Predicting an optimal function for diagnostic and prognostic analyses with gene expression data.

Victor Lih Jong

Predicting an optimal function for diagnostic and prognostic analyses with gene expression data.

PhD thesis. Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, the Netherlands.

The studies in this thesis were financially supported by the VIRGO consortium, which is funded by the Netherlands Genomics Initiative and by the Dutch Government (FES0908). Financial support by the Julius Center for Health Sciences and Primary Care, for the publication of this thesis is gratefully acknowledged.

ISBN:	978-90-393-6781-0
Author:	Victor Lih Jong
Cover Design:	Victor Lih Jong, www.canva.com/vicky_jong
Printing:	ProefschriftMaken

Predicting an optimal function for diagnostic and prognostic analyses with gene expression data.

Voorspellen van een optimaal functie voor diagnostische en prognostische analyses met gen expressie data.

(met een samenvatting in het Nederlands)

Prédire une fonction optimale pour les analyses diagnostiques et pronostiques avec des données d'expression génique.

(avec un résumé en français)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof. dr. G. J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op dinsdag 23 mei 2017 des middags te 12.45 uur

door

Victor Lih Jong

geboren op 25 oktober 1982 te Agong, Kameroen

Promotoren: Prof. dr. ir. M. J. C. Eijkemans

Prof. dr. C. B. Roes

"It is far better to foresee even without certainty than not to foresee at all."

Henri Poincare in The Foundations of Science, page 129.

Contents

Chapter 1:	General introduction.	1
PART I:	DATA CHARACTERISTICS ASSOCIATE TO THE PERFORMANCE OF PREDICTIVE FUNCTIONS.	7
Chapter 2:	Exploring homogeneity of correlation structures of gene expression datasets within and between etiological disease categories.	9
Chapter 3:	Factors affecting the accuracy of a class prediction model in gene expression data.	31
PART II:	PREDICTING AN OPTIMAL PREDICTIVE FUNCTION FOR A GIVEN DATASET.	51
Chapter 4:	Selecting a classification function for class prediction with gene expression data.	53
Chapter 5:	Selecting an optimal probabilistic classifier with gene expression data: does it differ from an optimal direct classifier?	73
Chapter 6:	Choosing a Cox's predictive function for survival analysis with gene expression data.	93
PART III:	IDENTIFICATION AND VALIDATION OF PREDICTIVE BIOMARKERS.	117
Chapter 7:	Transcriptome assists prognosis of disease severity in respiratory syncytial virus infected infants.	119
Chapter 8:	General discussion.	137
APPENDICES		
	Bibliography.	147
	Summary.	161
	Nederlands samenvatting.	167
	Résumé en français.	171
	List of publications.	177
	Acknowledgements.	179
	Curriculum vitae.	185

Chapter 1

General introduction.

Chapter 1

1.1 Background

The completion of the human genome project in 2003 and the advent of high-throughput technologies are widely considered to provide more understanding of the mechanisms of human diseases and hence improve the clinical management of patients. The type of high-throughput data, also known as -omics data depends on the biological molecules that are measured in an experiment. Metabolomics, proteomics, genomics and transcriptomics respectively stemming from the measure of metabolites, proteins, genes and transcripts are a few examples of such -omics data. Gene expression includes a wide range of applications relevant to the high-throughput analysis of expression of biological quantities, including microarrays (nucleic acid, protein, array CGH, genome tiling, and other arrays), RNA-seq, proteomics and mass spectrometry [http://www.oxfordjournals.org/, 03/01/2017]. Gene expression analysis, measures the expression of tens of thousands of genes (proteins) and has become a widely used tool to identify particular disease subpopulations and to perform diagnostic and prognostic predictions [van 't Veer et al., 2002; Huang et al., 2010]. In clinical practice, they are used in diagnostic and

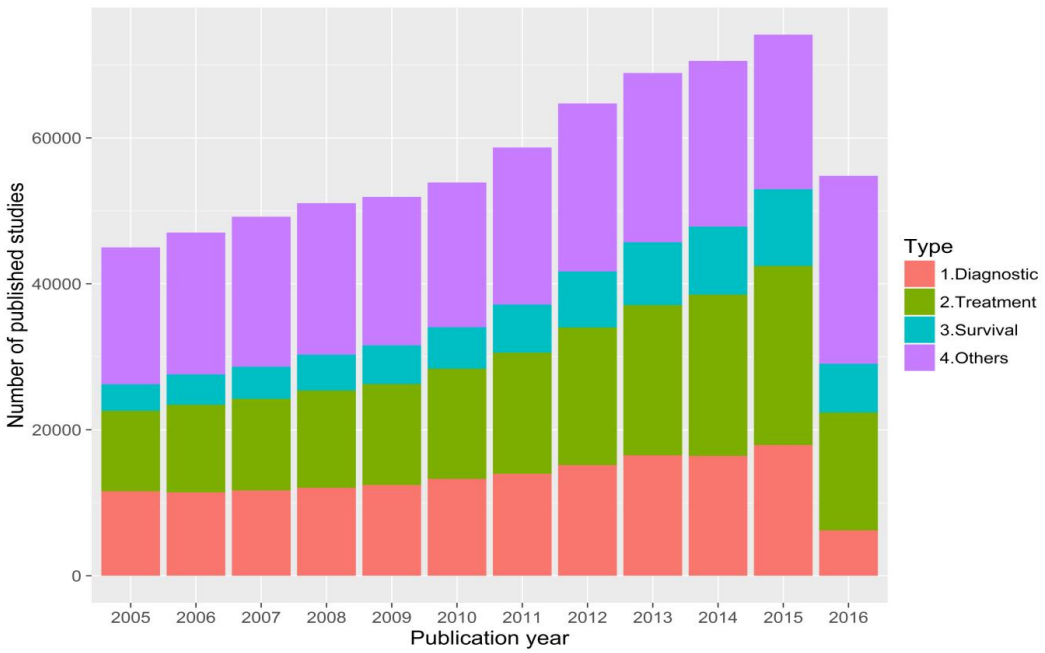

Figure 1.1: Number of published gene expression studies recorded in the United States National Library of Medicine - National Institute of Health (PubMed) repository from 2005-2016. *A quick search was performed with the keywords "Gene expression", "Gene expression" AND "diagnosis", "Gene expression" AND "treatment", "Gene expression" AND "survival". The low number of published studies in 2016 is possibly due to incomplete number of records at the time of the search.*

prognostic analyses while in preclinical studies (toxicogenomics), they involve predicting the toxicity of compounds in animal models with the goal of speeding up the evaluation of toxicity for new drug candidates [Shi et al., 2010]. Figure 1.1 illustrates a tremendous increase in the number of published gene expression studies recorded in the United States National Library of Medicine - National Institute of Health (PubMed) repository over a period of 12years (2005 - 2016).

A large number of statistical methods are available for the analysis of gene expression data and these methods vary depending on the biological question posed. The three main analyses performed on gene expression data includes: clustering, differential expression and prediction analyses. Clustering analysis does not include a priori information on samples' outcome. It is commonly aimed to group genes and/or samples that share a similar underlying biological mechanism. This might lead to the discovery of biological pathways within a disease and/or disease subtypes. The other two types of analyses are outcome-related analyses, where the outcome is taken into account in the analysis. Differential expression analysis on one hand, aims to extract informative genes that differentiate samples of two (or more) distinct groups. While predictive modeling on the other hand, aims to extract genes that together accurately predict the outcome of independent sample(s). Predictive analysis could be classified as diagnostic or prognostic (response to treatment and time to event or survival) analysis. Just like the overall number of published gene expression studies, the number of published studies on predictive analyses using gene expression data is larger than other analysis types, and increases over the years, as illustrated on Figure 1.1.

The diagnosis, treatment selection and prognosis of an individual patient can become more accurate with efficient statistical predictive models. Regular statistical prediction modeling with gene expression data suffers from the curse of dimensionality i.e. a very low number of samples (n) relative to the number of genes (p), also known as the curse of dimensionality problem ($p \gg n$). As such, several predictive functions (algorithms) have been developed to advert the curse of dimensionality problem. Nevertheless, these algorithms have been shown to perform differently across gene expression datasets [Lee et al., 2005; van Wieringen et al., 2009; Kim & Simon, 2011]. And that correlation is one of data characteristics that is associated to the performance of such functions [Kim & Simon, 2011]. In addition to data-specific characteristics, several study specific factors have been shown, particularly in cancer studies, to have an association with the performance of predictive functions, e.g. microarray type, clinical endpoint and predictive modeling technique [Ntzani and Ioannidis, 2003; Shi et al., 2010]. While substantial amount of

information is known about the characteristics of predictive functions and prediction modeling procedures, little is known about which data characteristics affect the performance of predictive functions. For instance, diagonal linear discriminant analysis is a predictive function that assumes no covariance and hence no correlation between variables and might fail if the data is highly correlated. On the other hand, linear discriminant analysis assumes a common covariance matrix for the classes and thus to some extent, accounts for correlations [Hastie et al., 2003]. In addition, penalized regressions like ridge, lasso, elastic net are capable to handle correlated variables in different ways. Tree-based methods are by nature designed to capture interactions between variables while neural networks might capture other complex structures within a given dataset.

Considering the above mentioned characteristics of the predictive functions, it is evident that the performance of these functions might depend on the characteristics of the data in question. Despite this, the literature on how to choose an optimal predictive function for a given dataset is sparse. A common practice is to compare several predictive functions and select the function with the smallest cross-validated error but this is often computationally intensive and even when feasible, leads to selection bias [Varma & Simon, 2006; Tibshirani and Tibshirani, 2009; Bernau et al., 2013; Ding et al., 2014] because there is a high probability that a function might have the smallest cross-validated error simply by chance. This probability increases with both the number of algorithms compared and a decrease in sample size. Thus, selection bias is often high in data with small sample sizes and when several least optimal algorithms are compared [Ding et al. 2014]. These authors stated as an example that one uses a small pilot data, compares several machine learning methods and selects the minimum error classifier (MEC) with a falsely small error because of the selection bias. When the model proceeds to a large cohort validation for translational research, it will likely fail. Hence, several bias correction methods have been proposed in the literature of class prediction with gene expression data. Nevertheless, no such method is 100% effective when several least optimal algorithms are compared on a dataset with a small sample size, as observed by Ding et al., (2014). As such, some experimenters adhere to one or a few predictive functions irrespective of the dataset, disease or medical question being addressed. While others choose a function for their datasets by affinity or familiarity without taking into account the characteristics of such data.

Another school of thought in the literature emphasizes the use of super-learners which combine the scores of several predictive functions to improve overall performance. Nevertheless, super-learners are less acceptable in medical applications because there is resistance of pathologists and

medical practitioners to use uninterpretable black box multivariate tests for making important treatment decisions [Simon, 2014]. Parsimonious models that can be presented as nomograms are increasingly popular in the medical literature [Kattan, 1998]. Simon, (2014) states that if features make biological and medical sense and the test provides well-calibrated forecasts that help patients and physicians make important medical decisions, then acceptance is more likely. While super-learners as a black box means its model composes of the entire genome which is often time consuming and costly to profile, traditional predictive functions yield a gene signature composing of a few tens or hundreds of genes that can easily be profiled at a low cost, for quick application in the clinic. Thus, traditional predictive functions that can yield a gene signature are often preferred than super-learners. That notwithstanding, there exists several traditional predictive functions in the literature and a key question that arises is, which predictive function should one utilize for a given gene expression dataset?

In this thesis, we would address this question by first identifying distinct data characteristics that associate to the performance of predictive functions, using real-life datasets and simulations. Secondly, we shall use our empirical results to construct predictive models to determine an optimal predictive function (algorithm) for a given gene expression dataset for different biological endpoints (i.e. for diagnostic and prognostic studies).

1.2 Outline

In Part I, we focus on determining data characteristics that associate to the performance of predictive functions using real-life datasets from public repositories. Correlation is one of the data characteristics that has been outlined in the literature to be associated with the performance of predictive functions. In Chapter 2, we assess the homogeneity of correlation structures across gene expression datasets and between disease categories. In Chapter 3, we outline and quantify data characteristics that could be associated to the performance (accuracy) of direct classification functions. The associations of these data characteristics to the accuracy of nine direct classification functions would be assessed using a random effects logistic regression model.

Part II of the thesis focuses on empirically utilizing identified (from Part I) data characteristics to construct predictive models for determining an optimal predictive function for a given gene expression data. In Chapter 4**,** such a predictive model for selecting an optimal direct classification function among ten functions often used for binary classification would be devised. While in Chapter 5, a similar model for probabilistic classifiers shall also be constructed using nine of the

ten functions that are designed (can self-post-process) to produce probabilities. We will then extend our study to survival (time to event) analysis and a predictive model to determine an optimal Cox's predictive function among seven Cox's proportional hazard functions would be constructed and presented in Chapter 6.

In Part III (Chapter 7), we would utilize our results to select a predictive function for gene expression profiles obtained 24 hours upon hospitalization of respiratory syncytial virus (RSV) infected infants, with goal to identify prognostic biomarkers for disease severity. A signature that might serve as the basis to develop a clinical predictive model that could help in the management of RVS patients in pediatric wards shall be presented.

Finally, a general discussion outlining the outcome of this thesis and perspectives for future research shall be presented in Chapter 8.

PART I

DATA CHARACTERISTICS ASSOCIATE TO THE PERFORMANCE OF PREDICTIVE FUNCTIONS.

Chapter 2

Exploring homogeneity of correlation structures of gene expression datasets within and between etiological disease categories.

Victor L. Jong, Putri W. Novianti, Kit C. B. Roes & Marinus J. C. Eijkemans

Stat. Appl. Genet. Mol. Biol. 2014; **13**(6): 717–732

Chapter 2

Abstract

Background: The literature shows that classifiers perform differently across datasets and that correlations within datasets affect the performance of classifiers. The question that arises is whether the correlation structure within datasets differ significantly across diseases.

Results: In this study, we evaluated the homogeneity of correlation structures within and between datasets of six etiological disease categories; inflammatory, immune, infectious, degenerative, hereditary and acute myeloid leukemia (AML). We also assessed the effect of filtering; detection call and variance filtering on correlation structures. We downloaded microarray datasets from ArrayExpress for experiments meeting predefined criteria and ended up with 12 datasets for non-cancerous diseases and six for AML. The datasets were preprocessed by a common procedure incorporating platform specific recommendations and the two filtering methods mentioned above. Homogeneity of correlation matrices between and within datasets of etiological diseases was assessed using the Box's M statistic on permuted samples.

Conclusion: We found that correlation structures significantly differ between datasets of the same and/or different etiological disease categories and that variance filtering eliminates more uncorrelated probesets than detection call filtering and thus renders the data highly correlated.

2.1 Introduction

Class prediction is one of the major experimental analyses in gene expression studies. Although a standard procedure to build a class prediction model has been published by Wessels et al., (2005) with additional guidelines by MAQC-II initiative [Shi et al., 2010], there is no standard guideline to choose an appropriate classification function for such a class prediction model for a particular dataset. As a result, investigators make a choice of classification method at random or by affinity or familiarity but it has been shown by Lee et al., (2005) that classifiers perform differently across datasets and in the MAQC-II initiative that classification functions explain a large portion of the variability in the performance of class prediction models. Thus, a classifier that works well in one dataset might not work well in another dataset. This is possibly due to the fact that datasets have different characteristics affecting the performance of classifiers. One of which is the correlation structure within the gene expression data as shown by Kim & Simon, (2011).

Considering that classifiers perform differently across datasets [Lee et al., 2005] and that classifiers and gene expression data have their characteristics of which those of classifiers are known, a study of gene expression data's characteristics should be able to aid investigators to choose the right classification method that is suitable for such data. Particularly, the correlation structure of gene expression data might help to choose the appropriate classification method. For instance, shrunken centroid discriminant analysis (SCDA) is one of the classification methods that work well in data with weak correlations. Also, diagonal linear discriminant analysis (DLDA) assumes zero covariance and hence zero correlation between variables and might fail if the data is highly correlated. On the other hand, linear discriminant analysis assumes a common covariance (correlation) matrix for the classes and thus takes into account correlations to an extent [Hastie et al., 2003]. In addition, penalized logistic regression performs well in a strongly correlated dataset or takes into account correlated groups of variables [Hastie et al., 2003; Kim & Simon, 2011]. The characteristics of these classification functions and this preliminary finding of Kim & Simon, (2011) illustrate the dependency of classification functions on the properties of gene expression data. Though they focused on probabilistic classifiers while using calibration and refinement scores as measures of evaluation, they showed that when using misclassification error to evaluate classifiers, the correlations within gene expression data do have an influence on the performance of these classifiers. However, they considered various simple correlation structures for their simulated datasets and did not systematically study the true correlation structures that exist in

real-life datasets. A question that therefore arises is, whether the correlation structures in real-life gene expression datasets differ across datasets/diseases.

In this study, we address this question by exploring the homogeneity of correlation structures within non-cancerous etiological diseases and one particular type of cancer with the aim to investigate whether the correlation structures between datasets are heterogeneous. A secondary goal is whether disease type might play a role in that, datasets of a particular disease might have a homogeneous structure. The effect of two filtering methods on the correlation structure within a given data was assessed as a third goal. We mainly focus on non-cancerous datasets because they have received low attention in the literature, whereas they might have varying complexities.

We limit ourselves to published microarray data available in online repositories in a wide range of human diseases outside the field of cancer and in one particular type of cancer, which were used or could be used for at least binary classification after eliminating healthy controls (if present). We excluded healthy controls because most classification analyses involve diseased patients. In diagnostic studies for instance, it is common to classify patients into disease subtype, e.g., type 1 or type 2 diabetes or patients suspected of a disease. Meanwhile, the controls in these studies were actually healthy controls and were never suspected of any disease. For prognostic and response to treatment cases, it is obvious that only patients with a particular disease are included. The details of the methodology are presented in Section 2.2. Results of the correlation analysis are provided in Section 2.3 and it is followed by a discussion in Section 2.4.

2.2 Methodology

2.2.1 Data extraction

We downloaded microarray datasets from ArrayExpress database [www.ebi.ac.uk/arrayexpress/, 01/05/2013]. The criteria for selecting the datasets are that the experiments 1) were conducted in humans, 2) outside the field of cancer, 3) have samples with class labels in at least two classes, 4) were published after 2005, 5) can be categorized as one of the disease types: inflammatory, immune, infectious, degenerative, and hereditary disease and 6) have raw datasets. Additionally, we excluded experiments entirely on normal or healthy samples. The five types of diseases were chosen to represent a broad spectrum of diseases outside the field of cancer. We expect that the diseases in the same type should have similar complexity, due to the fact that they are grouped based on etiology. In total, we downloaded 12 (3 inflammatory, 2 immune, 2 infectious, 3 degenerative and 2 hereditary) gene expression datasets on these non-cancerous diseases. In

addition, six datasets from one particular type of cancer, i.e., acute myeloid leukemia (AML) were downloaded. The datasets are briefly described below and summarized on Table 2.1.

- *Ulcerative colitis (UC1):* is a subset (cohort A) of data from two cohorts of patients who received their first treatment with infliximab for refractory ulcerative colitis. Pre-treatment colonic mucosal expression profiles were compared for responders (8) and non-responders (16) with the goal to identify mucosal gene signatures predictive of response to infliximab in patients with ulcerative colitis [Arijs et al., 2009].
- *Ulcerative colitis (UC2):* is a prospective accrued cohort of 128 children between 2 and 18 years of age, hospitalized for intravenous corticosteroid therapy for acute ulcerative colitis that were all treated with methylprednisolone. Blood samples were collected on day 3 after admission to corticosteroid treatment and on day 5 of corticosteroid treatment, patients were determined to be non-responders if they needed a second line of medical therapy (e.g., infliximab) or colectomy. Twenty corticosteroid-responsive and 20 corticosteroid-refractory patients were randomly selected for analysis of mRNA expression with the goal to determine whether early initiation of intravenous corticosteroid treatment is associated with response [Kabakchiev et al., 2010].
- *Asthma (AST):* is a subset (severe therapy-resistant and mild asthma) of data from a study in which the white blood cells of children with severe therapy-resistant asthma (17) and controlled or mild asthma (19) identified from a Swedish nation-wide study, together with recruited healthy controls (18) were profiled with the goal to identify global patterns of gene expression in severe therapy-resistant versus controlled asthma and healthy controls [www.ebi.ac.uk/arrayexpress/, 01/05/2013]. We leave out healthy controls to be able to get a binary class.
- *Periodic Fever, Aphthous Stomatitis, Pharyngitis and Adenitis (PFAPA):* is a common disease (whose pathogenesis is unknown) in children between the ages of 0 and 5. The data is a subset of a study conducted to discriminate PFAPA patients form associated diseases and healthy controls. This subset contains whole blood profiles of six FPAPA patients and six healthy controls. The gene expression datasets from six PFAPA patients during non- and flare were used [Stojanov et al., 2011].
- *Dystonia (DYS):* is from a DYT1 dystonia (an autosomal-dominantly inherited movement disorder) study in which whole blood gene profiles of manifesting and non-manifesting carriers were compared. The gene expression datasets were obtained by hybridization of cDNA from

whole blood samples from each patient using Affymetrix HG 1.0-ST chips. The training and testing datasets were combine and for this study we consider the 22 samples that belong to the carrier and 23 samples from the symptomatic group [Walter et al., 2010].

- *Psoriasis (PSO):* is from pre-treatment peripheral blood of moderate-to-severe psoriasis patients that were treated with Alefacept treatment and were classified at the end of the treatment as non-responders (7) or responders (9). The goal is to use pre-treatment gene expression profiles to predict whether or not a patient will respond to Alefacept treatment [Suarez-Farinas et al., 2010].
- *Kawasaki disease (KD):* Egami scoring system was used to group patients with kawasaki disease into intravenous immunoglobulin (IVIG) responsive (Group A) or resistant group (Group B) before starting the treatment. All patients in Group A received IVIG, meanwhile patients in Group B randomly assign into IVIG or intravenous immunoglobulin and methylprednisolone (IVIG+IVMP) treatment. Then 2.5 mL whole blood was taken from each patient to isolate its total RNA. Through microarray analysis, these samples were used to determine the response status of patients. In our study, we only consider the Group B patients, where the medical question was to compare IVIG (6) and IVIG+IVMP (5) treatments [Ogata et al., 2009].
- *Alzheimer (ALZ1):* Omega-3 fatty acids, e.g., docosahexaenoic acid (DHA) and eicosapentaneanoic acid (EPA) present in marine oils have been linked to reduce the risk of developing Alzheimer disease. In this study, gene expression profiles of blood samples obtained from 16 patients originating from a randomized double-blind, placebo-controlled trail where 174 Alzheimer disease patients received daily either 1.7 g of DHA plus 0.6 g of EPA (control) or isocaloric placebo oil for 6 months were assessed. The goal of the study was to investigate if there is a difference in response to treatment between the controlled and placebo groups [Vedin et al., 2012]. We used 15 of these samples because the cell file of 1 sample could not be downloaded.
- *Alzheimer (ALZ2):* is a diagnostic study in which hippocampus tissues were taken from patients with Alzheimer disease (7 incipients, 8 moderate, and 7 severe) as well as control people (9). The diagnosis was based on the MiniMental Status Examination (MMSE) test. The re-classification of Alzheimer patients was done to regroup the data, where we merged incipient and moderate patients, so that we have gene expression datasets from 15 non- and seven severe patients [Blalock et al., 2004].

Table 2.1: Descriptions of non-cancerous datasets.

Disease	Disease type	Tissue	Affymetrix Array	Paper*	Year	N	p1	p2	p2*	p3	p4
UC1	Inflammation	Colonic mucosal biopsies	HG U133 Plus 2.0	Yes	2009	24 (16,8)	54675	24792 (45.64%)	5251 (9.60%)	4650 (8.50%)	2500 (4.57%)
UC2	Inflammation	Whole Blood	HG 1.0 ST	Yes	2010	40 (20,20)	32321	23377 (72.33%)	4315 (13.35%)	3389 (10.49%)	2500 (7.73%)
AST	Inflammation	White blood	HG 1.0 ST	No	2011	36 (19,17)	32321	23505 (72.72%)	1509 (4.67%)	1293 (4.00%)	1293 (4.00%)
PFAPA	Infection	peripheral blood	HG U133A 2.0	Yes	2011	12 (6,6)	22277	18149 (81.47%)	3442 (15.45)	3340 (14.99%)	2500 (11.22%)
DYS	Infection	Whole blood	HG 1.0 ST	Yes	2010	45 (22,23)	32321	17802 (55.08%)	4057 (12.55%)	2811 (8.69%)	2500 (7.73%)
PSO	Immune	PBMCs	HG U95 Version 2	Yes	2010	16 (7,9)	12625	9383 (74.32%)	2007 (15.89%)	1987 (15.74%)	1987 (15.74%)
KD	Immune	Peripheral blood	HG U133 Plus 2.0	Yes	2010	11 (6,5)	54675	25946 (47.45%)	17504 (32.02%)	15438 (28.24%)	2500 (4.57%)
ALZ1	Degenerative	PBMCs	HG Focus	Yes	2012	15 (4,11)	8793	5817 (66.16%)	1419 (16.14%)	1355 (15.41%)	1355 (15.41%)
ALZ2	Degenerative	Hippocampus	HG-U133A	Yes	2007	22 (7,15)	22283	17796 (79.86%)	2302 (10.33%)	2295 (10.29%)	2295 (10.29%
HF	Degenerative	Left ventricle cardiac biopsies	HG 1.0 ST	Yes	2012	19 (7,12)	32321	22863 (70.74%)	2721 (8.42%)	2068 (6.39%)	2068 (6.39%)
GAU	Hereditary	Skin	HG U133A 2.0	No	2010	10 (5,5)	22277	16622 (74.61%)	2027 (9.09%)	2017 (9.05%)	2017 (9.05%)
CS	Hereditary	Skin	HG U133 Plus 2.0	Yes	2009	20 (10,10)	54675	23721 (43.38%)	5933 (10.85%)	5422 (9.92%)	2500 (4.57%)

Paper*: Paper availability, **p1**: The total number of probesets in the array, **p2**: The number of probesets whose expression values are greater than five in at least 10% of the total number of samples (Filter 1 only), **p2*:** The number of probesets whose standard deviation are > 0.5 (Filter 2 only), **p3**: The number of probesets after applying the first and the second filtering criterion and **p4**: number of probesets in the actual data used for clustering. Percentages are proportions of the initial number of probesets p1 and (,,.) represents the number of samples per class.

