF) and KEGG pathways of the expression data for both consistently low and high ranked ontologies using gene set tests implemented in `limma`, based on the robust logFC statistic. For the GO analysis the DE gene list obtained for the 48 samples using robust regression as well as the selected molecular signature, were analysed to see if the gene ontologies, or KEGG pathways observed were over-represented in the selected genes, for both up and down regulated genes using a hypergeometric tests (Beissbarth and Speed, 2004). Ontologies with an overall probe count of less than 5 were excluded from the analysis.

Validation of molecular signature and ontologies Validation of the molecular signature was performed using published gene signatures in a classification paradigm, discussed in the following section with complete details in Sarah-Jane Schramm et al. (2011). The signature found for Mann et al. (2011) is a highly immunologically driven gene list. Such findings agree with other published works including that of Göran Jönsson et al. (2010), who clustered samples with survival times based on immunological status. The most highly significant ontology in the BP network was *Immune:TCR signalling* with many other significant ontologies being linked to immune response. The link to immune response is well documented in melanoma research, it is also linked to other tumour types. For example Berezhnaya (2010); Finn (2008); Göran Jönsson et al. (2010) and Tiwari (2010) considered the link between the tumour and the immune system. More detail of this validation is shown in Mann et al. (2011).

5.3.5 Integration and validation of multiple molecular signatures

As discussed in Chapter 4 there are multiple ways of integrating data and performing meta- and mega-analysis. To add further rigour to the research being applied on the melanoma data, the gene lists obtained for classification were compared to other published gene lists where research purposes coincided with the current study. In Sarah-Jane Schramm et al. (2011) meta-analysis was used as a validation tool to compare the molecular signatures from multiple expression array experiments to establish if it were possible to construct a molecular classification signature similar to that obtained for breast cancer in van't Veer et al. (2002). Tímár et al. (2010) compared gene signatures from four melanoma studies (Bittner et al., 2000; Thomas John et al., 2008; Mandruzzato et al.,

2006; Winnepenninckx et al., 2006) and found that there was little overlap between the published signatures. In recent months, several new studies of metastasised melanomas have been published. Sarah-Jane Schramm et al. (2011) reviewed these recent works and performed a formal and systematic cross-validation meta-analysis study to compare the capacity of each signature to predict survival outcomes on the other examined data sets. This cross-validation meta-analysis is described below.

Meta-analysis method Five published data sets were obtained, which have been discussed in Chapter 2 and refered to here as the Bogunovic, John, Jönsson, Mann and Winnepenninckx data. The data was preprocessed according to platform type and requirements. Each data set consisted of two groups, a good prognosis group and a poor prognosis group. More details regarding the individual data sets, group definitions and preprocessing are shown in Sections 2.3.1 and 2.3.2. Each published data set was coupled with the published signature genes list, these signature gene lists were lists of Entrez IDs. The main question to be considered is whether the signature gene list had predictive capabilities outside of the original data set it was developed under. The predictive power of each of the five published signature gene lists in turn were evaluated based on the other four gene expression data sets, such that:

1. For each study (data set A), a published gene list was obtained from the presented signature; this is called the gene list obtained from data set A, resulting in the feature vector.

2. For another study (data set B), a published expression data set was obtained from a public repository.

3. A classification rule using SVM was developed using the expression results from data set B, but the feature list from data set A.

4. This classification rule was tested on data set B applying LOOCV, to estimate the misclassification error rate for patient outcome for the gene signature from data set A tested on data set B.

Figure 5.6 is a graphical representation of the comparison and classification process used.

Results Table 5.2 contains the LOOCV error rates for the different gene signatures tested on the other expression data. The greyed boxes represent the LOOCV error rate when the signature gene list is used to classify the original data set that it was developed on. Notably, several of the signature lists validated well when the other studies were classified by them, with misclassification error rates as low as 0.08 (when the Jönsson

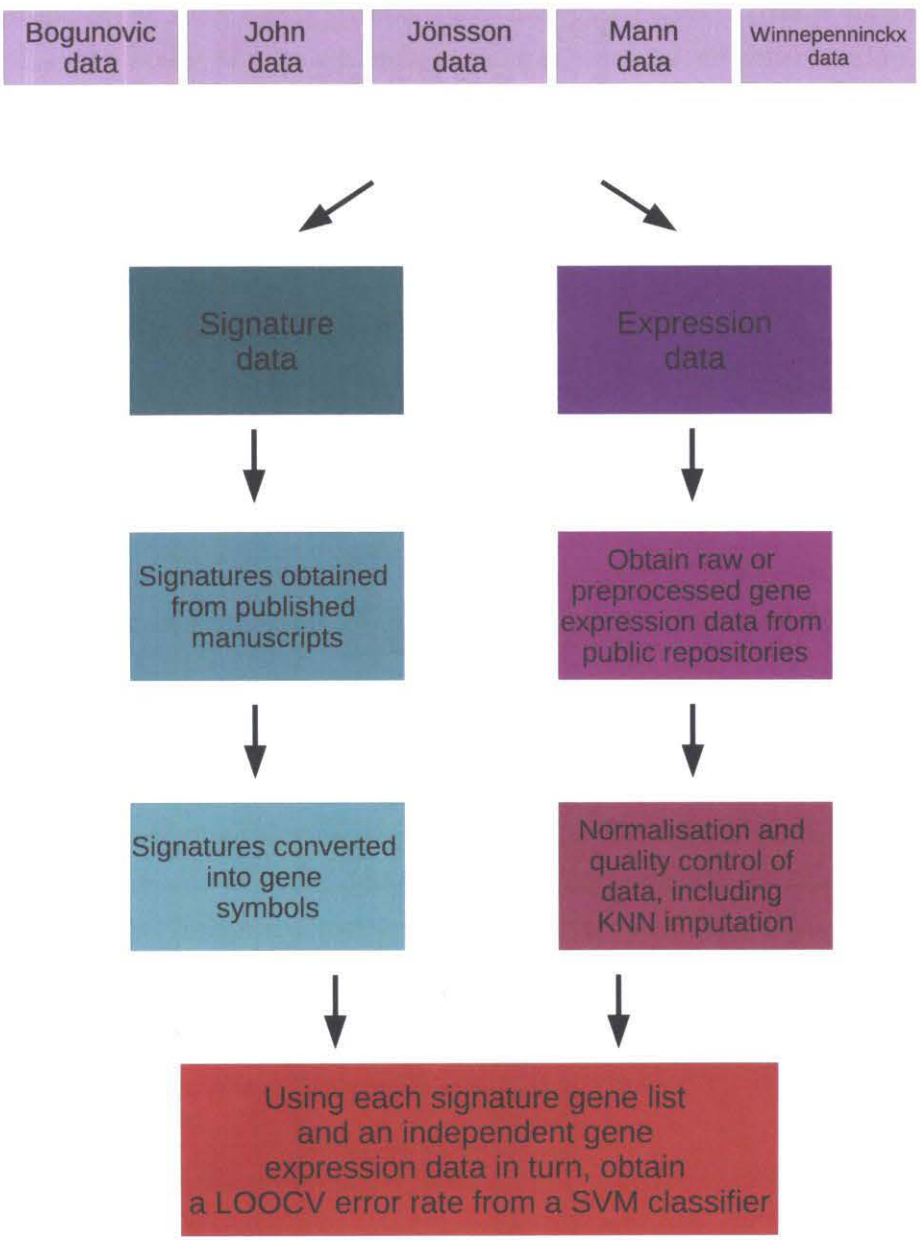


FIGURE 5.6: Graphical representation of steps involved when establishing how well a signature gene list classifies a different data set. The aim of this comparison is to establish if published gene lists are transferable across multiple data sets applying similar experimental questions.

Gene signature \ Expression data set	Bogunovic	John	Jönsson	Mann	Winnepenninckx	Mean
Bogunovic	0.18	0.17	0.20	0.54	0.48	0.31
John	0.24	0.13	0.20	0.27	0.48	0.26
Jönsson	0.42	0.08	0.17	0.38	0.74	0.36
Mann	0.36	0.13	0.30	0.35	0.35	0.30
Winnepenninckx	0.39	0.50	0.20	0.44	0.70	0.44

TABLE 5.2: LOOCV error rates for the misclassification of patients into outcome related classes in the cross-validation of gene signatures between independent data sets.

signature gene list was used to classify the John data set). Low error rates confirm the biological relevance of the signatures in question. Overall the study with the lowest average error rate for its published gene signature was John (0.26), which performed well except in the case of the Winnepenninckx expression data. This observation may be a reflection of the stage differences between these two data sets (the Winnepenninckx data involves the study of primary melanoma sites where as all other studies examine more advanced tumour stages and metastasised sites). Overall there is a surprising degree of agreement between all the studies when classification rules were developed using externally obtained gene lists.

5.4 Methods for data integration

Considered here are two methods of integration; (i) pre-validation of the microarray vector and then the application of logistic regression, and (ii) random forests. If there are n samples, let $\boldsymbol{X}^{(\mathrm{C})}$ be the matrix of clinical variables $(n \times Q)$ and let $\boldsymbol{X}^{(\mathrm{E})}$ be the matrix of gene expression values $(I \times n)$. If $\boldsymbol{y}$ is a vector of clinical outcomes, in this case good/bad survival prognosis, an integrative predictive model would combine these two elements.

In this specific context the clinical variables contain missing values. To this end, all integrative methods need to be able to be applied through multiple imputation. Combining models, other than in a regression context, is challenging for multiply imputed models so. As a results of these challenges modelling methods are limited. Although there are many other integrative methods recently developed, it is because of missingness that they have not been implemented.

5.4.1 Pre-validated vector and regression

'Pre-validation' is a method developed in Tibshirani and Efron (2002), where the gene expression molecular signature is used to make a prediction for the samples within a study. The method is similar to that of 'staking' detailed in Wolpert (1992) from machine learning. The development of pre-validation came about because of the concern of biasing the effects of the expression data while integrating clinical and expression data together in a regression context, as performed by van't Veer et al. (2002).

In the van't Veer et al. (2002) study a 70 gene molecular signature was designed and the samples were classified using this rule, producing $\boldsymbol{A}$, the vector of estimates with one estimate for each sample. The vector was added to a regression model containing

clinical variables, and the coefficients for each variables were assessed before and after the introduction of the vector $\boldsymbol{A}$ into the model. The significance of these variables and the adjusted size of the coefficients was used to argue for the importance of the molecular signature in class prediction over and above that of the clinical variables. Höfling and Tibshirani (2007) and Tibshirani and Efron (2002) argue that such a method favourably biases the vector of estimates developed by the molecular signature as the same samples were used to construct the signature and be predicted by the signature.

The aim of pre-validation is to construct a 'less biased' microarray predictor (Tibshirani and Efron, 2002), to be fit alongside the clinical variables. The process can be considered in five steps, which is graphically represented in Figure 5.7:

1. Divide the cases into k equal parts.
2. Set aside one part.
3. Using the other $k-1$ parts, obtain a molecular signature and classification rule.
4. Use this rule to predict the kth part.
5. Repeat steps 2–5 for all parts, resulting in a pre-validated vector of estimates for the microarray vector, $\boldsymbol{A}_{\mathrm{PV}}$.

To combine $\boldsymbol{A}_{\mathrm{PV}}$ with the Q selected clinical variables from $\boldsymbol{X}$, a regression model can be developed,

$$y_j = \beta_0 + \beta_1 x_{1,j}^{(\mathrm{C})} + \beta_2 x_{2,j}^{(\mathrm{C})} + ... + \beta_Q x_{Q,j}^{(\mathrm{C})} + \beta_{Q+1} A_{\mathrm{PV}j} \tag{5.4}$$

where β_q is the coefficient for the qth variable ($q = 1, ...Q$) in the regression, and $x_{q,j}$ is the jth sample's observation for the qth variable. Variables in the integrated regression model are typically selected independently from prior investigation, so no variable selection takes place. Pre-validation is not a substitute when an independent data set is available. Tibshirani and Efron (2002) advise that $k \neq n$, which would result in a highly variable LOOCV pre-validation vector. Multiple imputation can be applied to the regression model, with m regression models being constructed for the m imputations and then aggregated using an inclusion frequency. The pre-validated variable $\boldsymbol{A}_{\mathrm{PV}}$ is not included in the imputation process but is treated as a complete variable for modelling purposes. Hence this allows the production of an integrated regression model making use of the pre-validated microarray vector and the clinical variables.

5.4.2 Random forests

'Random forests', developed by Breiman (2001) is an ensemble method of classification trees. It is possible to use random forests in an integrative paradigm. This is possible because one of the strengths of random forests and classification trees is that variables with vastly different ranges or distributions do not dominate the developed model. This allows direct modelling of the expression values with the clinical values. The flexibility of such an approach allows for several options regarding the clinical and expression variables. For example, it is possible to construct random forests on the selected clinical variables as well as the selected expression variables. Further, it is possible to make use of the pre-validated vector, $\boldsymbol{A}_{\text{PV}}$, with the clinical variables. Out-of-bag error rates and importance scores (Breiman, 2001) can be used to evaluate the model's predictive capabilities and the importance of the used variables.

Although Breiman et al. (1984) propose that imputation is possible with classification trees, this is achieved through surrogate splits. Surrogate splitting is a form of single imputation and in this thesis it is conjectured that an adaptation of random forests with multiple imputation will in general improve its use in the presence of missing data, allowing for the required increase in variability (Chapter 3). Such conjecture is similar to Ding and Simonoff (2010) and Feelders (1999). To compensate for the shortcomings of single imputation the method of multiply imputed random forests was developed. This is a novel method not yet fully explored. If in a traditional random forest design B trees are constructed, if there are to be m multiple imputations, mB trees are built. To incorporate multiple imputation into random forests, an additional layer of randomness can be added into the random forest design. Here prior to the construction of the forest, one of the m imputed clinical data sets is selected, and a random forest is developed on only B trees. This process is repeated m times. The eventual product is m results incorporating the aggregation of out-of-bag error rates, importance scores or final voted classes. Such a method is illustrated in Figure 5.8.

5.5 Integration of clinical and expression data

To integrate the clinical and gene expression data, multiple variations of the pre-validated method and multiply imputed random forests were applied. Table 5.3 includes a summary of the seven used methods. These methods were applied because they could be implemented with multiple imputation resulting in a final aggregated model. The success of the logistic regression methods were very similar, with final 6-fold [3] cross-validation

[3] 6-fold pre-validation was used because five to ten-folds were suggested in Tibshirani and Efron (2002). Moreover, 48 is divisible by six, allowing for computational ease.

Data	Modeling method	Number of genes	Pre-validation of molecular signature
Clinical	Logistic regression	none	no
Clinical and expression	Logistic regression	10	yes
Clinical and expression	Logistic regression	60	yes
Clinical and expression	Multiply imputed random forests	10	no
Clinical and expression	Multiply imputed random forests	60	no
Clinical and expression	Multiply imputed random forests	10	yes
Clinical and expression	Multiply imputed random forests	60	yes

TABLE 5.3: Methods used to integrate clinical and expression data.

error rates being between 23% and 37%. The out-of-bag error rates obtained from the random forest methods with multiple imputation were highly unstable. Due to this instability, the method was not pursued to obtain a final model and is now the subject of further research.

5.5.1 Clinical and expression integration

The final method selected to integrate the clinical and gene expression data was the pre-validation method, with the pre-validated vector obtained using the 60 gene signature (the third method presented in Table 5.3). The pre-validation method was applied in a three fold manner:

1. In the first stage a regression model is obtained based only on the clinical data. Variables are selected, in this case making use of the BIC criterion, and multiple imputation.

2. The second stage makes use of a k-fold cross validation process ($k = 6$) based on the molecular signature and classification rule to obtain an expression data class estimate. This expression data class estimate, known as the 'pre-validation estimate' is a class vector (in this case consisting of 0's and 1's).

3. The selected clinical and pre-validated variables were integrated into a logistic regression model, under $m = 5$ multiple imputations using Amelia II, validated using a 6-fold cross validation.

Variable	Mean Coefficient
BRAF mutation (Yes/No)	-2.92
Cell Size	-2.46
NM	1.91
NRAS mutation (Yes/No)	-2.75
Pigment	-1.12
Primary Stage = 2	-3.45
Primary Stage = 3	-1.07
Pre-validation estimate	2.39

TABLE 5.4: Coefficients for the logistic regression model based on the clinical data and the pre-validated expression data vector, with $m = 5$ multiple imputations.

5.5.2 Final model and prediction

The final 6-fold cross-validation error rate for this model is 23%, and the model coefficients are reported in Table 5.4. Notably, none of the effects associated with the clinical, pathologic and mutation variables were weakened significantly by the incorporation of the gene expression pre-validated variables (compare Table 5.1 and Table 5.4). This indicates that the gene expression profiling signature does not specifically reflect any of the selected clinical variables, but has an independent prognostic value. For example it cannot simply be a molecular footprint of BRAF or NRAS pathway mutation.

To summarise, a model was derived that is effective in identifying patients with Stage III melanomas who have good long-term survival prognosis, after nodal resection. This integrative model takes into account clinical, pathologic, gene mutation and gene expression data. Through ontology investigations it is clear that the gene expression signature is indicative of immune response activation, agreeing with other work in the field (Berezhnaya, 2010; Finn, 2008; Göran Jönsson et al., 2010; Tiwari, 2010).

5.6 Publications

This chapter is the detailed analysis for results under review in Mann et al. (2011) and also covers work published in Sarah-Jane Schramm et al. (2011). Both analyses were conducted by the author with Professor Graham Mann's groups at the Westmead Millennium Institute, Australia and the Melanoma Institute Australia, in particular Professor Mann and Ms Sarah-Jane Schramm.

Publications

- S.-J. Schramm, **A.E. Campain**, R. Scolyer, Y.H. Yang and G.J. Mann (2011) Review and cross-validation of gene expression signatures and melanoma prognosis. **132**:274–283 *Journal of Investigative Dermatology.*

Manuscripts under review

- G.J. Mann, G.M. Pupo, **A.E. Campain**, C.A. Carter, S.-J. Schramm, A. Pianova, S. Gerega, C. De Silva, K. Lai, J. Wilmott, M. Synott, P. Hersey, R.F. Kefford, J.F. Thompson, Y.H. Yang and R.A. Scolyer (2011) BRAF mutation, NRAS mutation and absence of an immune-related expressed gene profile predict poor outcome in stage III melanoma. Under Review, *Journal of Clinical Oncology.*

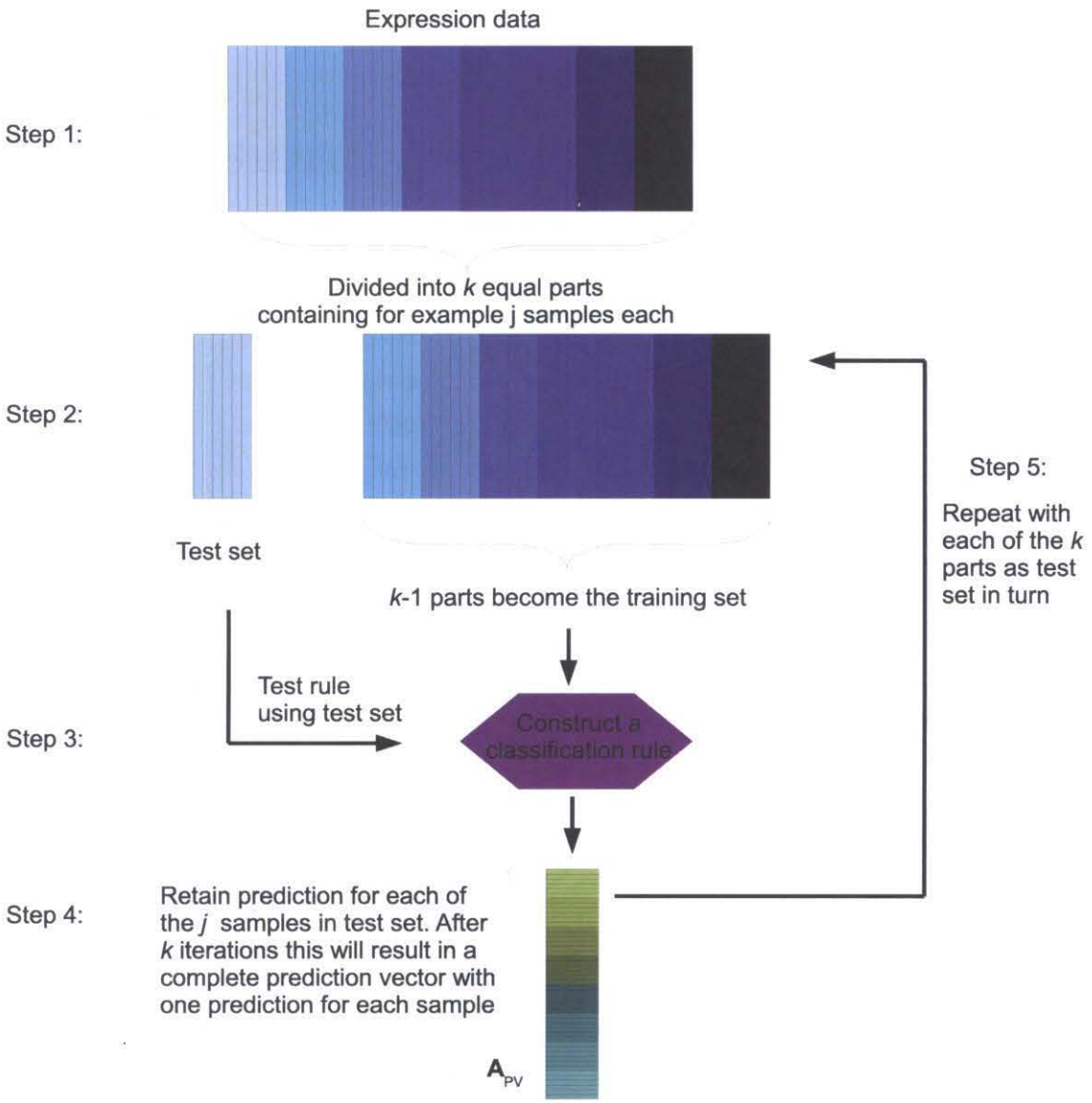


FIGURE 5.7: Graphical representation of the pre-validation method (Tibshirani and Efron, 2002).

