PROGNOSTIC METHODS FOR INTEGRATING DATA FROM COMPLEX DISEASES

KAUSHALA SAMUDINI JAYAWARDANA
NAIWALA PATHIRANNEHELAGE

A thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy

School of Mathematics and Statistics
Faculty of Science
The University of Sydney
22 January 2016

"The mind is everything. What you think, you become"

- Lord Buddha

Kaushala Samudini Jayawardana Naiwala Pathirannehelage: *Prognostic methods for integrating data from complex diseases,* Doctor of Philosophy, © 22 January 2016

This thesis is dedicated,

To my dearest ammi and thaththi,
For bringing me up to what I am today, for all the encouragement throughout the years, for your unflagging love and faith...

To my beloved husband Madawa,
For your unconditional love, unwavering support, inexhaustible patience and strength that made me go on...

This journey would not have been possible without you...

This thesis is also dedicated to our little sweetheart,
Oh how I look forward to the day we finally meet!!

DECLARATION

I declare that the PhD thesis titled, 'Prognostic methods for integrating data from complex diseases' and the work presented in it are my own. I confirm that the work described in this thesis was performed between August 2011 and March 2015 at the School of Mathematics and Statistics, Faculty of Science, The University of Sydney. Except where otherwise indicated, this thesis is entirely my own work. Any help and assistance that I have received in my research work and the preparation of the thesis itself have been appropriately acknowledged. Furthermore, this thesis contains no material that has been submitted previously, in whole or in part, for the award of any other academic degree, fellowship or other recognition.

Kaushala Samudini
Jayawardana Naiwala
Pathirannehelage

ABSTRACT

The elucidation of the complexity underlying diseases such as cancer, has gained a vast surge with the rapid advances in high-throughput data technologies that resulted in the generation of a multitude of data. With the increased data availability, bioinformaticians are presented with the strenuous task of developing improved methodologies to deliver biological insights. In the context of predicting the disease outcome, multilayered data hold the potential to provide complementary information and improve the disease prognosis. Data integration is increasingly becoming essential in this context, to address a wide range of problems such as increasing the power of studies, addressing inconsistencies between studies, obtaining more reliable biomarkers and gaining a broader understanding of the disease. In this setting, the development of statistical methods faces many challenges. This thesis focuses on addressing these challenges while contributing to the methodological advancements in the field.

Prior to the integration of multiple data sources, it is imperative to address the statistical problems associated with the analysis of individual data types. To this end, we propose a clinical data analysis framework to obtain a model with good prediction accuracy, addressing the common issues such as missing data and model instability. This proposed framework demonstrated the highest prediction accuracy when evaluated on real biological data, in a comparison study. A detailed pre-processing pipeline is proposed for miRNA data that removes unwanted noise, variations and offers improved concordance with qRT-PCR data. Furthermore, platform specific models are developed to uncover biomarkers that are predictive of the survival outcome using mRNA, protein and miRNA data, to identify the data source with the most important prognostic information.

This thesis deals with two types of data integration: horizontal data integration, which is the integration of multiple datasets of the same type, and vertical data integration, which is the integration of datasets from different platforms for the same set of patients. In exploring the horizontal data integration, we use miRNA datasets from multiple sources to develop a comprehensive meta-analysis framework. The proposed novel framework aids in identifying the inconsistencies among studies, while identifying a reproducible and robust set of biomarkers. The comprehensive analysis framework addresses the many challenges in horizontal data integration, such as the differences in study aims and designs, and the heterogeneity of patient cohorts in each study, favourably through a multi-step validation protocol.

Exploiting the availability of multi-layered data on the same set of patients, clinical, mRNA, miRNA and protein, we develop novel frameworks in the vertical data integration paradigm. This type of data integration also faces many challenges such as having more variables than observations in the omics data and the imbalance of variables in the integration setting, that are addressed in our frameworks. The first is the integration of clinical and high-throughput data extending the pre-validation principle. The proposed integration framework allows variable selection from among platforms and identifies dominant sources of prognostic information. Next, we derived platform dependent weights to develop a data integration framework with the weighted Lasso. The comparison of integration at various levels using the proposed frameworks revealed that integration of multi-layered data is instrumental in improving the prediction accuracy and to obtain more biologically relevant biomarkers. Using the proposed data integration frameworks we devise a visualisation technique to look at prediction accuracy at the patient level. This graphical device revealed important findings with translational impact to aid in personalised medicine.

PUBLICATIONS AND PRESENTATIONS

Some of the methods, concepts, analyses and results in this thesis have appeared previously in the following:

PUBLICATIONS

Jayawardana, K., Schramm, S., Tembe, V., Müller, S., Thompson, J., Scolyer, R., Mann, G., and Yang, J. (2016). Identification, review and systematic cross-validation of microRNA prognostic signatures in metastatic melanoma. Journal of Investigative Dermatology, 136(1), 245-254.

Jayawardana, K., Schramm, S. J., Haydu, L., Thompson, J. F., Scolyer, R. A., Mann, G. J., Müller, S., and Yang, J. Y. (2015). Determination of prognosis in metastatic melanoma through integration of clinico-pathologic, mutation, mRNA, microRNA, and protein information. International Journal of Cancer, 136(4), 863-874.

Jayawardana, K., Müller, S., Schramm, S. J., Mann, G. J., and Yang, J. Y. (2013). Vertical data integration for melanoma prognosis. Proceedings of the 59th World Statistics Congress of the International Statistical Institute, 3599-3604.

PRESENTATIONS

Jayawardana K, Müller S, Yang JYH (2014) Biomarker Discovery In Metastatic Melanoma Via Multi-Layered Data. *Australian Statistical Conference (IMS - ASC 2014).*

Jayawardana, K, Yang, J, Müller, S (2013) Prognostic biomarkers through integrative analysis of multiple data sources. *57th Annual Meeting of the Australian Mathematical Society, Sydney.*

Jayawardana, K. (2013). Biomarker discovery via multi-layered data. *Winton charitable foundation workshop day 2013.*

Jayawardana, K., Yang, J., Müller, S. (2013). Integrating Clinical and Omics Data for Melanoma Prognosis. *Young Statisticians Conference 2013, Melbourne.*

Jayawardana, K., Yang, J. (2012). Comparison Study on Integrating Data for Melanoma Prognosis. *ComBio 2012, Adelaide .*

ACKNOWLEDGEMENTS

As I approach the completion of my PhD, I realize that I am deeply indebted to the support and assistance of many people. I wish to thank everyone of those, with whom I shared experiences in life, who have helped me survive the birthing of this thesis.

Foremost, I would like to express my heartfelt gratitude to my supervisor, Professor Jean Yang, for her tremendous time and effort spent in leading, advising and encouraging me during the passage of time passed with this research. Her passion for challenges and her enthusiasm has given me inspiration, her vast knowledge has given me guidance, her perseverance in research has given me motivation. I feel extremely lucky to work under the supervision of such a gifted statistician.

I owe a special gratitude to Associate Professor Samuel Müller, not only for being a wonderful associate supervisor, but also for introducing me to Jean when I was looking for PhD opportunities. I am deeply indebted to him for the opportunity to do my PhD in a leading university in Australia by encouraging me to apply. He has been a constant source of knowledge and support, guiding me with tremendous patience throughout the years.

I gratefully acknowledge the support of everyone in the School of Mathematics and Statistics, specifically, the former Head of School Professor Neville Weber, all the academic staff including Associate Professor Shelton Peiris and Dr. John Ormerod, all the non-academic staff, especially Sonia Morr for her kindness and support in all administration issues and Paul Szabo for helping out with the technical difficulties in my PhD career. I would like to acknowledge the Australian government and the University of Sydney for providing me with necessary financial support mainly through the international postgraduate research scholarship (IPRS), Australian Postgraduate Award (APA) and Postgraduate Research Support Scheme (PRSS). I would also like to acknowledge the top-up scholarship in mathematical biology given through Winton charitable foundation and Associate Professor Mary Myerscough for coordinating it. I would also like to thank my fellow PhD candidates, colleagues and students at the University of Sydney. I especially thank everyone in the bioinformatics research group including Shila and Victor who were always very friendly, kind and willing to help, Vivek and Penghao for their support and time in helping me with miRNA and proteomics analysis during the very early days of my PhD and fellow PhD candidate Sanjaya who always offered help and provided information regarding the PhD and other issues. I feel very lucky to have the opportunity to work with a talented group of collaborators from the Westmead Millennium Institute for Medical Research and Melanoma Institute Australia, especially Professor Graham Mann and Dr. Sarah-Jane whose enthusiasm and thoroughness in research has encouraged and motivated me.

These past few years have been less stressful and more enjoyable because of the companionship of my friends. My friends in Sydney made the past three and half years more bearable through many gatherings, trips and food. I would like to thank them, the extended UWS crew and the Macquarie crew, for making it much easier to live so far from my beloved motherland. My closest friend Ramadha always made time for me even amongst her own PhD studies. I am forever grateful to her for the many chat ses-

sions carried out across the world from USA, in encouraging me, sharing many laughs and also tears and simply just listening to me that helped me get through my very stressful moments. I am tempted to thank all of my friends who from my childhood until this point, have joined me in the discovery of what life is about and how to make the best of it. Especially everyone from 'FriendsForever': Ramadha, Kaveesha, Thisari, Dilshani, Gayathri and Randima, who made the four years at UOC one of the most enjoyable and cherished periods of my life. Since then they have always cheered me up bringing me closer to my motherland and helped me get through the tough times by being the best friends I will ever find in my life and I am grateful to their friendship more than they will ever know.

I sincerely thank all my teachers since childhood for their contribution in shaping me up to what I am today, my school teachers and the lecturers from the University of Colombo who gave me confidence in my abilities that made me climb up the ladder of education.

I feel truly lucky to have many supportive relatives and I would like to acknowledge all of them who offered many words of encouragement throughout the years. I would especially like to express my deepest gratitude and love to my maternal grandmother, aachchi amma and my late paternal grandmother, kiriamma who always believed in me. I am lucky that my life has been touched by their love, kindness and support. There are many people who made my life better simply by being part of it. I would like to extend my gratitude to them, especially sil maniyo and upasakamma for touching my life with the teachings of the Buddha since childhood, that made me a better person.

An ocean of thanks goes out to my awesome family. My parents who were the guiding stars in my life, have always made me believe that education is the key to success. They have made so many sacrifices in life, put hold to so many plans of theirs to give way to ours and always made their children the first priority in life. I am forever indebted to all that they have done for me. Without their endless support and unconditional love I would not have achieved this much. My sister and brother-in-law for all the love and support, especially in the early few months in Australia in getting us settled here. My brother just for being the annoying younger brother you are, which made the tough situations seem lighter. I know you miss your lovely akkies in Sri Lanka. I truly miss the fights, the laughs and the wonderful days when all five of us were together, but I am thankful for all those times that made me the person I am today.

I thank my wonderful extended family, especially my mother-in-law and father-in-law for their love, continuous support and encouragement, but more specifically for their son, my beloved husband Madawa.

My husband has been a constant source of love, caring and support. He is my best friend, my soul-mate and the best thing that ever happened to me. He is tremendously patient throughout my many mood swings, he always listens to me and offers me a shoulder to cry on, he instils my confidence and faith whenever I loose it and loves me unconditionally during my good and bad times. His complete and unflagging love carries me through always. I get through my worst moments simply by thinking about him. He has always understood me the best and has taken care of whatever needed tending without any complaints so that I could complete my thesis in time. Words cannot simply express my love and gratitude for him. This PhD journey has not been an easy ride. Through his love, patience, support and unwavering faith in me I was able to make it out unscathed.

Last but not the least, my heartfelt gratitude and love goes out to our little sweetheart whom we still eagerly await. Last few months have not been easy for him because of the sleepless nights, endless hours of work and enormous amounts of stress. I thank you for bearing all the hard times and long hours, for keeping me company through your kicking and hiccups and for simply giving me hope and the will to carry on with the best thing in my life to look forward to: *your arrival*.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AIC	Akaike Information Criterion
AJCC	American Joint Committee on Cancer
BIC	Bayesian Information Criterion
B-MI	Bootstrap samples and Multiple Imputation
C-index	Concordance index
CV	Cross-validation
DE	Differentially expressed
DLDA	Diagonal Linear Discriminant Analysis
DNA	Deoxyribonucleic acid
GEO	Gene Expression Omnibus
GLM	Generalised Linear Model
GP	Good prognosis
KM-curve	Kaplan-Meier curve
KNN	k-Nearest Neighbours
Lasso	Least absolute shrinkage and selection operator
LOOCV	Leave One Out Cross Validation
LR	Logistic regression
mBMI	modified Bootstrap Multiple Imputation
MIA	Melanoma Institute Australia
miRNA	Micro-RNA
miRNA-Seq	Sequencing of miRNA molecules
mRNA	Messenger-RNA
mRNA-Seq	Sequencing of mRNA molecules
NGS	Next generation sequencing
NSC	Nearest Shrunken Centroids
PP	Poor prognosis
qRT-PCR	Quantitative real time polymerase chain reaction
RNA	Ribonucleic acid
RNA-Seq	Sequencing of RNA molecules
RUV	Removing Unwanted Variation
SKCM	Skin Cutaneous Melanoma
SVM	Support Vector Machine
TCGA	The cancer genome atlas
TMM	Trimmed means of M-values

1

INTRODUCTION

Statistics in medical research gained a vast surge with the development of high-throughput biotechnologies, which provided thousands of measurements for each patient. Over the years, high-throughput data generating technologies have evolved immensely from the initially developed microarrays to the next-generation sequencing (NGS) technologies (Su *et al.*, 2011; Git *et al.*, 2010; Willenbrock *et al.*, 2009). The cost associated with the sequencing of DNA has been markedly reduced, allowing more and more data to be generated (Grada and Weinbrecht, 2013; Metzker, 2010; Hurd and Nelson, 2009). This increased amount of data naturally required the assistance of statisticians and computer scientists to develop advanced methods in analysing data with mathematical, statistical, and algorithmic methods, hence the development of the field 'bioinformatics' (Luscombe *et al.*, 2001). Throughout the years efforts were made to make these data more publicly available (Hubble *et al.*, 2009; Barrett *et al.*, 2007; Parkinson *et al.*, 2007), which allowed the researchers in multiple fields to access them freely. This in turn facilitated the investigation of different biological questions that impacted on translational medicine (Robinson, 2014).

Predictive modelling is an important area in bioinformatics that focuses on the biologically relevant question of predicting the outcome (*e.g.*, survival outcome, cancer sub-type) of a set of patients. With the availability of the massive amount of data, the interest of researchers has been increasingly focused on constructing better predictive models to enhance the biomarker discovery. A biomarker is a set of features or variables that is objectively measured as an indicator of a biological or medical state such as the presence of some disease state (Strimbu and Tavel, 2010; Biomarkers-Definitions-Working-Group, 2001). Traditionally, clinical data alone has been used for the identification of biomarkers. However, in the vicinity of high-throughput data, also known as omics data, the current clinical management of critical diseases like cancer has clear potential to be improved (Jayawardana *et al.*, 2015a; Kim *et al.*, 2014; Chin and Gray, 2008; Hanash, 2004). Much work has been done, and is still ongoing in this context, using single data platforms (clinical and omics data) (Schramm *et al.*, 2012; Segura *et al.*, 2012; Tremante *et al.*, 2012; Caramuta *et al.*, 2010) and multiple data platforms (Kim *et al.*, 2014; Mann *et al.*, 2013; Daemen *et al.*, 2009; Boulesteix *et al.*, 2008; Gevaert *et al.*, 2006). Despite the gravity and the urgency of biomarker discovery and validation in a medical context, there are many statistical challenges that need to be addressed in the development of predictive methods. This thesis aims to address some of these challenges in clinical data and omics data, as well as in integrating these components.

In the following sections of this chapter we outline the different components of this thesis. In Section 1.1, we discuss the motivation behind this work and outline the flow of the remainder of the thesis. This chapter continues in Section 1.2 with describing the different types of data used to explore and develop the statistical methods in this thesis, followed by a discussion on data integration. In particular, we explore two types of data integration, horizontal integration and vertical integration. The methods, statistical challenges and advantages of the data integration will be outlined in Section 1.3. We conclude the chapter by providing a background on melanoma in Section 1.4, on which the datasets of this thesis will be based, setting the context for the biological implications of our findings in the subsequent chapters.

1.1 MOTIVATION AND OUTLINE

In many critical diseases like cancer, it has been observed that the conventional clinicopathologic parameters and the standard staging procedures are insufficient in assigning prognosis at individual patient level (Weigelt *et al.*, 2010; van't Veer and Bernards, 2008; John *et al.*, 2008; Rosenwald *et al.*, 2002; Watanabe *et al.*, 2001; Alizadeh *et al.*, 2000). Specifically in melanoma, the standard clinical factors have limited prognostic power at an individual level, as patients with tumours of similar morphology can have markedly different survival outcomes (Table 1.1) (Schramm, 2014; John *et al.*, 2008; Winnepenninckx *et al.*, 2006; Bittner *et al.*, 2000).

This thesis has been motivated by the need for new improved prognostic biomarkers to assist in personalised medicine, where the availability of multiple data types (clinical and omics data) holds the potential to significantly improve upon the current standards (van't Veer and Bernards, 2008; John *et al.*, 2008). Such a discovery of improved biomarkers is vital in assigning treatment therapies reliably at the patient level. The investigation in this thesis is strengthened by the availability of a complex dataset of melanoma patients (described in detail in Chapter 2), which enables to examine both molecular and clinical information across patients. Furthermore, there is prospect in improving upon the current methods used in the field using recent methodological advances such as the weighted Lasso (Bergersen *et al.*, 2011; Zou, 2006), to uncover improved prognostic markers with the use of the multiple datasets available.

With this motivation in mind, we proceed to develop statistical methods and frameworks in this thesis using single data platforms (clinical and omics data) and integrating multiple data platforms, in the context of predictive model building. Chapter 2 outlines the datasets used in this thesis. It aims to give a clear description on the types of data used, on the preliminary analysis performed including pre-processing and quality control, as well as on the performance evaluation procedures adapted in subsequent chapters.

Appropriate and sensible low-level analysis is critical in removing unwanted noise for downstream analysis and further complex methodological development. Therefore, in this thesis we propose platform specific methods, prior to developing data integration frameworks. In this context, Chapter 3 develops a framework for clinical data, mBMI (modified bootstrap multiple imputation framework), modifying and extending the bootstrap multiple imputation (B-MI) framework proposed by (Campain, 2012). The proposed framework, while addressing the common problems in clinical data as missing data and instability in final models, aims to construct a model with good predictive capabilities.

Chapters 4 and 5 proceed on to the timely concept of data integration, developing novel statistical frameworks for integrating multiple data sets to discover improved biomarkers. In particular, Chapter 4 explores horizontal data integration, outlining a comprehensive meta-analysis procedure to identify a robust set of biomarkers. This includes some of the work presented in (Jayawardana *et al.*, 2015b).

Chapter 5 proposes novel frameworks for vertical data integration using multiple types of data (clinical and omics). It makes conceptual advances in statistical bioinformatics with regard to personalised medicine, uncovering subsets of patients for which more dynamic and complex translational models (using both clinical and omics data) are needed. Some of the work in Chapter 5 is published in (Jayawardana *et al.*,

2013) and (Jayawardana *et al.*, 2015a). The thesis concludes with a general discussion and future research directions in Chapter 6.

1.2 BACKGROUND

1.2.1 *Clinical data*

Traditionally clinical data has been used and is still used in assigning prognosis to patients with critical diseases like cancer. Even with the advancement of high-throughput data technologies and the generation of a myriad of data, the importance of clinical data is not reduced (Jayawardana *et al.*, 2015a). The wide range of information contained within the clinical data, from clinical variables such as age, sex, to pathological and mutation information, might be one of the reasons behind this. Hence it is of utmost importance that this vital source of data is analysed accurately, addressing the challenges in statistical clinical research. Although clinical information from multiple data sources will be used throughout this thesis, the Mann clinical data (Mann *et al.*, 2013) will be used primarily to investigate multiple aspects of clinical data (Chapter 3).

In the analysis of clinical data to uncover important predictive variables, regression models are used to study the relationship between p explanatory/predictor variables (collected in the design matrix $\mathbf{X} \in \mathbb{R}^{(p+1)\times n}$, the first row of $\mathbf{X}$ is a row of 1's to model the intercept) and the response variable ($\mathbf{y} \in \mathbb{R}^n$). The generalised linear model (GLM) (Nelder and Wedderburn, 1972) extends the linear regression model to address a wider range of data. In general, in a GLM the expectation of the response ($E(y_j) = \mu_j$) is modelled as

$$g(\mu_j) = \mathbf{x}_j^T \boldsymbol{\beta},\ j = 1, 2, \dots, n,$$

where $g(\mu_j)$ is a monotone function called the link function that connects the expected value of the response variable (the random component) to a linear combination of explanatory variables (the systematic component), $\mathbf{x}_j^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \dots + \beta_p x_{pj}$, where $\boldsymbol{\beta}$ is the vector of regression parameters. In this thesis, the logistic regression model will be used, as the response variable (y) is binary, which is coded as 1 and

0 respectively, to represent good prognosis and poor prognosis groups. Here, the link function is *logit*, given as

$$g(\mu_j) = \ln\left(\frac{\mu_j}{1-\mu_j}\right).$$

More details are given in Chapter 3.

Despite the wide use and importance of clinical data, many apparent issues need to be dealt with in order to obtain a stable predictive model. Missing data is one such problem that is unavoidable in many clinical studies, and hence there is a substantial amount of literature that discusses missing data in clinical research, including (Little and Rubin, 2002; Schafer, 1999; Rubin, 1987). Most statistical methods require complete datasets and this has hindered their use in the presence of missing data. The data could be missing completely at random (MCAR) where missingness is unrelated to variables in the dataset, missing at random (MAR) where missingness is completely random when conditioned on data available and missing not at random (MNAR) where missingness is related to missing data (Little and Rubin, 2002). Over the years, statisticians have used many approaches in handling missing data. Complete case analysis is one such approach that can be used if data are MCAR. However, considering only the complete cases increases the loss of precision and power and can introduce bias. Imputation has become a more commonly used method in handling missing data, as it allows the construction of a complete dataset and continuation with the commonly used complete data methods, while having more power and potentially less bias than the complete case methods (Rubin, 1987).

In recent years dealing with missing data has become less problematic when methods such as multiple imputation (MI) became available in standard statistical software. Although there are many other techniques to handle missing data (Barnes *et al.*, 2008; Carpenter and Kenward, 2008; Ibrahim *et al.*, 2005; Robins *et al.*, 1994; Little, 1986), this thesis will use multiple imputation, as it is one of the most widely used methods in dealing with missing data in clinical studies. There is a vast array of literature that discusses various aspects of multiple imputation including (Sterne *et al.*, 2009) that addresses potential and pitfalls, (Harel and Zhou, 2007) that reviews theory behind MI and compares software available and (Campain, 2012) that compares different imputation algorithms. The use of MI algorithms depends on the structure of the missingness within the data. Most MI algorithms such as Amelia II (King *et al.*, 2001) assumes that the missingness is MAR, whereas there are other algorithms such as MICE (van Buuren and Groothuis-Oudshoorn, 2011) that can handle both MAR and MNAR. In this thesis we use AMELIA II and MICE to handle missingness in our clinical data. A detailed investigation of MI and MI algorithms is beyond the scope of this thesis.

Another problem in analysing clinical data is model instability. In the variable selection procedure, in selecting the most important variables to describe the response, a small perturbation in the original data set can potentially incur large changes in the final model in many instances (Steyerberg *et al.*, 2000; Breiman, 1996), hence the model instability. This can be explored by perturbing the data, taking sub-samples or re-samples via methods such as the bootstrap (Efron and Tibshirani, 1986; Efron, 1979) and considering multiple sets of predictors or variables in the selection of the final model. There are many reasons behind the model instability, including highly correlated predictors (Curto and Pinto, 2007; Kiers and Smilde, 2007) that lead to unreliable regression coefficients and hence erroneous conclusions. Several methods have been proposed to deal with model instability, such as methods that address multicollinearity in the predictors (Kiers and Smilde, 2007) and model averaging techniques that address the variability in the models developed from re-samples (Schomaker and Heumann, 2014; Campain, 2012; Heymans *et al.*, 2007).

Chapter 3 develops 'mBMI', modified bootstrap multiple imputation framework, to build a stable model for clinical data addressing missing data through multiple imputation and instability through bootstrap sampling. The mBMI introduces a new measure for selecting a sparse subset of variables depending on their predictive capability to be assessed for the final model, rather than considering the whole set of variables.

1.2.2 *High-throughput biomedical data*

Over the past decade layers of complexity in the cell has been uncovered, resulting in the generation of more data sources at different levels. DNA (deoxyribonucleic acid), the basic component of a chromosome, is the heredity material or the information carrier in all living organisms. It is made up of molecules called nucleotides assembled as a chain consisting of four chemical bases called adenine (A), thymine (T), cytosine (C) and guanine (G). A DNA molecule consists of two complementary strands of nucleotides arranged in a double helix. A hydrogen bond will be formed between complementary base pairs binding adenine with thymine (A to T) and cytosine with guanine (C to G) (Chargaff, 1951). The DNA molecules are packed into thread-like structures called chromosomes (23 pairs for humans) in the nucleus of each cell.

The process by which the inherent information embedded within the DNA, known as genes, are synthesised into physical or biological outcomes is called gene expression and this constitutes one of the central tenets of molecular biology (Crick, 1970). The central dogma of molecular biology (Crick, 1970) explains the flow of genetic information in a biological system and is often beneficial in understanding cell biology (Figure 1.1): the information coded in DNA is passed on to a type of RNA (ribonucleic acid) called messenger RNA (mRNA) and the mRNA can travel outside the nucleus into the cytoplasm to create functional proteins.

mRNA

Messenger RNA (mRNA) is created by a process called transcription, where the information of a gene is copied onto a complementary single stranded RNA molecule. This is governed by complementary base pairing between the DNA and RNA, where A in the DNA is transcribed to U (uracil) in the RNA, T to A, G to C and vice versa. After this process mRNA has the capacity to travel through the cell, undergoing post-transcriptional modifications to be translated into amino acids or proteins (Watson *et al.*, 2013).

In this thesis we primarily use mRNA data measured via microarray technology. Microarrays allow to measure the activity level or the expression level for thousands of genes in a single sample simultaneously. To achieve this purpose, microarrays make use of the hybridisation of DNA, the process by which a single stranded DNA obtains a correct complementary strand, by measuring the concentration of the gene's mRNA transcript in the cell's total RNA (Karakach *et al.*, 2010). To determine this, DNA probes are immobilised onto a slide, printed as spots or coded onto beads. Fluorescent labelled

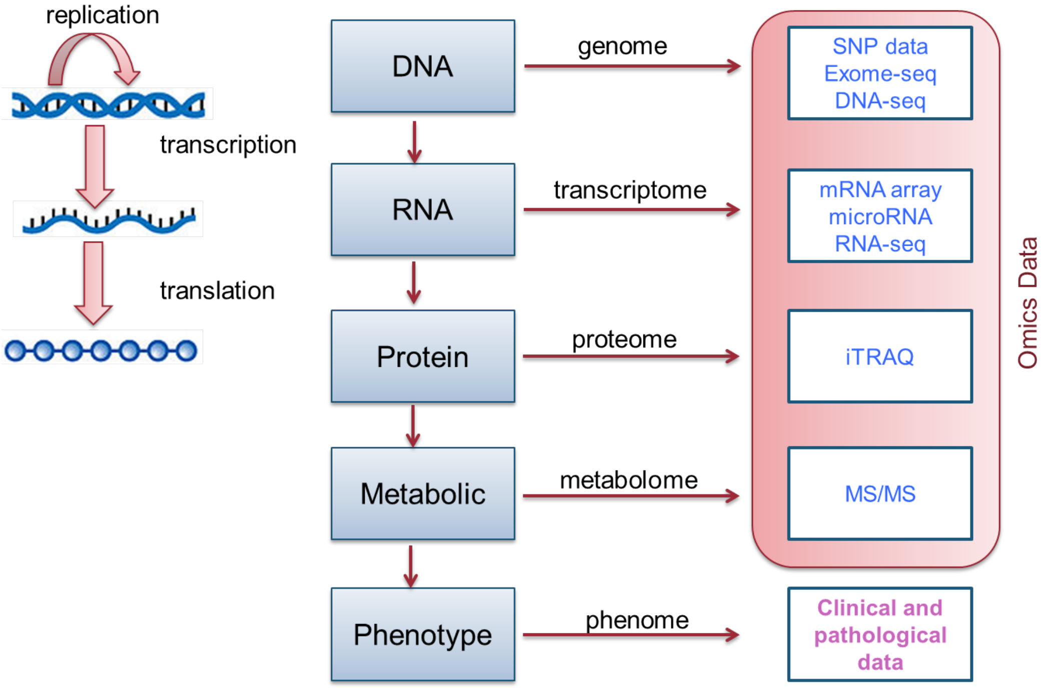


Figure 1.1: **Central dogma of molecular biology (Crick, 1970) and the generation of multiple datasets at different levels of the information flow:** DNA is created by self replication and the information is transcribed into messenger RNA which can be translated into proteins. At different levels of this process multiple omics data are obtained using high-throughput data technologies, that can obtain measurement of thousands of objects in one experiment in a limited time frame.

or pre-determined complementary DNA (cDNA) strands generated from mRNAs in samples are then placed onto the microarray slide. The labelled cDNAs that represent mRNAs in the cell will then hybridise (bind) to any complementary probes on the microarray slide, leaving its fluorescent tag, which can be measured by scanners to give the relative quantities of mRNA on the sample. The levels of intensity or the colors of the resulting slide are used to determine which genes are more expressed in normal tissues *vs.* which are more expressed in disease tissues such as cancer, which leads to a vast area of research uncovering disease causing genes.

Different types of microarrays have been designed since the first spotted array (Schena *et al.*, 1995). Among the many approaches used are spotted arrays, where the cDNA probes or long oligonucleotides are printed onto the slide (Diehl *et al.*, 2001), 'on-chip' arrays, where the probes are built directly onto the surface of the slide (Auer *et al.*, 2009), and bead arrays, where complementary oligonucleotide sequences are attached to microscopic beads (about 20-30 beads with the same probe sequence) combined and spread over the slide (Fan *et al.*, 2006). Microarray platforms can be identified by the company that constructed them such as Illumina, Affymetrix and Agilent.

Some of the other technologies used to measure gene expression are quantitative real time polymerase chain reaction (qRT-PCR) and next generation sequencing technologies (NGS). qRT-PCR is considered to be a gold standard for measuring gene expression as it is known to generate robust, quantitative expression data for single genes and offers rapid and reproducible results (Klein, 2002). High-throughput data generating technologies had a major breakthrough when next generation sequencing (NGS) technologies emerged in recent years. NGS technologies allowed the investigation of multiple genomes and transcriptomes at much lower costs in an extremely efficient way, producing much higher resolution and coverage than what was previously available, and an incredible volume of data (discussed in detail in (Patrick, 2014)).

Protein

Proteins are made of chains of amino acids. In the process of translation, mRNA leaves the nucleus, interacts with a complex called ribosome, and code for amino acids by reading the sequence of mRNA bases by codons (sequence of three bases in mRNA). Many amino acids are coded by more than one codon, such that the information stored in one gene could be translated to make many different proteins (Alberts *et al.*, 2002).

In this thesis we use protein data obtained via quantitative iTRAQ (isobaric Tags for Relative and Absolute Quantification), which is a chemical labelling method that can be used to quantify proteins and liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS). The proteins are sequenced using tandem mass spectrometry (MS/MS or MS^2), where the sequences are cut into smaller segments and ionised whereby the molecules separate based on mass. When sufficient hits are detected at the detector, these are picked for fragmentation and a spectrum based on the mass is created. The peptide sequence in the mass spectra is identified via methods such as database search and De-novo sequencing (Hughes *et al.*, 2010). The identified peptides are quantified via methods such as iTRAQ, where labels are add onto peptides that are broken apart in MS/MS. Often the relative abundance of a peptide is measured as the ratio of intensities of a peptide across conditions investigated. A review describing the technologies can be found in (Rauniyar and Yates, 2014).

MicroRNA

Despite the availability of genes or DNA segments that code for proteins, it has been found that there are many other parts on the genome that play a vital regulatory role, but are not involved in protein production (Lee *et al.*, 1993). MicroRNAs (miRNAs) are one such class of non-coding RNAs that are central in the regulation of gene expression (Kong and Han, 2005; Bartel, 2004). MiRNAs usually induce gene silencing, causing down regulation of gene expression. Through pairing, miRNAs prevent protein production by suppressing protein synthesis and/or by promoting the degradation of their target mRNAs. MiRNAs are known to be involved in a wide range of biological processes such as cell cycle control, apoptosis and several developmental and physiological processes as cardiac and skeletal muscle development, ageing and immune responses (Mo, 2012; Git *et al.*, 2010). Due to this regulatory role of the miRNAs, they have been implicated in a variety of human diseases like cancers, heart disease and neurological disorders (Mo, 2012). Because of this, miRNAs are intensely studied as potential candidates for diagnostic and prognostic biomarkers and predictors of drug response.

With the increasing interest in miRNAs, the technologies for measuring gene expression have been successfully transferred to measure miRNA expression. The three principal methods used in this context are microarrays (Yin *et al.*, 2008), qRT-PCR (Chen *et al.*, 2005) and NGS technologies (Hafner *et al.*, 2008). A comparison of these three technologies can be found in (Git *et al.*, 2010).

1.3 INTEGRATION OF DATA

Data integration has become a popular field of research among scientists, as it holds the potential to use multiple datasets (both clinical and omics data) to decipher the biological information they contain more effectively, rather than only using a single data source. The increasing availability of this multitude of data enables the researchers to perform analysis on diverse topics to yield a depth of valuable information. The overarching aims of data integration are to achieve improved precision, accuracy and statistical power over that from an individual data source (Hamid *et al.*, 2009; Hong and Breitling, 2008; Choi *et al.*, 2003; Normand, 1999). The information uncovered from multiple data sources are more likely to be robust and reliable than from a single data source, therefore integration is very useful in comparing and validating results from multiple studies as well (Hamid *et al.*, 2009; Hong *et al.*, 2006).

1.3.1 *Horizontal data integration*

In horizontal data integration, a set of statistical tools is used to combine multiple studies or data sources of the same type that answer related hypotheses for conclusive inference. In the more traditional sense, horizontal data integration deals with two levels of integration: 'mega-analysis', where the datasets are combined prior to analysis making a single 'mega' dataset which will then be analysed to investigate a biological question; and 'meta-analysis', where statistics from different studies are combined to make common conclusions.

Mega-analysis especially focuses on normalisation of the merged dataset using methods including null correction, quantile normalisation, ComBat (Johnson *et al.*, 2007) and RUV-2 (Gagnon-Bartsch and Speed, 2011) discussed in detail in (Campain, 2012). Comprehensive reviews on meta-analysis using microarrays and genome-wide association studies (GWAS) are provided in (Begum *et al.*, 2012) and (Tseng *et al.*, 2012).

Microarray meta-analysis studies answer a wide range of questions, where a majority deal with detection of differentially expressed (DE) genes (Tseng *et al.*, 2012). Examples of different methods used for DE genes detection are Fisher's method (Rhodes *et al.*, 2002; Fisher, 1950) for combining p-values, fixed and random effect models for combining effect sizes (Choi *et al.*, 2003), methods for combining ranks such as the RankProd algorithm (Hong *et al.*, 2006) and latent variable approaches such as the probability of expression (POE) (Parmigiani *et al.*, 2002).

Meta-analysis has also been widely used for inter-study prediction/classification analysis (Tseng *et al.*, 2012). Validation of biomarkers on external data (Simon, 2011;

Diamandis, 2010; Subramanian and Simon, 2010; Dupuy and Simon, 2007; Ransohoff, 2007) and inter-study prediction (Shen *et al.*, 2008; Beer *et al.*, 2002) belong to this external validation category. Validation on external data in the form of a meta-analysis is important in the sense that it provides confidence for the biomarkers to be used in clinical practise as it is based on multiple studies rather than a single study. This class of meta-analysis has been reviewed and used in many studies (Jayawardana *et al.*, 2015b; Waldron *et al.*, 2014; Schramm *et al.*, 2012; Schramm and Mann, 2011). Other purposes for which meta-analysis has been carried out include pathway analysis (Shen and Tseng, 2010; Setlur *et al.*, 2007; Manoli *et al.*, 2006), network and co-expression analysis (Wang *et al.*, 2009, 2006; Zhou *et al.*, 2005; Segal *et al.*, 2004) and reproducibility and bias analysis across multiple studies (Yang and Sun, 2007; Parmigiani *et al.*, 2004; Kuo *et al.*, 2002).

Horizontal data integration has a range of benefits including the added power to the analysis via increased sample size, 'integration-driven discovery' (Choi *et al.*, 2003), improved reproducibility and reliability (Hong *et al.*, 2006) and investigation of conflicting conclusions in multiple studies (Hong and Breitling, 2008; Normand, 1999). However, there are many challenges and difficulties, especially in inter-study prediction analysis. These difficulties include the differences in study aims, designs, experimental protocols, platforms and heterogeneous patient cohorts, causing discrepancies in populations of interest in multiple studies being considered (Tseng *et al.*, 2012; Campain and Yang, 2010). Different approaches have been used in the literature to overcome these obstacles, including directly merging studies of the same platform before constructing a prediction signature and developing sophisticated normalisation techniques to normalise data across studies to enable the application of the prediction model across studies (Tseng *et al.*, 2012). Chapter 4 directly deals with some of these challenges, where a comprehensive meta-analysis of melanoma miRNA signatures/biomarkers is performed using five miRNA signatures from four data sources (Tembe *et al.*, 2014; Segura *et al.*, 2010; Caramuta *et al.*, 2010).

1.3.2 *Vertical data integration*

In contrast to horizontal integration, the notion of 'vertical data integration' refers to the integration of information for a common set of subjects/patients from multiple sources of data. These data could be measured from a number of distinct platforms or different molecular events such as DNA, mRNA, miRNA, protein and clinical. This type of integration is often more challenging in the sense that data are obtained from very distinct platforms where the number of variables may differ extensively. For example, clinical data typically has about 100 variables and omics data has thousands of variables. Care is required such that variables from one platform do not overpower the other so that variables in one platform are ignored or not selected. For complex diseases, vertical integration could be tremendously advantageous. For example, understanding the complicated cancer genome requires investigating its dysregulation at multiple levels such as the genome, transcriptome and proteome (Kim *et al.*, 2014; Chin and Gray, 2008; Hanash, 2004).

Over the past decade many methods have been developed for vertical data integration in biological studies. However, most of them focus on pairs of molecular events, such as clinical and microarray data (Gevaert *et al.*, 2006; Tibshirani and Efron, 2002), microarray and proteomics data (Daemen *et al.*, 2009), gene expression and copy number data (Bergersen *et al.*, 2011), transcriptomics and proteomics data (Matheis *et al.*, 2011).

Vertical data integration has been used in many studies to answer a multitude of biological questions. Many of these studies use different sources of omics data for clinical decision support, such as survival outcome prediction or disease subgroup prediction. Some studies focus on combined predictive power of clinical and omics data. One of the earlier methods is 'pre-validation' (Tibshirani *et al.*, 2002), where a microarray predictor is constructed and included as one extra variable alongside clinical variables to predict survival outcome. A kernel-based approach is used in (Daemen *et al.*, 2009) for clinical decision support, integrating multiple genome-wide data sources. More recently, (Kim *et al.*, 2014) used a graph-based framework that integrates multiple omics data sources, using an intermediate integration approach to predict cancer clinical outcome. Bayesian approaches, such as those using Bayesian networks (Gevaert *et al.*, 2006) and 'iBAG' (Wang *et al.*, 2013), have also been used for integrative analysis to study the association with patient survival outcome. Vertical data integration has also been used in contexts other than clinical decision support. For instance (van Iterson *et al.*, 2013) integrated miRNA and mRNA expression data to improve miRNA target predictions.

Broadly, vertical data integration studies can be grouped into three categories based on the primary focus of the study (Wang *et al.*, 2013; Daemen *et al.*, 2009). In the first category, 'sequential integration', data from different sources/platforms are analysed sequentially. One type of data is analysed first and another type is used to confirm or clarify the findings, for understanding the biology underlying a disease (Qin, 2008; Tomioka *et al.*, 2008; Fridlyand *et al.*, 2006).

The second group is 'biological integration' (Wang *et al.*, 2013), where the datasets are often merged at database level by cross-referencing the identifiers for common analysis. Some examples are biological pathway analysis, studying regulatory mechanisms and studying inter-relationships and associations (van Wieringen *et al.*, 2012; Karpenko and D., 2010; Goble and Stevens, 2008; Waters *et al.*, 2006). This type of analysis is often hindered by mismatching issues between samples and the inconsistencies of the biological annotation databases.

The third group of integration, 'model-based integration studies' (Wang *et al.*, 2013), integrates the data from multiple layers of data sources in a mathematical/statistical model to answer common questions such as predicting clinical outcome. Data from multiple platforms are treated equally and the the most relevant features from all available data sources are selected (Daemen *et al.*, 2009; Lanckriet *et al.*, 2004). Most of the studies in this category ignored the inter-relationships among platforms, which was later solved by the introduction of canonical correlation based methods (Waaijenborg and Zwinderman, 2009; Witten and Tibshirani, 2009; Parkhomenko *et al.*, 2007). Supervised sparse canonical correlation analysis was introduced by (Witten and Tibshirani, 2009) to identify linear combinations of two sets of variables that are correlated with each other and associated with the outcome as well. The recently introduced weighted Lasso (Bergersen *et al.*, 2011; Shimamura *et al.*, 2007; Zou, 2006) also falls into this category, as it focuses on using additional data as variable weights to guide the variable selection procedure.

Despite the many advantages of vertical integration, there are many challenges and difficulties associated with this type of integration owing to the heterogeneous data sources it accommodates. This includes the processing of distinct platforms to obtain the optimum signal provided by each data source, imbalance of the number of variables such as that between clinical and high-throughput data, 'large p, small n' framework in which the integration is to be performed that limits the statistical methodologies available for such an integration and the mismatch between the samples which further reduces the sample size (n). Furthermore, performance evaluation is challenging in this multi-platform setting. The evaluation can be done at various levels of the process and which of these is most accurate is an open question. Chapter 5 of this thesis addresses some of these critical challenges and offers several solutions and possibilities for dealing with vertical data integration, which includes work published in collaborative research (Jayawardana *et al.*, 2015a, 2013).

1.4 MELANOMA

This section aims to give a brief overview of melanoma, which is the studied complex disease in our motivational dataset (elaborated in detail in Chapter 2).

Melanoma, the deadliest form of skin cancer, is a significant health problem causing approximately 50,000 deaths annually world-wide (Slipicevic and Herlyn, 2012), accounting for 0.1% of total global mortality (Lucas *et al.*, 2006). Moreover, both incidence and mortality continue to rise in many Western countries (Howlader *et al.*, 2012; Garbe and Leiter, 2009; Thompson *et al.*, 2005; Marrett *et al.*, 2001). It is also one of the most common cancers in young adults, exacting a disproportionate social and economic toll compared to other cancers (de Vries and Coebergh, 2004). Melanoma is one of the most common types of cancers diagnosed in Australia, the country with the world's highest incidence rate for this disease (http://www.melanoma.org.au).

Despite the gravity of melanoma there has been minimal success regarding new treatment therapies, the development of which has been hindered by the difficulty in identifying patients who could benefit from targeted and potentially aggressive systemic therapies (Mann *et al.*, 2013; Jönsson *et al.*, 2010). The treatment options are primarily based on the various stages of melanoma, which are defined by the American Joint Committee on Cancer (AJCC) (Balch *et al.*, 2009, 2001) and include four stages (Table 1.1). The survival outcome of melanoma is significantly variable, rendering the clinical management rather challenging. This problem is particularly apparent for patients with nodal metastatic disease (AJCC Stage III) where 5-year survival estimates range from 29% to 81.5% (Gershenwald *et al.*, 2010). The dominant prognostic factors and the five year survival estimates for various stages of the disease are detailed in Table 1.1.

Table 1.1: **AJCC staging of melanoma** – This table describes the AJCC staging system of melanoma, the dominant prognostic factors associated with various stages and the 5 year survival estimates.

AJCC stage	Dominant prognostic factors	5 year survival estimate
Stage I and II (Primary melanoma: clinically localised disease)	Tumour thickness, Ulcerative state, Mitotic rate	85% to 99%
Stage III (Nodal metastatic disease)	No. of metastatic nodes, Ulceration of primary, Tumour burden	29% to 81.5%
Stage IV (Distant metastatic disease)	Site of distant metastases, Serum LDH (lactate dehydrogenase) level	15%

Given this markedly different survival outcome between the four stages of melanoma and more specifically within Stage III melanoma (illustrated in Table 1.1), there is an urgent need to identify and validate accurate prognostic biomarkers that will assist rational treatment planning. The limited set of current prognostic factors (Table 1.1) is useful in assigning broad probabilities of relapse. However, these factors remain insufficient for personalising melanoma management, which aims to enable the provision of the most appropriate treatment to different subsets of patients to ensure optimal benefit.

In this thesis, we aim to address this biologically significant problem of uncovering improved prognostic biomarkers, via the analysis of data from different platforms individually and integratively. For this purpose, we use the motivational melanoma dataset and other external melanoma datasets available publicly (discussed in Chapter 2). The challenges investigated and the solutions offered in this thesis are generalisable to any similar dataset of complex diseases other than melanoma.

2

DATASETS, PRE-PROCESSING AND EVALUATION

This chapter describes the datasets used in this thesis to illustrate our statistical frameworks. These datasets are summarised in Table 2.1. The structure of the chapter is as follows. In Section 2.1 we outline the main melanoma dataset used in this thesis. The details associated with quality control and pre-processing of the data are also presented. Section 2.2 outlines the three external melanoma datasets, that were obtained from public repositories and from published manuscripts. These datasets are used primarily in Chapter 4 and for validation purposes in Chapter 5. In Section 2.3 we outline different cross-validation procedures used and perform a comparison study between them in Section 2.4.

Table 2.1: **Summary of the datasets used.** This table shows the summary of the datasets used in this thesis: Mann (data from Melanoma Institute Australia (MIA)), TCGA (The Cancer Genome Atlas (http:// cancergenome.nih.gov)), Segura (Segura *et al.*, 2010) and Caramuta (Caramuta *et al.*, 2010). The sample sizes in the complete datasets are also included. Abbreviations: n_{total}, Total sample size; $n_{StageIII}$, The no. of Stage III melanoma patients; $n_{withCPM}$, The no.of patients with matched clinical information; n, The sample size used for the analysis; n_{PP}, The no. of patients in Poor Prognosis group; n_{GP}, The no. of patients in Good Prognosis group.