- *Heart failure (HF):* is a subset (mRNA) of data from a study in which microRNAs (miRNAs) and mRNA expression profiles of type 2 diabetic ischemic heart failure (D-HF), non-diabetic ischemic heart failure (ND-HF) patients and healthy non-diabetic non-heart failure controls were compared in parallel with the goal to evaluate the impact of miRNA dysregulations and their potential pathogenetic roles. The control group was excluded so that all samples were in heart failure disease state, in which seven patients are diabetic and 12 are non-diabetic [Greco et al., 2012].
- *Gaucher (GAU):* is a subset (Type I and III) of data from an experiment that was conducted to provide insight into the unique pathogenesis of five Type I, 5 Type III Gaucher disease patients and four controls. Control and patient fibroblast cultures were established from the full-thickness skin biopsies obtained under IRB of the National Institute of Neurological Disorders and Stroke protocols and were compared with the goal to find a gene signature that could be used to improve diagnostic accuracy and potential novel therapies for patients [www.ebi.ac.uk/arrayexpress/, 01/05/2013]. Ten samples that develop this disease were used for our study.
- *Craniosynostosis syndrome (CS):* is a skull abnormality condition that can be caused by a gene mutation. This study was aimed to compare the gene expression among three monogenic syndromes; Apert (10), Munke (10), and Saethre-chotzen syndrome (10) and control group (10). The cRNA from each patient's skin was hybridized using Affymetrix HG U133 Plus 2.0 array. For further analyses, we took Apert and Munke patients [Bochukova et al., 2010].
- *Acute myeloid leukemia (AML):* six cancerous datasets of experiments conducted on patients with any form of AML addressing several medical questions were downloaded to assess if our assumption that the correlation structure might be homogeneous in cancerous diseases. We denote the datasets as AML1‑AML6 and are summarized on Supplementary Table S1 and are, respectively, described by Payton et al., (2009); Le Dieu et al., (2009); Majeti et al., (2009); Beghini et al., (2012); Bacher et al., (2012) and Stirewalt et al., (2012). In all the experiments the healthy samples were excluded to render the data strictly to disease patients as was the case with non-cancerous datasets.

2.2.2 Preprocessing

The downloaded raw datasets were normalized using quantile normalization, background correction performed according to manufacturer's platform recommended correction, using perfect match (pm), probesets were summarized with median polish summarization and log base

two transformed [Gautier et al., 2004]. For each dataset, we then filtered out non-informative probesets using two of the several filtering methods outlined by Marczyk et al., (2013). In one, we retained probesets that had expression values greater than five in at least ten percent (10%) of the total samples (we referred to this as detection call filtering). In the other, we retained the probesets with standard deviations greater than point five (SD > 0.5 and referred to as variance filtering) and compared the correlation distributions between these methods. We then combined the two filtering criteria and refer to this retained list of probsets as the actual gene expression data. For computational reasons, if the number of probesets in the actual gene expression data is > 2500 probesets, we sample 2500 probesets from the actual gene expression data, else we consider the actual expression data for further analysis. Henceforth, we refer to these sampled or actual data as simply data.

2.2.3 Clustering

To be able to have control of the correlation structure within each dataset, we grouped the probesets that are highly correlated based on their expression values together. In an attempt to group these probesets into clusters based on these correlations, we computed a matrix of pairwise Pearson correlations for all the probesets (cor) and computed one minus correlation matrix ($1-cor$) as the distant metric used in clustering. We then applied the Partitioning Around Medoids (PAM) clustering algorithm [Kaufman and Rousseeuw, 2005] and the gap statistic [Tibshirani et al., 2001] to determine the optimal number of clusters. The gap statistic based on five hundred bootstraps was used to evaluate the optimal number of clusters (maximum: 20 clusters). The optimal number of clusters (k) was chosen as the local maximum which is at least one standard error larger than the previous, that is which satisfies the following criterion [Tibshirani et al., 2001; Hastie et al., 2003]:

$$k = \underset{K}{Argmin}\{K|G(K) \geq G(K+1) - s_{K+1}\} \qquad K = 1,2,3,\ldots,20 \tag{2.1}$$

where G is the gap statistic and s is a function of standard deviation of log cluster dissimilarity.

Finally, the probesets were clustered using PAM with the optimal number of clusters k determined from the gap statistics, $1-cor$ as distant metric and allowing for swapping till convergence. While utilizing the gap statistics because of its ability to detect no clusters, we also implemented a clustering algorithm called weighted gene co-expression network analysis (WGCNA) [Langfelder & Horvath, 2008] which is capable of excluding “noise” probesets from the clusters. This second clustering with WGCNA serves as a confirmatory analysis of the clusters

produced by PAM. To implement WGCNA, the gene expression data was transformed into a similarity metric, $s_{ij} = \frac{1+cor(gene_i, gene_j)}{2}$. This transformation was preferred because two probesets with correlation 1 should have a similarity of 1 while those with a correlation of -1 will have a similarity value of 0 so that the clustering via WGCNA in principle, was done on the correlation metric as is the case with clustering via PAM. For the same reasoning, an adjacency metric was set to be the same as a similarity metric ($a_{ij} = s_{ij}$). Then, hierarchical clustering with average linkage and dynamic tree cut algorithm [Langfelder et al., 2008] was applied in the dissimilarity measurement based on topological overlap matrix [Zhang & Horvath, 2005]. The minimum cluster size was set to be 30, i.e., a cluster can be formed by a minimum of 30 probesets. To correct the clustering results, we combined similar clusters by evaluating their first variance component. Two or more clusters would be combined if their eigenprobesets dissimilarity does not exceed 0.2. The intra- and inter-clusters correlations were examined through exploratory data analyses of all the datasets after applying both clustering algorithms.

2.2.4 Statistical comparison of homogeneity of correlation structures

To assess whether or not correlation structures are homogeneous across datasets within and between disease categories, we used Box's M test. To make comparison possible, common probesets from the preprocessed unfiltered (non-sampled) datasets to be compared were extracted. And because Box's M test is a test for covariance matrices, we standardized expression values of the probesets within each dataset to yield unit variances for all the probesets, since we are interested in the correlations. The Pearson covariance matrix of the probesets in each dataset were computed, the modulus of the determinant computed by the LU decomposition to avoid singularity; where L is a lower triangular and U an upper triangular matrix [Becker et al., 1998]. Let for datasets $1, \dots, d$ with sample sizes $n_1, \dots, n_d$, the covariance matrix of dataset i be Σ_i, we assess the hypotheses $H_0: \Sigma_1 = \cdots = \Sigma_d$ v/s $H_1: \Sigma_i \neq \Sigma_j$ for some $i \neq j$, using the Box's M statistic defined as:

$$M = -2log\Lambda \tag{2.2}$$

where $\Lambda = \frac{\prod_{i=1}^{d} |\Sigma_i|^{(n_i-1)/2}}{|\Sigma|^{(N-d)/2}}$, $\Sigma = \sum_{i=1}^{d} \frac{(n_i-1)}{N-d} \Sigma_i$ and $N = \sum_{i=1}^{d} n_i$

With 1000 permutation samples under H_0 each of length equal the total sample size and ignoring the group labels, we computed the p-value from the empirical distribution of the M statistic to assess the null hypothesis of homogenous covariance matrices. An M value greater than expected

under the H_0 leads to a rejection of H_0 [Zhang & Boos, 1992]. The p-value of the statistic is computed as

$$p = \frac{\#\{M_B > M\}}{B + 1} \tag{2.3}$$

where # is the number of, M_B are the statistics computed for each bootstrap sample, M is the observed statistic and B is the total number of bootstraps.

2.3 Results

The basic characteristics of the twelve datasets are presented on Table 2.1 while those of the six AML datasets are presented on Supplementary Table S1. Other details on all the datasets are shown on Supplementary Table S2. All downloaded datasets came from Affymetrix chips. The filtering step in most datasets yielded relatively low numbers of probesets compared to the initial

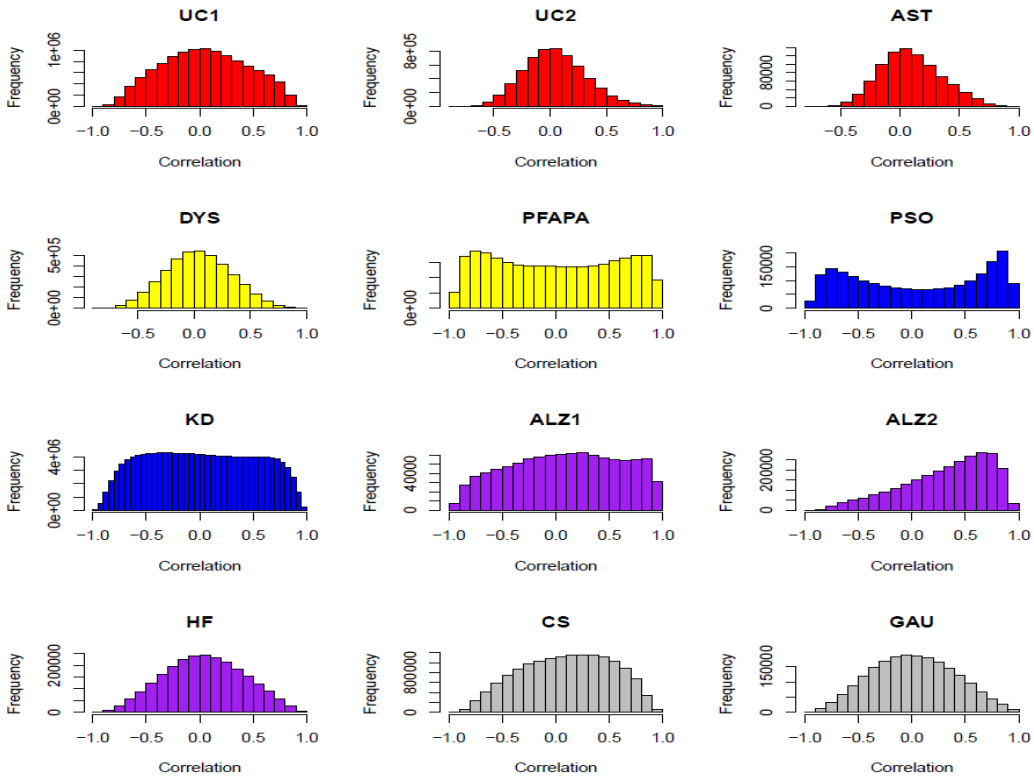

Figure 2.1: Distributions of the upper (lower) triangular values of the correlation matrix of probesets' pairwise Pearson correlations for the twelve datasets after filtering on both expression values > 5 in at least 10% of total samples and SD > 0.5. *Inflammatory (red), Infection (yellow), Immune (blue), Degenerative (purple) and Hereditary (gray) diseases.*

Table 2.2: Summary of absolute pair-wise correlations within each dataset before filtering and after each filtering method.

Filtering	Data	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Unfiltered	UC1	0.0000	0.0889	0.1871	0.2158	0.3143	0.9992
	UC2	0.0000	0.0883	0.1865	0.2162	0.3153	0.9923
	AST	0.0000	0.0837	0.1777	0.2096	0.3045	0.9948
	PFAPA	0.0000	0.1470	0.3074	0.3428	0.5089	0.9996
	DYS	0.0000	0.0971	0.2062	0.2385	0.3509	0.9876
	PSO	0.0000	0.1641	0.3368	0.3595	0.5343	0.9980
	KD	0.0000	0.1347	0.2824	0.3157	0.4671	0.9983
	ALZ1	0.0000	0.1592	0.3297	0.3566	0.5315	0.9981
	ALZ2	0.0000	0.1591	0.3258	0.3445	0.5124	0.9972
	HF	0.0000	0.1280	0.2677	0.2980	0.4411	0.9911
	GAU	0.0000	0.1244	0.2597	0.2911	0.4286	0.9998
	CS	0.0000	0.1082	0.2259	0.2547	0.3735	0.9976
Detection call filtering	UC1	0.0000	0.1028	0.2156	0.2456	0.3594	0.9992
	UC2	0.0000	0.1028	0.2157	0.2438	0.3583	0.9923
	AST	0.0000	0.0957	0.2024	0.2337	0.3425	0.9948
	PFAPA	0.0000	0.1690	0.3492	0.3764	0.5619	0.9996
	DYS	0.0000	0.1245	0.2609	0.2874	0.4277	0.9876
	PSO	0.0000	0.1913	0.3846	0.3962	0.5876	0.9980
	KD	0.0000	0.1744	0.3568	0.3777	0.5623	0.9997
	ALZ1	0.0000	0.1673	0.3444	0.3691	0.5501	0.9981
	ALZ2	0.0000	0.1581	0.3253	0.3457	0.5154	0.9972
	HF	0.0000	0.1456	0.3014	0.3268	0.4856	0.9911
	GAU	0.0000	0.1276	0.2662	0.2972	0.4380	0.9998
	CS	0.0000	0.1364	0.2811	0.3060	0.4517	0.9981
Variance filtering	UC1	0.0000	0.1402	0.2951	0.3275	0.4911	0.9992
	UC2	0.0000	0.0752	0.1600	0.1938	0.2761	0.9923
	AST	0.0000	0.0782	0.1677	0.2049	0.2943	0.9948
	PFAPA	0.0000	0.2620	0.5137	0.4954	0.7306	0.9996
	DYS	0.0000	0.0803	0.1709	0.2041	0.2945	0.9876
	PSO	0.0000	0.3430	0.6046	0.5542	0.7827	0.9980
	KD	0.0000	0.2086	0.4191	0.4262	0.6347	0.9997
	ALZ1	0.0000	0.2051	0.4222	0.4402	0.6684	0.9981
	ALZ2	0.0000	0.2340	0.4579	0.4514	0.6663	0.9972
	HF	0.0000	0.1153	0.2431	0.2794	0.4096	0.9907
	GAU	0.0000	0.1365	0.2838	0.3176	0.4667	0.9998
	CS	0.0000	0.1655	0.3373	0.3591	0.5325	0.9981

number of probesets. The detection call retained on average between 59% and 72% of the initial number of probesets for all datasets with hereditary having the lowest and degenerative the highest for non-cancerous and 64% for AML datasets. Meanwhile, the variance filtering yielded lower proportions of probesets than the detection call filtering, ranging from 9% (inflammatory) to 28% (AML) with AML and Immune seen as the diseases with large proportion of probesets with high variability. These results are displayed on Supplementary Table S3. We explored the distribution of the pairwise correlations by plotting the histograms of the upper (lower) triangular values of each correlation matrix for the different filtering methods. These distributions are displayed on Supplementary Figures S1 and S2. Also, the distributions from both filtering methods

Table 2.3: Summary of clusters within datasets from PAM algorithms.

Disease	# of cluster(s)	# of probesets within a cluster	Per cluster median correlation
UC1	5	318 408 835 635 304	0.454 0.336 0.610 0.412 0.331
UC2	9	376 360 339 231 315 265 340 179	0.287 0.360 0.284 0.290 0.520 0.457 0.477 0.178 0.256
AST	6	307 329 288 129 101 139	0.423 0.338 0.318 0.192 0.212 0.129
PFAPA	5	941 1017 157 191 194	0.647 0.732 0.436 0.380 0.361
DYS	5	381 549 393 818 359	0.260 0.276 0.253 0.353 0.255
PSO	5	502 265 819 141 260	0.727 0.606 0.823 0.477 0.695
KD	12	188 255 181 147 243 494 125 234 209 193 115 146	0.651 0.744 0.609 0.628 0.629 0.814 0.622 0.582 0.658 0.633 0.606 0.575
ALZ1	11	152 118 158 259 156 86 70 145 59 61 91	0.811 0.854 0.892 0.830 0.758 0.589 0.591 0.783 0.461 0.410 0.531
ALZ2	9	87 344 325 592 160 338 165 131 153	0.500 0.709 0.738 0.806 0.639 0.632 0.491 0.527 0.478
HF	8	323 165 511 169 157 277 171 295	0.444 0.386 0.578 0.337 0.346 0.507 0.374 0.587
GAU	8	327 428 212 211 201 248 118 272	0.717 0.501 0.447 0.417 0.390 0.474 0.466 0.493
CS	12	232 146 241 110 352 189 229 134 106 240 361 160	0.496 0.454 0.581 0.391 0.703 0.495 0.613 0.385 0.495 0.686 0.737 0.466

#: Number

combined are displayed on Figure 2.1. From these figures, we see that the correlation structures clearly differ between datasets. Comparing Figures S1 and S2 in the Supplementary material and looking at the summary of the absolute pair-wise correlations within each data displayed on Table 2.2, one sees that the different filtering methods revealed different correlation distributions. The variance filtering method eliminates huge numbers of lowly correlated probesets rendering the data highly correlated than the detection call filtering as shown by the increase in the mean/median absolute pair-wise correlations on Table 2.2. This phenomenon prevailed on the correlation distributions of PFAPA, PSO and ALZ2. Confirming that variance filtering dominates the detection calls filtering, a combination of both yields similar distributions (Figure 2.1) as variance filtering only. In the ALZ2 dataset, the correlation was dominated by positive correlations. The ALZ1 dataset also has a large number of positive correlations but moderate number of negative correlations. A similar distribution was also observed in the KD dataset. The other datasets have correlations centered around zero, but with different skewnesses. UC1, GAU and HF cases cover the entire correlation range of $[-1, 1]$ while UC2, DYS tend to have correlations that do not cover the entire range. Figure S3 shows the pairwise correlation distributions for the six acute myeloid leukemia (AML) datasets and it can clearly be seen from this figure that the distributions are alike and in general normally distributed around zero except for AML3. A possible fact for high absolute

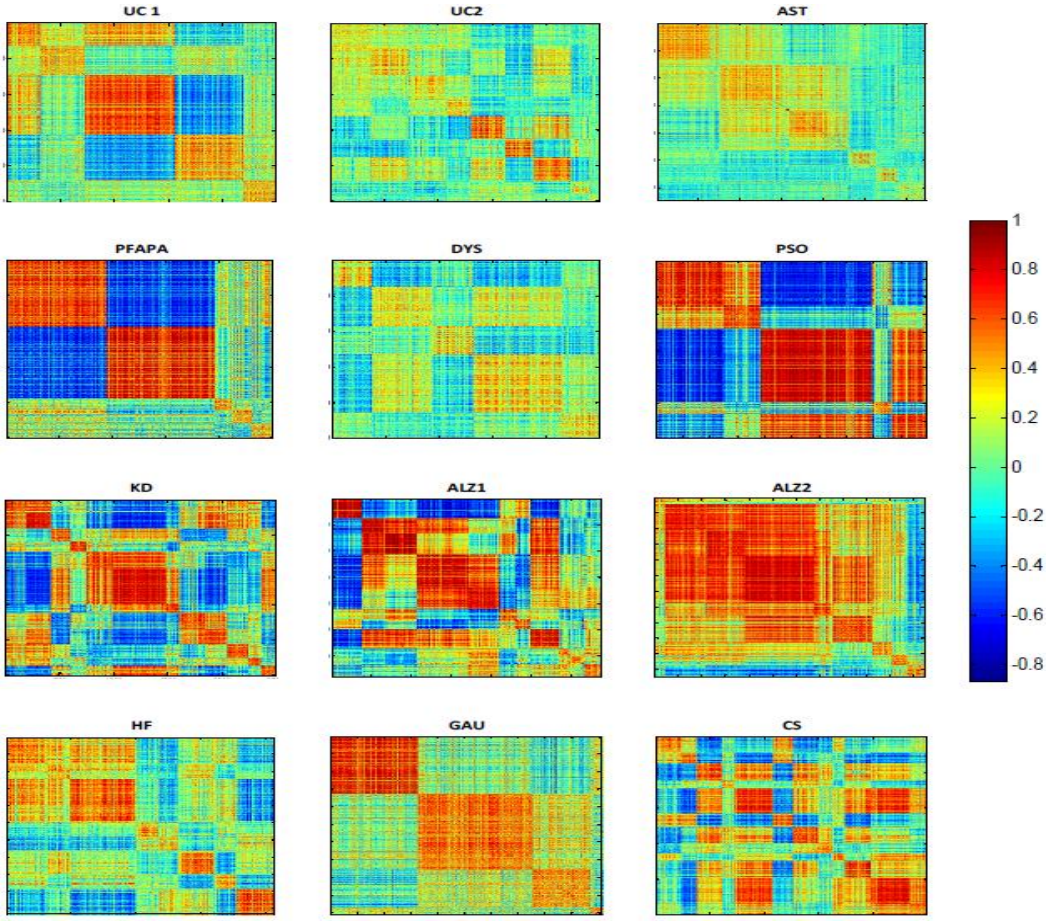

Figure 2.2: Correlation structure within each dataset based on clustering from PAM. *Values range from –1 (dark blue) to 1 (dark red). The diagonal rectangles represent the intra-clusters correlations while the off diagonal rectangles represent the inter-clusters correlations. A cluster size is represented by the size of the corresponding diagonal rectangle.*

correlations in this data could be due to the small sample size (n = 8) as compared to other datasets.

The clustering results from PAM are displayed on Table 2.3 and Figure 2.2. In this table, the number of clusters, the cluster sizes and the median values of the intra-cluster correlations for each dataset are presented. Almost all intra-clusters have positive correlations, indicating that probesets within each cluster are closely related. This is expected since clustering is on positive correlations. Some clusters are more closely related than others as indicated by the high median intra-cluster correlation values. As shown on the distributions of the pair-wise correlations (Figure 2.1), highly correlated data like ALZ1, ALZ2, CS, GAU, KD, PFAPA, PSO and UC1 contain at least one cluster with median correlation greater than or equal to 0.6.

Table 2.4: Summary of clusters within datasets from WGCNA algorithms.

Disease	# of cluster(s)	# of probesets within a cluster	Per cluster median correlation
UC1	4	545 242 1547 166	0.434 0.397 0.283 0.343
UC2	5	670 588 230 770 242	0.052 0.272 0.305 0.297 0.568
AST	2	490 803	0.209 0.142
PFAPA	2	1175 1325	0.626 0.460
DYS	3	277 126 2097	0.398 0.640 0.090
PSO	2	819 1168	0.567 0.714
KD	4	987 181 1195 137	0.457 0.492 0.520 0.519
ALZ1	3	485 224 646	0.662 0.616 0.441
ALZ2	1	2294 1	0.403
HF	4	382 358 1325 3	0.514 0.367 0.219 0.819
GAU	12	103 322 430 53 69 42 43 89 45 89 127 58 65 482	0.396 0.509 0.435 0.500 0.815 0.573 0.563 0.602 0.572 0.360 0.537 0.489
CS	3	422 2076 2	0.493 0.231 0.967

#: Number

In Figure 2.2, the intra-cluster correlations are represented by the diagonal rectangles from top-left to bottom-right that are dominated by dark red color showing highly positive correlations. Meanwhile, the off diagonal rectangles represent the inter-cluster correlations. Interestingly, the more positive intra-cluster correlations in two clusters, the more negative their inter-cluster correlations. This is obvious since we clustered based on correlations and it indicates that two highly positive correlated clusters should be negatively correlated otherwise they should have been a single cluster. This might be referred to as down and up regulated clusters. The phenomenon can clearly be observed in UC1 clusters 3 and 4, PFAPA clusters 1 and 2 and PSO clusters 1 and 3. On the other hand, if two clusters have low positive intra-cluster correlations they turn to have low negative or positive inter-cluster correlations as observed in AST clusters 1 and 2 and DYS clusters 2 and 3.

The probesets in each dataset were re-clustered using the WGCNA algorithm and the intra- and inter-clusters correlations are presented on Figure 2.3 and with cluster summaries on Table 2.4. In a similar manner, this table presents the number of clusters, the cluster sizes and the median values of the intra-cluster correlations for each dataset while Figure S4 displays the corresponding dendograms of the datasets. Each color represents a cluster with the gray color representing the cluster of noisy probesets. WGCNA produces relatively lower intra-cluster correlations in some datasets as compared to PAM. It can be clearly seen with the DYS dataset in Figure 2.3 that the

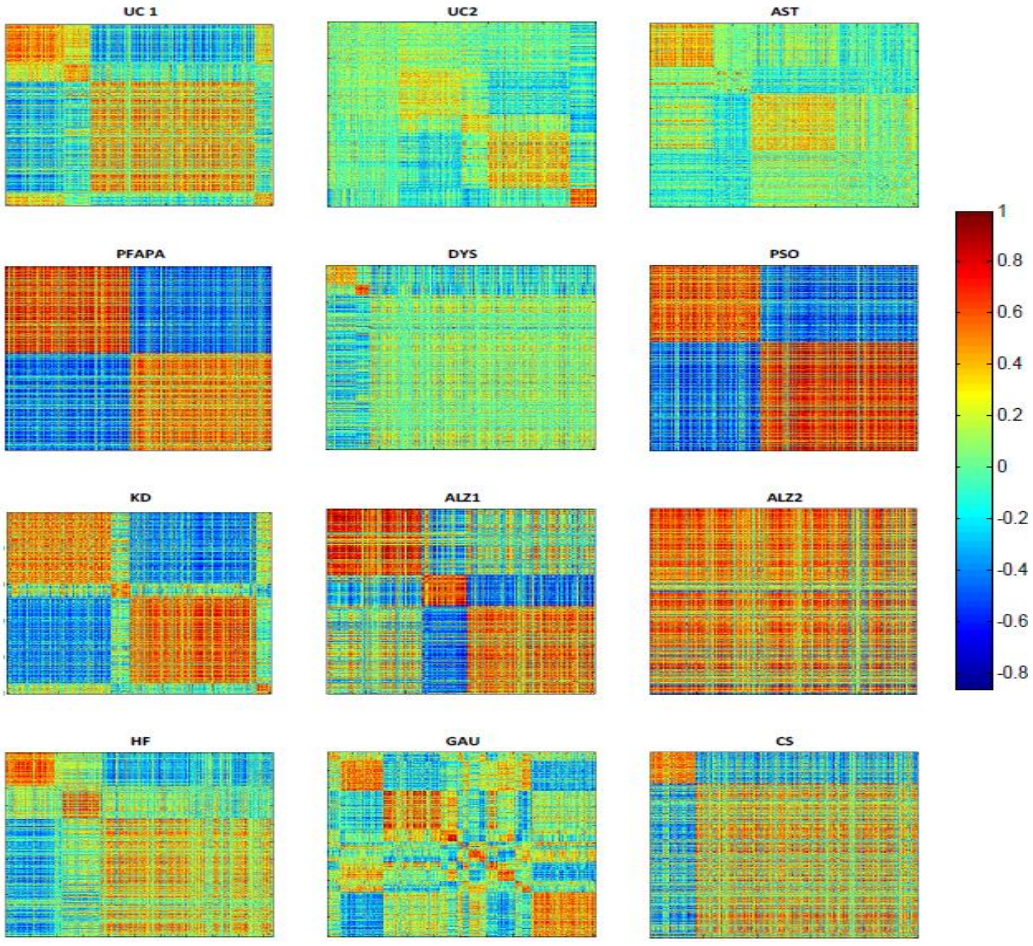

Figure 2.3: Correlation structure within each dataset based on clustering from WGCNA. *Values range from –1 (dark blue) to 1 (dark red). The diagonal rectangles represent the intra-clusters correlations while the off diagonal rectangles represent the inter-clusters correlations. A cluster size is represented by the size of the corresponding diagonal rectangle.*

third cluster (green cluster indicated in Figure S4) has very low intra-cluster correlation on the average. The same scenario holds for the first cluster in the UC2 dataset as shown on this figure. As in PAM, the more positive intra-cluster correlations in two clusters, the more negative their inter-cluster correlations. This can clearly be observed in the first two clusters in PFAPA and PSO datasets. Also, if two clusters have low positive intra-cluster correlations they tend to have low negative or positive inter-cluster correlations as in AST clusters 1 and 2 and DYS clusters 1 and 3. This serves as a confirmation that our clustering results from PAM are not unlikely. Since PAM and WGCNA produce similar results, we then perform clustering using PAM only, on the AML datasets and the results are displayed on Supplementary Table S4 and Figure S5. This table presents (in a

Table 2.5: Statistical test of homogeneity of correlations within degenerative diseases (ALZs); AMLs and between inflammatory and immune; inflammatory and hereditary; immune and hereditary; infectious and hereditary

Type	Comparison	Disease category(ies)	# probesets	Box's M Statistic	p-value
Non-cancerous	UC1 v/s KD	Inflammation v/s Immune	2646	37661.180	<0.001
	UC1 v/s CS	Inflammation v/s Hereditary	1708	44341.930	<0.001
	KD v/s CS	Immune v/s Hereditary	3032	32917.660	0.049
	PFAPA v/s GAU	Infection v/s Hereditary	582	9452.448	<0.001
	ALZ1 v/s ALZ2	Degenerative (Alzheimer)	298	22793.700	<0.001
Cancerous	Equality of AML1-6	AML	898	903539.900	0.746

v/s: versus #: Number of common

similar manner as in non-cancerous datasets) the number of clusters, the cluster sizes and the median values of the intra-cluster correlations for each AML dataset. The positive intra-cluster median correlations also represent the similarities of the probesets within a cluster as in non-cancerous datasets. From Figure S5, it is clear that AML datasets produce a larger number of clusters (average of 12.67 clusters) than non-cancerous data; with inflammatory having mean cluster size of 6.67, infectious 5, immune 8.5, degenerative 9.33 and hereditary 10.