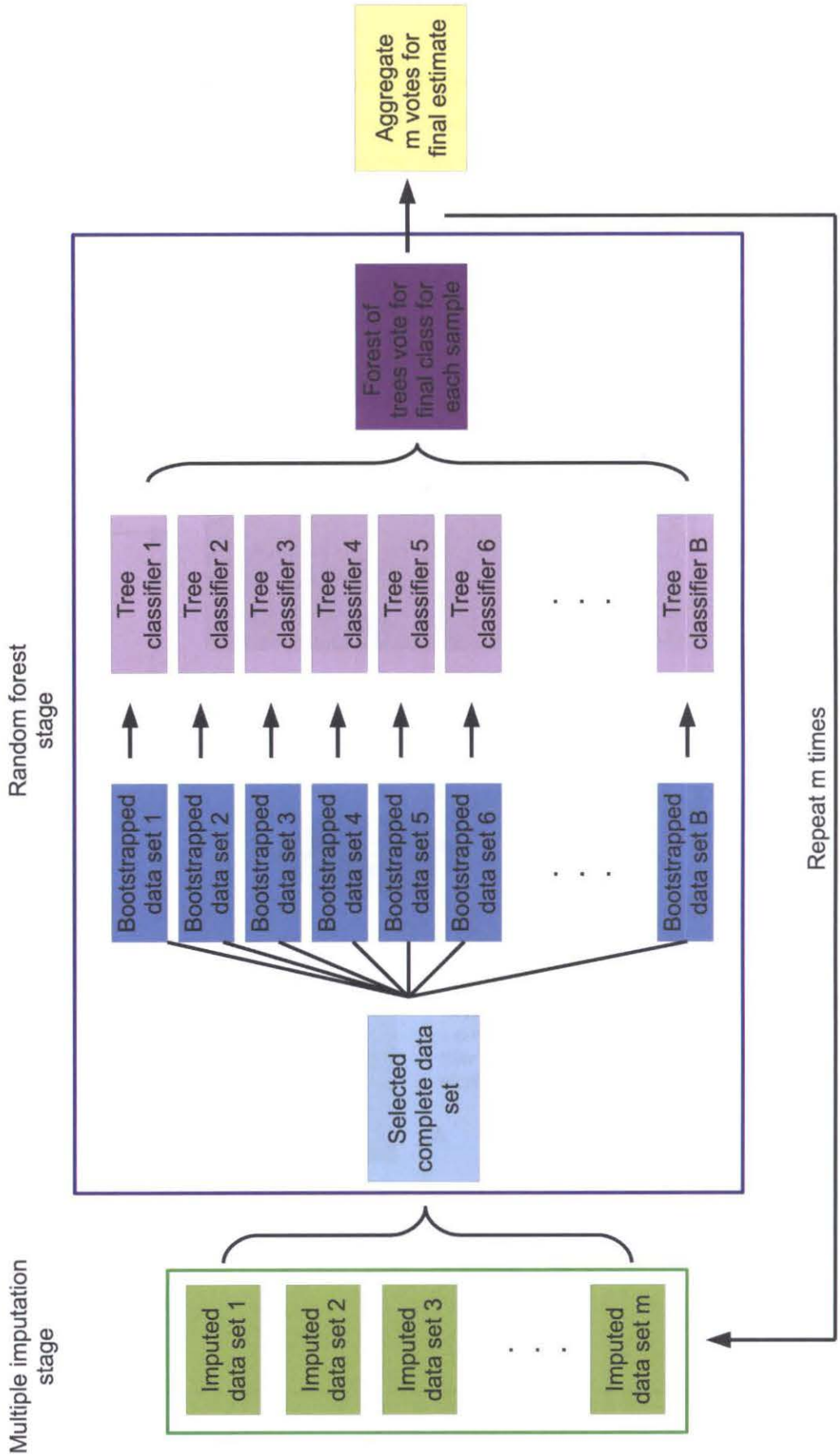


FIGURE 5.8: Graphical representation of multiply imputed random forests.

Chapter 6

Conclusion

The integration of gene expression and clinical data has great potential to deepen our understanding of complex diseases. However, there are still many issues relating explicitly to clinical or expression data that need to be matured before such goals can be reached. This thesis has considered some of these problems and offers solutions that are readily applicable. Studied in particular are the problems of; missing data, the importance of building a stable model as well as gene expression integration. Throughout this thesis, each of these statistical problems was motivated by real data, highlighting the relevance of such solutions to current research.

The concept of imputation and multiple imputation is well established in statistical research. Chapter 3 addressed three less well examined questions: how to compare multiple imputation methods, what is the best such method, and does a critical point exist beyond which missingness is too great to trust downstream analysis? Missing data was induced within a complete data set to simulate missingness at particular proportions. Assessments of the multiple imputation methods were made possible by examining the changes in distributions, as missingness increased, for each imputation method and considering the information lost through the degradation of classification accuracy. MICE was found to be the most appropriate multiple imputation method of those compared. It was confirmed that a missingness rate of 30% or less can be adequately handled by multiple imputation.

Construction of a stable model is an important concept in statistics. Unstable models are produced when small changes in the data set result in vast changes to the final model. The unstable nature of some models can be a result of many elements including; high amounts of correlation between variables, that is the occurrence of multicollinearity within the final models, and/or high levels of class imbalance within the response

data. The B-MI approach was presented within Chapter 3. This is a variable selection method that results in stable variables through the implementation of bootstraps. Stable variables are selected with an appropriate choice of τ_B, the inclusion frequency of variables based on bootstrapped samples. A simulated data set with missingness in the explanatory variables was developed from a known model. The B-MI method and other commonly used methods in the presence of missing data were compared. The B-MI method was shown to be the most effective method in selecting the original design model and resulted in stable variables being selected. The data set within this simulation was highly unbalanced, and it was shown that the use of weights within the regression modelling process enhanced both the predictive capabilities of the models as well as increased the frequency of selecting the known model coefficients. The B-MI approach, and the comparison of multiple imputation methods is presented also in Campain et al. (2011).

The analysis of the EPU data, in Chapter 3, showed how both missing data and the construction of a stable model can be appropriately handled in practice. The final model developed can be used in a clinic and agrees with current clinical understanding.

Integrating microarray data is becoming an important area of research. This is especially true as the amount of publicly available microarray data increases. Public data can be used in a number of ways, for example; (i) in a validation context, (ii) to increase the power of current projects, or (iii) to address a differing question from the original analysis. Chapter 4 explored the integration of microarray data sets at two levels; high level and low level integration, known as meta- and mega-analysis respectively. A new meta-analysis method, mDEDS (Campain and Yang, 2010; Yang et al., 2011), was explored within this chapter and was compared to other meta-analysis methods through a simulation study and a classification study. Meta- and mega-analysis methods were compared through a hypertension study, making use of four publicly available data sets (Campain et al., 2010). This comparison highlighted that there are different strengths to the different integration methods and the use of such approaches depends on the purpose of the study, the research question, as well as the data available to integrate.

Chapter 5 was a case study where the different issues addressed within this thesis came together. The Mann data contains clinical data, with missing values as well as expression data. The desire of the clinicians was to develop a predictive model that could incorporate these two elements with the potential for clinical use. Integrating such data, with the use of multiple imputation and a pre-validation vector, allowed for the development of a model that agreed with current clinical understanding. This model was able to be validated by external expression data and offered an increase in predictive capabilities in a clinical context (Mann et al., 2011).

This thesis contributed to the solutions of important challenges in the analysis of clinical and microarray data and their integration, including missing data, class imbalance, unstable model selection, meta- and mega-analysis and the exploration of integrative techniques that include multiple imputation. There are still many open questions in this field of research. Such questions include: the handling of mismatched probe sets within microarray integration and the notion of a many-to-one mappings of probes to genes and gene regions; the incorporation of multiple imputation into various classification methods as well as the integration of clinical and expression data and; the integration of these two data types with new technologies emerging in bioinformatics such as genome sequencing and mass-spectrometry data.

Appendix A

Further results from Chapter 2

A.1 Hypertensive data sets quality control and analysis plots

For each of the four hypertensive data sets, quality control (QC) and individual analysis was performed. Figures A.1–A.4 contain four of the QC images used to determine the quality of the data set in question and two graphical tools for interpretation. Each figure includes:

(a) Spatial images for each array in the analysis.

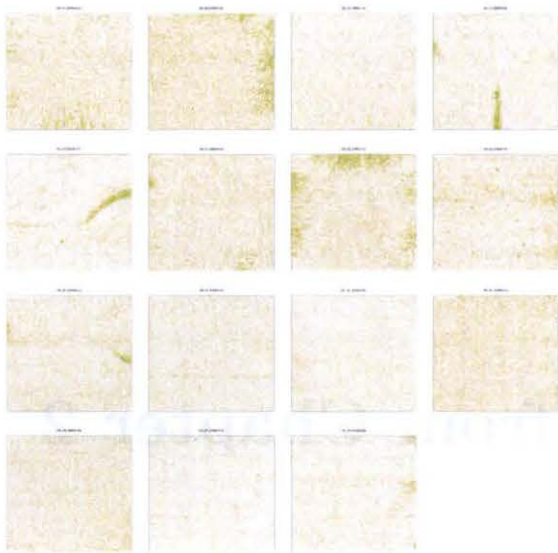
(b) Relative log expression boxplots for each of the arrays in the analysis (RLE plots), the hypertensive samples are in red and the normotensive samples are in green.

(c) NUSE plots for each of the arrays in the analysis, the hypertensive samples are in red and the normotensive samples are in green.

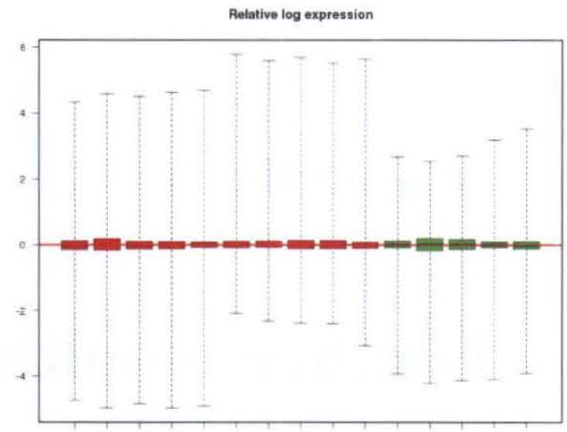
(d) Hierarchical clustering plots for individual hypertensive/normotensive studies. Top 500 most variable genes are used, judged on overall variability. The Euclidean distance was used as the distance metric with the complete agglomeration method.

(e) MA plots showing the average expression value (x-axis) against the $\log_2$(FC) values (y-axis). Genes with a positive FC greater than 1.5 are highlighted in red and genes with a negative FC less that -1.5 are highlighted in green.

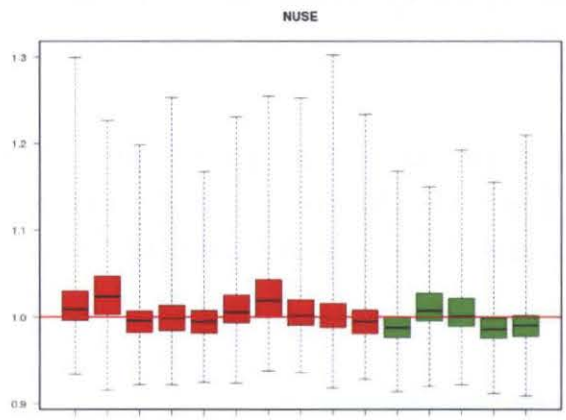
(f) Volcano plot showing the $\log_2$(FC) values versus the -$\log_{10}$(p-value).



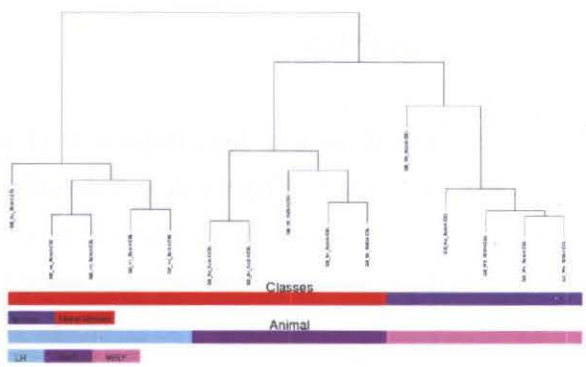
(a) Spatial image plots



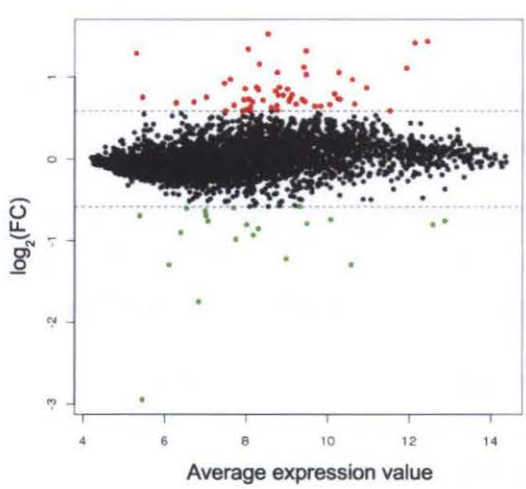
(b) RLE boxplots



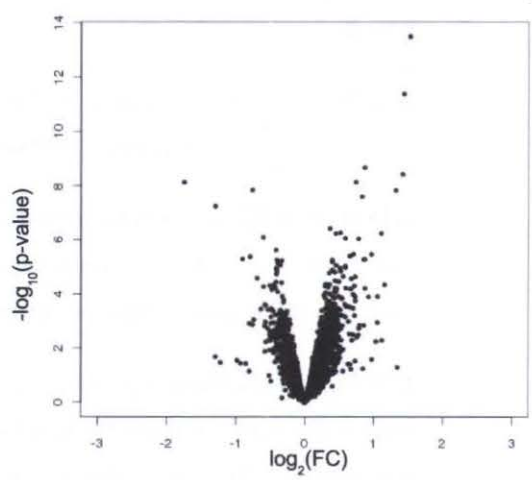
(c) NUSE boxplots



(d) Hierarchical clustering plot

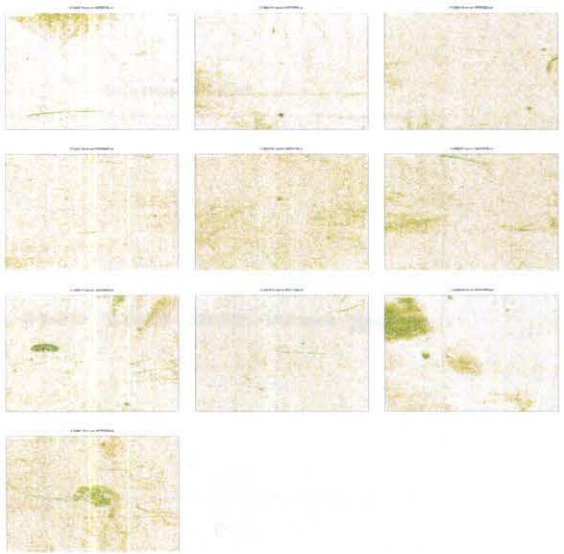


(e) MA

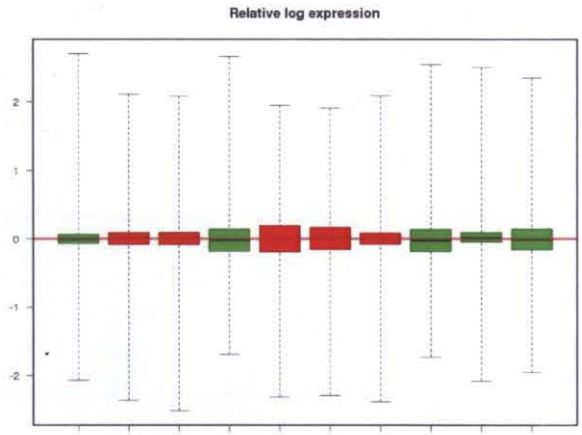


(f) Volcano plot

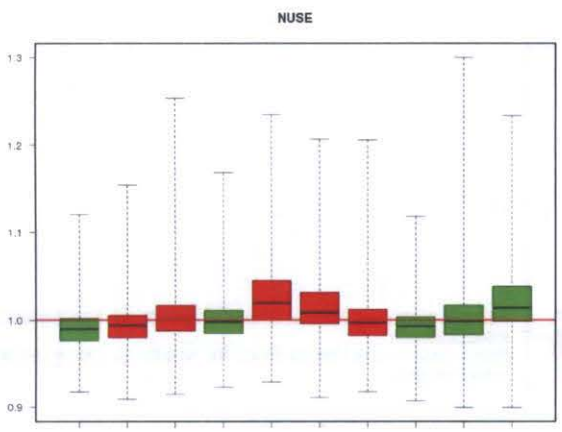
FIGURE A.1: QC and analysis plots for the 15 samples in the Cerutti data analysis.



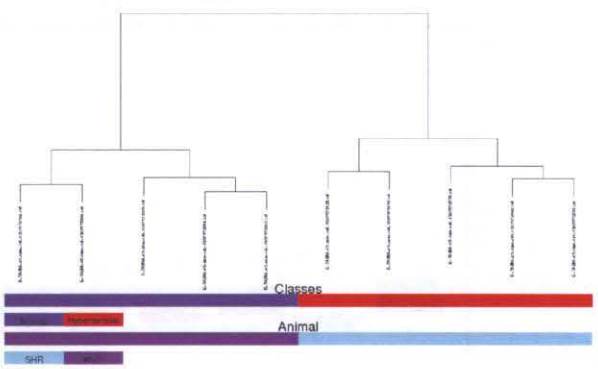
(a) Spatial image plots



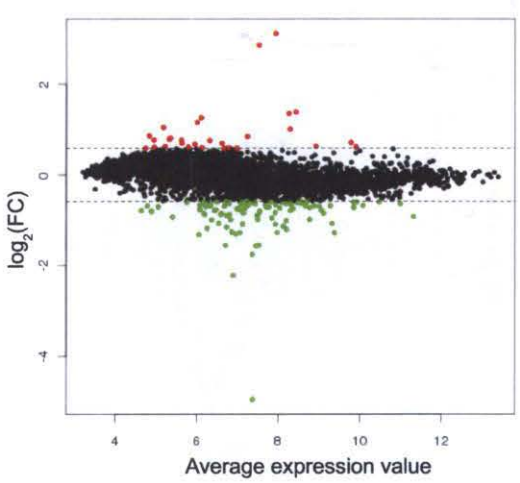
(b) RLE boxplots



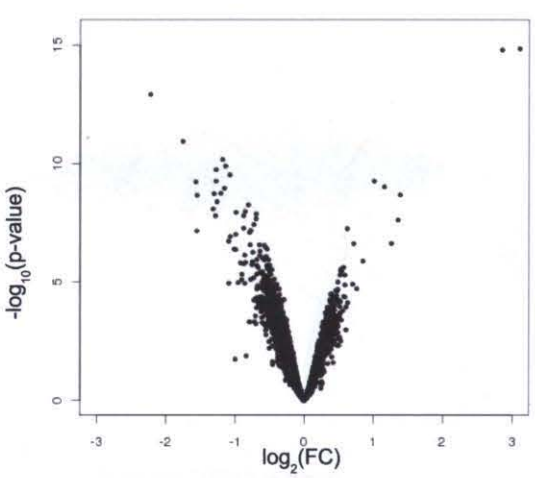
(c) NUSE boxplots



(d) Hierarchical clustering plot

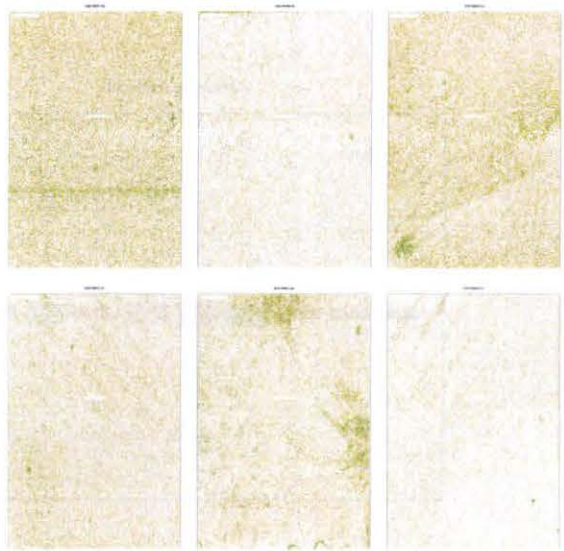


(e) MA plot

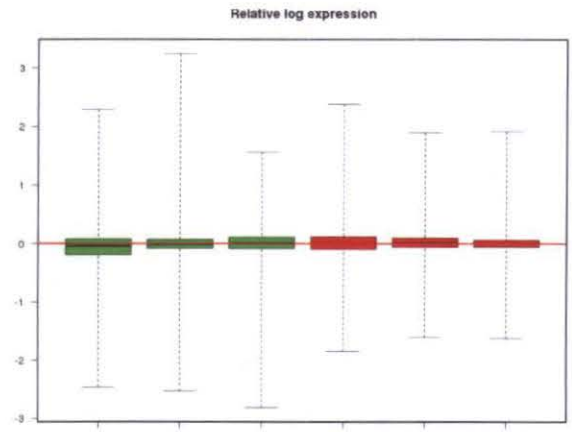


(f) Volcano plot

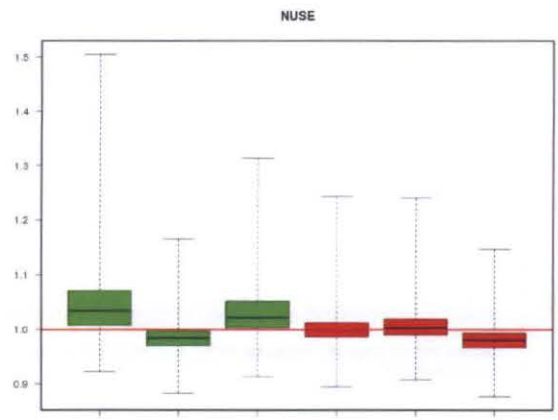
FIGURE A.2: QC and analysis plots for the 10 samples in the Clemitson data analysis..