Data type	Mann	TCGA	Segura	Caramuta
Clinical	$n_{total} = 105$ $n_{StageIII} = 84$ $n = 48$ $(n_{PP} = 22,$ $n_{GP} = 26)$			
mRNA	$n_{total} = 99$ $n_{withCPM} = 79$ $n = 47$ $(n_{PP} = 22,$ $n_{GP} = 25)$	$n_{StageIII} = 43$ $n = 27$ $(n_{PP} = 11,$ $n_{GP} = 16)$		
miRNA	$n_{total} = 95$ $n_{withCPM} = 75$ $n = 45$ $(n_{PP} = 22,$ $n_{GP} = 23)$	$n_{StageIII} = 41$ $n = 23$ $(n_{PP} = 12,$ $n_{GP} = 11)$	$n = 59$ $(n_{PP} = 23,$ $n_{GP} = 36)$	$n = 15$ $(n_{PP} = 7,$ $n_{GP} = 8)$
Protein	$n_{total} = 41$ $n = 33$ $(n_{PP} = 14,$ $n_{GP} = 19)$			

2.1 MANN DATA

This dataset has been provided by Professor Graham Mann's group, and will be referred to as 'Mann data'. The Mann data contains clinical data, mRNA (gene expression) data, miRNA data and protein data, with a common set of samples (matched samples) between them (Table 2.1). Tumour samples were obtained from the Melanoma Institute Australia (MIA) Biospecimen Bank, a prospectively collected repository of fresh-frozen tumours accrued with written informed patient consent and Institutional Review Board approval (Sydney South West Area Health Service institutional ethics review committee (Royal Prince Alfred Hospital Zone) Protocol No. X08-0155/HREC 08/RPAH/262, No. X11-0023/HREC 11/RPAH/32, and No. X07-0202/ HREC/07/RPAH/30). These samples were collected since 1996 through MIA, formerly the Sydney Melanoma Unit (Mann *et al.*, 2013). We use this dataset with the primary purpose of developing improved predictive models for Stage III patients, utilising the molecular and phenotype information.

2.1.1 *Details and pre-processing of the multiple datasets*

Clinical data

The clinical data component of the Mann data includes 84 Stage III melanoma patients (Table 2.1). A multitude of variables were observed for each patient, which included clinical variables, pathological variables and mutation variables (Table 2.2). The data from somatic mutation profiling (BRAF and NRAS mutation status) were identified via the Sequenom OncoCarta v1.0, MelaCarta v1.0 platform followed by MassARRAY25 mass spectroscopy (Mann *et al.*, 2013).

Initial pre-processing of the clinical data included removing variables that contained more than 50% missing data and removing the categorical variables with too small frequencies among categories. This resulted in 21 variables that are used in the clinical data analysis in Chapter 3. Table 2.2 summarises these selected variables along with the percentages and number of missing data for each variable. Table 2.3 shows the number of missing variables per sample. The average overall missingness is 10% for the clinical data.

From the survival data for the 84 Stage III patient samples, two extreme survival groups were identified (Mann *et al.*, 2013):

1. Group 1: Poor Prognosis (PP), survival <1 year after surgical resection and died due to melanoma.

2. Group 2: Good Prognosis (GP), survival >4 years after surgery with no sign of relapse.

This resulted in 48 samples for the clinical data with sample sizes of $n_{PP} = 22$ and $n_{GP} = 26$ in the two survival groups respectively (Table 2.1). Chapter 3 presents the details of the complete analysis conducted using these clinical data, where we propose a framework integrating multiple imputation, bootstrap sampling and logistic regression to build a stable model for determinants of survival outcome.

mRNA data

We use the gene expression microarray data (mRNA data) generated as described below. Total RNA was extracted from fresh-frozen Stage III melanoma tissues and assayed using Sentrix Human-6 v3 Expression BeadChips (Illumina, San Diego, CA). Quality control and data normalisation were performed using variance-stabilising transformation (VST) (Lin *et al.*, 2008) and quantile normalisation as implemented in the R package 'lumi' (Du *et al.*, 2008). The number of probes analysed was reduced from 48,802 to 26,085 after removing the unexpressed probes with a detection p-value less than 0.01. More details are provided in (Mann *et al.*, 2013) and the data are publicly available in GEO[1] (GSE54467). The total number of Stage III melanoma samples was 79 and this was reduced to the two extreme survival groups (GP and PP) producing $n_{PP} = 22$ and $n_{GP} = 25$ (Table 2.1). The difference in the sample sizes between the clinical data ($n = 48$) and the gene expression data ($n = 47$) was due to the unavailability of gene expression microarray data for one, otherwise eligible, sample (Jayawardana *et al.*, 2015a).

[1] Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo)

Table 2.2: **The variables used in the Mann clinical data** – This table shows the details of the variables used in the analysis of the Mann data; the names of the variables, the variable type, the percentage (and the number) of missing datum per each variable for the complete set of 84 Stage III patients and the 48 patients from the two survival groups and the number of patients in each category for factor variables and the range for numerical variables.

	Variable Name	Variable type	Description	Type	% (No.) Missing $n = 48$	% (No.) Missing $n = 84$	Categor-ies/Range
1	Person_Sex	Clinical	gender	Factor (nominal)	0 (0)	0 (0)	F:20, M:28
2	Age_Analysis	Clinical	age (at banking)	Numeric	0 (0)	0 (0)	
3	Tum_NumNodesInv	Pathological	number of metastatic nodes	Numeric	0 (0)	0 (0)	
4	Tum_MetSize	Pathological	size of nodal metastatic tumour (mm)	Numeric	2 (1)	6 (5)	
5	Tum_Extranodal	Pathological	extra-nodal invasion (present vs. absent)	Factor (nominal)	2 (1)	6 (5)	No:30, Yes:17
6	Tum_CellType	Pathological	cell shape (round vs. ovoid, elongated, and spindle)	Factor (nominal)	6 (3)	10 (8)	0:27, 1:18
7	Tum_CellSize	Pathological	cell size (small and medium vs. large)	Factor (ordinal)	6 (3)	10 (8)	0:18, 1:27
8	Tum_Necrosis	Pathological	necrosis (percentage)	Numeric	2 (1)	6 (5)	
9	Tum_Pigment	Pathological	degree of pigmentation (present vs. absent)	Factor (ordinal)	4 (2)	7 (6)	0:30, 1:16
10	Tum_BRAFmut	Mutation	BRAF mutation status (mutant vs. wild type)	Factor (nominal)	2 (1)	6 (5)	No:25, Yes:22
11	Tum_NRASmut	Mutation	NRAS mutation status (mutant vs. wild type)	Factor (nominal)	2 (1)	6 (5)	No:34, Yes:13
12	Prim_Site_SunExp	Clinical	sun exposure of anatomic site of the preceding primary (chronic vs. intermittent and nil)	Factor (nominal)	17 (8)	15 (13)	Chronic:12, Intermittent:29
13	Prim_Stage	Clinical	AJCC stage at diagnosis (I vs. II vs. III and IV)	Factor (ordinal)	2 (1)	1 (1)	StageI:14, II:19, III:14
14	Prim_Naevus	Clinical	nevus in association with primary melanoma (present vs. absent)	Factor (nominal)	35 (17)	29 (24)	Absent:24, Present:7
15	Prim_Breslow	Clinical	Breslow thickness (mm)	Numeric	17 (8)	12 (10)	
16	Prim_Mitos	Clinical	primary tumour mitotic rate (mm2)	Numeric	19 (9)	15 (13)	
17	Prim_Clark	Clinical	Clark level (II vs. III vs. IV vs. V)	Factor (ordinal)	15 (7)	12 (10)	2: 4, 3:9, 4:24, 5:4
18	Prim_Regress	Clinical	regression (absence vs. early (mild or focal) vs. intermediate (immature angiofibroplasia) and late (mature angiofibroplasia))	Factor (ordinal)	21 (10)	14 (12)	Absent:18, Early:11, Late:9
19	Prim_Ulc	Clinical	ulceration of primary (present vs. absent)	Factor (nominal)	17 (8)	13 (11)	No: 28, Yes:12
20	SSM_variable	Clinical	primary melanoma subtype (superficial spreading)	Factor (nominal)	21 (10)	17 (14)	0:18, 1:20
21	NM_variable	Clinical	primary melanoma subtype (nodular)	Factor (nominal)	21 (10)	17 (14)	0:21, 1:17

Table 2.3: **Missing datum per sample in the Mann clinical data** – This table shows the number of missing datum per each sample for the complete set of 84 Stage III patients and the 48 patients from the two survival groups.

Number of missing datum	0	1	2	3	4	5	6	7	8	9
Number of samples ($n = 48$)	28	4	2	2	4	0	0	0	1	7
Number of samples ($n = 84$)	51	7	3	3	6	0	0	1	1	12

miRNA data

The profiling of the miRNA expression data was conducted as follows. Total RNA was extracted from the fresh frozen Stage III melanoma tissues (Roche High Pure miRNA isolation kit Cat. no. 05080576001, Roche Diagnostic, Indianapolis, IN, USA). RNA extract quality and quantity was assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA). miRNA expression profiling was performed using Agilent Technologies' miRNA platform (version 16, Agilent Technologies, Santa Clara, CA). More details are provided in (Tembe *et al.*, 2014). This miRNA dataset is publicly available in GEO (GSE59334). The patient groups were compared for 45 samples with $n_{PP} = 22$ and $n_{GP} = 23$ (Table 2.1). The sample size differences are due to insufficient tissue and unavailability of miRNA expression data at the time of the analysis (Jayawardana *et al.*, 2015a).

We provide some basic biological background on the structure of miRNA data in Section 4.1, where a systematic pre-processing pipeline is elaborated. For the analysis in Chapter 5, the miRNA data were normalised using quantile normalisation (as implemented in the R package 'limma' (Smyth, 2005)), adjusted for the difference in overall mean and probe level mean, and aggregated at the miRNA level, producing 390 unique miRNAs eligible for the analysis.

Protein data

Protein data were obtained via quantitative iTRAQ and liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS). More details are provided in (Mactier *et al.*, 2014). All protein expression ratios were log-transformed. In comparing the two survival groups only 24 patients fell within these two groups. Due to the insufficient sample size, 9 samples from the original prospective collection that were initially ruled ineligible (due to falling outside the survival groups), were included in the analysis to constitute 33 samples in total (Jayawardana *et al.*, 2015a). This gives $n_{PP} = 14$ and $n_{GP} = 19$.

2.2 EXTERNAL MELANOMA DATASETS

In addition to the four datasets from the Mann data described in Section 2.1, three other external datasets are used in the subsequent chapters. These include two miRNA datasets from published studies (Segura *et al.*, 2010; Caramuta *et al.*, 2010) and mRNA and miRNA data from TCGA (Table 2.1). We use these datasets to develop and explore the statistical frameworks constructed in Chapter 4 and Chapter 5. The external miRNA datasets and the Mann miRNA data will be used in the meta-analysis in Chapter 4. More details of the data platforms and the survival classes compared are given in Appendix C (Table C.2). The TCGA mRNA and miRNA data will be used exclusively for the biological validation, illustrating the implications of our vertical data integration in Chapter 5. The cohort sizes are summarised in Table 2.1.

2.2.1 *Segura miRNA dataset*

miRNA data from (Segura *et al.*, 2010) involving assay of formalin-fixed paraffin-embedded melanoma metastases, were generously provided directly by Associate Professor Eva M Hernando-Monge, Ph.D. (Department of Pathology, NYU Langone Medical Center, New York, NY.). We obtained the raw data and pre-processed as described in (Segura *et al.*, 2010). Briefly, the steps followed are:

- The dataset was log-transformed ($\log_2$ scale) and quantile normalised, the latter implemented via the 'limma' R package (Smyth, 2005).
- Data were filtered to exclude miRNAs with low variance across samples (*i.e.,* having a coefficient of variation <1%) and miRNAs with row names equal to 'null'.

This pre-processing resulted in 614 miRNAs for the analysis.

2.2.2 *Caramuta miRNA dataset*

Data for the study by (Caramuta *et al.*, 2010), involving human fresh frozen regional lymph node metastases of patients with cutaneous melanoma, were available in Gene Expression Omnibus (GEO), Accession Number: GSE19387 and downloaded on 23/09/2013 containing 167 miRNA samples. We obtained the processed and normalised miRNA data. The original sample size was 16, with one patient considered as censored in the original survival analysis (Caramuta *et al.*, 2010). This patient with short survival ('M-4') had disease unrelated death and due to the inability to allocate this sample clearly to a survival class, this sample was removed from our analysis. Therefore, for the present analysis only 15 melanoma samples were considered (Table 2.1). Since the dataset contained missing data 2% overall missingness, the missing values were imputed using the k-Nearest Neighbours (KNN) imputation algorithm, implemented in R via the 'impute' package (Hastie *et al.*, 2015) with $k = 10$.

2.2.3 *Data from TCGA*

The publicly available mRNA and miRNA data for Skin Cutaneous Melanoma (SKCM) from TCGA were downloaded and are detailed below. Relatable protein information was not available from TCGA and could not be tested. TCGA case count of SKCM is very large (> 400 as at September 2014). However, the cohort is a heterogeneous mix of sample types and stages, among other factors. In this thesis, only the AJCC Stage III melanomas from the TCGA data were considered. The patient stage at the time of tumour banking was only available for samples from the Melanoma Institute Australia tissue bank, therefore these patients were filtered. The cases where the sample was not from melanoma lymph node metastases and where patient stage at the time of tumour banking was not Stage III were removed.

mRNA data

The normalised mRNA (UNC IlluminaHiSeq_RNASeqV2) expression information along with the corresponding clinicopathological annotations for the subset ($n = 43$) of AJCC Stage III samples available were used to validate findings in Chapter 5. The cohort with similar classes to that of the Mann data (*i.e.*, survival > 4yrs with no sign of relapse or survival < 1yr after resection of metastatic disease) was initially sought. However, partitioning the data in this way resulted in an untenable sample size. Therefore, the criteria for good and poor prognosis were adjusted as follows, which retained 27 samples (Table 2.1):

1. Group1: Poor Prognosis (PP), survival <4 years after surgical resection and dead ($n_{PP} = 11$, the range of survival times 0.4-3.9 years), and;
2. Group 2: Good Prognosis (GP), survival $\geqslant$ 4 years after surgery ($n_{GP} = 16$, the range of survival times 4.6-20.7 years).

Variance stabilised transformation (VST) was applied to the normalised dataset using the R package 'DESeq' (Anders and Huber, 2010).

miRNA data

miRNA (BSGSC IlluminaHiSeq_mRNASeq) data for SKCM were used in both Chapter 4 and 5. The sample cohort available at the time of the analysis in Chapter 5 was different from the cohort used in Chapter 4.

For the analysis in Chapter 5 we used the cohort of 43 samples (explained above in Section 2.2.3: mRNA data). The miRNA data was normalised using the trimmed mean of M-values method (TMM) (Robinson and Oshlack, 2010) and variance stabilised transformation was applied.

For the analysis in Chapter 4 we used the miRNA data downloaded on 23/10/2013, resulting in 41 samples eligible for inclusion in the analysis. We carried out a detailed pre-processing procedure on this miRNA data and the procedure is included in Chapter 4. Detailed consideration of the sample sizes in each survival group (more details are given in Appendix A) and the guidance from Professor Mann for the results to be of clinical relevance, resulted in choosing the following survival groups with a total of 23 samples (Table 2.1) for the analysis:

1. Group1: Poor Prognosis (PP), survival $\leqslant$ 2 years after surgical resection and died due to melanoma ($n_{PP} = 12$), and;
2. Group 2: Good Prognosis (GP), survival $\geqslant$ 3 years after surgery with no sign of relapse ($n_{GP} = 11$).

2.3 EVALUATION OF PROGNOSTIC OUTCOME: CROSS-VALIDATION

This thesis focuses on the predictive capability of the models/frameworks developed or compared. Therefore, in the majority of this work we use the prediction error rate or misclassification error to assess the performance. In this section we outline the different CV procedures used in this thesis to enhance the clarity.

As discussed in Chapter 1, performance evaluation is one of the key challenges in data integration, or more generally in bioinformatics method development. Only through repeated application in real world independent datasets that have similar characteristics to the training data, can a statistician be confident of the methods developed. Such an exposure to real data ensures that the statistical methods developed will yield meaningful results with impact in medical research.

A major challenge in performance evaluation is the unavailability of independent data that are similar to the data at hand, to validate the models. Such situations gave rise to a variety of methods for performance evaluation. The holdout method, where the data are partitioned into two mutually exclusive subsets (a training set and a test set), is one such method. However, the holdout method is known as a pessimistic estimator as it makes inefficient use of the data, because only a portion of data is used for training (Kohavi, 1995). The holdout method is also hindered by the limited availability of samples, especially in molecular data context. Bootstrap estimation (Efron, 1979), where a bootstrap sample is created by sampling with replacement from the data, addresses this limited sample availability issue. However, the bootstrap is known to have large bias and low variance (Kohavi, 1995; Bailey and Elkan, 1993).

The finite amount of sample availability demands the re-use of samples through repeated partitioning of the data into training and testing samples. Cross-validation (CV) is such a method that makes efficient use of the available samples while offering a straightforward way of performance evaluation. This is achieved by randomly splitting the available samples repeatedly into training sets to train models and testing sets to validate the model. There are many variations of the CV:

- k-fold CV: The total sample size (n) is split into k subsets, where in each of the k iterations, a subset is used as the test set and the remaining ($n - n/k$) samples are used as the training/learning set. By ensuring that this split is done randomly, one could eliminate or minimise the possible biases of samples of similar type grouping together. However, even though the sampling is done randomly, k-fold CV could still be biased if a seed is chosen to split the samples. The most commonly used k's include choosing $k = 5$ and $k = 10$.
  The repeated application of k-fold CV is shown to be a better approach than performing k-fold CV once, in dealing with bias and variance of the estimations (Efron *et al.*, 2004; Kohavi, 1995; Burman, 1989).

- Leave-one-out CV (LOOCV): LOOCV, also known as complete CV, is a special case of k-fold CV with $k = n$. LOOCV removes the possibility of samples clustering, however this is more computationally expensive. LOOCV also has low bias and high variability in estimation (Bailey and Elkan, 1993; Efron, 1983).

- Stratified CV: Similar to k-fold CV. However, the folds are stratified so that they contain the same proportions of labels as the outcome variable. This method is known to be better, both in terms of bias and variance in estimations than k-fold CV (Kohavi, 1995; Weiss and Indurkhya, 1994).

Apart from the different variations of the CV discussed above, a mixture of various CV procedures are used in the literature (van Vliet *et al.*, 2012; Varma and Simon, 2006; Wessels *et al.*, 2005; Michiels *et al.*, 2005). Direct comparison between the CV error rates from these procedures is difficult and sometimes invalid due to the differences in their layouts. Therefore, it is imperative to understand the differences in the procedures and adapt a CV procedure consistently throughout an analysis, when the aim is to compare between methods or models. It is clear that when developing a prediction model or a prediction framework, each aspect, such as the feature selection, classification algorithm or parameter selection for the classification algorithm, should be inside the CV loop. This ensures that the estimation has low estimation bias. However, this is not always feasible in the vicinity of complex data structures, the complexity in the framework as well as time constraints. Therefore, in most instances when dealing with real life data, one is limited to the validation of the most important aspects of the process. As already pointed out in the literature, 'to choose a classifier or to combine classifiers, the absolute accuracies are less important and we are willing to trade off bias for low variance, assuming the bias affects all classifiers similarly' (Kohavi, 1995).

To improve the clarity and readability of the remaining of the thesis, in the following we discuss and compare the different CV layouts used. The main focus is given to variable/feature selection and classification components of the frameworks. A detailed comparison of different tuning parameters within the classifiers is out of the scope of this thesis. In this thesis, whenever k-fold CV is used for performance evaluation, we use the repeated application of the CV (100 runs), which is shown to be a better approach as discussed above. The repeated CV also allows us to assess the variability of the error estimates, giving a better sense of the prediction accuracies compared in the subsequent chapters.

2.3.1 *CV procedure A: FullCV*

This procedure (Figure 2.1) involves the cross-validation of both feature selection and classification. An additional layer (or loop) of CV is introduced to obtain the optimum number of features to be included in the classification algorithm. Therefore, instead of having a fixed classifier, it includes a fixed classifier training algorithm (Varma and Simon, 2006). The FullCV method avoids the bias introduced by using all of the training data to choose the features (Reunanen, 2003; Simon *et al.*, 2003; Ambroise and McLachlan, 2002). This method is similar to the pre-validation approach (Jayawardana2014, Vliet2012, Tibshirani2002). FullCV can be used for the process of identifying novel biomarkers/prognostic signatures, and works as follows:

1. The data are randomly split into k folds ($k = 5$) with one part retained as the test set and the remaining $k - 1$ parts used as the training set in each fold.

2. For each training set, an internal CV is performed to select the optimum number of features (*i.e.,* the biomarker/signature) - using the training set, multiple sets of features are selected (*e.g.* $5, 10, 15, \ldots, 100$ number of features) and a CV is performed within the training set to select the optimum number of features out of these sets giving the lowest error rate.

3. This optimum set of features is then selected using the complete training set using the same feature selection method employed in the previous step.

4. A classifier is constructed on the training set, using the selected features.

5. The performance of the classifier is then evaluated on the test set, producing the predicted values for each observation in the test set.

6. Steps 2-5 are repeated for all allocations of the k folds into the training and test set, producing a complete prediction vector with one prediction for each sample.

7. Steps 1-6 are repeated for S runs ($S = 100$), where in each repetition the dataset is newly split into k folds, producing 100 predicted vectors.

8. Finally, these predicted vectors are compared against the actual classes (*e.g.* survival outcome: good and poor prognosis) to produce S prediction error rates.

Despite the optimal nature of the FullCV procedure, one cannot ignore the other possible methods of CV which could be comparably efficient in terms of computational time. Moreover, when a comparison between a set of procedures or models is made, if the scientist uses the same validation protocol, it will still be a valid comparison.

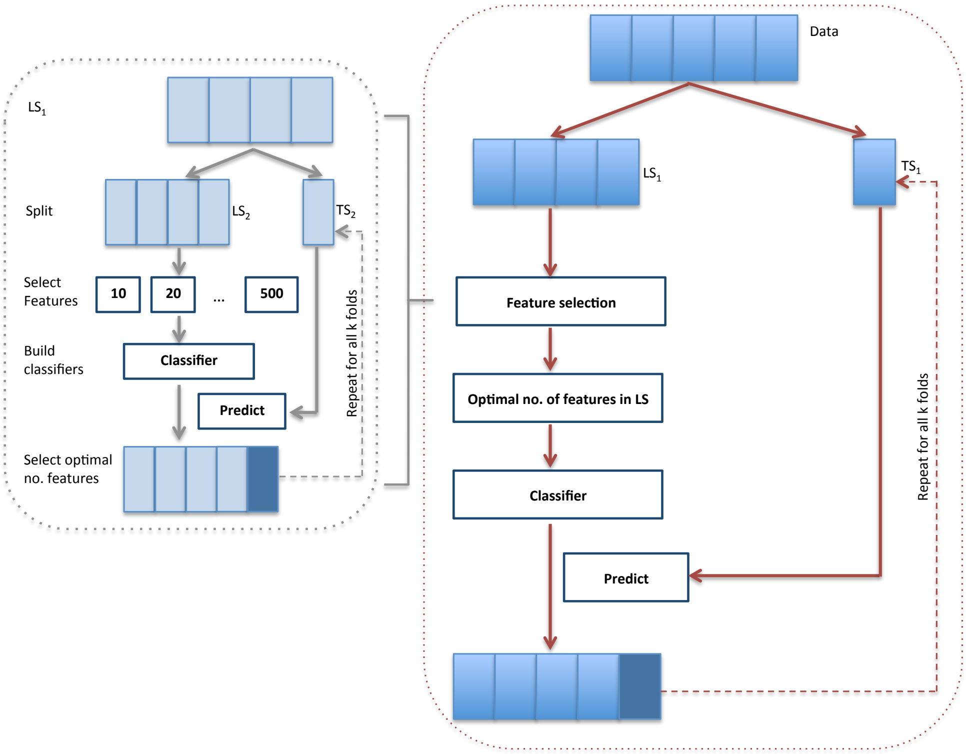


Figure 2.1: **CV procedure A: FullCV**

2.3.2 *CV procedure B: ClassifierCV*

In 'ClassifierCV', the CV is performed to evaluate the classifier only, but not the feature selection (Figure 2.2). This method is similar to pre-specifying the number of features to be selected, outside the CV loop. Hence, ClassifierCV is biased towards feature selection, possibly under-estimating the errors, as the test set takes part in the training process. The procedure is as follows;

1. The data are randomly split into k folds ($k = 5$) with one part retained as the test set and the remaining $k - 1$ parts used as the training set in each fold.

2. For each training set, multiple sets of features are selected (*e.g.* top$10, 20, \ldots$), and classifiers are trained for each of the feature sets.

3. The performance of the classifiers are then evaluated on the test set, producing the predicted values for each observation in the test set.

4. Steps 1-3 are repeated for all allocations of the k folds into the training and test set, producing complete prediction vectors with one prediction for each sample for each feature set.

5. These predicted vectors are compared against the actual classes (*e.g.* survival outcome: good and poor prognosis) to produce prediction error rates.

6. Optimal number of features is then selected as the feature set that produced the lowest error rate.

7. Steps 1-6 are repeated for S runs ($S = 100$), where in each repetition the dataset is newly split into k folds, producing 100 prediction error rates.

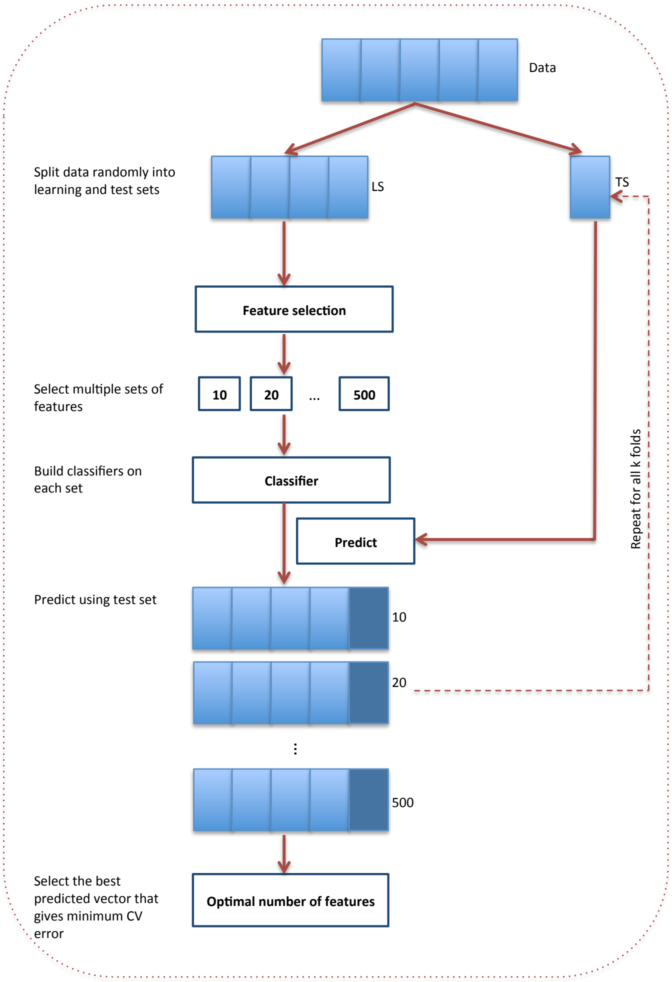

Data
Split data randomly into learning and test sets
LS
TS
Feature selection
Select multiple sets of features
10
20
...
500
Build classifiers on each set
Classifier
Predict
Repeat for all k folds
Predict using test set
10
20
⋮
500
Select the best predicted vector that gives minimum CV error
Optimal number of features

Figure 2.2: **CV procedure B: ClassifierCV**

2.3.3 *CV procedure C: ResubCV*

In this procedure ('ResubCV'), feature selection is performed by comparing re-substitution error rates (Figure 2.3). The procedure is as follows:

1. The data are randomly split into k folds ($k = 5$) with one part retained as the test set and the remaining $k - 1$ parts used as the training set in each fold.
2. For each training set, multiple sets of features are selected (*e.g.* $\text{top}10, 20, \ldots$) are selected, and classifiers are trained for each of the feature sets.
3. The performance of the classifiers are then evaluated on the same training set, producing the re-substitution error rates.
4. The optimal number of features is selected using the lowest re-substitution error rate, a classifier is trained using this number of features on the training set, and the performance is evaluated using the test set.
5. Steps 1-4 are repeated for all allocations of the k folds into the training and test set, producing a complete prediction vector with one prediction for each sample.
6. Steps 1-5 are repeated for S runs ($S = 100$), where in each repetition the dataset is newly split into k folds, producing 100 predicted vectors, which in turn will be compared against the actual classes to produce S prediction error rates.

2.3.4 *CV procedure D: FixedCV*

In this procedure ('FixedCV') a pre-determined number of features (*e.g.* top 20) is used (Figure 2.4). One could use some other procedure to determine the optimal number of features or use prior knowledge to determine this optimal number. This procedure violates the assumption that all training is done within the CV procedure, thus incurring a bias (Varma and Simon, 2006). Briefly, the dataset is split into learning and test sets as explained previously. The fixed number of features (pre-determined) is selected using the learning set and a classifier is trained on this set of features, which will

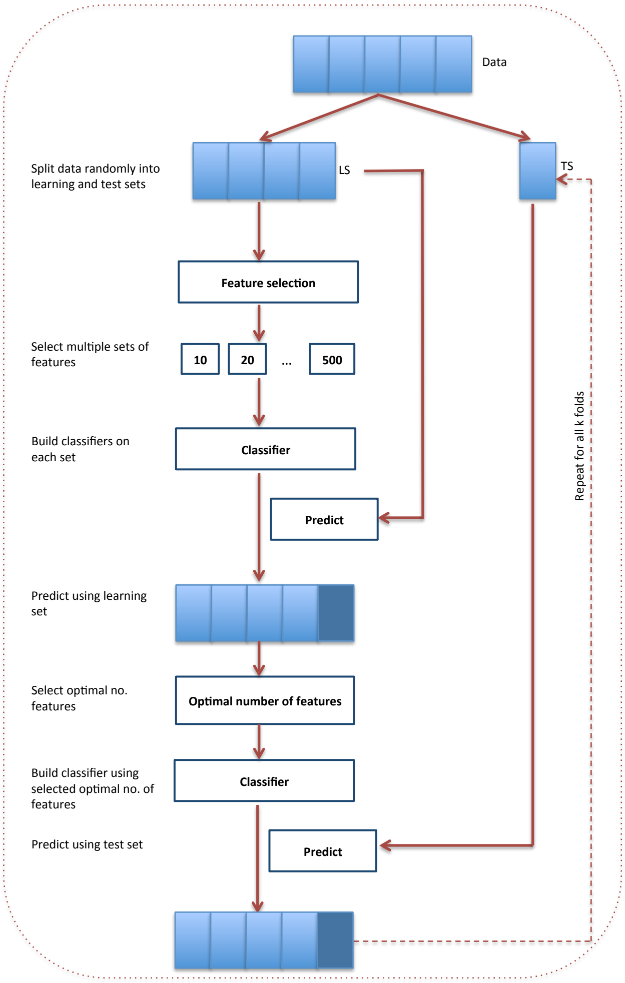

Data
Split data randomly into learning and test sets
LS
TS
Feature selection
Select multiple sets of features
10
20
...
500
Build classifiers on each set
Classifier
Predict
Repeat for all k folds
Predict using learning set
Select optimal no. features
Optimal number of features
Build classifier using selected optimal no. of features
Classifier
Predict using test set
Predict

Figure 2.3: **CV procedure C: ResubCV**

then be tested on the test set. Repeating the procedure for all k iterations will produce a complete prediction vector. Repeated S runs of the procedure result in S predicted vectors and thus, S prediction error rates.

2.3.5 *CV procedure E: FinalCV*

This procedure is often used for validating final models (Figure 2.5). The 'FinalCV' procedure is often useful when comparing the prediction accuracy of previously constructed biomarkers/signatures in an independent dataset. The procedure is similar to Fixed features CV and follows the same steps. However, in FinalCV the exact set of features/variables is also known apart from the number of features. Therefore, the exact features are selected using the learning set in each fold.

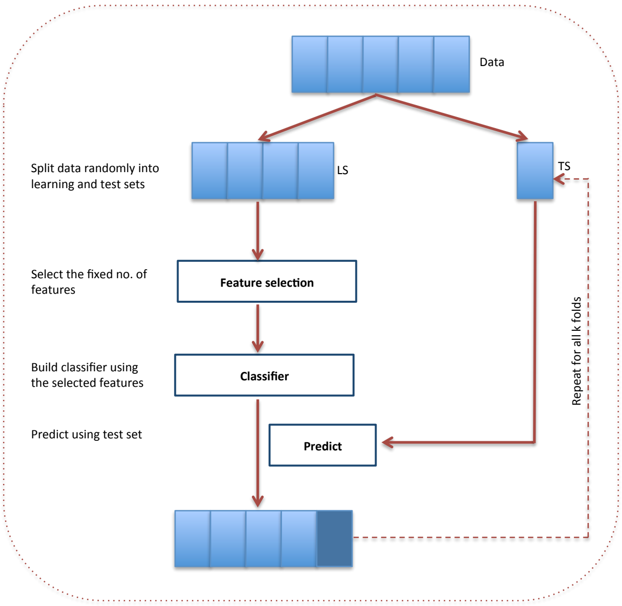

Data
Split data randomly into learning and test sets
LS
TS
Select the fixed no. of features
Feature selection
Build classifier using the selected features
Classifier
Predict using test set
Predict
Repeat for all k folds

Figure 2.4: **CV procedure D: FixedCV**

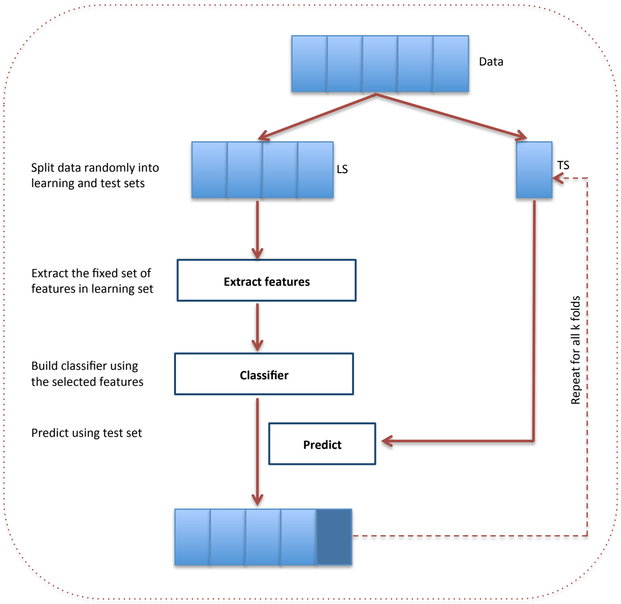

Data
Split data randomly into learning and test sets
LS
TS
Extract the fixed set of features in learning set
Extract features
Build classifier using the selected features
Classifier
Predict using test set
Predict
Repeat for all k folds

Figure 2.5: **CV procedure E: FinalCV**

In the next section, we carry out a comparison study of the CV layouts/procedures discussed above.

2.4 comparison study of the cv procedures

We compare the CV procedures in terms of their prediction capabilities, efficiency and the selected features. For this purpose we use the Mann mRNA data to illustrate the procedures in real data. Features are selected by ranking them using the differences in median expression values of the two groups (GP and PP) which is referred to as the 'median robust method' (Jayawardana *et al.*, 2015a; Campain, 2012) and the classification algorithm employed is support vector machines (SVM) as implemented in 'e1071' R package (Meyer *et al.*, 2014). We use the 5-fold CV repeatedly in $S = 100$ runs of the complete CV protocols, allowing to account for the variability in error rates. For the 'FixedCV' the fixed number of features selected from learning set was set to be 20 and this is an arbitrary choice. The features in the 'FinalCV' procedure were pre-selected using the median robust method (selected the top 20 features) using the complete dataset.

Table 2.4: **Comparison of CV procedures** – This table shows the results from the comparison study of the five CV procedures discussed above.

CV procedure	Mean 5-fold CV error rate	Percentage change from full CV	Total time	Features included in more than 50% of models
A: FullCV	34%	-	3.27hrs	55
B: ClassifierCV	26%	24%	43min	53
C: ResubCV	35%	3%	38min	31
D: FixedCV	33%	3%	28min	13
E: FinalCV	23%	29%	1min	20

We observed that the ClassifierCV and FinalCV procedures produced relatively low mean error rates of 26% and 24% respectively (Table 2.4, Figure 2.6). As discussed above, this could be an under-estimation of the true error rate (which is unknown), attributed to the lack of cross-validating the feature selection properly. The percentage change in CV error rates from the FullCV, which can be assumed to be a closer approximation to the true error rate, are 24% and 29%, respectively, for the two methods. The FullCV, resubCV and FixedCV procedures produced similar CV error rates (34%, 35% and 33%) with percentage changes in mean CV errors from FullCV being

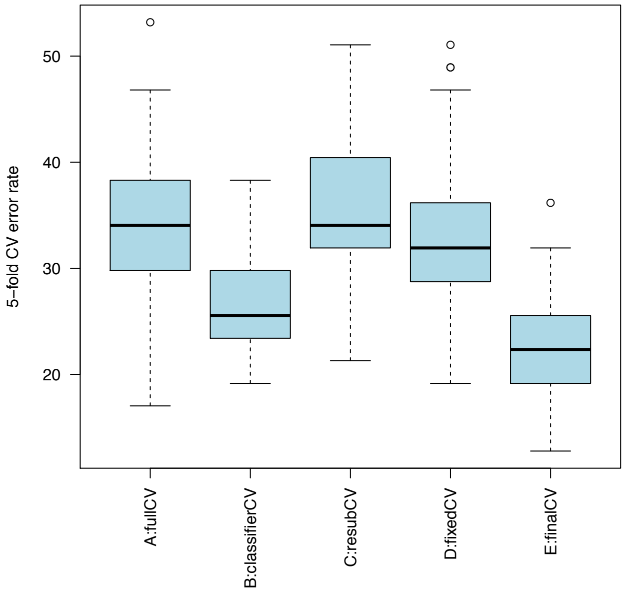


Figure 2.6: **Comparison of CV procedures: Final 5-fold CV error rates**

quite negligible (3%). When comparing the computation time, the FullCV procedure is significantly computationally expensive (3.27 hours) owing to its comprehensive CV procedure, while the others showed similar efficiency levels (43 minutes, 38 minutes and 28 minutes respectively for B, C and D procedures) with one exception, the FinalCV procedure (1 minute).

The final set of features were determined by taking the most frequently included variables (in more than 50% of the models) and these were 55, 53, 31, 13 and 20 (fixed) respectively (Table 2.4). The higher number of stable variables in FullCV (55) and ClassifierCV (53) compared to ResubCV (31), indicates that the FullCV and ClassifierCV are better at producing more stable variables. However, the numbers in FixedCV (13) and FinalCV (20) are limited by the arbitrary choice of fixed number of variables we allowed to be selected in each run ($p = 20$).

The variable inclusion frequencies of the procedures A–D showed a similar pattern (Appendix A). When considering the intersection between the final selected genes (Figure 2.7) we observed that all the genes intersected, with the only difference being the extra genes selected in each procedure. There were 13 such common genes among all procedures A–D, 7 genes common to all procedures except FixedCV, 11 genes common

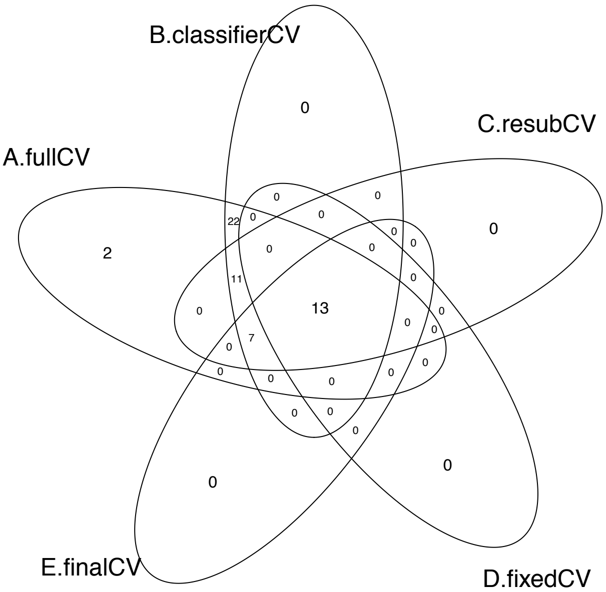


Figure 2.7: **Comparison of CV procedures: Intersection of the genes selected**

to procedures A–C, 22 genes common between A and B and 2 genes for A only (Figure 2.7). As discussed above, the selection of the CV procedure to be used in the evaluation of a process should be based on the most important element to be cross-validated as well as based on a cost-benefit analysis. This comparison also highlights the fact that when we are comparing published papers, we cannot simply compare the published error rates directly, because of the possible differences between the CV procedures.

3

MODEL BUILDING FOR CLINICAL DATA

Clinical data is widely used for decision making in many critical diseases. In diseases like cancer it is still considered to be the primary data source to determine the survival of patients and in allocating patients to various treatments (Balch *et al.*, 2009). Typically, clinical data consists of a combination of nominal, categorical and continuous variables pertaining to different characteristics of a patient.

Statistical research encounters many challenges when analysing clinical data, including missing data and unstable final models. In Chapter 3 of (Campain, 2012), the author proposed the B-MI, a procedure that incorporates bootstrap sampling and multiple imputation to address these challenges. However, the direct application of the B-MI procedure to our primary dataset, Mann data (introduced in Chapter 2), rendered unstable models with high prediction errors. This might be due to the different structure of the Mann data to the data that was used to develop B-MI. For example the sample size in (Campain, 2012) was $n = 416$ in contrast to $n = 48$ in the Mann data.

In this chapter, we extend B-MI to address the issues of model instability and high prediction errors. Our proposed approach, the mBMI (modified bootstrap multiple imputation) framework involves the following modifications and novel features:

1. Selection of the best performing subset of models.

   We select the best performing models that have a small prediction error rate (a pre-specified proportion from all of the constructed models on bootstrapped samples) to derive the final model. This step will ensure an informative prognostic model. In B-MI no such thresholding was used to obtain the final model, instead all of the selected models (using the BIC (Schwarz, 1978)) were considered to obtain the final model. In our data we found that the B-MI procedure did not possess good predictive power.

2. Selection of the stable variables from the best performing models.

   In (Campain, 2012) stability was ensured using an inclusion threshold; if the proportion of models that included a particular variable (inclusion frequency) exceeded this threshold, then that variable was used in the final model. In the mBMI a similar concept of thresholding is used. However, we use the above selected subset of the top models to calculate the inclusion frequency of variables.

This step ensures the stability of the final model, while warranting the predictive capability.

3. Cross-validation (CV) error rate to evaluate the models.

   To be consistent with the remainder of the thesis, we use the CV error rate to evaluate the prediction error of the models. This is in contrast to the area under the receiver operating characteristic curve (AUC) that was used in (Campain, 2012). Furthermore, we used CV to evaluate the mBMI procedure in contrast to the hold-out method in (Campain, 2012), which is known to make inefficient use of data (discussed in Chapter 2).

Although the proposed framework could be applied to any real clinical data set, we were in particular motivated by the detailed series of clinical information in the Mann data. The characteristics in the Mann data pose many additional challenges in constructing models. These include; (i) the smaller sample size ($n = 48$), this makes the missing data problem more prominent, contributing to the instability of the models, and (ii) the large number of variables ($p = 21$) relative to the sample size, that includes both categorical and numerical variables. The clinical variables considered herein are detailed in Chapter 2 (Section 2.1.1). In summary, the Mann data contains 10% missing data overall, and only approximately half of the samples are complete ($n_{complete} = 28$). Two extreme survival groups were identified from the survival times; the good prognosis group (GP) and the poor prognosis group (PP). Therefore, the dependent variable is essentially a binary variable, which is coded as 1 and 0 respectively to represent GP and PP groups in constructing models.

The analysis of the Mann data is carried out in a prognostic setting, the interest being to predict the survival outcome accurately. Therefore, logistic regression is utilised for constructing models as discussed in Section 1.2.1. The probability of an 'event' (*e.g.* death) occurring is μ_j, where $\mu_j = P(y_j = 0)$ and we model μ_j through

$$\ln(\tfrac{\mu_j}{1-\mu_j}) = \mathbf{x}_j^T \boldsymbol{\beta},\ j = 1, 2, \ldots, n,$$

for $\mathbf{x}_j = (\mathbf{x}_{0j}, x_{1j}, \ldots, x_{pj})$; the vector of p explanatory/predictor variables, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$; the vector of regression parameters.

This chapter elaborates the mBMI framework in Section 3.1, where the exploration of the framework constitutes of two parts, a comparison study and a simulation study. It is our primary focus to develop frameworks that work on real data. Therefore, the mBMI framework is applied to the Mann data in Section 3.2 to evaluate its effectiveness on real data. In order to compare the predictive capabilities of the proposed framework with a collection of other methods, a comparison study is performed in Section 3.2. However, a major drawback in using real data to evaluate the framework is that the stability of the variables cannot be assessed, as the true variables are unknown. Therefore, in Section 3.3 the performance of the mBMI framework is investigated using a simulation study, where the stability of the variables are explored to select optimum parameters for the mBMI framework.

3.1 THE MBMI FRAMEWORK

The novel framework developed in this chapter mainly addresses the common challenges when using clinical data for model building: missing data, stability of models and their ability to predict. These challenges are discussed next.

Missing data and multiple imputation

Missing data poses analysis challenges in a wide range of studies. Because of this, there exists a substantial amount of literature discussing missing data including (Little and Rubin, 2002), (Schafer, 1999) and (Rubin, 1987). Despite the presence of many advanced and efficient methods of dealing with missing data, multiple imputation (MI) has been widely accepted as a tool with good properties that can be readily applied in a wide range of datasets (Schafer, 1999). Complete case analysis greatly reduces the statistical efficiency of the analysis by the reduced sample sizes and also removes the underlying structure from the data. In contrast, imputation, where the missing data are filled-in, allows to use complete data analysis approaches while maintaining the original data structure, including the sample size. Furthermore, MI allows to take into account the variability associated with the missing values by imputing these multiple times. The estimated parameters can be aggregated using the average as per Rubin's rule (Rubin, 1987), where the overall estimate is;

$$\bar{\beta}_q = \frac{\sum_{r=1}^{m} \hat{\beta}_{rq}}{m},$$

where $\hat{\beta}_{rq}$ is the estimated value of the q^{th} parameter from the r^{th} imputed dataset. The variance is calculated by combining the between-imputation variance $B = (1-m)^{-1} \sum_{r=1}^{m} (\hat{\beta}_{rq} - \bar{\beta}_q)^2$, and the within-imputation variance $\bar{U} = m^{-1} \sum_{r=1}^{m} SE_{rq}^2$, where SE_{rq} is the standard error of the q^{th} estimated parameter. The estimated total variance is $T_q = (1 + \frac{1}{m})B + \bar{U}$.