Finally, the correlation structure of a few datasets on which we could extract common probesets based on a common microarray platform and/or disease type were used to compare within and between disease category homogeneity using the Box's M test. The number of common probesets in the comparisons made ranges from 298 to 3032 and the results of these comparisons are displayed on Table 2.5. The very highly significant results for non-cancerous diseases indicate that the correlation structures are heterogeneous across datasets. In a similar manner, we extracted a total of 898 common probesets in the six AML datasets and tested for homogeneity of correlation structures within these AML datasets using the Box's M test. The non-significant result indicates a homogeneous correlation structure. Supplementary Figure S6 shows the distribution of the pairwise correlations of the extracted probesets in each dataset with these common probesets and it is a clear representative of the overall distribution in each dataset. A visualization of the empirical distributions of the Box's M statistics for each and every comparison made can be observed on Supplementary Figure S7.

2.4 Discussion

We compared the correlation structures of twelve datasets across five non-cancerous disease types (inflammatory, immune, infectious, degenerative, and hereditary) and six datasets on one cancerous disease type (AML) and investigated the effect of filtering on the correlation structures

within datasets. We found that the correlation structures differ significantly between datasets. Furthermore, based on the selected gene expression data used in this study, datasets of the same disease (e.g., ALZ1 and ALZ2) have heterogeneous correlation structures. This could be associated to several reasons some of which include the medical question addressed, cell type, pathogen, organ source, etc., considered in the experiment. Nevertheless, there was no evidence not to assume homogeneity of correlation structures within AML datasets. This could be due to the fact that most AML studies utilized bone marrow as cell type and they address the same medical question (diagnosis). Nevertheless, this could also be due to the fact that most cancerous diseases are tissue specific and hence have less variability as compared to non-cancerous diseases. In addition, we found that filtering increases the proportion of highly correlated (in both directions) probesets as compared to the total number of initial probesets.

Classification functions like DLDA and SCDA are by nature designed to ignore correlations. Meanwhile, LDA and ridge regression (PLR) are designed to take into account correlations. As one of the most commonly used classification methods in gene expression analysis, support vector machine (SVM) though commonly understood as a method of finding the maximum-margin hyperplane, is generally seen as a regularization function estimation problem, corresponding to a hinge loss function with a quadratic penalty like that of ridge regression [Hastie et al., 2003; Ye et al., 2011]. There also exist variant types of SVM like lasso (L1 norm regularization) support vector machine [Zhu et al., 2004], which eliminates variables without taking into account correlated group of variables, and doubly (L1 and L2 norms) regularized support vector machine [Wang et al., 2006] which takes into account groups of correlated variables. In addition, Yang et al., (2006) pointed out that though support vector machine can study any non-linear relation, if a group of non-distinct (correlated) variables are selected as input variable set, the training time of support vector machine is lengthened and the errors become bigger. Hence, SVM and its variants will be affected by correlated variables. Furthermore, correlation patterns within gene expression data have been shown by Kim & Simon, (2011) to be a determinant of the accuracy of most classification functions used in building class prediction models. Nevertheless, the homogeneity or heterogeneity of correlation structures within gene expression datasets has not been studied in the literature.

We downloaded gene expression data across various disease types out of the field of cancer and six AML datasets, meeting predefined criteria from data repositories and then filtered out non-informative probesets in these datasets by first preprocessing the datasets with same standards

and then investigating the correlation structures from different filtering methods by visualizing the correlation distributions. It was observed that a particular filtering method may increase the relative proportion of highly correlated (in both directions) probesets in a gene expression dataset. Secondly, we combined two filtering methods and investigated the correlation structures within datasets and then clustered the probesets based on correlation using PAM and confirmed our clustering results using WGCNA. Though clustering of probesets with PAM based on correlations and gap statistic produced better correlated clusters than WGCNA algorithm, both algorithms revealed similar correlation structures for each and every dataset.

We observed the intra- and inter-clusters correlations within each and every dataset by visualizing the heatmaps of the correlation matrices. In other to investigate our hypothesis that correlation structures vary across gene expression datasets with possibility of similar correlations structures within data from the same disease type, we compared correlation matrices of independent datasets within and between disease categories. By visual inspection, we found that the correlation structures differ considerably across certain datasets, e.g., dystonia and psoriasis, while some correlation structures are very alike, e.g., UC2 and AST datasets. The correlation structures within inflammatory disease (UC1, UC2 and AST) are similar with moderately low numbers of high absolute correlation values and the correlation structure in AML datasets is homogeneous but this is not the case with infectious (PFAPA and DYS) and degenerative (ALZ1, ALZ2 and HF) diseases which turn to have differing correlation structures. In general, we found that the correlation structures within gene expression data vary from one experiment to another irrespective of the disease type. It is hard to generalize the correlation structure within datasets of the same disease type. Although our results show that there is homogeneity of the correlation structure in the inflammatory disease and AML, we averse to generalize this finding because we do not have a large number of independent datasets.

We found that correlation structures significantly differ across datasets even for datasets of the same disease while for AML datasets there seems to be no significant difference between the correlation structures. Since the literature shows that correlations do affect the performance of classification methods and giving our findings that correlation structures differ across datasets, it therefore becomes evident that the performance of classifiers that are sensitive to correlations will differ considerably across these datasets. As such, a classification method to be used in a particular data cannot be chosen at random or by familiarity but should be chosen based on the characteristics of the data under study (one of which is pairwise correlation between probesets)

and the classification function. We therefore recommend that researchers should at least explore the correlation structures within their data before making a choice of a classifier for building a predictive model. This could easily be done by plotting a distribution of the upper (lower) triangular of the pair-wise Pearson correlations matrix and assessing the scale and shape of the distribution. More especially how correlated the data is, together with classifier specific characteristics, might help in choosing an optimal classification function for the dataset at hand. Since AML datasets have been shown to have a homogenous correlation structure possibly due to a similar experimental design, it is worth stating that data from the same experimental procedure (disease under study, cell type considered, medical question addressed, etc.) might have a homogeneous correlation structure. Hence, a literature review of most classifiers that have been found to work well with similar data as the new data generated from the same procedure might be helpful. However, this does not mean that a correlation analysis of datasets should not be carried out to help choose an optimal classification function.

Given the very high dimension of microarray datasets, it is computationally intensive to compute the entire correlation matrix of all the probesets. As such, we recommend that in the exploratory data analysis phase, a subset of a specified number of probesets representing the entire dataset could be sampled from the data because employing strict filtering to reduce the number of probesets to a few but moderately informative probesets may reveal a different structure by increasing the relative proportion of highly correlated probesets in both directions. For researchers with high performing computing facilities, exploring the entire dataset will be a better option. Also, knowing fully well that correlation structures differ from one disease to another and that it has an impact on the accuracy of a class prediction model [Kim & Simon, 2011], it will be worth studying the effect of the relative proportion of correlated probesets amongst other gene expression data factors, on a classifier's performance. Though we have proposed a method on how to explore how correlated a gene expression data is, we could not find a statistical methodology on how to quantify into a single number the correlation structure within a given dataset. Hence this topic and other data characteristics that might possibly affect the performance of a classification function remain a subject of investigation.

Acknowledgments

This study was financially supported by the VIRGO consortium, which is funded by the Netherlands Genomics Initiative and by the Dutch Government (FES0908). The funding agencies in no way influenced the outcome or conclusions of the study.

Supplementary Material

The online version of this article (DOI: 10.1515/sagmb-2014-0003) offers supplementary material, available to authorized users, and include:

Figure S1: Distributions of the upper (lower) triangular values of the correlation matrix of probesets pairwise Pearson correlations for the twelve datasets after filtering on expression values >5 in at least 10% of total samples. *Inflammatory (red), Infection (yellow), Immune (blue), Degenerative (purple) and Hereditary (gray) diseases.*

Figure S2: Distributions of the upper (lower) triangular values of the correlation matrix of probesets' pairwise Pearson correlations for the twelve datasets after filtering on S.D. > 0.5. *Inflammatory (red), Infection (yellow), Immune (blue), Degenerative (purple) and Hereditary (gray) diseases.*

Figure S3: Distributions of the upper (lower) triangular values of the correlation matrix of probesets retained in the AML dataset after filtering on both expression values >5 in at least 10% of total samples and S.D. > 0.5.

Figure S4: The number of clusters within each dataset from the WGCNA algorithm. *Each color represents a cluster of correlated (co-expressed) probesets. Gray color represents noisy probesets. Most datasets lack the gray colors indicating that there were no noisy probesets due to our strict filtering criteria.*

Figure S5: Correlation structure within AML datasets based on clustering from PAM. *Values range from -1 (dark blue) to 1 (dark red). The diagonal rectangles represent the intra-clusters correlations while the off diagonal rectangles represent the inter-clusters correlations. A cluster size is represented by the size of the corresponding diagonal rectangle.*

Figure S6: Distributions of the upper (lower) triangular values of the correlation matrix of 898 common probesets pairwise Pearson correlations for the AML datasets used for the Box's M test.

Figure S7: Empirical distributions of Box's M statistic using 1000 permutation samples under the null hypothesis for each and every comparison. *The red vertical line indicates the observed Box's M statistic.*

Table S1: Descriptions of AML datasets

Table S2: Other data characteristics

Chapter 2

Table S3: Average percentage of probesets retained from each filtering in each disease category

Table S4: Summary of Clusters within AML datasets from PAM algorithm

Chapter 3

Factors affecting the accuracy of a class prediction model in gene expression data.

Putri W. Novianti, **Victor L. Jong**, Kit C. B. Roes & Marinus J. C. Eijkemans

BMC Bioinformatics 2015; **16**(199)

Abstract

Background: Class prediction models have been shown to have varying performances in clinical gene expression datasets. Previous evaluation studies, mostly done in the field of cancer, showed that the accuracy of class prediction models differs from dataset to dataset and depends on the type of classification function. While a substantial amount of information is known about the characteristics of classification functions, little has been done to determine which characteristics of gene expression data have impact on the performance of a classifier. This study aims to empirically identify data characteristics that affect the predictive accuracy of classification models, outside of the field of cancer.

Results: Datasets from twenty five studies meeting predefined inclusion and exclusion criteria were downloaded. Nine classification functions were chosen, falling within the categories: discriminant analyses or Bayes classifiers, tree based, regularization and shrinkage and nearest neighbors methods. Consequently, nine class prediction models were built for each dataset using the same procedure and their performances were evaluated by calculating their accuracies. The characteristics of each experiment were recorded, (i.e., observed disease, medical question, tissue/cell types and sample size) together with characteristics of the gene expression data, namely the number of differentially expressed genes, the fold changes and the within-class correlations. Their effects on the accuracy of a class prediction model were statistically assessed by random effects logistic regression. The number of differentially expressed genes and the average fold change had significant impact on the accuracy of a classification model and gave individual explained-variation in prediction accuracy of up to 72% and 57%, respectively. Multivariable random effects logistic regression with forward selection yielded the two aforementioned study factors and the within class correlation as factors affecting the accuracy of classification functions, explaining 91.5% of the between study variation.

Conclusion: We evaluated study- and data-related factors that might explain the varying performances of classification functions in non-cancerous datasets. Our results showed that the number of differentially expressed genes, the fold change, and the correlation in gene expression data significantly affect the accuracy of class prediction models.

3.1 Introduction

As one of the major types of analyses for gene expression studies, supervised learning or classification has received high attention. Studies vary from the application of supervised methods to real-life problems [Bansard et al., 2011; Kabakchiev et al., 2010; Scian et al., 2011], methods comparisons [Dudoit et al., 2002; Statnikov et al., 2005] and methods development [Guyon et al., 2002; Friedman et al., 2010]. Methods to build predictive models are widely available in the literature and it had been shown that the performance of a classification method varies, depending on the dataset to which the method is applied [Lee et al., 2005]. The characteristics of a dataset that naturally could be handled by a classification function might be one of the underlying reasons accounting for this variability. A classical method like linear discriminant analysis works under an assumption of the equality of covariance matrices between classes; while penalized logistic regression could handle a dataset with strongly correlated variables. Other specific study factors had also been shown to determine the predictive ability of a classification model, such as model building technique, array platform, clinical problem and sample size [Shi et al., 2010; Ntzani & Ioannidis, 2003]. Most of these characteristics are related to the technology or procedure and not to the specific data at hand. The characteristics of a gene expression dataset together with the nature of a classification function may play a key role in yielding a good class prediction model for gene expression data.

Evaluation studies on the aforementioned factors were based on classification models within the field of cancer. The effect of these factors might differ on datasets from non-cancerous diseases. This is because most cancerous diseases are often tissue specific unlike non-cancerous diseases that might involve the entire system and hence have different complexities. As one of gene expression data characteristics that has been proven by Kim & Simon, (2011) to have an effect on the performance of probabilistic classifiers when calibration and refinement scores were used as model evaluation measurements, correlation structures have been shown to differ significantly between datasets from both cancerous and non-cancerous diseases [Jong et al., 2014]. These findings had led to the question, what factors do affect the performance of class prediction models on datasets from non-cancerous diseases. As such, a literature review study to quantify the association between study factors and the performance of classification methods outside the field of cancer was initiated [Novianti et al., 2014]. The study, however, was limited to the characteristics of the microarray experiment, without investigating the effect of gene expression data characteristics such as the correlation between genes.

In this study, we outline potential study and data specific factors and assess their contribution to the accuracy of classification functions using real-life gene expression data. The factors were chosen from both the experimental settings of the studies (i.e., disease, medical questions, tissue/cell types and sample size) and the characteristics of the gene expression data (i.e., the number of differentially expressed genes, the fold changes and the within-class correlations).

3.2 Methodology

3.2.1 Data extraction

We downloaded microarray gene expression datasets from the ArrayExpress data repository. The criteria for selecting the datasets were that the experiments 1) had been conducted in humans; 2) outside the field of cancer; 3) had samples with class labels in at least two classes; 4) were published after 2005; and 5) provided raw cell files. To reduce the source of variability of classification model performances because of the array used in the experiments, we retained studies conducted with only Affymetrix arrays. This additional exclusion criterion was also motivated by the wide used of Affymetrix array by studies that were recorded in the ArrayExpress repository. Out of 54169 recorded studies in the ArrayExpress, 21284 (39.2%), 4436 (8.2%) and 3896 (7.2 %) studies used Affymetrix, Illumina and Agilent array, respectively (last checked in November 18, 2014). We took only two disease classes or dichotomized the outcomes if there were more than two classes in a study. In total, we downloaded twenty five gene expression datasets [Kabakchiev et al., 2010; Arijs et al., 2009; Toedter et al., 2011; Lee et al., 2011; Olsen et al., 2011; Wu et al., 2007; Walter et al., 2010; Hyrcza et al., 2007; Suarez-Farinas et al., 2010; Ogata et al., 2009; Mootha et al., 2003; Blalock et al., 2004; Bronner et al., 2009; Scherzer et al., 2007; Greco et al., 2012; Bochukova et al., 2010] briefly described in the Supplementary Material (Additional file 1) and summarized on Table 3.1. In addition to the extracted datasets, the following study characteristics were recorded: medical question addressed, disease type, tissue/cell type, microarray platform, paper availability, year of publication and sample size. The twenty five gene expression datasets came from microarray studies that were conducted in thirteen different diseases. We grouped the diseases based on etiology resulting in five major types namely; inflammatory (10), infectious (4), immune (4), degenerative (4), and hereditary (3) diseases. The disease grouping was aimed to evaluate the potential effect of the disease complexity to the performance of the classification methods.

Table 3.1: Characteristic of the gene expression experiments

Disease	ID+	Medical question	Disease class	Cell/Tissue type	Affymetrix platform	Citation *	N	p	Ndeg	fc	cc
UC1	E-GEOD-14580	Response to treatment (non-/responder)	Inflammation	Colonic mucosal biopsies	HG U133 Plus 2.0	yes	24 (16,8)	4650	623	1.551	0.162
UC2	E-GEOD-21231	Response to treatment (non-/responder)	Inflammation	Blood	HG 1.0 ST	yes	40 (20,20)	3388	0	0.207	0.112
UC3	E-GEOD-36807	Diagnostic (UC/CD)	Inflammation	Intestinal biopsy	HG U133 Plus 2.0	no	28 (15,13)	6541	21	2.222	0.305
UC4	E-GEOD-23597	Response to treatment (non-/responder)	Inflammation	Colonic biopsy	HG U133 Plus 2.0	yes	14 (7,7)	4793	0	1.119	0.298
UC5	E-MTAB-331	Diagnostic (UC/CD)	Inflammation	CD8+ T cell	HG 1.0 ST and HG 1.1 ST	yes	59 (30,29)	1402	312	0.714	0.164
UC6	E-GEOD-9452	Diagnostic (with/without inflammation)	Inflammation	Colon	HG U133 Plus 2.0	yes	17 (8,9)	3702	2401	3.697	0.165
UC7	E-GEOD-6731	Diagnostic (UC/CD)	Inflammation	Colon	HG U95AV2	yes	30 (11,19)	1055	0	0.485	0.228
AST1	E-GEOD-27011	Diagnostic (mild/severe)	Inflammation	Blood	HG 1.0 ST	no	36 (19,17)	1293	39	0.302	0.113
AST2	E-GEOD-51392	Diagnostic (asthma/rhinitis)	Inflammation	Bronchial epithelial cells	HG U133 Plus 2.0	no	11 (6,5)	3969	0	1.805	0.171
AST3	E-GEOD-31773	Diagnostic (non/severe)	Inflammation	CD4 T cells	HG U133 Plus 2.0	no	12 (4,8)	18321	14488	16.964	0.317
DYS	E-GEOD-19419	Diagnosis (carrier/symp)	Infection	Blood	HG 1.0 ST	yes	45 (22,23)	2811	0	0.182	0.153
HIV1	E-GEOD-35864	Diagnostic (HIV/HIV with complication)	Infection	Basal ganglia	HG U133 Plus 2.0	no	18(6,12)	8737	0	1.14	0.346
HIV2	E-GEOD-14278	Prognostic (resistant/susceptible)	Infection	Peripheral blood	HG U133 Plus 2.0	no	18 (9,9)	11286	4	0.58	0.12
HIV3	E-GEOD-6740	Diagnostic (chronic/non chronic)	Infection	CD4 T cell	HG U133A	yes	15 (10,5)	865	5	0.74	0.168
PSO	E-GEOD-18948	Response to treatment (non-/responder)	Immune	Blood	HG U95	yes	16 (7,9)	1987	34	1.131	0.369
KD	E-GEOD-16797	Response to treatment (IVIG responsive /non)	Immune	Blood	HG U133 Plus 2.0	yes	12 (6,6)	11043	5	1.688	0.224
Dia1	E-GEOD-18732	Diagnostic (type 2 diabetes/intolerant)	Immune	Skeletal muscle	HG U133 Plus 2.0	no	71 (45,26)	2038	10	0.279	0.16
Dia2	E-CBIL-30	Diagnostic (diabetes type 2 / abnormal glucose)	Immune	Skeletal muscle	HG U133A	yes	26 (18,8)	1749	0	0.269	0.435
ALZ1	E-GEOD-1297	Diagnostic (severe/not severe)	Degenerative	Hippocampus	HG U133A	yes	22 (7;15)	2295	13	0.693	0.287
ALZ2	E-MEXP-2280	Diagnostic (Alz/Pick's disease)	Degenerative	Medial temporal lobe	HG U133 Plus 2.0	yes	19 (7,12)	6899	1592	1.086	0.231
PARKI	E-GEOD-6613	Diagnostic (Parkinson/non-Parkinson)	Degenerative	Blood	HG U133A	yes	83 (50,33)	638	0	0.192	0.361
HF	E-GEOD-26887	Diagnostic (with/-out Diabetes)	Degenerative	Left ventricle cardiac biopsies	HG 1.0 ST	yes	19 (7,12)	2068	0	0.374	0.131
GAU	E-GEOD-21899	Diagnostic (type 1/ 3)	Hereditary	Skin	HG U133A 2.0	no	10 (5,5)	2017	4	1.807	0.143
CS	E-MEXP-2236	Diagnostic (Apert/Muenke)	Hereditary	Skin	HG U133 Plus 2.0	yes	20 (10;10)	5422	21	0.59	0.255
CF	E-GEOD-10406	Diagnostic (Chronic rhinosinusitis/+Cystic fibrosis)	Hereditary	Sinus mucosa	HG U133 Plus 2.0	no	15 (9,6)	7604	0	0.786	0.206

\+ : The ArrayExpress accessing ID
* : Paper availability
Ndeg : The number of differentially expressed probesets
fc : The average fold change from all probesets
cc :The average within class correlation values from all probesets

3.2.2 Preprocessing

The raw datasets were normalized using quantile normalization, background correction performed according to manufacturer's platform recommended correction and log base two transformed [Ogata et al., 2009]. Median polish was used as a summarization method to quantify expression values because of its ability to deal with outlying probes [Mootha et al., 2003]. For each dataset, we filtered out non-informative probesets using two filtering criteria. First, we retained probesets that had expression value greater than five in at least ten percent (10%) of the total samples. Secondly, we filtered the retained probesets whose standard deviations were greater than 0.5 (sd > 0.5). We refer to the retained list as the actual expression data.

3.2.3 Classifier building

We built and evaluated in each dataset class prediction models with the set of nine classifiers described in the classification functions subsection. Since we are only equipped with a finite sample and the underlying distribution is unknown, the empirical counterpart to the generalization accuracy of a classification function *f* is estimated as

$$R_{emp}[f] = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) \tag{3.1}$$

where n is the number of available samples and $L(.,.)$ is a loss function with $L(u,v) = 1\ if\ u = v$, $L(u,v) = 0$ otherwise [Slawski et al., 2008].

Though this empirical counterpart to the generalization accuracy can be used to evaluate classifiers, it usually overfits the sample $\mathcal{S}$. A general practice is to split the samples into a learning set $\mathcal{L}$ and a testing set $\mathcal{T}$. Predicted value from a classification function $\hat{f}(.)$ is constructed from a learning set $\mathcal{L}$ only and evaluated using a testing set $\mathcal{T}$ [Slawski et al., 2008]. In case sample sizes are very small, a good practice is to generate several learning and testing sets from the available sample, construct a classifier with each learning set and using the corresponding testing set, estimate the empirical generalization accuracy. The final empirical generalization accuracy is the average across the testing sets. Suppose B learning sets $\mathcal{L}_b$ $(b = 1, \dots B)$ are generated from sample $\mathcal{S}$ and the corresponding testing sets $\mathcal{T} = \mathcal{S}\backslash\mathcal{L}_b$ with $\hat{\mathrm{f}}_{\mathrm{b}}(.)$ obtained from $\mathcal{L}_b$ $(b = 1, \dots B)$ then an estimate of the accuracy is calculated by

$$acc = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{|\mathcal{T}_b|}\sum_{i\epsilon T_b} L(y_i, \widehat{f_b}(x_i)) \tag{3.2}$$

where $|\cdot|$ is the cardinality of the considered set [Slawski et al., 2008].

As such, each dataset was split into two-thirds for the learning set and one-third for the testing set taking into account the number of samples per class (i.e. stratified sampling), using Monte Carlo cross-validation (MCCV) [Slawski et al., 2008] and the probesets were ranked using the moderated t-statistic [Smyth, 2004] on the learning set. The learning set was further split into an inner-learning set and an inner-testing set using leave one out cross-validation (LOOCV)., The parameter(s) of the classification functions (if any) were tuned by ranking the probesets on the moderated t-statistic and building the classifier with different values of the parameter(s) using the inner-learning set and evaluated with the out of bag inner-testing set as proposed by Wessels et al., (2005). The number of top probesets to be included in the classification function was also determined among $\mathrm{p} = 5, 10, 15, 20, 25, 50, 55$ for non-discriminant and $\mathrm{p} = 2, 3, 4, 5$ (except for the GAU dataset, $\mathrm{p} = 2, 3$) for linear discriminant analysis (LDA) and diagonal linear discriminant analysis (DLDA) using the corresponding inner-learningand inner-testing sets. The restriction of the top probesets for the discriminant functions is due to the inability of these functions to accommodate a number of probesets greater than the number of samples. With the optimal probeset(s) and number of top probesets (p) for each classification function, the class prediction model was built for each classification function using the learning set and then evaluated within the testing set. The process was repeated $\mathrm{B} = 100$ times. The numbers of correctly-classified and misclassified samples in both learning and testing sets were then recorded.