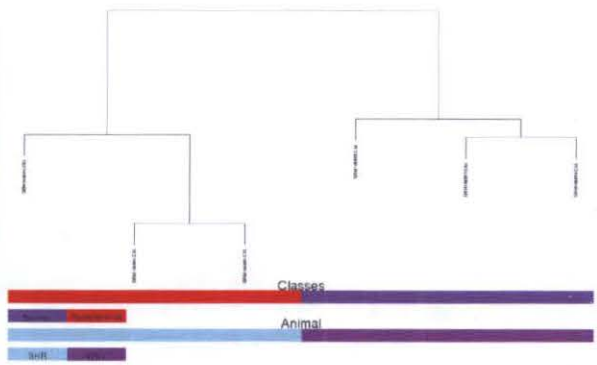
(a) Spatial image plots



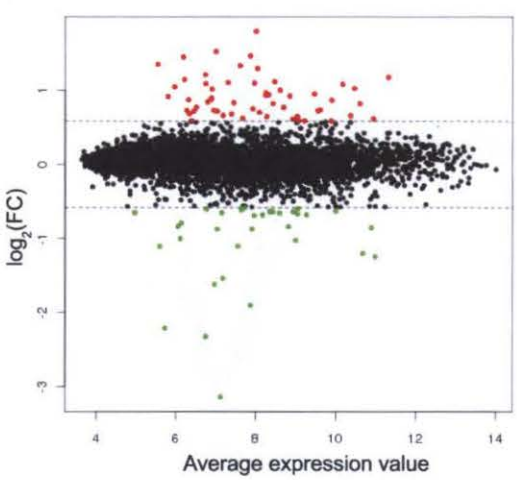
(b) RLE boxplots



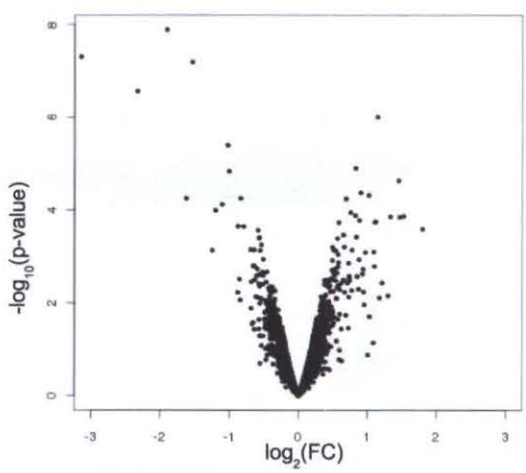
(c) NUSE boxplots



(d) Hierarchical clustering plot



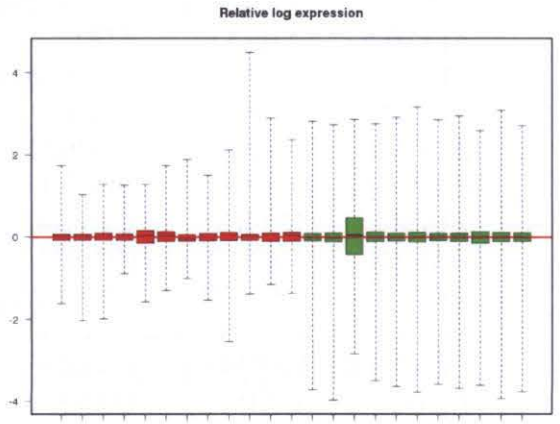
(e) MA plot



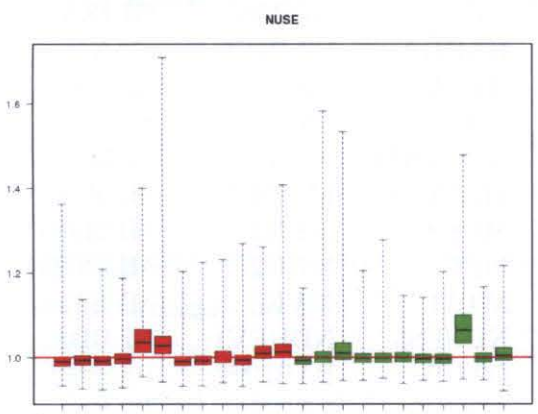
(f) Volcano plot

FIGURE A.3: QC and analysis plots for the 6 samples in the Grayson data analysis.

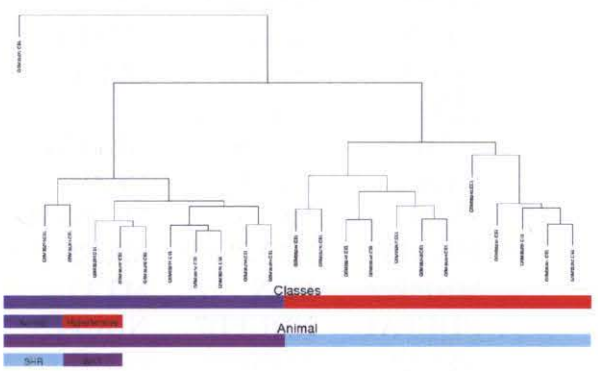
(a) Spatial image plots



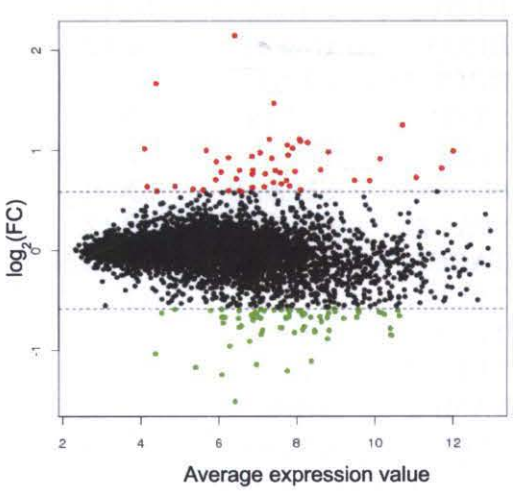
(b) RLE boxplots



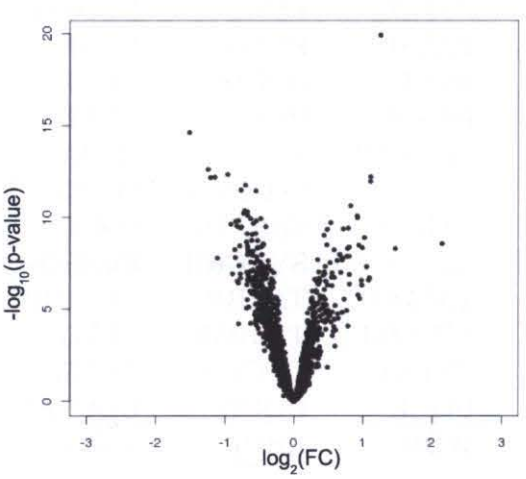
(c) NUSE boxplots



(d) Hierarchical clustering plot



(e) MA plot



(f) Volcano plot

FIGURE A.4: QC and analysis plots for the 23 samples in the Rysä data analysis.

House-keeping gene symbols					
ABL1	ADAM15	ADAMTSL2	ADAR	ADRBK1	AES
AFG3L2	AHSA1	AIF1	ALG3	AP2M1	AP2S1
API5	ARAF	ARHGDIA	ARHGEF7	ARL2	ARPC2
ARPC3	ARPC4	ATF4	ATP5A1	ATP5G3	ATP5I
ATP5J2	ATP5O	ATP6AP1	ATP6V0B	ATP6V1E1	ATP6V1F
B2M	BANF1	BECN1	BLOC1S1	BMI1	BSG
BUD31	C1QL1	C21orf33	C2orf24	CANX	CAPNS1
CAPZB	CASC3	CCBP2	CCT3	CCT7	CD40
CDA	CENPB	CHMP2A	CIZ1	CLOCK	CLSTN1
CLTA	CNTN1	COL6A1	COPE	COPS6	COX6A1
COX7A2L	COX8A	CSF1	CSTB	CTBP1	CTNNB1
DAD1	DAP	DAXX	DAZAP2	DDOST	DDT
DHCR7	DKK4	DNPEP	DRAP1	DULLARD	E2F4
EFNA3	EIF3C	EIF3D	EIF3F	EIF3G	EIF3I
EIF3K	EIF4A2	ERH	ERP29	EXTL3	FBL
FBXO7	FCER2	FOLR1	FOXM1	FUS	GAS1
GM2A	GNB2L1	GOT2	GPI	GPR56	GRIK5
GTPBP6	GUK1	H2AFY	HADHA	HADHB	HAX1
HDGF	HINT1	HSPA5	HYOU1	ID3	IDH3B
IER2	ILK	ISLR	JAG1	JAK1	JTB
JUND	KARS	KIAA0174	KIAA0494	KIF1C	LAMP1
LASP1	LMTK2	LTBP4	MANF	MAP4	MAZ
MC2R	MCL1	MCM3AP	MDH1	MFN2	MFSD10
MGAT1	MLEC	MLF2	MPG	MRC2	MRPL23
MRPS12	MSN	MT3	MTA1	MVK	MYST2
NDUFA1	NDUFB7	NDUFC1	NDUFS5	NDUFV1	NEDD8
NFKBIA	NONO	ODC1	OTUB1	PABPN1	PAK4
PAX8	PCGF2	PDAP1	PDCD6	PFDN5	PFN1
PHB2	PHF1	PICK1	PIN1	PITPNM1	POLR2A
POLR2F	POLR2L	PPP1R10	PRKCSH	PRKD2	PRPF8
PRPH	PSMB1	PSMB2	PSMB4	PSMB7	PSMD11
PSMD2	PSMD3	PTDSS1	PTOV1	PTTG1IP	PUF60
RASSF7	RBM8A	RERE	RHOA	RNF44	RNH1
RPL13	RPL18	RPL3	RPL36AL	RPN1	RPS10
RPS16	RPS5	RUVBL2	SAFB	SAP18	SARS
SCAMP3	SDC3	SDHA	SEPT2	SEPT7	SFRS17A
SGSH	SLC25A1	SLC25A11	SLC6A7	SLC6A8	SND1
SNRPA	SNRPD2	SOD1	SPAG7	SREBF1	SSR2
ST5	SYNCRIP	SYNPO	TACC1	TADA3L	TAGLN
TALDO1	TAPBP	TAX1BP3	TBCB	TCF25	TEX261
TIMM44	TMBIM6	TPMT	TRAP1	TSFM	TSHZ1
TSTA3	TTC1	TUFM	UBE2D2	UBE2I	UBE2M
UQCR	UQCRC1	UQCRFS1	VAMP3	VARS	VEGFB
WDR1	XBP1	YARS	ZFPL1	ZNF592	ZNHIT1

TABLE A.1: 264 human house-keeping genes with rat homologs used in the RUV mega-method adjustment and as evaluation genes when comparing different mega-analysis methods.

Appendix B

Further results from Chapter 3

B.1 Further results from imputation comparison study

In the main text for the simulated imputation results in Section 3.2.4, results from two variables, 'past miscarriages' and 'clots' were presented. For completeness, the simulation results for the remaining 10 variables are included.

Variables which are consistently redundant in the final model yield results similar to 'past-miscarriages', include:

Presence of abdominal pain,

Maternal age,

CRL,

Consistent with menstrual dates,

Number of natural deliveries and

Smoker.

Variables which are important in the final model produce results similar to 'clots', include:

Presence of bleeding,

FHR,

mean GS and

Gestational age in days.

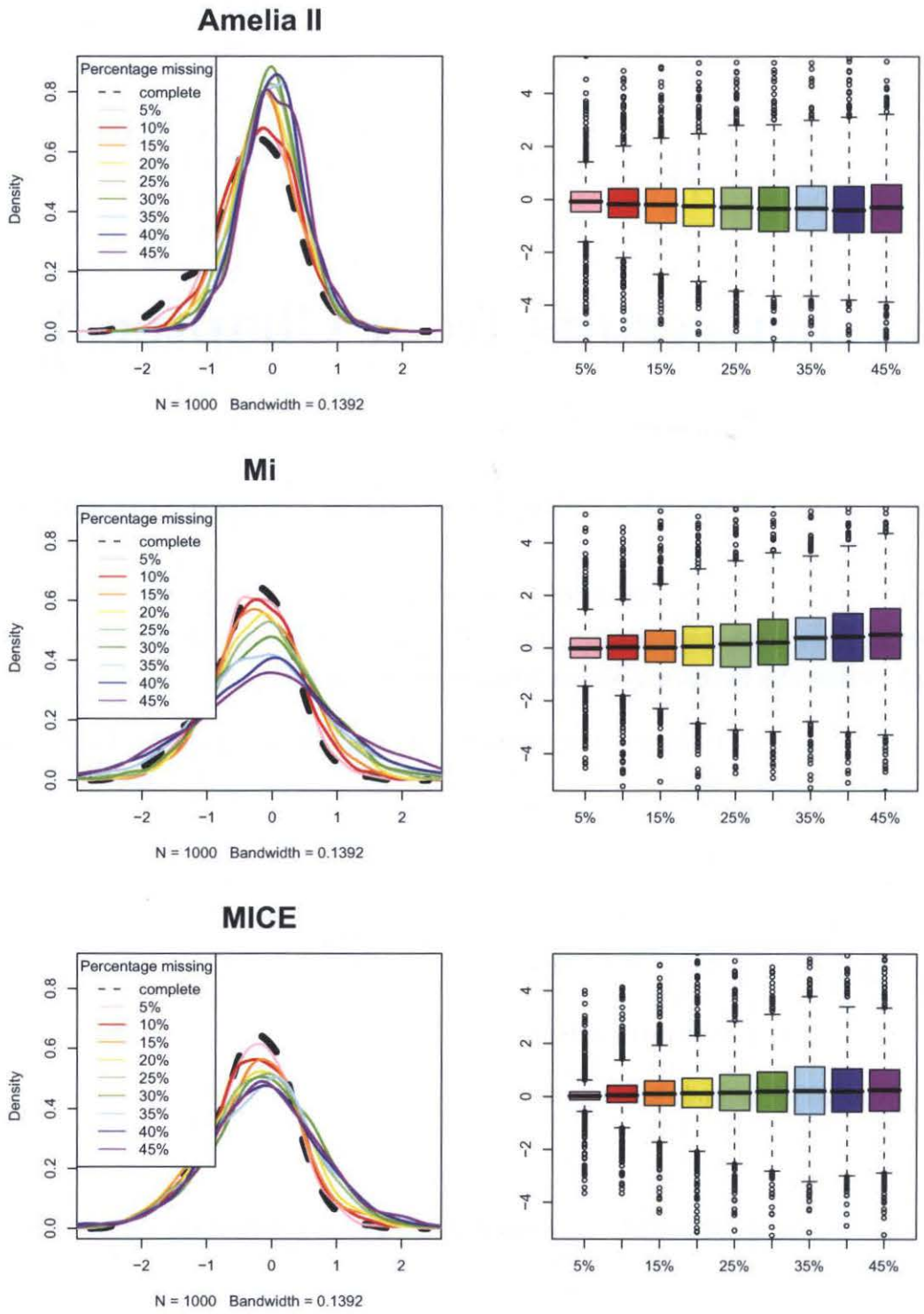


FIGURE B.1: Bootstrapped distributions and boxplots for estimated coefficients for the **'presence of abdominal pain'** variable as the amount of missingness increases. Boxplots depict the log-ratios for imputed data set coefficients to complete data set coefficients, paired by bootstrap draw. Imputation methods used include Amelia II, Mi and MICE.

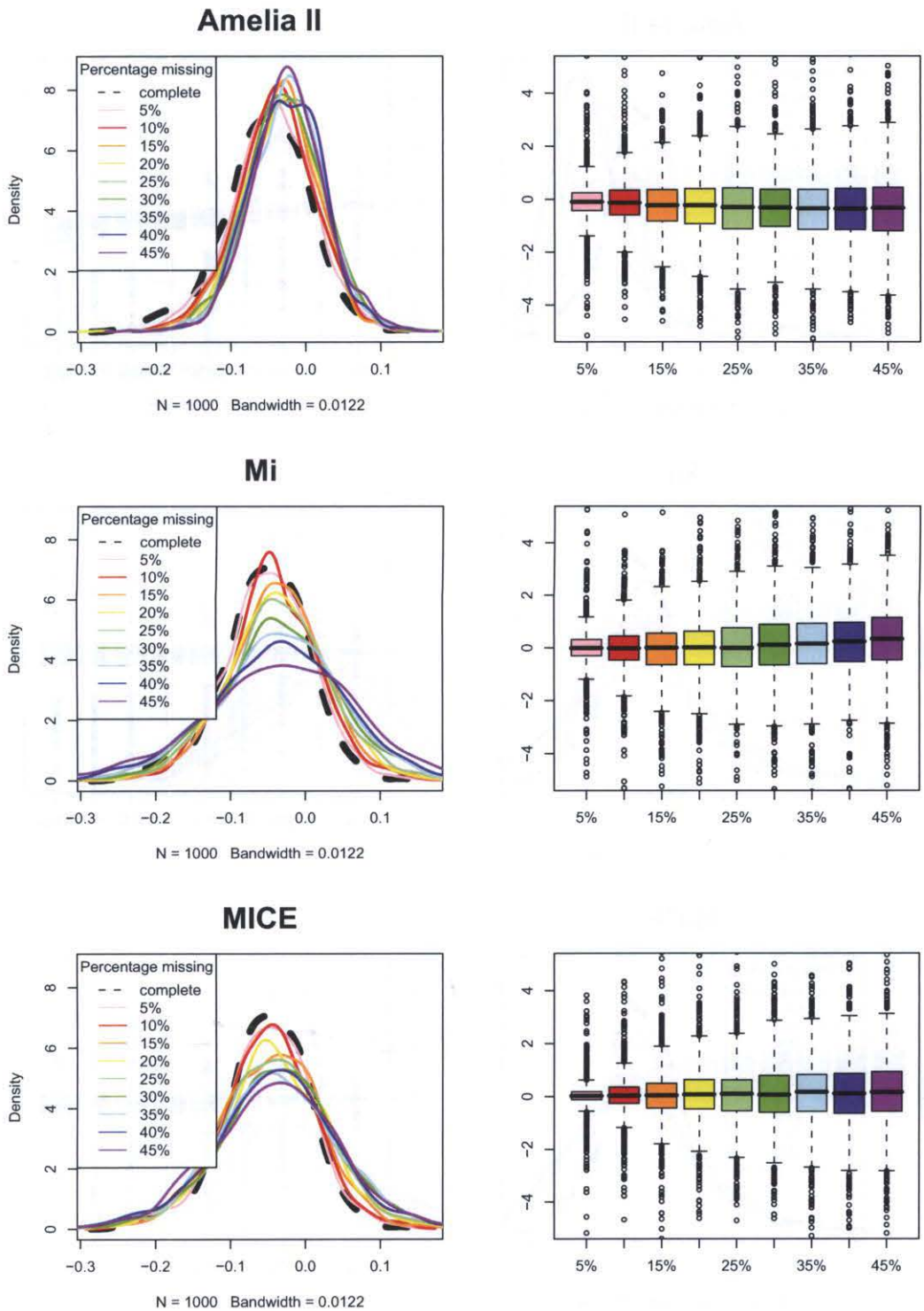


FIGURE B.2: Bootstrapped distributions and boxplots for estimated coefficients for the **'maternal age'** variable as the amount of missingness increases. Boxplots depict the log-ratios for imputed data set coefficients to complete data set coefficients, paired by bootstrap draw. Imputation methods used include Amelia II, Mi and MICE.