When the interest is on the selection of variables, an inclusion frequency (how often the variable is selected out of the multiple models constructed) can be involved in aggregating the parameter estimates. If a particular variable is prevalent in more models than a pre-determined inclusion threshold (Heymans *et al.*, 2007; Austin and Tu, 2004), the variable is included in the final model. The multiple imputation inclusion frequency (τ_{MI}) can be incorporated into the overall parameter estimate through

$$\bar{\beta}_q = \frac{\sum_{r=1}^{m} \hat{\beta}_{rq}}{m} I(\hat{\rho}_q \geqslant \tau_{MI}),$$

where $\hat{\rho}_q$ is the estimated inclusion frequency of variable q. In this thesis, τ_{MI} was set to 50% as a good balance between an overly sparse model and a model with coefficient estimates that incorporate a large number of zero estimates. We did not explore the sensitivity of this choice in detail.

MI algorithms are readily available in statistical software such as R. In this thesis, Amelia II (King *et al.*, 2001) and MICE (van Buuren and Groothuis-Oudshoorn, 2011) are used for imputing clinical data, as these algorithms can handle both continuous and categorical data (for more details on the imputation algorithms we refer to (Honaker *et al.*, 2011) and (van Buuren and Groothuis-Oudshoorn, 2011)). These are popular procedures and we refer for a comparison of various MI algorithms in our context to (Campain, 2012). Amelia II (King *et al.*, 2001) assumes that the variables are jointly multivariate normal and that the data structure is missing at random (MAR; see also Chapter 1). The missing data are imputed using a bootstrap and expectation maximisation (EM) approach. For our analysis we use the R package 'Amelia' (Honaker *et al.*, 2011). Multiple imputation by chained equations (MICE) (van Buuren and Groothuis-Oudshoorn, 2011) is an iterative procedure using the Gibbs sampler, where random variables are generated from a marginal distribution directly without calculating the density (Casella and George, 1992). MICE is available in the R package 'mice' (van Buuren and Groothuis-Oudshoorn, 2011).

Instability of regression models

We call the regression models unstable when a small perturbation in the original dataset results in large changes in the final model. Instability occurs due to many reasons including highly correlated variables (Kiers and Smilde, 2007; Curto and Pinto, 2007), instability of parameter estimates or imbalanced class distributions. This chapter mainly focuses on the instability of parameter estimates that is more common when the sample size is small as in the Mann data. Since this thesis investigates the construction of predictive models, the stability of the model is of utmost importance as unstable models that are not reproducible could cause high prediction errors on independent data.

Predictive modelling

In order to achieve high prediction accuracy, it is imperative to select a stable set of features that are reproducible and perform well on independent data. Therefore, the predictive modelling and the construction of stable models are highly related. In medical research, special interest focuses on building a parsimonious model that is interpretable and applicable in a clinical sense to determine patient prognosis. Therefore, we aim to select a subset of variables that is stable and achieves high prediction accuracy through developing our mBMI framework.

The mBMI framework combines bootstrap sampling and multiple imputation, where multiple subsets of variables with high predictive capability are used to obtain a stable model with high prediction accuracy. The framework is graphically represented in Figure 3.1, and detailed below.

The data set $\mathbf{Z} = (\mathbf{X}, \mathbf{y})$, where $\mathbf{X} = [\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_p]^T$, a $(p+1) \times n$ matrix of explanatory variables, where n is the sample size, p is the number of predictors, and $\mathbf{y}$ is the binary response vector of length n with classes coded as 0 (PP) and 1 (GP).

1. Stage 1: Multiple Imputation

   The data set undergoes m multiple imputations producing m complete datasets denoted by $\mathbf{Z}^1, \mathbf{Z}^2, \ldots, \mathbf{Z}^m$.

2. Stage 2: Bootstrapping

   The b^{th} bootstrap sample is drawn, where $b = 1, 2, \ldots, B$. $\mathbf{Z}_b^r$ denotes the bootstrapped imputed data for the b^{th} bootstrap sample and the r^{th} imputation. The bootstrap sample index vector is fixed over the m datasets. That is $\mathbf{Z}_b^r$ has the same sampled columns for all $r = 1, 2, \ldots, m$. These datasets are paired in that they share the same observed values but have different imputed values. Stratified bootstrap samples are drawn such that the class distribution is consistent and the proportion of complete and imputed samples are maintained in each bootstrap sample drawn. This approach is consistent with (Campain, 2012).

3. Stage 3: Feature selection

   A logistic regression is applied to each of the imputed and bootstrapped datasets, where variable selection is performed using a selection criterion such as Akaike's information criterion (AIC) (Akaike, 1974) or the Bayesian information criterion (BIC) (Schwarz, 1978). This gives for each of the m datasets the estimated bootstrap parameters vector,

   $$\hat{\boldsymbol{\beta}}_b^r = (\hat{\beta}_{0,b}^r, \hat{\beta}_{1,b}^r, \ldots, \hat{\beta}_{p,b}^r),$$

where those components not retained by the variable selection procedure have zero estimates. Any other variable selection criterion could be employed here and the procedure is flexible enough to incorporate feature selection methods as stepwise procedures or regularisation methods such as the Lasso (Tibshirani, 1996) and Elastic net (Zou and Hastie, 2005).

4. Stage 4: Aggregate over MI data sets

   Variables are aggregated across the m data sets using an inclusion frequency, where the variables that are included in more than a certain percentage (here 50%) of the m models are selected and aggregated by averaging the m estimates, thus producing one model.

   The inclusion frequency of the b^{th} bootstrap sample for the q^{th} variable (the proportion of times variable q is retained in the bootstrap sample for m datasets) is

   $$\hat{\rho}_{q,b} = \frac{1}{m} \sum_{r=1}^{m} 1\{\hat{\beta}^{r}_{q,b} \neq 0\}.$$

   The aggregated coefficient for the q^{th} variable is

   $$\bar{\beta}_{q,b} = \frac{\sum_{r=1}^{m} \hat{\beta}^{r}_{q,b}}{m} 1\{\hat{\rho}_{q,b} \geqslant \tau_{MI}\},$$

   where τ_{MI} is the multiple imputation inclusion frequency threshold.

5. Stage 5: Repeat the procedure

   Repeat for $b = 1, 2, \ldots, B$ bootstrap samples and obtain a $(p+1) \times B$ matrix of parameter estimates $\bar{\boldsymbol{\beta}}$, where in each column the variables selected in one bootstrap run will have a non-zero estimate, thus producing B models.

6. Stage 6: Identify models with good prediction accuracy

   For each of the B models, the prediction error rate is calculated, using the CV error rate ('FinalCV' is used; see Section 2.3). The best performing feature combinations (models) are then chosen based on a prediction error threshold (τ_{CV}), resulting in s models. The best performing models can be chosen in two ways:

   - Setting a prediction error cut-off

   $$\textit{Prediction error cut-off} = \tau_{CV}.$$

   $$\textit{Best performing models} = \textit{Models with CV error rate} \leqslant \tau_{CV}.$$

- Selecting the top performing models

  Prediction error threshold $= \tau_{CV}$.

  Best performing models $=$ *Top* $\tau_{CV} \times 100\%$ *of the models ranked according to the prediction error rate.*

7. Stage 7: Identify best performing variables

   Considering the best performing feature combinations (the s models from Stage 6), the features/variables are then ranked based on their inclusion frequency, $\hat{\rho}_q$, where

   $$\hat{\rho}_q = \tfrac{1}{s} \textstyle\sum_{i=1}^{s} 1\{\hat{\beta}_{q,i} \neq 0\},\ q = 1, 2, \ldots, p.$$

8. Stage 8: Final model fitting

   The ranked variable list is then considered in a logistic regression model based on a final inclusion frequency threshold (τ_{FM}). A variable is included in the final model if $\hat{\rho}_q \geqslant \tau_{FM}$, and the final model is obtained by applying a logistic regression model and model aggregation on this set of variables.

   $$\bar{\beta}_q^* = \frac{\sum_{r=1}^{m} \hat{\beta}_q^{*r}}{m},$$

   where $\hat{\beta}_q^{*r}$ is the parameter estimate of the final model for the q^{th} variable in the r^{th} imputed dataset.

This mBMI procedure focuses on the stability of the variables that are included among the best prognostic models to devise a final model. This produces a stable model with good prognostic capabilities. Similar procedures were developed in (Heymans *et al.*, 2007), (Campain, 2012) and (Schomaker and Heumann, 2014), but without focusing on top performing models.

3.2 COMPARISON STUDY

There are numerous procedures and methods for clinical data analysis in the medical statistics literature. It was of interest to assess a selection of these methods relative to our mBMI approach in terms of the prediction error, as predictive modelling is the main focus in this thesis.

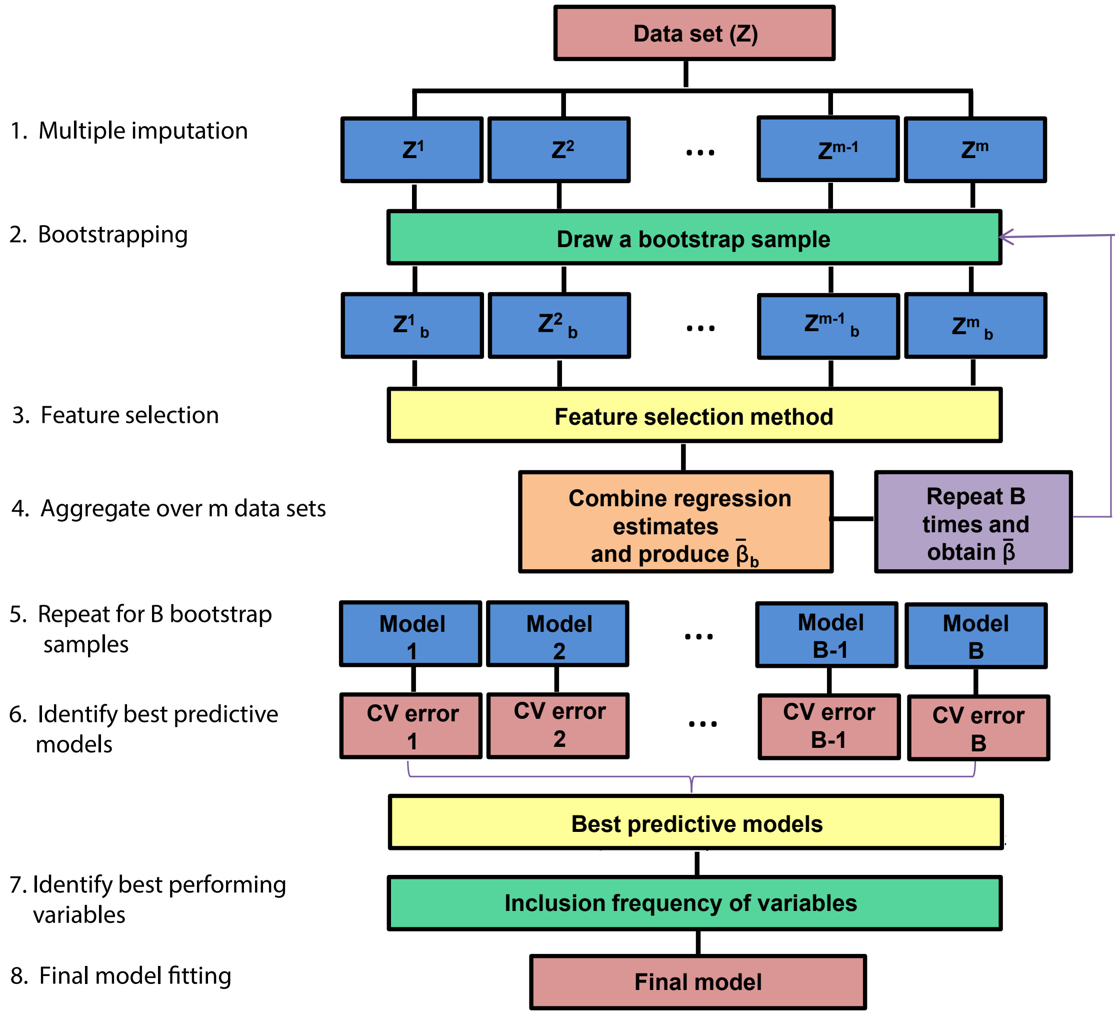


Figure 3.1: **The schematic of the mBMI framework.**

Despite the merits of the mBMI procedure in building stable predictive models for any clinical dataset with missing values, our procedure was primarily constructed to answer biological questions underpinning the Mann dataset. Hence our comparison is based on the evaluation of the final models from each procedure from the Mann data.

3.2.1 *Mann data: The mBMI*

The mBMI framework was applied to the Mann clinical dataset. As discussed in Section 2.1.1, the Mann dataset has clinical, pathological and mutation variables ($p = 21$ after pre-processing) with a relatively low sample size of $n = 48$, when the two survival groups (GP *vs.* PP) are considered.

We applied the mBMI framework with the following parameters held fixed after discussion with biologists and clinicians in the melanoma research group according to the accuracy levels they require.

- Number of multiple imputations: $m = 25$ using Amelia II.
- Variable selection: stepwise selection with BIC as the selection criteria.
- Number of bootstrap samples: $B = 500$.
- MI inclusion threshold: $\tau_{MI} = 0.50$. This value is maintained in the remainder of the thesis.
- Prediction error threshold (cut-off): $\tau_{CV} = 0.30$. This value was set after discussion with the biologists as they required the final model to have prediction accuracy of at least 70%.
- Final inclusion frequency threshold (τ_{FM}). Instead of setting a fixed value for the inclusion frequency in the selection of the variables to be used in the final model, we further expanded this step as follows:

  Step 7 of the mBMI determines the ranked list (according to their inclusion frequency) of best performing variables. The ranked variables are then included in a logistic regression model in a forward algorithm, where in the first model the highest ranked (most frequently appearing) variable is included, and in the second model the top two highest ranked variables are included and so on, until the full set of variables are included into the model. These models are then validated using CV. The final model was selected to be the top set of variables that gave the lowest prediction error.

Results: The final clinical model is the model with the 10 top performing variables (Table 3.1). The error rate for predicting patient outcome (final model fitting step) initially increased as the number of variables included in the final model changes from 1 to 2, and then steadily decreased until it reached its minimum of 20% (Figure 3.2) incorporating 10 variables, before increasing again.

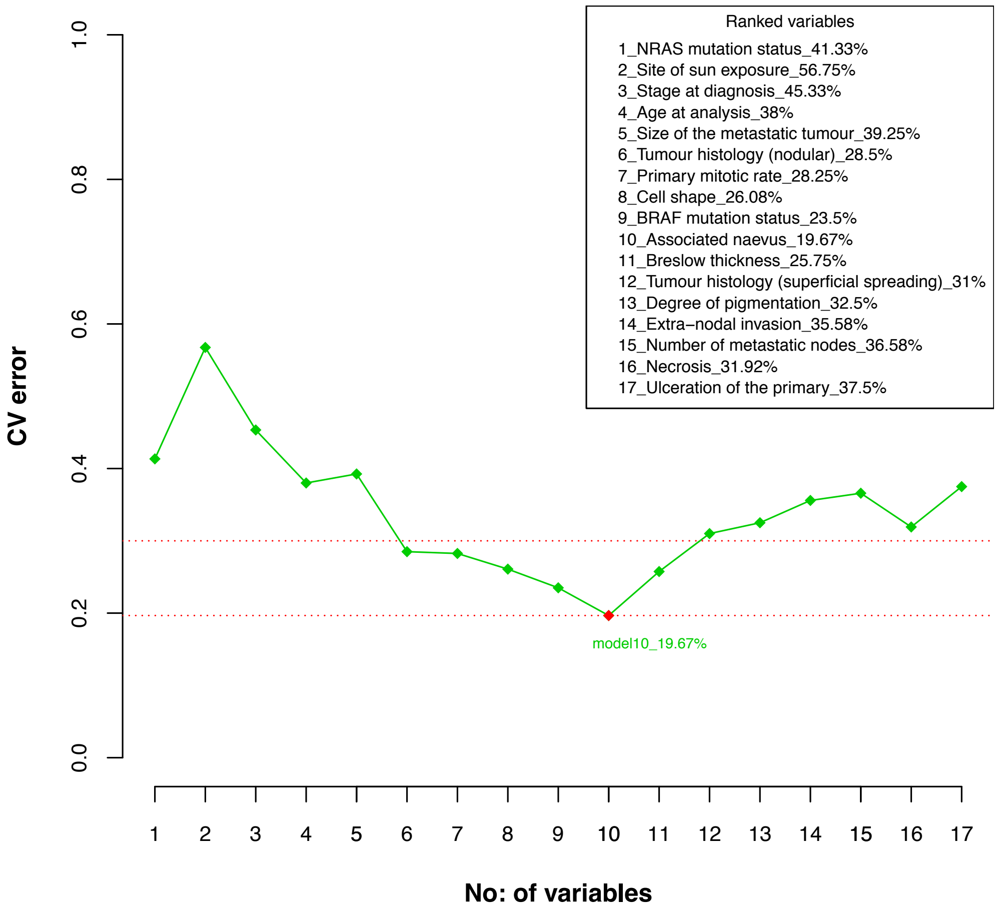


Figure 3.2: **Final variable selection in mBMI.** The application of the mBMI procedure to the Mann data. The fluctuation of the 5-fold CV error rate with the number of variables in the final model in the last stage of the mBMI.

In addition to earlier disease stage (Stage I cf. Stages II/III) at initial presentation, the following clinico-pathologic and mutation factors were associated with improved outcome: intermittent sun exposure of the primary site (cf. chronic), decreased size of the metastatic tumour, younger age, the presence of a nodular component in the primary melanoma, tumour cell type ovoid, elongated or spindle (cf. round), and higher mitotic rate of the preceding primary tumour primary, absence of BRAF mutation or absence of NRAS mutation, and, absence of naevus in association with the primary tumour (Table 3.1). In most cases the sign of the mean coefficient value was as expected and in some cases is supported by prior literature, *e.g.*, lower stage (Mann *et al.*, 2013; Balch

Table 3.1: **Final model from mBMI.** Coefficients for top-ranking clinical variables chosen as the final model via the mBMI framework.

	Variable	Mean Coefficient	Standard Error
1	Tum_NRASmut (Yes/No)	-5.18	2.60
2	Prim_Site_SunExp (Intermittent/Chronic)	4.30	2.48
3	Prim_Stage (Stage II/Stage I)	-8.70	4.27
4	Prim_Stage III	-7.04	3.90
5	Age_Analysis	-0.17	0.09
6	Tum_MetSize	-0.05	0.04
7	NM_variable (1/0)	3.80	2.37
8	Prim_Mitos	0.32	0.18
9	Tum_CellType (1/0)	2.76	1.93
10	Tum_BRAFmut (Yes/No)	-2.11	1.83
11	Prim_Naevus (Present/Absent)	-1.78	2.44

et al., 2009), cell shape not round (Shaw *et al.*, 2006), smaller sized regional node metastases (Balch *et al.*, 2009), absence of naevus in association with the primary melanoma (Kakavand *et al.*, 2014), and absence of BRAF or absences NRAS mutation (Mann *et al.*, 2013). More details are shown in (Jayawardana *et al.*, 2015a).

3.2.2 *Mann data: Other modelling procedures*

The modelling methods compared to mBMI are briefly outlined below. Throughout the comparison study $m = 25$ multiple imputations were used with $\tau_{MI} = 0.50$ for model aggregation.

1. Multiple imputation and logistic regression.
   - Stepwise model selection:
     Variable selection was carried out using a stepwise selection procedure using BIC as the selection criterion. A final logistic regression model is fitted to the data using the selected variables.
   - Lasso and Elastic net model selection:
     For Lasso based model selection the R package 'glmnet' was used (Friedman

et al., 2010). Here a generalised linear model is fitted via penalised maximum likelihood. The objective function for the penalised logistic regression uses the negative binomial log-likelihood, and is minimised:

$$min_{\beta \in \mathbb{R}^{p+1}} \left(-\left[\frac{1}{n} \sum_{j=1}^{n} y_j(\mathbf{x}_j^T \beta) - \log(1 + e^{\mathbf{x}_j^T \beta}) \right] + \lambda \left[\left(\frac{1-\alpha}{2} \right) ||\beta||_2^2 + \alpha ||\beta||_1 \right] \right).$$

The regularisation path is computed for the Elastic net penalty $((1-\alpha)/2||\beta||_2^2 + \alpha||\beta||_1)$ at a grid of values for the regularisation parameter λ. The Elastic net mixing parameter α is between 0 and 1, and when $\alpha = 1$ the Elastic net penalty is the Lasso penalty and when $\alpha = 0$ it is the Ridge penalty.

We use the misclassification error as the criterion to select λ, and $\alpha = 1$ (Lasso model selection) and $\alpha = 0.5$ (Elastic net model selection).

2. Random Forest:

   Variable selection was carried out using the package 'varSelRF' (Diaz-Uriarte, 2010). It uses the Out-Of-Bag (OOB) error as minimisation criterion and eliminates the least important variables successively using variable importance scores from Random Forest.

3. Clustering:

   Hierarchical clustering was performed on the variables using the R package 'ClustOfVar' (Chavent *et al.*, 2013). One imputed dataset was used to select the number of clusters and then it is applied over the m datasets. The CV error rate for inclusion of each variable in the cluster was computed after the optimum number of clusters was determined. The variable that results in the lowest error rate is selected from each cluster to be included in the final model.

4. Principal Components Analysis:

   For each imputed dataset, the principal components (PCs) were retained such that the cumulative proportion of variation explained by the PCs was at least 75%. CV was carried out on this new dataset of PC scores.

5. B-MI:

   The B-MI (Campain, 2012) was used to select the variables.

6. Weighted BMI:

   The B-MI procedure was modified such that a rank is associated with each model. The modification is as follows;

   - In Stage 6 of the mBMI framework, the CV error of all of the B models are calculated.
   - Rank the B models based on their CV error such that the highest CV error model has rank 1 and we denote the rank from the bootstrap model b by r_b.
   - Give a score to all the variables based on these ranks.

     For the q^{th} variable,

     $$c_q = \sum_{b=1}^{B} \frac{r_b \times 1\{\bar{\beta}_{q,b} \neq 0\}}{B(B+1)/2}.$$
   - Rank the variables based on this score (higher score for variables included in models with low CV error).
   - The ranked variables are then included in a logistic regression model in a forward algorithm, where in the first model the highest ranked variable is included, and in the second model the top two highest ranked variables are included and so on, until the full set of variables are included into the model. The final set of variables is the set that gives the lowest CV error.

Results: Comparison of methods on Mann data. The CV error rates of the final models were compared and the results are summarised in Table 3.2. The mBMI method gave the lowest prediction error rate with a parsimonious model (24% with 10 variables). A comparative error rate could be achieved by weighted BMI with Lasso selection (26%) with 11 variables. All the other methods had a CV error between 34% to 50%. LOOCV error rates were used in this comparison study as required by the biologists for the error rates to be comparable with a previous study (Mann *et al.*, 2013) using the Mann data.

Table 3.2: **Comparison of modelling procedures for clinical data.** The number of variables selected are given in square brackets

Method	**Variable selection**	**Prediction error rate**
Logistic Regression	Stepwise selection	38% [12]
	Lasso	39% [19]
	Elastic net	42% [18]
	Ridge	41% [21-all]
Random forest		45%
Clustering		35% [7]
Principal Components Regression		50%
BMI	Elastic net (best)	35%[13]
Weighted BMI	Stepwise	34% [6]
Weighted BMI	Lasso	26% [11]
mBMI	Stepwise	24% [10]

3.3 EVALUATION OF THE MBMI FRAMEWORK: SIMULATION STUDY

To explore additional features of the mBMI framework, a simulation study was carried out where optimal parameters are investigated for the mBMI framework. Through this simulation study, we attempt to give a clearer picture of the utility of the mBMI framework, where the parameters within the mBMI could be optimised according to the dataset used. Simulating data for a logistic regression is often more challenging than for a linear regression model, hence a much larger dataset is used than the Mann data with a simpler data structure. A total of 20 explanatory variables were generated independently from a uniform distribution (U(-1,1)) and standardised to have zero mean and unit variance. The response variable was randomly generated from the binomial distribution, where the probability of success vector (π) was obtained using the pre-determined parameter vector (β) and simulated regression data ($\mathbf{X}$),

$$\pi_j = \frac{1}{1+e^{-x_j^T \beta}}.$$

Since the outcome classes were relatively balanced in the Mann data, we attempted to have similar class structure in the simulated data. The simulated response variable

Table 3.3: **The regression coefficients for the simulated variables.** The true parameter vector (β), and the 90% confidence intervals of the parameter estimates using a logistic regression model are given in the table. The confidence intervals in bold contain zero.

Variable	**True beta**	90% Confidence interval	
		5%	95%
A*	1.0	0.65	1.06
B*	1.0	0.29	0.68
C*	1.0	0.37	0.76
D*	1.0	0.36	0.76
E*	1.0	0.34	0.73
F*	-1.0	-0.52	-0.14
G*	-1.0	-0.69	-0.30
H*	-1.0	-0.51	-0.13
I*	-1.0	-0.75	-0.37
J*	-1.0	-0.76	-0.37
K	0.0	**-0.34**	**0.04**
L	0.0	**-0.21**	**0.17**
M	0.0	**-0.25**	**0.12**
N	0.0	**-0.33**	**0.04**
O	0.0	**-0.28**	**0.10**
P	0.0	-0.42	-0.05
Q	0.0	**-0.12**	**0.26**
R	0.0	**-0.35**	**0.02**
S	0.0	0.01	0.38
T	0.0	**-0.29**	**0.09**

had 259 one's (*i.e.* good prognosis) and 241 zero's (*i.e.* poor prognosis), where the total sample size was 500.

A logistic regression model is applied to the simulated data. Table 3.3 shows the 90% confidence interval obtained for the parameter estimates. It is apparent from the results that model instability is present even when we use the complete dataset, as two of the variables with zero coefficients produce confidence intervals without including zero.

Selection of the parameters

Within the mBMI, we can vary the following parameters as explained in Section 3.1;

- Prediction error threshold (τ_{CV}).

  In the mBMI after ranking the B models by their prediction error rate, the best performing models are selected (Stage 6 of the mBMI) using the prediction error threshold. That is, the models showing the highest prediction accuracy are selected for the next steps of the analysis. In this simulation study, τ_{CV} was varied to consider top 2.5%, 5%, 10-50% models ($\tau_{CV} = 0.025, 0.05, 0.10, 0.20, \dots, 0.50$).

- Final inclusion frequency threshold (τ_{FM}).

  The inclusion frequency of the variables in the best performing models are considered (Stage 8 of the mBMI) in selecting the final model. We varied τ_{FM} from 0 to 1. That is if $\tau_{FM} = 0$, then all the variables that are included in at least one of the models are used in the final model. If $\tau_{FM} = 1$, then the variables that are included in all of the models are used in the final model.

Results: The results are summarised in Figure 3.3 and 3.4. The patterns of prediction error rates are similar for all τ_{CV} (see Appendix B). To select the τ_{FM}, the inclusion frequencies of variables for all τ_{CV} are considered. One instance is illustrated in Figure 3.3, where the prediction error threshold, τ_{CV} is 0.025 (*i.e.,* top 2.5% models). The separation between the variables in the true model and the variables that are not in the true model can be clearly seen after $\tau_{FM} = 0.65$. Therefore in this instance the final inclusion frequency threshold was selected to be 0.65. Similarly the best τ_{FM} was selected for each τ_{CV} (see Appendix B). To select the best τ_{CV} and τ_{FM} combination the prediction error rates of the models with selected τ_{FM} were considered (Figure 3.4). The error rate from $\tau_{CV} = 0.025$ was slightly lower than the others. Therefore, from this simulation study, the parameters selected were,

- $\tau_{CV} = 0.025$ (prediction error threshold: consider the top 2.5% models).

- $\tau_{FM} = 0.65$ (final inclusion frequency threshold: the variables that are included in at least 65% of the top models).

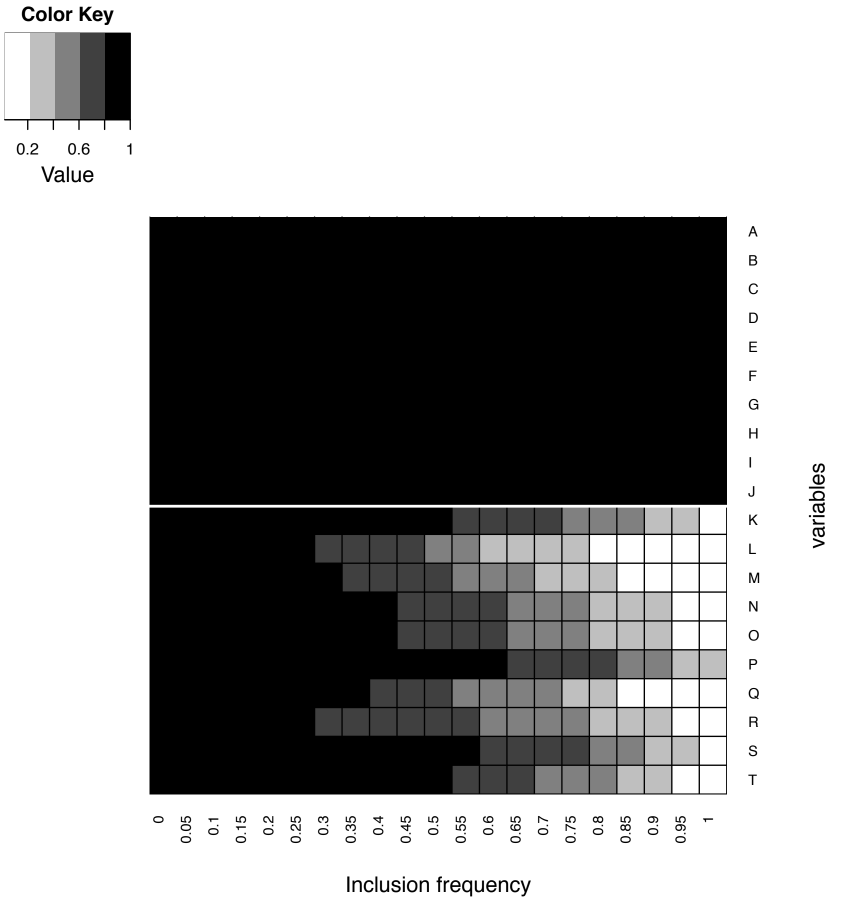


Figure 3.3: **Stability of the variables.** The stability measures of the variables when $\tau_{CV} = 0.025$. The rows represent the variables and the columns represent different τ_{FM} values. A variable is considered to be stable if they are included in the majority of the fitted models. In this figure the color indicates the stability, where darkness is proportional to the inclusion frequency. The scale varies from 0 (the variable is not included in any of the fitted models) to 1 (the variable is included in all of the fitted models, hence highly stable).

3.4 discussion and conclusions

Missing data, instability of final models and predictive modelling are some of the key challenges when model building for clinical data. This chapter focused on developing a framework that addresses all of the above mentioned challenges via the mBMI framework.

To construct a stable model for the purpose of achieving high prediction accuracy, the stabilising steps in the mBMI were proposed. For a model to be stable, the same set of variables should be consistently chosen even with perturbations in the data.

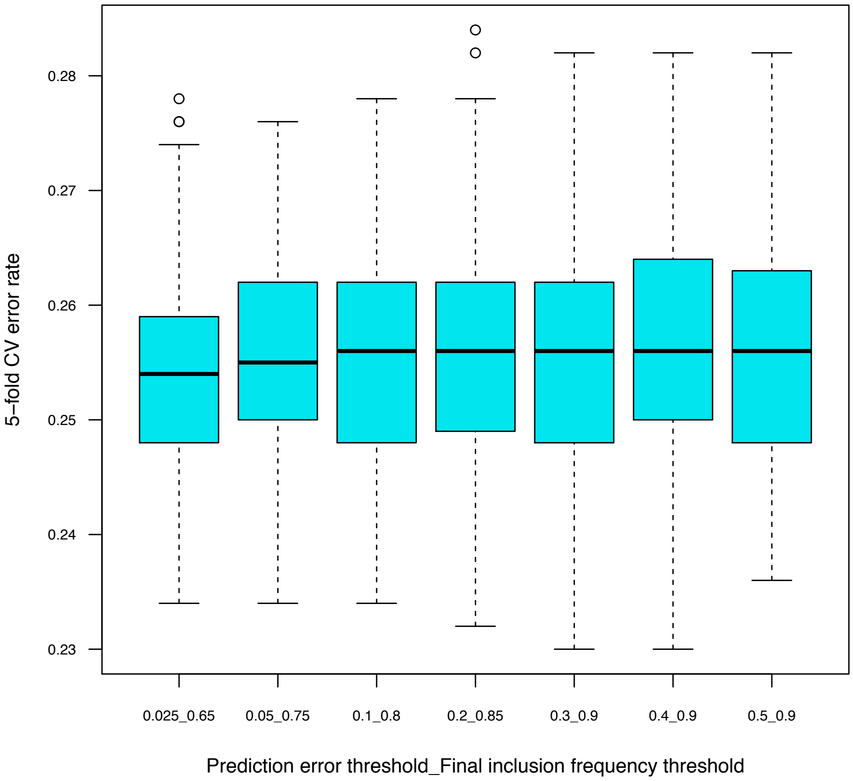


Figure 3.4: **The prediction error rates of the selected models.** The 5-fold CV error rate for τ_{FM} selected for each τ_{CV}.

These perturbations in the data were created by taking bootstrap samples and selecting variables over the many re-samples of the data. Our framework facilitated the selection of best performing subsets of variables, focussing on those models with small CV error rate. To ensure that the final model is stable, the variables that had an inclusion frequency above a pre-specified threshold were considered.

The tuning of the parameters, prediction error threshold and final inclusion frequency threshold can be regarded as data dependent steps. The simulation of data that clearly demonstrates the features of the mBMI procedure proved to be rather challenging, however the main features of the mBMI have been demonstrated via the simulation study conducted in this chapter. Due to the complexity of the Mann data, simulating data that possess similar characteristics was much harder. Furthermore, with smaller sample sizes similar to that of the Mann data, it was rather difficult to illustrate the components of the mBMI. Therefore, although such datasets were simulated and explored using the mBMI, to give a clearer picture, a much simpler simulation study was presented in Section 3.3. Choosing $\tau_{CV} = 0.025$ and $\tau_{FM} = 0.65$ proved to produce better results for this particular simulation set-up, also because of its ability to distinguish between the true variables and other variables clearly as the stability of the true variables were significantly higher. These parameters also produced comparable CV error rates for the final models constructed. No significant distinction between the CV error rates were noticed, which might be due to the simple structure of the simulation set-up.

The application of the mBMI in the Mann data enabled us to assess its validity in real clinical data. The results produced an interpretable model in the clinical context with a good prediction error rate. A comparison study of the mBMI method with a collection of other methods popular in clinical data modelling was carried out using the Mann data. Stability could not be assessed due to the data being a real dataset. We conjecture that the stability measures in the mBMI produced a final model that is stable as well. The CV error rates of the final model indicate that the mBMI framework produced the best model in terms of its prediction capability.

Overall, the mBMI framework can be regarded as a very useful and effective method that addresses missing data, unstable models and predictive model construction. Our parrallel implementation of the mBMI in a multi-core architecture made the exploration of the features within the mBMI much easier and less time consuming.

4

HORIZONTAL DATA INTEGRATION

Biomarker discovery and the evaluation of its accuracy in a prognostic setting has been one of the key questions in the field of bioinformatics over the years. Traditionally, a single data source, either clinical or high-throughput omics data, has been used to address this. The availability of multiple datasets through public repositories such as GEO (Barrett *et al.*, 2007)[1], ArrayExpress (Parkinson *et al.*, 2007)[2] and TCGA[3] provides unprecedented opportunities: to use the many available data sources effectively in the biomarker discovery process to identify more effective biomarkers, to improve the power of current studies via the availability of more data samples and/or variables, and to validate the current biomarkers. One such field that emerged utilising the availability of multiple data sources is 'horizontal data integration', where multiple datasets of the same type are used in an integrative analysis (Tseng *et al.*, 2012). As discussed in Chapter 1, horizontal integration can be broadly categorised into meta-analysis; the integration of statistics from different studies, and mega-analysis; combining the datasets prior to the analysis. In this chapter, we examine methods and approaches in horizontal data integration in a meta-analysis setting, using multiple datasets of the same type (*e.g.,* same disease).

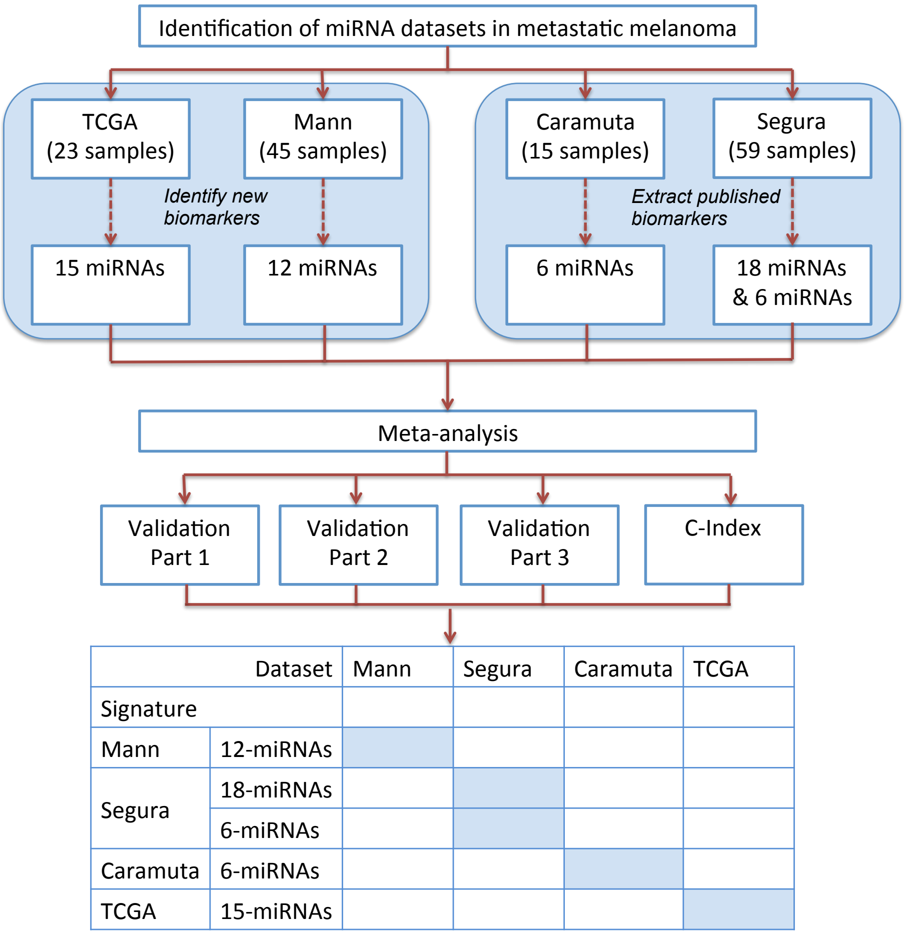
Classically, meta-analysis has been used to combine datasets to provide more power to the analysis. This is done by combining statistics to detect a common set of important features and effect sizes, where usually all the studies share a common null hypothesis. In contrast, we use the meta-analysis in a prognostic context in this chapter, to better estimate the prediction accuracy of biomarkers as well as to identify a robust set of biomarkers from multiple studies. In biomedical applications this class of meta-analysis often includes studies with different aims and designs. However, integrating data in such a meta-analysis is still imperative to gain new information. This meta-analysis allows to evaluate the biomarkers using independent data, in the light of increasing the confidence on these biomarkers in clinical use. The conclusions are more robust as they are based on multiple studies rather than a single data source. In Chapter 1 we discussed a number of challenges and issues associated with meta-analysis. Two of the challenges encountered are: (i) the differences in study aims and designs that

[1] Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo).
[2] http://www.ebi.ac.uk/arrayexpress/
[3] The cancer genome atlas (http://cancergenome.nih.gov).

hinder the comparison of different biomarkers, as all variables being considered are not assayed in each of the multiple studies, and (ii) the heterogeneous patient cohorts in each study.

In this chapter, we use the miRNA datasets from multiple studies (discussed in Section 2.2) in developing a comprehensive meta-analysis procedure, while addressing the above challenges. The miRNA biomarkers are of particular interest here, as they are a new type of biomarkers that emerged recently. In melanoma, there are only a handful of studies (Segura *et al.*, 2010; Caramuta *et al.*, 2010) that investigate this class of markers. As such, these datasets highlight the challenges associated with horizontal data integration and lend themselves to illustrate the advantages in meta-analysis to comprehensively evaluate the efficacy of signatures in a biologically relevant context.

Section 4.1 describes pre-processing of the datasets, as careful analysis at this stage is essential to obtain optimal signal from each data source. In Section 4.2, novel biomarkers are identified for the studies where no published prognostic biomarkers are available. In Section 4.3 we utilise all tissue-based miRNA prognostic signatures in metastatic melanomas proposed in the literature to date (Jayawardana *et al.*, 2015b; Segura *et al.*, 2010; Caramuta *et al.*, 2010) in a systematic cross-validation (CV) of the biomarkers. The aim of the above analysis is to assess the prognostic utility of the biomarkers in independent datasets, thereby highlighting a robust set of biomarkers. Furthermore, the C-index (Uno *et al.*, 2011) will be used as an alternative method to the CV error rate to evaluate the signatures. The use of the C-index allows to assess the signatures for the actual event times (survival times), rather than for the prediction of binary outcome (GP *vs.* PP) as in CV. The signatures are then evaluated relative to equivalent but random variable sets to assess whether they have better predictive power. In Section 4.4, a robust and reproducible set of miRNAs are evaluated in the validation datasets. The meta-analysis procedure explored in this chapter (Figure 4.1) yields insights into the translational potential of this class of markers.



	Dataset	Mann	Segura	Caramuta	TCGA
Signature					
Mann	12-miRNAs	▨			
Segura	18-miRNAs		▨		
	6-miRNAs		▨		
Caramuta	6-miRNAs			▨	
TCGA	15-miRNAs				▨

Figure 4.1: **Schematic of the meta-analysis procedure.** The figure outlines our comprehensive meta-analysis procedure. The shaded cells in the table denote the instances where the signatures are validated on their own training data.

4.1 Pre-processing and normalisation of the datasets

In this chapter, we use four miRNA datasets for meta-analysis. The first dataset is the Mann miRNA data (Section 2.1), for simplicity we refer to it as ‘Mann data’ in this chapter. Apart from this in-house miRNA data, three other external miRNA datasets are used in the analysis. These are the datasets from TCGA, the Segura data (Segura *et al.*, 2010) and the Caramuta data (Caramuta *et al.*, 2010). The latter two were identified after a comprehensive literature search for miRNA microarray expression profiling studies in clinical samples of metastatic melanoma. The details of the Segura and Caramuta datasets, pre-processing and low-level analysis were detailed in Section 2.2. The pre-processing and the normalisation of the Mann data, which includes a comprehensive comparison of normalisation procedures, and the pre-processing of TCGA data are detailed in this section. This low-level analysis is imperative in removing unwanted noise in the data to assist in downstream analysis.

The Mann data

This data is used as the primary miRNA dataset and we use this to develop a pre-processing approach for miRNA data, which includes a comparison study to determine the optimal normalisation method. Because of the comprehensive nature of the pre-processing involved, the details are given in this chapter and were not shown in Chapter 2. The pipeline developed here involves careful pre-processing, low-level analysis and normalisation required to ensure a more accurate and clean dataset for further analysis (Figure 4.2). This procedure can be adopted for any similar miRNA microarray dataset.

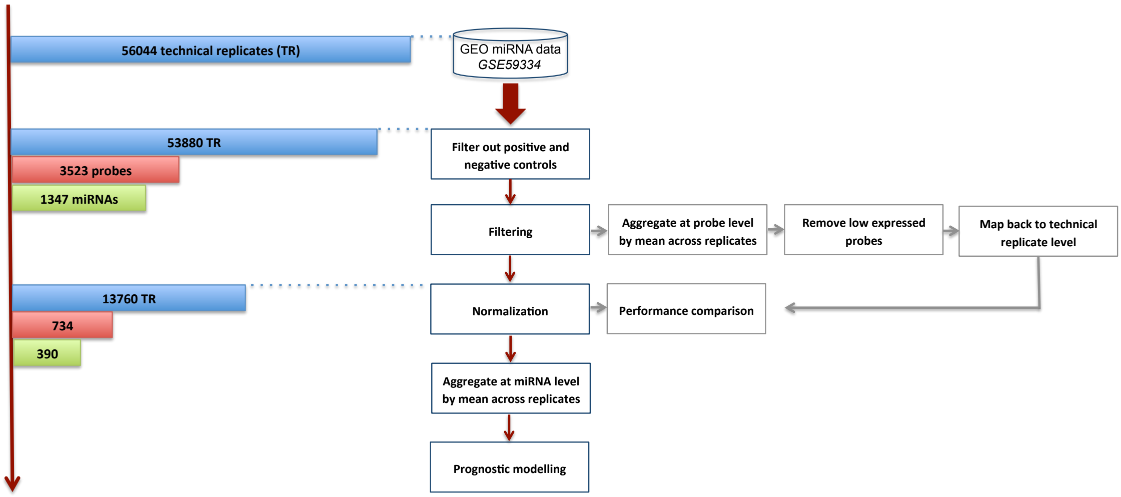

56044 technical replicates (TR)
GEO miRNA data
GSE59334
53880 TR
3523 probes
1347 miRNAs
Filter out positive and negative controls
Filtering
Aggregate at probe level by mean across replicates
Remove low expressed probes
Map back to technical replicate level
13760 TR
734
390
Normalization
Performance comparison
Aggregate at miRNA level by mean across replicates
Prognostic modelling

Figure 4.2: **Schematic of the processing and analysis of the Mann data.** On the left hand side of the plot the numbers of miRNAs (probes and technical replicates) at various stages of the process are shown.

This dataset can be downloaded from Gene Expression Omnibus (GEO, GSE59334; 3523 probes and 1347 miRNAs) (Tembe *et al.*, 2014). Actual miRNAs were selected, where all positive and negative controls were filtered out. The data contained 1347 miRNAs with each miRNA consisting of multiple probes (1-4), and each probe consisting of multiple technical replicates (10-40). A filtering protocol comprising of three key steps was applied:

- For each probe, the average expression value over all technical replicates were taken.
- The probes having a maximum expression value greater than or equal to 7.5 across all the samples (patients) were selected (the probes having low expression with maximum expression level across samples < 7.5 were filtered out). This process resulted in a filtered dataset of 734 probes pertaining to 390 miRNAs (Figure 4.2).
- The filtered data were mapped to the level of technical replicates (13,760 entries).

Typically miRNA data contain much less signal than mRNA data. Therefore, it is critical to apply a filtering protocol as described to enable successful downstream analysis. Such a filtering was applied with the intention of improving upon the signal in the data, as the majority of the miRNAs exhibited low expression values relative to others (Figure 4.3). The boxplots of the data before and after filtering (Figure 4.4) show the effect of filtering. It is apparent that the data before filtering has much lower expression at the sample level (45 samples) than after filtering. Also, the batch effect is much more evident in the data before filtering (Figure 4.4a), and this has much less impact in the filtered data (Figure 4.4b) as the samples are on the same scale with similar variability.