3.2.4 Classification functions

The nine classification functions were chosen to represent the broad list in the literature that falls within the categories: discriminant analyses or Bayesian (linear discriminant analysis (LDA), diagonal linear discriminant analysis (DLDA), and shrunken centroid discriminant analysis (SCDA)), tree base (random forest (RF) and tree-based boosting (TBB)), regularization and shrinkage (RIDGE, LASSO and support vector machines (SVM)), and *k*-nearest neighbors (*k*NN) methods *K*-nearest neighbour (*K*NN).

- *Linear discriminant analysis (LDA):* Discriminant analyses are Bayes optimal classifiers, which assume that the conditional distributions of predictors given the classes are multivariate normally distributed and the within-class covariance matrices are equal for all classes [Slawski et al., 2008]. In order to get an optimum LDA classifier, we optimized the number of probesets to be included in the model.

- *Diagonal linear discriminant analysis (DLDA):* As LDA, DLDA also works under the assumption of multivariate normality of class densities and a diagonal within-class covariance matrix for each class [Slawski et al., 2008]. The optimum number of probesets was tuned by cross-validation.
- *Shrunken centroid discriminant analysis (SCDA):* It is also well-known as the prediction analysis of microarray (PAM) and it is specially developed to handle the high-dimensionality of gene expression microarray data. The method works by shrinking the class centroids to the overall centroid. For binary classification, the mean for each probesest j in each class k is calculated, and is called the class centroid. The class centroids are first normalized by overall mean, pooled standard deviation and sample size. This normalized class centroid is denoted by $\mathrm{d_{jk}}$. The goal of this method is to shrink d_{jk} towards zero by reducing d_{jk} by an amount of Δ. A large Δ value implicitly means excluding more probesets, which lead to a reduction in the model complexity. On the other hand, less number of probesets in a model would increase the risk of excluding informative probesets [Tibshirani et al., 2003]. To balance this trade-off, parameter Δ was optimized amongst the following values: $0.1, 0.25, 0.5, 1, 2, \text{and } 5$. SCDA is categorized as an embedded filtering method because of its ability to do filtering and model building simultaneously [Saeys et al., 2007].
- *Random forest (RF):* Random forest is a classification method designed for decision tree classifiers. It combines the predictions made by multiple decision trees to yield the final prediction of a test sample. Supposed the sample size of the training set is N, each tree is constructed by: (i) sampling with replacement a random sample of cases of size $\frac{2}{3}N$ and (ii) at each node, a random sample of predictor variables m sampled from all predictor variables is selected and the predictor variable with the best split based on a given objective function is used. Step (ii) above is repeated until the tree is grown to terminal nodes with minimum size k. The out-of-bag (oob) samples are used to evaluate the constructed tree. Randomization helps to reduce the correlations among decision trees so that the generalization accuracy of the classifier can be improved. A higher value for the minimum terminal node size k would possibly lead to smaller grown trees. Once multiple trees have been built, they are then combined by voting; that is each tree cast a vote at its terminal nodes [Breiman, 2001]. The parameters m and k are often optimized using cross-validation. In this study, we fixed the number of trees in a forest at 500 and the number of random probesets at each split m and

the minimum terminal nodes size k were tuned within the values $\left((0.1, 0.25, 0.5, 1, 2) * \sqrt{p}\right)$ and $(1,2,3)$, respectively. Where p is the total number of probesets.

- *Tree-based boosting (TBB):* Boosting is a classification method that combines the output of several "weak" classifiers to produce a powerful "committee" [Hastie et al., 2003]. It is an iterative procedure used to adaptively change the distribution of the training samples so that the base classifiers focus on samples that are hard to classify. Boosting assigns a weight to each learning sample and may adaptively change the weight at the end of each boosting round. These weights are then used either as a sampling distribution or can be used by the base classifier to learn a model that is biased toward higher-weight samples. The idea is to give all observations the same weights at the start, draw a bootstrap sample and build a classifier, which in this case is a classification tree (hence tree-based boosting) then test the classifier with all the subjects. The weights of misclassified subjects are increased in the next bootstrap sample thereby given them higher chances to be sampled. We optimized the number of trees (bootstrap samples) that falls within these following values: 50, 100, 200, 500 and 1000.
- *Ridge regression (RIDGE):* The L_2-penalization is used in logistic regression to shrink the less significant coefficients toward zero. The amount of shrinkage is controlled by a parameter λ, where larger λ implies a larger degree of shrinkage [Hastie et al., 2003]. The parameter λ of the penalization is a tuning parameter obtained by cross-validation $(\lambda = 0.0625, 0.125, 0.25, 0.5, 1, 2, 4, 8, \text{and } 16)$.
- *LASSO*: As in ridge regression, LASSO uses a penalization parameter (λ) to estimate the coefficients of logistic regression, this time using L_1-penalization. λ is interpreted as truncating the less significant coefficients, so that LASSO also works as a method for variable selection. We selected the optimum λ parameter within the range 0.1:0.9 by 0.1 using cross-validation [Hastie et al., 2003].
- *Support vector machines (SVM):* SVM classification [Schölkopf & Smola, 2002] is a binary classification method that fits an optimal hyperplane between two classes by maximizing the margin between the classes' closest points. The points lying on the boundaries are called support vectors, and the middle of the margin is the optimal separating hyperplane. Data points on the "wrong" side of the discriminant margin are weighted down to reduce their influence and it is controlled by the cost parameter C. For the nonlinear case, SVM uses a nonlinear mapping (via kernels) to transform the original training data into a higher dimension. Within this new dimension, it searches for the linear optimal separating hyperplane that is, a

"decision boundary" separating the tuples of one class from another. The SVM finds this hyperplane using the support vectors ("essential" training tuples) and margins are defined by the support vectors [Han and Kamber, 2006]. We used a linear kernel and the optimal cost parameter was obtained from $0.1, 1, 5, 10, 50, 100, 500$ using crossvalidation .

- *K-nearest neighbor (KNN):* For a sample $\mathcal{S}$, the KNN algorithm classifies this sample $\mathcal{S}$ based on a measure of distance between $\mathcal{S}$ and other learning samples. It finds the K samples in the learning set closest to $\mathcal{S}$ and then predicts the class of $\mathcal{S}$ by majority votes. The value K is usually specified by the user. It should be noted that if K is too small, then the nearest-neighbor classifier may be susceptible to over-fitting. On the other hand, if K is too large, the nearest-neighbor classifier may misclassify the test instance, because its list of nearest neighbors may include data that are located far away from its neighborhood [Tan et al., 2005]. The optimal value of K is chosen by cross-validating amongst $K = 1{:}\,10$ by 1.

3.2.5 Predictive factors

The study characteristics (referred to as "study factors") were evaluated for their effect on the performance of the classification methods. The factors were chosen from both the experimental settings of the studies and the characteristics of the gene expression data. We selected study factors that have been proven in the literature or intuitively have association with the performance of classification models. To represent the experimental setting, we chose study factors like medical question, sample size and cell/tissue type used in the experiment. The gene expression data were explored further to find the characteristics that might contribute to the performance of classification methods, namely the number of differentially expressed genes, fold changes and within-class pairwise correlations. The study factors are described as follows:

- *Medical question:* The medical questions were of different types: diagnostic, prognostic and response to treatment related studies. Diagnostic studies tend to have higher classification model performance than prognostic or response to a treatment studies, as experienced by e.g. [Willenbrock et al., 2004]. This factor also came out as one of the factors that was associated with classification model performance outside the field of cancer [Novianti et al., 2014]. We classified the medical questions of the experiments as either diagnostic or non-diagnostic.
- *Sample size:* Microarray datasets suffer from a severe curse of dimensionality. The impact of the number of samples used in the analysis was therefore investigated, particularly in the field of cancer by Ntzani & Ioannidis, (2003). The class imbalance is another point of consideration when building a classification model. It may introduce bias towards the majority class in a

prediction model and the classification performance will be overestimated, especially when the accuracy is used to evaluate the model [Blagus and Lusa, 2010]. The class imbalance factor is calculated as the number of samples in the majority class divided by the total sample size.

- *Cell type:* The tissue or cell type used in the experiment is likely to be dissimilar between studies and may impact the resolution of information and also the performance of classifiers. In a specific cancer case, like in acute myeloid leukemia (AML), the findings could be greatly affected by the cell type used in the experiment (e.g. in [E-GEOD-12662, E-GEOD-14924, E-GEOD-35010]). We therefore considered the cell type as one of the factors. We used a broad categorization of blood versus non-blood cell types.
- *The number of differentially expressed genes (pDEG):* For each dataset, we performed a differential expression analysis by fitting a linear model for microarray data (well-known as limma) [Symth, 2005] and controlling the false discovery rate (FDR at 5%) defined as expected proportion of false rejection among the rejected hypotheses using the Benjamini and Hochberg (BH) procedure [Benjamini & Hochberg, 1995].
- *The within-class correlation level (withincor):* We constructed the within-class correlation matrices for each dataset. A shrinkage approach was applied to estimate the correlation matrix to deal with the high dimensionality and sparsity [Schafer & Strimmer, 2005]. We took the average of absolute pairwise correlations within each class and averaged those values over the two classes to represent the level of the within-class correlation coefficient for a dataset.
- *The fold change (fc):* We calculated the fold change for each actual probeset as the absolute difference of the mean of log_2 expressions between samples in two groups, divided by the pooled standard deviation. We summarized the fold changes in each dataset as the mean fold changes from all probesets.

3.2.6 Random effects logistic regression

The nine classification models were built in the twenty five gene expression microarray datasets. We considered these datasets as clustered data, where the selected studies and the classification methods act as clusters. Further, in each study, we treated the accuracy as a grouped binomial variable, for which we had the number of samples that were correctly and incorrectly classified. We therefore evaluated the six aforementioned predictive factors for classification accuracy by a logistic random intercept regression model [Stijnen et al., 2010]. The logistic random effects model is the generalization of the linear mixed model to binomial outcomes. In this case, the sigmoid

logistic link function is applied to the common linear mixed model and the error distribution is binomial instead of normal.

As the accuracy is well known to be biased towards the majority class, the random intercept logistic model was corrected by the class imbalance level, which was always included in the regression model. For the l^{th} study factor, the random effects model is written as

$$log\left(\frac{\pi(x_{isM})}{1-\pi(x_{isM})}\right) = (\beta_0 + \vartheta_{0s} + \vartheta_{0M}) + \beta_1 class\_imbalance_S + \beta_2 predictive\_factor_{lS} \quad (3.3)$$

where $\pi(x_{isM})$ is the probability of a sample i in study S to be correctly classified with the classification model M; ϑ_{0S} and ϑ_{0M} are the random intercepts with respect to study S ($\vartheta_{0S} \sim N(0, \sigma_{0S}^2)$); and classification method M ($\vartheta_{0M} \sim N(0, \sigma_{0M}^2)$). All the aforementioned study factors were evaluated by simple and multiple logistic random intercept regression models. Multiple regression evaluation was done by a forward selection approach. In each step, two nested models, with and without a particular study factor, were compared by Akaike's information criterion (AIC). Each factor l was also evaluated by its explained-variation of the random intercept variance term,

$$var_l = \frac{\sigma_{null}^2 - \sigma_l^2}{\sigma_{null}^2} \quad (3.4)$$

where σ_{null}^2 is the random intercept variance from a model with "class imbalance" only (referred to as null model). Since the logistic models have two random effects variables, σ_{null}^2 is the combined variance of the study (σ_{0S}^2) and the classification method (σ_{0M}^2) random effect from a null model. Meanwhile, σ_l^2 is the combined variance from a random effects model with the lth factor. The explained variation of all significant factors in the model (we refer to as "final model") was also evaluated. It was calculated by replacing the σ_l^2 in Eq.3.4 with the combined variance in the final model.

We evaluated the stability of the simple and multiple random effect logistic regression models by performing Jackknife resampling analysis. In each iteration, one study was left out and the model building process was repeated using the retained studies.

3.2.7 Software

All statistical analyses were performed in `R` software by using these following packages: `affy` for preprocessing procedures [Gautier et al., 2004], `CMA` for predictive modeling [Slawski et al., 2008], `limma` for fitting a linear model for microarray data [Smyth, 2005], `lme4` for random effects

Table 3.2: Individual random effect meta-regression

Study Factor	Coef*	AIC	P value	Individual explained-variation
Cell type	0.24^{+}	137.9	0.44	4.87%
Medical question	-0.32^{++}	137.8	0.38	2.55%
Sample size	-0.01	135.9	0.10	12.06%
The number of differentially expressed genes	0.21	116.0	<0.001	72.16%
Fold change	1.42	118.1	<0.001	57.31%
Within class correlation	1.74	137.5	0.31	5.80%

* :Coefficient of the corresponding study factor in the random effects logistic regression
\+ :Coefficient for the non-blood category in the Cell Type study factor
++ :Coefficient for the non-diagnostic category in the Medical Question study factor

linear model [Bates & Maechler, 2009] and `ggplot2` for data visualization [Wickham, 2009]. The R scripts are available in the Supplementary Material (Additional File 10).

3.3 Results

On average, most classification methods performed better on hereditary disease. Meanwhile, the highest variability of the classification performance was observed on infectious disease (Additional file 4: Figure S1). Of the 25 experiments selected, 19 experiments addressed a diagnostic study. Diagnostic studies tend to be easily classified and hence yield higher accuracies than other (prognostic or response to a treatment) studies, as experienced by Willenbrock et al., (2004). Despite this, the factor medical question is not significantly associated to accuracy (Additional file 5: Figure S2). A similar insignificant effect is also shown by cell type used in the experiment (Additional file 6: Figure S3). A more formal individual evaluation of the effect of each study factor to the predictive ability of a classification method was assessed by a random effects regression model as described in the Method section. The results of the individual evaluations are summarized in Table 3.2 and the individual explained-variation is depicted on Figure 3.1.

The *fc* and *pDEG* study factor were positively associated to accuracy in their respective univariate random effects models. This intuitive finding confirms that a classification model could possibly achieve a good performance as the genes' fold change or the number of differentially expressed genes increases (Additional file 7: Figure S4 and Additional file 8: S5). We transformed the *pDEG* to the log_2-scale to deal with the high variability of the number of differentially expressed genes among studies, which ranged from 0 to 14,488.

Further, *pDEG* and *fc* had a relatively high individually explained-variation, i.e., 72% and 57%, respectively. Given its highest individual effect on the performance of classification model, we then used *pDEG* as the first factor entering the multiple regression model that was constructed by

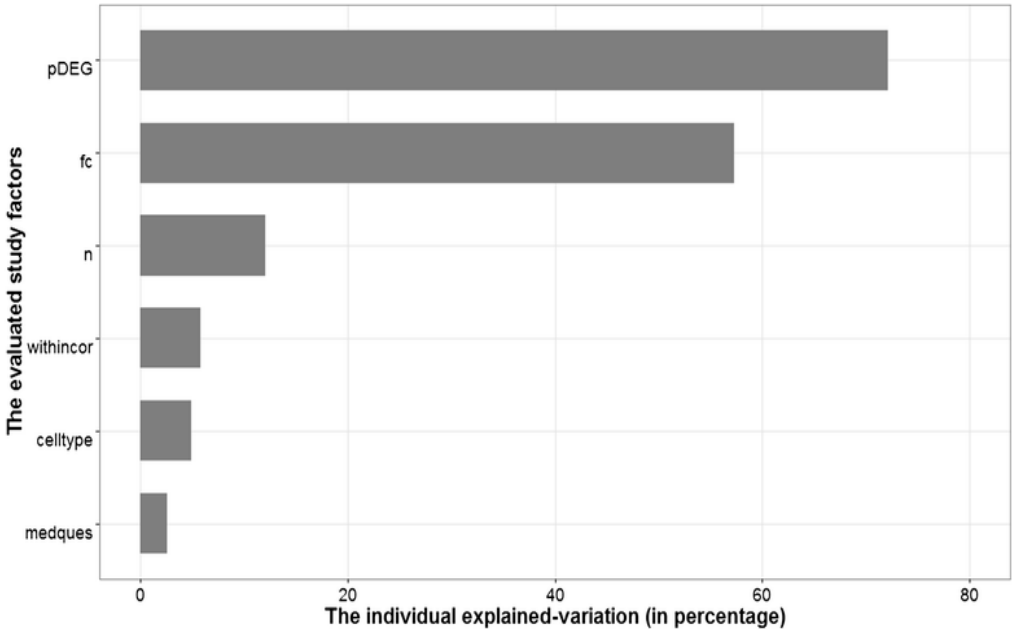

Figure 3.1: The individual explained-variation of study factors. *Abbreviations: the number of differentially expressed genes on the log scale (pDEG), the fold change (fc), the sample size (n), the average within-class correlation coefficient (withincor), the cell type (celltype), and the medical question (medques).*

the forward selection approach. We stopped the modeling process when there was no additional study factor that improved the multiple regression model, conditional on the previously selected study factors in the model. The forward selection procedure yielded *pDEG*, *fc* and the within class average correlation (*withincor)*as the factors that simultaneously associated to the classification models accuracy. We referred this model as the final model of the multiple random effects logistic regression. The three study factors in the final model explained 91.5% of the random between study variation relative to the null model. As in the univariable case, *pDEG* and *fc* have positive effects on the accuracy of classification methods. Interestingly, *withincor* turned out to be one of the study factors that significantly improved the multiple regression model, although it was not significant univariately.

Despite a relatively small number of studies, the random effects logistic regression model was stable, as shown by the high agreement of the random effects logistic regression models in the Jackknife resampling analyses. The Jackknife analysis was done by leaving out one study at a time and rebuilding the random effects regression model in the remaining studies. In the univariable evaluation of Jacknife resampling, the *fc* and *pDEG* study factors were always found to be significant in the random effects models. The *sample size*, however, came as one of significant

study factors five times, i.e., when *UC4*,*UC5*, *HIV3*, *KD* AND *HF* studies were left out from the random effect models (Additional file 2: Table S1). In the multivariable evaluation, the significant study factors in the final model were selected 19 times out of 25 Jackknife samples yielding a robustness of 76%. The *pDEG*, *withincor,* and *fc* were in the model for 25 times (100 %), 24 times (96 %) and 19 times (76 %), respectively (Additional file 3: Table S2).

3.4 Discussion

We enumerated possible characteristics of gene expression data and investigated their impact on the predictive accuracy of nine chosen classification methods using twenty-five downloaded gene expression datasets. While a substantial amount of information is known about the characteristics of classification methods, little has been done to determine which characteristics of gene expression data affect the performance of a classifier. Classification methods have been shown to have varying performances in gene expression datasets. The classification methods, on average, performed differently across the different disease types (Additional file 4: Figure S1), but the random effects logistic regression model failed to show a significant relationship between disease type and the accuracy of classification models. This might be as a result of the limited number of samples available to evaluate such a factor with five categories.

In general, we might have an issue of statistical power and model over-fitting when considering this variable. A solution could be to increase the number of studies by adding cancer studies to increase the statistical power and possibly lead to a comparison in different behavior of the study factors between cancerous and non-cancerous diseases. However, supervised learning on gene expression studies in the field of cancer have been studied extensively [Shi et al., 2010; Ntzani & Ioannidis, 2003; Dupuy & Simon, 2007]. As such, we chose to focus on microarray gene expression experiments outside the field of cancer. We assessed the stability of the results from both univariable and multivariable random effects logistic regression via Jackknife resampling. We excluded one dataset for each sampling and repeated the random effects modeling process. We then recorded P values of each study factor in univariable models and the study factors that were included in the model in multivariable evaluation. Large number of datasets needs to be included in order to yield more generalizable results and also to avoid underpowered findings, particularly in an evaluation or comparison study [Boulesteix, 2013]. Nevertheless, the evaluation of our results by Jackknife resampling shows high stability of our results and high agreement as compared to the findings by using full datasets

A similar study that was based on a quantitative review was conducted to evaluate study factors that were associated with the performance of classification models in the non-cancer field [Novianti et al., 2014]. That study had found that the cross-validation technique considerably affected the predictive ability of classification models, in line with the finding of MAQC II consortium study [Shi et al., 2010]. In the current study, we then controlled for the effect of cross-validation technique to observe the effect of other study factors that could not be observed earlier by Novianti et al., (2014). The same predictive modeling technique, including cross-validation, feature selection and classification functions, was applied to the preprocessed gene expression datasets. The performance of the optimum classification models were measured by calculating the proportion of correctly classified samples and total sample size. Random effects logistic regression models showed that gene expression data characteristics such as fold changes, the number of differentially expressed genes and the correlation between genes, contribute to the performance of classification models.

We used classification accuracy as the outcome of analysis. Although it is well-known to be a rough measure for the performance of a classification model, accuracy is widely used in practice due to its straightforward interpretation. In highly imbalanced datasets, accuracy may yield overoptimistic results, because a classification model might easily send all samples to the majority class. The class imbalance should therefore be taken into account when interpreting prediction accuracy [Dupuy & Simon, 2007]. A meaningful classification model necessarily should have higher accuracy than the proportion of the majority class. To deal with the problem of class imbalance when using accuracy, we corrected our random effects models for the class imbalance level.

We showed that the number of the differentially expressed genes, genes' fold changes and the average within-class pairwise correlations significantly affected the accuracy of classification models. The positive coefficient of the number of differentially expressed genes (*pDEG*) both in the simple and multiple random effects models shows that the classification models performed better if the number of differentially expressed genes present in a dataset is increased. Similarly, fold change (*fc*) was significant in both univariable and multivariable evaluations with positive effects. These intuitive findings were mentioned earlier by the MAQC II consortium study [Shi et al., 2010], where the number of informative genes had relatively high degree of explained variability of the classification model performance in cancer studies.

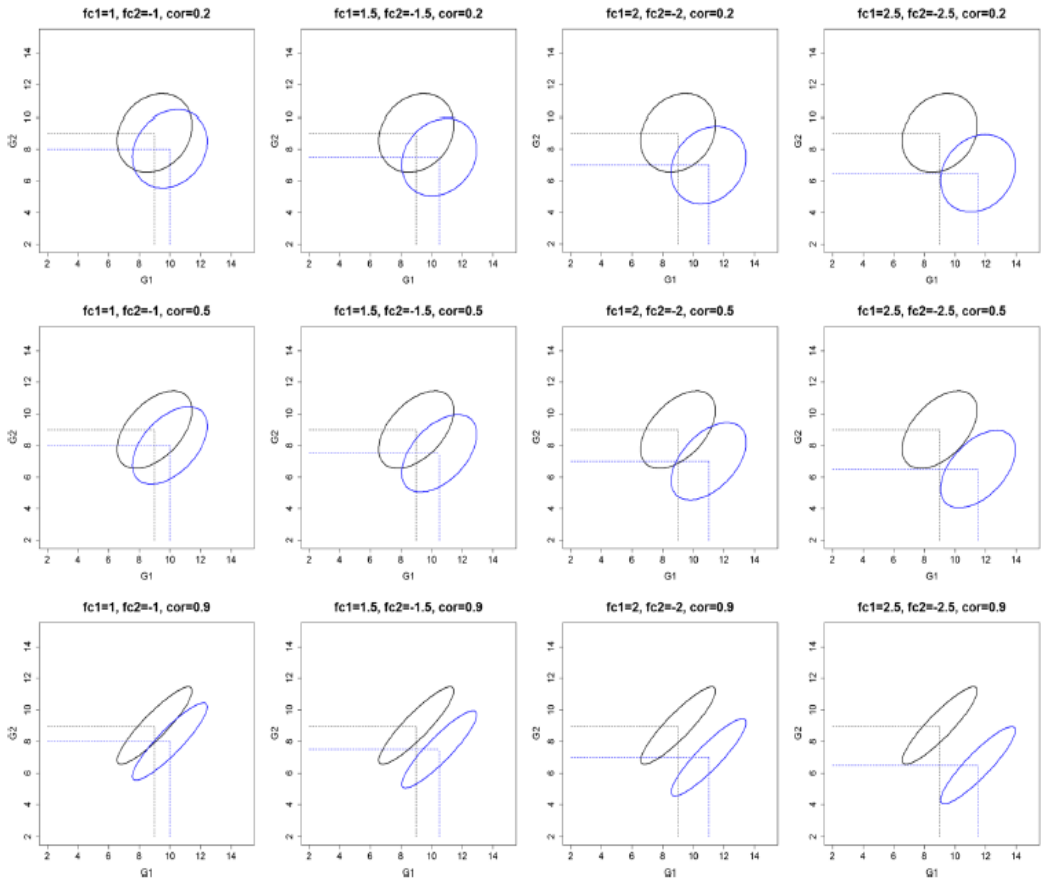

Figure 3.2: Visualization of the generated gene expression datasets with the scenario of fc1 = +, fc2 = −, cc1 = cc2 = +. *Abbreviations: fc1(2): fold change of gene 1 (2); cc1(2): correlation coefficient of gene 1 (2)*

The within-class correlation (*withincor*) has a positive effect on the accuracy of classification models together with *pDEG*and *fc* in the final random effect model. The positive effect of the *withincor* study factor to the classification model performance, is in contrast to knowledge from linear models that correlated variables bring no additional information to the model and therefore tend to reduce the predictive ability of the model. Our results show that the relationship between *withincor* and model accuracy is confounded by the *pDEG* and *fc.* To explain this finding, let's first consider the within class correlation between two genes, both with a certain fold change. The two classes are more separable when the pairwise within class correlation between two genes becomes stronger (Figure 3.2: one gene up- and the other down regulated and positive within class correlation and S7: both genes up regulated and negative within class correlation). Meanwhile, we hardly observe an effect of the within class correlation in the other possible

scenarios (Additional file 9: Figure S8: one gene up- and the other down regulated and negative within class correlation and S9: both genes up regulated and positive within class correlation). Thus, there are two possible effects of the within class correlation to the classification model's performance, i.e., either positive or no effect, which might be the reason for an overall significant positive coefficient of the *withincor* study factor.

The theoretical examples given above concern probesets with relatively high fold changes, reflecting the probesets that were involved in the classification models. In our classification approach, we ranked probesets based on the limma feature selection methods and used top-K probesets to feed the classifiers, as commonly done in practice [Kabakchev et al., 2010; Scian et al., 2011; Arijs et al., 2009; Menke et al., 2012; Rasimas et al., 2012; Lunnon et al., 2013] in non-cancer and [Willenbrock et al., 2004] in cancerous diseases. By using this approach, we ensured that the probesets involved in the classification models had considerable fold changes. Thus, it supports the confounding effect of the *fc* study factor to the *withincor* in the multivariable random effect regression model.

The correlation structure in gene expression data had been proven to have a negative impact on the performance of probabilistic classifiers [Kim & Simon, 2011]. This could possibly be due to the measure of evaluation and/or the fact that all genes were used and not a top number from a ranked list. In the non-probabilistic classifier, its effect has not been studied yet. The result of this study could be a preliminary proof of the effect of correlation between genes (or probesets) to the performance of general classification models (for both probabilistic and non-probabilistic classifiers). Given our results, a similar simulation study as that of Kim & Simon, (2011) by considering broad range of combination values of fold changes, the number of informative genes and correlation structure of a gene expression dataset, is worth initiating by applying both probabilistic and direct classification functions.