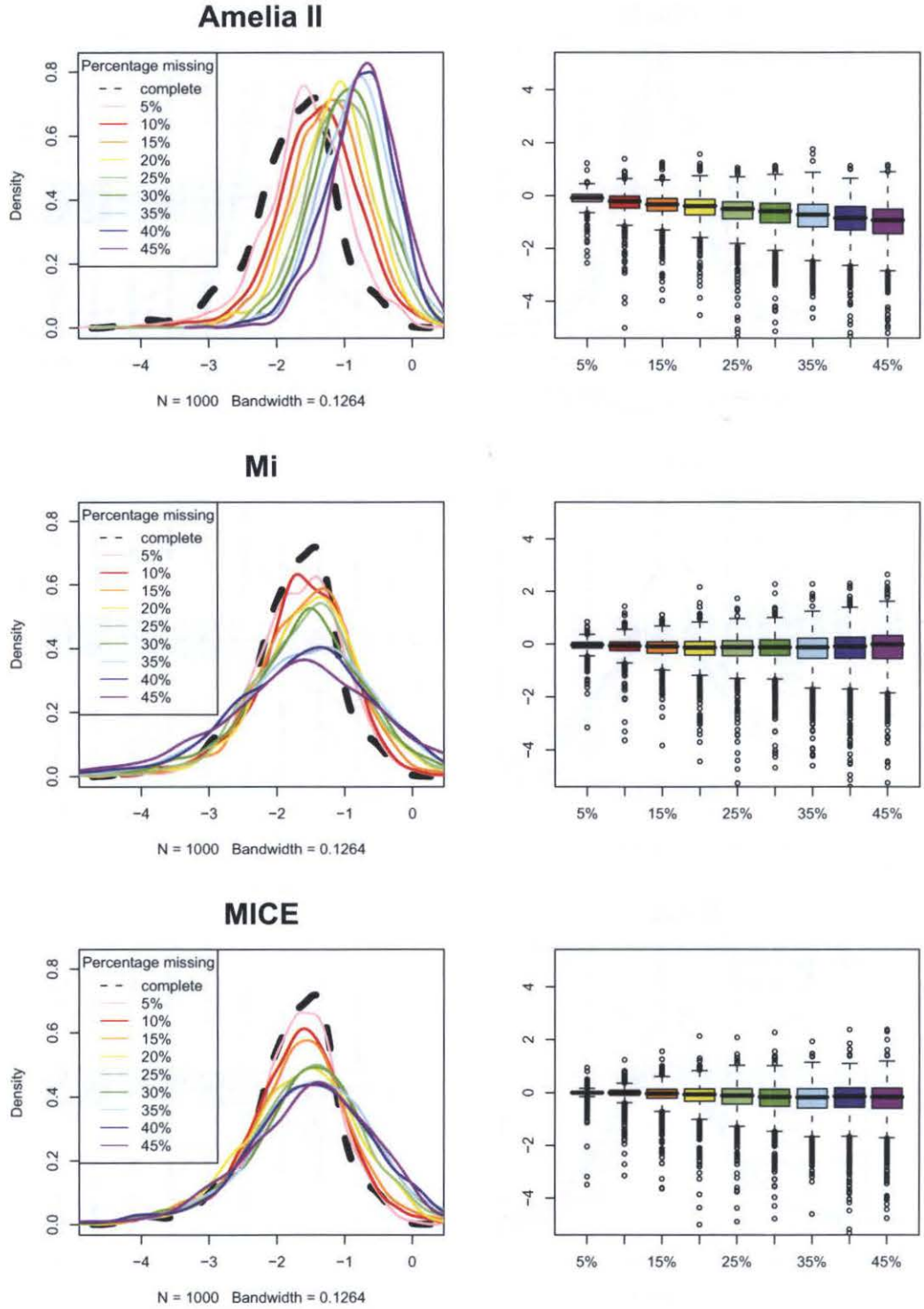


FIGURE B.3: Bootstrapped distributions and boxplots for estimated coefficients for the **'presence of bleeding'** variable as the amount of missingness increases. Boxplots depict the log-ratios for imputed data set coefficients to complete data set coefficients, paired by bootstrap draw. Imputation methods used include Amelia II, Mi and MICE.

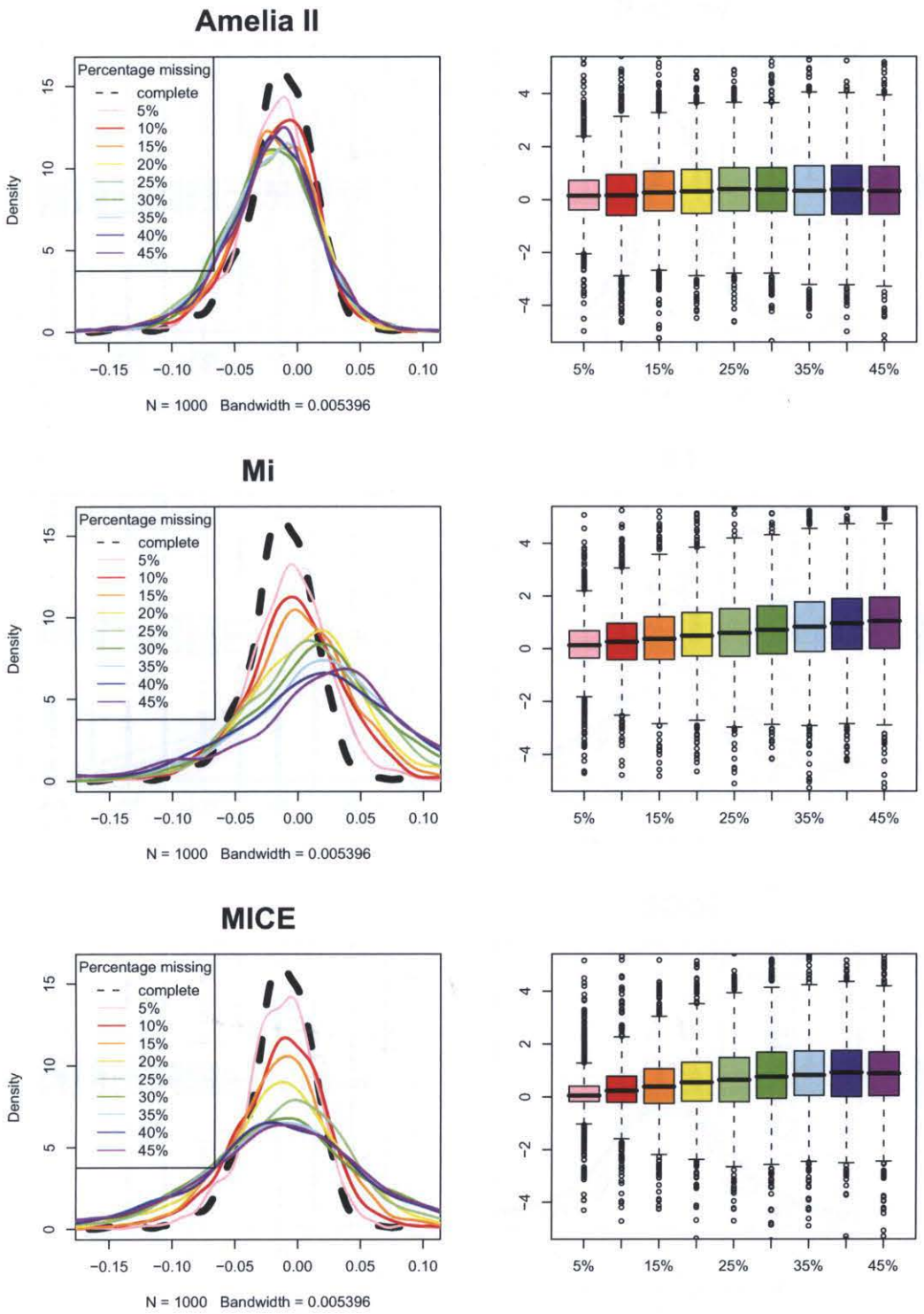


FIGURE B.4: Bootstrapped distributions and boxplots for estimated coefficients for the **'CRL'** variable as the amount of missingness increases. Boxplots depict the log-ratios for imputed data set coefficients to complete data set coefficients, paired by bootstrap draw. Imputation methods used include Amelia II, Mi and MICE.

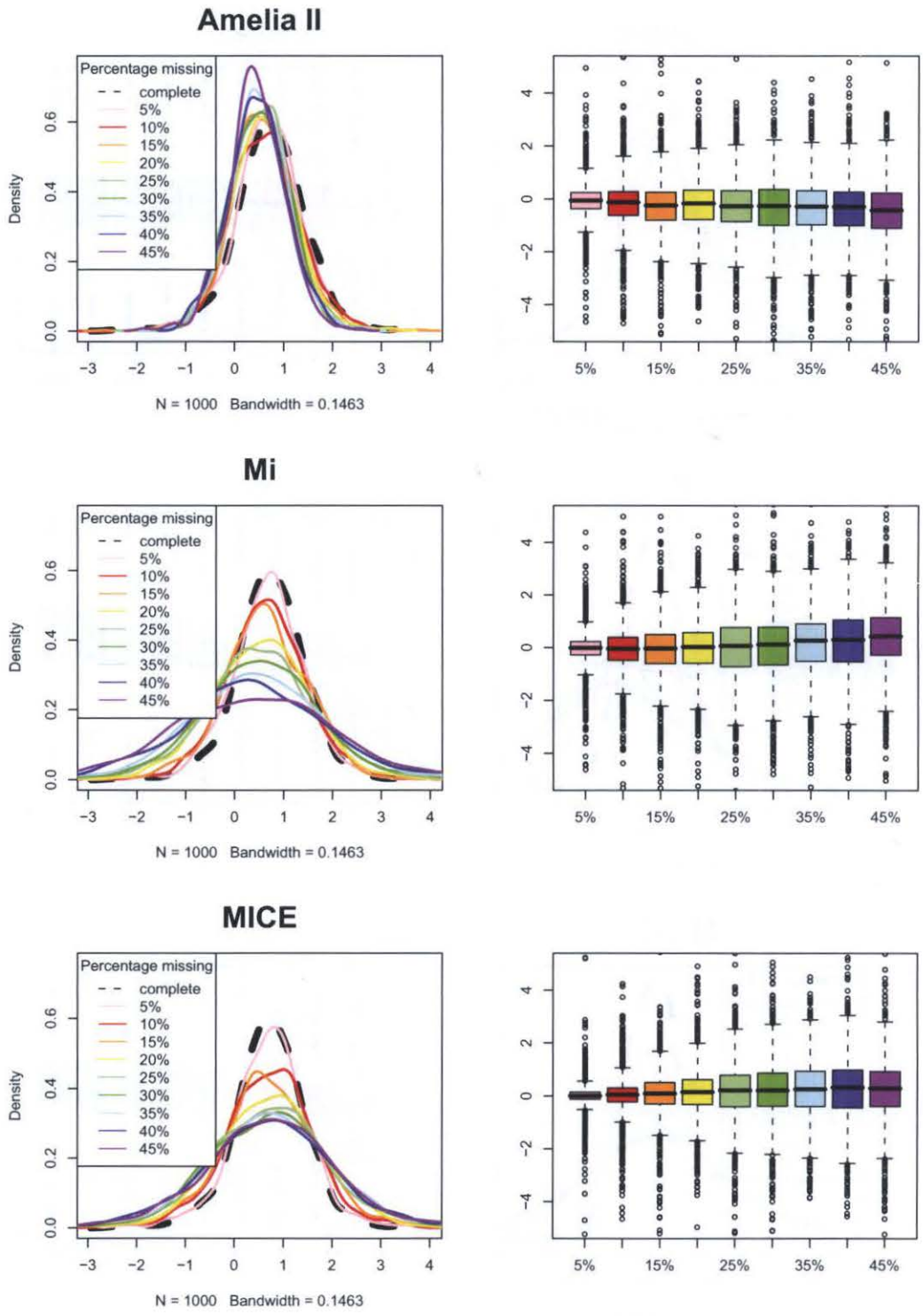


FIGURE B.5: Bootstrapped distributions and boxplots for estimated coefficients for the **'consistent with menstrual dates'** variable as the amount of missingness increases. Boxplots depict the log-ratios for imputed data set coefficients to complete data set coefficients, paired by bootstrap draw. Imputation methods used include Amelia II, Mi and MICE.

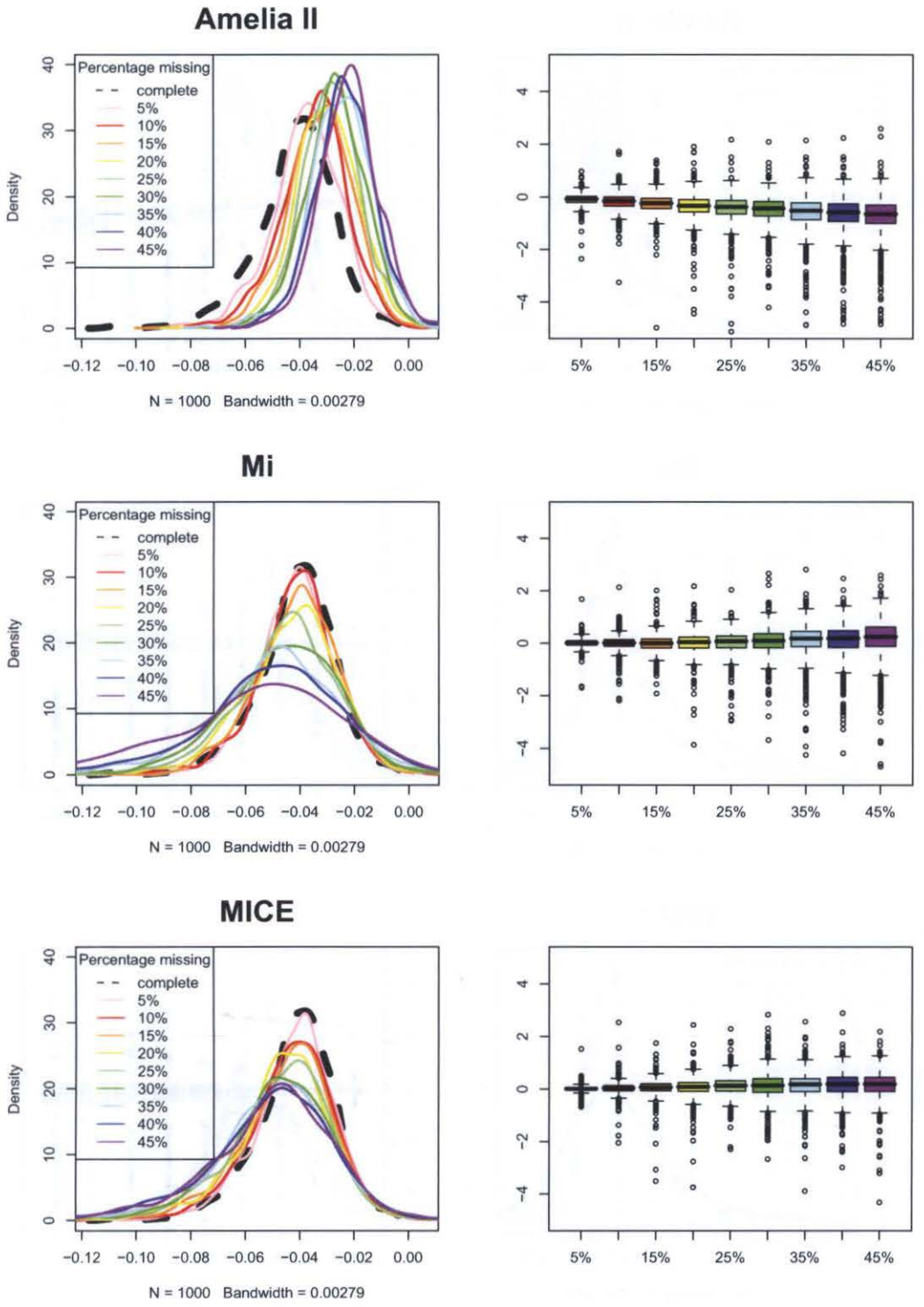


FIGURE B.6: Bootstrapped distributions and boxplots for estimated coefficients for the **'FHR'** variable as the amount of missingness increases. Boxplots depict the log-ratios for imputed data set coefficients to complete data set coefficients, paired by bootstrap draw. Imputation methods used include Amelia II, Mi and MICE.

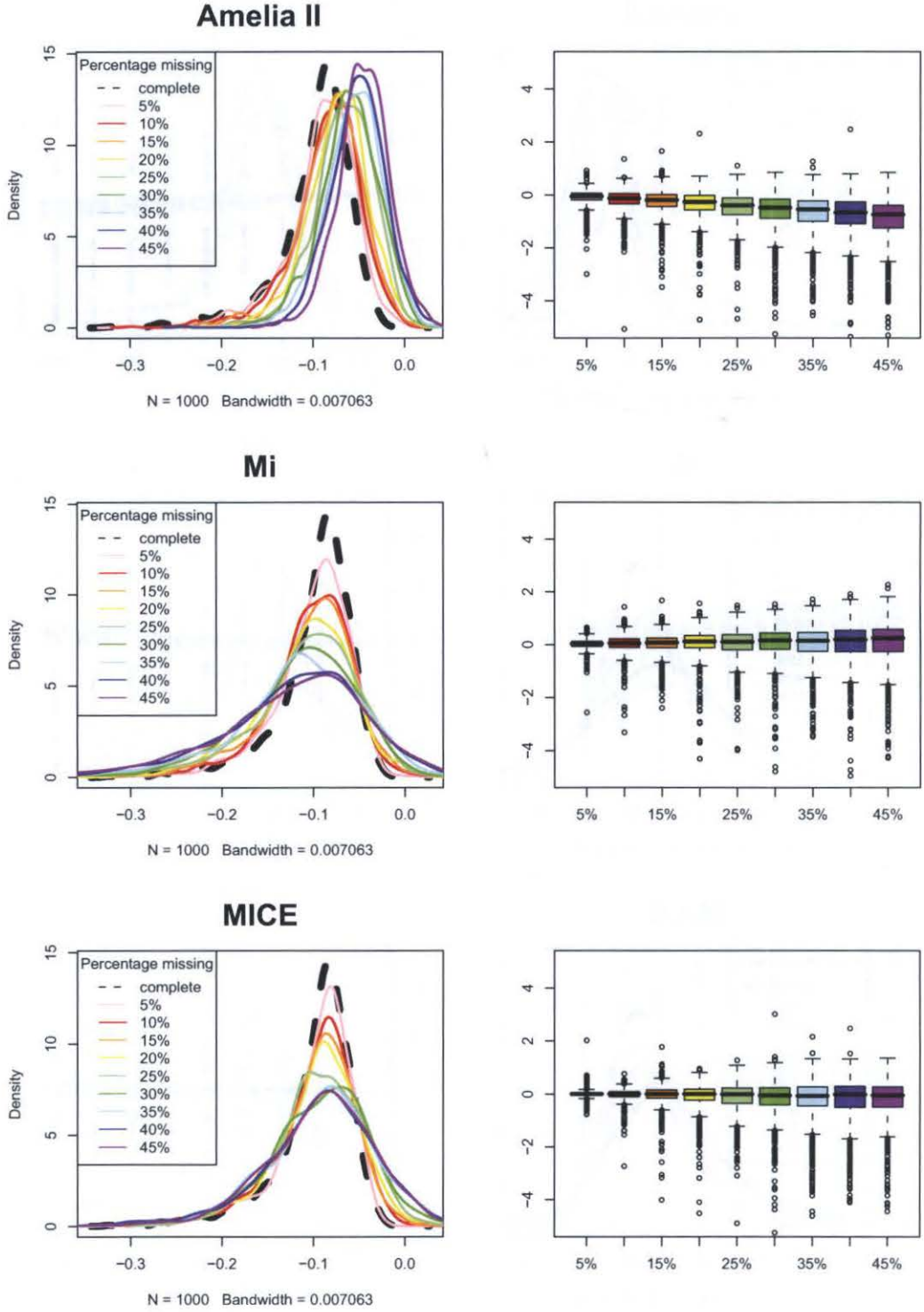


FIGURE B.7: Bootstrapped distributions and boxplots for estimated coefficients for the **'mean GS'** variable as the amount of missingness increases. Boxplots depict the log-ratios for imputed data set coefficients to complete data set coefficients, paired by bootstrap draw. Imputation methods used include Amelia II, Mi and MICE.