Comparison of normalisation procedures. The normalisation of data is an imperative step in assuring stronger signal in the data. It aims to remove batch effects and unwanted noise or variation from external factors, both technical and biological. In this section, we perform a comparison study of different normalisation procedures to select the most appropriate normalisation procedure for this data. There exist comparison studies in the normalisation procedures literature, for example for miRNA microarray data (Rao *et al.*, 2008; Hua *et al.*, 2008) and for miRNA-Seq data (Garmire and Subramaniam, 2012). However, none of these studies compared the effect of utilising a more recently developed method of removing unwanted variation, RUV (Gagnon-Bartsch and Speed, 2012) in normalisation. In this thesis we use RUV in combination with the more traditionally used quantile normalisation, which is considered as a more general ‘global adjustment’ method (Gagnon-Bartsch and Speed, 2012). Evaluation strategies similar

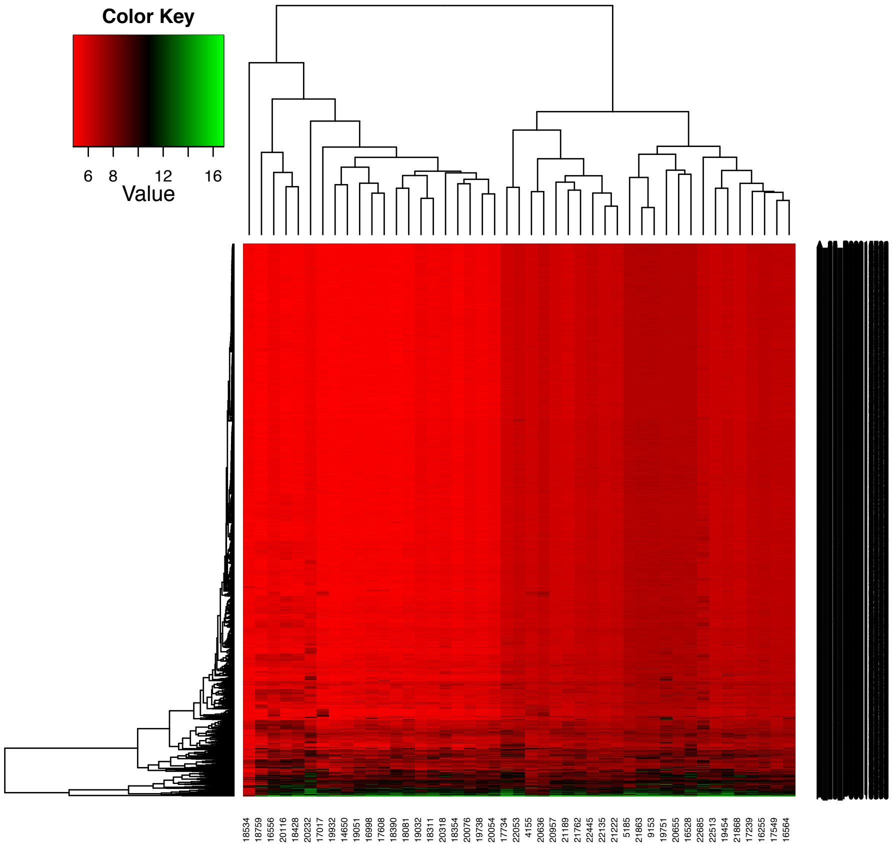


Figure 4.3: **Heatmap of the probe-level expression values in the Mann data.** Many probes had very low expression values and as a result are not likely to add value to any subsequent analysis; rather, they would be expected to add noise to the data.

to those described in (Gagnon-Bartsch and Speed, 2012) were used for this comparison study as follows:

1. Number of differentially expressed miRNAs

   The number of differentially expressed (DE) miRNAs discovered are expected to increase after proper normalisation, as it is an indication of the increase in distinction between the groups (*e.g.* survival groups; GP and PP) being compared. We used a linear mixed effects model to find the number of DE miRNAs, owing to the structure of the data (every miRNA in this dataset corresponded to one or more probes and each probe had one or more technical replicates). This approach takes into account the variability between conditions (e.g. good vs. poor

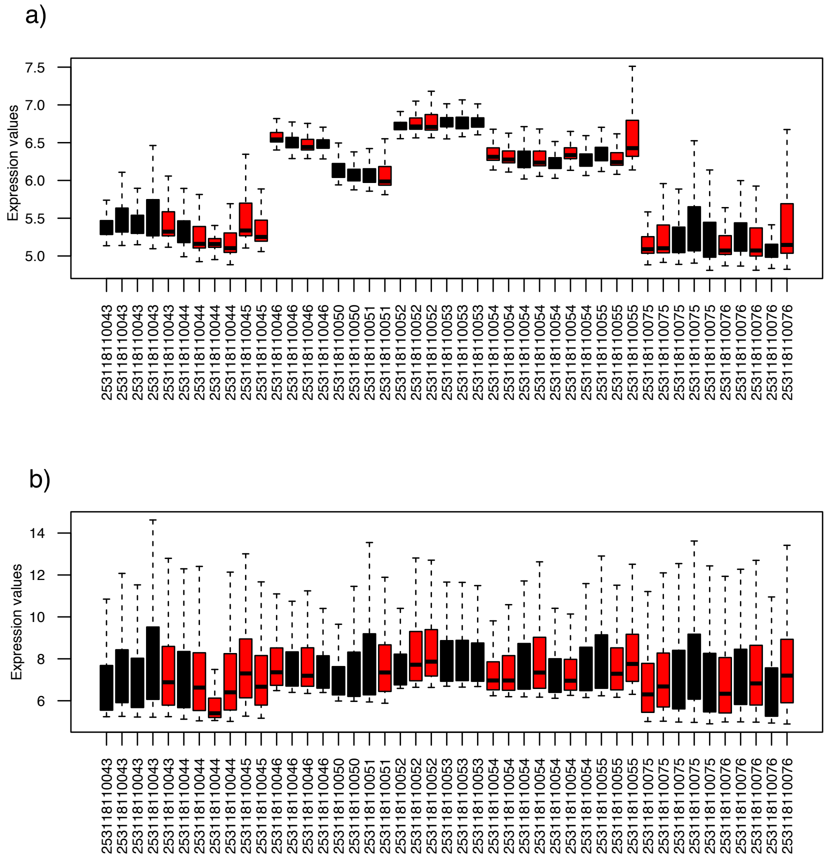


Figure 4.4: **Boxplots for the expression values of the samples in the Mann data.** a) before filtering the data, b) after filtering the data. x-axis shows the 45 samples ordered by the batch number.

prognosis), as well as within probes and samples (more details are given in Appendix C). The model used was

$$y_{ijkl} = \mu + \alpha_i + \eta_k + \gamma_{ijk} + \epsilon_{ijkl}.$$

Here, y_{ijkl} denotes the normalised $\log_2$ expression for the i^{th} condition, j^{th} sample, k^{th} probe and l^{th} replicate, μ denotes the baseline expression, α_i denotes the effect of the i^{th} condition, η_k denotes the effect of the k^{th} probe, γ_{ijk} denotes the variability in the i^{th} condition for the j^{th} sample and the k^{th} probe, and ϵ_{ijkl} denotes the measurement error. The fixed effect coefficients of the model are μ and the α's, and the random terms are the η's, γ's and ϵ's. It is assumed that all random terms are Gaussian as follows: $\eta_k \sim N(0, \sigma^2_\eta)$ independent of $\gamma_{ijk} \sim N(0, \sigma^2_\gamma)$ independent of $\epsilon_{ijkl} \sim N(0, \sigma^2_\epsilon)$. The R package 'nlme' (Pinheiro *et al.*, 2014) is used to fit this mixed model (using the function 'lme'). A condition refers to any comparison of interest. For the purpose of comparing normalisation methods here: 1) the survival groups (good vs. poor prognosis); and, 2) BRAF mutation status (BRAF V600E/K mutant vs. wild type) are considered. Candidate DE miRNAs are defined by controlling for 5% false discovery rate (FDR).

2. Distribution of p-values

   The distribution of p-values from a DE analysis, when shown through a histogram, is expected to be almost uniform over the interval (0,1) with a spike near zero (Gagnon-Bartsch and Speed, 2012; Leek and Storey, 2008, 2007). This is because by definition, the non-DE features (miRNAs here) are uniformly distributed over the unit interval, as it is assumed that only a fraction of features are associated with the comparison of interest (condition). The miRNAs associated with the comparison of interest will ideally have approximately zero p-values. A dataset that has been processed well, removing unwanted variations, is expected to exhibit this ideal situation.

3. Correlation with qRT-PCR data

   As explained in Chapter 1, qRT-PCR data are considered as the gold standard for measuring expression and hence are frequently used for validation purposes (Hua *et al.*, 2008). For this comparison, the qRT-PCR data for a subset of ten miRNAs described in (Tembe *et al.*, 2014), were obtained and normalised via RUV. This data is then compared with the same subset of ten miRNAs in the Mann data normalised via different methods using Pearson's correlation coefficients.

Different combinations of quantile normalisation (implemented in 'limma' package (Smyth, 2005) and a modified version of RUV (Removing Unwanted Variation (Gagnon-Bartsch and Speed, 2012); 'randomRUV' with nuCoeff being a regularisation factor and k being the number of unobserved factors (for more details we refer to Appendix C), were compared with different parameters ($k = 2, 10$ and $\mathtt{nuCoeff} = 0.00001, 0.001, 0.1$), before and after filtering the data (filtering explained above). The following normalisations were compared in this section:

1. Q (Quantile normalisation then filtering)

2. subQ (Quantile normalisation after filtering)
3. RUV (RUV then filtering; $k = 10, nuCoeff = 0.001$)
4. subRUV (RUV after filtering; $k = 10, nuCoeff = 0.001$)
5. subRUV1 (RUV after filtering; $k = 10, nuCoeff = 0.1$)
6. subRUV5 (RUV after filtering; $k = 10, nuCoeff = 0.00001$)
7. QRUV (Quantile normalisation then RUV before filtering; $k = 2, nuCoeff = 0.001$)
8. subQRUV (Quantile normalisation then RUV after filtering; $k = 2, nuCoeff = 0.001$)
9. subQRUV1 (Quantile normalisation then RUV after filtering; $k = 10, nuCoeff = 0.1$)
10. subQRUV5 (Quantile normalisation then RUV after filtering; $k = 10, nuCoeff = 0.00001$).

Results: Filtered data that were quantile normalised followed by RUV under different parameters showed the best performance overall. Specifically, the subQRUV approach yielded the highest number of DE miRNAs for survival group distinction (good and poor prognosis) and a comparably higher number of DE miRNAs for BRAF mutation (mutant and wild type) (Table 4.1). The p-value histograms under both conditions showed the best distribution (near uniform, with peak at zero) under the subQRUV normalisation protocol (Figure 4.5). The comparison with the qRT-PCR data using the correlation coefficients showed high consistency with the QRUV-normalised data, where the highest absolute value of the correlation coefficient was exhibited for the majority of the miRNAs (Table C.1). Therefore, the subQRUV normalised data (data filtered then normalised by quantile normalisation and RUV with nuCoeff=0.001 and k=2) was used for the subsequent analysis. Finally, the data were aggregated at the miRNA level by taking the average across all probes and technical replicates for each miRNA (390 miRNAs) (Figure 4.2).

Table 4.1: **Number of differentially expressed miRNAs from each normalisation method.** The number of DE miRNAs for the separation of two survival groups (good vs. poor prognosis) and for BRAF mutation (mutant vs. wild type), for some of the normalisation techniques compared in this study. Option 4 was ultimately used since it gave the highest number of DE miRNAs for survival groups, as well as comparably higher number of DE miRNAs for BRAF mutation.

	Method	Differentially expressed miRNAs for survival groups	Differentially expressed miRNAs for *BRAF* mutation
1	Quantile → filter (Q)	7	5
2	Filter → Quantile (subQ)	7	0
3	Filter → RUV (subRUV)	3	19
4	Filter → Quantile → RUV (subQRUV)	10	19
5	Filter → Quantile → RUV (nuCoeff=0.1) (subQRUV1)	7	4
6	Filter → Quantile → RUV (nuCoeff=0.0001) (subQRUV5)	0	37

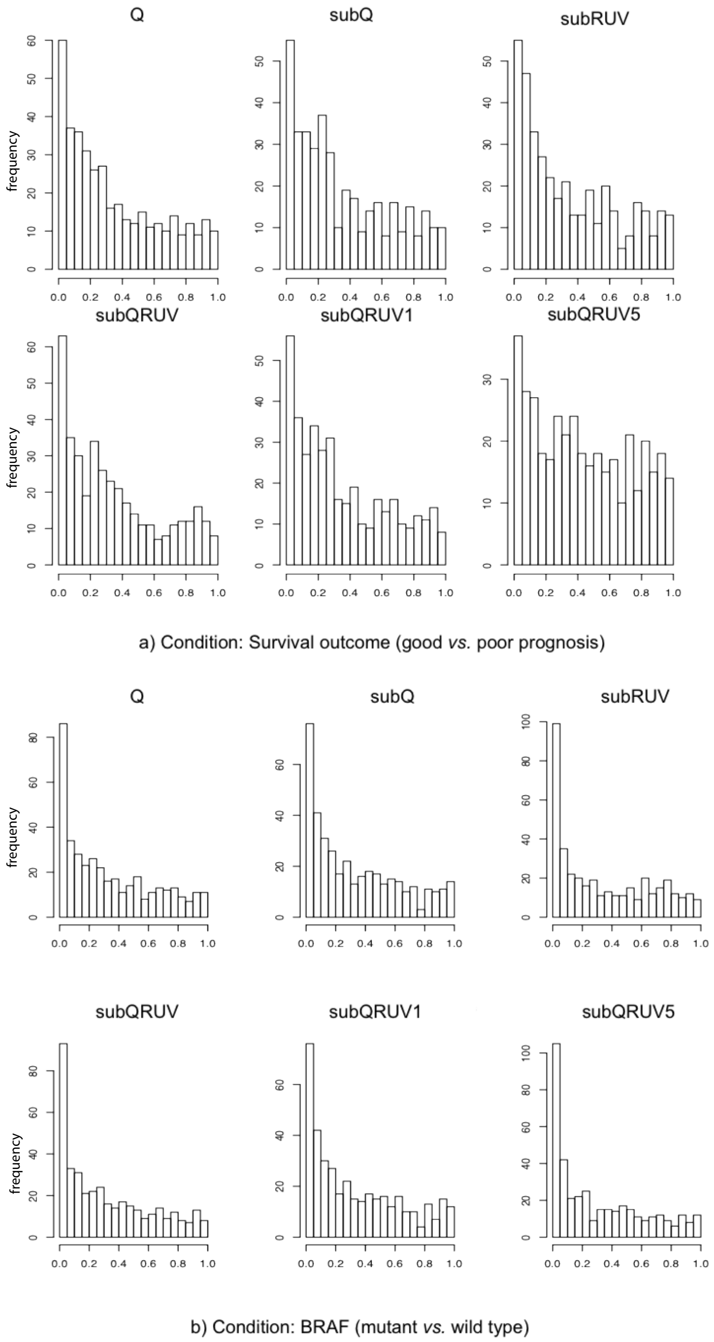
TCGA miRNA data

In the miRNA expression dataset downloaded from TCGA, there were miRNA-Seq counts for 1134 miRNA regions that were annotated as 'mature', 'star', 'stemloop', 'unannotated' and 'precursor', along with their respective accession number (*e.g.* mature, MIMAT0000062). The following filtering steps were applied:

- For direct comparability with the remaining microarray-based expression datasets evaluated in this study, all sequences other than those annotated as 'mature' or 'star' regions were excluded.

- The read counts for each miRNA (*e.g.*, the sum of the read counts for all sequences under the unique identifier 'hsa-let-7a-1_MIMAT0000062') were aggregated, this produced 775 miRNAs.

- The data for the 'expressed' miRNAs were filtered, *i.e.* all miRNAs with row sum of counts equal to zero across the samples were excluded, resulting in 774 miRNAs (Figure 4.6).



a) Condition: Survival outcome (good *vs.* poor prognosis)

b) Condition: BRAF (mutant *vs.* wild type)

Figure 4.5: **Histograms of p-values from analysis of differential expression (DE).** The histograms of p-values for the DE analysis on the comparison of interests (conditions) show, a) the survival group comparison between good and poor prognosis and b) the mutation status comparison between BRAF mutant and wild type.

This filtered dataset was normalised using TMM, the trimmed mean of M-values method (Robinson and Oshlack, 2010), to account for differences in library sizes. Data were then transformed using 'variance stabilised transformation' in the R package 'DESeq' (Anders and Huber, 2010) (Figure 4.6).

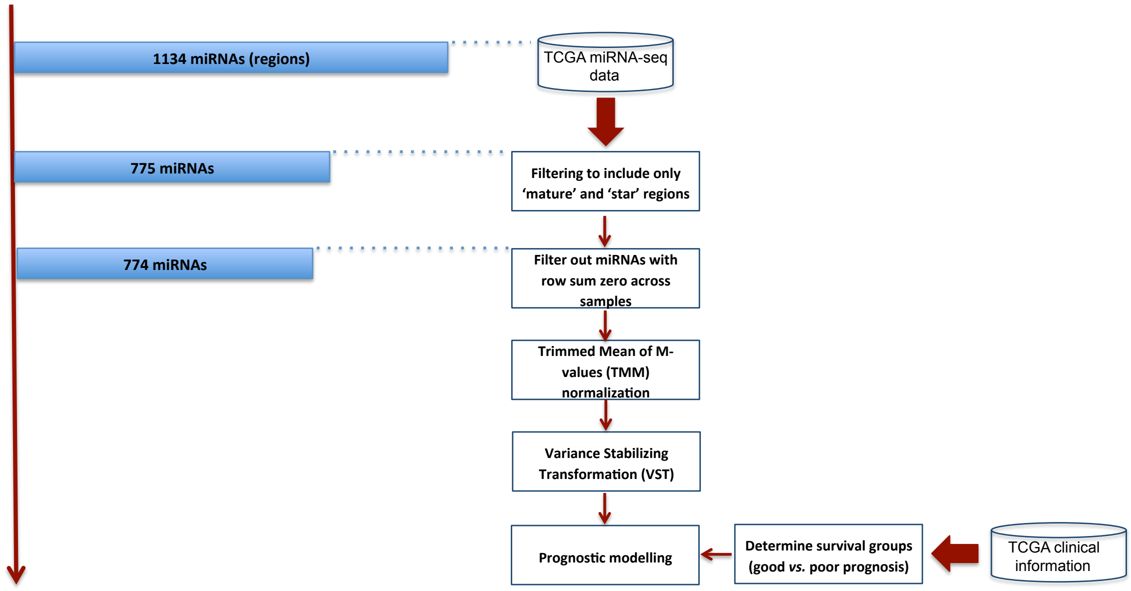

1134 miRNAs (regions)
TCGA miRNA-seq data
775 miRNAs
Filtering to include only 'mature' and 'star' regions
774 miRNAs
Filter out miRNAs with row sum zero across samples
Trimmed Mean of M-values (TMM) normalization
Variance Stabilizing Transformation (VST)
Prognostic modelling
Determine survival groups (good *vs.* poor prognosis)
TCGA clinical information

Figure 4.6: **miRNA-Seq data analysis pipeline for data from TCGA.** On the left hand side of the plot the numbers of miRNAs at various stages of the process are shown.

4.2 IDENTIFICATION OF NEW MIRNA-BASED PROGNOSTIC SIGNATURES

From the broadly identified miRNA datasets in melanoma (Mann, TCGA, Segura, Caramuta), only (Segura *et al.*, 2010) and (Caramuta *et al.*, 2010) contain signatures. Therefore, we concentrated on identifying prognostic signatures for the Mann and TCGA data as part of our meta-analysis procedure. For this purpose, the CV procedure 'A: Full CV' outlined in Section 2.4.1 was used with the following parameters:

- Feature selection method: 'Median robust' (Jayawardana *et al.*, 2015a; Campain, 2012).

  In the median robust method, the miRNAs are ranked based on the difference between the median expression of the two groups (PP and GP). A set of molecular signatures are devised such that the kth molecular signature contains the k top-ranked features. The final expression value for miRNA q is

$$e_q = \tilde{x}_{q,GP} - \tilde{x}_{q,PP},$$

where $\tilde{x}_{q,GP}$ and $\tilde{x}_{q,PP}$ represent the group median for the q^{th} miRNA for GP and PP groups, for $q = 1, 2, \ldots, 390$.

- Classification method: Nearest shrunken centroids (NSC) (Tibshirani *et al.*, 2002) as implemented in 'pamr' R package (Hastie *et al.*, 2014).

This procedure was applied to the Mann and TCGA datasets. In the Mann data, the set of features evaluated (Step 2 of Full CV) was varied over top $5, 10, \ldots, 100$ and in TCGA over the top $10, 20, \ldots 500$, depending on the varying total number of features in each dataset (390 miRNAs and 774 miRNAs, respectively). The prognostic signature comprised of the set of miRNAs included in more than 50% of the models (500 models in total; 5-folds x 100 runs).

Results: The prediction error rates of the identified prognostic biomarkers/signatures were evaluated via the 'Final CV' procedure (CV procedure E, Section 2.4.5) using leave-one-out cross-validation (LOOCV) to be comparable with the other error rates. The signatures for the Mann and TCGA data comprised of 12 and 15 miRNAs, respectively (Table C.2) which produced LOOCV error rates of 33% and 39% (Table 4.2).

Abiding by the standards for validation of biomarkers, REMARK (REporting recommendations for tumour MARKer prognostic studies)-criteria (McShane *et al.*, 2005), the prediction accuracies of these biomarkers were compared with the prediction accuracy of the four most statistically significant clinico-pathologic prognostic parameters in Stage III melanomas. These standard-of-care parameters are the number of tumour-positive lymph nodes, tumour burden at the time of staging (microscopic v. macroscopic), presence or absence of primary tumour ulceration, and thickness of the primary melanoma (Balch *et al.*, 2009). These four variables were used in a logistic regression model to assess their prognostic power and the results are presented in Table 4.2. The results showed an improved predictive performance for the Mann signature (33% compared with 36% for standard-of-care variables). However, this improvement was not shown for TCGA signature (39% compared with 37%).

Finally, the performance after the integration of these signatures with the standard-of-care variables was evaluated. Here, the pre-validated vectors (Tibshirani and Efron,

2002) from each signature were derived and these were used together with the clinical variables in a logistic regression model (Jayawardana *et al.*, 2015a). The prediction error rates (Table 4.2) in this integrative framework were very similar (41% and 42%, respectively) for the two datasets, indicating that integration did not add value in this instance.

Table 4.2: **LOOCV error rates for the prognostic signatures identified using the Mann and TCGA data.**

	LOOCV error rate	**Method of assessment**
Mann		
12-miRNA prognostic signature	33%	NSC
Standard-of-care variables[1]	36%	Logistic regression
Standard-of-care variables[1] and 12-miRNA signature combined	41%	Pre-validated vector in a logistic regression framework
TCGA		
15-miRNA prognostic signature	39%	NSC
Standard-of-care variables[1]	37%	Logistic regression
Standard-of-care variables[1] and 15-miRNA signature combined	42%	Pre-validated vector in a logistic regression framework

[1] Number of tumour-positive lymph nodes, tumour burden at the time of staging (microscopic vs. macroscopic), presence or absence of primary tumour ulceration, and thickness of the primary melanoma (Balch *et al.*, 2009).

Abbreviations: NSC, nearest shrunken centroids; TCGA, The Cancer Genome Atlas.

4.3 cross-platform meta-analysis

We used two approaches to perform the meta-analysis in this chapter. The primary method used is CV, where the signatures are assessed for their ability to predict the survival groups (GP *vs.* PP) in every validation dataset. The C-index is also used as an alternative method to CV, because it allows us to compare the performance of the signatures in predicting the actual survival times. More details of the datasets and their signatures are given in Appendix C (Table C.2).

4.3.1 *Evaluation using the CV error*

Meta-analysis was perfomed via a systematic cross-validation to assess the signature performance in independent datasets. The five signatures (3 published and 2 identified in Section 4.2) from the Segura, Caramuta, Mann and TCGA data were used in this meta-analysis. Similar analysis has been performed in (Schramm *et al.*, 2012; Campain, 2012). This cross-validation meta-analysis is shown in Figure 4.1 and is as follows:

1. For each study a miRNA list (say signature X) was obtained which formed the feature vector.

2. For each of the four datasets (say dataset A), miRNA expression data were obtained and pre-processed as already described.

3. LOOCV was performed on the expression dataset A, where one sample was randomly assigned to the test set and the remaining samples were treated as the training set;

    - Using the expression dataset A and the miRNA list from signature X, a NSC classifier was constructed on the training set.

    - The classification rule constructed was then examined for its capacity to predict patient survival outcome (GP and PP) of the test set (dataset A).

    - The procedure was repeated for the assignment of each sample into a test set, which gave a complete predicted vector, thus producing a LOOCV prediction error rate for signature X on dataset A.

One of the main challenges in horizontal data integration is the differences in the study aims and designs. Our cross-validation meta-analysis also encounter this issue, which resulted in not all miRNAs of a given signature being present in the other independent datasets. The reasons behind this are that these miRNAs are either not detected, not measured, or below filtering limits. In these instances assessment proceeded using the smaller number of miRNAs actually available for analysis.

Another challenge encountered was the heterogeneity of the patient cohorts from multiple studies, which resulted in untenable sample sizes when filtered according to some criteria. To address this, we performed the validation procedure with filtering according to the biomarkers and also without filtering, using the complete validation datasets.

Our comprehensive validation protocol constitutes of three parts and these are described below:

1. Part 1: Validating the biomarkers and the survival endpoints associated with the biomarkers

   The accuracy of each biomarker in predicting patient clinical outcome was tested in each of the other expression datasets using equivalent survival endpoints (used in determining the survival outcome groups: GP and PP) associated with the signature. The validation datasets were filtered to be comparable to the setting where each signature was derived originally, including the determination of survival classes (GP and PP). However, in some instances these survival endpoints were not strictly identifiable or resulted in too small sample sizes to evaluate effectively in the validation data.

2. Part 2: Validating the biomarkers only

   We evaluated the performance of the biomarkers using the survival classes of the validation dataset. This allowed the use of the full validation dataset.

3. Part 3: Evaluating the performance of biomarkers relative to random feature sets

   This approach is a newly emerging standard for evaluating the performance of gene expression microarray signatures (Waldron *et al.*, 2014; Beck *et al.*, 2013), based on (Venet *et al.*, 2011) who found that equivalent random features could cluster patients into prognostically different subgroups. This challenged the tendency to interpret signatures that are significantly associated with survival, as having biological and/or clinical relevance. Therefore, to deal with this issue in this study we compared the predictive power of each signature with that of random miRNA sets of the same size obtained from within the same expression dataset. The comparison entailed:

   - The predictive power of the random miRNA sets (mean error rate over the error rates of 100 random sets of the same size as the signature of interest).

- Improvement over random signatures (IOR) score.

  Assessed by the fold change in the CV error for random signature compared to that of the signature of interest in each validation dataset. This produced 100 such improvement scores, one for each set of the random signatures.

  $$\text{Improvement over random signature} = \frac{\text{CV error of the random signature}}{\text{CV error of the signature of interest}}.$$

  A score greater than one indicates an improvement in prediction accuracy for the signature of interest compared to the random gene set. The higher the score, the greater the improvement and the greater the predictive capability of the signature with respect to a random set of miRNAs of the same size. A score below 1 indicates that the random miRNA set performs better on the respective dataset than the actual signature of interest. Similar scores have been computed to serve this purpose in the literature (Waldron *et al.*, 2014).

Results: The predictive power of each of the signature miRNA lists was assessed in the validation datasets (Table 4.3 and 4.4). The NSC classifier was used for this purpose, except in the case when only 1 miRNA was being assessed, and the support vector machine (SVM) classifier was used instead. The prediction accuracy of the final models was evaluated using LOOCV. The sample limitation in most of the datasets did not allow to use 5-fold CV.

Part 1 - Validation of each signature and its classes (Table 4.3): The table shows the CV error rates, the available miRNAs to be assessed in each validation dataset (square brackets) and the number of samples assessed in each survival class (parentheses; GP:PP). For example, in the validation of the Mann 12-miRNA signature on the Caramuta data, the CV error rate was 54% and out of the 12 miRNAs, only 7 were present in the Caramuta data (indicated by '[7]'). In this instance the total sample size of the Caramuta data when filtered according to the criteria of the Mann biomarker was 13 with 6 and 7 samples in GP and PP groups respectively (indicated by '(6:7)').

Table 4.3: **Part 1: Validation of the biomarker and their survival classes.** Summary of LOOCV error rates for independent validation of miRNA prognostic signatures in metastatic melanoma.

Biomarker		Mann	Segura	Caramuta	TCGA
Mann (12 miRNAs)		33% [12] (23:22)	Not evaluable[1] (1:1)	54% [7] (6:7)	38% [12] (5:8)
Segura	(18 miRNAs)	31% [16] (23:22)	22% [18] (36:23)	27% [7] (8:7)	42% [18] (25:11)
	(6 miRNAs)	31% [6] (23:22)	27% [6] (36:23)	40% [1] (8:7)	33% [6] (25:11)
Caramuta (6 miRNAs)		48% [5] (18:22)	Not evaluable[1] (1:2)	13% [6] (8:7)	Not evaluable[1] (2:8)
TCGA (15 miRNAs)		53% [5] (23:22)	Not evaluable[1] (1:7)	62% [3] (6:7)	39% [15] (11:12)

[1] Insufficient sample size for analysis.
Number of miRNAs able to be assessed in each validation dataset is indicated in square brackets. Number of samples assessed in each class (longer survival:shorter survival) is given in parentheses.

In four cases the error rate evaluation procedure could not be completed due to small sample size (shown in parentheses in Table 4.3). This is because in this case we restricted the validation dataset to survival classes of the signatures. The lowest error rate observed (13% for the Caramuta signature in their own data) did not validate when examined in the larger, independent sample size of 40 from Mann (5 out of 6 miRNAs present produced an error rate of 48%). The 18-miRNA and 6-miRNA signatures from the Segura data achieved estimated error rate s of 22% and 27% via their own data. Although the error rates were much higher in the independent validation, these signatures showed best independent validation results overall (27% in the Caramuta data for the 18-miRNA signature and 31% in the Mann data for both signatures). Of the newly proposed signatures, the Mann signature performed best in its own data (33%), but did not validate well in the independent datasets (38% in TCGA data being the best independent validation). TCGA signature did not perform well even in its own data (39%). However, a critical caveat was that very few of the 15 miRNAs were identifiable in independent datasets.

Table 4.4: **Part 2: Validation of the signature only.** Summary of LOOCV error rates for independent validation of miRNA prognostic signatures, using the survival classes of the validation dataset.

Biomarker		Mann (45)	Segura (59)	Caramuta (15)	TCGA (23)
Mann (12 miRNAs)		33% [12]	29% [11]	53% [7]	52% [12]
Segura	(18 miRNAs)	31% [16]	22% [18]	27% [7]	43% [18]
	(6 miRNAs)	31% [6]	27% [6]	40% [1]	39% [6]
Caramuta (6 miRNAs)		51% [5]	31% [6]	13% [6]	35% [6]
TCGA (15 miRNAs)		53% [5]	39% [10]	53% [3]	39% [15]

Number of miRNAs able to be assessed in each validation dataset is indicated in square brackets. Number of samples assessed in each validation dataset is given in parentheses.

Part 2 - Validation of the signature only (Table 4.4): This validation process was carried out because the examination of signatures and their associated classes as in Part 1 validation often lead to ineffectual sample size. The small sample sizes in the filtered datasets are due to differences among cohorts in survival distribution, tissue type, and/or other factors. Therefore, to overcome this challenge, the survival endpoints associated with the validation expression datasets were used, which resulted in increased sample sizes (Table 4.4, parentheses) relative to part 1. The prognostic utility of the signature per se was examined in relatable, although not identical survival class conditions. Some of the LOOCV errors remained the same due to samples being apportioned to good and bad survival classes in the same manner, despite the different follow-up definitions used. Therefore, the 18-miRNA signature from Segura still showed the best independent validation result (27% in the Caramuta data) with the Mann signature following closely behind (29% in the Segura data). The other signatures also validated in at least one independent dataset (31% for the 6-miRNA Segura signature in the Mann data, 31% for the Caramuta signature in the Segura data and 39% for TCGA signature in the Segura data), although this result could not be observed consistently in every independent validation dataset. In some cases these higher error rates may be attributed to the small number of miRNAs that could be evaluated in independent cohorts (*e.g.* out of the 15 miRNAs from TCGA signature, only a small number of miRNAs could be evaluated in other data; 3 in the Caramuta data and 5 in the Mann data).

Part 3 - Evaluation of prognostic miRNA signatures relative to equivalent random gene sets: When random signature validation proceeded via the Mann and Caramuta datasets, random gene sets produced the expected error rates of approximately 50% (Table 4.5). Notably, the same observation could not be made for signature assessments using the Segura expression data, where random gene sets predicted accuracy better than what would be expected by chance (40-41%). When considering the improvement over random signatures (IOR) scores, all but one signature - the 6-miRNA signature from Segura - showed the largest gains in accuracy over random equivalent gene sets when assessed in their own expression data (Figure 4.7, Table 4.6). In contrast, this signature displayed greater gains in accuracy over random sets when evaluated using the data from Mann (Figure 4.7). In terms of validation in independent datasets overall, the two signatures from Segura showed the largest improvement over random sets. The smallest gains were observed for TCGA signature (Figure 4.7, Table 4.6). It is noteworthy that the Caramuta signature showed high variability of improvement scores (1 to 7) when assessed via its own expression data. In terms of the validation datasets, TCGA dataset and the data from Mann showed the largest improvement scores for independent validations (Table 4.6).

Table 4.5: **Part 3: CV error rates from random miRNA sets.** The performance of the random signatures in each validation dataset using mean LOOCV error rates.

Biomaker		Mann (45)	Segura (59)	Caramuta (15)	TCGA (23)
Mann (12 miRNAs)		53%	41%	51%	62%
Segura	(18 miRNAs)	51%	40%	44%	61%
	(6 miRNAs)	59%	40%	58%	66%
Caramuta (6 miRNAs)		59%	40%	58%	66%
TCGA (15 miRNAs)		52%	40%	47%	61%

4.3.2 *Evaluation using the concordance index*

In this chapter, we used CV as the main approach of meta-analysis, where it is used as a validation tool for evaluating performance of the signatures in independent data. The error rates produced used patient clinical outcome groups (good vs. poor prognosis)

Table 4.6: **Part 3: Improvement in the accuracy of signatures over random signature scores.** The table gives the mean scores for the improvement over 100 random gene sets. The average score for signatures (using independent validation datasets) and the average scores for datasets are also presented in the table.

		Mann (45)	Segura (59)	Caramuta (15)	TCGA (23)	Average score for signatures
Mann (12 miRNAs)		1.62	1.41	0.96	1.20	1.19
Segura	(18 miRNAs)	1.66	1.82	1.64	1.57	1.57
	(6 miRNAs)	1.89	1.49	1.45	1.68	1.68
Caramuta (6 miRNAs)		1.15	1.30	4.45	1.88	1.44
TCGA (15 miRNAs)		0.99	1.03	0.88	1.57	0.97
Average score[1]for dataset		1.42	1.25	1.23	1.55	
Average score[2]for dataset		1.46	1.41	1.88	1.55	

[1] Average based on independent validation scores (where the dataset was not the training dataset of the respective signatures).

[2] Average across all signatures.

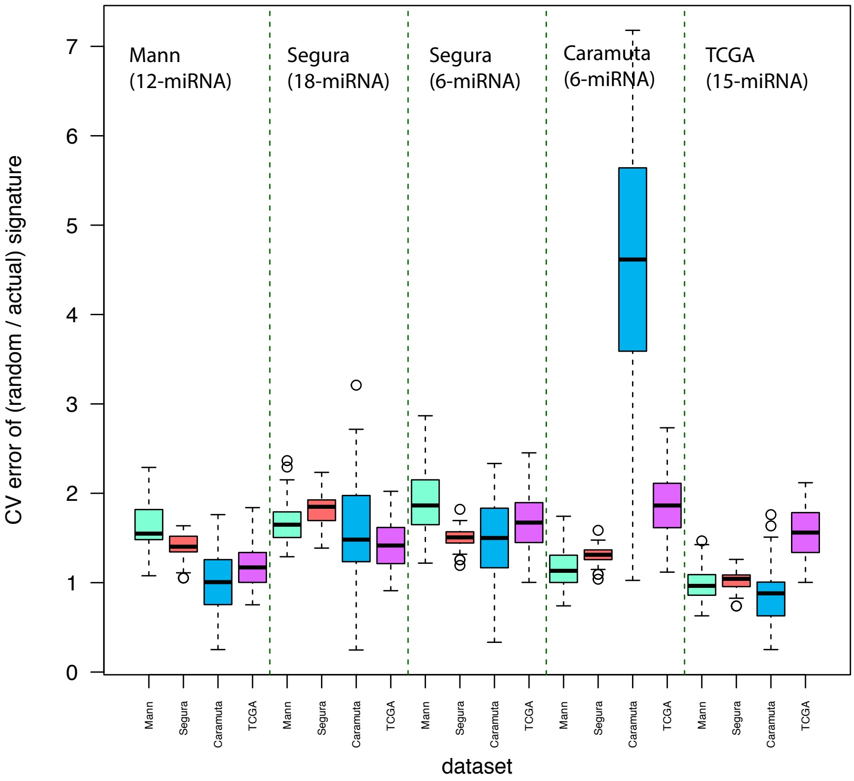


Figure 4.7: **Improvement over random signature scores.** The improvement in prediction error of the signatures relative to the prediction errors of equivalently sized random miRNA sets, for each of the 100 random miRNA sets generated, ordered by the signature of interest.

and are essentially estimating the ability of the model to predict the survival group a patient belongs to. Extending the analysis, we assessed all signatures using Uno's concordance index (C-index) (Uno *et al.*, 2011). The C-index is routinely used in the medical literature to quantify the capacity of a given biomarker to discriminate among subjects with different event times (Lee *et al.*, 2014; Pennells *et al.*, 2014; van Klaveren *et al.*, 2014; Waldron *et al.*, 2014). The C-index can be interpreted as 'the probability that a patient predicted to be at lower risk than another patient will survive longer than that patient' (Waldron *et al.*, 2014). Expected values of the C-index are 0.5 for random predictions and 1 for perfect risk models (Waldron *et al.*, 2014). This phase of the analysis comprised of two components:

1. Calculation of the C-index and standard errors.
   This proceeded via the 'survC1' package (Uno, 2013) in R and used similar parameters to previous related work by Waldron and colleagues (Waldron *et al.*, 2014) in its calculation. *tau* (truncation time) was the combined median survival time (21 months in this analysis) and *itr* (iterations) was 1000.

2. Synthesis of the results (C-indices) of each signature via a meta-analysis.
   Following the approach described in (Waldron *et al.*, 2014), this meta-analysis was implemented using the R package 'rmeta' (Lumley, 2012). The meta-analysis produced a synthesised C-index for each signature as a weighted average of C-index in each dataset. Weights of each C-index were calculated using the inverse of the estimated variance, corresponding to a fixed effects meta-analysis. In the synthesis of the C-index, the C-index based on the training dataset of each signature was excluded, so as to reflect the validation in independent datasets. Signatures were then ranked by this synthesised estimate of the C-index.

Results: The results revealed that all signatures predicted accuracy better than random predictions, as shown by C-index>0.5 (Table 4.7). Overall, TCGA signature showed the best predictive power via its own expression data with a near perfect risk score (C-index =0.99). However, this predictive power was not retained in independent validation using the other datasets (0.63-0.67). All other signatures performed well in their own training data with the lowest C-index of 0.72 reported in the case of the Mann signature, however, this signature performed better in independent datasets (C-index: 0.81-0.91). Due to the higher number of miRNAs in the Segura 18-miRNA signature compared to the sample size in TCGA dataset ($n = 23$), the Segura 18-miRNA signature could not be evaluated in TCGA dataset.

The Mann signature and the Segura 18-miRNA signature showed the highest concordance between the predicted and actual patient survival times (synthesised C-index score = 0.83; Table 4.7). This indicates that these two signatures exhibited the best independent validation performance. To remove a source of bias, the training dataset was excluded in the synthesis of the C-index of the signatures. The remaining three signatures did not yield high synthesised C-indices (0.65-0.68) demonstrating their poor performance in independent validation datasets.

The validation datasets were ranked by the average C-index across signatures (i.e., column-wise ranking). From the average C-index across signatures, TCGA dataset rendered to be the best validation dataset, closely followed by the Caramuta dataset. The performance of these two datasets (TCGA and Caramuta) opposed the results of the evaluation using CV errors (Table 4.3 and Table 4.4). When the CV was used (Section 4.3.1), these datasets failed to provide good validation results for most of the signatures, which could be attributed to the relatively small sample sizes in each of the survival classes.

4.4 ROBUST SET OF BIOMARKERS

We examined the direct overlap among the independently derived signatures analysed herein (summarised in Table C.2) to ascertain a more robust set of biomarkers. While there were no miRNAs common to all biomarkers, 5 miRNAs intersected between the Mann 12-miRNA and the Segura 18-miRNA signatures and 2 miRNAs intersected between the Segura 18-miRNA and the Caramuta 6-miRNA signatures (Figure 4.8). However, the meta-analysis showed that overall, the Segura signature and the Mann signature performed the best in independent data. Therefore, we consider the intersection between the Mann and the Segura signatures as a more robust set of biomarkers.

Table 4.7: **Meta-analysis via the C-index.** The synthesised C-index to assess the performance of each signature among independent datasets are given in the table, together with its 95% confidence interval, and the ranking of each signature. The last three rows show the average score for each validation dataset and the rank of each dataset based on that value.

	Mann	Segura	Caramuta	TCGA	Synthesised C-index	95% CI-lower	95% CI-upper	Rank of the signatures
Mann (12 miRNAs)	0.72	0.81	0.83	0.91	0.83	0.71	0.95	1
Segura (18 miRNAs)	0.73	0.85	0.92	N/E	0.83	0.67	0.98	1
Caramuta (6 miRNAs)	0.63	0.69	0.88	0.73	0.68	0.51	0.81	3
Segura (6 miRNAs)	0.66	0.83	0.63	0.70	0.66	0.58	0.78	4
TCGA (18 miRNAs)	0.63	0.66	0.67	0.99	0.65	0.56	0.74	5
Average score for dataset[1]	0.66	0.72	0.76	0.78				
Average score for dataset[2]	0.67	0.77	0.79	0.83				
Rank of the datasets	4	3	2	1				

[1] Average based on independent validation scores (where the dataset was not the training dataset of the respective signatures).

[2] Average across all signatures.

To evaluate the performance of these 5 miRNAs (miR-142-5p, miR-150-5p, miR-342-3p, miR-155-5p and miR-146b-5p), we generated the prediction error rates for this set in the four datasets using the CV (similar to Section 4.3.1).

Results: The 5-miRNA signature validated in the Mann and the Segura data (LOOCV error rates of 31% and 24% respectively) but not on other datasets (Table 4.8). The mean error rate for the signature was 40%. However, in the Caramuta data only 2 of the 5 miRNAs were present and the Caramuta and TCGA datasets did not provide good validation results in previous cases as well (Section 4.3.1). Therefore, further investigation on these 5 miRNAs using other independent data are required to assess their validity.

Table 4.8: **Validation of the 5-miRNA biomarker.** This table shows the LOOCV error rates for the identified set of miRNAs in every validation dataset.

Data	Mann	Segura	Caramuta	TCGA	Mean
LOOCV	31% [5]	24% [5]	67% [2]	39% [5]	40%

Number of miRNAs able to be assessed in each validation dataset is indicated in square brackets.

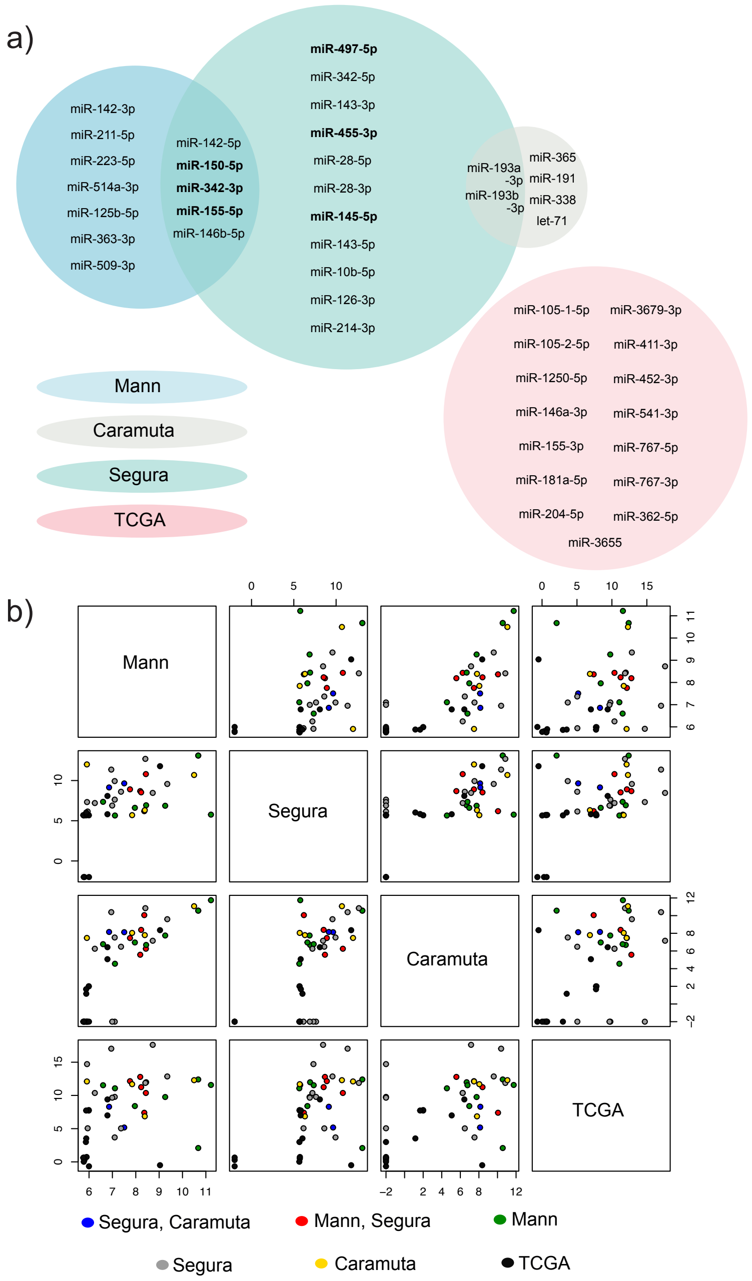


Figure 4.8: **Overlap among the biomarkers evaluated.** a) Venn diagram for the overlap among the miRNA-based signatures evaluated in this study. Circles sizes represent the actual relative size of each signature. The 6 miRNA signature from Segura is shown in bold. b) Scatter plots of the union of all miRNAs present in at least one of the signatures among the 4 validation datasets. Expression values are the raw values of the miRNA expression data in each dataset, transformed for ease of comparison (Table C.3). Colours represent miRNAs common between signatures and the value of -2 was used to represent miRNAs that were not present in the raw data even before any filtering was done.