3.4.1 Conclusion

We evaluated factors that possibly had an impact on the performance of classification models in gene expression experiments outside the field of cancer. The factors were categorized into two main groups: the study- and the data-related factors. Our study showed that the data-related factors 'number of differentially expressed genes', 'fold change', and 'within-class correlation' significantly affect the accuracy of classification functions.

Acknowledgments

This study was financially supported by the VIRGO consortium, which is funded by the Netherlands Genomics Initiative and by the Dutch Government (FES0908). The funding agencies in no way influenced the outcome or conclusions of the study. The authors would like to thank T Debray (Epidemiology, Julius Center for Health Sciences and Primary Care, UMC Utrecht) and M Marinus (HPC-team, UMC Utrecht) who assisted the Authors in running the statistical analysis on the high performance computing cluster owned by the UMC Utrecht, the Netherlands. The Authors would also like to thank the anonymous Reviewers for their critical comments and constructive suggestions to the article.

Supplementary Material

The open access online version of this article (DOI: 10.1186/s12859-015-0610-4) offers supplementary material which include:

Additional file 1: A brief description of the selected microarray gene expression experiments.

Additional file 2: Table S1. Stability analysis of univariable random effects logistic regression models via Jackknife resampling.

Additional file 3: Table S2. Study factors that were included in the multivariable random effect logistic regression models via Jackknife resampling.

Additional file 4: Figure S1. Boxplot of Disease type against the classification model accuracy.

Additional file 5: Figure S2. Boxplot of Medical question against the classification model accuracy.

Additional file 6: Figure S3. Boxplot of Cell Type against the classification model accuracy.

Additional file 7: Figure S4. Plot of the Fold Change against the classification model accuracy.

Additional file 8: Figure S5. Plot of the Number of Differentially Expressed Genes (in the log scale) against the classification model accuracy.

Additional file 9: Figure S6. The visualization of the generated gene expression datasets with the scenario of fc1 = +, fc2 = +, cc1 = cc2 = −. Abbreviations: fc1(2): fold change of gene 1 (2); cc1(2): correlation coefficient of gene 1 (2).

Additional file 10: R script.

Chapter 3

PART II

PREDICTING AN OPTIMAL PREDICTIVE FUNCTION FOR A GIVEN DATASET.

Chapter 4

Selecting a classification function for class prediction with gene expression data.

Victor L. Jong, Putri W. Novianti, Kit C. B. Roes & Marinus J. C. Eijkemans

Bioinformatics 2016; **23**(12):1814-1822

Abstract

Background: Class predicting with gene expression is widely used to generate diagnostic and/or prognostic models. The literature reveals that classification functions perform differently across gene expression datasets. The question, which classification function should be used for a given dataset remains to be answered. In this study, a predictive model for choosing an optimal function for class prediction on a given dataset was devised.

Results: To achieve this, gene expression data were simulated for different values of gene-pairs correlations, sample size, genes' variances, deferentially expressed genes and fold changes. For each simulated dataset, ten classifiers were built and evaluated using ten classification functions. The resulting accuracies from 1152 different simulation scenarios by ten classification functions were then modeled using a linear mixed effects regression on the studied data characteristics, yielding a model that predicts the accuracy of the functions on a given data. An application of our model on eight real-life datasets showed positive correlations (0.33–0.82) between the predicted and expected accuracies.

Conclusion: The here presented predictive model might serve as a guide to choose an optimal classification function among the 10 studied functions, for any given gene expression data.

4.1 Introduction

Microarray gene expression profiling has become a widely used tool to identify particular disease subpopulations and to perform diagnostic and prognostic predictions [van 't Veer et al., 2002; Huang et al., 2010]. In clinical practice, they are used in diagnostic and prognostic analyses while in preclinical studies (toxicogenomics), they involve predicting the toxicity of compounds in animal models with the goal of speeding up the evaluation of toxicity for new drug candidates [Shi et al., 2010]. Though class prediction analysis is a common practice, the question that remains to be addressed is, given the wide availability of classification functions nowadays, which classification function do we use for a particular dataset? Classification functions have been shown to perform differently across gene expression datasets [Lee et al., 2005]. Moreover, the MAQC-II initiative has pointed out that classification function is one of the variables that explains the variability between gene expression class prediction performance [Shi et al., 2010].

While substantial amount of information is known about the characteristics of classification functions and class prediction building procedures, little is known about which data characteristics have impact on the performance of a class prediction model. For instance, diagonal linear discriminant analysis (DLDA) assumes no covariances and hence no correlations between variables and might fail if the data is highly correlated. On the other hand, linear discriminant analysis (LDA) assumes a common covariance matrix for the classes and thus to some extent, accounts for correlations [Hastie et al., 2003]. In addition, penalized regressions like ridge, lasso, elastic net are capable to handle correlated variables. Support Vector Machine (SVM), though commonly understood as a method of finding the maximum-margin hyperplane, may also be seen as a regularization function estimation problem, corresponding to a hinge loss function with a quadratic penalty as that of ridge regression [Hastie et al., 2003; Ye et al., 2011]. And it has been shown by Yang et al., (2006) that if a group of non-distinct variables are selected as input variable set, its training time lengthened and the errors become bigger. On the other hand, tree-based methods are by nature designed to capture interactions between variables while neural networks might capture other complex structures within a given dataset.

Given the above observations, it is obvious that the performance of these functions depends on the characteristics of the data in question. Despite this, the literature on how to choose a classification function for a given dataset is sparse. A common practice is comparing several classification functions and selecting the one with the minimum error rate but this has been pointed by Varma et al., (2006); Tibshirani &Tibshirani, (2009); Bernau et al., (2013) and Ding et

al., (2014) to lead to selection bias. As such, some experimenters adhere to one or a few classification functions irrespective of the dataset, disease or medical question being addressed. While others choose a classification function for their datasets by affinity or familiarity without taking into account the characteristics of such data.

A simulation study by Kim & Simon, (2011) shows that correlation is one of the data characteristics that affect the performance of most probabilistic classification functions. In addition, Jong et al., (2014) showed that correlation structures differ across gene expression data of different etiological diseases. The study by Novianti et al., (2015) shows that microarray gene expression data characteristics like $\log_2$ fold change of expression values, number of deferentially expressed genes and pairwise correlations between genes are associated to the accuracy of classification functions. However, this study was conducted in real-life gene expression datasets, where the magnitude and/or direction of association might have been confounded by unobserved data characteristics.

In this study, we aim to provide a guideline for making a choice of a classification function for a binary class prediction problem based on observed magnitudes and directions of the data characteristics, using accuracy as a measure of evaluation. We investigate the effect of sample size, proportion of deferentially expressed (DE) genes, genes' variances, log fold changes, pairwise correlations between DE and noisy genes on the accuracy of classification functions using extensive simulations.

The remainder of this article is organized as follows: methodology to simulate data, classification functions considered and the building and evaluation of class prediction models are presented in Section 4.2; Section 4.3 contains a predictive summary of the results of class prediction models for different simulated scenarios; Section 4.4 provides an application of our predictive model from the simulated results on real-life microarray gene expression datasets and Section 4.5 presents a discussion.

4.2 Methodology

4.2.1 Simulated data (scenarios)

To simulated gene expression data, we hypothesized that sample size, proportion of DE genes, genes' variances, log fold changes, pairwise correlations between DE and noisy genes might be associated to the performance of classification functions. These six variables were to be systematically varied in our simulations.

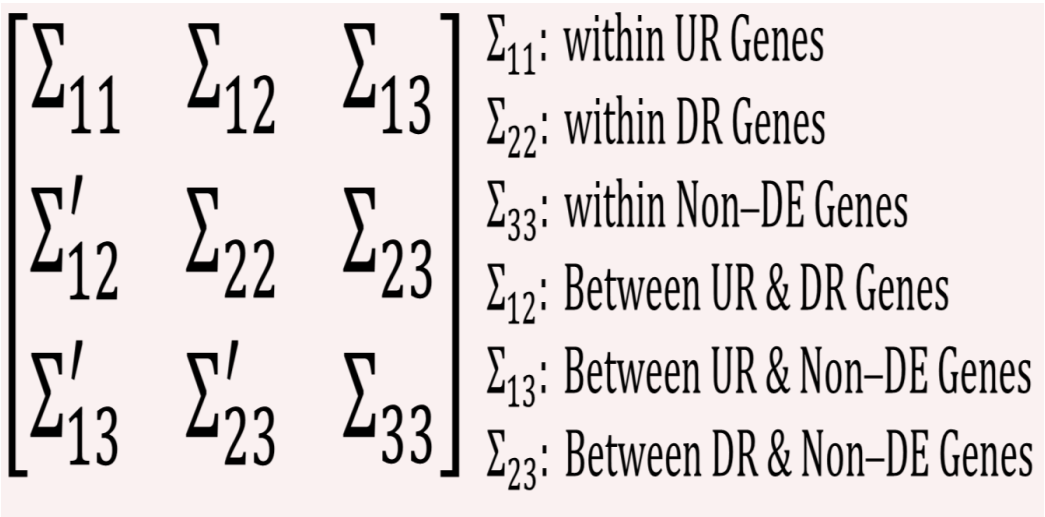

Figure 4.1: Assumed correlation structure. *Contains 3 clusters of up-regulated (UR), down-regulated (DR) and noisy (Non-DE) genes.*

From observed correlation structures in real-life gene expression datasets [Jong et al., 2014], we generalized the structure as shown on Figure 4.1, containing three clusters referred to as up-regulated (UR), down-regulated (DR) and noisy genes. The absolute values of pairwise correlation for DE genes (ρ) were varied as 0.00, 0.25, 0.50 and 0.75 with UR cluster taking oppositely-signed correlation values for DR cluster. The pairwise correlations both within the noisy cluster and between the noisy and the DE clusters were per gene-pair randomly drawn from a normal distribution centered at zero with a standard deviation θ i.e. $N(0, \theta)$ where $\theta = 0.00, 0.25, 0.50, 0.75$. The scenario ρ = θ = 0.00 corresponds to complete independence. Resulting correlation values lying outside the interval $[-1, 1]$ were uniformly converted to the intervals $[-1, -0.15]$ and $[0.15, 1]$ for negative and positive values respectively. The variances of the genes ($\sigma^2 = \frac{1}{\lambda}$) were drawn from an exponential distribution i.e. $exp(\lambda)$ where $\lambda = 0.25, 0.50, 1.00, 1.50$. The distributional assumptions were made based on observation from real-life datasets as experienced by Jong et al., (2014) and Novianti et al., (2015). With the correlation values and the variances, the within covariance matrices Σ_0 and Σ_1 were constructed for the two classes. In addition, the proportion of DE genes (π) was also allowed to take up 1%, 3% and 5% of the total number of genes, as values. This resulted to 192 different complex covariance matrices that were used to simulate the data for different values of other variables.

Table 4.1: Simulated gene expression data characteristics

Data characteristics	Values
Sample size (n)	20, 50, 100
Proportion of DE genes (π)	1%, 3%, 5%
log_2 fold change of DE genes (Δ)	0.5, 1
Pairwise correlations of DE genes (ρ)	0, 0.25, 0.5, 0.75
Gene' variances ($\sigma^2 = \frac{1}{\lambda}$) ~Exp(λ)	$\lambda = 0.25, 0.50, 1, 1.5$
Pairwise correlations of noisy genes $(\gamma) \sim N(0, \theta)$	$\theta = 0, 0.25, 0.5, 0.75$

50% of π were each up- and down-regulated.

Finally, two different values of absolute $\log_2$ fold change (Δ) and three different sample sizes (n) were considered (Table 4.1). For a fixed number of genes ($p = 1000$) and n samples, the samples' labels (0,1) were generated from a Bernoulli distribution with a probability 0.5 and the gene expression data of $p \times n$ dimension was generated from a multivariate normal distribution with mean vectors from a uniform distribution, $U(6,10)$ of length p and the covariance matrices corresponding to the above description, using Cholesky decomposition [Gulob & van Loan, 1996] as a method to determine the root of the covariance matrix. The mean log_2 expression values of DE genes were incremented or decremented with the corresponding log_2 fold change value for samples in class 1. The choice of multivariate normal distribution and mean vector corresponds to the practical assumption that gene expression data are normally distributed in log_2 scale and based on observation that the log_2 expression values often fall in the interval $(0, 16)$. For each combination of the values of the data characteristics, the dataset was simulated as shown in Figure 4.2 (Algorithm 1), yielding 1152 different simulation scenarios, each of which was randomly replicated 1000 times.

4.2.2 Classification functions

Ten elective choices of classification functions were chosen to represent the broad list in the literature that falls within the categories: discriminant analyses or Bayes classifiers, tree-based, regularization and shrinkage, nearest neighbors and neural networks methods. For discriminant analyses, linear discriminant analysis (LDA), quadratic discriminant analysis [McLachlan, 1992] and shrunken centroid discriminant analysis (SCDA) or prediction analysis of microarrays (PAM) [Tibshirani et al., 2002] were selected. Random forest (RF) [Brieman, 2002] was chosen as tree-based method while support vector machines (SVM) [Schölkopf and Smola, 2002], L_1 penalized logistic regression or Lasso (PLR1) [Tibshirani, 1996], L_2 penalized logistic regression or Ridge (PLR2) [Zhu, 2004] as well as L_1 and L_2 penalized logistic regression or Elastic net (PLR12) [Zou et

Algorithm 1

For proportion π *in 1%, 3% and 5% of* $p = 1000$ *as differentially expressed{*

For log_2 *fold change* Δ *in 0.5, 1{*

For absolute pairwise correlation of differentially expressed genes ρ *in 0.00, 0.25, 0.50, 0.75{*

For pairwise correlation of other genes $(\gamma) \sim \boldsymbol{N}(0, \theta)$; θ *in 0.00, 0.25, 0.50, 0.75{*

For genes' variance $(\sigma^2) \sim \boldsymbol{Exp}(\lambda)$; λ = *0.25, 0.50, 1.00, 1.50{*

For sample size n *in 20, 50, 100 {*

Let $p_2 = \pi \times p$ be the number of DE genes of which $1, \ldots, p_1 = \frac{p_2}{2}$ are UR and $p_1 + 1, \ldots . p_2$ are DR

1. **Construct covariance matrices from** σ^2, ρ & γ

$$\boldsymbol{\Sigma_0} = \boldsymbol{\Sigma_1} = \boldsymbol{\Sigma} = \begin{matrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma'_{12} & \Sigma_{22} & \Sigma_{23} \\ \Sigma'_{13} & \Sigma'_{23} & \Sigma_{33} \end{matrix}$$

For iteration B in 1, …, 1000 {

2. **Construct mean** log_2 **expressions for both classes**

$$\boldsymbol{\mu_0} = \mathbf{U}[6,10]; \quad \boldsymbol{\mu_1} = \begin{cases} \mu_0 - \Delta & 1, \ldots, p_1 \\ \mu_0 + \Delta & p_1 + 1, \ldots, p_2 \\ \mu_0 & p_2 + 1, \ldots, p \end{cases}$$

3. **Generate learning set:**

learningLabels = $\mathbf{Bin}(n, 0.5)$; $n_{0L} = sum(\text{learningLabels} == 0)$; $n_{1L} = n - n_{0L}$

$\boldsymbol{learn_0} \sim \boldsymbol{mvN}(n_{0L}, \boldsymbol{\mu_0}, \boldsymbol{\Sigma})$; $\boldsymbol{learn_1} \sim \boldsymbol{mvN}(n_{1L}, \boldsymbol{\mu_1}, \boldsymbol{\Sigma})$; **learningSet=** $rbind(\boldsymbol{learn_0}, \boldsymbol{learn_1})$

4. **Generate test set:**

testLabels = $\mathbf{Bin}(5000, 0.5)$; $n_{0T} = sum(\text{testLabels} == 0)$; $n_{1T} = 5000 - n_{0T}$

$\boldsymbol{test_0} \sim \boldsymbol{mvN}(n_{0T}, \boldsymbol{\mu_0}, \boldsymbol{\Sigma})$; $\boldsymbol{test_1} \sim \boldsymbol{mvN}(n_{1T}, \boldsymbol{\mu_1}, \boldsymbol{\Sigma})$; **testSet=** $rbind(\boldsymbol{test_0}, \boldsymbol{test_1})$

5. ***Build and*** **evaluate classifiers with 10 classification functions as described in the text**

}

}

}

}

}

}

}

Figure 4.2: Algorithm to simulate data, build and validate class prediction models*. For each value of the six variables, the covariance matrix was constructed in step 1, the learning and test data were simulated at steps 2-4 and class prediction models were built and validated in step 5. Steps 2-5 were then repeated 1000* times.

al., 2005] were considered for regularized and shrinkage methods. Finally, k-nearest neighbors (KNN) and feed-forward neural network (NNET) [Ripley, 1996] were the lone choices for nearest neighbors and neural networks respectively.

In machine learning, opinions are that super-learners might provide good class predictions but model complexities of these learners are usually high. As such, super learners might not be useful in clinical practice where physicians often want simple class prediction models, that might yield a subset of genes (and possibly coefficients) for easy interpretation. This is because given a subset of genes, focus can be geared toward these genes rather than the entire genome for which experiments are often costly and time consuming. Thus, our choices of classification functions were driven by the choices often made and considered useful in clinical practice.

4.2.3 Building and evaluating classifiers

To assess the dependency of the chosen classification functions on characteristics of the simulated gene expression data, we built on each simulated dataset, class prediction models with all the classification functions listed above. The simulated dataset was considered as a learning set and for classifiers that require pre-selection of genes because of their limitation to accommodate a number of parameters greater than the number of samples (i.e. LDA, QDA and NNET), the genes were ranked by their moderated t statistics [Smyth, 2004] using the learning set. The learning set was split into a $\frac{1}{3}$ inner- test set and $\frac{2}{3}$ learning set using 5-fold Monte-Carlo-cross-validation (MCCV) with stratification.

The parameter(s) of the classification functions were subsequently tuned using the inner-learning set and evaluated with the inner-test set. These tuning parameters were: number of genes (top k) for LDA and QDA; shrinkage intensity of class centroids for SCDA; with a fixed forest size of 500 trees, the number of variables randomly sampled as candidates at each split and minimum size of terminal nodes for RF; with a linear kernel, the cost of regularization for SVM; L_1 penalty for Lasso; L_2 penalty for Ridge; L_1 & L_2 penalties for Elastic net; number of nearest neighbors for KNN and finally, the number of genes (top k), number of units in a hidden layer and decay weights for NNET. With the optimal parameter(s) for each classification function, the class prediction models were built using the learning set. The resulting models were evaluated on a test set consisting of 5000 samples generated from the same model as the learning set (see Figure 4.2). The error rates of the classification functions on this test set were recorded. The process was repeated 1000 times (sampling both learning and test sets) for each simulation scenario and the resulting error rates over the 1000 replications were used for further analyses.

4.2.4 Random effects linear regression

An average of the error rates of each and every classification function over 1000 replications for each simulated scenario was computed yielding 11520 data points resulting from the 1152 different simulation scenarios by 10 classification functions. The error rates were then transformed to accuracies ($1 - [error\ rate\ +\ 0.001]$) and these accuracies were modeled using a linear random effects regression model with the classification function as the random effects clustering variable, by transforming the accuracies to an unbounded range using the logit function. For the ℓ^{th} standardized study factor, the random effects model is written as:

$$log\left(\frac{\pi(x_{ij})}{1-\pi(x_{ij})}\right) = Y_{ij} = \beta_0 + \vartheta_{0j} + (\beta_1 + \vartheta_{1j})X_{ij}^{\ell} + \varepsilon_{ij} \quad (4.1)$$

where $0 < \pi(x_{ij}) < 1$ is the average accuracy in scenario i for classification function j, $\vartheta_j = (\vartheta_{0j}, \vartheta_{1j})' \sim N(0, D)$ are respectively the random intercepts and slopes of the classification functions while $\varepsilon_{ij} \sim N(0, \sigma^2)$ are the independent and identically distributed residuals, also independent from the random effects ϑ_j. D is a 2×2 covariance matrix of the random effects. All the aforementioned study factors were evaluated by univariate and multivariate linear random effects regression models. Multivariate regression evaluation was done by a backward selection approach. In each step, two nested models, with and without a particular study factor, were compared by log-likelihood ratio test at 5% significance. Each factor ℓ was also evaluated by its explained-variation defined as:

$$\mathrm{Var}_{\ell} = \frac{\mathrm{MSE}_{\mathrm{null}} - \mathrm{MSE}_{\ell}}{\mathrm{MSE}_{\mathrm{null}}} \quad (4.2)$$

where $\mathrm{MSE}_{\mathrm{null}}$ and MSE_{ℓ} are the mean square errors of the null (random intercept only) and the ℓ^{th} standardized study factor models respectively. The explained variation of the selected multivariate model was also evaluated.

4.2.5 Software

All statistical analyses were performed in `R software version 3.2.0`, and `Bioconductor` [Gentleman et al., 2004] using the following packages: `mvtnorm` [Genz and Bretz, 2009] for simulating data, `limma` [Ritchie et al., 2015] for ranking genes via linear models, `CMA` [Slawski et al., 2008] for predictive classification modeling, `lattice` [Sarkar, 2008] for visualization and `lme4` [Bates et al., 2015] for linear random effects modeling. Additionally, we have developed an R package called "`SPreFuGED`: **S**electing a **Pre**dictive **Fu**nction for a given **G**ene **E**xpression **D**ata", that allows researchers to determine an optimal function for any dataset.

4.3 Results

Figure 4.3 shows the average error rates over the 1000 random replicates (y-axis) of the functions (x-axis) for different combinations of variances, pairwise correlations of noisy (non-DE) genes and DE genes for a fixed sample size ($n = 100$), proportion of DE genes ($\pi = 5\%$) and log_2 fold change ($\Delta = 1$). From this figure, one sees that, the error rates for all functions increase with increasing variances (from top- to bottom-row), pairwise correlation values of non-DE genes (from left- to right-column) and pairwise correlation values of DE genes (different colored lines).

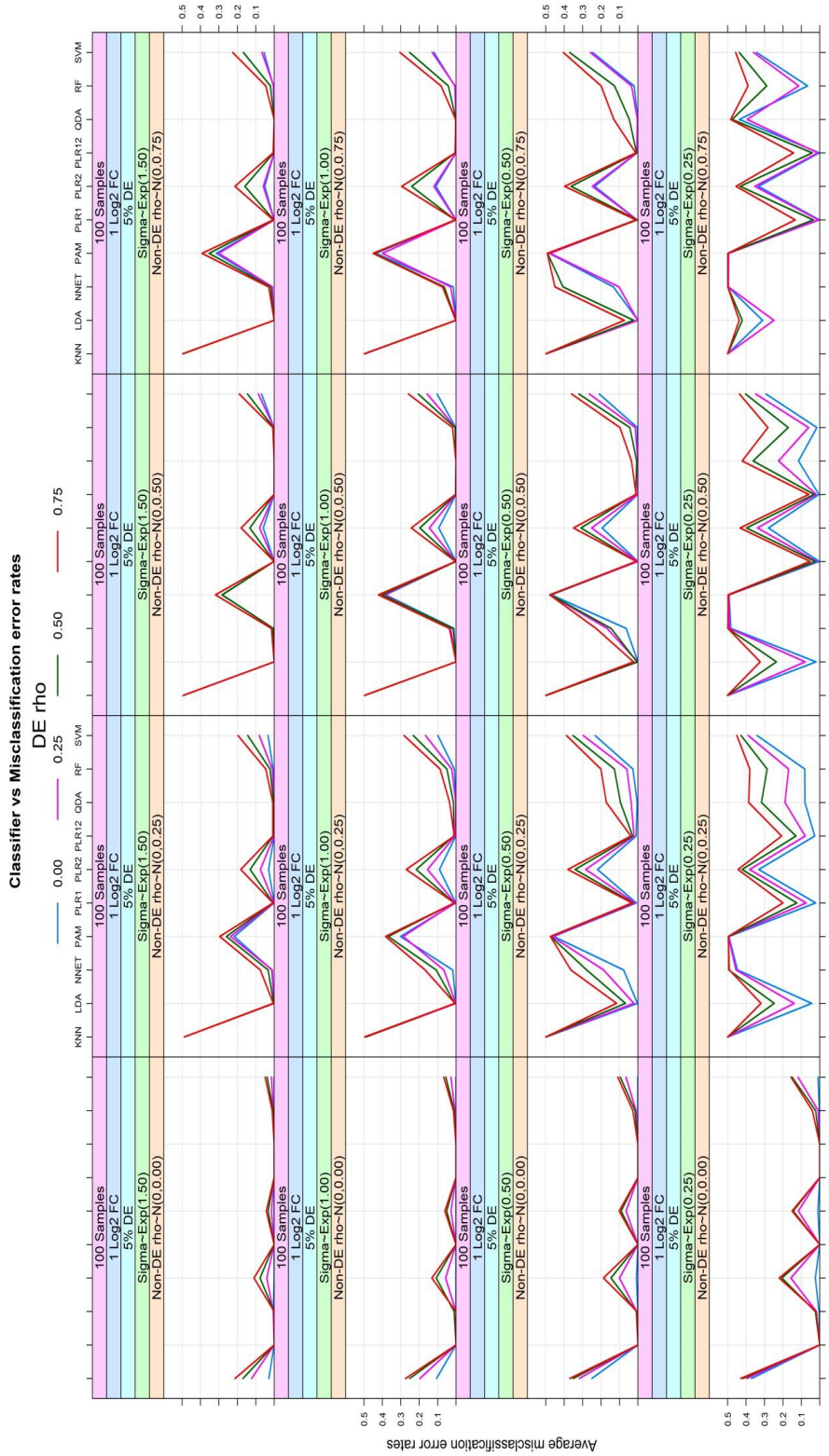

Figure 4.3: Average misclassification error rates of the ten classification functions for sample size of 100, log_2 fold change of 1 and 5% DE genes. *Top-row to bottom row indicate increase in variance ($\frac{1}{4}$) while from left-column to right-column indicate increase in the pairwise correlation of Non-DE genes and the different colored lines from (blue – red) indicate increase in the pairwise correlations of DE genes.*

Table 4.2: Structure of the performance data generated from evaluating the classification functions on the simulated data

ID	Classifier	SampSize	propDE	Variance	deCorr	otherCorr	log2FC	Acc
1	SVM	100	5	0.667	0.00	0.00	1	0.999
2	SVM	100	5	0.667	0.25	0.00	1	0.984
3	SVM	100	5	0.667	0.50	0.00	1	0.961
4	SVM	100	5	0.667	0.75	0.00	1	0.950
5	KNN	100	5	0.667	0.00	0.25	1	0.970
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
11515	LDA	20	1	4	0.50	0.75	0.5	0.498
11516	LDA	20	1	4	0.75	0.75	0.5	0.499
11517	QDA	20	1	4	0.00	0.75	0.5	0.500
11518	QDA	20	1	4	0.25	0.75	0.5	0.499
11519	QDA	20	1	4	0.50	0.75	0.5	0.499
11520	QDA	20	1	4	0.75	0.75	0.5	0.498

On the other hand, other scenarios for different values of sample size, proportion of DE genes and log_2 fold change (Supplementary Figures S1A-C) indicate a negative association of sample size, proportion of DE gene and log_2 fold change to the error rates. The non-constant variability of the error rates between classification functions across scenarios indicates a scenario-specific optimality for each and every classification function.