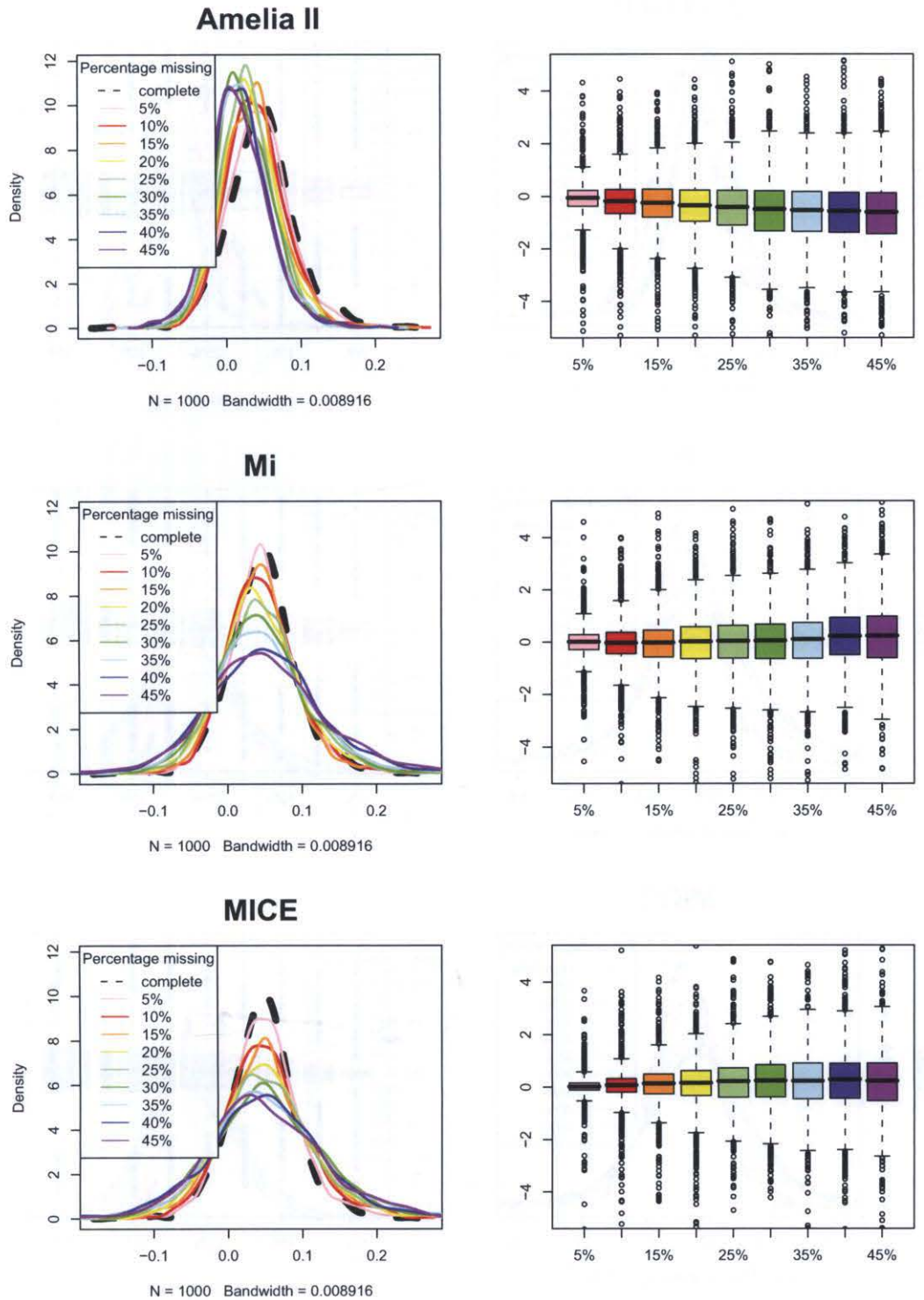


FIGURE B.8: Bootstrapped distributions and boxplots for estimated coefficients for the **'gestational age in days'** variable as the amount of missingness increases. Boxplots depict the log-ratios for imputed data set coefficients to complete data set coefficients, paired by bootstrap draw. Imputation methods used include Amelia II, Mi and MICE.

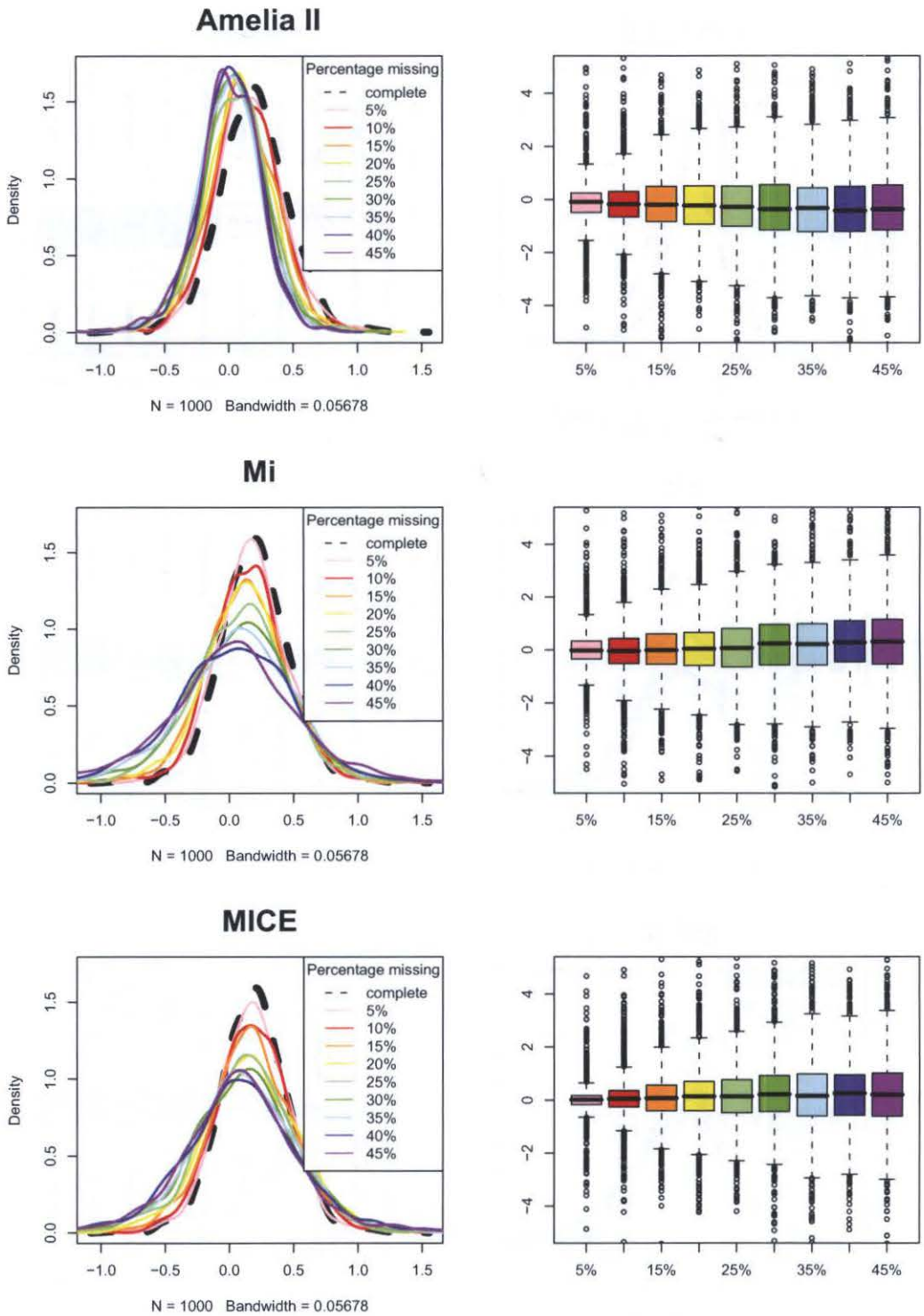


FIGURE B.9: Bootstrapped distributions and boxplots for estimated coefficients for the **'number of natural deliveries'** variable as the amount of missingness increases. Boxplots depict the log-ratios for imputed data set coefficients to complete data set coefficients, paired by bootstrap draw. Imputation methods used include Amelia II, Mi and MICE.

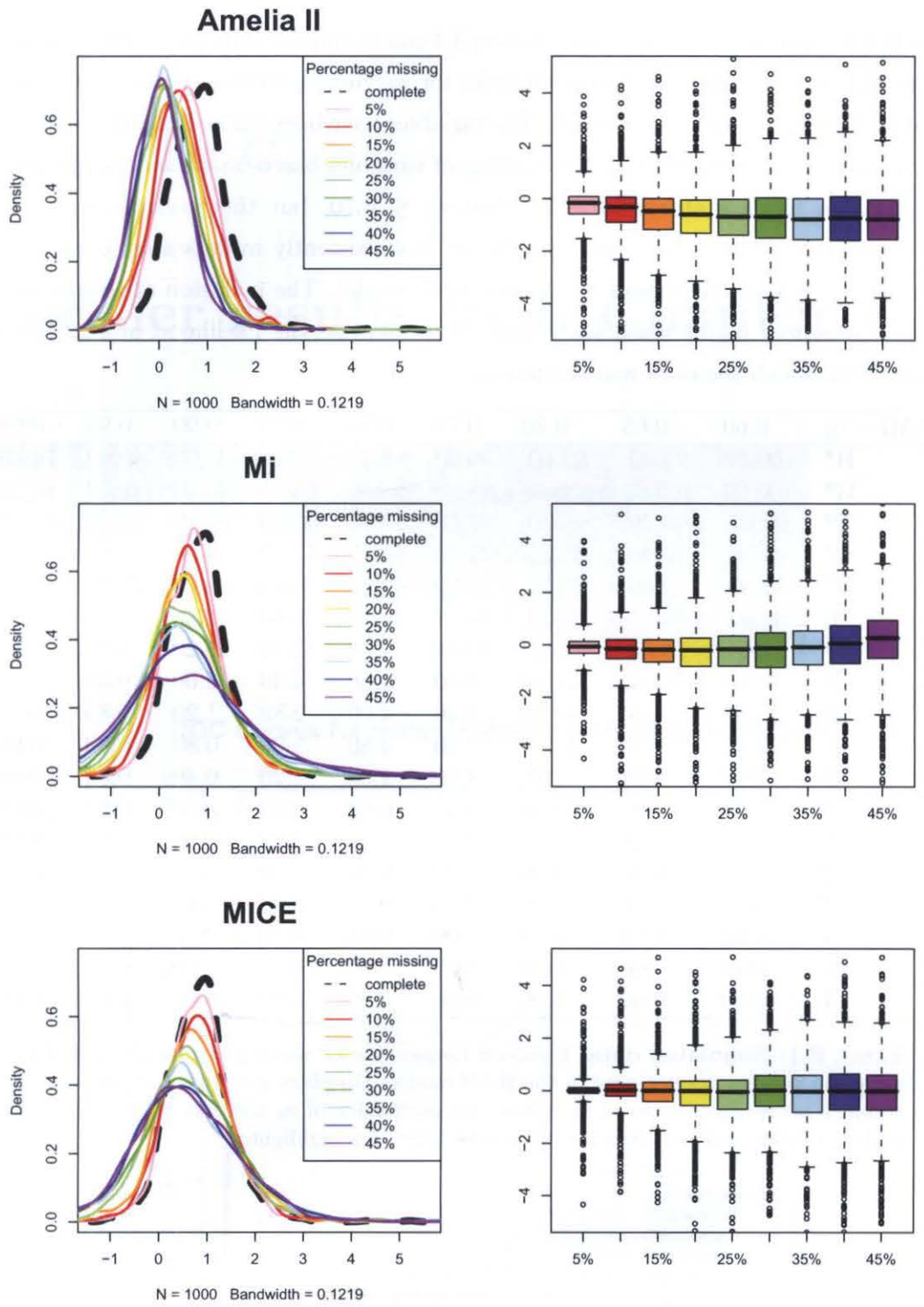


FIGURE B.10: Bootstrapped distributions and boxplots for estimated coefficients for the **'smoker'** variable as the amount of missingness increases. Boxplots depict the log-ratios for imputed data set coefficients to complete data set coefficients, paired by bootstrap draw. Imputation methods used include Amelia II, Mi and MICE.

B.2 Further results from the B-MI approach

The B-MI approach as presented in Section 3.4 can be applied with and without weights. Table B.1 contains the inclusion frequencies for varying τ_B values when weights are not used in the logistic regression model. The variables have been ranked in order of stability across all considered τ_B values. The ranking of variables based on stability suggests that the variable 'E' is a stable variable when $\tau_B \leq 0.70$, but this variable is not from the simulated model. Moreover, variable 'E' is consistently more stable, overall, than variable 'P', a variable present in the simulated model. The inclusion of weights in the model, presented in the main text, appears to increase the likelihood of selecting the model from which the data was simulated.

B-MI – τ_B	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.98	1.00
H*	100.00	99.60	99.60	99.20	99.20	99.20	98.00	89.60	70.40	0.00
M*	100.00	100.00	100.00	98.80	98.80	96.40	92.40	76.00	46.40	0.00
T*	99.60	99.20	99.20	97.60	94.80	90.40	84.40	66.40	38.40	0.00
C*	96.00	92.40	85.20	77.20	66.80	52.80	33.20	14.80	3.60	0.00
G*	88.80	86.80	81.60	74.80	64.00	50.80	38.00	22.40	7.60	0.00
E	82.00	74.00	67.60	54.40	46.80	35.60	19.60	9.60	3.20	0.00
P*	82.80	76.80	65.20	53.60	44.80	32.40	25.20	9.60	2.00	0.00
D	28.00	20.40	14.80	10.80	8.80	4.40	2.00	0.40	0.00	0.00
B	18.80	14.40	10.00	6.40	4.00	2.80	1.20	0.80	0.40	0.00
L	24.00	20.40	12.40	7.20	4.80	1.20	0.80	0.00	0.00	0.00
U	16.00	9.20	4.80	4.00	2.00	1.20	0.40	0.00	0.00	0.00
S	1.60	1.20	0.40	0.40	0.00	0.00	0.00	0.00	0.00	0.00
F	2.40	1.20	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q	1.20	0.80	0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00
J	1.60	0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N	0.80	0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
K	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

TABLE B.1: **Simulated data:** Inclusion frequencies for varying inclusion threshold τ_B, when weights are not used in the B-MI model. Variables are ranked in order of stability across all considered τ_B values. For each value of τ_B the variables considered stable, as their inclusion frequency is above $1/2\tau_B$ are highlighted.

Appendix C

Further results from Chapter 4

C.1 Additional results for Case study 1

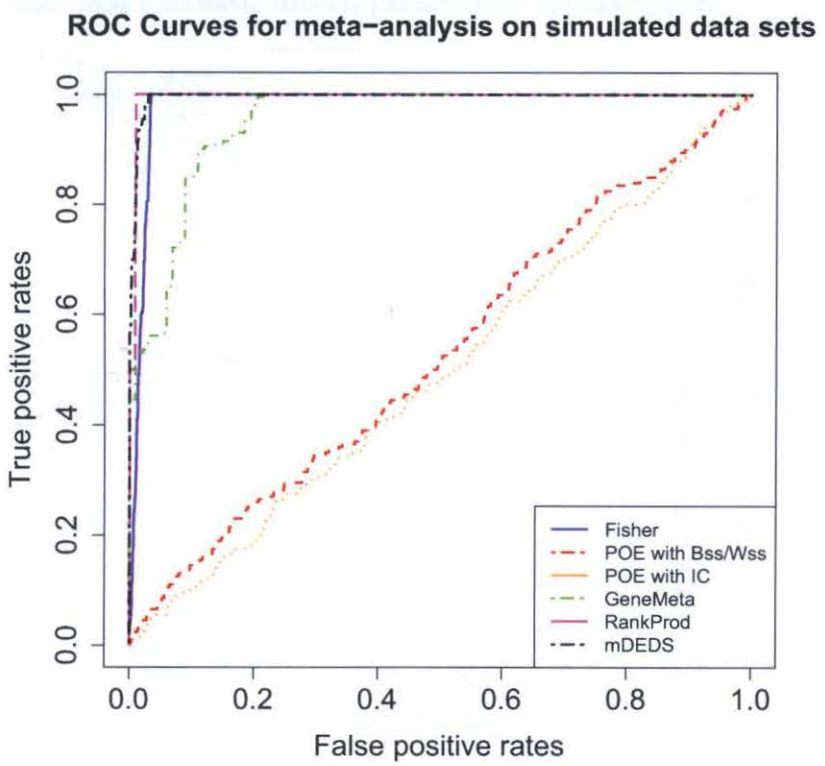


FIGURE C.1: ROC plots for simulated data using different meta-analysis methods for the 10% DE gene level (5% true, 5% platform specific DE genes) simulation. For further detail see Section 4.4.

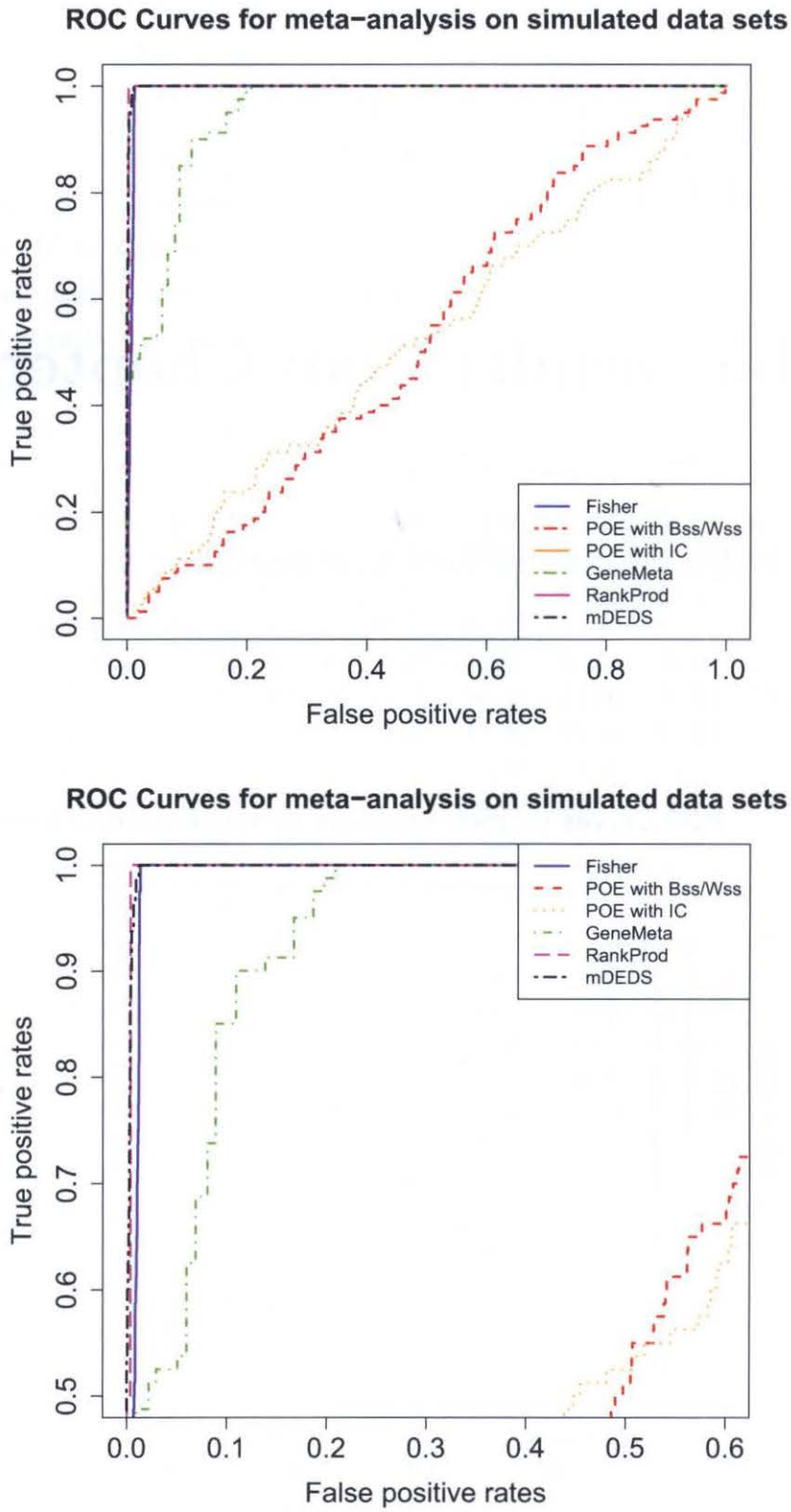


FIGURE C.2: ROC plots for simulated data using different meta-analysis methods for the 2.5% DE gene level (1.25% true, 1.25% platform specific DE genes) simulation. The lower plot is a zoomed in version of the upper plot. For further detail see Section 4.4.