4.5 DISCUSSION AND BIOLOGICAL IMPLICATIONS

In this chapter meta-analysis has been explored as a validation tool to evaluate the performance of biomarkers in independent datasets. Through our comprehensive meta-analysis procedure we have addressed many challenges associated with this form of horizontal integration and highlighted the distinct advantages.

In our meta-analysis procedure we developed a pre-processing approach for miRNA data prior to the data integration. This pre-processing step is imperative to ensure that there is adequate signal in the data to proceed to any downstream analysis. The signal in our miRNA data is much less than in the mRNA data, where approximately 50% signal is found. Therefore, mRNA data analysis methods cannot be applied to miRNA data directly. To address this issue of low signal, in our pre-processing pipeline we apply filtering on the data which is critical in enabling successful downstream analysis. The cut-off we apply (7.5) is arbitrary and depends on the dataset. The structure of the miRNA data shows that there is much variability between different probes within a miRNA (Figure C.1). Therefore this variability is accounted for in our analysis using a mixed effects model (Section 4.1). Currently there is no standard way of performing normalisation in miRNA data, in contrast to the more established methods in mRNA data. Our pre-processing pipeline addresses this dearth through a systematic comparison of normalisation methods to select the most appropriate method for miRNA arrays, that leads to perform meta-analysis and obtain meaningful downstream interpretations from the data.

The assessment of the signatures/biomarkers in multiple independent datasets assures robustness in the validation of signatures, allowing for high confidence in the error rate estimates along with their associated conclusions. This is one of the main advantages of the meta-analysis performed in this chapter. Such confidence is imperative in medical research, for the proposed and published biomarkers to be of value in clinical practise. This meta-analysis is also advantageous in identifying inconsistencies among multiple studies that use similar types of data of the same disease. For instance in this chapter we used melanoma miRNA data, but the published and proposed signatures from multiple studies did not contain common miRNAs across all signatures.

The low amount of intersection among signatures highlights the inconsistencies among studies that aim to address similar biological questions. Meta-analysis has been used frequently in the literature to address such inconsistencies. Some examples are (Qu *et al.*, 2015; Wei *et al.*, 2014; Wu *et al.*, 2014; Zeng *et al.*, 2014). Furthermore, meta-analysis is beneficial in highlighting reproducible features across multiple datasets, that would guide future research and extensive investigations on these features.

Despite the many merits of our meta-analysis we also encountered numerous challenges and in the following we address these.

One of the main challenges is the differences in the study aims and designs. This resulted in not all of the miRNAs in the signatures being present in the other validation datasets. Therefore, the signatures failed to be directly evaluated in other studies. In these instances the analysis was performed with the available, albeit low number of features.

To further investigate the possible reasons behind the absence of miRNAs across datasets, we examined the raw expression data prior to the application of any data pre-processing filters under each study (Figure 4.8, Appendix Table C.3). This was to ascertain whether the missing miRNAs were either not detected, not measured or below filtering thresholds of the pre-processing protocols. This investigation revealed that a number of missing miRNAs were removed during filtering, while others were not assayed or detected under each dataset. This detailed breakdown of the expression information serves to illustrate that while certain miRNAs play a significant role in their own training data, they might have relatively low impact in other validation datasets (Figure 4.8b). It also highlights the not insubstantial challenge to validation presented by among platform differences *e.g.*, more than half of the sequences captured via the RNA-seq technology used in TCGA assays were not assessed in the previous generation platform (Agilent's Human miRNA Microarray system) used by Caramuta and colleagues (Caramuta *et al.*, 2010). More details are provided in Appendix C.

Another challenge encountered in our analysis was the heterogeneity of the patient cohorts across multiple datasets. Because of the differences in the study cohorts, the validation of signatures together with the signature survival classes, was restricted (validation part 1). This is due to the untenable sample sizes that resulted after filtering the samples according to different survival endpoints of the studies. While our validation proceeded where possible in this restricted setting (validation part 1), to increase the sample size available for the analysis we instead used the survival endpoints associated with the validation expression dataset (validation part 2). Use of this approach addressed the challenge in dealing with ineffectual sample sizes while the prognostic utility of the signature was examined in relatable, although not identical, survival classes.

In general, although all of the studies were on metastatic melanoma, many biological differences in the study designs and cohorts were noticed. It could be due to this reason that none of the biomarkers globally validated in all independent cohorts, and this also might explain the reasons behind the absence of all miRNAs in each study. For instance, the Mann study was smaller in terms of sample size ($n = 45$) when compared with the Segura study ($n = 59$). The former was restricted to an analysis of AJCC Stage III regional lymph node metastases while the latter included both Stage III and Stage IV samples from among different tissue sites (brain, distant skin, local recurrence, regional lymph node, visceral, and regional skin). The more restrictive approach used in the Mann data sought to reduce the potential for confounding effects due to sample heterogeneity.

In spite of these hurdles, our study revealed many biologically relevant implications. We identified two new miRNA-based prognostic signatures. For compliance with the REMARK criteria, we compared the performance of these signatures with the predictive accuracy of standard-of-care clinico-pathologic markers (Balch *et al.*, 2009). It is interesting that the analysis of recently available data from TCGA did not produce a signature of high accuracy, nor was that signature validated in the independent data. Small sample size seems a probable explanation. It is also possible that cohort differences may have contributed to the result. For example, the survival data from TCGA were less mature than the data from (Tembe *et al.*, 2014), reflected in differences in the overall distribution of survival times between them.

We undertook the first systematic meta-analysis of all tissue-based prognostic biomarkers derived from studies of miRNA expression profiling in metastatic melanoma reported in the literature so far. The comprehensive nature of our meta-analysis procedure enabled the handling of the above discussed challenges while highlighting the advantages. Therefore, the approach is also prospectively applicable to upcoming signatures of interest and/or new potential validation datasets. Due to these reasons, this analysis is of high significance in the field of melanoma. Accurate prognostic information is essential for clinicians to be able to reliably stratify patients for a comparative assessment of treatment therapies and to inform patients of their likely future clinical outcome.

Despite the low intersection between the 5 signatures assessed in our study, the intersection of 5 miRNAs in Segura and Mann (miR-142-5p, miR-150-5p, miR-342-3p, miR-155-5p and miR-146b-5p) is encouraging from a reproducibility perspective. These same miRNAs being observed in independent expression profiling experiments will direct future research, suggesting that they should be immediately prioritised for further biomarker validation and functional analyses. Ongoing issues in the dearth of independent cohorts available for testing and validation of prognostic biomarkers, often hinders a meta-analysis. High quality validation data with reduced heterogeneity is urgently required to validate the prognostic biomarkers to be of clinical use.

Conclusion

In conclusion, our research revealed that the two signatures from (Segura *et al.*, 2010) and the 12-miRNA signature from the Mann dataset could indeed be validated in independent data. Moreover, our comparison of signatures with equivalent random gene sets showed that not all evaluations produced the expected random set error rate of 50%. In the case of the Segura expression data, where random sets of miRNAs achieved error rates of 40%, cohort heterogeneity (a mixture of patient with Stage III and IV disease as well as several tissue types) may have been responsible. Also of note, the large range of improvement scores observed using the biomarker from Caramuta and colleagues (Caramuta *et al.*, 2010), as assessed in its own data, shows that while for some random sets the signature performance is significantly better, random gene sets with predictive power similar to the signature itself could be found. Our meta-analysis procedure, which involves several approaches to validation, offers solutions to the critical limitations of this integrative analysis, and highlights the potential translational value in the biomarkers assessed while guiding future research directions.

5

VERTICAL DATA INTEGRATION

In this chapter we examine methods in integrating data vertically and develop integration frameworks to effectively incorporate information from multiple data sources (platforms) for the same patient. In contrast to horizontal data integration (Chapter 4), the focus of vertical data integration is on integrating multiple datasets from different platforms such as clinical data and multiple high-throughput (omics) data. In a medical context, it has been observed in many critical diseases that the patients with the same morphology have significantly different survival outcome (Schramm, 2014). Therefore, a single data source is unlikely to completely reflect the biology underlying a disease, and might be limiting in distinguishing patients with different outcomes (Jayawardana *et al.*, 2015a; Kim *et al.*, 2014; Chin and Gray, 2008; Hanash, 2004). Specifically in cancer, explaining cancer clinical outcomes remains challenging because of the complexity in the cancer genome (Kim *et al.*, 2014). Analysis at multiple levels of the biological system remains a necessity to fully elucidate tumour behaviour (Hanash, 2004). Vertical data integration has been motivated by these observations and expectations on the independent or complementary information different data types may provide. It holds the potential to significantly improve the prognosis of disease outcome with more insight into a patient's innate characteristics.

In the last decade, organised efforts have been made to generate a vast amount of 'vertical data'; that is, data from multiple levels of the biological systems (Chapter 1) obtained for the same sample, to facilitate the above investigations. With the advancements of high-throughput data technologies, the generation of data from multiple platforms has become quicker and cheaper (Grada and Weinbrecht, 2013; Metzker, 2010). The large scale efforts by consortia such as TCGA[1] made matched samples across different datasets publicly available. However, the development of improved statistical methods and frameworks in this paradigm are required to close the gap between the availability of a vast amount of data and the biological questions that need to be addressed. In this chapter, we aim to address and close this gap by developing statistical frameworks with impact in the medical context.

In addition to the majority of the issues in horizontal data integration (Chapter 4) such as those associated with individual platforms (data types), this chapter deals with

[1] The cancer genome atlas (http://cancergenome.nih.gov).

the new challenges that arise from vertically integrating data from distinct platforms (different data types). One of the key challenges here is having more variables than observations in the omics data, the large p small n problem, and therefore we have to deal with a substantially increased number of variables in the integration. Furthermore, different platforms have significantly different number of variables, such as the clinical data with $p < n$ and omics data with $p \gg n$, causing an imbalance of variables in the integration setting. Moreover, the mismatch of the samples between different data types further reduces the effective sample size, which hinders the development of statistical methods. The integration of different data types to determine which data source contains more influential prognostic information and to investigate whether data integration improves upon the current clinical standards, remains a challenge in the presence of the above concerns. Despite the many issues, the potential advantages of this class of data integration, such as the complementary information from different data types, has inspired many researchers to further explore this area of research. Furthermore, if the dominant prognostic variables could be identified from each data type, the cost of data generation could be significantly reduced.

This chapter aims to address these challenges of vertical data integration and explores its advantages via the development of novel statistical frameworks in a predictive setting. We exploit the availability of multiple data sources; clinical, pathological and mutation information, mRNA, miRNA and protein information, on the same set of patients from the Mann data (introduced in Chapter 2). The use of this motivating dataset facilitates to assess the actual biological relevance of the methods developed.

This chapter begins with the individual platform analysis in Section 5.1, which is an imperative step before proceeding to data integration. In Section 5.2, the principle of pre-validation (Tibshirani and Efron, 2002) is extended to develop novel frameworks in vertical data integration, incorporating the mBMI framework (developed in Chapter 3). Figure 5.1 summarises the data and the procedure followed in Sections 5.1 and 5.2. In Section 5.3, platform dependent weights are constructed to develop a vertical data integration framework, utilising the weighted Lasso (Bergersen *et al.*, 2011). The cross-validation (CV) procedures discussed in Chapter 2 are utilised to generate a novel form of a visualisation technique, that can be used in the vertical data paradigm, in

Section 5.4. This leads to the biologically-based discovery of the dominant sources of prognostic information out of the multiple data sources at the sample level.

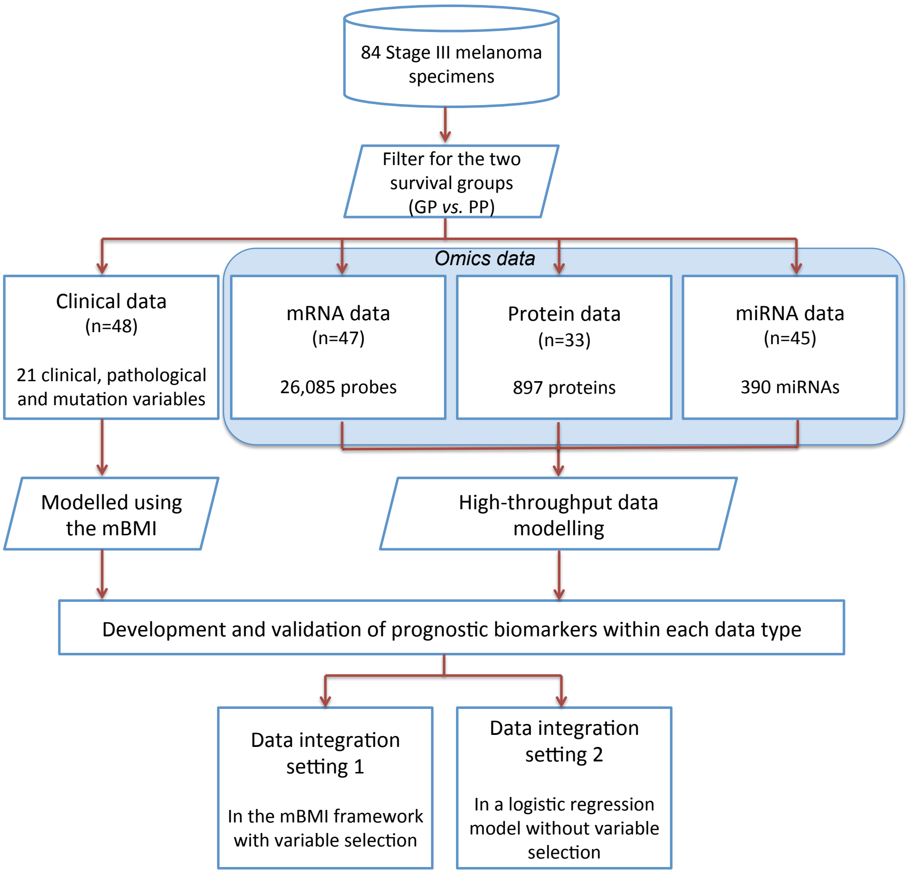

84 Stage III melanoma specimens
Filter for the two survival groups (GP *vs.* PP)
Omics data
Clinical data (n=48)
21 clinical, pathological and mutation variables
mRNA data (n=47)
26,085 probes
Protein data (n=33)
897 proteins
miRNA data (n=45)
390 miRNAs
Modelled using the mBMI
High-throughput data modelling
Development and validation of prognostic biomarkers within each data type
Data integration setting 1
In the mBMI framework with variable selection
Data integration setting 2
In a logistic regression model without variable selection

Figure 5.1: **Schematic of the data integration procedure using pre-validated vectors.**

5.1 INDIVIDUAL PLATFORM ANALYSIS

In this section, we model the different data types (platforms) individually to identify the most accurate prognostic biomarker through each data platform. This would enable the comparison of data platforms and to evaluate whether a single platform was capable of achieving the desired accuracy or whether the data integration undeniably adds value. The analysis at individual platform level is also important in understanding the noise level associated with each platform. The clinical, mRNA, protein and miRNA data from the Mann dataset will be used in this chapter. The details of the pre-processing steps followed are given in Section 2.1.1.

5.1.1 *Clinico-pathologic and mutation data modelling*

The clinical data component of the Mann data comprises of clinical, pathological and mutation variables. We used the mBMI framework developed in Chapter 3 to model the clinical data; its modelling procedure is detailed in Section 3.2.1.

5.1.2 *High-throughput data modelling*

Data from each of the high-throughput (omics) data platforms were analysed using a two-step procedure. In the first step the optimal number of features (genes, proteins, miRNAs) is selected based on the CV error rate and then, a prognostic classifier is built on the selected features. For this purpose, the CV procedure 'B: Classifier CV' outlined in Section 2.3.2 was used. The specific choice of the parameters is detailed below.

Feature selection of high-throughput data

mRNA and miRNA data: The feature selection (devising molecular signatures) in mRNA and miRNA data was performed using the 'median robust' method (Jayawardana *et al.*, 2015a; Campain, 2012). In the median robust method, the genes/miRNAs are ranked based on the difference between the median expression of the two groups (PP and GP), and a set of molecular signatures are devised such that the k^{th} molecular signature contained the k top-ranked features.
The final expression value for gene/miRNA q is

$$\mathrm{expr}_q = \tilde{x}_{q,GP} - \tilde{x}_{q,PP},$$

where $\tilde{x}_{q,GP}$ and $\tilde{x}_{q,PP}$ represent the group median for the q^{th} gene/miRNA for GP and PP groups, for $q = 1, 2, \ldots, p$ ($p = 26085$ in the mRNA data and $p = 390$ in the miRNA data).

Protein data: A fixed effects model was used to select features. For each protein, peptides with less than 24 missing values across samples ($n = 33$) were considered and a

fixed effects model was applied with the group effect (whether each sample belongs to group PP or GP) and peptide effect (which takes into account the fact that each protein contains multiple peptides). The peptides were ranked based on the coefficient estimates of the group effect from the fixed effects model. A set of molecular signatures was devised such that the k^{th} molecular signature contained the k top-ranked features. For each molecular signature, the features (peptides) in that molecular signature were filtered from the data set, adjusted for the sample means and aggregated on protein level using the mean.
The final expression value for protein q is;

$$\mathrm{expr}_q = \mathrm{group}_q = \text{Estimated group effect } q \text{ from the fixed effects model,}$$
$$x_{qjk} = \mathrm{group}_q + \mathrm{peptide}_j + e_{qjk},$$

where group_q represents the effect of two groups GP and PP and $\mathrm{peptide}_j$ represents the peptide effect.

Classification methods

Prior to performing the actual analysis and obtaining final prediction error rates for the high-throughput data, we compared a number of classification algorithms and the best performing classification algorithm was chosen to conduct the final modelling. A brief summary of the classification methods employed in this study is given in the following.

1. Diagonal Linear Discriminant Analysis (DLDA)
   DLDA (Hastie *et al.*, 2003) is commonly used in high dimensional data settings and preferred over LDA (Linear Discriminant Analysis). It is preferred because DLDA can have more variables (p) than samples (n). DLDA assumes that the features are independent within each class, that is, the within-class covariance matrices are diagonal, other than having a common covariance matrix across classes as in LDA. New samples are classified to the class that gives the largest value for the discriminant score. DLDA was implemented in this thesis using the R package 'supclust' (Dettling and Maechler, 2011).

2. Nearest Shrunken Centroids (NSC)

   NSC (Tibshirani *et al.*, 2002) identifies subsets of variables that best characterise each class. This method computes a standardised centroid for each class, which is the average expression for each variable divided by the within class standard deviation for that variable and shrinks each of the class centroids towards the overall centroid using soft thresholding. For a new sample, it takes the expression profile and compares it to each of the shrunken class centroids, and classifies the sample into the class whose centroid that it is closest to. The R package 'pamr' (Hastie *et al.*, 2014) was used to implement NSC in this section.

3. k-Nearest Neighbours (KNN)

   KNN (Ripley, 1996) identifies the k nearest neighbours or observations to a new sample based on some distance measure (usually the Euclidean distance) and classifies the new observation using the majority decision rule. In this section results for $k = 5$ are shown, as results for $k = 1$ and $k = 10$ did not produce greater accuracy. KNN was implemented using the R package 'class' (Venables and Ripley, 2002) in this section.

4. Support Vector Machine (SVM)

   SVMs were originally developed by (Cortes and Vapnik, 1995) for binary classification. SVM aims to find the optimal separating hyper-plane between two classes by maximising the margin between the closest points of the classes. The points on the boundaries are called support vectors. When a linear separator cannot be found, the data are projected to a higher dimensional space via kernel techniques, where the data becomes linearly separable. This whole task involves solving a quadratic optimisation problem. The results using the linear kernel are shown in this section. Other kernels have also been used and are not detailed here as they did not produce greater accuracy. SVM was implemented using the R package 'e1071' (Meyer *et al.*, 2014) here.

Results: Comparison of classification methods on high-throughput data. The above mentioned classification methods were employed in building a classifier and these classifiers were then compared in a single run of the CV process to select the best performing classifier that gives the minimum CV error rate for each data type. Figure 5.2 illustrates this comparison of different classification methods for the three high-throughput data platforms. As shown, there were no significant differences between the four classification methods over the range of molecular signatures in each data type. Through this comparison, the following classification methods were chosen as the best performing classification methods:

- mRNA data: SVM
- protein data: NSC
- miRNA data: NSC

5.2 INTEGRATIVE PROGNOSTIC MODELLING USING PRE-VALIDATED VECTORS

'Pre-validation' was initially developed in (Tibshirani and Efron, 2002), with the aim of constructing a less biased microarray predictor to be modelled alongside the clinical variables. Many applications and extensions of pre-validation are available in the literature. For example, (Boulesteix *et al.*, 2008) assessed the additional significance of microarray data compared to clinical data in breast cancer and colorectal cancer data, (Segura *et al.*, 2010) compared the prediction accuracy of a miRNA signature for melanoma patients to that of other predictors on clinical and demographic variables and (van Vliet *et al.*, 2012) compared three different integration strategies (early, intermediate, and late integration) on breast cancer data containing gene expression data and clinical parameters.

This chapter makes use of the principle of pre-validation, where a molecular signature from each omics data platform is used to obtain a single variable (the pre-validated vector) which is modelled in combination with one another and also with the clinical variables.

The pre-validation procedure of (Tibshirani and Efron, 2002) is detailed as follows and it is graphically represented in Figure 5.3.

1. Divide the samples into k equal parts.

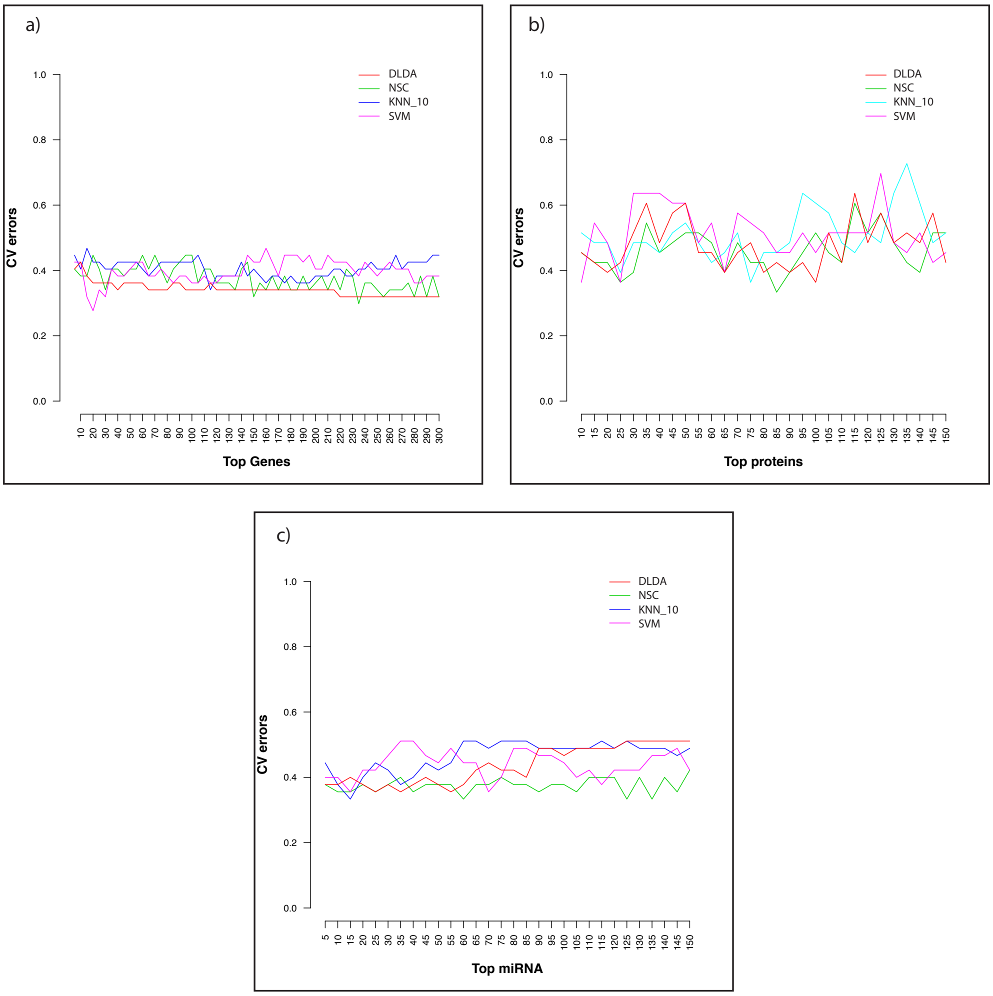


Figure 5.2: **Comparison of classification methods in high-throughput data.** The x-axis represents the number of features in the molecular signature, and the y-axis represents the 5-fold CV error rate. a) Comparison for mRNA data. b) Comparison for protein data. c) Comparison for miRNA data.

2. Set aside one part as the test set component.

3. A molecular signature (set of features) is obtained using the other $k - 1$ parts (learning set) and a classifier is trained on the learning set on the molecular signature.

4. Use this classifier to predict the survival class of the k^{th} part.

5. Repeat steps 2-4 for all k parts, resulting in a pre-validated vector of estimates for the omics data. This pre-validated vector (denoted as A_{PV}) is a complete prediction vector with one prediction for each sample.

6. This pre-validated vector is used together with other types of variables (*e.g.*, clinical variables) in a logistic regression model to get the final prediction error rate.

When an independent data set is available, pre-validation is not a substitute. If $k = n$ this would result in a highly variable LOOCV pre-validation vector (Tibshirani and Efron, 2002).

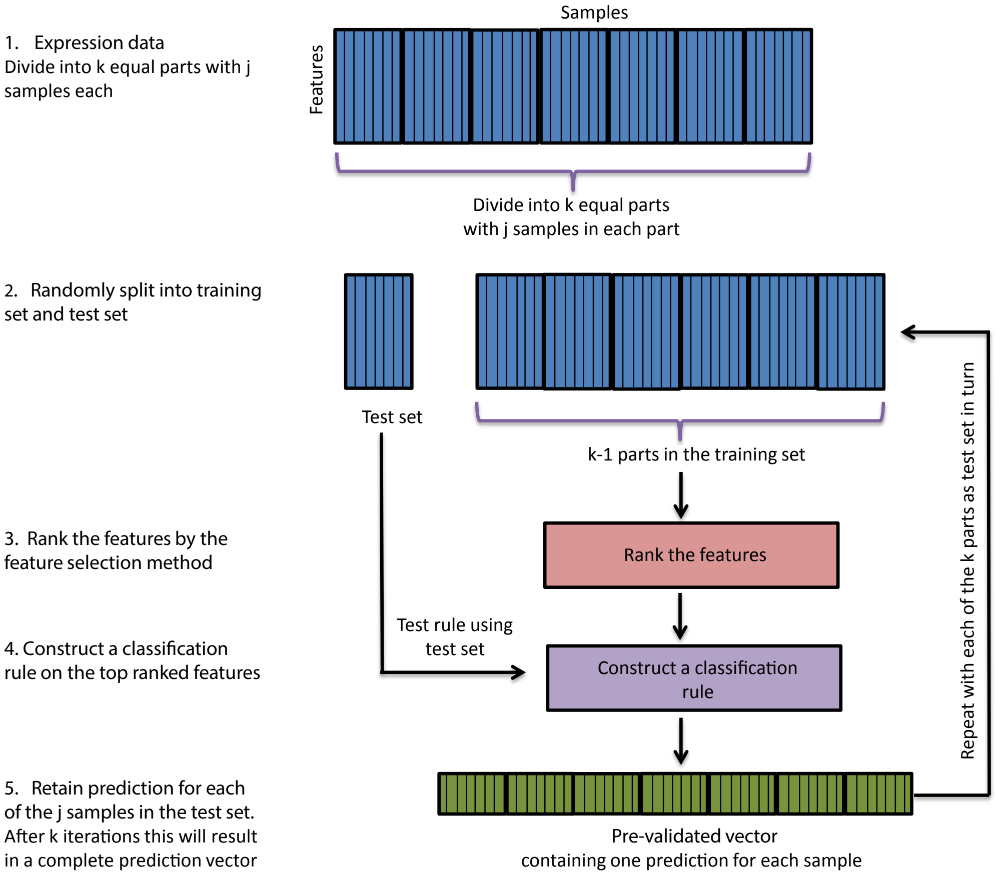

Samples
1. Expression data
Divide into k equal parts with j samples each
Features
Divide into k equal parts
with j samples in each part
2. Randomly split into training set and test set
Test set
k-1 parts in the training set
Repeat with each of the k parts as test set in turn
3. Rank the features by the feature selection method
Rank the features
4. Construct a classification rule on the top ranked features
Test rule using test set
Construct a classification rule
5. Retain prediction for each of the j samples in the test set. After k iterations this will result in a complete prediction vector
Pre-validated vector
containing one prediction for each sample

Figure 5.3: **The pre-validation procedure.** The graph illustrates the pre-validation procedure in Tibshirani and Efron (Tibshirani and Efron, 2002).

In the integrative analysis, when the p clinical variables are integrated with the pre-validated vector from the omics data (A_{PV}), a logistic regression model can be developed as follows:

$$\text{logit}(\pi_j) = \beta_0 + \beta_1 x_{1,j} + \beta_2 x_{2,j} + \ldots + \beta_p x_{p,j} + \beta_{p+1} A_{PV_j}$$

Where π_j is the probability of an 'event' occurring ($\pi_j = P(Y_j = 1)$), β_q is the coefficient for the qth variable ($q = 0, 1, \ldots, p+1$) in the regression, and $x_{q,j}$ is the j^{th} sample's observation for the q^{th} clinical variable. The pre-validation procedure could be easily adapted to include multiple pre-validated vectors (*e.g.* $A_{1,PV}, A_{2,PV}$). In the integrative settings, this pre-validated vector is treated as a complete variable and can be coupled with the m imputed clinical data sets as an additional variable for modelling purposes leading to m regression models. These models can then be aggregated using an inclusion frequency (as explained in Chapter 3). Hence this procedure allows the production of an integrated regression model making use of the pre-validated high-throughput data vector and the clinical variables.

5.2.1 *Data integration framework*

To identify the principal sources of prognostic information from among the different data types - clinical, pathologic, mutation, gene, protein and miRNA information - we performed an integrative analysis under two settings. Both settings involved the principle of pre-validation (Tibshirani and Efron, 2002). The two integrative prognostic modelling procedures developed are detailed below.

- ***Setting 1: With variable selection (mBMI)***
  The mBMI method (Chapter 3) was used, where all of the clinical variables ($p = 21$) and pre-validated vectors from the high-hroughput data sources were used as initial variables. The pre-validated vector for each of the omics data types used was the one that gave the lowest prediction error rate out of the 100 that were obtained in high-throughput data modelling as described above. The mBMI procedure selected from among these variables (clinical variables and pre-validated vector/s) to generate a final model focussing on the prognostic capability of the selected group of variables.

- ***Setting 2: Without variable selection***
  Top-performing clinical variables (from Chapter 3, Section 3.2.1) were modelled together with the pre-validated vectors from the various high-throughput data types in a logistic regression model. The prognostic capability of these models was then assessed using 5-fold CV error rates.

Results: In these integrative settings, the patients were matched across datasets according to their identification numbers. The highest number of possible samples were considered in each integration. To account for the variability, the analysis was repeated 100 times and the mean error rates are reported in Table 5.1. Overall, the data type associated with the lowest mean CV error rate was the mRNA data (25%). The other omics data (protein and miRNA) did not show any improvement over the clinical data (mean error rates 35% and 37% respectively compared to 30% in the clinical data).

The integrative analysis aided in identifying dominant sources of prognostic information that could allow for improvements to individualised risk assessment. In integration setting 1, the lowest error rate of 18% was reported for the combination of clinical information with a gene expression pre-validated vector, showing that the integrated platforms could perform better than individual platforms. Relative to this result, prognostic accuracy was reduced for combinations of protein and clinical information (32%) as well as miRNA and clinical information (33%). Although the clinical data alone performed better, this indicated an improvement over using protein and miRNA data alone in prognosis. Integrating clinical data together with signatures from all three high-throughput platforms produced an error rate of 29% (Table 5.1, Figure 5.4).

In integration setting 2, the mean 5-fold CV error rates for the multiple permutations analysed were on average higher compared with those returned under integration setting 1: clinical and mRNA data (26%), clinical and protein data (33%), clinical and miRNA data (33%), and clinical data with signatures from all three platforms (33%) (Table 5.1 and Figure 5.4).

5.2.2 *Validation*

The validation of the results from individual platform analysis and integrative analysis consist of 2 components. Firstly to comply with the standards of biomarker validation, the constructed models were compared with the current standard among the clinicians for prognosis of Stage III melanoma patients. Due to the unavailability of a similar independent dataset, the results were validated on the full cohort of the present dataset itself, prior to any restrictive filtering of the samples. These are detailed below.

Table 5.1: **Five-fold cross-validation error rates for the final models.** Error rates in predicting patient outcome for each of the different data types analysed in this study, as well as for various combinations thereof, with the model applied.

		Mean 5-fold CV error rate		Classification method	Details
1	clinical	30%		mBMI with LR[1]	Individual platform analysis
2	mRNA	25%		SVM	
3	protein	35%		NSC	
4	miRNA	37%		NSC	
		Integration setting 1 (mBMI)	Integration setting II (without selection)		
5	clinical + mRNA	18%	26%	Pre-validated LR[1]	Integrative analysis
6	clinical + protein	32%	33%	Pre-validated LR[1]	
7	clinical + miRNA	33%	33%	Pre-validated LR[1]	
8	clinical + mRNA + protein + miRNA	29%	33%	Pre-validated LR[1]	

[1] LR: logistic regression

1. Comparison with the standard-of-care

   To compare the prediction accuracy of the models constructed above with the existing standard-of-care factors, the four most statistically significant clinicopathologic prognostic parameters in patients with Stage III melanomas (*i.e.*, number of tumour-positive lymph nodes, tumour burden at the time of staging (microscopic v. macroscopic), presence or absence of primary tumour ulceration, and thickness of primary melanoma) were evaluated in relation to the GP and PP classes of this study. A logistic regression model constructed using these 4 variables was used for this purpose and the prediction error rate was assessed by 5-fold CV error rate.

2. Survival analyses

   Kaplan-Meier curves were constructed for the prognostic classifiers developed under the various combinations of data types, based on observed survival times (in years) and survival status (dead (event) or alive (censored)) using the R package 'KMsurv' (Klein *et al.*, 2012). Because of the limited availability of similar independent data, the full Stage III cohort ($n = 84$) from the Mann data was

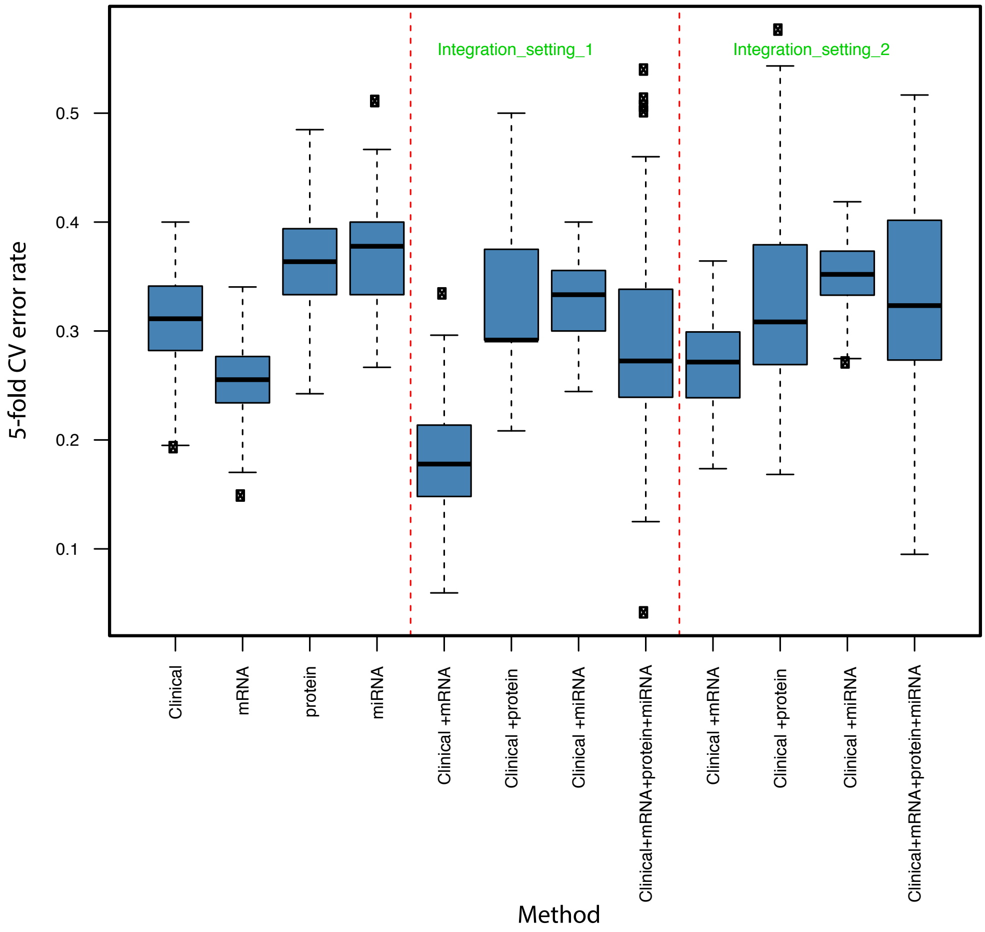


Figure 5.4: **Prediction error rates for the prognostic models considered in this analysis.** 5-fold CV error rates for the different data types, individually and in combination with each other, when modelled under two different integration settings: i.e. with and without variable selection.

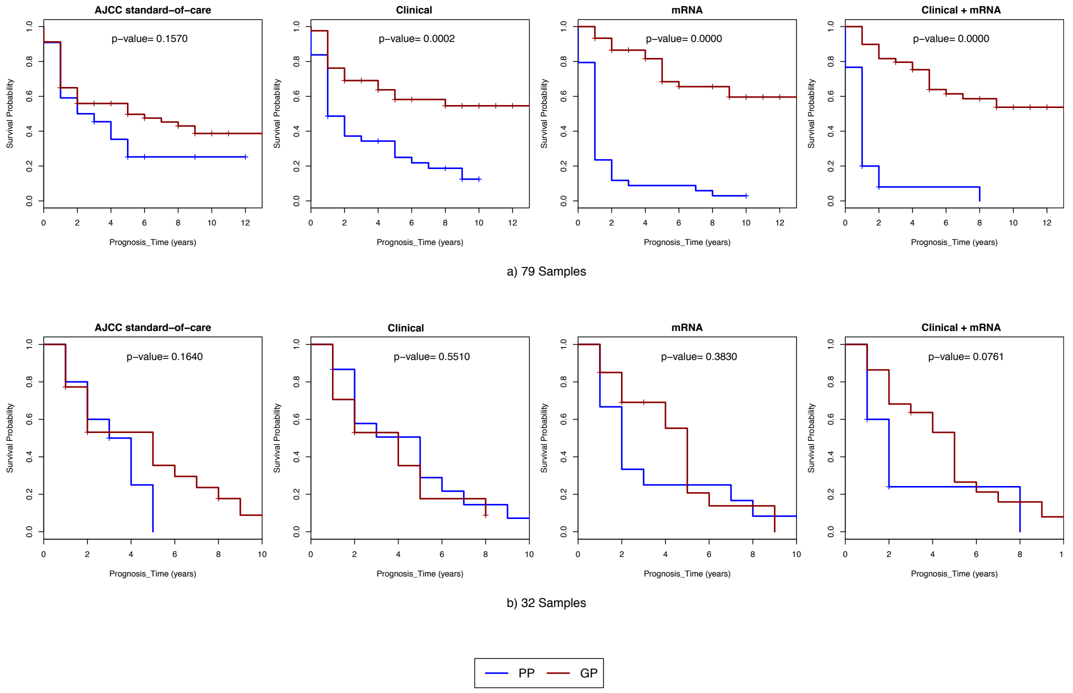
used without the samples for which data were unavailable. *e.g.,* for the model combining clinical and mRNA information, Kaplan-Meier curves were constructed for the subset of patients ($n = 79$) for whom clinical and mRNA data were available. A log-rank test was used to examine whether there were significant ($p - value < 0.05$) differences between groups (GP and PP) being compared.

Results: The prognostic models built in this study were more accurate than the existing standard-of-care. This is illustrated by the mean 5-fold CV error rate of 52% via a logistic regression model built on these 4 variables, compared to the error rates from all the other models (Table 5.1).

The lowest prediction error (18%) was observed when the clinical and mRNA data were integrated under integration setting 1 (Section 5.2.1). This resulted in an improvement over the individual platform modelling of both clinical and mRNA data. Therefore, the KM curves were constructed for the clinical, mRNA and the integrated clinical and mRNA classifiers. Figure 5.5 illustrates the KM curves for these data using all patients in the AJCC Stage III cohort for whom data were available ($n = 79$, Figure 5.5A) and for the independent subset of these patients ($n = 32$, Figure 5.5B). All Kaplan-Meier plots generated using the 79 AJCC Stage III patient data displayed significant among-group differences ($p - value < 0.001$), except for the standard-of-care variables ($p - value = 0.157$). Notably, the p-value for the combined clinical and mRNA classifier was much lower (6.95×10^{-13}) than the p-value from clinical individually (2.43×10^{-4}). Significance was not retained for the independent subset of 32 patients. However, the p-value (0.08) for the integrated clinical and mRNA classifier was noticeably lower than the p-values observed for the two data types evaluated individually, as well as for the standard-of-care classifier ($0.55, 0.38$ and 0.16, respectively).

5.3 integration via platform dependent weights

In vertical data integration it is crucial that equal credence is given to all data sources, such that the modelling procedure is not dominated by the platform size. An important procedure that we can adapt in this context of vertical data integration is the use of platform dependent weights to guide the variable selection in a data platform. We can use one platform (data type) to derive weights, exploiting a particular relationship with the other data platform, thereby assuring that the final variables selected are indicative of both data types. This intuitive method of integration avoids the extreme variable reduction in one platform (for example the pre-validation approach) and also the risk of one data platform being completely dominated by the other (for example direct integration of two data types). Rather, the information from multiple data types is utilised effectively, that is flexible upon the biological context where we perform the data integration.



a) 79 Samples

b) 32 Samples

Figure 5.5: **Kaplan-Meier curves of the different classifiers.** Kaplan-Meier curves were constructed for good prognosis (GP) and poor prognosis (PP) patient groups based on predictions using: 1) Standard-of-care variables; 2) Clinical information; 3) MRNA-based molecular signature; and, 4) The integrated classifier comprising clinical data and a pre-validated mRNA classifier. The top panel a) includes all 79 AJCC Stage III patient samples in the cohort, while the bottom panel b) includes only the subset of the cohort not used to construct the classifier (*i.e.*, independent subset of samples). P-values reflect the log-rank test.

This approach is based on the adaptive Lasso proposed by (Zou, 2006), where the authors proposed the use of weights to guide the variable selection in a standard Lasso (Tibshirani, 1996) procedure. (Bergersen *et al.*, 2011) adapted this procedure to be used in a vertical data integration framework, where additional data enters the model indirectly by acting on the penalty parameter of each variable. This approach naturally assumes there exists a primary platform and this notion is consistent with many cancer prognosis studies (Segura *et al.*, 2010; Bogunovic *et al.*, 2009). The weighted Lasso approach thus avoids a further increase in the number of variables and holds the promise to be an innovative method in integrating omics data, as it allows data dependent weights to be chosen. To date, a number of methods used variations of the weighted Lasso to introduce variable specific penalisation. This includes (Shimamura *et al.*, 2007) in graphical Gaussian models, (Charbonnier *et al.*, 2010) in time course expression data and (Garcia *et al.*, 2013) in structured variable selection.

In the following, we provide a background on the weighted Lasso procedure, and detail the data integration framework we develop adapting the weighted Lasso.

5.3.1 *The weighted Lasso*

In high dimensional omics data settings where there are more variables than samples, many of the standard statistical methods fail to work. Lasso (Tibshirani, 1996) based methods became widely used, because they allowed variable selection using fast algorithms in a high dimensional setting. The weighted Lasso is one variation of the Lasso, where data dependent weights are used in the penalisation such that the penalty parameter varies for each covariate/variable. Suppose $\mathbf{y}$ is the response vector, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p]$ is the predictor matrix, $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is the parameter vector, λ is a regularisation parameter and w_q are data specific weights. The objective function to be minimised is:

$$||y - X\beta||^2 + \lambda \textstyle\sum_{q=1}^{p} w_q |\beta_q|.$$

The weighted Lasso is a generalised version of the standard Lasso ($w_q = 1 \forall q$). The adaptive Lasso (Zou, 2006) is a special case of the weighted Lasso, where the weights $w_q = \frac{1}{|\hat{\beta}_q|}$; and $\hat{\beta}_q, q = 1, 2, \ldots, p$ are the ordinary least squares estimates after regressing $\mathbf{y}$ on $\mathbf{X}$. The R package 'glmnet' (Friedman *et al.*, 2010) facilitates the use of variable specific weights in the Lasso selection.

The use of extra knowledge in inference has been prevalent in the field. Grouping of the genes in Group Lasso (Yuan and Lin, 2006) and prior structural information on genes in the Elastic Net Procedure (Slawski *et al.*, 2010) are some examples. Furthermore, recent studies show that more stable results can be generated for the Lasso, when weights based on relevant external information or prior knowledge of the variables are used in the penalty parameter (Bergersen *et al.*, 2011; Charbonnier *et al.*, 2010).

5.3.2 *Data integration framework*

The data integration framework we develop in this section facilitates the use of multiple data sources available for the same set of samples, so that the features selected are more relevant and informative in the prediction of the outcome. We do this by constructing weights from one data platform, using the association between the two data platforms considered. The information across various platforms then enters the model indirectly as weights, to guide the feature selection procedure of another platform. The integration framework developed here addresses the omnipresent challenge in the vertical integration setting of more variables than samples ($p \gg n$), by avoiding a further increase in the number of features.

Our data integration framework emulates the CV procedure 'A: Full CV', where in step 2, the weighted Lasso described above (Section 5.3.1) will be used as the feature selection procedure. To illustrate the framework developed herein, mRNA, protein and miRNA data from the Mann data (Chapter 2) will be utilised. Furthermore, the mRNA dataset will be considered as the primary platform as it was observed that gene expression data have the best prognostic information (Section 5.1). The information from other datasets will be integrated via weights.

The integration methods compared are:

- Lasso (no weights, feature (gene) selection using only mRNA data)
- WL_GP (weighted Lasso gene selection using weights based on protein data; integration of mRNA data and protein data).
- WL_GM (weighted Lasso gene selection using weights based on miRNA data; integration of mRNA data and miRNA data).
- WL_GPM (weighted Lasso gene selection using weights based on protein and miRNA data; integration of mRNA data, protein and miRNA data).

5.3.3 *Correlation based weights*

In the integration framework the weighted Lasso is used to select genes that contain the information from protein and miRNA data. We investigated the effect of using the inverse correlation of mRNA data with other data types as weights. The use of this weighting scheme was associated with the assumption that, genes highly correlated with selected proteins/miRNAs are more likely to be associated with patient survival. Therefore, such genes were given less weight to have a higher chance of being selected.