The average accuracies $(1 - [average\ error\ rates + 0.001])$ of the simulations were summarized to a data matrix as shown on Table 4.2. For each of the predictive variables, a linear random effects regression model was fitted as described in the method section. The individually explained variances of the study factors are depicted on Figure 4.4. This figure shows that sample size, pairwise correlations of non-DE genes and the proportion of DE genes are the leading factors respectively accounting for approximately 17%, 14% and 13% of the null variance. While genes' variances and fold change respectively account for 8% and 7% of the null variance, pairwise correlations between DE genes accounts for simply 1%. As observed graphically, the univariate models (results not shown) confirmed a positive association of sample size, proportion of DE genes and fold change, and a negative association of pairwise correlations of non-DE, DE and the genes' variances to the accuracies.

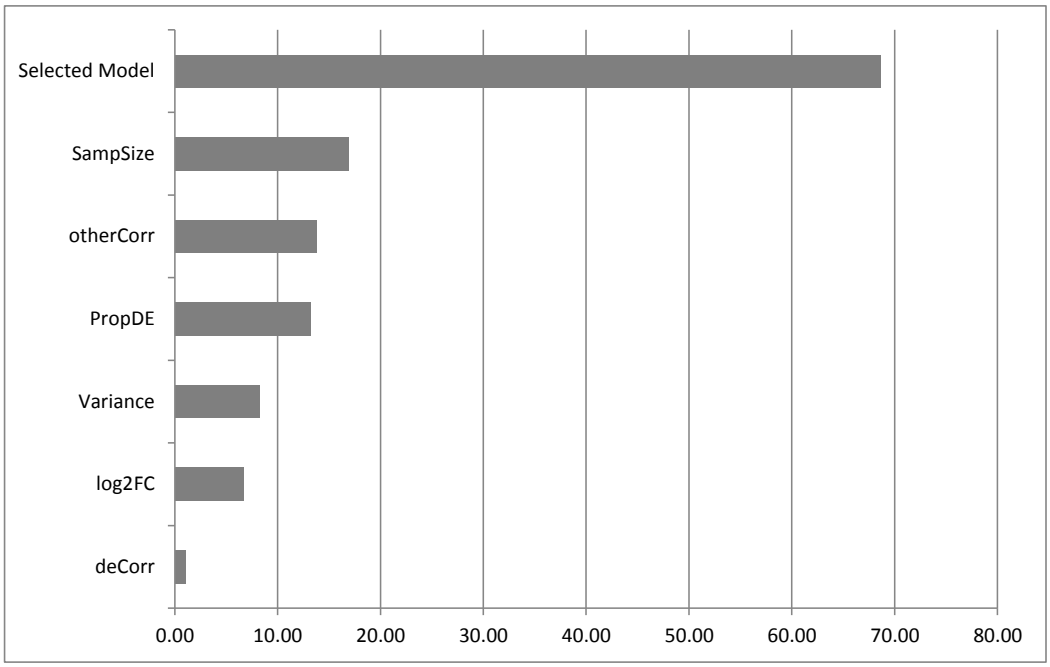

Figure 4.4: Proportion of the null variance explained by each and every studied factor. *The selected model refers to the predictive model presented on Table 4.3.*

For the multivariate linear random effects regression model, we started with a complex model of random intercepts and slopes and three ways interactions of the predictive factors. Starting with pairwise correlation between DE genes because of its low individually explained variance, we eliminated variables using the log-likelihood ratio test. We ended up with the model presented on Table 4.3 consisting of the fixed effects two ways interactions of all the six predictive factors, random intercepts and slopes. This model explains approximately 70% of the null variance as illustrated on Figure 4.4. The left panel of Table 4.3 presents the estimates of fixed effects, the standard errors and the t statistics while the top-right panel presents the net effect of a standard deviation (SD) unit increase of a given factor conditional on common values of other factors. Finally, the bottom-right panel presents the performances of the classification functions at different values of the predictive factors.

From the top-right panel of this table, one notices that a 1 SD unit increase in sample size, corresponding to n = 89.67 will lead to an increase in the Log odds (accuracy), with the highest increase observed when other variables are at their highest values. A similar effect is observed for a 1 SD unit increase in the proportion of DE genes. Though a 1 SD unit increase in fold change leads to an crease in the Log odds, as sample size and proportion of DE genes, its effect is highest

Table 4.3: Fixed effects estimates (left panel) and their conditional on other factors net effects (right panel)

Fixed effects

Parameter	Estimate	Std. Error	t value
Intercept	1.026	0.220	4.663
StdSampSize ($\tilde{n}$)	0.573	0.123	4.660
StdPropDE ($\tilde{\pi}$)	0.508	0.108	4.695
StdVariance ($\tilde{\sigma}^2$)	-0.400	0.081	-4.915
StdDECorr ($\tilde{\rho}$)	-0.158	0.027	-5.878
StdOtherCorr ($\tilde{\theta}$)	-0.560	0.085	-6.557
StdLog2FC ($\tilde{\Delta}$)	0.378	0.068	5.565
StdSampSize*StdLog2FC	0.109	0.009	12.551
StdPropDE*StdLog2FC	0.145	0.009	16.718
StdVariance*StdLog2FC	-0.109	0.009	-12.588
StdDECorr*StdLog2FC	-0.053	0.009	-6.130
StdOtherCorr*StdLog2FC	-0.151	0.009	-17.451
StdSampSize*StdOtherCorr	-0.091	0.009	-10.542
StdPropDE*StdOtherCorr	-0.138	0.009	-15.927
StdVariance*StdOtherCorr	0.102	0.009	11.819
StdDECorr*StdOtherCorr	0.019	0.009	2.182
StdSampSize*StdDECorr	-0.067	0.009	-7.706
StdPropDE*StdDECorr	-0.119	0.009	-13.745
StdVariance*StdDECorr	0.048	0.009	5.519
StdSampSize*StdVariance	-0.161	0.009	-18.571
StdPropDE*StdVariance	-0.172	0.009	-19.815
StdSampSize*StdPropDE	0.249	0.009	28.729

Conditional net effects of study factors

1 SD unit increase

Other variables		$\tilde{n}$	$\tilde{\pi}$	$\tilde{\sigma}^2$	$\tilde{\rho}$	$\tilde{\theta}$	$\tilde{\Delta}$
2 SD	Log odds	0.650	0.715	-0.983	-0.502	-1.078	0.259
1 SD		0.612	0.473	-0.691	-0.330	-0.819	0.318
0 SD		0.573	0.508	-0.400	-0.158	-0.560	0.378
-1 SD		0.534	0.543	-0.108	0.015	-0.300	0.438
-2 SD		0.496	0.578	0.184	0.187	-0.041	0.498

1 SD corresponds to		n=89.67	π =4.63	σ^2=3.22	ρ=0.65	θ=0.65	Δ=1.00

Classification functions

Other variables		KNN	LDA	NNET	PAM	PLR1	PLR12	PLR2	QDA	RF	SVM
2 SD	Log odds	-1.797	1.237	-0.511	-1.540	2.117	2.037	-0.977	0.957	0.265	-1.018
1 SD		-0.445	1.835	0.550	-0.232	2.451	2.396	0.178	1.632	1.085	0.150
0 SD		0.089	1.617	0.794	0.258	1.968	1.938	0.517	1.491	1.088	0.500
-1 SD		-0.193	0.580	0.221	-0.070	0.667	0.662	0.038	0.532	0.273	0.033
-2 SD		-1.294	-1.273	-1.170	-1.214	-1.451	-1.430	-1.258	-1.245	-1.359	-1.251

when the other variables are at their lowest values. While on the average a 1 SD unit increase in the genes' variances, pairwise correlations of non-DE and DE genes will lead to a decrease in the accuracy, these effects become very severe when other variables are at their highest values. For very low values of other variables, a 1 SD unit increase of pairwise correlations between DE genes could even lead to an increase (a positive effect) on the accuracy as was previously observed and illustrated diagrammatically by Novianti et al., (2015). A similar effect is observed for a 1 SD unit increase in the genes' variances at very low values of other variables. These varying effects, indicate the complex interactions between the study factors and hence illustrate why classification functions will perform differently on different datasets.

Lastly, the bottom-right panel of the table shows that all classification functions will perform reasonably well if the predictive factors are at their average values (0 SD) with PLR1, PLR12, LDA and QDA having outstanding performances. For extremely small values (-2 SD) of the studied factors, all functions fail. An indication that the positively associated factors (sample size, proportion of DE genes and fold change) have a high combined net effect than the negatively associated factors (pairwise correlations between non-DE and DE genes and genes' variances). Additionally, for extremely large values (2 SD) of all predictive variables, classification functions like PLR1 and PLR12 clearly demonstrate their abilities to handle correlated variables and higher variances. That notwithstanding, the optimality of a function is scenario specific as illustrated on supplementary Table S1 where both PLR1 and PLR12 fail when other variables are fixed at -2SD and otherCorr or DECorr is varied from -1SD to 2SD. It must be noted however that the combination of all other variables simultaneously being at -2, or at +2, is highly unlikely.

4.4 Application

To evaluate the predictive ability of the here presented random effects regression model on real-life data, eight Affymetrix gene expression datasets of the 25 non-cancerous datasets described in one of our previous study [Novianti et al., 2015] were used. These datasets were selected to include a variety of Array platforms, both class-balance and class-imbalance, number of DE probesets, as well as various sample sizes. Three of these datasets were preprocessed without filtering while the other five were preprocessed and filtered as described by Novianti et al., (2015). We quantified the data characteristics studied and presented on Table 4.4 as follows: (i) sampSize, by counting the samples in the study, (ii) propDE, by ranking the probesets using `limma` [Ritchie et al., 2015] and computing the proportion of DE probesets based on a $\log_2$ fold change cutoff of 1 if the number of DE is ≥10 or 0.5 otherwise, (iii) variance, was determined as the mean of the

Table 4.4: Characteristics of the 8 datasets used for evaluating the predictive model.

No	Study	ID+	Affymetrix Platform	Probesets	SampSize	propDE	Variance	deCorr	otherCorr	log2FC
1	CF*	E-GEOD-10406	HG U133 Plus 2.0	54675	15 (09, 06)	0.267	0.143	1.205	0.414	0.408
2	CS	E-MEXP-2236	HG U133 Plus 2.0	5422	20 (10, 10)	1.881	0.654	1.200	0.368	0.418
3	Dia2	E-CBIL-30	HG U133A	1749	26 (18, 08)	2.859	0.444	0.604	0.611	0.434
4	HIV2	E-GEOD-14278	HG U133 Plus 2.0	11286	18 (09, 09)	1.435	0.523	1.157	0.616	0.393
5	UC2*	E-GEOD-21231	HG 1.0 ST	32321	40 (20, 20)	0.402	0.139	0.631	0.257	0.268
6	UC3	E-GEOD-36807	HG U133 Plus 2.0	6541	28 (15, 13)	6.849	0.735	1.381	0.715	0.503
7	UC5	E-MTAB-331	HG 1.0 ST /HG 1.1 ST	1402	59 (30, 29)	1.427	0.461	1.390	0.951	0.286
8	UC7*	E-GEOD-6731	HG U95AV2	12625	28 (11, 19)	0.135	0.116	1.280	0.474	0.311

* : No filtering was performed + : ArrayExpress accessing ID
The sixth to the eleventh columns correspond to the variables under study. (.,.) represent the sample sizes for each class.

variances of all the probesets, (iv) log2FC, computed as the mean log_2 fold changes of the DE probesets, (v) deCorr as the mean of the elements of the upper- (lower-) triangular of the correlation matrix of the DE probesets and (vi) otherCorr, was computed as the standard deviation (SD) of the elements of the upper- (lower-) triangular of the correlation matrix of non-DE probesets. This matrix was computed from all non-DE probesets if they were less than 20000 or a sample of 20000 from these non-DE probesets otherwise.

These data characteristics were standardized using the mean and SD of the respective variables from the simulated data. And our model was used to predict the accuracies for all classification functions in each dataset (supplementary Figure S2). We then built and evaluated classifiers using the classification functions by splitting the data into $\frac{2}{3}$ learning set and $\frac{1}{3}$ test set with stratification and a 3-fold inner cross-validation on the learning set for parameters optimization. This step was repeated a hundred times, each time predicting the accuracies of classification functions on the learning set using the random effects model and also recording the expected (observed) accuracies on the test set. These predicted and observed accuracies over the 100 repetitions are respectively presented on Supplementary Figure S3A & B. To compare the predicted to observed accuracies, and considering that we are interested in the ordering of performance (i.e. determining an optimal function for a given data), we used the ranked base Spearman correlation between the average predicted accuracies and the average observed accuracies.

The results of this comparison for each dataset are presented on Figure 4.5. The positive correlation values on this figure indicate agreement between our predicted and observed accuracies. Though these correlations are not very high in some datasets, our model more or less

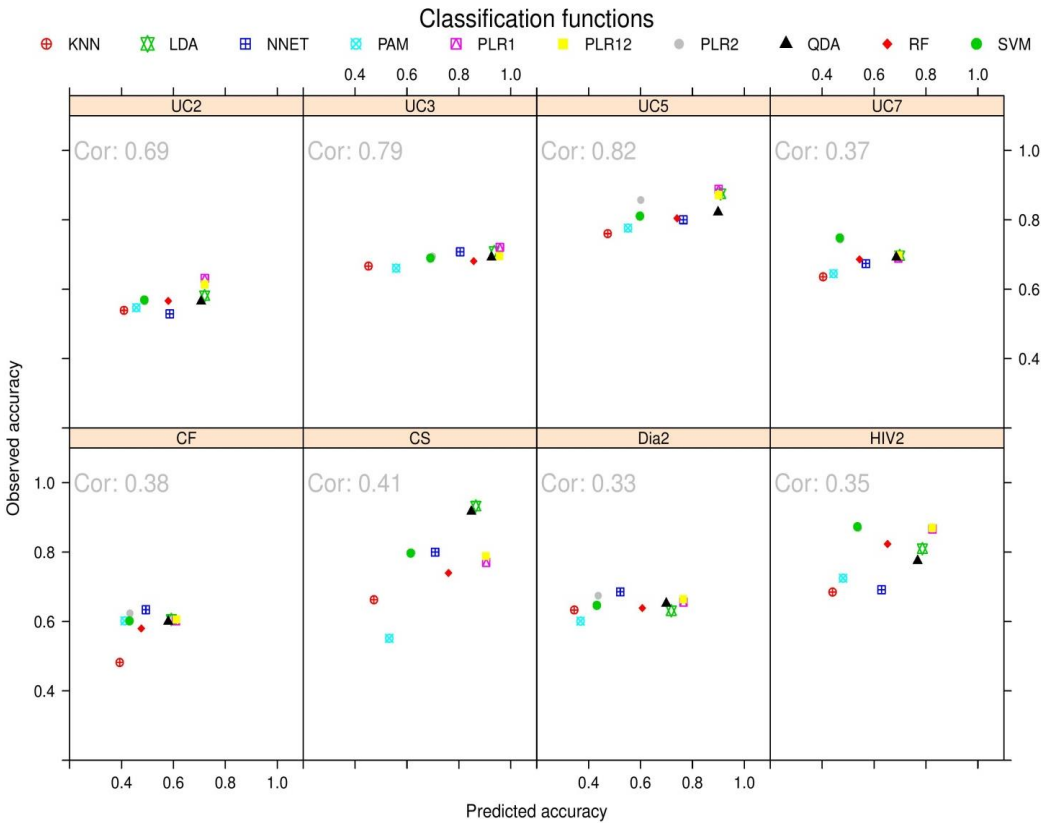

Figure 4.5: Predicted vs Expected (Observed) accuracies. *Cor represents Spearman correlations between the predicted and observed accuracies.*

determined an optimal classification function for all the datasets except for UC7 where Ridge regression and SVM emerged first instead of fourth as predicted (i.e. 87.5% sensitivity). Nevertheless, the model was able to rule out on which classification(s) will perform poor on a given dataset, with approximately 100% certainty. As expected, the performance of the functions deteriorate on CF (small sample size and low proportion of DE probesets), Dia2 (high class-imbalance and small fold changes), UC2 (low proportion of DE probesets and small fold changes), UC3 (large variances and high correlations) and UC7 (low proportion of DE probesets). From Figure 4.5 and supplementary Figure S3A & B, one sees that except on the UC3 data, our model's accuracies are less than or equal to observed accuracies. The model performs well on dataset with large sample sizes and balanced classes (UC2, UC3 and UC5). It attained its lowest performance on Dia2 where there is high class-imbalance and hence few samples of the small class in the learning set and on HIV2 and CF datasets with small sample sizes.

4.5 Discussion

We hypothesized that the performance of classification functions on gene expression data depends on sample size, proportion of DE genes, genes' variances, $\log_2$ fold changes between DE genes and magnitude of the pairwise correlation within DE genes and non-DE genes, and showed their association to the accuracies of ten often used and clinically relevant classification functions using simulations. Additionally, we built a predictive model to determine an optimal classification function among the studied functions using the simulation results. An application of the predictive model on eight non-cancerous real-life gene expression datasets predicted optimal function(s) for seven out of the eight and was able to rule out function(s) that will perform poor on almost all the datasets. This model may serve as a guide for choosing a classification function for a given gene expression data.

Classification functions have been shown to perform differently across gene expression datasets [Lee et al., 2005] and data characteristics have been shown to differ across datasets and are associated to the performance of classification functions [Jong et al., 2014; Novianti et al., 2015]. While sufficient knowledge is available on the properties of most classification functions and procedures to build class prediction models using gene expression data have been outlined by [Wessels et al., 2005], little is known about data characteristics that accounts for the variability in the performance of classification functions and how to use these characteristics to choose an optimal classification function for a specific dataset. As such, most researchers adhere to specific classification function(s) or randomly choose a classification for their class prediction models irrespective of the disease or data under study. A common practice is to evaluate several classification functions and select the one with smallest misclassification error but this leads to selection bias [Varma & Simon, 2006; Tibshirani & Tibshirani, 2009; Bernau et al., 2013; Ding et al., 2014].

In this study, we outlined data characteristics together with clinically relevant and often used classification functions and investigated their effects on classification performance using simulation studies. Based on these simulation studies, we provided a guide for choosing an optimal classification function for a specific dataset using the data's characteristics and the studied classification functions through a linear random effects predictive model. As a meta-model one would expect it to explain close to 100% of the variance in the simulated data but our predictive model accounts for approximately 70% of the variability in the simulated data. The remaining 30% unexplained variance may be associated to sampling variability stemming from the several (192)

random covariance matrices used to generate both learning and test sets as well as the different learning and test sets generated at each iteration.

Although we used different classification functions and evaluated these functions using accuracy, our simulation results confirm the findings of Kim & Simon, (2011) that classifiers tend to have poor performance on highly correlated data. Our results also agree with those of Novianti et al., (2015) that correlations, the absolute $\log_2$ fold changes and the number of DE probesets are associated to the accuracy of a class prediction model. In addition, these results specify clearly the directions of the association and point out the effects of other data characteristics like sample size, genes' variances that were not previously identified.

Most importantly, we have provided a predictive model that can serve as a guide to choose a classification function for a given dataset and its application on eight real-life datasets (both filtered and unfiltered) indicated a good predictive ability of the model. Although our model was reasonably good in its prediction on real-life data, we want to point out that it might have failed in some datasets because of the following reasons: (i) most of the eight non-cancerous datasets had small sample sizes and splitting these datasets to learning and test sets yielded even smaller sample sizes of the learning sets and hence might have led to poor estimates of the characteristics under study and (ii) the observed accuracies might not be the true accuracies because of the few Bootstrap samples. It could have been better if we had the means to perform several Bootstraps but due to the small sample sizes, the number of independent Bootstrap samples is limited. The fact that our predictions were most often slightly lower than the observed accuracies for almost all classification functions might indicate the general trend that the performance of a model usually decreases on an independent dataset. Hence, our model's predictions might reflect expected accuracies on independent datasets.

In the simulated data, we assumed exponential and normal distributions for the variances and pair-wise correlation of non-DE genes respectively. These distributional assumptions might be violated in some datasets. As such, it will be worth trying different distributions. Also, we used accuracy as a measure of evaluation by minimizing the loss function but in clinical applications, probabilities are more informative than simple yes or no predictions because they quantify the uncertainty of a prediction [Pepe, 2005]. As such, it is worth evaluating these data characteristics on probabilistic classification functions where by the log-likelihood function is optimized, this might possibly provide a predictive model that will be most useful in clinical applications. Despite these limitations, our model was found to work well with data containing reasonably large and

balanced sample sizes (n≥30). As such, our results apply to balanced class data. For data with class-imbalance some classification functions will have deteriorating performance, for which several solutions are proposed. However, this topic is outside the focus of the current study. In summary, our results serve as a guide to use data characteristics to choose an optimal classification function for a given dataset.

Acknowledgements

This work has been supported by the VIRGO consortium, which is funded by the Netherlands Genomics Initiative and by the Dutch Government (FES0908). The funding agencies in no way influenced the outcome or conclusions of the study. The authors would like to thank the VIRGO consortium and its sponsors, the Netherlands Genomics Initiative and the Dutch Government for the financial support. Next to this, we thank Martin Marinus and the HPC-team at UMC Utrecht for the high performing computing facilities.

Supplementary Material

The online version of this article (DIO: 10.1093/bioinformatics/btw03) offers supplementary material, available to authorized users. And include:

Figure S1: Average misclassification error rates of the ten classification functions for: (A) sample size of 100, log_2 fold change of 1 and 1% DE genes; (B) sample size of 100, log_2 fold change of 0.5 and 5% DE genes; (C) sample size of 50, log_2 fold change of 1 and 5% DE genes . *Top-row to bottom row of each figure indicate increase in variance (1/λ) while from left-column to right-column indicate increase in the pairwise correlation of Non-DE genes and the different colored lines from (blue –red) indicate increase in the pairwise correlations of DE genes.*

Figure S2: Predicted accuracies of the 8 real-life dataset by our linear random effects regression model using the characteristics shown on Table 4.

Figure S3: 100 cross-validated (CV) accuracies for each classification function; (A) predicted using the linear random effects regression model on CV learning set & (B) using out-of-bag (OOB) CV test set on each classifier built on the CV learning set.

Table S1: Predicted log odds of accuracy for different combinations of SD unit values for the studied variables: *accompanying excel file*

R source code and R package SPreFuGED

Chapter 4

Chapter 5

Selecting an optimal probabilistic classifier with gene expression data: does it differ from an optimal direct classifier?

Victor L. Jong, Kit C. B. Roes & Marinus J. C. Eijkemans

Submitted

Abstract

Motivation: Class prediction with gene expression is widely becoming a regular task in generating diagnostic and prognostic models through machine learning algorithms. The literature reveals that probabilistic classifiers are of high relevance in clinical applications and that classification functions perform differently across datasets. The question, which probabilistic classification function should be used for a given dataset remains unascertained. In this study, we focused on probabilistic classifiers, where we compared their performance to direct classifiers and devised a predictive model for determining an optimal probabilistic function for class prediction with a given dataset.

Results: Gene expression data were simulated for different values of gene-pairs correlations, sample size, genes' variances, number of deferentially expressed genes and fold changes. For each simulated dataset, nine probabilistic classifiers were built and evaluated using nine probabilistic classification functions. The resulting Brier scores of 10368 data points from 1152 different simulation scenarios by 9 classification functions were then modeled using a linear random effects regression on the studied data characteristics, yielding a model that predicts the Brier score of the functions on a given data. An application of our model on twelve real-life datasets showed high positive correlations (0.26–0.97) between the predicted and expected Brier scores.

Conclusion: We have shown that the optimality of a function on a given dataset depends on whether it is trained as a direct or probabilistic classifier. We present a predictive model that might serve as a guide for determining an optimal probabilistic function among the nine studied functions, for any given gene expression data.

5.1 Introduction

Gene expression profiling, particularly microarray gene expression profiling has become a widely used tool to identify disease subpopulations and to perform diagnostic and prognostic predictions in clinical applications [van 't Veer et al., 2002; Huang et al., 2010]. Genomic data usually have a huge number of parameters relative to the number of samples ($p \gg n$) and several classification functions have been proposed in machine learning to tackle this curse of dimensionality. Classification can be considered as direct or probabilistic. Whereas direct classifiers assign a new sample to a particular class, probabilistic classifiers assign probabilities to a sample's class membership. In machine learning, most class prediction analyses with high-dimensional data focus on direct classification. However, for medical decision making, it is more valuable to have a probabilistic classifier because medical decision making is complex and misclassification costs are often high [Kim & Simon, 2011]. Hence, probabilistic classifiers that provide an estimate of the probability of class membership for new cases are considered to be more useful than classification rules that simply assign cases to a class [Pepe, 2005; Malley et al., 2012]. These probabilities in conjunction with other patients' information might be used in making complex integrated clinical decisions.

Although class prediction analysis is a common practice, both direct and probabilistic classification functions have been shown to perform differently across datasets [Lee et al., 2005; Shi et al., 2010; Kim & Simon, 2011; Kruppa et al., 2014a; Kruppa et al., 2014b]. Comparing several classification functions and selecting the best (based on minimum error) is computationally intensive and when feasible, leads to selection bias [Varma & Simon, 2006; Tibshirani & Tibshirani, 2009; Bernau et al., 2013; Ding et al., 2014] because a function might have the smallest error simply by chance. This "capitalization on chance" probability increases with both the number of algorithms compared and a decrease in sample size. Thus, selection bias is often high in data with small sample sizes and when several algorithms are compared [Ding et al. 2014]. Ding et al., (2014) stated as an example that one uses a small pilot data, compares several machine learning methods and selects the minimum error classifier (MEC) with a falsely small error because of the selection bias. When the model proceeds to a large cohort validation for translational research, it will likely fail. Hence, several bias correction methods have been proposed for classification with gene expression data. Nevertheless, no such method is 100% effective when several least optimal algorithms are compared on a dataset with a small sample size, as observed by Ding et al., (2014) for classification. Thus, to choose an optimal classification function for a given gene expression

data remains a challenge. There is a philosophy of voting and crowd machines (super-learners) which emphasizes that combining the scores of all machines can improve the overall performance. Nevertheless, super-learners are often considered less relevant in medical applications because for medical studies, there is resistance of pathologists and medical practitioners to use uninterpretable black box multivariate tests for making important treatment decisions [Simon, 2014]. Hence, traditional classifiers that can yield a gene signature are often preferred than super-learners, since such a signature composing of a subset of genes can be easily profiled within a reasonable time frame and is also cost effective.