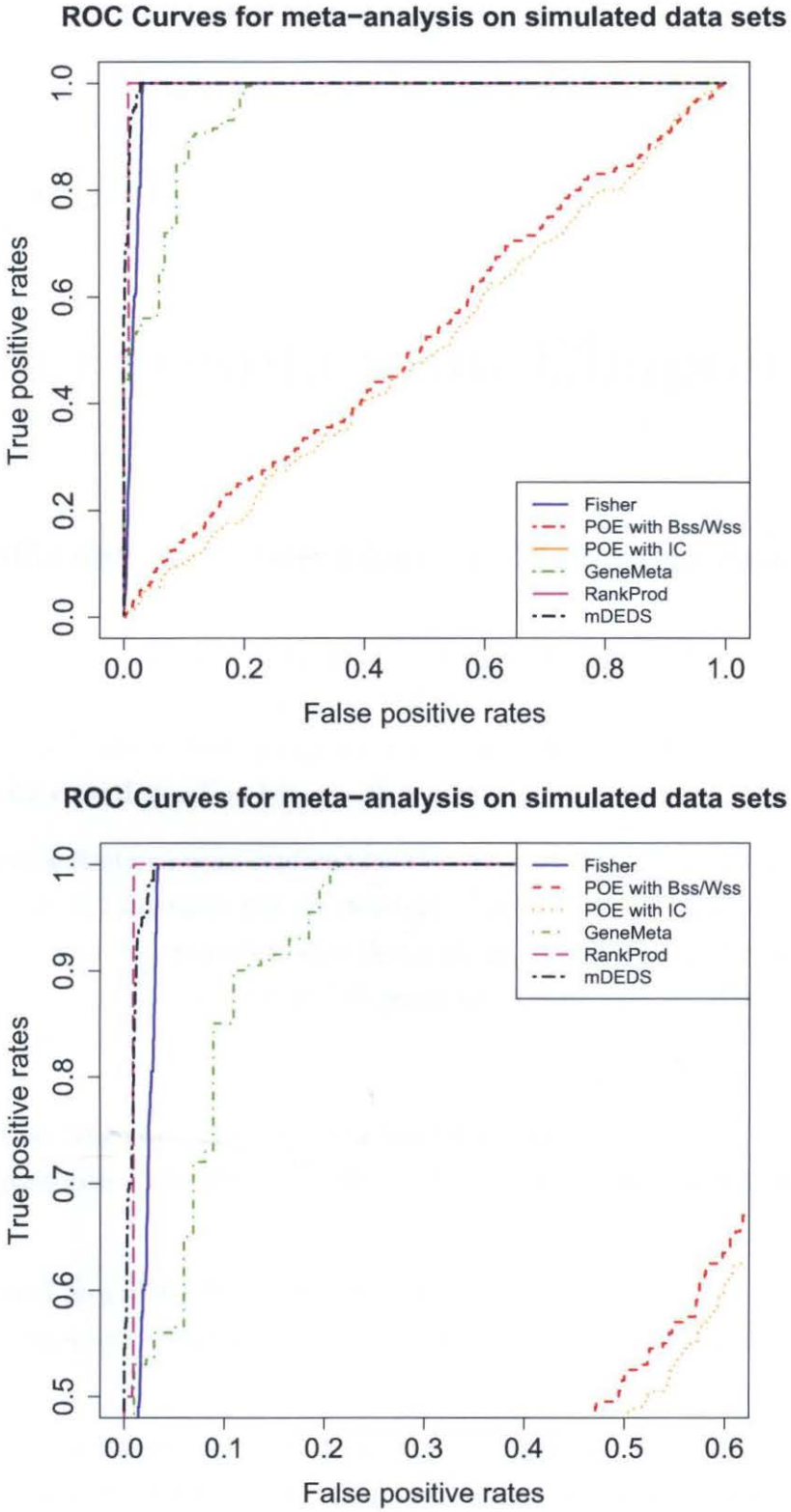


FIGURE C.3: ROC plots for simulated data using different meta-analysis methods for the 4% DE gene level (2% true, 2% platform specific DE genes) simulation. The lower plot is a zoomed in version of the upper plot. For further detail see Section 4.4.

Appendix D

Further results from Chapter 5

D.1 Different class definitions for Melanoma case study

In Chapter 5, the samples were reduced to the two extreme cases. Samples were defined as having a 'good' prognosis if patients survived more than four years with no sign of relapse ($n = 23$), and a 'bad' prognosis if they died within a year due to melanoma ($n = 25$). This class distinction left $n = 48$ samples.

Different class definitions were explored for this data set during the initial phase of the analysis. Table D.1 contains the six different class definitions explored. A stringent class definition, case (f), was selected so that a strong molecular signature was obtained, resulting in a reasonable number of DE genes for downstream analysis. Six different class definitions were used:

(a) Good prognosis - Survived four years ($n = 34$);
Poor prognosis - Died for any reason within two years (less than 24 months) ($n = 37$).

(b) Good prognosis - Survived four years ($n = 34$);
Poor prognosis - Died for any reason within one year (less than 12 months) ($n = 26$).

(c) Good prognosis - Survived four years ($n = 34$);
Poor prognosis - Died due to melanoma within two years (less than 24 years) ($n = 34$).

(d) Good prognosis - Survived four years ($n = 34$);
Poor prognosis - Died due to melanoma within one year (less than 12 months) ($n = 25$).

(e) Good prognosis - Survived four years with no relapse; ($n = 23$)
Poor prognosis - Died due to melanoma within two years (less than 24 months) ($n = 34$).

(f) Good prognosis - Survived four years with no relapse; ($n = 23$)
Poor prognosis - Died due to melanoma within one year (less than 12 months) ($n = 25$).

	Definition				
	Good prognosis	Bad prognosis	FDR < 0.05	FDR < 0.1	|FC| > 2
(a)	Survived four years	Died for any reason within two years	0	0	31
(b)	Survived four years	Died for any reason within one year	62	127	82
(c)	Survived four years	Died due to melanoma within two years	0	14	44
(d)	Survived four years	Died due to melanoma within one year	34	115	74
(e)	Survived four years with no relapse	Died due to melanoma within two years	238	512	122
(f)	**Survived four years with no relapse**	**Died due to melanoma within one year**	**253**	**512**	**141**

TABLE D.1: Number of genes selected as DE using different class definitions and selection criterion.

D.2 Different feature selection methods for expression data

In Chapter 5, feature selection of the classification of the expression data was performed using a median robust method, where genes were ranked based on the differences between the two group medians, that is $(\tilde{x}_{\text{good}} - \tilde{x}_{\text{poor}})$, where $\tilde{x}$ represents the median of a group. Two other common feature selection methods were used; (i) genes were ranked based on their robust linear coefficients, applied through `limma` (Smyth and Wettenhall, 2003) and; (ii) genes were ranked based on the ratio of their between sum of squares value over their within sum of squares value (Bss/Wss) (Dudoit and Fridlyand, 2003). Results of these feature selection methods are seen in Figure D.1. For each of these three methods

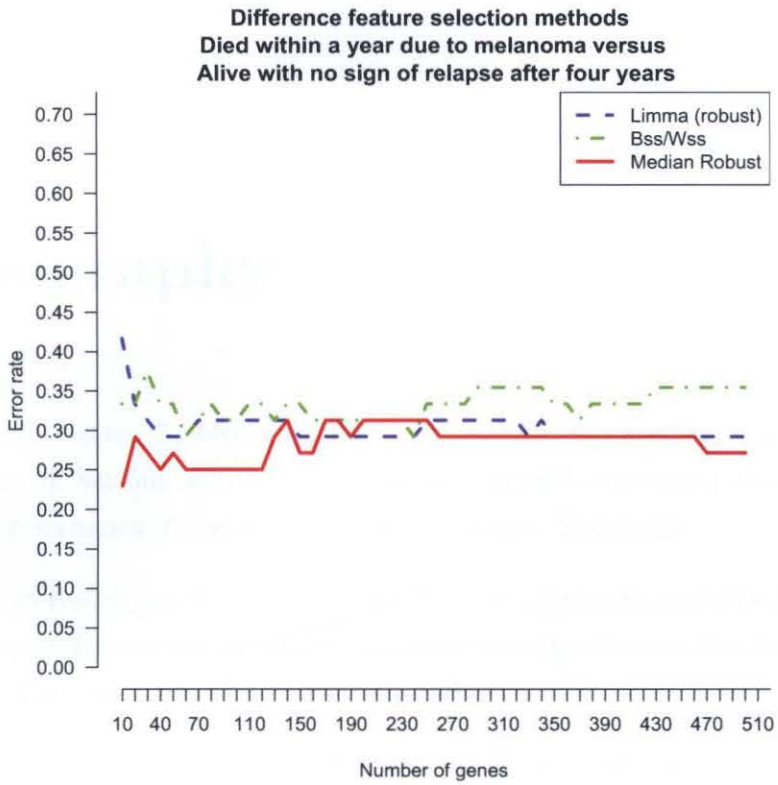


FIGURE D.1: Varying feature selection methods for the classification of the expression samples from the melanoma case study, where a good prognosis is defined as surviving with no sign of relapse for four or more years, and a bad prognosis is defined as dying due to melanoma within a year.

the DLDA classifier was used to construct the discriminant rule based on the top ranked genes. The number of genes used varied from 10 to 500 along the horizontal axis, and error rates were obtained using LOOCV, along the vertical axis.

Bibliography

Abrahantes, J., Sotto, C., Molenberghs, G., Vromman, G., and Biernckx, B. (2011). A comparison of various software tools for dealing with missing data via imputation. *Journal of Statistical Computation and Simulation*, To appear.

Acuna, E. and Rodriguez, C. (2004). *Classification, clustering and data mining applications*, chapter The treatment of missing values and its effect in the classifier accuracy, pages 639–648. Springer-Verlag, Heidelberg Berlin.

Agresti, A. (2007). *An Introduction to categorical data analysis.* Wiley, Hoboken, NJ, 2nd edition.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511.

Allison, D. B., Cui, X., Page, G. P., and Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, 7:55–65.

Altman, D. G. and Andersen, P. K. (1989). Bootstrap investigation of the stability of a Cox regression model. *Statistics in Medicine*, 8:771–783.

Archer, S. L., Marsboom, G., Kim, G. H., Zhang, H. J., Toth, P. T., Svensson, E. C., Dyck, J. R. B., Gomberg-Maitland, M., Thébaud, B., Husain, A. N., Cipriani, N., and Rehman, J. (2010). Epigenetic attenuation of mitochondrial superoxide dismutase 2 in pulmonary arterial hypertension: a basis for excessive cell proliferation and a new therapeutic target. *Circulation*, 121:2661–2671.

Auer, H., Newsom, D. L., and Kornacker, K. (2009). Expression Profiling Using Affymetrix GeneChip Microarrays. *Methods in Molecular Biology*, 509:35–46.

Austin, P. C. and Tu, J. V. (2004). Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *Journal of Clinical Epidemiology*, 57:1138–1146.

Azur, M., Stuart, E., Fragakis, C., and Leaf, P. (2011). Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20:40–49.

Balch, C. M., Buzaid, A. C., Soong, S. J., Atkins, M. B., Cascinelli, N., Coit, D. G., Fleming, I. D., Gershenwald, J. E., Houghton, A., Kirkwood, J. M., McMasters, K. M., Mihm, M. F., Morton, D. L., Reintgen, D. S., Ross, M. I., Sober, A., Thompson, J. A., and Thompson, J. F. (2001). Final version of the American Joint Committee on Cancer staging system for cutaneous melanoma. *Journal of Clinical Oncology*, 19:3635–3648.

Balch, C. M. and Soong, S.-J. (2008). Predicting outcomes in metastatic melanoma. *Journal of Clinical Oncology*, 26:168–169.

Barnes, S. A., Mallinckrodt, C. H., Lindborg, S. R., and Carter, M. K. (2008). The impact of missing data and how it is handled on the rate of false-positive results in drug development. *Pharmaceutical Statistics*, 7:215–225.

Barrett, G. L. and Mullins, J. J. (1992). Studies on blood pressure regulation in hypertensive REN-2 transgenic rats. *Kidney International*, 37:S125–S128.

Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W.-C., Ledoux, P., Rudnev, D., Lash, A. E., Fujibuchi, W., and Edgar, R. (2005). NCBI GEO: mining millions of expression profiles–database and tools. *Nucleic Acids Research*, 33:D562–D566.

Beissbarth, T. and Speed, T. P. (2004). GOstat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, 20:1464–1465.

Berezhnaya, N. M. (2010). Interaction between tumor and immune system: the role of tumor cell biology. *Experimental Oncology*, 32:159–166.

Bertram, C. E. and Hanson, M. A. (2002). Prenatal programming of postnatal endocrine responses by glucocorticoids. *Reproduction*, 124:459–467.

Beyene, J., Atenafu, E. G., Hamid, J. S., To, T., and Sung, L. (2009). Determining relative importance of variables in developing and validating predictive models. *BMC Medical Research Methodology*, 9:64.

Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., Gillanders,

E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., and Sondak, V. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406:536–540.

Bogunovic, D., O'Neill, D. W., Belitskaya-Levy, I., Vacic, V., Yu, Y.-L., Adams, S., Darvishian, F., Berman, R., Shapiro, R., Pavlick, A. C., Lonardi, S., Zavadil, J., Osman, I., and Bhardwaj, N. (2009). Immune profile and mitotic index of metastatic melanoma lesions enhance clinical staging in predicting patient survival. *Proceedings of the National Academy of Sciences*, 106:20429–20434.

Bosotti, R., Locatelli, G., Healy, S., Scacheri, E., Sartori, L., Mercurio, C., Calogero, R., and Isacchi, A. (2007). Cross platform microarray analysis for robust identification of differentially expressed genes. *BMC Bioinformatics*, 8 Suppl 1:S5.

Boulesteix, A.-L., Porzelius, C., and Daumer, M. (2008). Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. *Bioinformatics*, 24(15):1698–1706.

Boulesteix, A. L. and Slawski, M. (2009). Stability and aggregation of ranked gene lists. *Briefings in Bioinformatics*, 10:556–568.

Bramer, W. L., White, A., Thompson, S., and M.A. (1997). *Advances in intelligent data analysis*, chapter Techniques for dealing with missing values in classification, pages 527–536. Springer-Verlag, Heidelberg Berlin.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annuals of Statistics*, 24:2350–2383.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. Chapman & Hall/CRC, New York, 1st edition.

Breitling, R., Armengaud, P., Amtmann, A., and Herzyk, P. (2004). Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters*, 573:83–92.

Butte, A. J., Dzau, V. J., and Glueck, S. B. (2001). Further defining housekeeping, or 'maintenance' genes: focus on 'a compendium of gene expression in normal human tissues'. *Physiological Genomics*, 7:95–96.

Buuren, S. V. and Oudshoorn, C. (2007). *MICE: multivariate imputation by chained equations*. R package version 1.16.

Buuren, S. V., Oudshoorn, K. C., and van Rijckevorsel, J. (1999). Flexible multivariate imputation by chained equations of the AVO-95 Survey. Technical report, TNO Prevention and Health.

Cahan, P., Rovegno, F., Mooney, D., Newman, J. C., St Laurent, G., and McCaffrey, T. A. (2007). Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization. *Gene*, 401:12–18.

Campain, A., Müller, S., and Yang, Y. H. (2011). Stable logistic regression models in the presence of missing values and class imbalances. *Biostatistics*, Under review.

Campain, A. and Yang, Y. H. (2010). Comparison study of microarray meta-analysis methods. *BMC Bioinformatics*, 11:408.

Campain, A. E., Marques, F. Z., Yang, Y. H. J., and Morris, B. J. (2010). Meta-analysis of genome-wide gene expression differences in onset and maintenance phases of genetic hypertension. *Hypertension*, 56:319–324.

Canales, R. D., Luo, Y., Willey, J. C., Austermiller, B., Barbacioru, C. C., Boysen, C., Hunkapiller, K., Jensen, R. V., Knight, C. R., Lee, K. Y., Ma, Y., Maqsodi, B., Papallo, A., Peters, E. H., Poulter, K., Ruppel, P. L., Samaha, R. R., Shi, L., Yang, W., Zhang, L., and Goodsaid, F. M. (2006). Evaluation of DNA microarray results with quantitative gene expression platforms. *Nature Biotechnology*, 24:1115–1122.

Cardoso, F., Piccart-Gebhart, M., Van't Veer, L., and Rutgers, E. (2007). The MINDACT trial: the first prospective clinical validation of a genomic tool. *Molecular Oncology*, 1:246–251.

Carlström, M., Brown, R. D., Sällström, J., Larsson, E., Zilmer, M., Zabihi, S., Eriksson, U. J., and Persson, A. E. G. (2009). SOD1 deficiency causes salt sensitivity and aggravates hypertension in hydronephrosis. *American Journal of Physiology - Regulatory, Integrative and Comparative Physiology*, 297:R82–R92.

Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46:167–174.

Cerutti, C., Kurdi, M., Bricca, G., Hodroj, W., Paultre, C., Randon, J., and Gustin, M.-P. (2006). Transcriptional alterations in the left ventricle of three hypertensive rat models. *Physiological Genomics*, 27:295–308.

Cessie, S. L. and Houwelingen, J. C. V. (1992). Ridge Estimators in Logistic Regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41:191–201.

Chan, P. and Stolfo, S. (1988). Toward scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection. In *Proceedings of the fourth international conference on knowledge discovery and data mining*, pages 164–168, Menlo Park CA.

Chargaff, E. (1951). Some recent studies on the composition and structure of nucleic acids. *Journal of Cellular Physiology*, 38:41–59.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A*, 158:419–466.

Chawla, N. V., Bowyer, K. W., and Kegelmeyer, P. W. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Chen, J. J., Hsueh, H.-M., Delongchamp, R. R., Lin, C.-J., and Tsai, C.-A. (2007). Reproducibility of microarray data: a further analysis of microarray quality control (MAQC) data. *BMC Bioinformatics*, 8:412.

Chen, Y., Wen, G., Rao, F., Zhang, K., Wang, L., Rodriguez-Flores, J. L., Sanchez, A. P., Mahata, M., Taupenot, L., Sun, P., Mahata, S. K., Tayo, B., Schork, N. J., Ziegler, M. G., Hamilton, B. A., and O'Connor, D. T. (2010). Human dopamine beta-hydroxylase (DBH) regulatory polymorphism that influences enzymatic activity, autonomic function, and blood pressure. *Journal of Hypertension*, 28:76–86.

Choi, J. K., Yu, U., Kim, S., and Yoo, O. J. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19 Suppl 1:84–90.

Choong, S., Rombauts, L., Ugoni, A., and Meagher, S. (2003). Ultrasound prediction of risk of spontaneous miscarriage in live embryos from assisted conceptions. *Ultrasound in Obstetrics and Gynecology*, 22:571–577.

Claeskens, G. and Hjort, N. L. (2008). Cambridge University Press.

Clemitson, J.-R., Dixon, R. J., Haines, S., Bingham, A. J., Patel, B. R., Hall, L., Lo, M., Sassard, J., Charchar, F. J., and Samani, N. J. (2007). Genetic dissection of a blood pressure quantitative trait locus on rat chromosome 1 and gene expression analysis identifies SPON1 as a novel candidate hypertension gene. *Circulation Research*, 100:992–999.

Consentino, F. and Claeskens, G. (2011). Missing covariates in logistic regression, estimation and distribution selection. *Statistical Modelling*, 11:159–183.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227:561–563.

Cubeddu, L. X., Davila, J., Zschaeck, D., Barbella, Y. R., Ordaz, P., and Dominguez, J. (1981). Cerebrospinal fluid and plasma dopamine-beta-hydroxylase activity in human hypertension. *Hypertension*, 3:448–455.

De Leeuw, F.-E., Richard, F., De Groot, J. C., Van Duijn, C. M., Hofman, A., Van Gijn, J., and Breteler, M. M. B. (2004). Interaction between hypertension, apoE, and cerebral white matter lesions. *Stroke*, 35:1057–1060.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7:177–188.

Dettling, M. and Bühlmann, P. (2004). Finding Predictive Gene Groups from Microarray Data. *Journal of Multivariate Analysis*, 90:106–131.

Diehl, F., Grahlmann, S., Beier, M., and Hoheisel, J. D. (2001). Manufacturing DNA microarrays of high spot homogeneity and reduced background signal. *Nucleic Acids Research*, 29:E38.

Dikalova, A., Clempus, R., Lassègue, B., Cheng, G., McCoy, J., Dikalov, S., San Martin, A., Lyle, A., Weber, D. S., Weiss, D., Taylor, W. R., Schmidt, H. H. H. W., Owens, G. K., Lambeth, J. D., and Griendling, K. K. (2005). Nox1 overexpression potentiates angiotensin II-induced hypertension and vascular smooth muscle hypertrophy in transgenic mice. *Circulation*, 112:2668–2676.

Ding, Y. and Simonoff, J. (2010). An investigation of missing data methods for classification trees applied to binary response data. *Journal of Machine Learning Research*, 11:131–170.

Du, P., Kibbe, W. A., and Lin, S. M. (2008). lumi: a pipeline for processing Illumina microarray. *Bioinformatics*, 24:1547–1548.

Dudoit, S. and Fridlyand, J. (2003). *Statistical analysis of gene expression microarray data / edited by Terry Speed*, chapter Classification in Microarray Experiments. Chapman & Hall/CRC.

Dudoit, S., van der Laan, M. J., and Pollard, K. S. (2004). Multiple testing. Part I. Single-step procedures for control of general type I error rates. *Statistical Applications in Genetics and Molecular Biology*, 3.

Dudoit, S., Yang, Y. H., Callow, M., and Speed, T. P. (2000). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical Report 578, University of California at Berkeley.

Dufva, M. (2005). Fabrication of high quality microarrays. *Biomolecular Engineering*, 22:173–184.

Dufva, M. (2009). Fabrication of DNA microarray. *Methods in Molecular Biology*, 529:63–79.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annuals of Statistics*, 7:1–26.

Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals and other measure of statistical accuracy. *Statistical Science*, 1:54–77.

Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall, New York.

Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E. (2005). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21:171–178.

Ein-Dor, L., Zuk, O., and Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences*, 103:5923–5928.

Fan, J. and Ren, Y. (2006). Statistical analysis of DNA microarray data in cancer research. *Clinical Cancer Research*, 12:4469–4473.

Fan, J.-B., Gunderson, K. L., Bibikova, M., Yeakley, J. M., Chen, J., Wickham Garcia, E., Lebruska, L. L., Laurent, M., Shen, R., and Barker, D. (2006). Illumina universal bead arrays. *Methods in Enzymology*, 410:57–73.