For the integration of two data types (*i.e.,* WL_GP and WL_GM explained above), we use weights w_q, where

$$w_q = \frac{1}{\rho_q},$$
$$\rho_q = \mathrm{median}(|cor_s(G_q, X_1)|, |cor_s(G_q, X_2)|, \ldots, |cor_s(G_q, X_p)|).$$

The ρ_q is the median across absolute values of the Spearman correlation coefficients (cor_s) for q^{th} gene (G_q) and the p proteins/miRNAs ($X_1, X_2, \ldots, X_p$) (Jayawardana *et al.,* 2013). Similar correlation based weights have been used by (Bergersen *et al.,* 2011), where they used the Spearman correlation coefficient to construct weights in integrating gene expression and copy number data.

We construct a composite weight for the integration of all three omics platforms (WL_GPM),

$$w_q = \frac{1}{\rho_q},$$
$$\rho_q = \max(\rho_q(GP), \rho_q(GM)),$$

where $\rho_q(GP)$ and $\rho_q(GM)$ are the median correlation measures obtained as described above for each gene based on protein and miRNA information respectively. It is conjectured that the genes whose expression values are highly correlated with at least one of the other two platforms are more likely to better explain the outcome.

5.3.4 *Comparison study*

Prior to performing the integrative analysis, firstly the three data types were used to investigate the different parameters that could be optimised within our data integration framework. The main criterion of evaluation used in this comparison was the CV error rate produced via the 5-fold CV procedure. The following aspects in the CV procedure ('Full CV') were investigated:

1. The set of variables to be considered

   Lasso feature selection facilitates the feature selection in high-throughput data settings where there are more variables than samples ($p \gg n$). Therefore, Lasso based methods can be used to select features in the mRNA expression data that has thousands of variables. However, the number of features that is selected will be approximately 0.2% of the initial number of features ($p = 26085$ and $n = 47$ in mRNA dataset). Hence a high degree of variability will likely be observed in the feature selection. Therefore, in this comparison study different combinations of variables were considered for the prediction accuracy.

   - 1: Using the full set of genes.

     The complete set of genes ($p = 26085$ in mRNA data and $p = 200$ in protein/miRNA data) were used in the analysis.

   - 2: Using subset of genes.

     A subset of genes ($p = 200$ in mRNA data and $p = 200$ in protein/miRNA data) were used in the analysis.

   In both cases, subsets ($p = 200$) of the secondary data platform variables (protein and miRNA data) were used to ensure that the external information used as weights to aid in gene selection are indeed 'relevant external information'.

   Subset selection:

   The selection of subsets was based on the differentially expressed (DE) features. The features were ranked according to their adjusted p-value in a DE analysis for the separation of two outcome groups, GP and PP, and the top 200 features were selected. The DE analysis was carried out using the R package 'limma' (Smyth, 2005) where a linear model was used to identify DE features between the two groups. The expression of feature q could be modelled as;

$$\mathbf{y}_q = \mathbf{X}\boldsymbol{\beta}_q + \boldsymbol{\epsilon},$$

where $\mathbf{y}_q$ is the vector of expression values for feature q, $\mathbf{X}$ is the design matrix, $\boldsymbol{\beta}_q$ is a vector of coefficients which included estimable parameters for the intercept (α) and the group effect (γ) at two levels with 1 indicating a patient in GP group and 0 indicating a patient in PP group for our data. $\boldsymbol{\epsilon}$ is the vector of normal errors. In the context of the Mann mRNA dataset this linear model can be presented as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{47} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \gamma \end{pmatrix} + \boldsymbol{\epsilon}$$

2. Different classification algorithms

   Once the optimum number of features is determined, a classifier is built on the selected optimal features. The classifiers compared in this study were: logistic regression, SVM (support vector machine), diagonal linear discriminant analysis (DLDA), nearest shrunken centroids (NSC) and k-nearest neighbour (KNN) with $k = 1, 5, 10$.

3. Selection criteria of the penalty parameter

   In the internal CV loop where the optimum number of features is determined, a criterion should be utilised to compare the performance of different sets of features selected by the weighted Lasso. The criteria compared in this study were misclassification error rate (CV error), AIC (Akaike, 1974) and BIC (Schwarz, 1978).

4. Relative strength of weights

   The anti-correlation based weights (weights have an inverse relationship with correlation measures, such that when the correlation value is high the weights tend to be low) utilised in this study were further investigated. The purpose here was to compare whether the strength of the weights to the penalty parameter ($\lambda_q = w_q \times \lambda$) in the weighted Lasso feature selection, has any impact on the prediction error rates the models produced. The following functions of correlation coefficients were compared.

a) Inverse correlation: $w_q = \frac{1}{\rho_q}$

b) $w_q = \frac{1}{1+\rho_q}$

c) $w_q = \frac{1}{1+\sqrt{\rho_q}}$

d) $w_q = \frac{1}{1+\rho_q^2}$

e) $w_q = 1 - \rho_q$

5. Different forms of penalties

Since the introduction of Lasso feature selection, there have been many developments associated with the penalty function used. Elastic net (Zou and Hastie, 2005) is one such penalty, given by,

$$(1-\alpha)/2||\beta||_2^2 + \alpha||\beta||_1.$$

Elastic net is known to be particularly useful in $p \gg n$ situation, or when there are correlated predictor variables. It is a combination between the Ridge regression penalty (when $\alpha = 0$) and the Lasso penalty (when $\alpha = 1$) (Friedman *et al.*, 2010). More details are given in Section 3.2.2. In this comparison study, we compared the effect of different penalty functions on the CV error rate.

Results: It was not feasible to compare all parameters simultaneously, because of the higher computational time involved due to the high number of combinations. Therefore, in our comparison study the parameters were selected in a forward manner, where in each stage a certain parameter was compared while the others remained fixed. The results are presented here for the integration of mRNA (gene expression) and protein data (WL_GP integration method). Similar results were obtained for the other integration methods as well.

1. The comparison of two feature combinations using full set of genes and the subset of genes (comparison 1) suggested that combination 2 with subsets of both

types of variables performs better in terms of the prediction accuracy (Figure 5.6). Furthermore, DLDA classifier (comparison 2) performed best (Figure 5.6). These observations held consistently for all integration set-ups.

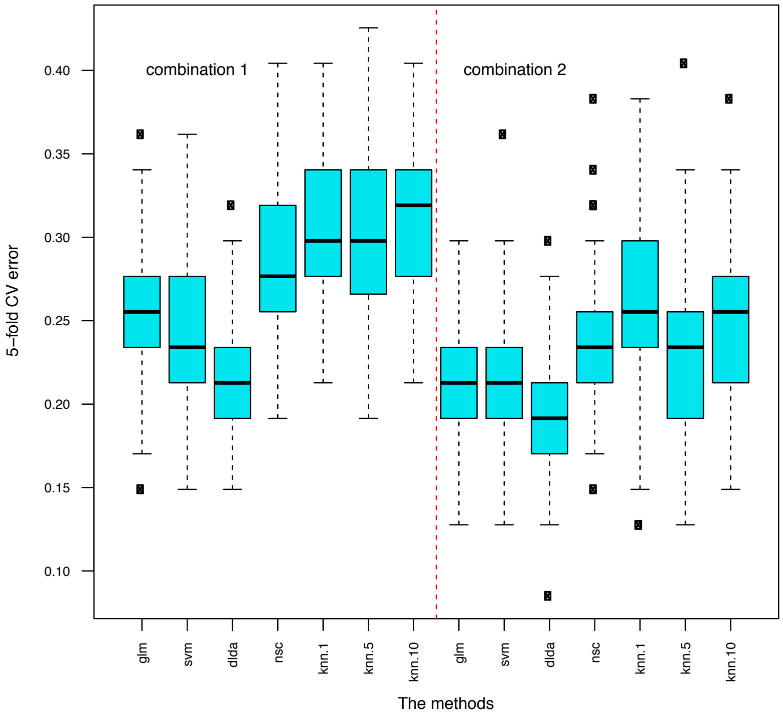


Figure 5.6: **Comparison of the feature combinations using different classification methods.** The prediction errors for the weighted Lasso gene selection in the integration of mRNA and protein data.

2. In the internal loop of the CV procedure ('Final CV'; Chapter 2), CV error, BIC and AIC were used as the selection criteria of the penalty parameter (λ) to determine the optimum number of features (comparison 3). AIC and CV (the misclassification error rate) performed similarly and both produced better prediction accuracy than BIC (Figure 5.7). Hence, CV was selected.

3. The comparison of different anti-correlation based weights (comparison 4) rendered that inverse correlation performed better, while the other weights showed much similar predictive performance (Figure 5.8).

4. The comparison between the Lasso and the Elastic net penalties (comparison 5) gave similar mean error rates and so the Lasso penalty was used for subsequent

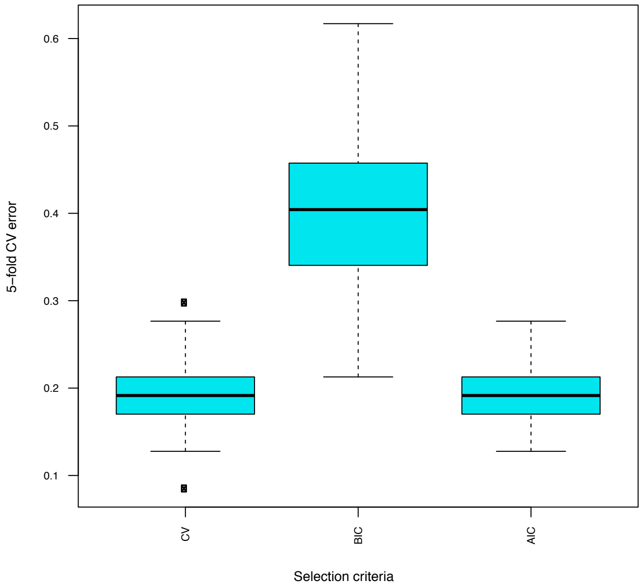


Figure 5.7: **Comparison of different model selection criteria.** Misclassification error (CV) and penalised likelihood methods (BIC and AIC) were compared in terms of prediction accuracy.

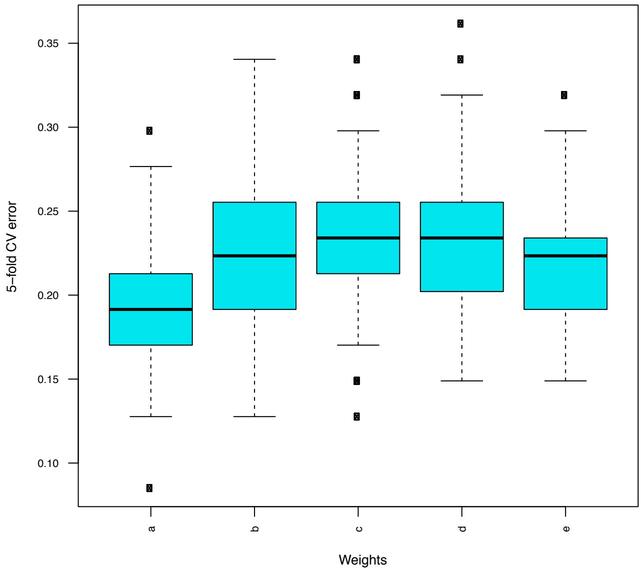


Figure 5.8: **Comparison of different forms of anti-correlation based weights.** The weights are, a: $w_q = \frac{1}{\rho_q}$, b: $w_q = \frac{1}{1+\rho_q}$, c: $w_q = \frac{1}{1+\sqrt{\rho_q}}$, d: $w_q = \frac{1}{1+\rho_q^2}$, e: $w_q = 1 - \rho_q$.

analysis.

Selected parameters: The parameters chosen to be used in the subsequent analysis were:

- Feature combination: Use subsets of variables (combination 2).

- Classification algorithm: DLDA classifier.
- Selection of optimal lambda/penalty parameter: Misclassification error rate (CV).
- Weights: Inverse correlation.
- Penalty in feature selection: Lasso.

5.3.5 *Performance evaluation*

We assessed the performance of our data integration framework to evaluate the efficacy of vertical data integration. The parameters selected in the comparison study (Section 5.3.4) was used in the CV procedure.

1. Prognostic capability

   It is of interest to see whether the incorporation of weights in the Lasso variable selection improves the prediction error rate, when compared to the standard Lasso.

   We observed that the prediction accuracy improved (lower 5-fold CV error) in all 3 integration settings (Figure 5.9). The lowest error rate was achieved when all 3 data platforms were integrated (mean 5-fold CV error rate of 17%), indicating that data integration helps in this setting.

2. Stability of the features

   To assess the stability of feature selection, the variables included in majority (more than 50%) of the models were considered.

   All 4 settings being compared consisted of 10 genes each, that had an inclusion frequency of 50%. However, the inclusion frequencies were slightly higher in all 3 integration set-ups when compared with the Lasso (Figure 5.10). That is, there were higher number of genes with inclusion frequency in the range 80-100% when the data were integrated.

3. Biological validation

   The melanoma gene cards (Rappaport *et al.*, 2013) consist of scores for genes

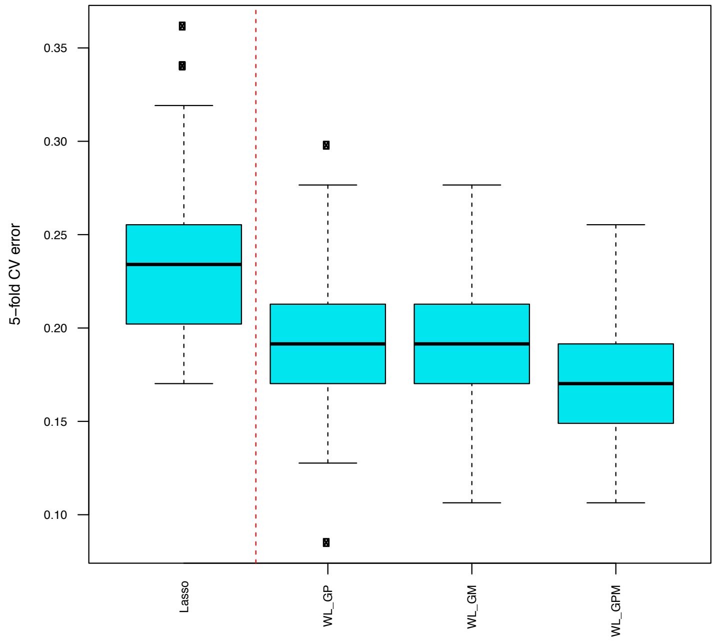


Figure 5.9: **Comparison of the Lasso with the weighted Lasso.** The evaluation of prediction performance when weights based on relevant external information (protein and miRNA data) are used to guide the variable selection. Here, the 3 data integration set-ups are compared with no integration ('Lasso') setting.

known to be related with melanoma progression (from literature and other studies). Therefore, these scores can be used to assess the biological relevance of the features selected as the more biologically relevant features will achieve a higher gene card score. These scores were downloaded from (http://www.malacards.org) and the cumulative scores for each of the 4 gene lists were calculated. The cumulative scores were used to assess the biological relevance of the genes selected via the Lasso (no integration, only using mRNA data) and the integration set-ups (WL_GP, WL_GM and WL_GPM). For this evaluation, we ranked the genes selected under each setting based on their inclusion frequency. The gene with the highest inclusion frequency was ranked first, and the cumulative gene card scores at different top number of genes were compared.

We observed that the cumulative score was higher for the 3 integration set-ups a majority of the time, when compared with no integration (Figure 5.11). The WL_GP, where mRNA and protein data were integrated, performed best and achieved higher scores than the Lasso throughout the entire range of gene sets.

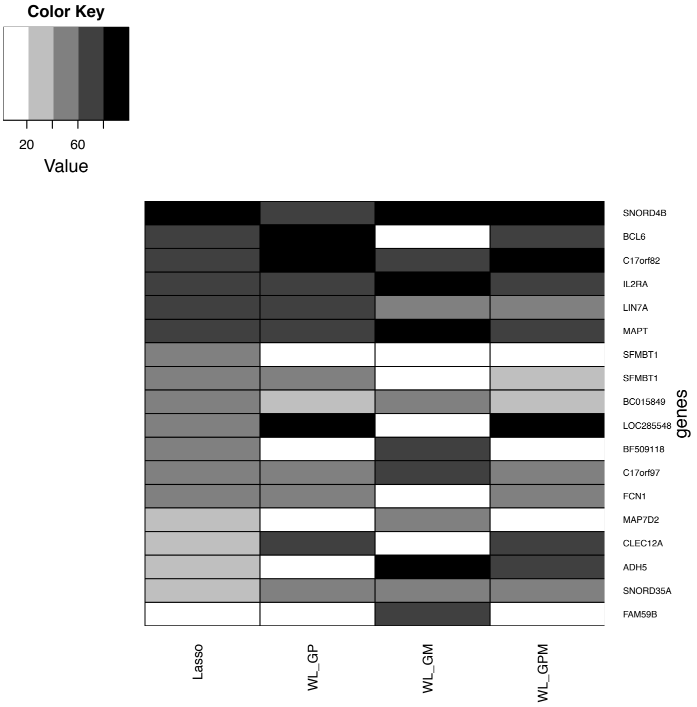


Figure 5.10: **Heatmap of inclusion frequencies.** The inclusion frequencies of the genes selected from the 4 set-ups compared. Each gene list consisted of genes included in more than 50% of the models under each set-up, and the union of genes from all 4 lists are presented.

This indicated that the genes selected under data integration produced more biologically relevant feature lists.

5.3.6 *Network based weights for data integration*

The weighted Lasso can be used innovatively in vertical data integration, as any meaningful weight from other data platforms could be adopted into this setting. Intuitive weights could be applied here from the background knowledge about the data being integrated or by consultation with biologists and clinicians proficient in the relevant field.

In this thesis, the main form of weight considered was the inverse correlation, where it was assumed that genes that are highly correlated with DE proteins/miRNAs are more likely to be relevant in survival outcome prediction. Hence such genes were given less weight. We also considered other forms of weights after discussion with the

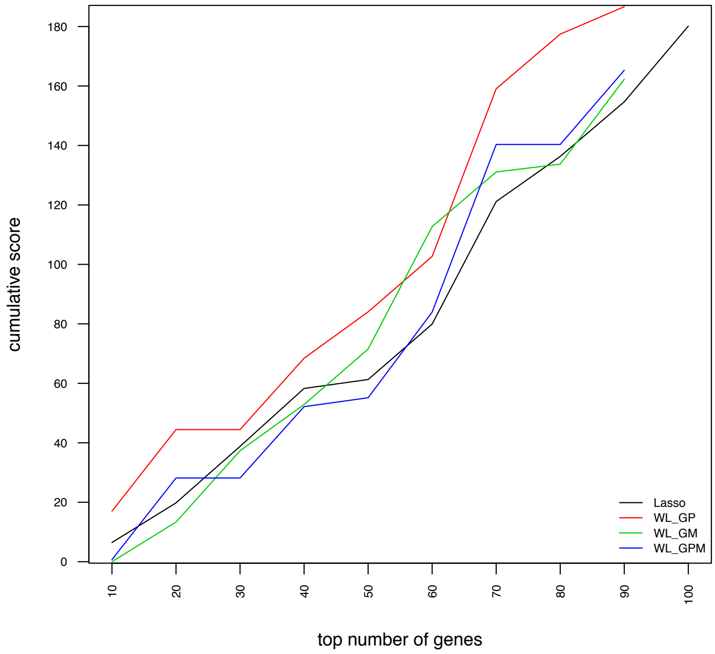


Figure 5.11: **The melanoma gene card scores.**The genes included in atleast one model (inclusion frequency$>$ 0%) was ranked according to their inclusion frequencies and the gene card scores were considered at different top number of genes.

biological collaborators. One such weight was the protein expression network based weights, and the results are presented in this section.

The protein expression network was obtained (Barter *et al.*, 2014) from the inverse covariance matrix of the protein dataset (using R package 'QUIC' (Hsieh *et al.*, 2011)). If there exists an edge between two proteins in the network, the inverse covariance matrix has a non zero entry. The degree (d_q) of a protein could be defined as the number of edges a protein has in the network. The proteins could be mapped to genes and hence a weight could be obtained for each gene.

The weights considered in this section were:

1. $w_q = d_q$
2. $w_q = \frac{1}{d_q}$

These two weights were computed using all of the proteins in the network (weights a and b), and using the subset of proteins with at least five edges (weights c and d).

We observed (Figure 5.12) that when mRNA and protein data are integrated via weights based on protein data (from protein expression network), the weight a produced much lower prediction error rate than when no weights are employed in gene selection. However, the other weights did not produce better accuracy compared to the no integration setting.

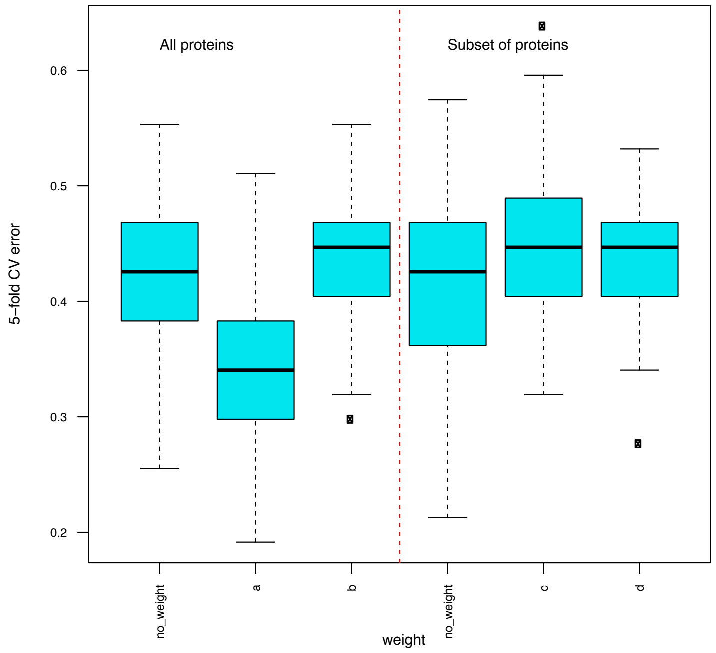


Figure 5.12: **Protein expression network based weights.** The 5-fold CV error rates when four different weights utilising protein-based information from protein expression networks are utilised, when compared with the no integration setting. Error rates are given when all proteins were considered (weights a and b) and a filtered subset of proteins (with no. of edges $\geqslant 5$) were considered (weights c and d).

5.4 VISUALISATION OF SAMPLE LEVEL PROGNOSTIC ACCURACY

In recent years, increasing attention has been drawn to personalised medicine (Scolyer and Thompson, 2013; Yu *et al.*, 2013; Sewell *et al.*, 2012; John *et al.*, 2008), where it was identified that treatments should be allocated with care to individual patients because of the heterogeneity of sample cohorts. Therefore, it is imperative that the prognostic capability of clinical and high-throughput data are evaluated at the patient level. To address this, we investigated the patient level accuracies of the prognostic models developed, to obtain a visualisation technique for identifying such heterogeneous subgroups of patients. Such a visualisation technique would enable the identification of most dominant sources of prognostic information for patient subgroups. Our CV protocol facilitated the evaluation of variability associated with the prediction each sample (patient) achieved, because it involved 100 runs of the 5-fold CV.

In the following graphical output of the sample level accuracy, the survival outcome of the patients (GP *vs.* PP) is shown as a sidebar. Patients are represented in the columns (tumour ID) and are ordered based on the overall extent to which they are misclassified (most to least). The different prognostic models under which each sample is classified are given in rows. The scale varies from 0 (all 100 runs of the model misclassified this sample) to 100 (all 100 runs of the model correctly classified this sample).

1. Sample plots for integration via pre-validation

   The Sample plots (Figure 5.13 and Figure 5.14) highlighted informative patterns of classification accuracy, where they illustrated the degree of concordance between the different prognostic models for the samples common among data types. Furthermore, the sample accuracy was compared with the AJCC standard-of-care variables as well.

   29% of the 24 samples common among all data types appeared more likely to be classified correctly under integration setting 1 compared to integration setting 2 (Figure 5.13). For other samples, the reverse was true. In addition, the classification accuracy of one sample (ID: 343) was markedly improved in most cases where both the protein information and clinical variables were present in the model. In this instance, almost all other classifiers performed poorly (<25% accuracy) (Figure 5.13).

   In the 45 samples for which clinical, mRNA, and miRNA information were available, 15% were correctly classified in more than 75% of the runs within any model (excluding the standard-of-care variables). There were no samples for which miRNA information alone outperformed other data types or combinations. One sample, ID: 350, for which there were no protein data available, was incorrectly classified more than 75% of the time irrespective of which data, or combination thereof, were used. Specifically, this sample was rarely identified as being derived from a patient with a good outcome (GP: survival > 4yr post resection of metastatic disease with no sign of relapse) despite the patient follow up status being alive with no sign of relapse at 2948 days (8+ years) post-surgical resection.

This investigation revealed valuable information, especially to clinicians, regarding the differences in prognostic capability of various data sources at patient level. It indicated that although some data sources such as protein data had poor predictive performance overall, for certain groups of patients they contain important information that could assist in their individualised treatments. Hence, the overall predictive capability does not always make the data sources redundant.

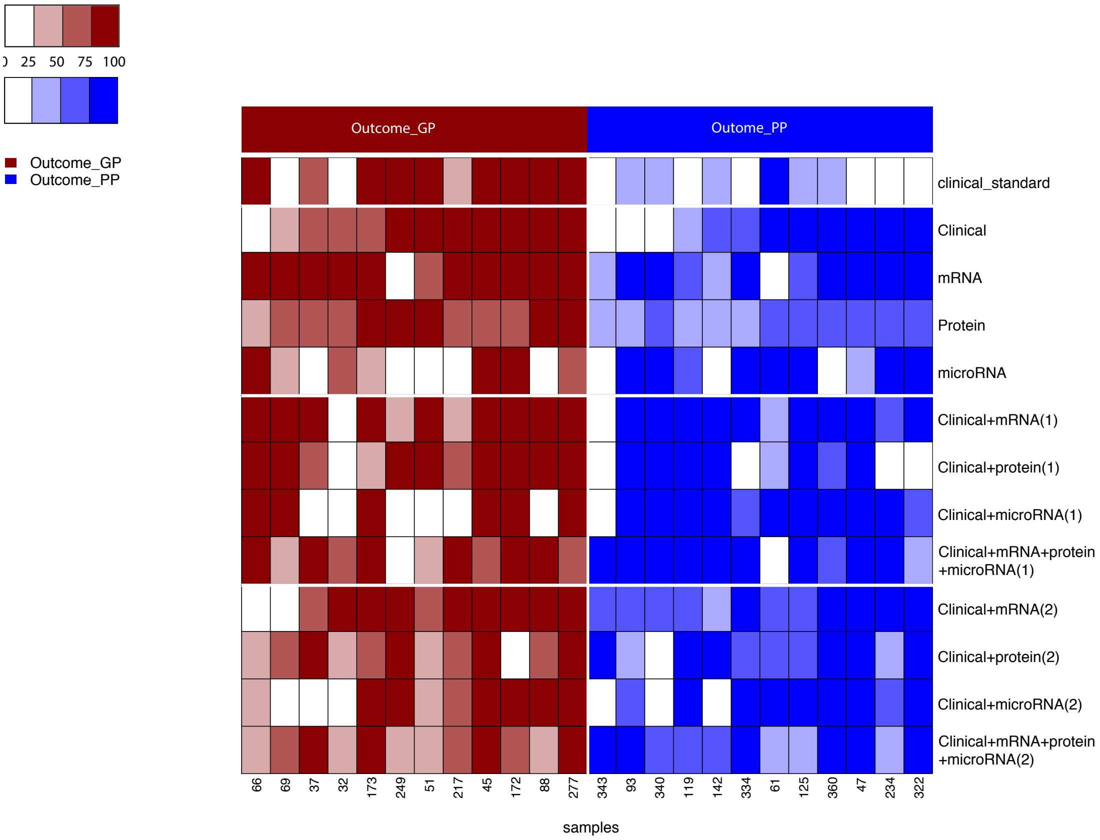


Figure 5.13: **Patterns of sample misclassification for 24 patients.** Investigating whether there was any consistency among the different prognostic models with respect to the misclassification of the 24 samples common across all platforms.

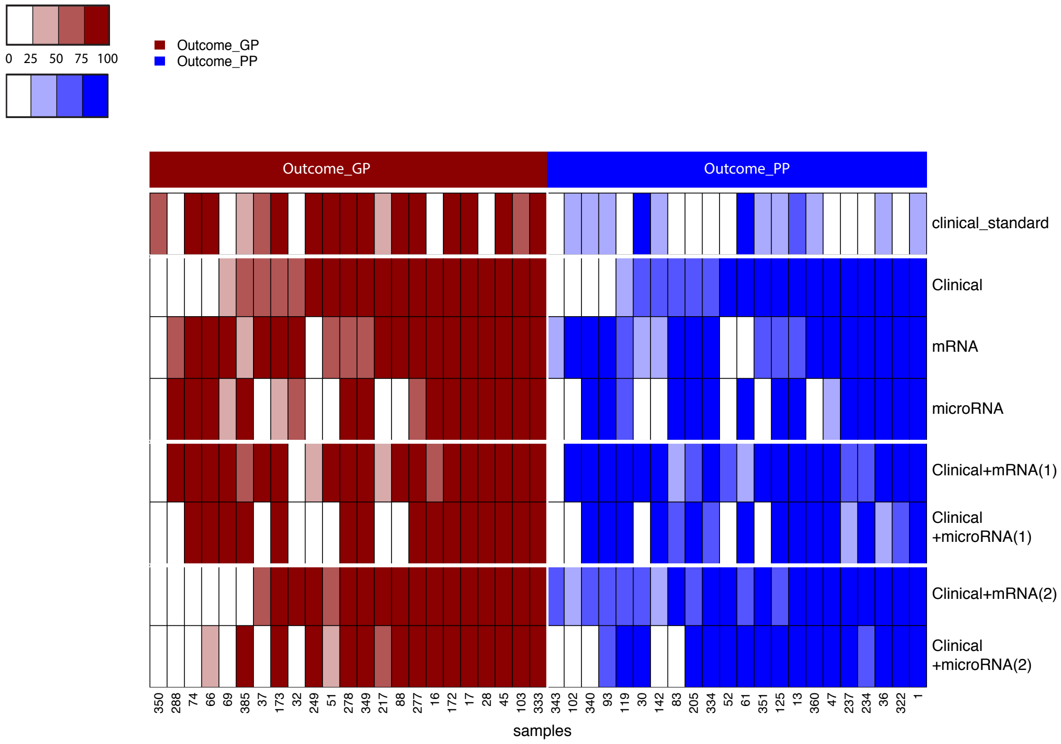


Figure 5.14: **Patterns of sample misclassification for 45 patients.** Investigating for consistencies among the different prognostic models generated for 45 samples evaluated in all approaches excluding protein assays.

2. Sample plots for integration via the weighted Lasso

   Similar to the sample plots for integration via pre-validation, here we considered whether integration adds value in predicting the individual patients, who did not achieve accurate predictions when only individual data sources were considered.

   We observed that while for most patients the data platforms performed similarly, there were some patients for whom the data integration improved the prediction performance (Tumour IDs: 249, 217, 396, 61, 343, 36) (Figure 5.15). Therefore, this is a clear indication that from the individual patient perspective, data integration is useful.

3. Validation in independent data

   This validation was performed to affirm the findings from the original study using the Mann data, that different platforms perform better for different subsets of patients. Hence, the subset of samples from The Cancer Genome Atlas (TCGA) with the necessary clinico-pathologic annotation (Section 2.2.3), was used to eval-

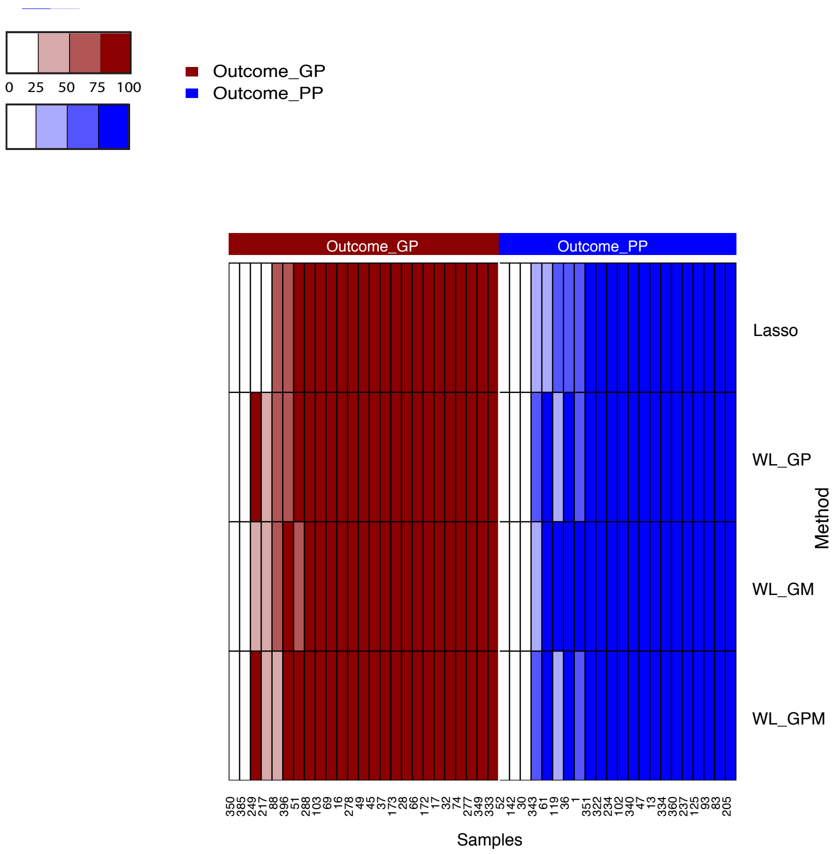


Figure 5.15: **Patterns of sample misclassification.** The comparison of patient level prediction accuracies for the Lasso (no integration) and the 3 integration settings using the weighted Lasso.

uate the potential benefit of using multiple data sources rather than a single data source for predicting the survival outcome of individual patients. Prognostic modelling was performed as described in Section 5.1.2 (high-throughput data modelling).

Principal observations are supported by validation in an independent cohort; see Figure 5.16, which illustrates the degree of concordance between the different prognostic models for the 27 samples evaluated. Consistent with the original findings, both mRNA and miRNA data perform well for a subset of patients. For other groups and/or individuals, mRNA is the more accurate prognostic indicator compared to miRNA, and vice versa (Figure 5.16A). There are approximately 9 samples (33%) that fall off the diagonal, which further clarifies that for these patients one data platform out-performed the other (Figure 5.16B). Although no data integration was performed on this data, this observation further implicates the potential utility of multiple data platforms rather than one, when predicting the survival outcome of individual patients.

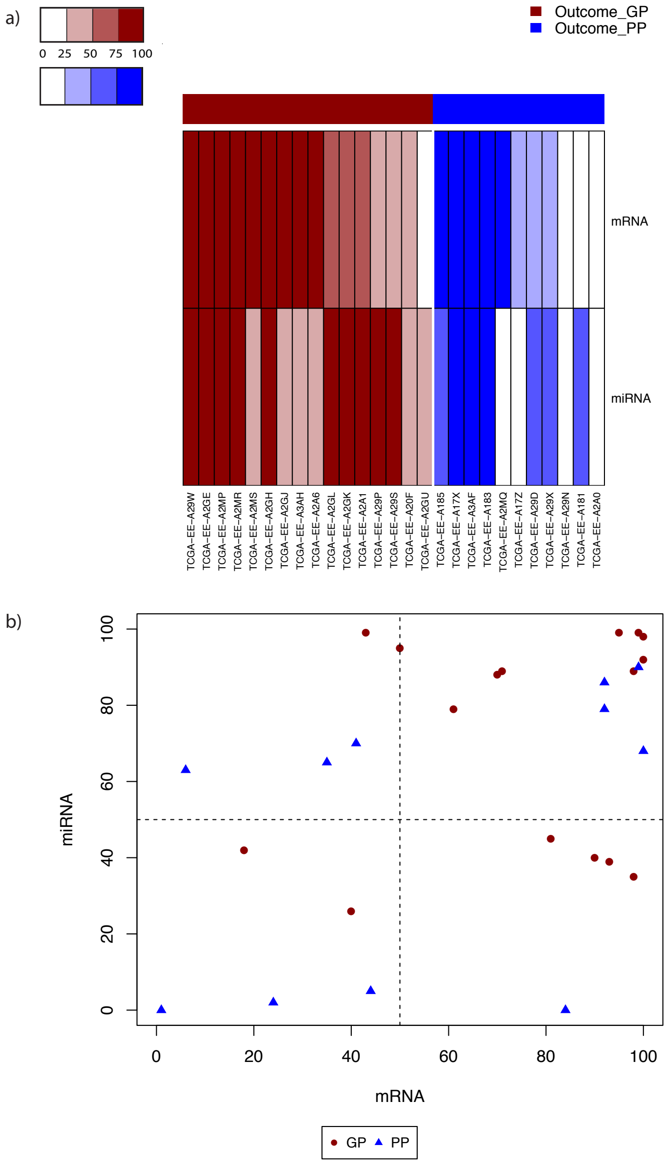


Figure 5.16: **Patterns of sample misclassification among the different prognostic models generated using TCGA data.** a) the heat map of the values. Patients are represented in the columns and are ordered based on the overall extent to which they are misclassified (least to most). b) the scatter plot of the values. The scale varies from 0 (all 100 runs of the model misclassified this sample) to 100 (all 100 runs of the model correctly classified this sample).

5.5 DISCUSSION AND BIOLOGICAL IMPLICATIONS

In this chapter we proposed vertical data integration frameworks, extending the principle of pre-validation and the weighted Lasso. The proposed frameworks were developed using multiple sources of data on the same set of patients, to address the clinically important question of predicting survival outcomes for patients with metastatic melanoma. The frameworks produced results that highlighted the advantages of using multiple data types, while dealing with the challenges encountered favourably in real biological data.

The use of the principle of pre-validation in a data integration framework proved to be an effective method, as it is capable of addressing the intrinsic structural disparity of the different data types. It provides a successful way of handling a large number of variables in high-throughput data sources, by reducing thousands of omics variables to a single pre-validated vector. This is an extreme variable reduction, however, the simplicity of the approach and the ease of its adaptability to use more than two data types makes it an extremely important approach to be used in vertical data integration.

Our framework is capable of handling the integration of multiple omics data types with clinical data, using two data integration settings (with and without variable selection). Most studies in the literature used the pre-validation principle to integrate two data sources only (Segura *et al.*, 2010; Höfling and Tibshirani, 2008; Tibshirani and Efron, 2002) and mostly to compare the prediction accuracy of data sources in separate settings (van Vliet *et al.*, 2012; Bogunovic *et al.*, 2009) rather than integrating the information directly as in our framework. Our framework also allows to choose variables from among platforms inside the integration framework with the mBMI (setting 1). This provides a much more sensible way of variable selection from among clinical and omics data, because the thousands of omics variables are now reduced to few variables, and therefore are in a comparative scale with the clinical data.

Many other complex statistical methods are also available in the literature (discussed in Chapter 1). However, it is of note that complex methods do not automatically yield better results in real data (Campain, 2012; Dudoit *et al.*, 2002).

The results from our vertical data integration using pre-validated vectors have many biological implications, in particular in the present context of metastatic melanoma. In the modelling of high-throughput data, only the mRNA signature information outperformed the more expansive model of clinical variables. This observation appears to suggest that the most valuable prognostic information is contained within the mRNA transcript component of the tumour. However, another explanation may be that information can be extracted more easily via gene (mRNA) expression microarray platforms than with the less well developed miRNA and proteome-based technologies. Ongoing examinations of the contribution of miRNA and proteome-based work to improving biomarkers of prognosis in melanoma are essential work (explored in relation to miRNA data in Chapter 4). Furthermore, the features identified were shown to be biologically relevant in (Jayawardana *et al.*, 2015a). This integrated analysis of multiple data types using the same tumour specimens (samples), is the first of its kind in melanoma and among the earliest in any cancer (Jayawardana *et al.*, 2015a).

One of the main challenges we faced in validating the framework in independent biological data was the limited availability of high quality specimens with linked, well-annotated clinical data. Therefore, we chose to use the extended cohort of the Mann data to provide a comparable, although not completely independent, validation of the framework. This dearth of independent validation data is an ongoing challenge in many diseases (including cancers) and systematic efforts to meet this critical requirement are underway (Scolyer and Thompson, 2013).

The weighted Lasso approach also proved to be an effective method to be used in a vertical data integration framework. The ability to integrate secondary data types via weights provides an excellent opportunity to prevent a further increase in the number of variables in an already large p small n paradigm. Furthermore, this method allows to treat one platform as the primary platform and maintain a hierarchical structure in the data integration process. Our framework assumed that mRNA data is the primary platform and the protein and miRNA data were integrated indirectly via weights. The data integration also relied on the assumption that the genes highly correlated with the proteins and miRNAs are more likely to be relevant in survival outcome prediction, and were given less weight. The use of subsets of the secondary data platforms (protein and miRNA) chosen from a DE analysis, aimed to ensure that the external information we used were relevant in the prediction of outcome.

A particular advantage of the data integration using weights is that these weights can be selected by incorporating prior knowledge from biologists. Our choice of correlation based and network based weights was inspired by such discussions with the biologists, and hence the results are more inclined to be biologically relevant.

Our framework using the platform dependent weights is flexible enough to integrate clinical information or other variables. We propose a two stage feature selection approach which focuses on an intermediate variable reduction, so that the mRNA data and clinical data has an equal standing. In the first stage, we select the genes based on protein and miRNA information using the proposed data integration framework (Section 5.3.2) and in the second stage we integrate the selected genes and clinical variables to find the final predictive model. More details are shown in (Jayawardana *et al.*, 2013).

In all frameworks of this chapter, the variability associated with the final models was considered. The mean error rates over 100 runs were used to indicate the prediction error, instead of a single value for error that ignores the variability in real data. This CV protocol also enabled us to propose a novel visualisation technique in Section 5.4, to investigate the prediction accuracy at patient level. This graphical outcome is of high significance in the recently emerging area of personalised medicine (Scolyer and Thompson, 2013; Yu *et al.*, 2013; Sewell *et al.*, 2012), which allows to identify dominant sources of prognostic information for heterogeneous subsets of patients.

Conclusion

In conclusion, all of the proposed data integration frameworks showed that vertical data integration is beneficial, illustrated through the application in real biological data. The use of the pre-validation principle in integration showed that the integration of clinical and mRNA information held the most potential in assisting patient prognosis. Our framework using the weighted Lasso also resulted in improved prediction error rate (17% compared to 24% in no integration setting), indicating the potential value of the multi-layered omics data. Both frameworks exhibited biological relevance in terms of the features selected using multi-platform data. Furthermore, the proposed graphical visualisation tool of the patient level accuracies revealed the potential utility of multiple data sets, rather than a single data source in predicting the survival outcome of individual patients. Importantly, these observations were supported by validation in an independent cohort of samples from TCGA. Our investigation at the patient level is of high relevance in the medical context, as the identification of heterogeneous subsets of patients and the data sources that are most relevant in predicting their outcome is imperative in personalised medicine. This identification eventually aids in assigning the patients to the most appropriate treatment option (John *et al.*, 2008; van't Veer and Bernards, 2008).

6

DISCUSSION AND FUTURE RESEARCH

Developing statistical methodologies that are applicable in real life medical data is a major challenge in statistical bioinformatics. The surge in the amount of publicly available high-throughput data has encouraged researchers to integrate these with in-house experimental data in order to seek answers for a myriad of biological and medical questions. However, statistical research faces numerous challenges in the high-throughput data setting, such as having large number of variables than samples, commonly referred to as large p small n problem, platform differences and inconsistencies among studies. The development and validation of statistical methods addressing these challenges remains a necessity to bridge the gap between increased data availability and possible elucidations to biological queries. This thesis contributed to the development of such statistical methodologies and frameworks for real biological problems. We achieved this purpose through detailed analysis at individual platform (data type) level and through the integration of multiple datasets.

This thesis was motivated by the Mann data, which allowed us to exploit the availability of a detailed series of clinical data and omics data (mRNA, protein and miRNA data) with matched samples in metastatic melanoma (Chapter 2). Through this dataset and other external data of melanoma patients, we explored and extended methods that have translational impact in melanoma research. Using real data allowed us to observe and demonstrate the applicability and impact of the developed frameworks in the medical context. More sophisticated statistical methodologies, although rich in their theoretical formulation, may not perform well in real data. This observation, that the complex methods may not necessarily out-perform simple methods in real data analysis, has been previously perceived in bioinformatics (Dudoit *et al.*, 2002).

The major contributions in this thesis could be broadly categorised into three distinct areas: the development of statistical methodologies and frameworks, (i) at the individual platform level, (ii) for horizontal data integration and (iii) for vertical data integration. In the following we provide further discussions on the chapters of this thesis, highlighting the contributions and future directions.

One of the key challenges in developing a statistical method in medical research is the evaluation of methods developed and the demonstration that they have biological impact. While the datasets are increasingly becoming publicly accessible, the difficulty in finding high quality independent data that closely resemble the training data, hinder such validation that will justify their applicability in a medical context. As a result, in this thesis we used evaluation approaches based on cross-validation. In the literature different forms of CV procedures have been used to evaluate the results in the absence of independent data (Browne, 2000; Burman, 1989). These procedures concentrate on different aspects of validation, such as the feature selection and/or classification, and it is important that we use comparable CV approaches, when the aim is to compare between methods. However, a comparison study is warranted to assess the advantages and disadvantages. Therefore, in Chapter 2 we conducted a comparison study of the many forms of CV used in this thesis.

The development of a framework to address missing data, model instability and predictive capability of constructed models is an important concept in clinical data analysis. The mBMI framework (Jayawardana *et al.*, 2015a) proposed in Chapter 3 especially focuses on constructing a predictive model, while addressing the more common problems of missing data and model instability via the utilisation of multiple imputation and the bootstrap. We showed that the novel framework proposed is an effective method in dealing with these obstacles and demonstrated that it results in the best predictive performance in real data. The simulation study in Chapter 3 exhibited the parameters within our framework that could be further investigated and optimised. A more detailed simulation study based on real data, that proved to be challenging, is warranted for a clearer demonstration. Future work would involve such a simulation study and also disseminating the R package for mBMI, that is currently under construction.

The integration of information from multiple data sources long held the promise of improved solutions to biological questions (Kim *et al.*, 2014; Tseng *et al.*, 2012; Hamid *et al.*, 2009; Hanash, 2004), however is often statistically challenging because of the large number of variables and the incompatibilities between platforms. In Chapters 4 and 5, we investigated data integration, exploring and addressing the challenges and highlighting the advantages in the horizontal and the vertical data integration paradigm.