Although a substantial amount of information is known about the various classification functions and how most machine learning algorithms (functions) can be tailored to produce probabilities [Kruppa et al., 2014a & 2014b], little is known about data characteristics that affect the performance of these functions. Recently, Kim & Simon, (2011) showed using simulations that correlation, which was further determined to be significantly different across real-life gene expression data by Jong et al., (2014), is one of the data characteristics that affects the performance of both direct and probabilistic classification functions. In addition to correlation, sample size, number of differentially expressed genes, genes' fold changes and/or variances were recently determined using real-life data by Novianti et al., (2015) and simulations by Jong et al., (2016) to be associated to the performance of direct classifiers. While Jong et al., (2016) presented a mixed effects regression model that could be used to determine an optimal classification function for a given dataset, they focused exclusively on direct classifiers. Additionally, the choices for building and evaluating models differ between direct and probabilistic classifiers. For instance, optimizing direct classifiers might focus on minimizing the error rate or maximizing the accuracy while optimizing probabilistic classifiers involves maximizing a log-likelihood function as shown by Kim & Simon, (2011). As was observed by these authors, an optimal direct classifier might not necessarily be an optimal probabilistic classifier on a given dataset. Thus, different optimization functions might lead to different associative effects of data characteristics to the performance of classifiers.

In this study, we focus on probabilistic classifiers because of their relevance in biomedical applications. We investigate the effects of correlations, sample size, proportion of differentially expressed genes, genes' fold changes and variances on the Brier scores of probabilistic classifiers using extensive simulations and compared the results to those of direct classifiers [Jong et al., 2016]. We aim to provide a model for determining an optimal probabilistic classifier for a given

gene expression dataset. The rest of this article consists of Section 5.2 which contains the methodology to simulate data, the probabilistic classification functions considered and the building and evaluation of the probabilistic classifiers; Section 5.3 contains a summary of the results of the class prediction models for different simulated scenarios; Section 5.4 provides an application of the predictive model from our simulated results to real-life microarray gene expression datasets and Section 5.5 contains some discussions.

5.2 Methodology

5.2.1 Simulated data (scenarios)

To simulate gene expression data, we assumed that sample size, proportion of differentially expressed (DE) genes, genes' variances, fold changes, pairwise correlations between DE and noisy genes might be associated to the performance of probabilistic classifiers as was previously observed by Jong et al., (2016) for direct classifiers. Different values of the six variables were systematically varied and both training and test data were simulated as extensively described by Jong et al., (2016) and presented in supplementary material of this article.

5.2.2 Probabilistic classification functions

Nine of the ten elective choices of classification functions used by Jong et al., (2016) can produce probabilities either by design or self-post-processing. They were chosen from discriminant analyses, tree-based, regularization and shrinkage and nearest neighbors methods. For discriminant analyses, linear discriminant analysis (LDA), quadratic discriminant analysis [McLachlan, 1992] and shrunken centroid discriminant analysis (SCDA) also known as prediction analysis of microarrays (PAM) [Tibshirani et al., 2002] were selected. Random forest (RF) [Breiman, 2001] was chosen as tree-based method, while support vector machines (SVM) [Schölkopf & Smola, 2002], L_1 penalized logistic regression (Lasso or PLR1)) [Tibshirani, 1996], L_2 penalized logistic regression (Ridge or PLR2) [Zhu, 2004] and $L_1 \& L_2$ penalized logistic regression (Elastic net or PLR12) [Zou and Trevo, 2005] were considered for regularized and shrinkage methods. Finally, a variant of k-nearest neighbors called probabilistic nearest neighbors (KNN) [Slawski et al., 2008] was the lone choice for nearest neighbors.

5.2.3 Building and evaluating probabilistic classifiers

To assess the dependency of the chosen functions on characteristics of the simulated gene expression data, we built probabilistic classifiers on each simulated dataset with all the functions. The simulated dataset was considered as learning set and for functions that require pre-selection

of genes because of their limitation to accommodate a number of parameters greater than the number of samples (i.e. LDA and QDA), the genes were ranked by their moderated t statistics [Smyth, 2004] using the learning set. The learning set was split into an inner test set and a cross-validation set using 5-fold cross-validation. Supposed the binary response variable (two classes) are recoded to class 0 and 1, the parameter(s) of the functions were subsequently optimized using the cross-validation set and evaluated with the out-of-bag (OOB) inner test set by maximizing the binomial log-likelihood function:

$$l(\eta) = \sum_{i=1}^{n} y_i \, log\big(\hat{p}_{(-i),\eta(x_i)}\big) + (1 - y_i) \, log\big(1 - \hat{p}_{(-i),\eta(x_i)}\big) \tag{5.1}$$

where y_i is the observed class (0 or 1) of the i^{th} observation and $\hat{p}_{(-i),\eta(x_i)}$ is the predicted probability of the i^{th} observation to be in class 1 computed in a cross-validation set without the i^{th} case and using a vector of tuning parameters η. While n is the total number of OOB samples over the 5-fold cross-validation. This vector of tuning parameters η depends on the type of classification function and may consist of: number of genes (top k) for LDA and QDA; the shrinkage intensity of class centroids for SCDA; the number of variables randomly sampled as candidates at each split and minimum size of terminal nodes for RF; the cost of regularization for a linear kernel SVM; L_1 penalty for Lasso; L_2 penalty for Ridge; L_1 & L_2 penalties for Elastic net and the number of nearest neighbors for KNN. It is worth to mention that the same parameters space that was used by Jong et al., (2016) was utilized. With the optimal parameter(s) for each classification function, the class prediction models were built using the entire learning set.

The resulting models were evaluated on a test set consisting of 5000 samples generated from the same model as the learning set as shown on supplementary Figure S1. The functions were evaluated using the Brier score defined as:

$$BS = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{p}_i)^2 \tag{5.2}$$

where y_i is the observed (0 or 1) class of sample i and $\hat{p}_i$ is the predicted probability of class 1 for sample i and $N = 5000$ is the total number of test samples. The process was repeated 1000 times for each simulation scenario by replicating both the training and test sets and the resulting averaged Brier scores over the 1000 replications were used for downstream analyses.

Since some classification functions like SCDA, penalized logistic regressions (Ridge, Lasso and Elastic net) etc., are by nature designed to produce probabilities, these probabilities are often

used for evaluation even when these functions' parameters were optimized by maximizing the accuracy. To show that a different optimization criterion may lead to different performance, we randomly chose a few data scenarios and for the nine probabilistic functions, we compared the averaged Brier scores obtained by maximizing accuracy, to the averaged Brier scores obtained by maximizing log-likelihood function.

5.2.4 Random effects linear regression

An average of the Brier scores of each and every classification function over the 1000 replications for each simulated scenario was computed yielding 10368 (1152 scenarios by 9 classification functions) data points. The Brier scores which range from 0 to 0.5 (with 0.5 corresponding to the classifier that randomly assigns samples to categories and 0 to a perfect classifier), were then transformed to an accuracy-like value that lies in the interval [0, 1] with 0 representing the classifier with random assignment and 1 to a perfect classifier and referred to as transformed Brier scores (trBS). Let $\pi(x_{ij})$ be the average BS of classification function j in scenario i, its trBS is as follows:

$$\tilde{\pi}(x_{ij}) = 1 - \frac{\pi(x_{ij})}{0.5} \tag{5.3}$$

The trBS were modeled using a linear random effects regression analysis with the classification functions as the clustering variable, by transforming the trBS to an unbounded range using the logit function. For the ℓ^{th} standardized study factor, the random effects model is written as:

$$\varphi = log\left(\frac{\tilde{\pi}(x_{ij})}{1-\tilde{\pi}(x_{ij})}\right) = Y_{ij} = \beta_0 + \vartheta_{0j} + (\beta_1 + \vartheta_{1j})X_{ij}^{\ell} + \varepsilon_{ij} \tag{5.4}$$

where $\beta = (\beta_0, \beta_1)'$ are the fixed effects, $0 < \tilde{\pi}(x_{ij}) < 1$ is the expected trBS of classification function j in scenario i, $\vartheta_j = (\vartheta_{0j}, \vartheta_{1j})' \sim N(0, D)$ are respectively the random intercepts and slopes of the classification functions while $\varepsilon_{ij} \sim N(0, \sigma^2)$ are the independent and identically distributed residuals, also independent from the random effects ϑ_j. D is a 2×2 covariance matrix of the random effects [Jong et al., 2016]. All the aforementioned study factors were evaluated by univariate and multivariate linear random effects regression models. Multivariate regression evaluation was done by a backward selection approach using the log-likelihood ratio test. In each step, two nested models with and without a particular study factor were tested. Each factor ℓ was also evaluated by its explained null variance defined as:

$$\mathrm{Var}_{\ell} = \frac{\mathrm{MSE}_{\mathrm{null}} - \mathrm{MSE}_{\ell}}{\mathrm{MSE}_{\mathrm{null}}} \tag{5.5}$$

where $\mathrm{MSE}_{\mathrm{null}}$ and MSE_{ℓ} are the mean square errors of the null (random intercept only) and the ℓ^{th} standardized study factor models respectively. The explained null variance of the selected multivariate model was also evaluated.

5.2.5 Software

All statistical analyses were performed in `R software version 3.2.2` and `Bioconductor` [Gentleman et al., 2004] using the following packages: `mvtnorm` [Genz and Bretz, 2009] for simulating data, `limma` [Ritchie et al., 2015] for ranking genes via linear models, `CMA` [Slawski et al., 2008] for classification modeling of some functions, `lattice` [Sarkar, 2008] for visualization and `lme4` [Bates et al., 2015] for linear random effects modeling. Additionally, we have incorporated the functions to replicate and apply these results in the R package `SPreFuGED` [Jong et al., 2016].

5.3 Results

Figure 5.1 shows the average (over the 1000 random replicates) Brier scores (y-axis) by absolute pairwise correlations within non-DE genes (x-axis) of the classification functions (colored lines) for different combinations of genes' variances and pairwise correlations of DE genes for a fixed sample size ($n = 100$), proportion of DE genes ($\pi = 5\%$) and log_2 fold change ($\Delta = 1$). From this figure, one clearly sees that in general, the Brier scores for all classification functions increase with increasing pairwise correlations of non-DE genes (left to right of x-axis), variances (from top- to bottom-row), pairwise correlation values of DE genes (from left- to right-column).

On the other hand, other scenarios for different values of sample size, proportion of DE genes and log_2 fold change (Supplementary Figures S2A-C) indicate a negative association of sample size, proportion of DE gene and log_2 fold change to the Brier scores. This shows that in general, probabilistic classifiers performed better with increasing proportion of differentially expressed genes, log_2 fold changes and sample sizes but worse with increasing pairwise correlations between genes and genes' variances as was previously observed for direct classifiers by Jong et al., (2016). Nevertheless, the differences in the performance pattern of classification functions when trained as direct classifiers (Supplementary Figure S3) or probabilistic classifiers (Figure 5.1) indicate that the optimality of a classification function in a given scenario, depends on the way it is trained, as was previously observed by Kim & Simon, (2011).

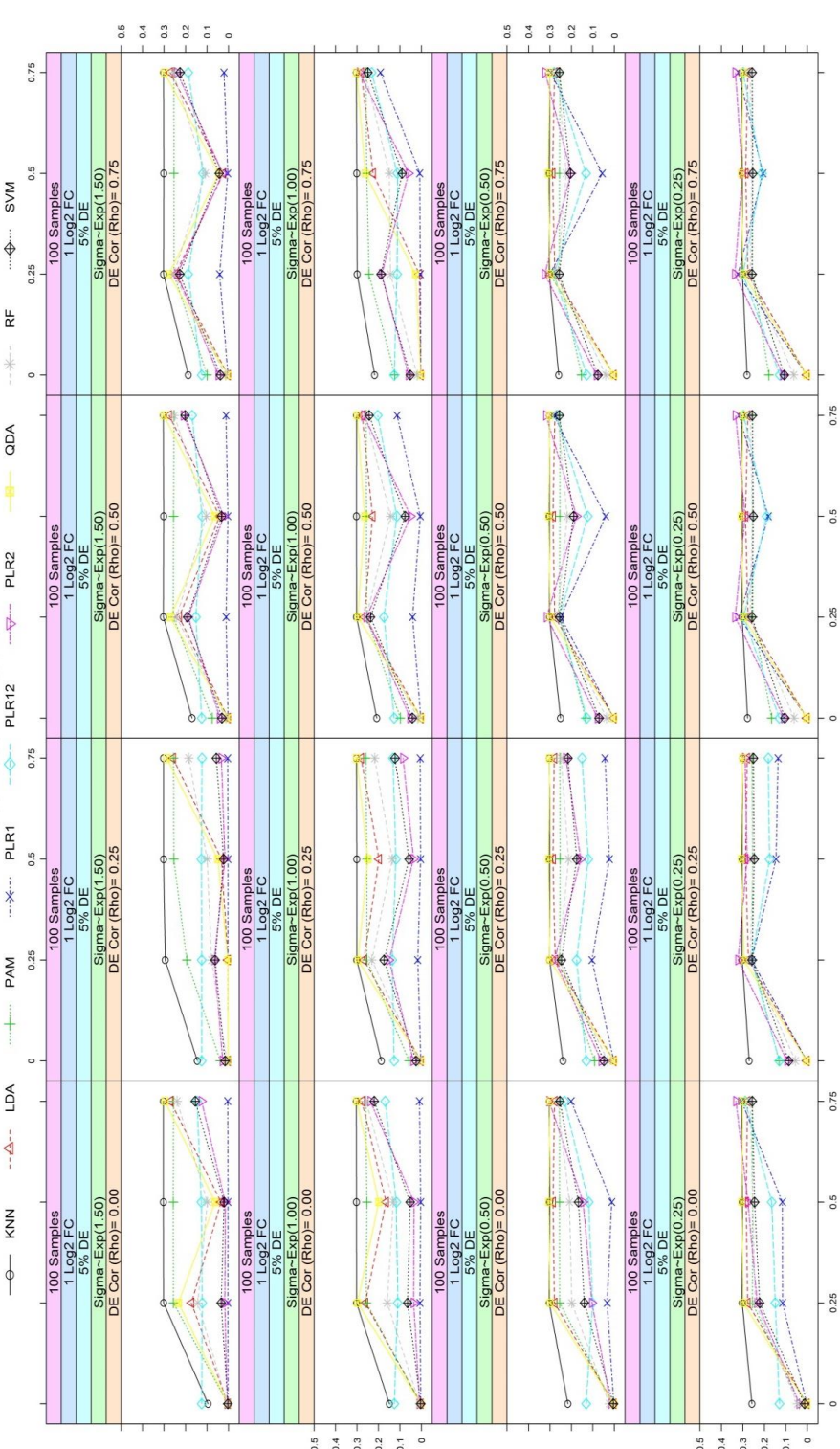

Figure 5.1 Average Brier Scores of the nine classification functions for sample size of 100, log_2fold change of 1 and 5% DE genes: *the x-axis indicate the value of absolute pairwise correlations with non-DE genes, top-row to bottom-row indicate increase in variance* $(\frac{1}{\chi})$*, left-column to right-column indicate increase in the pairwise correlation of DE genes and the different colored lines represent the classification functions.*

Since some classification functions are by design capable of producing probabilities, some researchers often optimize these functions on accuracy but use the predicted probabilities of test set for evaluation. We computed the Brier scores of the functions optimized on accuracy for a fixed sample size ($n = 50$), proportion of DE genes ($\pi = 5\%$) and $\log_2$ fold change ($\Delta = 1$) by training the nine classification functions as probabilistic classifiers optimizing the binomial log-likelihood function (Supplementary Figure S4A) and based on accuracy (supplementary Figure S4B). These figures show different patterns of the averaged Brier scores. Specifically, one will clearly notice that in most scenarios, elastic net (PLR12) has very similar performance to Lasso (PLR1) and might be considered among the optimal functions when optimized base on accuracy, but will clearly be distinguished from Lasso when optimized by maximizing the log-likelihood function. This further indicates that the performance of a probabilistic function on a given dataset also depends on whether it is optimized by maximizing accuracy or log-likelihood function.

The average Brier scores (BS) of the simulations were summarized to a data matrix as shown on Table 5.1. For each of the predictive variables, a linear random effects regression model was fitted to the transformed Brier scores (trBS) as described in the method section. The individually explained null variance of the study factors are depicted on Figure 5.2. This figure shows that pairwise correlations of non-DE genes and sample size are the leading factors respectively accounting for approximately 18% and 17% of the null variance. While the proportion of

Table 5.1: Structure of the performance data generated from evaluating the classification functions on the simulated data

ID	Classifier	SampSize	propDE	Variance	log2FC	otherCorr	deCorr	BS
1	KNN	100	5	0.667	1	0.00	0.00	0.0952
2	KNN	100	5	0.667	1	0.00	0.25	0.1447
3	KNN	100	5	0.667	1	0.00	0.50	0.1693
4	KNN	100	5	0.667	1	0.00	0.75	0.1869
5	LDA	100	5	0.667	1	0.00	0.00	0.0025
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
10363	RF	20	1	4	0.5	0.75	0.50	0.2531
10364	RF	20	1	4	0.5	0.75	0.75	0.2528
10365	SVM	20	1	4	0.5	0.75	0.00	0.2763
10366	SVM	20	1	4	0.5	0.75	0.25	0.2764
10367	SVM	20	1	4	0.5	0.75	0.50	0.2714
10368	SVM	20	1	4	0.5	0.75	0.75	0.2744

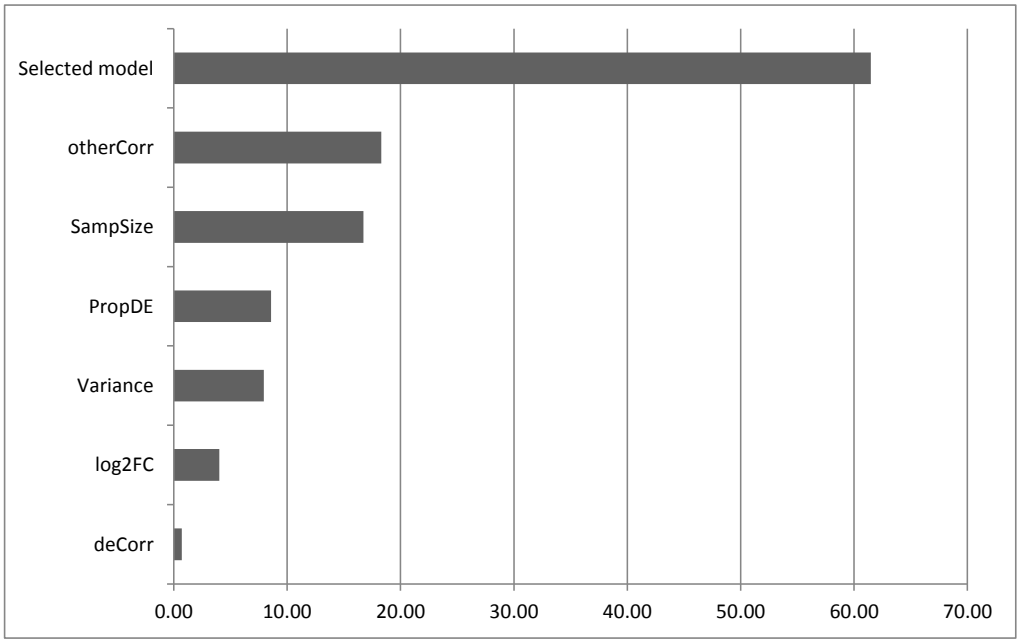

Figure 5.2: Proportion of the null variance explained by each and every studied factor. *The selected model refers to the predictive model presented on Table 5.2.*

differentially expressed genes, genes' variances and $\log_2$ fold change respectively account for approximately 9%, 8% and 5% of the null variance, pairwise correlations between DE genes accounts for simply 1%. As observed graphically, the univariate models (results not shown) confirmed a positive association of sample size, proportion of DE genes and fold change, and a negative association of pairwise correlations of non-DE, DE and the genes' variances to the transformed Brier scores. These results are in line with those observed by Jong et al., (2016) except for the fact that pairwise correlations of non-DE genes accounts for the highest explained null variance for probabilistic classifiers as opposed to sample size that was observed by these authors to be the factor accounting for the most explained null variance for direct classifiers. This could be that highly correlated variables lead to random; uncertain predicted probabilities (closed to 0.5) or extreme predicted probabilities (closed to 0 or 1) for probabilistic classifiers. Both of which are related to calibration and refinement scores as parts of the Brier score [Kim & Simon, 2011].

For the multivariate linear random effects regression model, we started with a complex model of random intercepts and slopes and three ways interactions of the predictive factors. Starting with pairwise correlation between DE genes because of its low individually explained null variance, we

Table 5.2: Fixed effects estimates (left panel) and their conditional on other factors net effects (right panel)

Fixed effects

Parameter	Estimate	Std. Error	t value
Intercept	0.328	0.162	2.028
StdSampSize ($\tilde{n}$)	0.402	0.126	3.184
StdPropDE ($\tilde{\pi}$)	0.308	0.082	3.748
StdVariance ($\tilde{\sigma}^2$)	-0.289	0.082	-3.507
StdDECorr ($\tilde{\rho}$)	-0.094	0.025	-3.829
StdOtherCorr ($\tilde{\theta}$)	-0.448	0.121	-3.713
StdLog2FC ($\tilde{\Delta}$)	0.220	0.052	4.190
StdSampSize*StdLog2FC	0.007	0.008	0.858
StdPropDE*StdLog2FC	0.034	0.008	4.263
StdVariance*StdLog2FC	-0.061	0.008	-7.582
StdDECorr*StdLog2FC	-0.027	0.008	-3.318
StdOtherCorr*StdLog2FC	-0.128	0.008	-15.976
StdSampSize*StdOtherCorr	-0.077	0.008	-9.616
StdPropDE*StdOtherCorr	-0.123	0.008	-15.306
StdVariance*StdOtherCorr	0.071	0.008	8.873
StdDECorr*StdOtherCorr	0.011	0.008	1.410
StdSampSize*StdDECorr	-0.032	0.008	-4.022
StdPropDE*StdDECorr	-0.071	0.008	-8.807
StdVariance*StdDECorr	0.032	0.008	4.014
StdSampSize*StdVariance	-0.108	0.008	-13.456
StdPropDE*StdVariance	-0.110	0.008	-13.624
StdSampSize*StdPropDE	0.101	0.008	12.534

Conditional net effects of study factors

1 SD unit increase

Other variables	$\Phi = \log\left(\frac{\tilde{\pi}(x_{ij})}{1-\tilde{\pi}(x_{ij})}\right)$	$\tilde{n}$	$\tilde{\pi}$	$\tilde{\sigma}^2$	$\tilde{\rho}$	$\tilde{\theta}$	$\tilde{\Delta}$
2 SD		0.182	-0.029	-0.639	-0.267	-0.940	-0.130
1 SD		0.292	0.139	-0.464	-0.181	-0.694	0.045
0 SD		0.402	0.308	-0.289	-0.094	-0.448	0.220
-1 SD		0.513	0.476	-0.114	-0.008	-0.202	0.395
-2 SD		0.623	0.645	0.061	0.078	0.044	0.569

1 SD corresponds to	n=89.67	π =4.63	σ^2=3.22	ρ=0.65	θ=0.65	Δ=1.00

Classification functions

Other variables	$\Phi = \log\left(\frac{\tilde{\pi}(x_{ij})}{1-\tilde{\pi}(x_{ij})}\right)$	KNN	LDA	PAM	PLR1	PLR12	PLR2	QDA	RF	SVM
2 SD		-2.021	-1.715	-1.810	-1.634	-1.705	-1.948	-1.724	-1.695	-1.872
1 SD		-0.690	-0.128	-0.390	0.367	-0.145	-0.523	-0.191	-0.193	-0.369
0 SD		-0.319	0.498	0.069	1.406	0.453	-0.059	0.381	0.348	0.172
-1 SD		-0.910	0.162	-0.433	1.485	0.091	-0.557	-0.007	-0.072	-0.247
-2 SD		-2.461	-1.134	-1.896	0.602	-1.233	-2.015	-1.357	-1.453	-1.628

eliminated variables using the log-likelihood ratio test. This yielded the model presented on Table 5.2 consisting of the fixed effects two ways interactions of all the six predictive factors, random intercepts and slopes. This model explains approximately 62% of the null variance as illustrated on Figure 5.2. The left panel of Table 5.2 presents the estimates of fixed effects, the standard errors and the t statistics while the top-right panel presents the net effect of a standard deviation (SD) unit increase of a given factor conditional on common values of other factors. Finally, the bottom-right panel presents the performance of the classification functions at different values of the predictive factors. From the top-right panel of this table, one notices that on one hand, a 1 SD unit crease in sample size, corresponding to n = 89.67 will lead to an increase in the trBS (decrease in the BS), with the highest increase observed when other variables are at their lowest values. A similar effect is observed for a 1 SD unit increase in the proportion of DE genes and fold change. Except for fold change, this is completely different for direct classifiers where the highest increase of these variables is achieved when other variables are at their highest values, as previously observed by Jong et al., (2016). On the other hand, a 1 SD unit increase in the genes' variances, pairwise correlations of non-DE and DE genes will lead to a decrease in the predictive performance, these effects become very severe when other variables are at their highest values.

Lastly, the bottom-right panel of the table shows that most classification functions will perform reasonably well if the predictive factors are at their average values (0 SD) except for KNN and PLR2. For extreme small values (-2 SD) of the studied factors, all classification functions fail except for PLR1, possibly because of its ability to eliminate correlated variables. Additionally, for extremely large values (2 SD) of all predictive variables, all classification functions fail ($\varphi < 0$). This is an indication that the positively associated factors (sample size, proportion of DE genes and fold change) have a lower combined net effect on the performance of the probabilistic classifiers than the negatively associated factors (pairwise correlations between non-DE and DE genes and genes' variances). This is indeed different for direct classifiers where the positively associated factors had a higher combined net effect for most classification functions than the negatively associated factors [Jong et al., 2016]. As previously stated by Jong et al., (2016), the combination of all other variables simultaneously being at -2, or at +2, is highly unlikely.

5.4 Application

To evaluate the predictive ability of the here presented random effect regression model on real-life data, twelve Affymetrix gene expression datasets of the 25 non-cancerous datasets described in one of our previous study [Novinati et al., 2015] were used. We selected these datasets to

Table 5.3: Characteristics of the 12 datasets used for evaluating the predictive model.