Fare, T. L., Coffey, E. M., Dai, H., He, Y. D., Kessler, D. A., Kilian, K. A., Koch, J. E., LeProust, E., Marton, M. J., Meyer, M. R., Stoughton, R. B., Tokiwa, G. Y., and Wang, Y. (2003). Effects of atmospheric ozone on microarray data quality. *Analytical Chemistry*, 75:4672–4675.

Farhangfar, A., Kurgan, L., and Dy, J. (2008). Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition*, 41:3692–3705.

Fecher, L. A., Cummings, S. D., Keefe, M. J., and Alani, R. M. (2007). Toward a molecular classification of melanoma. *Journal of Clinical Oncology*, 25:1606–1620.

Feelders, A. (1999). Handling missing data in trees: surrogate splits or statistical imputation? *Principles of Data Mining and Knowledge Discovery*, 1704:329–334.

Fierro, A. C., Vandenbussche, F., Engelen, K., Van de Peer, Y., and Marchal, K. (2008). Meta analysis of gene expression data within and across species. *Current Genomics*, 9:525–534.

Finn, O. J. (2008). Tumor immunology top 10 list. *Immunological Reviews*, 222:5–8.

Fisher, R. A. (1950). *Statistical methods for research workers*. Oliver & Boyd, Edinburgh, 11th (rev.) edition.

Gagnon-Bartsch, J. and Speed, T. (2011). Using control genes to correct for unwanted variation in microarray data. Technical report, Department of Statistics, University of California, Berkeley.

García-Castro, M., Reguero, J. R., Batalla, A., Díaz-Molina, B., González, P., Alvarez, V., Cortina, A., Cubero, G. I., and Coto, E. (2003). Hypertrophic cardiomyopathy: low frequency of mutations in the beta-myosin heavy chain (MYH7) and cardiac troponin T (TNNT2) genes among Spanish patients. *Clinical Chemistry*, 49:1279–1285.

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20:307–315.

Geller, A. C., Miller, D. R., Annas, G. D., Demierre, M.-F., Gilchrest, B. A., and Koh, H. K. (2002). Melanoma incidence and mortality among US whites, 1969-1999. *Journal of the American Medical Association*, 288:1719–1720.

Gelman, A., Hill, J., Yajima, M., Su, Y.-S., and Pittau, M. G. (2009). *Mi: missing data imputation and model checking*. R package version 0.08-04.02.

Gentleman, R., Ruschhaupt, M., Huber, W., and Lusa, L. (2008). *Meta-analysis for microarray experiments*. Bioconductor.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80.

Gevaert, O., De Smet, F., Timmerman, D., Moreau, Y., and De Moor, B. (2006). Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics*, 22:e184–e190.

Göran Jönsson, Christian Busch, Stian Knappskog, Jürgen Geisler, Hrvoje Miletic, Markus Ringnér, Johan R. Lillehaug, Ake Borg, and Lonning, P. E. (2010). Gene Expression Profiling–Based Identification of Molecular Subtypes in Stage IV Melanomas with Different Clinical Outcome. *Clinical Cancer Research*, 16(3356).

Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, 8:206–213.

Gray-Schopfer, V., Wellbrock, C., and Marais, R. (2007). Melanoma biology and new targeted therapy. *Nature*, 445:851–857.

Grayson, T. H., Ohms, S. J., Brackenbury, T. D., Meaney, K. R., Peng, K., Pittelkow, Y. E., Wilson, S. R., Sandow, S. L., and Hill, C. E. (2007). Vascular microarray profiling in two models of hypertension identifies caveolin-1, Rgs2 and Rgs5 as anti-hypertensive targets. *BMC Genomics*, 8:404.

Greco, A. V., Porcelli, G., Magalhaes, J. F., and Altomonte, L. (1978). Urinary kallikrein excretion and plasma DBH activity in hypertension. *Agents Actions*, 8:572–575.

Gregory, A. W., Roayaei, J. A., Quiñones, O. A., and Schneider, K. T. (2007). A microarray analysis for differential gene expression in the soybean genome using Bioconductor and R. *Briefings in Bioinformatics*, 8:415–431.

Grützmann, R., Boriss, H., Ammerpohl, O., Lüttges, J., Kalthoff, H., Schackert, H. K., Klöppel, G., Saeger, H. D., and Pilarsky, C. (2005). Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. *Oncogene*, 24:5079–5088.

Grzymala-Busse, J. W. and Hu, M. (2001). A Comparison of Several Approaches to Missing Attribute Values in Data Mining. In *RSCTC '00: Revised Papers from the Second International Conference on Rough Sets and Current Trends in Computing*, pages 378–385, London, UK. Springer-Verlag.

Guerra, R., Allison, D. B., and Goldstein, D. (2008). *Meta-analysis and combining information in genetics and genomics (interdisciplinary statistics)*, chapter Comparison of meta-analysis to combined analysis of a replicated microarray study. Chapman & Hall/CRC.

Guo, L., Lobenhofer, E. K., Wang, C., Shippy, R., Harris, S. C., Zhang, L., Mei, N., Chen, T., Herman, D., Goodsaid, F. M., Hurban, P., Phillips, K. L., Xu, J., Deng, X., Sun, Y. A., Tong, W., Dragan, Y. P., and Shi, L. (2006). Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nature Biotechnology*, 24:1162–1169.

Guo, X., Cheng, S., Taylor, K. D., Cui, J., Hughes, R., Quiñones, M. J., Bulnes-Enriquez, I., De la Rosa, R., Aurea, G., Yang, H., Hsueh, W., and Rotter, J. I. (2005). Hypertension genes are genetic markers for insulin sensitivity and resistance. *Hypertension*, 45:799–803.

Ha, K. C., Coulombe-Huntington, J., and Majewski, J. (2009). Comparison of Affymetrix Gene Array with the Exon Array shows potential application for detection of transcript isoform variation. *BMC Genomics*, 10:519.

Hanley, J. A. (2005). *Receiver Operating Characteristic (ROC) Curves.* John Wiley & Sons, Ltd.

Harel, O. and Zhou, X.-H. (2007). Multiple imputation: review of theory, implementation and software. *Statistics in Medicine*, 26:3057–3077.

Hartmann, O. (2005). Quality control for microarray experiments. *Methods of Information in Medicine*, 44:408–413.

Hartner, A., Porst, M., Klanke, B., Cordasic, N., Veelken, R., and Hilgers, K. F. (2006). Angiotensin II formation in the kidney and nephrosclerosis in Ren-2 hypertensive rats. *Nephrology Dialysis Transplantation*, 21:1778–1785.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *Elements of statistical learning: data mining, inference, and prediction.* Springer series in statistics. Springer, New York, NY, 2nd edition.

Heymans, M. W., van Buuren, S., Knol, D. L., van Mechelen, W., and de Vet, H. C. W. (2007). Variable selection under multiple imputation using the bootstrap in a prognostic study. *BMC Medical Research Methodology*, 7:33.

Höfling, H. and Tibshirani, R. (2007). A Study of Pre-Validation.

Honaker, J., King, G., and Blackwell, M. (2008). *Amelia: Amelia II: a program for missing data.* R package version 1.1-33.

Hong, F. and Breitling, R. (2008). A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, 24:374–382.

Horton, N. and Kleinman, K. (2007). Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, 61:79–90.

Horton, N. and Laird, N. (2000). Maximum likelihood analysis of logistic regression models with incomplete covariate data and auxiliary information. *Biometrics*, 57:34–42.

Horton, N. and Lipsitz, S. (2001). Multiple imputation in practice: comparison of software packages for regression models with missing variables. *The American Statistician*, 55:244–254.

Hoshida, Y., Villanueva, A., Kobayashi, M., Peix, J., Chiang, D. Y., Camargo, A., Gupta, S., Moore, J., Wrobel, M. J., Lerner, J., Reich, M., Chan, J. A., Glickman, J. N., Ikeda, K., Hashimoto, M., Watanabe, G., Daidone, M. G., Roayaie, S., Schwartz, M., Thung, S., Salvesen, H. B., Gabriel, S., Mazzaferro, V., Bruix, J., Friedman, S. L., Kumada, H., Llovet, J. M., and Golub, T. R. (2008). Gene expression in fixed tissues and outcome in hepatocellular carcinoma. *New England Journal of Medicine*, 359:1995–2004.

Hsiao, L. L., Dangond, F., Yoshida, T., Hong, R., Jensen, R. V., Misra, J., Dillon, W., Lee, K. F., Clark, K. E., Haverty, P., Weng, Z., Mutter, G. L., Frosch, M. P., MacDonald, M. E., Milford, E. L., Crum, C. P., Bueno, R., Pratt, R. E., Mahadevappa, M., Warrington, J. A., Stephanopoulos, G., and Gullans, S. R. (2001). A compendium of gene expression in normal human tissues. *Physiol Genomics*, 7:97–104.

Hubble, J., Demeter, J., Jin, H., Mao, M., Nitzberg, M., Reddy, T. B. K., Wymore, F., Zachariah, Z. K., Sherlock, G., and Ball, C. A. (2009). Implementation of GenePattern within the Stanford Microarray Database. *Nucleic Acids Research*, 37:D898–D901.

Hughes, T. R., Mao, M., Jones, A. R., Burchard, J., Marton, M. J., Shannon, K. W., Lefkowitz, S. M., Ziman, M., Schelter, J. M., Meyer, M. R., Kobayashi, S., Davis, C., Dai, H., He, Y. D., Stephaniants, S. B., Cavet, G., Walker, W. L., West, A., Coffey, E., Shoemaker, D. D., Stoughton, R., Blanchard, A. P., Friend, S. H., and Linsley, P. S. (2001). Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnology*, 19:342–347.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., and Herring, A. H. (2005). Missing-data methods for generalized linear models: a comparative review. *Journal of the American Statistical Association*, 100:332–346.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249–264.

Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., Frank, B. C., Gabrielson, E., Garcia, J. G. N., Geoghegan, J., Germino, G., Griffin, C., Hilmer, S. C., Hoffman, E., Jedlicka, A. E., Kawasaki, E., Martínez-Murillo, F., Morsberger, L., Lee, H., Petersen, D., Quackenbush, J., Scott, A., Wilson, M., Yang, Y., Ye, S. Q., and Yu, W. (2005). Multiple-laboratory comparison of microarray platforms. *Nature Methods*, 2:345–350.

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8:118–127.

Ke, L. D., Chen, Z., and Yung, W. K. (2000). A reliability test of standard-based quantitative PCR: exogenous vs endogenous standards. *Molecular and Cellular Probes*, 14:127–135.

Kerr, K. F. (2007). Extended analysis of benchmark datasets for Agilent two-color microarrays. *BMC Bioinformatics*, 8:371.

Kiec-Wilk, B., Dembinska-Kiec, A., Olszanecka, A., Bodzioch, M., Schmitz, G., and Kawecka-Jaszcz, K. (2007). A724A polymorphism of sarco(endo)plasmic reticulum Ca2+-ATPase 2 (SERCA2) in hypertensive patients. *Clinical Chemistry and Laboratory Medicine*, 45:467–470.

Kimura, T., Yokoyama, T., Matsumura, Y., Yoshiike, N., Date, C., Muramatsu, M., and Tanaka, H. (2003). NOS3 genotype-dependent correlation between blood pressure and physical activity. *Hypertension*, 41:355–360.

King, G., Honaker, J., Joseph, A., and Scheve, K. (2001). Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Political Science Association*, 95:49–69.

Kreil, D. P. and Russell, R. R. (2005). There is no silver bullet–a guide to low-level data transforms and normalisation methods for microarray data. *Briefings in Bioinformatics*, 6:86–97.

Kuo, W. P., Jenssen, T.-K., Butte, A. J., Ohno-Machado, L., and Kohane, I. S. (2002). Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics*, 18:405–412.

Lai, Y., Eckenrode, S. E., and She, J.-X. (2009). A statistical framework for integrating two microarray data sets in differential expression analysis. *BMC Bioinformatics*, 10 Suppl 1:S23.

Langley-Evans, S. C. (1997). Intrauterine programming of hypertension by glucocorticoids. *Life Sciences*, 60:1213–1221.

Langley-Evans, S. C. (2000). Critical differences between two low protein diet protocols in the programming of hypertension in the rat. *International Journal of Food Sciences and Nutrition*, 51:11–17.

Larsson, O., Wennmalm, K., and Sandberg, R. (2006). Comparative microarray analysis. *A Journal of Intergrative Biology*, 10:381–397.

Lausted, C., Dahl, T., Warren, C., King, K., Smith, K., Johnson, M., Saleem, R., Aitchison, J., Hood, L., and Lasky, S. R. (2004). POSaM: a fast, flexible, open-source, inkjet oligonucleotide synthesizer and microarrayer. *Genome Biology*, 5:R58.

Le, T. H., Fogo, A. B., Salzler, H. R., Vinogradova, T., Oliverio, M. I., Marchuk, D. A., and Coffman, T. M. (2004). Modifier locus on mouse chromosome 3 for renal vascular pathology in AT1A receptor-deficiency. *Hypertension*, 43:445–451.

Le, T. H., Kim, H.-S., Allen, A. M., Spurney, R. F., Smithies, O., and Coffman, T. M. (2003). Physiological impact of increased expression of the AT1 angiotensin receptor. *Hypertension*, 42:507–514.

Lê Cao, K.-A., Meugnier, E., and McLachlan, G. J. (2010). Integrative mixture of experts to combine clinical factors and gene markers. *Bioinformatics*, 26:1192–1198.

Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J., and Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. *Genome Research*, 14:1085–1094.

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3:1724–1735.

Lévesque, S., Moutquin, J.-M., Lindsay, C., Roy, M.-C., and Rousseau, F. (2004). Implication of an AGT haplotype in a multigene association study with pregnancy hypertension. *Hypertension*, 43:71–78.

Lin, S. M., Du, P., Huber, W., and Kibbe, W. A. (2008). Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Research*, 36:e11.

Lindsey, J. K. (1997). *Applying generalized linear models.* Springer texts in statistics. Springer, New York.

Listgarten, J., Kadie, C., Schadt, E. E., and Heckerman, D. (2010). Correction for hidden confounders in the genetic analysis of gene expression. *Proceedings of the National Academy of Sciences*, 107:16465–16470.

Little, R. J. A. (1992). Regression with missing X's: a review. *Journal of the American Statistical Association*, 87:1227–1237.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical analysis with missing data.* Wiley Series in Probability and Statistics. Wiley, New York, 1st edition.

Lorigan, P., Eisen, T., and Hauschild, A. (2008). Systemic therapy for metastatic malignant melanoma–from deeply disappointing to bright future? *Experimental Dermatology*, 17:383–394.

Lu, X., Murphy, T. C., Nanes, M. S., and Hart, C. M. (2010). PPARγ regulates hypoxia-induced Nox4 expression in human pulmonary artery smooth muscle cells through NF-κB. *American Journal of Physiology - Lung Cellular and Molecular Physiology*, 299(4):L559–L566.

MacKie, R. M., Bray, C. A., Hole, D. J., Morris, A., Nicolson, M., Evans, A., Doherty, V., and Vestey, J. (2002). Incidence of and survival from malignant melanoma in Scotland: an epidemiological study. *Lancet*, 360:587–591.

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2011). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 39:D52–D57.

Malhotra, S., Poole, J., Davis, H., Dong, Y., Pollock, J., Snieder, H., and Treiber, F. (2004). Effects of NOS3 Glu298Asp polymorphism on hemodynamic reactivity to stress: influences of ethnicity and obesity. *Hypertension*, 44:866–871.

Malin, B., Karp, D., and Scheuermann, R. H. (2010). Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. *Journal of Investigative Medicine*, 58:11–18.

Mandruzzato, S., Callegaro, A., Turcatel, G., Francescato, S., Montesco, M. C., Chiarion-Sileni, V., Mocellin, S., Rossi, C. R., Bicciato, S., Wang, E., Marincola, F. M., and Zanovello, P. (2006). A gene expression signature associated with survival in metastatic melanoma. *Journal of Translational Medicine*, 4:50.

Mann, G., Pupo, G., Campain, A., Carter, C., Schramm, S., Pianova, A., Gerega, S., Silva, C. D., Lai, K., Wilmott, J., Synott, M., Hersey, P., Kefford, R., Thompson, J., Yang, Y., and Scolyer, R. (2011). BRAF mutation, NRAS mutation and absence of an immune-related expressed gene profile predict poor outcome in Stage III melanoma. *Journal of Clinical Oncology*, Under review.

Marinho, C., Alho, I., Arduíno, D., Falcão, L. M., Brás-Nogueira, J., and Bicho, M. (2007). GST M1/T1 and MTHFR polymorphisms as risk factors for hypertension. *Biochemical and Biophysical Research Communications*, 353:344–350.

Markey, M. K., Tourassi, G. D., Margolis, M., and DeLong, D. M. (2006). Impact of missing data in evaluating artificial neural networks trained on complete data. *Computers in Biology and Medicine*, 36:516–525.

Marques, F. Z., Campain, A. E., Davern, P. J., Yang, Y. H. J., Head, G. A., and Morris, B. J. (2011a). Genes influencing circadian differences in blood pressure in hypertensive mice. *PLoS One*, 6:e19203.

Marques, F. Z., Campain, A. E., Davern, P. J., Yang, Y. H. J., Head, G. A., and Morris, B. J. (2011b). Global identification of the genes and pathways differentially expressed in hypothalamus in early and established neurogenic hypertension. *Physiol Genomics*, 43:766–771.

Marques, F. Z., Campain, A. E., Zukowska-Szczechowska, E., Tomaszewski, M., Yang, Y. H. J., Charchar, F. J., and Morris, B. J. (2011c). Gene expression profiling reveals renin mRNA overexpression in human hypertensive kidneys and a role for microRNAs. *Hypertension*, 58:1093–1098.

Massie, B. (1998). 15 years of heart-failure trials: what have we learned? *Lancet*, 352 Supplement 1:S1229–S133.

Matsuno, K., Yamada, H., Iwata, K., Jin, D., Katsuyama, M., Matsuki, M., Takai, S., Yamanishi, K., Miyazaki, M., Matsubara, H., and Yabe-Nishimura, C. (2005). Nox1 is involved in angiotensin II-mediated hypertension: a study in Nox1-deficient mice. *Circulation*, 112:2677–2685.

Matwin, M. K., Holte, R., and Stan (1998). Machine learning for the detection of oil spills in staellite radar images. *Machine Learning*, 30:195–215.

McBride, M. W., Brosnan, M. J., Mathers, J., McLellan, L. I., Miller, W. H., Graham, D., Hanlon, N., Hamilton, C. A., Polke, J. M., Lee, W. K., and Dominiczak, A. F. (2005). Reduction of Gstm1 expression in the stroke-prone spontaneously hypertension rat contributes to increased oxidative stress. *Hypertension*, 45:786–792.

McCullagh, P. P. and Nelder, J. A. (1998). *Generalized linear models.* Monographs on statistics and Applied Probability; 37. Chapman & Hall/CRC, Boca Raton, 2nd edition.

McNeil, K. A., Newman, I., and Kelly, F. J. (1996). *Testing research hypotheses with the general linear model.* Southern Illinois University Press, Carbondale.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72:417–473.

Michiels, S., Koscielny, S., and Hill, C. (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, 365:488–492.

Michiels, S., Koscielny, S., and Hill, C. (2007). Interpretation of microarray data in cancer. *British Journal of Cancer*, 96:1155–1158.

Miklos, G. L. G. and Maleszka, R. (2004). Microarray reality checks in the context of a complex disease. *Nature Biotechnology*, 22:615–621.

Min, L.-J., Mogi, M., Iwanami, J., Li, J.-M., Sakata, A., Fujita, T., Tsukuda, K., Iwai, M., and Horiuchi, M. (2008). Angiotensin II type 2 receptor deletion enhances vascular senescence by methyl methanesulfonate sensitive 2 inhibition. *Hypertension*, 51:1339–1344.

Mook, S., Van't Veer, L. J., Rutgers, E. J. T., Piccart-Gebhart, M. J., and Cardoso, F. (2007). Individualization of therapy using Mammaprint: from development to the MINDACT Trial. *Cancer Genomics Proteomics*, 4:147–155.

Moreau, Y., Aerts, S., De Moor, B., De Strooper, B., and Dabrowski, M. (2003). Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends in Genetics*, 19:570–577.

Müller, S. and Welsh, A. H. (2005). Outlier Robust Model Selection in Linear Regression. *Journal of the American Statistical Association*, 100:1297–1310.

Müller, S. and Welsh, A. H. (2009). Robust model selection in generalized linear models. *Statistica Sinica*, 19:1155–1170.

Müller, S. and Welsh, A. H. (2010). On Model Selection Curves. *International Statistical Review*, 78:240–256.

Mullins, J. J., Peters, J., and Ganten, D. (1990). Fulminant hypertension in transgenic rats harbouring the mouse Ren-2 gene. *Nature*, 344:541–544.

Murtaugh, P. A. (1998). Methods of variable selection in regression modeling. *Communications in Statistics - Simulation and Computation*, 27:711–734.