In Chapter 4, we proposed a framework to integrate data of the same type in a horizontal data integration context. In the proposed framework, meta-analysis was used as a validation tool to validate biomarkers in independent data. Our comprehensive approach involved three main stages that are imperative in addressing the challenges encountered and highlighting the importance of such an integrative analysis. Firstly, careful pre-processing at individual platform level enabled to obtain the optimal signal, while minimising unwanted sources of variation. In the next step, we identified novel biomarkers from two publicly available datasets, of which one performed better than the existing standard-of-care variables. The pivotal step of meta-analysis exposed the many inconsistencies among biomarkers that have been proposed to address similar biological questions, while highlighting the potential translational value of the biomarkers in melanoma research. Our meta-analysis procedure demonstrated the utility of many external data sources pertaining to the same disease, in validating the existing biomarkers to be of clinical use and in identifying a more robust set of biomarkers. The analysis also pointed out future directions on the features where more extensive validations and investigations should be immediately prioritised in the field of melanoma. This work is also presented in (Jayawardana *et al.*, 2015b).

We extended the method of pre-validation and proposed a novel veritical data integration framework in Chapter 5. Through our analysis we demonstrated that the use of multiple types of data, such as clinical and omics data, can be leveraged to improve upon the prediction accuracy of disease prognosis and to identify among those the dominant sources of prognostic information. Such discoveries hold the potential in significantly reducing the cost of data generation and to enable the researchers to focus more specifically on those identified sources. Furthermore, through the incorporation of platform specific weights we explored methods in data integration extending the weighted Lasso. Our data integration, combined with the CV procedure we used, enabled the visualisation of sample level prediction accuracy. Through this, we demonstrated for the first time that there exist subsets of samples which could benefit from more dynamic and complex translational models. This inspired the ongoing work by colleagues (Patrick *et al.*, 2015), to investigate a multi-stage classifier, where different combinations of samples benefit from a different data type. Some of the results from this chapter have been published in (Jayawardana *et al.*, 2015a).

In summary, despite the many challenges encountered, our proposed statistical procedures demonstrated their validity in real biological data and are able to deliver revelations with translational impact.

A

SUPPLEMENTARY MATERIAL FOR CHAPTER 2

A.1 SURVIVAL CLASSES FOR TCGA DATA

We considered different criteria to determine the survival groups (GP *vs.* PP) for the TCGA cohort used in the analysis of Chapter 4 (introduced in Section 2.2.3: miRNA data). The final selection was option 7 (12 in poor prognosis group and 11 in good prognosis group). This selection was guided by the sample sizes in the two survival groups and was made after discussions with biologists in Professor Mann's group (Table A.1).

Table A.1: **The different samples sizes for various survival-based splits of the data from TCGA** – This table shows the different criteria considered prior to the final selection of survival groups from data from TCGA. Option 7 was ultimately used since it provided the best 'trade-off' between sample size (including between-class sample size balance) and relatability to the good and poor prognosis classes of the Mann miRNA data that was considered in Chapter 4. Abbreviations: ANSR, alive no sign of relapse; DM, dead melanoma; FU, follow-up.

	Poor prognosis group	n_{pp}	Good prognosis group	n_{pp}
1	Survival time $\leqslant$ 1yr (DM)	8	Survival time $\geqslant$ 2yrs (ANSR)	16
2	$\leqslant$ 1yr (DM)	8	$\geqslant$ 3yrs (ANSR)	11
3	$\leqslant$1yr (DM)	8	$\geqslant$ 4yrs (ANSR)	5
4	<2yrs (DM)	12	$\geqslant$ 2yrs (ANSR)	16
5	<2yrs (Dead, any cause)	14	$\geqslant$ 2yrs (Alive, any FU status)	18
6	<2yrs (DM)	12	$\geqslant$ 2yrs (Alive)	18
7	<2yrs (DM)	12	$\geqslant$ 3yrs (ANSR)	11
8	<2yrs (DM)	12	$\geqslant$ 4yrs (ANSR)	5
9	<3yrs (DM)	14	$\geqslant$ 4yrs (ANSR)	5
10	<4yrs (DM)	17	$\geqslant$ 4yrs (ANSR)	5

A.2 COMPARISON STUDY OF THE CV PROCEDURES

A.2.1 *Inclusion frequencies of variables*

We considered the variables included at least once in the repeated runs of the CV procedures. The variable inclusion frequencies of the procedures A–D showed a similar pattern (Figure A.1) when considering the variables with inclusion frequencies > 0%.

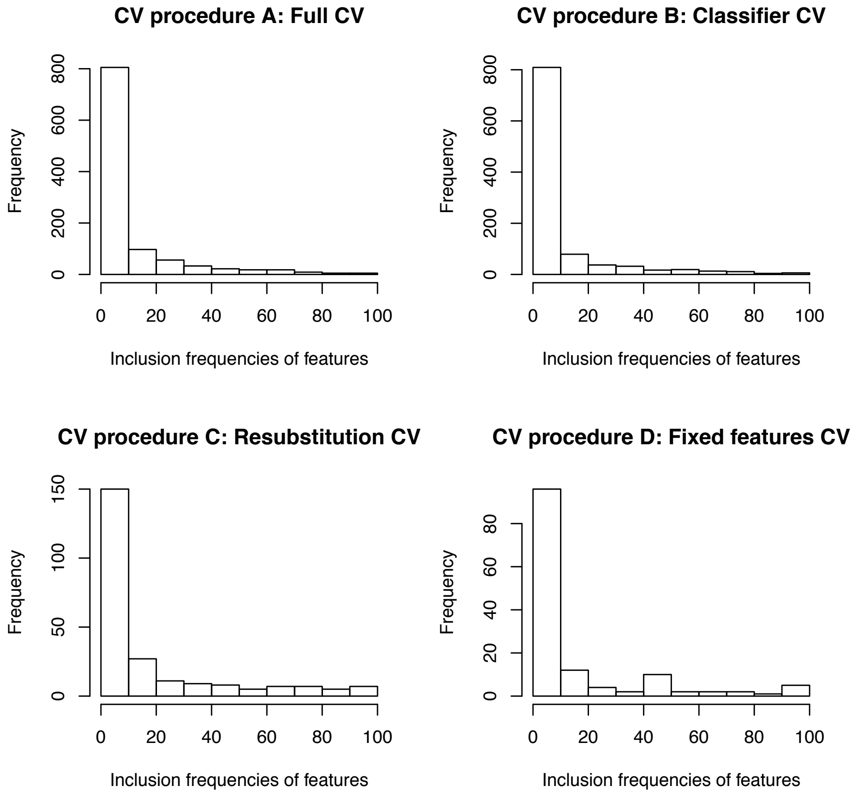


Figure A.1: **Comparison of CV procedures: Gene inclusion frequencies**

The 'FinalCV' procedure is not compared because the variables selected in each run was fixed.

B

SUPPLEMENTARY MATERIAL FOR CHAPTER 3

B.1 ADDITIONAL FIGURES FROM THE SIMULATION STUDY

The CV error rates for all τ_{FM} (final inclusion frequency threshold: $0.00, 0.05, 0.10, \ldots,$ 1.00) at each τ_{CV} (prediction error threshold: $0.025, 0.05, 0.10, 0.20, \ldots, 0.50$) considered are shown in Figure B.1. No clear identifiable pattern can be seen throughout the range of considered τ_{CV}.

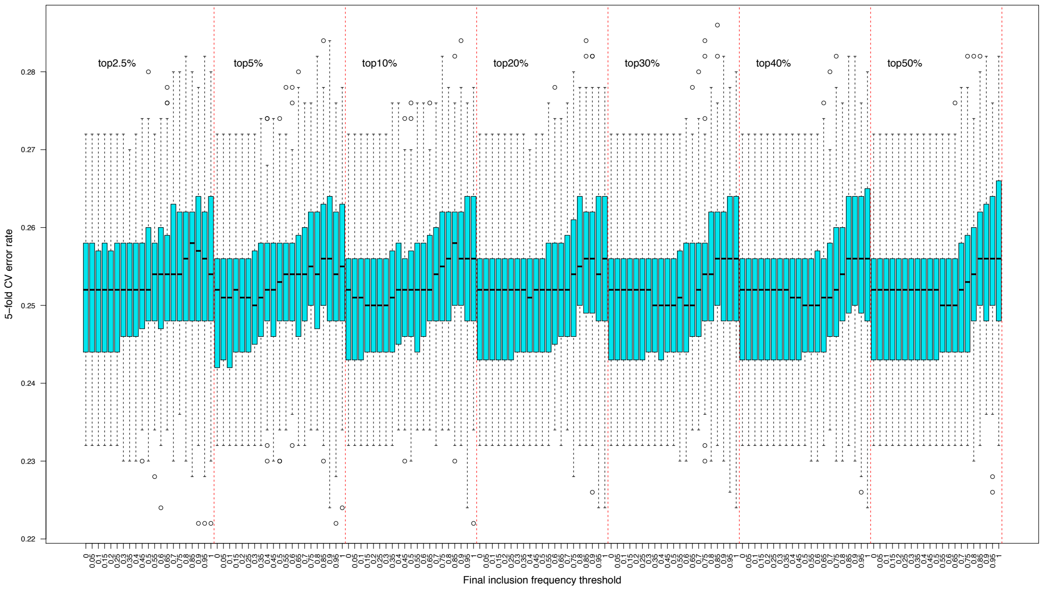


Figure B.1: **Prediction error rates.** The 5-fold CV error rates for the τ_{CV} and τ_{FM} (x-axis) considered in this study.

To select the τ_{FM}, the inclusion frequencies of variables for all τ_{CV} are considered. One instance was illustrated in Figure 3.3 in Chapter 3. The remaining instances are illustrated in Figures B.2-B.7. The τ_{FM} threshold was selected based on the separation between true variables and other variables from the inclusion frequency plots, as follows:

- When $\tau_{CV} = 0.05$, $\tau_{FM} = 0.75$
- When $\tau_{CV} = 0.10$, $\tau_{FM} = 0.80$
- When $\tau_{CV} = 0.20$, $\tau_{FM} = 0.85$
- When $\tau_{CV} = 0.30$, $\tau_{FM} = 0.90$
- When $\tau_{CV} = 0.40$, $\tau_{FM} = 0.90$
- When $\tau_{CV} = 0.50$, $\tau_{FM} = 0.90$

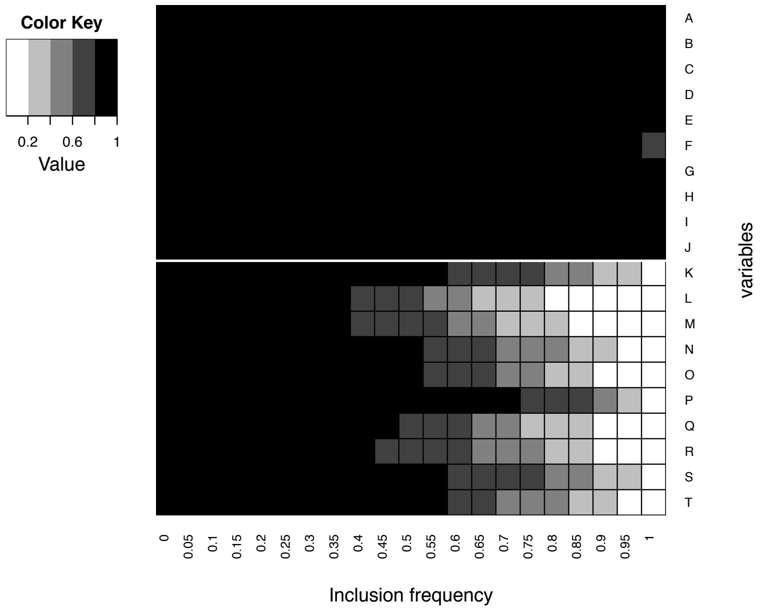


Figure B.2: **Stability of the variables.** The stability measures of the variables when $\tau_{CV} = 0.05$. The rows represent the variables and the columns represent different τ_{FM} values. The color indicates the stability, where darkness is proportional to the inclusion frequency.

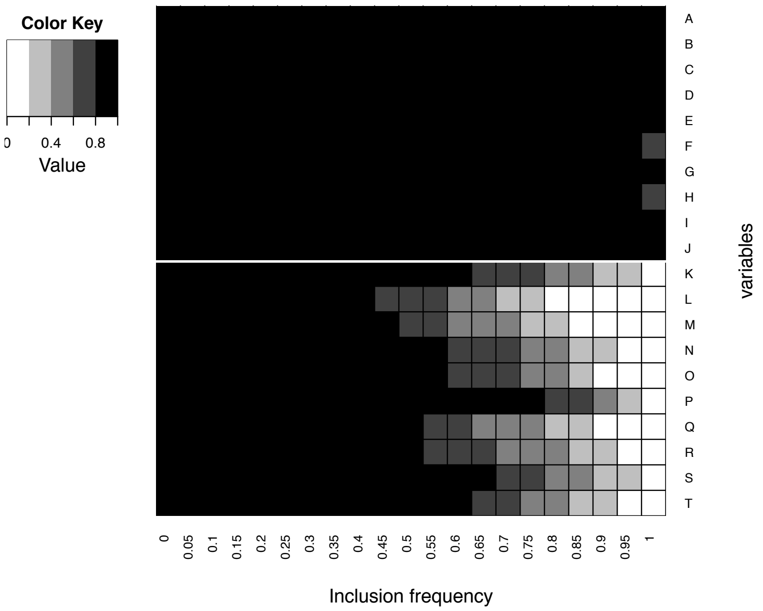


Figure B.3: **Stability of the variables.** The stability measures of the variables when $\tau_{CV} = 0.10$. The rows represent the variables and the columns represent different τ_{FM} values. The color indicates the stability, where darkness is proportional to the inclusion frequency.

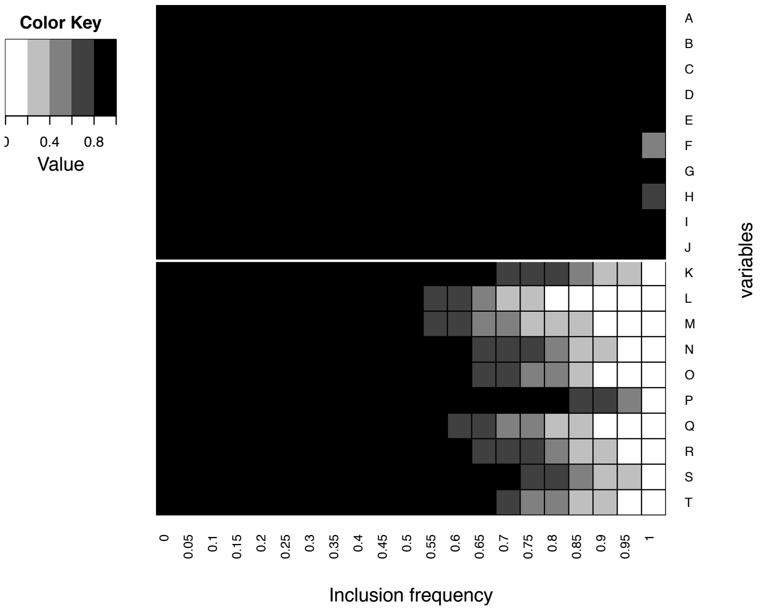


Figure B.4: **Stability of the variables.** The stability measures of the variables when $\tau_{CV} = 0.20$. The rows represent the variables and the columns represent different τ_{FM} values. The color indicates the stability, where darkness is proportional to the inclusion frequency.

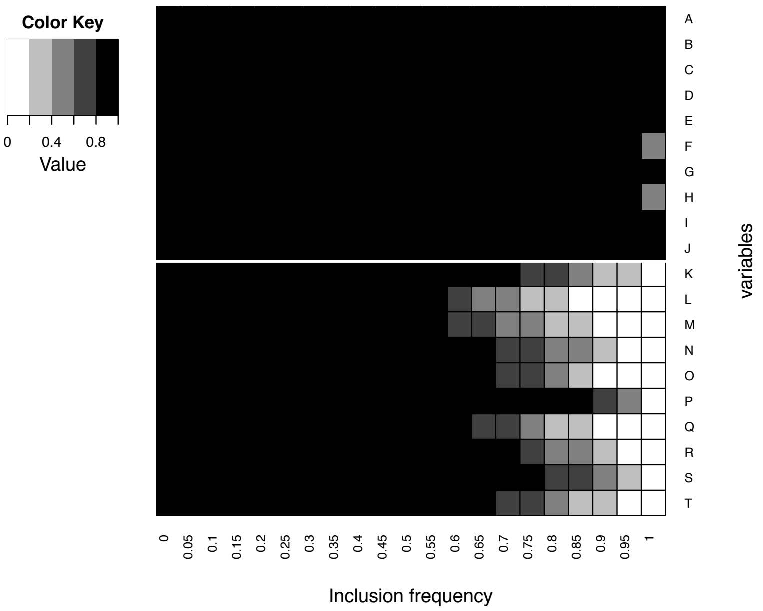


Figure B.5: **Stability of the variables.** The stability measures of the variables when $\tau_{CV} = 0.30$. The rows represent the variables and the columns represent different τ_{FM} values. The color indicates the stability, where darkness is proportional to the inclusion frequency.

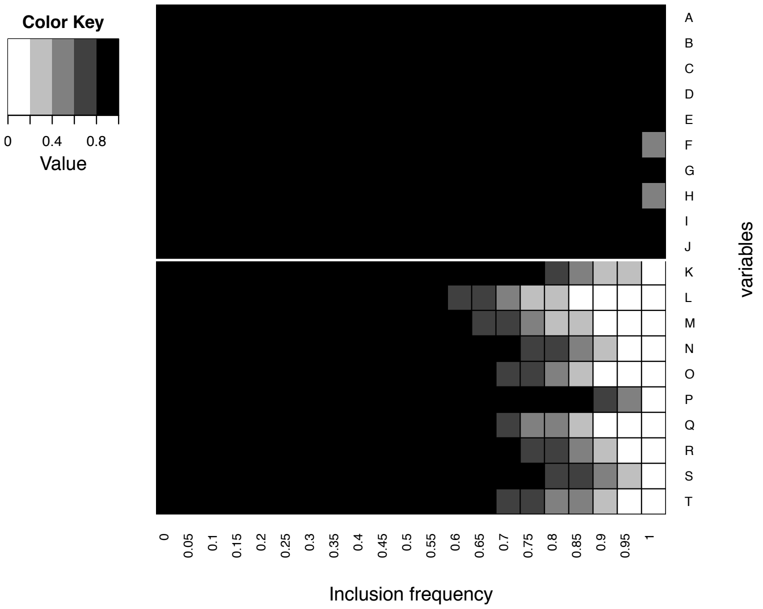


Figure B.6: **Stability of the variables.** The stability measures of the variables when $\tau_{CV} = 0.40$. The rows represent the variables and the columns represent different τ_{FM} values. The color indicates the stability, where darkness is proportional to the inclusion frequency.

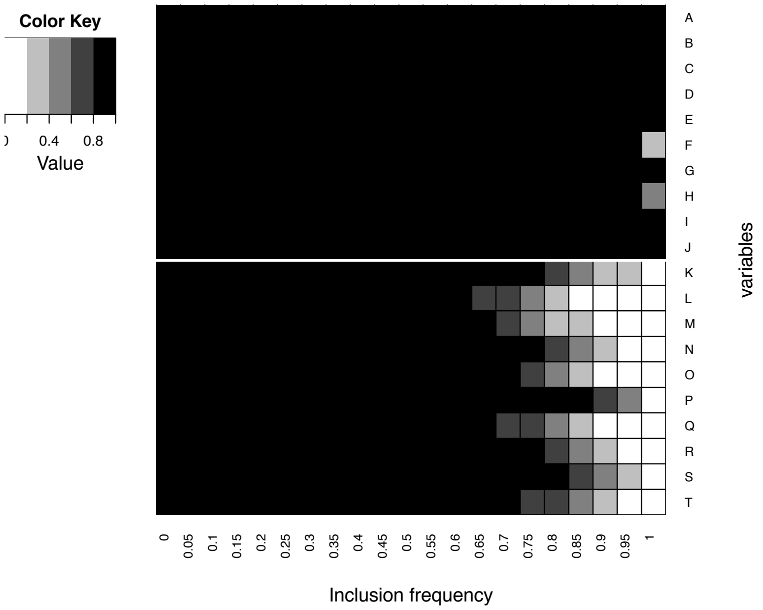


Figure B.7: **Stability of the variables.** The stability measures of the variables when $\tau_{CV} = 0.50$. The rows represent the variables and the columns represent different τ_{FM} values. The color indicates the stability, where darkness is proportional to the inclusion frequency.

C

SUPPLEMENTARY MATERIAL FOR CHAPTER 4

C.1 additional information for section 4.1

C.1.1 *Normalization methods: randomRUV*

The following R code for randomRUV was used for normalising the Mann miRNA data.

```r
naiveRandRuv <- function(Y, cIdx, nuCoeff=1e-3, k=m){

  ## W is the square root of the empirical covariance on the control genes.
  svdYc <- svd(Y[, cIdx])
  W <- svdYc$u[, 1:k] %*% diag(svdYc$d[1:k])

  ## Regularization heuristic: nu is a fraction of the largest eigenvalue
  of WW'
  nu <- nuCoeff*svdYc$d[1]^2

  ## Naive correction: ridge regression of Y against W
  nY <- Y - W %*% solve(t(W)%*%W + nu*diag(k), t(W) %*% Y)

  return(nY)
}
```

C.1.2 *Comparison of normalization methods*

Mixed model to find the number of DE miRNAs

In the analysis of the Mann data, a linear mixed effects model was used to find the number of DE miRNAs, to take into account the variability due to probes and replicates for each miRNA (every miRNA in this dataset corresponded to one or more probes and each probe had one or more technical replicates). The Figure C.1 shows that although the expression values of the samples for technical replicates do not show much variation between them, the values for the different probes show significant variation between the probes (indicated by the grouping of expression values of each sample into 1-4 groups in Figure C.1). This illustrates the necessity of accounting for the variability between the different probes when modelling the miRNA expression data.

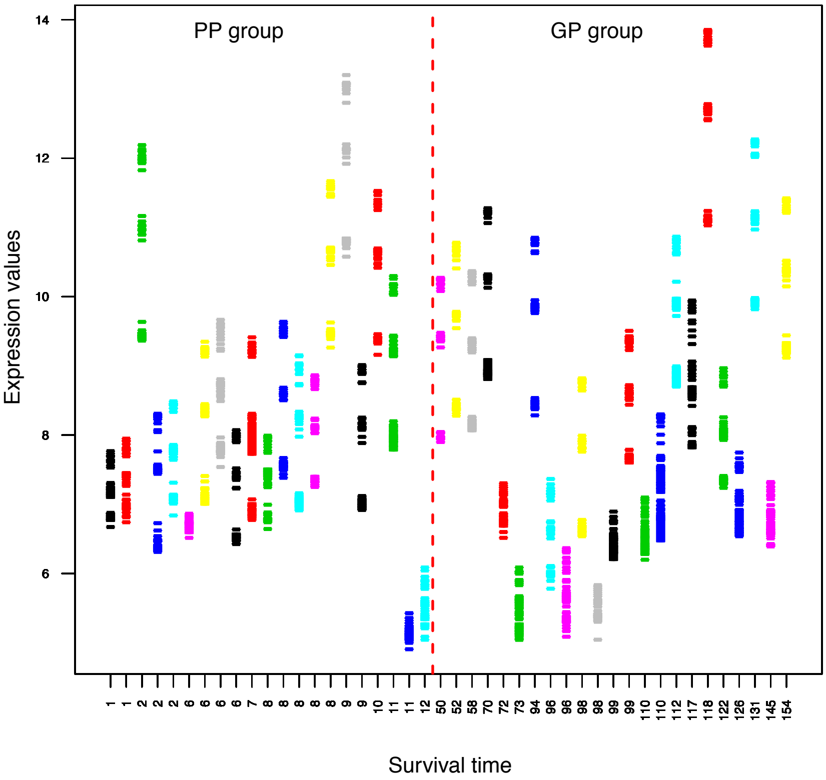


Figure C.1: **The expression values for the 45 samples of miRNA 'hsa-miR-150'.** The boxplots for the expression values of hsa-miR-150. The x-axis shows the survival times for the 45 samples (ordered by the survival times) and the y-axis shows the expression values.

Correlation with qRT-PCR data

The comparison of the normalised Mann miRNA data with the qRT-PCR data using the Pearson correlation coefficient showed high consistency with the QRUV-normalized data (using different parameters), where the highest absolute value of the correlation coefficient was exhibited for the majority of the miRNAs (Table C.1).

Table C.1: **Correlation of the Mann and qRT-PCR expression information.** The absolute value of the Pearson correlation coefficient for the data normalized via different techniques with qRT-PCR data.

Normalization method	hsa-miR-125b	hsa-miR-142-3p	hsa-miR-142-5p	hsa-miR-146b-5p	hsa-miR-150	hsa-miR-155	hsa-miR-1915	hsa-miR-211	hsa-miR-29c	hsa-miR-342-3p
Q	0.08	0.23	0.51	0.14	0.49	0.05	0.07	0.89	0.30	0.22
subQ	0.07	0.21	0.51	0.15	0.48	0.05	0.08	0.89	0.30	0.21
RUV	0.02	0.08	0.45	0.16	0.41	0.03	0.10	0.89	0.24	0.25
subRUV	0.01	0.04	0.41	0.05	0.35	0.02	0.11	0.89	0.22	0.26
subRUV1	0.06	0.12	0.48	0.20	0.45	0.03	0.07	0.89	0.27	0.22
subRUV5	0.21	0.05	0.12	0.32	0.01	0.09	0.02	0.53	0.01	0.24
QRUV	0.08	0.19	0.52	0.15	0.50	0.03	0.08	0.89	0.28	0.22
subQRUV	0.08	0.16	0.51	0.15	0.49	0.03	0.12	0.90	0.25	0.20
subQRUV1	0.08	0.20	0.51	0.16	0.48	0.05	0.08	0.89	0.31	0.21
subQRUV5	0.24	0.16	0.17	0.18	0.08	0.12	0.08	0.47	0.17	0.25
Best method	**subQRUV5**	**Q**	**QRUV**	**subRUV5**	**QRUV**	**subQRUV5**	**subQRUV**	**subQRUV**	**subQRUV1**	**subRUV**

C.2 ADDITIONAL INFORMATION FOR SECTION 4.3

C.2.1 *Summary of the melanoma miRNA datasets and the biomarkers used in the meta-analysis*

Table C.2 details the datasets used in Chapter 4, and the information about the published and identified biomarkers using these datasets and the survival classes evaluated in each dataset.

C.2.2 *Performance of biomarkers relative to random feature sets*

Figure C.2 shows the improvement over random signatures (IOR) scores for the validation of random miRNA sets in each validation dataset, ordered by the validation dataset to facilitate the comparison among signatures within a particular dataset.

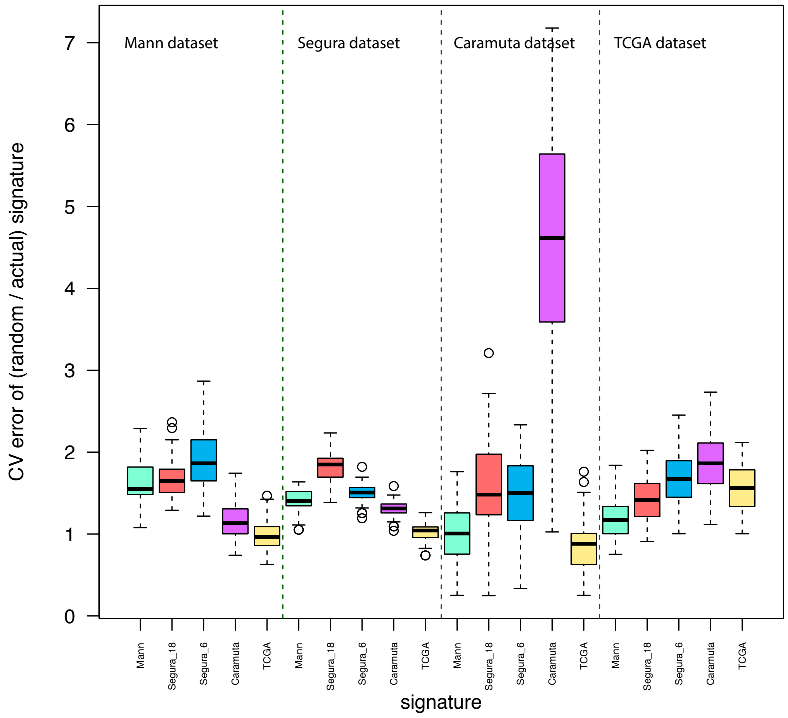


Figure C.2: **Improvement over random signature scores.** The improvement in prediction error of the signatures relative to the prediction errors of equivalently sized random miRNA sets, for each of the 100 random miRNA sets generated, ordered by the validation dataset.

Table C.2: **Summary of features of microRNA-based prognostic signatures reviewed and cross-validated.**

Expression data source	**Biomarker**	**miRNAs**	**Sample size**	**Sample characteristics**	**Platform**	**Classes compared**	**Performance in a multivariate setting**	**Data link**
Caramuta[1] (Caramuta *et al.*, 2010)	6 mRNAs predictive of short survival post diagnosis of regional lymph node metastases	miR-338, let-7, miR-365, miR-191, miR-193b-3p, miR-193a-3p	16	FF	Agilent-016436 Human miRNA Microarray 1.0 G4472A (470 miRNAs), miRBase release 9.1	< 13 months cf. >60 months survival from metastasis detection	Differentially expressed miRNAs were not significantly associated with age at diagnosis, gender, or Breslow tumor thickness of the primary tumors	GEO Accession Number: GSE19387
Segura[1] (Segura *et al.*, 2010)	18 miRNAs predictive of longer post-recurrence survival in metastatic patients	miR-214-3p, miR-126-3p, miR-143-5p, miR-28-5p, miR-342-5p, miR-10b-5p, miR-28-3p, miR-143-3p, miR-145-5p, miR-497-5p, miR-455-3p, miR-146b-5p, miR-155-5p, miR-342-3p, miR-150-5p, miR-142-5p, miR-193b-3p, miR-193a-3p	59	FFPE	Rosetta Genomics custom microarrays (911 miRNAs)	Longer survival (those who survived 18mo or more from the date of resection of the metastatic tumor) cf. shorter survival (patients who survived less than 18mo after same)	When the 6-miRNA signature and AJCC stage were included in the model, other variables such as age, sex, or time to first recurrence were not significant. Some miRNAs from the signature were related to stage and site of metastasis	Data generously provided directly by the authors of Segura et al. 2010
	6-miRNA predictor of longer post-recurrence survival in metastatic patients	miR-145-5p, miR-497-5p, miR-455-3p, miR-155-5p, miR-342-3p, miR-150-5p						
Mann[2] (Tembe *et al.*, 2014)	12-miRNA predictor of good prognosis, relapse-free survival after resection of metastatic lymph node disease	miR-509-3p, miR-363-3p, miR-125b-5p, miR-514a-3p, miR-223-5p, miR-211-5p, miR-142-3p, miR-146b-5p, miR-155-5p, miR-342-3p, miR-150-5p, miR-142-5p	45	FF	Agilent Human miRNA Microarray Release 16.0, 8x60K	Good prognosis, defined as time from surgery to death from melanoma >4yr with no sign of relapse cf. poor prognosis (patients surviving <1yr after same)	The 12-miR signature performed better (lower error rate) compared with the 4 dominant standard-of-care clinico-pathologic variables[3]	GEO Accession Number: GSE59334
TCGA [2]	15-miRNA predictor of better prognosis, relapse-free survival after resection of metastatic lymph node disease	hsa-miR-105-5p[4] hsa-miR-105-5p[4] hsa-miR-1250-5p, hsa-miR-146a-3p, hsa-miR-155-3p, hsa-miR-181a-5p, hsa-miR-204-5p, hsa-miR-362-5p, hsa-miR-3655, hsa-miR-3679-3p, hsa-miR-411-3p, hsa-miR-452-3p, hsa-miR-541-3p, hsa-miR-767-5p,hsa-miR-767-3p	23	FF	BSGSC Illumina HiSeq mRNASeq	Better prognosis, defined as time from surgery to death from melanoma >3yr with no sign of relapse) cf. worse prognosis (patients surviving <2yr after same)	The 15-miR signature performed slightly worse (higher error rate) compared with the 4 dominant standard-of-care clinico-pathologic variables[3]	TCGA data portal, SKCM

[1] Biomarker identified in a previous study.

[2] Signature identified in the present study.

[3] The four variables are: tumor-positive lymph nodes, tumor burden at the time of staging (microscopic v. macroscopic), presence or absence of primary tumor ulceration, and thickness of the primary melanoma (Balch et al., 2009).

[4] Refers to hsa-miR-105-1 mature (MIMAT0000102) and hsa-miR-105-2 mature (MIMAT0000102): identical miR sequence, different genomic loci.

Abbreviations: FFPE, formalin-fixed paraffin-embedded; FF, fresh frozen; mo, months; GEO: Gene Expression Omnibus; GSE, Gene Expression Omnibus Series; SKCM, Skin Cutaneous Melanoma; TCGA, The Cancer Genome Atlas; TNM, tumor-node-metastasis; yr, year.

C.2.3 *Intersection of the miRNA prognostic signatures*

The analysis of direct overlap among signatures did not allow for the observation of intersections that may be present prior to the application of the data pre-processing filters. Therefore, to ascertain any intersections prior to pre-processing the four validation datasets, we also examined the raw expression data for overlap among miRNAs appearing in at least one signature (Figure 4.8, Table C.3). For this aspect of the study, we considered the four main signatures in the four datasets since the 6-miRNA signature from (Segura *et al.*, 2010) was encompassed by their 18-miRNA signature.

While none of the signature miRNAs were common to all 5 biomarkers (Figure 4.8a), we observed some overlap at the raw data level. To begin, hsa-miR-514, from the 12-miRNA signature derived from the Mann expression data, was present in both the Segura and Caramuta datasets but did not pass our filtering thresholds and was therefore ineligible for analysis in our systematic cross-validation setting. An additional four miRNAs (hsa-miR-142-5p, hsa-miR-150, hsa-miR-155, hsa-miR-142-3p) from that same signature were present in the Caramuta but removed on filtering. This was the reason for only 11 and 7 miRNAs from the 12-miRNA signature being available to be assessed in Segura and Caramuta data respectively in our validation (Table 4.3 and 4.4, the number of miRNAs available in each case is presented within square brackets).

Similarly of the 18 miRNAs from the signature proposed in (Segura *et al.*, 2010) two of them (hsa-miR-28-3p and hsa-miR-143*) were excluded on filtering of the Mann expression data and ten were not present in the data from Caramuta (hsa-miR-142-5p, hsa-miR-150, hsa-miR-155, hsa-miR-455-3p (0), hsa-miR-145, hsa-miR-497, hsa-miR-143, hsa-miR-28-3p (0), hsa-miR-28-5p (0), hsa-miR-143*(0), hsa-miR-214). Note that miRNAs annotated with a '(0)' were not present in the raw data even before any filtering was performed. In analysis of the Caramuta signature (Caramuta *et al.*, 2010) only one miRNA was removed by filtering (hsa-miR-191, Mann expression data).

Finally, validation of the 15-miRNA signature from TCGA expression data presented the largest challenge with 10 of the signature miRNAs removed in filtering of the Mann expression data (hsa-miR-1250-5p, hsa-miR-146a-3p, hsa-miR-155-3p, hsa-miR-3655, hsa-miR-3679-3p, hsa-miR-411-3p, hsa-miR-452-3p, hsa-miR-541-3p, hsa-miR-767-

5p, hsa-miR-767-3p). In addition, 5 miRNAs were absent from the filtered Segura data (hsa-miR-1250-5p (0), hsa-miR-146a-3p, hsa-miR-3655 (0), hsa-miR-3679-3p (0), hsa-miR-541-3p), and 12 miRNAs were excluded from the expression data from Caramuta or else not assayed in the first place (hsa-miR-105-5p, hsa-miR-1250-5p(0), hsa-miR-146a-3p(0), hsa-miR-155-3p(0), hsa-miR-3655(0), hsa-miR-3679-3p (0), hsa-miR-411-3p (0), hsa-miR-452-3p (0), hsa-miR-541-3p (0), hsa-miR-767-5p, hsa-miR-767-3p). This detailed breakdown highlights the challenge to validation via a meta-analysis caused by the platform differences.

Table C.3: **Expression values for the collection (union) of miRNAs of the five signatures evaluated in the present study:** The table gives the transformed[1] expression values from each of the four validation datasets for each of the miRNAs in the signatures considered in this study. Expression values were extracted from the raw (unfiltered) data since, in some instances, miRNAs were excluded via our data pre-processing and normalisation protocols. For example, their expression was relatively low or contained missing information across multiple samples. Grey shading identifies the miRNAs that were common in at least two of the signatures while red shading shows the miRNAs that were removed from the processed dataset (the one eventually used in the cross-validation analysis) during filtering. Red shading in cells that are empty indicates miRNAs that were not present even in the respective raw data before any filtering was applied. The 18-miRNA signature from Segura encompasses the 6-miRNAs from that same study which are indicated by an asterisk.

	miRNA	miRbase name	previous IDs	Accession	Transformed[1]average expression values of unfiltered (raw) data								
					Mann		Segura		Caramuta		TCGA		
					A	Signature	A	Signature	A	Signature	A	Signature	Remarks
1	hsa-miR-142-5p	hsa-miR-142-5p		MIMAT0000433	8.3654	✓	6.1904	✓	10.0593		7.3926		
2	hsa-miR-150*	hsa-miR-150-5p		MIMAT0000451	8.1913	✓	8.6834	✓	5.5748		12.7868		
3	hsa-miR-342-3p*	hsa-miR-342-3p	hsa-miR-342	MIMAT0000753	8.4335	✓	10.8006	✓	6.2541		10.3697		
4	hsa-miR-155*	hsa-miR-155-5p		MIMAT0000646	7.7530	✓	8.9042	✓	7.4665		12.1464		
5	hsa-miR-146b-5p	hsa-miR-146b-5p	hsa-miR-146b	MIMAT0002809	8.2314	✓	8.5273	✓	8.3875		11.2458		
6	hsa-miR-193a-3p	hsa-miR-193a-3p	hsa-miR-193, hsa-miR-193a	MIMAT0000459	7.5063		9.6458	✓	8.1260	✓	5.1727		
7	hsa-miR-193b	hsa-miR-193b-3p	hsa-miR-193b	MIMAT0002819	6.8577		9.1404	✓	8.1523	✓	8.2952		
8	hsa-miR-142-3p	hsa-miR-142-3p		MIMAT0000434	11.2187	✓	5.7608		11.7466		11.5765		
9	hsa-miR-211	hsa-miR-211-5p		MIMAT0000268	8.4438	✓	6.9249		6.6836		11.9841		
10	hsa-miR-223	hsa-miR-223-3p		MIMAT0000280	9.2596	✓	6.8641		7.7528		9.7618		
11	hsa-miR-514	hsa-miR-514a-3p	hsa-miR-514	MIMAT0002883	7.1094	✓	5.6520		4.5558		11.0724		Average of -1,-2,-3

12	hsa-miR-125b	hsa-miR-125b-5p	hsa-miR-125b	MIMAT0000423	10.6743	✓	13.1041		10.5647		12.4031		hsa-mir-125b-1
				MIMAT0000424							2.0862		hsa-mir-125b-2
13	hsa-miR-363	hsa-miR-363-3p		MIMAT0000707	7.9617	✓	6.6087		6.9649		8.4015		
14	hsa-miR-509-3p	hsa-miR-509-3p	hsa-miR-509	MIMAT0002881	6.5996	✓	7.3519		6.7711		11.5354		Average of -1,-2,-3
15	hsa-miR-455-3p*	hsa-miR-455-3p		MIMAT0004784	7.0997		7.6271	✓			9.7798		
16	hsa-miR-145*	hsa-miR-145-5p	hsa-miR-145	MIMAT0000437	8.4111		12.6934	✓	10.8450		11.8788		
17	hsa-miR-497*	hsa-miR-497-5p	hsa-miR-497	MIMAT0002820	7.3686		8.6059	✓	6.4808		5.0416		
18	hsa-miR-342-5p	hsa-miR-342-5p	hsa-miR-342	MIMAT0000753	6.2475		7.2016	✓	6.2541		10.3697		
19	hsa-miR-143	hsa-miR-143-3p	hsa-miR-143	MIMAT0000435	6.9516		11.3604	✓	10.3931		17.0369		
20	hsa-miR-28-3p	hsa-miR-28-3p		MIMAT0004502	5.9138		7.3223	✓			14.7135		
21	hsa-miR-10b	hsa-miR-10b-5p	hsa-miR-10b	MIMAT0000254	8.7255		8.4616	✓	7.1581		17.6134		
22	hsa-miR-28-5p	hsa-miR-28-5p	hsa-miR-28	MIMAT0000085	6.9917		6.8890	✓			9.6582		
23	hsa-miR-143*	hsa-miR-143-5p	hsa-miR-143*	MIMAT0004599	5.9406		6.1481	✓			5.0572		
24	hsa-miR-126	hsa-miR-126-3p	hsa-miR-126	MIMAT0000445	9.3490		9.5731	✓	9.5966		12.8580		
25	hsa-miR-214	hsa-miR-214-3p	hsa-miR-214	MIMAT0000271	7.0981		9.9302	✓	7.5272		3.6813		
26	hsa-miR-191	hsa-miR-191-5p	hsa-miR-191	MIMAT0000440	5.9085		12.0039		7.4745	✓	12.1050		
27	hsa-miR-365	hsa-miR-365a-3p	hsa-miR-365	MIMAT0000710	8.3861		6.3127		7.7953	✓	6.8536		Average of -1,-2
28	hsa-miR-338-3p	hsa-miR-338-3p	hsa-miR-338	MIMAT0000763	7.8442		5.6963		8.0461	✓	11.6930		
29	hsa-let-7i	hsa-let-7i-5p	hsa-let-7i	MIMAT0000415	10.4968		10.6895		11.0673	✓	12.2967		
30	hsa-miR-105-5p[2]	hsa-miR-105-5p	hsa-miR-105	MIMAT0000102	5.9894		5.6806		2.0013		7.7246	✓	
31	hsa-miR-105-5p[2]	hsa-miR-105-5p	hsa-miR-105	MIMAT0000102	5.9894		5.6806		2.0013		7.7943	✓	
32	hsa-miR-1250-5p	hsa-miR-1250-5p		MIMAT0005902	5.8027						0.2840	✓	

33	hsa-miR-146a-3p	hsa-miR-146a-3p	hsa-miR-146a*	MIMAT0004608	5.8651		5.6674				2.9891	✓	
34	hsa-miR-155-3p	hsa-miR-155-3p	hsa-miR-155*	MIMAT0004658	5.8959		5.7237				0.7362	✓	
35	hsa-miR-181a-5p	hsa-miR-181a-5p	hsa-miR-181a	MIMAT0000256	9.0379		11.7842		8.3553		-0.5015	✓	hsa-mir-181a-2
36	hsa-miR-204-5p	hsa-miR-204-5p	hsa-miR-204	MIMAT0000265	6.7872		5.8107		5.0718		6.9938	✓	
37	hsa-miR-362-5p	hsa-miR-362-5p	hsa-miR-362	MIMAT0000705	6.7881		8.0799		6.4299		9.3987	✓	
38	hsa-miR-3655	hsa-miR-3655		MIMAT0018075	5.7720						0.6529	✓	
39	hsa-miR-3679-3p	hsa-miR-3679-3p		MIMAT0018105	5.9973						-0.6665	✓	
40	hsa-miR-411-3p	hsa-miR-411-3p	hsa-miR-411*	MIMAT0004813	5.7942		5.7169				0.4955	✓	
41	hsa-miR-452-3p	hsa-miR-452-3p	hsa-miR-452*	MIMAT0001636	5.7518		5.6970				0.6176	✓	
42	hsa-miR-541-3p	hsa-miR-541-3p	hsa-miR-541	MIMAT0004920	5.7751		5.6537				0.0230	✓	
43	hsa-miR-767-5p	hsa-miR-767-5p		MIMAT0003882	5.8764		5.8363		1.6717		7.7299	✓	
44	hsa-miR-767-3p	hsa-miR-767-3p		MIMAT0003883	5.8807		6.0252		1.1591		3.5284	✓	
						12		18		6		15	

[1] log transformed raw data for Mann, Segura and Caramuta, and Variance Stabilization Transformed data for TCGA.

[2] Refers to hsa-miR-105-1 mature (MIMAT0000102) and hsa-miR-105-2 mature (MIMAT0000102): identical miR sequence, different genomic loci.

D

SUPPLEMENTARY MATERIAL FOR CHAPTER 5

D.1 FURTHER RESULTS FROM SECTION 5.3

D.1.1 *Use of weights in the median robust method*

In Section 5.3 we investigated the effect of using platform dependent weights in the Lasso feature selection. We observed that when modelling mRNA data, using external information from other data platforms (protein and miRNA) in the feature selection improved the prediction accuracy. It is of interest to explore whether the weights based on relevant external information aid to select features with higher predictive accuracy in other feature selection methods apart from the Lasso. Inverse correlation based weights were used in the feature selection when using median robust (MR) method (MR method explained in Section 5.1.2). The standard procedure can be used here by replacing each entry in the expression data matrix (x_{qj}, the expression value for the q^{th} variable and j^{th} sample) by x_{qj}/w_q for each variable ($q = 1, 2, \ldots, p$) and each sample ($j = 1, 2, \ldots, n$).

The prediction error rates (Figure D.1) dropped slightly when weights based on protein information was utilized (mean 5-fold CV error = 23% compared to 24%) and when weights based on both protein and miRNA platforms were utilized (23%). However, this result did not hold for the integration of mRNA and miRNA data. Overall the errors were similar and these weights did not seem to have a significant impact on MR feature selection method.

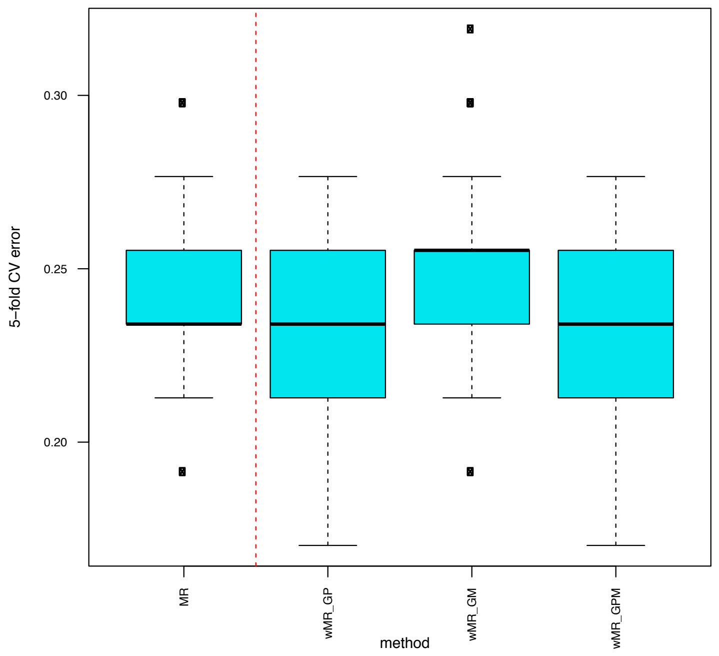


Figure D.1: **Effect of incorporating weights in median robust feature selection.** The 5-fold CV error rates in data integration setting with median robust feature selection. *MR:* No integration, *wMR_GP:* Integration of mRNA and protein data, *wMR_GM:* Integration of mRNA and miRNA data, *wMR_GPM:* Integration of mRNA, protein and miRNA data.

BIBLIOGRAPHY

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans Autom Control*, **19**(6), 716–723. (Cited on pages 46 and 112.)

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). *Molecular Biology of the Cell*. New York: Garland Science, 4 edition. (Cited on page 10.)

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**(6769), 503–511. (Cited on page 3.)

Ambroise, C. and McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci USA*, **99**(10), 6562–6566. (Cited on page 30.)

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol*, **11**(10), R106. (Cited on pages 27 and 73.)

Auer, H., Newsom, D. L., and Kornacker, K. (2009). Expression Profiling Using Affymetrix GeneChip Microarrays. *Methods Mol Biol*, **509**, 35–46. (Cited on page 9.)

Austin, P. C. and Tu, J. V. (2004). Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *J Clin Epidemiol*, **57**(11), 1138–1146. (Cited on page 44.)

Bailey, T. L. and Elkan, C. (1993). Estimating the accuracy of learned concepts. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, pages 895–900. Morgan Kaufmann. (Cited on pages 28 and 29.)

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. *J Clin Oncol*, **19**(16), 3635–3648. (Cited on page 17.)

Balch, C. M., Gershenwald, J. E., Soong, S. J., Thompson, J. F., Atkins, M. B., Byrd, D. R., Buzaid, A. C., Cochran, A. J., Coit, D. G., Ding, S., Eggermont, A. M., Flaherty, K. T., Gimotty, P. A., Kirkwood, J. M., McMasters, K. M., Mihm, M. C., Morton, D. L., Ross, M. I., Sober, A. J., and Sondak, V. K. (2009). Final version of 2009 AJCC melanoma staging and classification. *J Clin Oncol*, **27**(36), 6199–6206. (Cited on pages 17, 41, 51, 52, 74, 75, and 89.)

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. *Pharm Stat*, **7**(3), 215–225. (Cited on page 6.)

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I. F., Soboleva, A., Tomashevsky, M., and Edgar, R. (2007). NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res*, **35**(Database issue), D760–765. (Cited on pages 2 and 62.)

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**(2), 281–297. (Cited on page 11.)

Barter, R. L., Schramm, S.-J., Mann, G. J., and Yang, Y. H. (2014). Network-based biomarkers enhance classical approaches to prognostic gene expression signatures. *BMC Syst Biol*, **8 Suppl 4**, S5. (Cited on page 119.)

Beck, A. H., Knoblauch, N. W., Hefti, M. M., Kaplan, J., Schnitt, S. J., Culhane, A. C., Schroeder, M. S., Risch, T., Quackenbush, J., and Haibe-Kains, B. (2013). Significance analysis of prognostic signatures. *PLoS Comput Biol*, **9**(1), e1002875. (Cited on page 77.)

Beer, D. G., Kardia, S. L., Huang, C. C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G., Lizyness, M. L., Kuick, R., Hayasaka, S., Taylor, J. M., Iannettoni, M. D., Orringer, M. B., and Hanash, S. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med*, **8**(8), 816–824. (Cited on page 13.)

Begum, F., Ghosh, D., Tseng, G. C., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic Acids Res*, **40**(9), 3777–3784. (Cited on page 12.)

Bergersen, L. C., Glad, I. K., and Lyng, H. (2011). Weighted lasso with data integration. *Stat Appl Genet Mol Biol*, **10**(1). (Cited on pages 4, 14, 15, 94, 107, 108, and 110.)

Biomarkers-Definitions-Working-Group (2001). Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clin Pharmacol Ther*, **69**(3), 89–95. (Cited on page 2.)

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**(6795), 536–540. (Cited on page 3.)

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. *Proc Natl Acad Sci USA*, **106**(48), 20429–20434. (Cited on pages 107 and 126.)

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. (Cited on pages 2 and 99.)

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Ann Stat*, **24**(6), pp. 2350–2383. (Cited on page 7.)

Browne, M. W. (2000). Cross-validation methods. *J Math Psychol*, **44**(1), 108–132. (Cited on page 132.)

Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, **76**(3), pp. 503–514. (Cited on pages 29 and 132.)

Campain, A. (2012). *Challenges associated with clinical studies and the integration of gene expression data*. Ph.D. thesis, Faculty of Science, School of Mathematics and Statistics, The University of Sydney. (Cited on pages 4, 6, 7, 12, 37, 41, 42, 44, 46, 48, 53, 73, 76, 96, and 126.)

Campain, A. and Yang, Y. H. (2010). Comparison study of microarray meta-analysis methods. *BMC Bioinformatics*, **11**, 408. (Cited on page 13.)

Caramuta, S., Egyhazi, S., Rodolfo, M., Witten, D., Hansson, J., Larsson, C., and Lui, W. O. (2010). MicroRNA expression profiles associated with mutational status and survival in malignant melanoma. *J Invest Dermatol*, **130**(8), 2062–2070. (Cited on pages 2, 13, 20, 25, 26, 63, 64, 73, 88, 90, 148, and 149.)

Carpenter, J. R. and Kenward, M. G. (2008). A critique of common approaches to missing data. in: Missing data in randomised controlled trials– a practical guide. Technical report, Birmingham: National Institute for Health Research. (Cited on page 6.)

Casella, G. and George, E. I. (1992). Explaining the gibbs sampler. *Am Stat*, **46**(3), pp. 167–174. (Cited on page 45.)

Charbonnier, C., Chiquet, J., and Ambroise, C. (2010). Weighted-LASSO for structured network inference from time course data. *Stat Appl Genet Mol Biol*, **9**, Article 15. (Cited on page 108.)

Chargaff, E. (1951). Some recent studies on the composition and structure of nucleic acids. *J Cell Physiol Suppl*, **38**(Suppl. 1), 41–59. (Cited on page 8.)

Chavent, M., Kuentz, V., Liquet, B., and Saracco, J. (2013). *ClustOfVar: Clustering of variables*. R package version 0.8. (Cited on page 53.)

Chen, C., Ridzon, D. A., Broomer, A. J., Zhou, Z., Lee, D. H., Nguyen, J. T., Barbisin, M., Xu, N. L., Mahuvakar, V. R., Andersen, M. R., Lao, K. Q., Livak, K. J., and Guegler, K. J. (2005). Real-time quantification of microRNAs by stem-loop RT-PCR. *Nucleic Acids Res*, **33**(20), e179. (Cited on page 11.)

Chin, L. and Gray, J. W. (2008). Translating insights from the cancer genome into clinical practice. *Nature*, **452**(7187), 553–563. (Cited on pages 2, 14, and 93.)

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. (Cited on pages 12 and 13.)

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Mach Learn*, **20**(3), 273–297. (Cited on page 98.)

Crick, F. (1970). Central dogma of molecular biology. *Nature*, **227**(5258), 561–563. (Cited on pages 8 and 9.)

Curto, J. D. and Pinto, J. C. (2007). New multicollinearity indicators in linear regression models. *Int Stat Rev*, **75**(1), 114–121. (Cited on pages 7 and 45.)

Daemen, A., Gevaert, O., Ojeda, F., Debucquoy, A., Suykens, J. A., Sempoux, C., Machiels, J. P., Haustermans, K., and De Moor, B. (2009). A kernel-based integration of genome-wide data for clinical decision support. *Genome Med*, **1**(4), 39. (Cited on pages 2, 14, and 15.)

de Vries, E. and Coebergh, J. W. (2004). Cutaneous malignant melanoma in Europe. *Eur J Cancer*, **40**, 2355–2366. (Cited on page 16.)

Dettling, M. and Maechler, M. (2011). *supclust: Supervised Clustering of Predictor Variables such as Genes*. R package version 1.0-7. (Cited on page 97.)

Diamandis, E. P. (2010). Cancer biomarkers: can we turn recent failures into success? *J Natl Cancer Inst*, **102**(19), 1462–1467. (Cited on page 13.)

Diaz-Uriarte, R. (2010). *varSelRF: Variable selection using random forests*. R package version 0.7-3. (Cited on page 53.)

Diehl, F., Grahlmann, S., Beier, M., and Hoheisel, J. D. (2001). Manufacturing DNA microarrays of high spot homogeneity and reduced background signal. *Nucleic Acids Res*, **29**(7), E38. (Cited on page 9.)

Du, P., Kibbe, W. A., and Lin, S. M. (2008). lumi: a pipeline for processing Illumina microarray. *Bioinformatics*, **24**(13), 1547–1548. (Cited on page 22.)

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Stat Assoc*, **97**(457), pp. 77–87. (Cited on pages 126 and 131.)

Dupuy, A. and Simon, R. M. (2007). Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Inst*, **99**(2), 147–157. (Cited on page 13.)

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann Statist*, **7**(1), 1–26. (Cited on pages 7 and 28.)

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *J Am Stat Assoc*, **78**(382), 316–331. (Cited on page 29.)

Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat Sci*, **1**(1), pp. 54–75. (Cited on page 7.)

Efron, B., Burman, P., Denby, L., Landwehr, J. M., Mallows, C. L., Shen, X., Huang, H.-C., Ye, J., Ye, J., and Zhang, C. (2004). The estimation of prediction error: Covariance penalties and cross-validation [with comments, rejoinder]. *J Am Stat Assoc*, **99**(467), pp. 619–642. (Cited on page 29.)

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. *Meth Enzymol*, **410**, 57–73. (Cited on page 10.)

Fisher, R. A. (1950). *Statistical methods for research workers*. Oliver & Boyd, Edinburgh. (Cited on page 12.)

Fridlyand, J., Snijders, A. M., Ylstra, B., Li, H., Olshen, A., Segraves, R., Dairkee, S., Tokuyasu, T., Ljung, B. M., Jain, A. N., McLennan, J., Ziegler, J., Chin, K., Devries, S., Feiler, H., Gray, J. W., Waldman, F., Pinkel, D., and Albertson, D. G. (2006). Breast tumor copy number aberration phenotypes and genomic instability. *BMC Cancer*, **6**, 96. (Cited on page 15.)

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*, **33**(1), 1–22. (Cited on pages 52, 108, and 113.)

Gagnon-Bartsch, J. A. and Speed, T. P. (2011). Using control genes to correct for unwanted variation in microarray data. Technical report, Department of Statistics, University of California, Berkeley. (Cited on page 12.)

Gagnon-Bartsch, J. A. and Speed, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, **13**(3), 539–552. (Cited on pages 66, 67, and 69.)

Garbe, C. and Leiter, U. (2009). Melanoma epidemiology and trends. *Clini Dermatol*, **27**(1), 3 – 9. Melanoma and Pigmented Lesions, Part 1. (Cited on page 16.)

Garcia, T. P., Muller, S., Carroll, R. J., Dunn, T. N., Thomas, A. P., Adams, S. H., Pillai, S. D., and Walzem, R. L. (2013). Structured variable selection with q-values. *Biostatistics*, **14**(4), 695–707. (Cited on page 108.)

Garmire, L. X. and Subramaniam, S. (2012). Evaluation of normalization methods in mammalian microRNA-Seq data. *RNA*, **18**(6), 1279–1288. (Cited on page 66.)

Gershenwald, J. E., Soong, S.-j., and Balch, C. M. (2010). 2010 tnm staging system for cutaneous melanoma and beyond. *Ann Surg Oncol*, **17**(6), 1475–1477. (Cited on page 17.)

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**(14), e184–190. (Cited on pages 2, 14, and 15.)

Git, A., Dvinge, H., Salmon-Divon, M., Osborne, M., Kutter, C., Hadfield, J., Bertone, P., and Caldas, C. (2010). Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. *RNA*, **16**(5), 991–1006. (Cited on pages 2 and 11.)

Goble, C. and Stevens, R. (2008). State of the nation in data integration for bioinformatics. *J Biomed Inform*, **41**(5), 687–693. (Cited on page 15.)

Grada, A. and Weinbrecht, K. (2013). Next-generation sequencing: methodology and application. *J Invest Dermatol*, **133**(8), e11. (Cited on pages 2 and 93.)

Hafner, M., Landgraf, P., Ludwig, J., Rice, A., Ojo, T., Lin, C., Holoch, D., Lim, C., and Tuschl, T. (2008). Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods*, **44**(1), 3–12. (Cited on page 11.)

Hamid, J. S., Hu, P., Roslin, N. M., Ling, V., Greenwood, C. M., and Beyene, J. (2009). Data integration in genetics and genomics: methods and challenges. *Hum Genomics Proteomics*, **2009**. (Cited on pages 12 and 132.)

Hanash, S. (2004). Integrated global profiling of cancer. *Nat Rev Cancer*, **4**(8), 638–644. (Cited on pages 2, 14, 93, and 132.)

Harel, O. and Zhou, X. H. (2007). Multiple imputation: review of theory, implementation and software. *Stat Med*, **26**(16), 3057–3077. (Cited on page 6.)

Hastie, T., Tibshirani, R., and Friedman, J. (2003). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition. (Cited on page 97.)

Hastie, T., Tibshirani, R., Narasimhan, B., and Chu, G. (2014). *pamr: Pam: prediction analysis for microarrays*. R package version 1.55. (Cited on pages 74 and 98.)

Hastie, T., Tibshirani, R., Narasimhan, B., and Chu, G. (2015). *impute: impute: Imputation for microarray data*. R package version 1.40.0. (Cited on page 26.)

Heymans, M., van Buuren, S., Knol, D., van Mechelen, W., and de Vet, H. (2007). Variable selection under multiple imputation using the bootstrap in a prognostic study. *BMC Med Res Methodol*, **7**(1), 33. (Cited on pages 7, 44, and 48.)

Höfling, H. and Tibshirani, R. (2008). A study of pre-validation. *Ann Appl Stat*, **2**(2), 643–664. (Cited on page 126.)

Honaker, J., King, G., and Blackwell, M. (2011). Amelia II: A program for missing data. *J Stat Softw*, **45**(7), 1–47. (Cited on pages 44 and 45.)

Hong, F. and Breitling, R. (2008). A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, **24**(3), 374–382. (Cited on pages 12 and 13.)

Hong, F., Breitling, R., McEntee, C. W., Wittner, B. S., Nemhauser, J. L., and Chory, J. (2006). RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, **22**(22), 2825–2827. (Cited on pages 12 and 13.)

Howlader, N., Noone, A. M., Krapcho, M., Neyman, N., Aminou, R., Waldron, W., Altekruse, S. F., Kosary, C. L., Ruhl, J., Tatalovich, Z., Cho, H., Mariotto, A., Eisner, M. P., Lewis, D. R., Chen, H. S., Feuer, E. J., and Cronin, K. A. (2012). Seer cancer statistics review, 1975-2009. Technical report, National Cancer Institute. Bethesda, MD. (Cited on page 16.)

Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., and Ravikumar, P. (2011). Sparse inverse covariance matrix estimation using quadratic approximation. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2330–2338. http://nips.cc/. (Cited on page 119.)

Hua, Y. J., Tu, K., Tang, Z. Y., Li, Y. X., and Xiao, H. S. (2008). Comparison of normalization methods with microRNA microarray. *Genomics*, **92**(2), 122–128. (Cited on pages 66 and 69.)

Hubble, J., Demeter, J., Jin, H., Mao, M., Nitzberg, M., Reddy, T. B., Wymore, F., Zachariah, Z. K., Sherlock, G., and Ball, C. A. (2009). Implementation of GenePattern within the Stanford Microarray Database. *Nucleic Acids Res*, **37**(Database issue), 898–901. (Cited on page 2.)

Hughes, C., Ma, B., and Lajoie, G. A. (2010). De novo sequencing methods in proteomics. *Methods Mol Biol*, **604**, 105–121. (Cited on page 10.)

Hurd, P. J. and Nelson, C. J. (2009). Advantages of next-generation sequencing versus the microarray in epigenetic research. *Brief Funct Genomic Proteomic*, **8**(3), 174–183. (Cited on page 2.)

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., and Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *J Am Stat Assoc*, **100**(469), pp. 332–346. (Cited on page 6.)

Jayawardana, K., Müller, S., Schramm, S.-J., Mann, G. J., and Yang, J. Y. (2013). Vertical data integration for melanoma prognosis. In *Proceedings of the 59th World Statistics Congress of the International Statistical Institute*, pages 3599–3604. (Cited on pages 4, 16, 110, and 128.)

Jayawardana, K., Schramm, S.-J., Haydu, L., Thompson, J. F., Scolyer, R. A., Mann, G. J., Müller, S., and Yang, J. Y. (2015a). Determination of prognosis in metastatic melanoma through integration of clinico-pathologic, mutation, mRNA, microRNA, and protein information. *Int J Cancer*, **136**(4), 863–874. (Cited on pages 2, 5, 16, 22, 24, 25, 37, 52, 73, 75, 93, 96, 127, 132, and 134.)

Jayawardana, K., Schramm, S.-J., Tembe, V., Müller, S., Thompson, J. F., Scolyer, R. A., Mann, G. J., and Yang, J. Y. (2015b). Identification, review and systematic cross-validation of microrna prognostic signatures in metastatic melanoma. Under review. J Invest Dermatol. (Cited on pages 4, 13, 63, and 133.)

John, T., Black, M. A., Toro, T. T., Leader, D., Gedye, C. A., Davis, I. D., Guilford, P. J., and Cebon, J. S. (2008). Predicting clinical outcome through molecular profiling in stage III melanoma. *Clin Cancer Res*, **14**(16), 5173–5180. (Cited on pages 3, 120, and 129.)

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**(1), 118–127. (Cited on page 12.)

Jönsson, G., Busch, C., Knappskog, S., Geisler, J., Miletic, H., Ringnér, M., Lillehaug, J. R., Borg, A., and Lønning, P. E. (2010). Gene expression profiling-based identification of molecular subtypes in stage IV melanomas with different clinical outcome. *Clin Cancer Res*, **16**(13), 3356–3367. (Cited on page 17.)

Kakavand, H., Crainic, O., Lum, T., O'Toole, S. A., Kefford, R. F., Thompson, J. F., Wilmott, J. S., Long, G. V., and Scolyer, R. A. (2014). Concordant BRAFV600E mutation status in primary melanomas and associated naevi: implications for mutation testing of primary melanomas. *Pathology*, **46**(3), 193–198. (Cited on page 52.)

Karakach, T. K., Flight, R. M., Douglas, S. E., and Wentzell, P. D. (2010). An introduction to {DNA} microarrays for gene expression analysis. *Chemometr Intell Lab*, **104**(1), 28 – 52. {OMICS}. (Cited on page 8.)

Karpenko, O. and D., Y. (2010). Relational database index choices for genome annotation data. In *Bioinformatics and Biomedicine Workshops (BIBMW), 2010 IEEE International Conference on*, pages 264–268. (Cited on page 15.)

Kiers, H. A. L. and Smilde, A. K. (2007). A comparison of various methods for multivariate regression with highly collinear variables. *Stat Methods Appl*, **16**(2), 193–228. (Cited on pages 7 and 45.)

Kim, D., Shin, H., Sohn, K. A., Verma, A., Ritchie, M. D., and Kim, J. H. (2014). Incorporating inter-relationships between different levels of genomic data into cancer clinical outcome prediction. *Methods*, **67**(3), 344–353. (Cited on pages 2, 14, 93, and 132.)

King, G., Honaker, J., Joseph, A., and Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *Am Polit Sci Rev*, **95**(1), 49–69. (Cited on pages 7 and 44.)

Klein, Moeschberger, and modifications by Jun Yan (2012). *KMsurv: Data sets from Klein and Moeschberger (1997), Survival Analysis*. R package version 0.1-5. (Cited on page 104.)

Klein, D. (2002). Quantification using real-time PCR technology: applications and limitations. *Trends Mol Med*, **8**(6), 257–260. (Cited on page 10.)

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, pages 1137–1143, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. (Cited on pages 28, 29, and 30.)

Kong, Y. and Han, J. H. (2005). MicroRNA: biological and computational perspective. *Genomics Proteomics Bioinformatics*, **3**(2), 62–72. (Cited on page 11.)

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**(3), 405–412. (Cited on page 13.)

Lanckriet, G. R., De Bie, T., Cristianini, N., Jordan, M. I., and Noble, W. S. (2004). A statistical framework for genomic data fusion. *Bioinformatics*, **20**(16), 2626–2635. (Cited on page 15.)

Lee, M., Kim, W., Choi, Y., Kim, S., Kim, D., Yu, S. J., Lee, J. H., Kim, H. Y., Jung, Y. J., Kim, B. G., Kim, Y. J., Yoon, J. H., Lee, K. L., and Lee, H. S. (2014). Spontaneous evolution in bilirubin levels predicts liver-related mortality in patients with alcoholic hepatitis. *PLoS ONE*, **9**(7), e100870. (Cited on page 83.)

Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14. *Cell*, **75**(5), 843–854. (Cited on page 11.)

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, **3**(9), 1724–1735. (Cited on page 69.)

Leek, J. T. and Storey, J. D. (2008). A general framework for multiple testing dependence. *Proc Natl Acad Sci USA*, **105**(48), 18718–18723. (Cited on page 69.)

Lin, S. M., Du, P., Huber, W., and Kibbe, W. A. (2008). Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Res*, **36**(2), e11. (Cited on page 22.)

Little, R. J. A. (1986). Survey nonresponse adjustments for estimates of means. *Int Stat Rev*, **54**(2), pp. 139–157. (Cited on page 6.)

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., New York, NY, USA, 2 edition. (Cited on pages 6 and 43.)

Lucas, R., McMichael, T., Smith, W., and Armstrong, B. K. (2006). Solar ultraviolet radiation: Global burden of disease from solar ultraviolet radiation. Technical report, World Health Organization Public Health and the Environment. (Cited on page 16.)

Lumley, T. (2012). *rmeta: Meta-analysis*. R package version 2.16. (Cited on page 83.)

Luscombe, N. M., Greenbaum, D., and Gerstein, M. (2001). What is bioinformatics? A proposed definition and overview of the field. *Methods Inf Med*, **40**(4), 346–358. (Cited on page 2.)

Mactier, S., Kaufman, K. L., Wang, P., Crossett, B., Pupo, G. M., Kohnke, P. L., Thompson, J. F., Scolyer, R. A., Yang, J. Y., Mann, G. J., and Christopherson, R. I. (2014). Protein signatures correspond to survival outcomes of AJCC stage III melanoma patients. *Pigment Cell Melanoma Res*, **27**(6), 1106–1116. (Cited on page 24.)

Mann, G. J., Pupo, G. M., Campain, A. E., Carter, C. D., Schramm, S.-J., Pianova, S., Gerega, S. K., De Silva, C., Lai, K., Wilmott, J. S., Synnott, M., Hersey, P., Kefford, R. F., Thompson, J. F., Yang, Y. H., and Scolyer, R. A. (2013). BRAF mutation, NRAS mutation, and the absence of an immune-related expressed gene profile predict poor outcome in patients with stage III melanoma. *J Invest Dermatol*, **133**(2), 509–517. (Cited on pages 2, 5, 17, 21, 22, 51, 52, and 54.)

Manoli, T., Gretz, N., Gröne, H. J., Kenzelmann, M., Eils, R., and Brors, B. (2006). Group testing for pathway analysis improves comparability of different microarray datasets. *Bioinformatics*, **22**(20), 2500–2506. (Cited on page 13.)

Marrett, L. D., Nguyen, H. L., and Armstrong, B. K. (2001). Trends in the incidence of cutaneous malignant melanoma in New South Wales, 1983-1996. *Int J Cancer*, **92**(3), 457–462. (Cited on page 16.)

Matheis, K. A., Com, E., Gautier, J. C., Guerreiro, N., Brandenburg, A., Gmuender, H., Sposny, A., Hewitt, P., Amberg, A., Boernsen, O., Riefke, B., Hoffmann, D., Mally, A., Kalkuhl, A., Suter, L., Dieterle, F., and Staedtler, F. (2011). Cross-study and cross-omics comparisons of three nephrotoxic compounds reveal mechanistic insights and new candidate biomarkers. *Toxicol Appl Pharmacol*, **252**(2), 112–122. (Cited on page 14.)

McShane, L. M., Altman, D. G., Sauerbrei, W., Taube, S. E., Gion, M., and Clark, G. M. (2005). REporting recommendations for tumor MARKer prognostic studies (REMARK). *Nat Clin Pract Urol*, **2**(8), 416–422. (Cited on page 74.)

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat Rev Genet*, **11**(1), 31–46. (Cited on pages 2 and 93.)

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2014). *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. R package version 1.6-4. (Cited on pages 37 and 98.)

Michiels, S., Koscielny, S., and Hill, C. (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, **365**(9458), 488–492. (Cited on page 29.)

Mo, Y. Y. (2012). MicroRNA regulatory networks and human disease. *Cell Mol Life Sci*, **69**(21), 3529–3531. (Cited on page 11.)

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *J Roy Stat Soc A Gen*, **135**(3), pp. 370–384. (Cited on page 5.)

Normand, S. L. (1999). Meta-analysis: formulating, evaluating, combining, and reporting. *Stat Med*, **18**(3), 321–359. (Cited on pages 12 and 13.)

Parkhomenko, E., Tritchler, D., and Beyene, J. (2007). Genome-wide sparse canonical correlation of gene expression with genotypes. *BMC Proc*, **1 Suppl 1**, S119. (Cited on page 15.)

Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., Holloway, E., Kolesnykov, N., Lilja, P., Lukk, M., Mani, R., Rayner, T., Sharma, A., William, E., Sarkans, U., and Brazma, A. (2007). ArrayExpress–a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res*, **35**(Database issue), D747–750. (Cited on pages 2 and 62.)

Parmigiani, G., Garrett, E. S., Anbazhagan, R., and Gabrielson, E. (2002). A statistical framework for expression-based molecular classification in cancer. *J R Stat Soc Ser B Stat Methodol*, **64**(4), pp. 717–736. (Cited on page 12.)

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. *Clin Cancer Res*, **10**(9), 2922–2927. (Cited on page 13.)

Patrick, E. (2014). *Statistical methods for the analysis and interpretation of RNA-Seq data*. Ph.D. thesis, Faculty of Science, School of Mathematics and Statistics, The University of Sydney. (Cited on page 10.)

Patrick, E., Ormerod, J. T., Schramm, S.-J., Scolyer, R. A., Mann, G. J., Müller, S., and Yang, Y. H. (2015). A multi-step classifier identifies cohort heterogeneity in cancers leading to improved accuracy of prognostic biomarkers. In preparation. (Cited on page 133.)

Pennells, L., Kaptoge, S., White, I. R., Thompson, S. G., Wood, A. M., Tipping, R. W., Folsom, A. R., Couper, D. J., Ballantyne, C. M., Coresh, J., Goya Wannamethee, S., Morris, R. W., Kiechl, S., Willeit, J., Willeit, P., Schett, G., Ebrahim, S., Lawlor, D. A., Yarnell, J. W., Gallacher, J., Cushman, M., Psaty, B. M., Tracy, R., Tybj?rg-Hansen, A., Price, J. F., Lee, A. J., McLachlan, S., Khaw, K. T., Wareham, N. J., Brenner, H., Schottker, B., Muller, H., Jansson, J. H., Wennberg, P., Salomaa, V., Harald, K., Jousilahti, P., Vartiainen, E., Woodward, M., D'Agostino, R. B., Bladbjerg, E. M., J?rgensen, T., Kiyohara, Y., Arima, H., Doi, Y., Ninomiya, T., Dekker, J. M., Nijpels, G., Stehouwer, C. D., Kauhanen, J., Salonen, J. T., Meade, T. W., Cooper, J. A., Cushman, M., Folsom, A. R., Psaty, B. M., Shea, S., Doring, A., Kuller, L. H., Grandits, G., Gillum, R. F., Mussolino, M., Rimm, E. B., Hankinson, S. E., Manson, J. E., Pai, J. K., Kirkland, S., Shaffer, J. A., Shimbo, D., Bakker, S. J., Gansevoort, R. T., Hillege, H. L., Amouyel, P., Arveiler, D., Evans, A., Ferrieres, J., Sattar, N., Westendorp, R. G., Buckley, B. M., Cantin, B., Lamarche, B., Barrett-Connor, E., Wingard, D. L., Bettencourt, R., Gudnason, V., Aspelund, T., Sigurdsson, G., Thorsson, B., Kavousi, M., Witteman, J. C., Hofman, A., Franco, O. H., Howard, B. V., Zhang, Y., Best, L., Umans, J. G., Onat, A., Sundstrom, J., Michael Gaziano, J., Stampfer, M., Ridker, P. M., Michael Gaziano, J., Ridker, P. M., Marmot, M., Clarke, R., Collins, R., Fletcher, A., Brunner, E., Shipley, M., Kivimaki, M., Ridker, P. M., Buring, J., Cook, N., Ford, I., Shepherd, J., Cobbe, S. M., Robertson, M., Walker, M., Watson, S., Alexander, M., Butterworth, A. S., Di Angelantonio, E., Gao, P., Haycock, P., Kaptoge, S., Pennells, L., Thompson, S. G., Walker, M., Watson, S., White, I. R., Wood, A. M., Wormser, D., and Danesh, J. (2014). Assessing risk prediction models using individual participant data from multiple studies. *Am J Epidemiol*, **179**(5), 621–632. (Cited on page 83.)

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2014). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-118. (Cited on page 68.)

Qin, L. X. (2008). An integrative analysis of microRNA and mRNA expression–a case study. *Cancer Inform*, **6**, 369–379. (Cited on page 15.)

Qu, S., Guan, J., and Liu, Y. (2015). Identification of microRNAs as novel biomarkers for glioma detection: a meta-analysis based on 11 articles. *J Neurol Sci*, **348**(1-2), 181–187. (Cited on page 88.)

Ransohoff, D. F. (2007). How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines, and study design. *J Clin Epidemiol*, **60**(12), 1205–1219. (Cited on page 13.)

Rao, Y., Lee, Y., Jarjoura, D., Ruppert, A. S., Liu, C. G., Hsu, J. C., and Hagan, J. P. (2008). A comparison of normalization techniques for microRNA microarray data. *Stat Appl Genet Mol Biol*, **7**(1), Article22. (Cited on page 66.)

Rappaport, N., Nativ, N., Stelzer, G., Twik, M., Guan-Golan, Y., Stein, T. I., Bahir, I., Belinky, F., Morrey, C. P., Safran, M., and Lancet, D. (2013). MalaCards: an integrated compendium for diseases and their annotation. *Database (Oxf)*, **2013**, bat018. (Cited on page 116.)

Rauniyar, N. and Yates, J. R. (2014). Isobaric labeling-based relative quantification in shotgun proteomics. *J Proteome Res*, **13**(12), 5293–5309. (Cited on page 11.)

Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. *J Mach Learn Res*, **3**, 1371–1382. (Cited on page 30.)

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 Res*, **62**(15), 4427–4433. (Cited on page 12.)

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press. (Cited on page 98.)

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*, **89**(427), pp. 846–866. (Cited on page 6.)

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*, **11**(3), R25. (Cited on pages 27 and 72.)

Robinson, P. (2014). Genomic data sharing for translational research and diagnostics. *Genome Med*, **6**(9), 78. (Cited on page 2.)

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltnane, J. M., Hurt, E. M., Zhao, H., Averett, L., Yang, L., Wilson, W. H., Jaffe, E. S., Simon, R., Klausner, R. D., Powell, J., Duffey, P. L., Longo, D. L., Greiner, T. C., Weisenburger, D. D., Sanger, W. G., Dave, B. J., Lynch, J. C., Vose, J., Armitage, J. O., Montserrat, E., Lopez-Guillermo, A., Grogan, T. M., Miller, T. P., LeBlanc, M., Ott, G., Kvaloy, S., Delabie, J., Holte, H., Krajci, P., Stokke, T., and Staudt, L. M. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med*, **346**(25), 1937–1947. (Cited on page 3.)

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Chapman & HallCRC, New York. (Cited on pages 6, 43, and 44.)

Schafer, J. L. (1999). Multiple imputation: a primer. *Stat Methods Med Res*, **8**(1), 3–15. (Cited on pages 6 and 43.)

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**(5235), 467–470. (Cited on page 9.)

Schomaker, M. and Heumann, C. (2014). Model selection and model averaging after multiple imputation. *Comput Statist Data Anal*, **71**(0), 758 – 770. (Cited on pages 7 and 48.)

Schramm, S.-J. (2014). *Molecular determinants of melanoma progression and prognosis*. Ph.D. thesis, Sydney Medical School, The University of Sydney. (Cited on pages 3 and 93.)

Schramm, S.-J. and Mann, G. J. (2011). Melanoma prognosis: a REMARK-based systematic review and bioinformatic analysis of immunohistochemical and gene microarray studies. *Mol Cancer Ther*, **10**(8), 1520–1528. (Cited on page 13.)

Schramm, S.-J., Campain, A. E., Scolyer, R. A., Yang, Y. H., and Mann, G. J. (2012). Review and cross-validation of gene expression signatures and melanoma prognosis. *J Invest Dermatol*, **132**(2), 274–283. (Cited on pages 2, 13, and 76.)

Schwarz, G. (1978). Estimating the dimension of a model. *Ann Stat*, **6**(2), 461–464. (Cited on pages 41, 46, and 112.)

Scolyer, R. A. and Thompson, J. F. (2013). Biospecimen banking: the pathway to personalized medicine for patients with cancer. *J Surg Oncol*, **107**(7), 681–682. (Cited on pages 120, 127, and 128.)

Segal, E., Friedman, N., Koller, D., and Regev, A. (2004). A module map showing conditional activity of expression modules in cancer. *Nat Genet*, **36**(10), 1090–1098. (Cited on page 13.)

Segura, M. F., Belitskaya-Lévy, I., Rose, A. E., Zakrzewski, J., Gaziel, A., Hanniford, D., Darvishian, F., Berman, R. S., Shapiro, R. L., Pavlick, A. C., Osman, I., and Hernando, E. (2010). Melanoma MicroRNA signature predicts post-recurrence survival. *Clin Cancer Res*, **16**(5), 1577–1586. (Cited on pages 13, 20, 25, 63, 64, 73, 90, 99, 107, 126, 148, and 149.)

Segura, M. F., Greenwald, H. S., Hanniford, D., Osman, I., and Hernando, E. (2012). MicroRNA and cutaneous melanoma: from discovery to prognosis and therapy. *Carcinogenesis*, **33**(10), 1823–1832. (Cited on page 2.)

Setlur, S. R., Royce, T. E., Sboner, A., Mosquera, J. M., Demichelis, F., Hofer, M. D., Mertz, K. D., Gerstein, M., and Rubin, M. A. (2007). Integrative microarray analysis of pathways dysregulated in metastatic prostate cancer. *Cancer Res*, **67**(21), 10296–10303. (Cited on page 13.)

Sewell, A., Ogbeide, E., Issaeva, N., Lovly, C., Boyd, A., Gilbert, J., and Yarbrough, W. G. (2012). A personalized medicine approach to treat malignant proliferating trichilemmal tumors. *Otolaryngol Head Neck Surg*, **147**(2 suppl), P145. (Cited on pages 120 and 128.)

Shaw, H. M., Quinn, M. J., Scolyer, R. A., and Thompson, J. F. (2006). Survival in patients with desmoplastic melanoma. *J Clin Oncol*, **24**(8), e12; author reply e13. (Cited on page 52.)

Shen, K. and Tseng, G. C. (2010). Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics*, **26**(10), 1316–1323. (Cited on page 13.)

Shen, R., Chinnaiyan, A. M., and Ghosh, D. (2008). Pathway analysis reveals functional convergence of gene expression profiles in breast cancer. *BMC Med Genomics*, **1**, 28. (Cited on page 13.)

Shimamura, T., Imoto, S., Yamaguchi, R., and Miyano, S. (2007). Weighted lasso in graphical Gaussian modeling for large gene network estimation based on microarray data. *Genome Inform*, **19**, 142–153. (Cited on pages 15 and 107.)

Simon, R. (2011). Genomic biomarkers in predictive medicine: an interim analysis. *EMBO Mol Med*, **3**(8), 429–435. (Cited on page 12.)

Simon, R., Radmacher, M. D., Dobbin, K., and McShane, L. M. (2003). Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst*, **95**(1), 14–18. (Cited on page 30.)

Slawski, M., zu Castell, W., and Tutz, G. (2010). Feature selection guided by structural information. *Ann Appl Stat*, **4**(2), 1056–1080. (Cited on page 108.)

Slipicevic, A. and Herlyn, M. (2012). Narrowing the knowledge gaps for melanoma. *Ups J Med Sci*, **117**(2), 237–243. (Cited on page 16.)

Smyth, G. (2005). limma: Linear models for microarray data. In R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, and S. Dudoit, editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Statistics for Biology and Health, pages 397–420. Springer New York. (Cited on pages 24, 25, 69, and 112.)

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, **338**, b2393. (Cited on page 6.)

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. *Stat Med*, **19**(8), 1059–1079. (Cited on page 7.)

Strimbu, K. and Tavel, J. A. (2010). What are biomarkers? *Curr Opin HIV AIDS*. (Cited on page 2.)

Su, Z., Li, Z., Chen, T., Li, Q. Z., Fang, H., Ding, D., Ge, W., Ning, B., Hong, H., Perkins, R. G., Tong, W., and Shi, L. (2011). Comparing next-generation sequencing and microarray technologies in a toxicological study of the effects of aristolochic acid on rat kidneys. *Chem Res Toxicol*, **24**(9), 1486–1493. (Cited on page 2.)

Subramanian, J. and Simon, R. (2010). Gene expression-based prognostic signatures in lung cancer: ready for clinical use? *J Natl Cancer Inst*, **102**(7), 464–474. (Cited on page 13.)

Tembe, V., Schramm, S.-J., Stark, M. S., Patrick, E., Jayaswal, V., Tang, Y. H., Barbour, A., Hayward, N. K., Thompson, J. F., Scolyer, R. A., Yang, Y. H., and Mann, G. J. (2014). MicroRNA and mRNA expression profiling in metastatic melanoma reveal associations with BRAF mutation and patient prognosis. *Pigment Cell Melanoma Res*. (Cited on pages 13, 24, 65, 69, 89, and 148.)

Thompson, J. F., Scolyer, R. A., and Kefford, R. F. (2005). Cutaneous melanoma. *The Lancet*, **365**, 687–701. (Cited on page 16.)

Tibshirani, R. (1996). A comparison of some error estimates for neural network models. *Neural Comput*, **8**(1), 152–163. (Cited on pages 47, 107, and 108.)

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci USA*, **99**(10), 6567–6572. (Cited on pages 14, 74, and 98.)

Tibshirani, R. J. and Efron, B. (2002). Pre-validation and inference in microarrays. *Stat Appl Genet Mol Biol*, **1**, Article1. (Cited on pages 14, 74, 94, 99, 101, 102, and 126.)

Tomioka, N., Oba, S., Ohira, M., Misra, A., Fridlyand, J., Ishii, S., Nakamura, Y., Isogai, E., Hirata, T., Yoshida, Y., Todo, S., Kaneko, Y., Albertson, D. G., Pinkel, D., Feuerstein, B. G., and Nakagawara, A. (2008). Novel risk stratification of patients with neuroblastoma by genomic signature, which is independent of molecular signature. *Oncogene*, **27**(4), 441–449. (Cited on page 15.)

Tremante, E., Ginebri, A., Lo Monaco, E., Frascione, P., Di Filippo, F., Terrenato, I., Benevolo, M., Mottolese, M., Pescarmona, E., Visca, P., Natali, P. G., and Giacomini, P. (2012). Melanoma molecular classes and prognosis in the postgenomic era. *Lancet Oncol*, **13**(5), e205–211. (Cited on page 2.)

Tseng, G. C., Ghosh, D., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res*, **40**(9), 3785–3799. (Cited on pages 12, 13, 62, and 132.)

Uno, H. (2013). *survC1: C-statistics for risk prediction models with censored survival data*. R package version 1.0-2. (Cited on page 83.)

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med*, **30**(10), 1105–1117. (Cited on pages 63 and 83.)

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *J Stat Softw*, **45**(3), 1–67. (Cited on pages 7, 44, and 45.)

van Iterson, M., Bervoets, S., de Meijer, E. J., Buermans, H. P., 't Hoen, P. A., Menezes, R. X., and Boer, J. M. (2013). Integrated analysis of microRNA and mRNA expression: adding biological significance to microRNA target predictions. *Nucleic Acids Res*, **41**(15), e146. (Cited on page 15.)

van Klaveren, D., Steyerberg, E. W., Perel, P., and Vergouwe, Y. (2014). Assessing discriminative ability of risk models in clustered data. *BMC Med Res Methodol*, **14**, 5. (Cited on page 83.)

van Vliet, M. H., Horlings, H. M., van de Vijver, M. J., Reinders, M. J. T., and Wessels, L. F. A. (2012). Integration of clinical and gene expression data has a synergetic effect on predicting breast cancer outcome. *PLoS ONE*, **7**(7), e40358. (Cited on pages 29, 99, and 126.)

van Wieringen, W. N., Unger, K., Leday, G. G., Krijgsman, O., de Menezes, R. X., Ylstra, B., and van de Wiel, M. A. (2012). Matching of array CGH and gene expression microarray features for the purpose of integrative genomic analyses. *BMC Bioinformatics*, **13**, 80. (Cited on page 15.)

van't Veer, L. J. and Bernards, R. (2008). Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature*, **452**(7187), 564–570. (Cited on pages 3 and 129.)

Varma, S. and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, **7**, 91. (Cited on pages 29, 30, and 34.)

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, 4 edition. ISBN 0-387-95457-0. (Cited on page 98.)

Venet, D., Dumont, J. E., and Detours, V. (2011). Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput Biol*, **7**(10), e1002240. (Cited on page 77.)

Waaijenborg, S. and Zwinderman, A. H. (2009). Correlating multiple SNPs and multiple disease phenotypes: penalized non-linear canonical correlation analysis. *Bioinformatics*, **25**(21), 2764–2771. (Cited on page 15.)

Waldron, L., Haibe-Kains, B., Culhane, A. C., Riester, M., Ding, J., Wang, X. V., Ahmadifar, M., Tyekucheva, S., Bernau, C., Risch, T., Ganzfried, B. F., Huttenhower, C., Birrer, M., and Parmigiani, G. (2014). Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. *J Natl Cancer Inst*, **106**(5). (Cited on pages 13, 77, 78, and 83.)

Wang, K., Narayanan, M., Zhong, H., Tompa, M., Schadt, E. E., and Zhu, J. (2009). Meta-analysis of inter-species liver co-expression networks elucidates traits associated with common human diseases. *PLoS Comput Biol*, **5**(12), e1000616. (Cited on page 13.)

Wang, W., Baladandayuthapani, V., Morris, J. S., Broom, B. M., Manyam, G., and Do, K. A. (2013). iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*, **29**(2), 149–159. (Cited on page 15.)

Wang, Y., Joshi, T., Zhang, X. S., Xu, D., and Chen, L. (2006). Inferring gene regulatory networks from multiple microarray datasets. *Bioinformatics*, **22**(19), 2413–2420. (Cited on page 13.)

Watanabe, T., Wu, T. T., Catalano, P. J., Ueki, T., Satriano, R., Haller, D. G., Benson, A. B., and Hamilton, S. R. (2001). Molecular predictors of survival after adjuvant chemotherapy for colon cancer. *N Engl J Med*, **344**(16), 1196–1206. (Cited on page 3.)

Waters, K. M., Pounds, J. G., and Thrall, B. D. (2006). Data merging for integrated microarray and proteomic analysis. *Brief Funct Genomic Proteomic*, **5**(4), 261–272. (Cited on page 15.)

Watson, J. D., Baker, T. A., Bell, S. P., Gann, A., Levine, M., and Losick, R. (2013). *Molecular Biology of the Gene*. Benjamin Cummings, 7 edition. (Cited on page 8.)

Wei, W. J., Shen, C. T., Song, H. J., Qiu, Z. L., and Luo, Q. Y. (2014). MicroRNAs as a potential tool in the differential diagnosis of thyroid cancer: a systematic review and meta-analysis. *Clin Endocrinol (Oxf)*. (Cited on page 88.)

Weigelt, B., Baehner, F. L., and Reis-Filho, J. S. (2010). The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. *J Pathol*, **220**(2), 263–280. (Cited on page 3.)

Weiss, S. M. and Indurkhya, N. (1994). Decision tree pruning : Biased or optimal. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages 626–632. AAAI Press and MIT Press. (Cited on page 29.)

Wessels, L. F., Reinders, M. J., Hart, A. A., Veenman, C. J., Dai, H., He, Y. D., and van't Veer, L. J. (2005). A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics*, **21**(19), 3755–3762. (Cited on page 29.)

Willenbrock, H., Salomon, J., Søkilde, R., Barken, K. B., Hansen, T. N., Nielsen, F. C., Møller, S., and Litman, T. (2009). Quantitative miRNA expression analysis: comparing microarrays with next-generation sequencing. *RNA*, **15**(11), 2028–2034. (Cited on page 2.)

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., 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. *J Natl Cancer Inst*, **98**(7), 472–482. (Cited on page 3.)

Witten, D. M. and Tibshirani, R. J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat Appl Genet Mol Biol*, **8**, Article28. (Cited on page 15.)

Wu, K., Li, L., and Li, S. (2014). Circulating microRNA-21 as a biomarker for the detection of various carcinomas: an updated meta-analysis based on 36 studies. *Tumour Biol*. (Cited on page 88.)

Yang, X. and Sun, X. (2007). Meta-analysis of several gene lists for distinct types of cancer: a simple way to reveal common prognostic markers. *BMC Bioinformatics*, **8**, 118. (Cited on page 13.)

Yin, J. Q., Zhao, R. C., and V., M. K. (2008). Profiling microrna expression with microarrays. *Trends Biotechnol*, **26**(2), 70–76. (Cited on page 11.)

Yu, J. H., Dong, J. T., Jia, Y. Q., Jiang, N. G., Zeng, T. T., Xu, H., Mo, X. M., and Meng, W. T. (2013). Individualized leukemia cell-population profiles in common B-cell acute lymphoblastic leukemia patients. *Chin J Cancer*, **32**(4), 213–223. (Cited on pages 120 and 128.)

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J Roy Stat Soc*, **68**, 49–67. (Cited on page 108.)

Zeng, W., Tu, Y., Zhu, Y., Wang, Z., Li, C., Lao, L., and Wu, G. (2014). Predictive power of circulating miRNAs in detecting colorectal cancer. *Tumour Biol*. (Cited on page 88.)

Zhou, X. J., Kao, M. C., Huang, H., Wong, A., Nunez-Iglesias, J., Primig, M., Aparicio, O. M., Finch, C. E., Morgan, T. E., and Wong, W. H. (2005). Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nat Biotechnol*, **23**(2), 238–243. (Cited on page 13.)

Zou, H. (2006). The adaptive lasso and its oracle properties. *J Am Stat Assoc*, **101**(476), 1418–1429. (Cited on pages 4, 15, 107, and 108.)

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J Roy Stat Soc B*, **67**, 301–320. (Cited on pages 47 and 113.)