Newton-Cheh, C., Larson, M. G., Vasan, R. S., Levy, D., Bloch, K. D., Surti, A., Guiducci, C., Kathiresan, S., Benjamin, E. J., Struck, J., Morgenthaler, N. G., Bergmann, A., Blankenberg, S., Kee, F., Nilsson, P., Yin, X., Peltonen, L., Vartiainen, E., Salomaa, V., Hirschhorn, J. N., Melander, O., and Wang, T. J. (2009). Association of common variants in NPPA and NPPB with circulating natriuretic peptides and blood pressure. *Nature Genetics*, 41:348–353.

Nielsen, T. F. and Hökegaard, K. H. (1984). The course of subsequent pregnancies after previous cesarean section. *Acta Obstetricia et Gynecologica Scandinavica*, 63:13–16.

Nishisato, S. and Ahn, H. (1995). When not to analyze data: decision making on missing responses in dual scaling. *Annals of Operations Research*, 55:361–378.

Normand, S. L. (1999). Meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in Medicine*, 18:321–359.

Nuwaysir, E. F., Huang, W., Albert, T. J., Singh, J., Nuwaysir, K., Pitas, A., Richmond, T., Gorski, T., Berg, J. P., Ballin, J., McCormick, M., Norton, J., Pollock, T., Sumwalt, T., Butcher, L., Porter, D., Molla, M., Hall, C., Blattner, F., Sussman, M. R., Wallace, R. L., Cerrina, F., and Green, R. D. (2002). Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome Research*, 12:1749–1755.

Ohta, K., Kobashi, G., Hata, A., Yamada, H., Minakami, H., Fujimoto, S., Kondo, K., and Tamashiro, H. (2003). Association between a variant of the glutathione S-transferase P1 gene (GSTP1) and hypertension in pregnancy in Japanese: interaction with parity, age, and genetic factors. *Seminars in Thromosis and Hemostasis*, 29:653–659.

Okamoto, T., Suzuki, T., and Yamamoto, N. (2000). Microarray fabrication with covalent attachment of DNA using bubble jet technology. *Nature Biotechnology*, 18:438–441.

Owzar, K., Barry, W. T., Jung, S.-H., Sohn, I., and George, S. L. (2008). Statistical challenges in preprocessing in microarray experiments in cancer. *Clinical Cancer Research*, 14:5959–5966.

Paliege, A., Pasumarthy, A., Parsumathy, A., Mizel, D., Yang, T., Schnermann, J., and Bachmann, S. (2006). Effect of apocynin treatment on renal expression of COX-2, NOS1, and renin in Wistar-Kyoto and spontaneously hypertensive rats. *American Journal of Physiology - Regulatory, Integrative and Comparative Physiology*, 290:R694–R700.

Paravicini, T. M., Chrissobolis, S., Drummond, G. R., and Sobey, C. G. (2004). Increased NADPH-oxidase activity and Nox4 expression during chronic hypertension is associated with enhanced cerebral vasodilatation to NADPH in vivo. *Stroke*, 35:584–589.

Park, Y. R., Lee, H. W., and Kim, J. H. (2005). Integrating microarray gene expression object model and clinical document architecture for cancer genomics research. *AMIA Annual Symposium Proceedings*, page 1073.

Parkinson, H., Sarkans, U., Shojatalab, M., Abeygunawardena, N., Contrino, S., Coulson, R., Farne, A., Lara, G. G., Holloway, E., Kapushesky, M., Lilja, P., Mukherjee, G., Oezcimen, A., Rayner, T., Rocca-Serra, P., Sharma, A., Sansone, S., and Brazma, A. (2005). ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, 33:D553–D555.

Parmigiani, G., Garrett, E. S., Anbazhagan, R., and Gabrielson, E. (2002). A statistical framework for expression-based molecular classification in cancer. *Journal of the Royal Statistical Society Series B*, 64:717–736.

Parmigiani, G., Garrett-Mayer, E. S., Anbazhagan, R., and Gabrielson, E. (2004). A cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clinical Cancer Research*, 10:2922–2927.

Patterson, T. A., Lobenhofer, E. K., Fulmer-Smentek, S. B., Collins, P. J., Chu, T.-M., Bao, W., Fang, H., Kawasaki, E. S., Hager, J., Tikhonova, I. R., Walker, S. J., Zhang, L., Hurban, P., de Longueville, F., Fuscoe, J. C., Tong, W., Shi, L., and Wolfinger, R. D. (2006). Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. *Nature Biotechnology*, 24:1140–1150.

Pravenec, M., Landa, V., Zídek, V., Musilová, A., Kazdová, L., Qi, N., Wang, J., St Lezin, E., and Kurtz, T. W. (2003). Transgenic expression of CD36 in the spontaneously hypertensive rat is associated with amelioration of metabolic disturbances but has no effect on hypertension. *Physiological Research*, 52:681–688.

Pravenec, M., Landa, V., Zidek, V., Musilova, A., Kren, V., Kazdova, L., Aitman, T. J., Glazier, A. M., Ibrahimi, A., Abumrad, N. A., Qi, N., Wang, J. M., St Lezin, E. M., and Kurtz, T. W. (2001). Transgenic rescue of defective CD36 ameliorates insulin resistance in spontaneously hypertensive rats. *Nature Genetics*, 27:156–158.

Provost, F., Fawcett, T., and Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. In *In proceedings of the fifteenth international conference on machine learning*, pages 445–453. Morgan Kaufmann.

R Development Core Team (2005). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramasamy, A., Mondry, A., Holmes, C. C., and Altman, D. G. (2008). Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Medicine*, 5:e184.

Rangel, J., Nosrati, M., Torabian, S., Shaikh, L., Leong, S. P. L., Haqq, C., Miller, J. R., Sagebiel, R. W., and Kashani-Sabet, M. (2008). Osteopontin as a molecular prognostic marker for melanoma. *Cancer*, 112:144–150.

Ravo, M., Mutarelli, M., Ferraro, L., Grober, O. M. V., Paris, O., Tarallo, R., Vigilante, A., Cimino, D., De Bortoli, M., Nola, E., Cicatiello, L., and Weisz, A. (2008). Quantitative expression profiling of highly degraded RNA from formalin-fixed, paraffin-embedded breast tumor biopsies by oligonucleotide microarrays. *Laboratory Investigation*, 88:430–440.

Rhodes, D. R., Barrette, T. R., Rubin, M. A., Ghosh, D., and Chinnaiyan, A. M. (2002). Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Research*, 62:4427–4433.

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., and Chinnaiyan, A. M. (2004). Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proceedings of the National Academy of Sciences*, 101:9309–9314.

Riemke, J., Campain, A., Bignardi, T., Casikar, I., Alhamdan, D., Fauchon, D., Benzie, R., Müller, S., Yang, J., Mongelli, M., and Condous, G. (2011). Development of a new model to predict viability at the end of the 1st trimester after a single visit to an Early Pregnancy Unit. Preprint.

Ritchie, M., Diyagama, D., Neilson, J., van Laar, R., Dobrovic, A., Holloway, A., and Smyth, G. (2006). Empirical array quality weights in the analysis of microarray data. *BMC Bioinformatics*, 7.

Roberts, P. C. (2008). Gene expression microarray data analysis demystified. *Biotechnology Annual Review*, 14:29–61.

Rodriguez-Iturbe, B., Sepassi, L., Quiroz, Y., Ni, Z., Wallace, D. C., and Vaziri, N. D. (2007). Association of mitochondrial SOD deficiency with salt-sensitive hypertension and accelerated renal senescence. *Journal of Applied Physiology*, 102:255–260.

Rogus, J. J., Moczulski, D., Freire, M. B., Yang, Y., Warram, J. H., and Krolewski, A. S. (1998). Diabetic nephropathy is associated with AGT polymorphism T235: results of a family-based study. *Hypertension*, 31:627–631.

Rubin, D. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91:473–489.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Chapman & Hall/CRC, New York.

Russ, J. and Futschik, M. E. (2010). Comparison and consolidation of microarray data sets of human tissue expression. *BMC Genomics*, 11:305.

Rysä, J., Leskinen, H., Ilves, M., and Ruskoaho, H. (2005). Distinct upregulation of extracellular matrix genes in transition from hypertrophy to hypertensive heart failure. *Hypertension*, 45:927–933.

Sanchez, O., Marcos, E., Perros, F., Fadel, E., Tu, L., Humbert, M., Dartevelle, P., Simonneau, G., Adnot, S., and Eddahibi, S. (2007). Role of endothelium-derived CC chemokine ligand 2 in idiopathic pulmonary arterial hypertension. *American Journal of Respiratory and Critical Care Medicine*, 176:1041–1047.

Sarah-Jane Schramm, Anna Campain, Richard Scolyer, Yee Hwa Yang, and Graham Mann (2011). Review and cross-validation of gene expression signatures and melanoma prognosis. *Journal of Investigative Dermatology*, 132:274–283.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* Chapman & Hall/CRC, London.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8:3–15.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470.

Schmid, C. H., Terrin, N., Griffith, J. L., D'agostino, R. B. D., and Harry, P. S. (2001). Predictive performance of missing data methods for logistic regression, classification trees and neural networks. *Journal of Statistical Computation and Simulation*, pages 115–140.

Schomaker, M., C., H., and H., T. (2007). New approaches for model selection under missing data. Technical report, Department of Statistics, Ludwig Maximilians Universität, München.

Schwarz, E., Leweke, F. M., Bahn, S., and Liò, P. (2009). Clinical bioinformatics for complex disorders: a schizophrenia case study. *BMC Bioinformatics*, 10 Suppl 12:S6.

Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., Collins, P. J., de Longueville, F., Kawasaki, E. S., Lee, K. Y., Luo, Y., Sun, Y. A., Willey, J. C., Setterquist, R. A., Fischer, G. M., Tong, W., Dragan, Y. P., Dix, D. J., Frueh, F. W., Goodsaid, F. M., Herman, D., Jensen, R. V., Johnson, C. D., Lobenhofer, E. K., Puri, R. K., Schrf, U., Thierry-Mieg, J., Wang, C., Wilson, M., Wolber, P. K., Zhang, L., Amur, S., Bao, W., Barbacioru, C. C., Lucas, A. B., Bertholet, V., Boysen, C., Bromley, B., Brown, D., Brunner, A., Canales, R., Cao, X. M., Cebula, T. A., Chen, J. J., Cheng, J., Chu, T.-M., Chudin, E., Corson, J., Corton, J. C., Croner, L. J., Davies, C., Davison, T. S., Delenstarr, G., Deng, X., Dorris, D., Eklund, A. C., Fan, X.-h., Fang, H., Fulmer-Smentek, S., Fuscoe, J. C., Gallagher, K., Ge, W., Guo, L., Guo, X., Hager, J., Haje, P. K., Han, J., Han, T., Harbottle, H. C., Harris, S. C., Hatchwell, E., Hauser, C. A., Hester, S., Hong, H., Hurban, P., Jackson, S. A., Ji, H., Knight, C. R., Kuo, W. P., LeClerc, J. E., Levy, S., Li, Q.-Z., Liu, C., Liu, Y., Lombardi, M. J., Ma, Y., Magnuson, S. R., Maqsodi, B., McDaniel, T., Mei, N., Myklebost, O., Ning, B., Novoradovskaya, N., Orr, M. S., Osborn, T. W., Papallo, A., Patterson, T. A., Perkins, R. G., Peters, E. H., Peterson, R., Philips, K. L., Pine,

P. S., Pusztai, L., Qian, F., Ren, H., Rosen, M., Rosenzweig, B. A., Samaha, R. R., Schena, M., Schroth, G. P., Shchegrova, S., Smith, D. D., Staedtler, F., Su, Z., Sun, H., Szallasi, Z., Tezak, Z., Thierry-Mieg, D., Thompson, K. L., Tikhonova, I., Turpaz, Y., Vallanat, B., Van, C., Walker, S. J., Wang, S. J., Wang, Y., Wolfinger, R., Wong, A., Wu, J., Xiao, C., Xie, Q., Xu, J., Yang, W., Zhang, L., Zhong, S., Zong, Y., and Slikker, W. (2006). The microarray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24:1151–1161.

Shippy, R., Fulmer-Smentek, S., Jensen, R. V., Jones, W. D., Wolber, P. K., Johnson, C. D., Pine, P. S., Boysen, C., Guo, X., Chudin, E., Sun, Y. A., Willey, J. C., Thierry-Mieg, J., Thierry-Mieg, D., Setterquist, R. A., Wilson, M., Lucas, A. B., Novoradovskaya, N., Papallo, A., Turpaz, Y., Baker, S. C., Warrington, J. A., Shi, L., and Herman, D. (2006). Using RNA sample titrations to assess microarray platform performance and normalization techniques. *Nature Biotechnology*, 24:1123–1131.

Siddiqi, T. A., Caligaris, J. T., Miodovnik, M., Holroyde, J. C., and Mimouni, F. (1988). Rate of spontaneous abortion after first trimester sonographic demonstration of fetal cardiac activity. *American Journal of Perinatology*, 5:1–4.

Sims, A. H. (2009). Bioinformatics and breast cancer: what can high-throughput genomic approaches actually tell us? *Journal of Clinical Pathology*, 62:879–885.

Singh-Gasson, S., Green, R. D., Yue, Y., Nelson, C., Blattner, F., Sussman, M. R., and Cerrina, F. (1999). Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nature Biotechnology*, 17:974–978.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3.

Smyth, G. K. and Speed, T. (2003). Normalization of cDNA microarray data. *Methods*, 31:265–273.

Smyth, G. K., Yang, Y. H., and Speed, T. (2003). Statistical issues in cDNA microarray data analysis. *Methods in Molecular Biology*, 224:111–136.

Smyth, G. K. T. N. P. and Wettenhall, J. (2003). *Limma: linear models for microarray data user's guide.*

Speed, T. P. (2003). *Statistical analysis of gene expression microarray data.* Interdisciplinary statistics. Chapman & Hall/CRC, Boca Raton, FL.

Sreenivasulu, N., Sunkar, R., Wobus, U., and Strickert, M. (2010). Array platforms and bioinformatics tools for the analysis of plant transcriptome in response to abiotic stress. *Methods in Molecular Biology*, 639:71–93.

Steyerberg, E. W., Eijkemans, M. J., Harrell, F. E., and Habbema, J. D. (2000). Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Statistics in Medicine*, 19:1059–1079.

Stuart, E. A., Azur, M., Frangakis, C., and Leaf, P. (2009). Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative. *American Journal of Epidemiology*, 169:1133–1139.

Su, Y.-S., Gelman, A., Hill, J., and Yajima, M. (2011). Multiple imputation with diagnostics (mi) in R: opening windows into the black box. *Journal of Statistical Software*.

Suzuki, T., Higgins, P. J., and Crawford, D. R. (2000). Control selection for RNA quantitation. *Biotechniques*, 29:332–337.

Tambascia, R. C., Fonseca, P. M., Corat, P. D., Moreno, H., Saad, M. J., and Franchini, K. G. (2001). Expression and distribution of NOS1 and NOS3 in the myocardium of angiotensin II-infused rats. *Hypertension*, 37:1423–1428.

Tan, P. K., Downey, T. J., Spitznagel, E. L., Xu, P., Fu, D., Dimitrov, D. S., Lempicki, R. A., Raaka, B. M., and Cam, M. C. (2003). Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Research*, 31:5676–5684.

Tanner, M. A. (1991). *Tools for statistical inference*. Springer-Verlag, New York.

Thellin, O., Zorzi, W., Lakaye, B., De Borman, B., Coumans, B., Hennen, G., Grisar, T., Igout, A., and Heinen, E. (1999). Housekeeping genes as internal standards: use and limits. *Journal of Biotechnology*, 75:291–295.

Thomas John, Michael A. Black, Tumi T. Toro, Debbie Leader, Craig A. Gedye, Ian D. Davis, Parry J. Guilford, and Cebon, J. S. (2008). Predicting Clinical Outcome through Molecular Profiling in Stage III Melanoma. *Clinical Cancer Research*, 14.

Thompson, J. F., Scolyer, R. A., and Kefford, R. F. (2005). Cutaneous melanoma. *Lancet*, 365:687–701.

Tibshirani, R. J. and Efron, B. (2002). Pre-validation and inference in microarrays. *Statistical Applications in Genetics and Molecular Biology*, 1.

Tímár, J., Gyorffy, B., and Rásó, E. (2010). Gene signature of the metastatic potential of cutaneous melanoma: too much for too little? *Clinical and Experimental Metastasis*, 27:371–387.

Tiwari, M. (2010). From tumor immunology to cancer immunotherapy: miles to go. *Journal of Cancer Research and Therapeutics*, 6:427–431.

Touyz, R. M., Endemann, D., He, G., Li, J. S., and Schiffrin, E. L. (1999). Role of AT2 receptors in angiotensin II-stimulated contraction of small mesenteric arteries in young SHR. *Hypertension*, 33:366–372.

Tower, C. L., Strachan, B. K., and Baker, P. N. (2000). Long-term implications of caesarean section. *Journal of Obstetrics and Gynaecology*, 20:365–367.

Truntzer, C., Maucort-Boulch, D., and Roy, P. (2008). Comparative optimism in models involving both classical clinical and gene expression information. *BMC Bioinformatics*, 9:434.

Tsao, H., Atkins, M. B., and Sober, A. J. (2004). Management of cutaneous melanoma. *New England Journal of Medicine*, 351:998–1012.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98:5116–5121.

van Buuren, S., Boshuizen, H. C., and Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18:681–694.

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S. Fourth Edition*. Springer, New York, NY.

Warrington, J. A., Nair, A., Mahadevappa, M., and Tsyganskaya, M. (2000). Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiological Genomics*, 2:143–147.

Watanabe, T., Arita, S., Shiraishi, Y., Suguro, T., Sakai, T., Hongo, S., and Miyazaki, A. (2009). Human urotensin II promotes hypertension and atherosclerotic cardiovascular diseases. *Current Medicinal Chemistry*, 16:550–563.

Watanabe, T., Kanome, T., Miyazaki, A., and Katagiri, T. (2006). Human urotensin II as a link between hypertension and coronary artery disease. *Hypertension Research*, 29:375–387.

Watson, J. D. (2008). *Molecular biology of the gene.* Pearson/Benjamin Cummings Cold Spring Harbor Laboratory Press, San Francisco, 6th edition.

Weichert, W., Paliege, A., Provoost, A. P., and Bachmann, S. (2001). Upregulation of juxtaglomerular NOS1 and COX-2 precedes glomerulosclerosis in fawn-hooded hypertensive rats. *American Journal of Physiology - Renal Physiology*, 280:F706–F714.

Weiss, G. M. and Provost, F. (2001). The Effect of Class Distribution on Classifier Learning. Technical report, Department of Computer Science, Rutgers University.

Wenquan Niu, Yue Qi, Yueshe Qian, Pingjin Gao, and Zhu, D. (2009). The relationship between apolipoprotein E 2/3/4 polymorphosms and hypertension: a meta-analysis of six studies comprising 1812 cases and 1762 controls. *Hypertension Research*, 32:1060–1066.

Whistler, T., Unger, E. R., Nisenbaum, R., and Vernon, S. D. (2003). Integration of gene expression, clinical, and epidemiologic data to characterize Chronic Fatigue Syndrome. *Journal of Translational Medicine*, 1:10.

Winnepenninckx, V., Lazar, V., Michiels, S., Dessen, P., Stas, M., Alonso, S. R., Avril, M.-F., Ortiz Romero, P. L., Robert, T., Balacescu, O., Eggermont, A. M. M., Lenoir, G., Sarasin, A., Tursz, T., van den Oord, J. J., and Spatz, A. (2006). Gene expression profiling of primary cutaneous melanoma and clinical outcome. *Journal of the National Cancer Institute*, 98:472–482.

Wolpert, D. H. (1992). Stacked Generalization. *Neural Networks*, 5:241–259.

Wood, A. M., White, I. R., and Royston, P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in Medicine*, 27:3227–3246.

Yang, X., Belin, T. R., and Boscardin, W. J. (2005a). Imputation and variable selection in linear regression models with missing covariates. *Biometrics*, 61:498–506.

Yang, Y. H., Campain, A., and Speed, T. P. (2011). Finding differentially expressed genes in microarray data. *Nature Reviews Genetics*, Under review.

Yang, Y. H., Xiao, Y., and Segal, M. R. (2005b). Identifying differentially expressed genes from microarray experiments via statistic synthesis. *Bioinformatics*, 21:1084–1093.

Yauk, C. L., Berndt, M. L., Williams, A., and Douglas, G. R. (2004). Comprehensive comparison of six microarray technologies. *Nucleic Acids Research*, 32:e124.

Yeakley, J. M., Fan, J.-B., Doucet, D., Luo, L., Wickham, E., Ye, Z., Chee, M. S., and Fu, X.-D. (2002). Profiling alternative splicing on fiber-optic arrays. *Nature Biotechnology*, 20:353–358.

Yuen, T., Wurmbach, E., Pfeffer, R. L., Ebersole, B. J., and Sealfon, S. C. (2002). Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Research*, 30:e48.

Zhang, M., Yao, C., Guo, Z., Zou, J., Zhang, L., Xiao, H., Wang, D., Yang, D., Gong, X., Zhu, J., Li, Y., and Li, X. (2008). Apparently low reproducibility of true differential expression discoveries in microarray studies. *Bioinformatics*, 24:2057–2063.

Zhang, Y., Szustakowski, J., and Schinke, M. (2009). Bioinformatics analysis of microarray data. *Methods in Molecular Biology*, 573:259–284.

- 8 NOV 2012