N°	Study	ID+	Affymetrix Platform	Probesets	sampSize	propDE	varaince	deCorr	otherCorr	log2FC	Range of trBS.
1	Alz2*	E-MEXP-2280	HG U133 Plus 2.0	6899	19(07,12)	8.987	0.532	0.589	0.396	1.204	[0.698, 0.913]
2	Ast1	E-GEOD-27011	HG 1.0 ST	32321	36(19,17)	0.517	0.070	0.382	0.261	0.610	[0.552, 0.795]
3	CF	E-GEOD-10406	HG U133 Plus 2.0	54675	15(09,06)	0.267	0.143	0.414	0.408	1.205	[0.163, 0.510]
4	Dia2	E-CBIL-30	HG U133A	22283	26(18,08)	0.242	0.091	0.569	0.359	0.599	[0.178, 0.498]
5	Dys	E-GEOD-19419	HG 1.0 ST	32321	45(22,23)	0.430	0.138	0.247	0.292	0.619	[0.410, 0.524]
6	HF	E-GEOD-26887	HG 1.0 ST	32321	19 (07,12)	0.028	0.107	0.439	0.361	1.189	[0.323, 0.612]
7	PARKI*	E-GEOD-6613	HG U133A	638	83(50,33)	28.527	0.395	0.713	0.341	0.323	[0.467, 0.639]
8	Pso*	E-GEOD-18948	HG U95	1987	16(07,09)	7.952	0.532	0.808	0.602	1.185	[0.378, 0.720]
9	UC1	E-GEOD-14580	HG U133 Plus 2.0	54675	24(16,08)	0.966	0.124	0.675	0.266	1.321	[0.490, 0.647]
10	UC2	E-GEOD-21231	HG 1.0 ST	32321	40(20,20)	0.402	0.139	0.257	0.268	0.631	[0.370, 0.641]
11	UC3	E-GEOD-36807	HG U133 Plus 2.0	54675	28(15,13)	0.819	0.145	0.664	0.287	1.381	[0.419, 0.585]
12	UC5	E-MTAB-331	HG 1.0 ST /HG 1.1 ST	32321	59(30,29)	0.062	0.086	0.941	0.368	1.390	[0.631, 0.958]

* : No filtering was performed + : ArrayExpress accessing ID

The sixth to the eleventh columns correspond to the variables under study. (.,.) represent the sample sizes for each class and last column represents the range of the observed trBS from the nine classifiers.

include a variety of Array platforms, both class balance and class-imbalance, number of DE probesets, as well as various sample sizes. Nine of these datasets were simply preprocessed without filtering while the other three were preprocessed and filtered as described by Novianti et al., (2015). Data characteristics presented on Table 5.3 were then estimated as previously described by Jong et al., (2016)) and presented in the supplementary material or this article.

These data characteristics were standardized using the mean and standard deviations of the respective variables on the simulated data and the linear random effects model was used to predict the trBS of a classification function on each data. We then built and evaluated probabilistic classifiers using the nine classification functions by splitting the data into learning and test sets using leave-one-out cross-validation (LOOCV) with a 5-fold inner cross-validation on the learning set for parameters optimization based on a binomial log-likelihood function. The Brier scores from the LOOCV predicted probabilities were transformed to trBS as described in the method section. To compare the predicted to expected trBS, we used the ranked Spearman correlation between predicted and expected trBS.

The range of the expected trBS of the nine classifiers on each dataset is presented in the last column of Table 5.3 while the results of the predicted by expected trBS comparison, on each dataset are presented on Figure 5.3. If the correlations were random, we will expect both negative and positive correlations between the predicted and expected trBS. The strictly positive correlation values (cor) on this figure indicate a strong agreement between the predicted and expected trBS for most datasets. Low correlations were only observed on (CF and Dys) datasets on

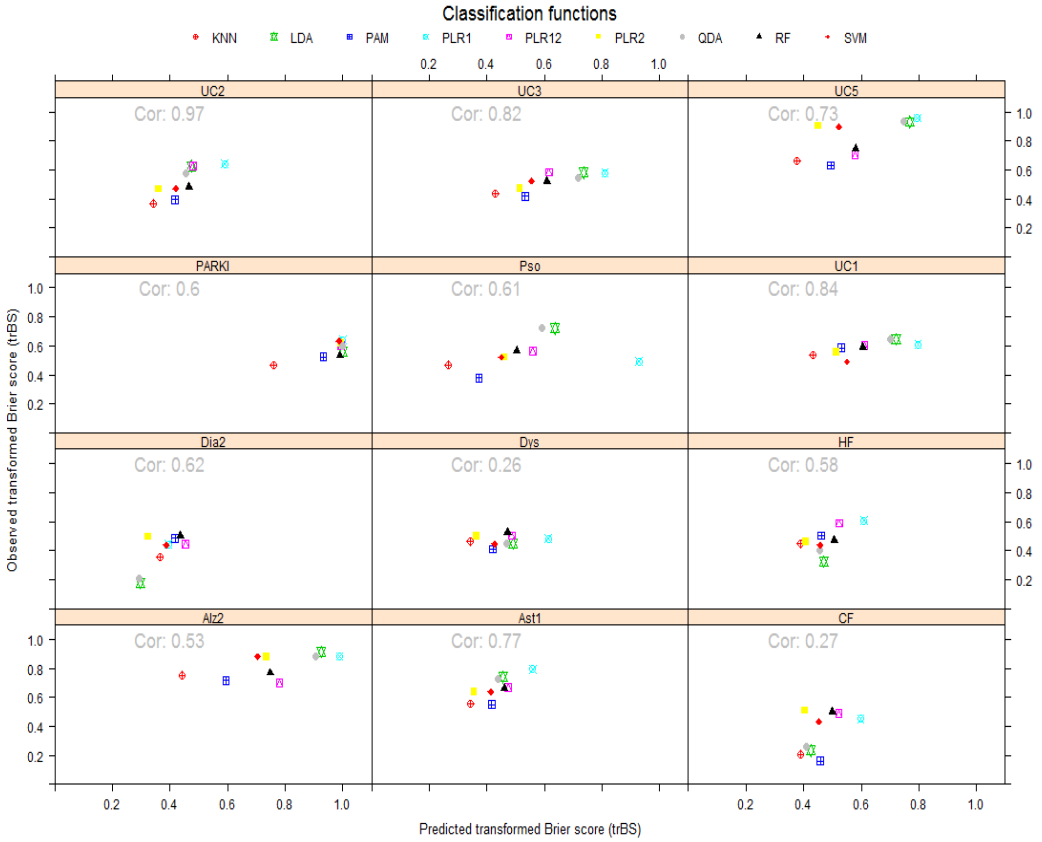

Figure 5.3: Predicted vs Expected (Observed) transformed Brier scores (trBS). *Cor represents Spearman correlations between the predicted and observed trBS.*

which, it was difficult to construct a perfect probabilistic classifier as demonstrated by very low expected trBS (less than or equal to 0.5) for all of the classification functions. For small sample sizes ($n \leq 16$) as in CF and Pso and/or very high correlation between noisy genes (as in Pso), the predicted accuracies for PLR1 (Lasso) are over estimated. Contrary to the theoretical expectation and observations of Novianti et al., (2015) and Jong et al., (2016) that class-imbalance leads to poor performance of the predictive model for direct classifiers, this was not observed in the Dia2 and UC1 datasets that had high class-imbalance. Theoretically, one will envisage an expected accuracy above 0.5 for all classification functions on these datasets. This can be achieved by sending all samples to the majority class but this is not the case for probabilistic classifiers, with as low as 0.178 expected trBS (loosely considered as a measure of accuracy in this case) on Dia2. This indicates that class-imbalance might have less effect on probabilistic classifiers (when maximizing

the log-likelihood function) as compared to direct classifiers (when minimizing the misclassification error rate).

5.5 Discussion

5.5.1 Summary

We hypothesized that the performance of probabilistic classification functions on gene expression data depends on the data's characteristics. Using simulation studies we showed that sample size, proportion of differentially expressed (DE) genes, genes' variances, DE genes' fold changes and magnitude of the pairwise correlation within DE genes and non-DE genes significantly affect the performance of probabilistic classifiers. And we have provided a predictive model with these data characteristics that predicts among the studied probabilistic classification functions, which will yield an optimal probabilistic classifier on a given gene expression data. An application of our predictive model on twelve real-life gene expression datasets showed very high agreement between the predicted and expected Brier scores of the probabilistic classifiers. Thus, our predictive model might serve as a basis for determining an optimal probabilistic classification function for a given gene expression data.

5.5.2 Current knowledge in the field

Classification functions have been shown to perform differently across datasets [Lee et al., 2005; Shi et al., 2010; Kim & Simon, 2011; Kruppa et al., 2014a; Kruppa et al., 2014b] and data characteristics as well have been shown to differ across diseases/datasets [Jong et al., 2014] and are associated to the performance of direct classification functions [Novianti et al., 2015; Jong et al., 2016]. While sufficient knowledge is available on the properties of most classification functions and procedures to build class prediction models using gene expression data have been outlined by Wessels et al., (2005) and Shi et al., (2010), little is known about data characteristics that accounts for the variability in the performance of probabilistic classification functions and how to use these characteristics to choose an optimal function for a specific dataset.

Since gene expression data often suffers from the curse of dimensionality, classification with such data utilizes cross-validation (CV) to estimate the error rate that is generalizable to independent data. Nevertheless, since there is no universally best machine learning algorithm, it is common practice to compare several algorithms and report the algorithm (function) that produces the smallest cross-validation error. This approach leads to selection bias because there is a high probability that an algorithm has the smallest error simply by chance. This probability increases

with both the number of algorithms compared and a decrease in sample size. Thus, selection bias is often high in data with small sample sizes and when several algorithms are compared [Ding et al. 2014]. Ding et al., (2014) stated as an example that one uses a small pilot data, compares several machine learning methods and selects the minimum error classifier (MEC) with a falsely small error because of the selection bias. When the model proceeds to a large cohort validation for translational research, it will likely fail. As such, several bias correction methods have been proposed in the literature of class prediction with gene expression data. Nevertheless, no such method is 100% effective when several least optimal algorithms are compared on a dataset with a small sample size, as observed by Ding et al., (2014).

Therefore, most researchers adhere to specific classification function(s) or randomly choose a classification functions for their class prediction analysis irrespective of the disease or data under study. Recently, Jong et al., (2016) attempted to subvert this challenge by providing a predictive model to select amongst ten classification functions, an optimal direct classifier for a given gene expression data using the data characteristics. Nevertheless, in clinical applications, direct classification is of less value relative to probabilistic classification that is often preferred because of its ability to quantify the uncertainty around predictions [Pepe, 2005; Kim & Simon, 2011; Malley et al., 2012; Simon, 2014].

In this study, we focused on probabilistic classifiers and we evaluated the effects of data characteristics on probabilistic classifiers with the goal to provide a model for determining an optimal probabilistic classifier for a given gene expression dataset. We employed an approach of Jong et al., (2016) that is aiming to be free of biases, to choose an algorithm that is optimal for a given dataset. For a given dataset, this approach quantifies data characteristics, uses a model to predict which algorithm will be optimal for this data and then uses that algorithm for predictions. One will immediately think of two types of biases in this approach namely, selection bias and optimistic bias.

(i) Selection bias, because through our prediction model we are implicitly comparing algorithms using their predicted errors and choosing the one with the smallest error. Nevertheless, our predicted errors are expected values based on large independent datasets. Hence, cannot occur by chance as CV errors. Theoretical properties of bias show that for a fixed number of algorithms, bias approaches zero as sample size grows large enough [Ding et al., 2014]. Hence, our approach is free of selection bias since the test sets sample sizes in our simulations were extremely large (5000 samples by 1000 iterations).

(ii) Optimistic bias because using the data to determine an optimal algorithm introduces bias in the assessment of predictive performance resulting from overfitting. Nevertheless, our approach is less prone to optimistic bias because: (a) most important variables (e.g. correlations) in our model do not utilize class labels and those utilizing the class labels for their estimations are summarized single values; an approach which is completely different from resubstitution that account for most if not all of optimistic bias. (b) though used for algorithm selection, the samples have not been seen by any such algorithm. Hence, the hypothesis chosen from the hypotheses class by the selected algorithm is entirely based on the training set making the test set an independent (unseen) set to such a hypothesis.

That notwithstanding, should there be any optimistic bias from our approach, it happens only on the experimental data and not on the mandatory independent validation data often required nowadays. More so, such optimistic bias on a generalizable algorithm is preferable to a less generalizable algorithm falsely chosen due to selection bias. Generalizability is plausible because the generalized performance of an algorithm (i.e. hypotheses class) depends on its interaction with the data characteristics (i.e. assumed data distribution) and since the pilot study is often a random sample from the population, the estimates of the data characteristics should be close to expected values in the population. Hence, our approach comes handy in that it allows one to use exclusively (without selection bias) the data characteristics from the pilot study to select an algorithm that can be generalized to larger cohorts.

5.5.3 Comparison between direct and probabilistic classifiers

Nine of the ten elective choices of classification functions used by Jong et al. (2016) can produce probabilities either by design or self-post-processing. These functions were selected an optimized as probabilistic classifiers using extensive simulations with same parameters space as outlined by Jong et al., (2016). The effects of the data characteristics on the Brier scores of these functions were investigated using a linear mixed effects model. Although we focused on probabilistic classifiers, we found that sample size, proportion of DE gene and DE genes' fold changes are positively associated to the transformed Brier score (trBS) which is a surrogate measure for accuracy, while genes' variances and the pairwise correlations within DE and non-DE genes are negatively associated to the trBS as was observed for the accuracies of direct classifiers. Nevertheless, we found that:

(i) for the same values of the studied variables, the optimality of a function depends on whether it is trained as a direct or probabilistic classifier.

(ii) the performance of probabilistic classifiers on a given dataset depends on whether the functions' parameters were optimized using accuracy or log-likelihood function.

(iii) pairwise correlations within non-DE genes accounted for the highest explained null variance for probabilistic classifiers as opposed to sample size that accounts for the most explained null variance for direct classifiers. This indicates that most probabilistic classifiers are highly sensitive to pairwise correlations while most direct classifiers are highly sensitive to sample size compared to other factors.

(iv) based on the application of our linear mixed effects model on real-life data, one might conclude that class-imbalance has less effect on probabilistic classifiers as compared to direct classifiers. Whereas direct classifiers are theoretically expected to be affected by class-imbalance when using accuracy as a measure of evaluation, probabilistic classifiers might not be severely affected by class-imbalance because it might simply serve as a prior probability for all classification functions. Nevertheless, class-imbalance was not considered in our simulations and might be a factor worth investigating for both direct and probabilistic classifiers.

5.5.4 Conclusion

We have shown that the optimality of a classification function on a given dataset depends on whether it is trained as a direct or probabilistic classifier. We have provided a guide for choosing an optimal probabilistic classifier for a specific data, using the data's characteristics and the studied probabilistic functions through a linear random effects regression model. Our model might serve as the basis for selecting an optimal probabilistic classification function for diagnostic and prognostic analysis with gene expression data.

Acknowledgements

This work has been supported by the VIRGO consortium, which is funded by the Netherlands Genomics Initiative and by the Dutch Government (FES0908). The funding agencies in no way influenced the outcome or conclusions of the study. The authors would like to thank the VIRGO consortium and its sponsors, the Netherlands Genomics Initiative and the Dutch Government for the financial support. Next to this, we thank Martin Marinus and the HPC-team at UMC Utrecht for the high performing computing facilities.

Chapter 5

Supplementary Material

Supplementary material for this manuscript is available upon request to the authors and include:

Gene expression data simulation and quantification of real-life data characteristics for application

Figure S1: (A) Assumed correlation structure; **(B)** Algorithm to simulate data, build and validate class prediction models

Figure S2: Average Brier scores of the nine classification functions for: **(A)** sample size of 20, log_2 fold change of 1 and 5% DE genes; **(B)** sample size of 100, log_2 fold change of 1 and 1% DE genes; **(C)** sample size of 100, log_2 fold change of 0.5 and 5% DE genes

Figure S2: Average Brier scores of the nine classification functions for: **(A)** sample size of 20, log_2 fold change of 1 and 5% DE genes; **(B)** sample size of 50, log_2 fold change of 1 and 1% DE genes; **(C)** sample size of 50, log_2 fold change of 0.5 and 5% DE genes.

Figure S3: Average misclassification error rates of the 9 classification functions for sample size of 100, log_2 fold change of 1 and 5% DE genes.

Figure S4: Average Brier scores of the nine classification functions for sample size of 50, log_2 fold change of 1 and 5% DE genes optimizing on **(A)** log-likelihood function and **(B)** accuracy functions.

Source code has been incorporated in the R package *SPreFuGED*

Chapter 6

Choosing a Cox's predictive function for survival analysis with gene expression data.

Victor L. Jong, Kit C. B. Roes & Marinus J. C. Eijkemans

submitted

Chapter 6

Abstract

Motivation: Survival prediction with gene expression data is widely becoming a regular analysis in personalized medicine and prognostic modeling. The literature reveals that Cox's predictive functions perform differently across gene expression datasets. The questions; (i) what gene expression data characteristics affect the performance of Cox's predictive functions and (ii) which predictive function should be used for a given dataset remains unascertained. In this study, we hypothesized possible data characteristic and investigate their association to the performance of Cox's predictive functions. Additionally, we aim to provide a guide for determining an optimal function for any given gene expression data.

Results: Gene expression data were simulated for different values of sample size, proportion of events, proportion of informative genes, genes' variances, log_2 effect sizes of informative genes and absolute values of the pairwise correlation within and between informative and non-informative genes. For each simulated dataset, seven Cox's predictive functions were trained and evaluated using integrated Brier score (IBS). The resulting IBS of 13608 data points from 1944 different simulation scenarios by 7 predictive functions were then modeled using a linear random effects regression on the studied data characteristics, yielding a model that predicts the IBS of the functions on a given data. An application of our model on two real-life datasets showed high correlations (0.67 & 0.89) between the predicted and expected IBS.

Conclusion: We have shown that sample size, proportion of events, proportion of informative genes, genes' variances, absolute values of the pairwise correlation within and between informative and non-informative genes associate to the performance of Cox's predictive functions. And we have presented a predictive model that might serve as a guide for determining an optimal Cox's predictive function for any given gene expression data.

6.1 Introduction

With the rapid advancement of biotechnology that enables genome-wide measurement of DNA sequence, RNA abundance and gene copy number, there has been an explosion of interest in predictive modeling with these data [Simon et al, 2011]. These technologies allow for the quantification of expression of thousands of genes at once. Microarray gene expression profiling in particular, has become a widely used tool to identify particular disease subpopulations and to perform diagnostic or prognostic predictions [van 't Veer et al., 2002; Huang et al., 2010] and survival predictions [Bøvelstad et al., 2007]. In survival analysis for instance, one studies survival time, which is defined as the time length from the beginning of observation until death (or some other event) of the observed patient or until the end of observation. The main goal is to predict the time to event [van Wieringen et al., 2009] or to classify patients into two or more risk groups [Simon et al., 2011], using the expression of the genes as explanatory variables.

For classical time-to-event modeling, the number of samples (n) exceeds the number of parameters (p) and the effective sample size n is the number of events, where $\frac{n}{p}$ ratios of 10 or even 20 are frequently recommended for the development of stable models [Simon et al., 2011]. While for genomic data where the number of parameters greatly exceed the number of samples $(p \gg n)$, this rule fails. As such, several methods have been proposed in the literature of survival modeling with high-dimensional data that perform dimension reduction and prevent overfitting. Nevertheless, these methods have been shown to perform differently across gene expression datasets irrespective of the measure of evaluation used [Bøvelstad et al., 2007; van Wieringen et la., 2009]. Thus, the question, which survival function should one choose for a given dataset remains a challenge.

Though a substantial amount of information is known about the characteristics of the predictive functions and predictive model building procedures, little is known about which data characteristics affect the performance of a survival function. For instance, penalized regressions like Ridge, Lasso, Elastic net are capable of handling correlated variables in one way or the other and might take into account the correlations that often exist between genes. It is therefore necessary to study how correlations and other data characteristics affect the performance of survival predictive functions. A common practice in classification analysis that could be applied to survival prediction is to compare several predictive functions and select the best (function with the smallest prediction error), but it has been reported in classification analysis that it leads to

selection bias [Varma & Simon, 2006; Tibshirani & Tibshirani, 2009; Bernau et al., 2013; Ding et al., 2014] because a function might have the smallest error simply by chance. This "capitalization on chance" probability increases with both the number of algorithms compared and a decrease in sample size. Thus, selection bias is often high in data with small sample sizes and when several algorithms are compared [Ding et al. 2014]. Ding et al., (2014) stated as an example that one uses a small pilot data, compares several machine learning methods and selects the minimum error predictor (MEP) with a falsely small error because of the selection bias. When the model proceeds to a large cohort validation for translational research, it will likely fail. Hence, several bias correction methods have been proposed in the literature of prediction with gene expression data. Nevertheless, no such method is guaranteed to be effective when several least optimal algorithms are compared on a dataset with a small sample size, as observed by Ding et al., (2014) for classification.

As such, some researchers adhere to one or a few functions irrespective of the dataset, disease or medical question being addressed. While others choose a function for their datasets by affinity or familiarity without taking into account the characteristics of such data [Jong et al., 2016]. The literature stipulates the use of super-learners which combine the scores of several predictive functions, to improve overall prediction. Nevertheless, super-learners are less acceptable in medical applications because there is resistance of pathologists and medical practitioners to use uninterpretable black box multivariate tests for making important treatment decisions [Simon, 2014]. Additionally, super-learners often utilize the entire genome instead of a selected profile, making practical application time consuming and costly. Thus, traditional predictive functions that can yield interpretable models are still preferred over super-learners.

In this study, we determine gene expression data's characteristics that affect the performance of Cox's predictive functions with a secondary goal to provide a guideline for choosing an optimal Cox's predictive function for a given gene expression data. We investigate the effect of sample size, proportion of events, proportion of informative genes, genes' variances, effect sizes, pairwise correlations between informative and noisy genes on the integrated Brier Scores of Cox's survival functions using extensive simulations.

The remainder of this article is organized as follows: a methodology to simulate data, survival functions considered and the building and evaluation of survival models are presented in Section 6.2; Section 6.3 contains a summary of a predictive model constructed from the results of the Cox's survival functions for different simulated scenarios; Section 6.4 provides an application of

our predictive model from the simulated results on real-life microarray gene expression datasets and a discussion is presented in Section 6.5.

6.2 Methodology

6.2.1 Simulated data (scenarios)

To simulate gene expression data, we hypothesized that sample size, proportion of events, proportion of informative genes, genes' variances, effect sizes, pairwise correlations between informative and noisy genes might be associated to the performance of survival functions as was previously observed by Jong et al., (2016) for classification functions. These seven variables were to be systematically varied in our simulations. From observed correlation structures in real-life gene expression datasets [Jong et al., 2014], we generalized the structure into three clusters as was previously reported by Jong et al., (2016). These clusters consist of positively, negatively and non-associated genes to the survival time. The absolute values of pairwise correlations of informative genes (ρ) were varied as 0.00, 0.25, 0.50 and 0.75 with positively associated cluster taking oppositely-signed correlation values to the negatively associated cluster. The pairwise correlations within the non-associated (noisy) cluster and between the noisy and the informative (both positively and negatively associated) clusters were per gene-pair randomly drawn from a normal distribution centered at zero with a standard deviation θ i.e. $N(0, \theta)$ where $\theta = 0.00, 0.25, 0.50$. The scenario $\rho = \theta = 0.00$ corresponds to complete independence. Resulting correlation values lying outside the interval [-1, 1] were uniformly converted to the intervals [-1, -0.15] and [0.15, 1] for negative and positive values respectively. The variances of the genes ($\sigma^2 = \frac{1}{\lambda}$) were drawn from an exponential distribution i.e. $exp(\lambda)$ where $\lambda = 0.5, 1, 1.5$. The distributional assumptions were made based on observation from real-life datasets as experienced by Jong et al., (2014) and Novianti et al., (2015) and as was previously utilized by Jong et al., (2016). With the correlation values and the variances, the covariance matrix Σ was constructed. In addition, the proportion of informative genes (π) was also allowed to take up 1%, 3% and 5% of the total number of genes (p), as values. This resulted to 108 different complex covariance matrices that were used to simulate the data for different values of other variables. Finally, different values of absolute log_2 effect size (β), proportion of events (τ) and different sample sizes (n) were considered as shown on Table 6.1.

For a fixed number of genes ($p = 1000$) and n samples, the gene expression data of $n \times p$ dimensions was generated from a multivariate normal distribution with mean vector from a

Table 6.1: Simulated gene expression data characteristics

Data characteristics	Values
Sample size (n)	100, 200
Proportion of events (τ)	25%, 50%, 75%
Proportion of informative genes (π)	1%, 3%, 5%
log_2 effect size of informative genes (β)	0.5, 1.0, 2.0
Pairwise correlations of informative genes (ρ)	0.0, 0.25, 0.50, 0.75
Gene' variances $(\sigma^2 = \frac{1}{\lambda}) \sim exp(\lambda)$	$\lambda = 0.5, 1.0, 2.0$
Pairwise correlations of noisy genes $(\gamma) \sim N(0, \theta)$	$\theta = 0.0, 0.25, 0.50$

50% of π were each positively and negatively associated to survival time.

uniform distribution $U(6,10)$ of length p and the covariance matrix Σ constructed as described above, using Cholesky decomposition as a method to determine the root of the covariance matrix. The choice of multivariate normal distribution and mean vector corresponds to the practical assumption that gene expression data are normally distributed in log_2 scale and based on observation that the log_2 expression values often fall in the interval (0, 16) [Jong et al., 2014]. The survival time (T) for the samples were generated from a Weibull distribution as follows:

$$T = \left(-\frac{log(U)}{\xi \times exp(\beta' X)} \right)^{1/v} \tag{6.1}$$

where U is a variable following a uniform distribution on the open interval (0, 1); $\xi = 0.01$ is a constant baseline hazard; β is a vector of the effect sizes, X is a matrix of the gene expression data and v was determine as $0.1 \times p$ for independent scenario or $(\rho + \theta) \times p$ for other scenarios. These values were chosen to restrict the survival time to a comparable range across scenarios. Varying v and ξ will give different ranges for survival time. The Weibull distribution was utilized because it is one of the distributions that shares the assumption of proportional hazards with the Cox model [Bender et al., 2005]. Lastly, a sample's status label $C = 0 \; or \; 1$ was generated from a Bernoulli distribution with a probability (τ). For each combination of the values of the data characteristics, the dataset was simulated as shown in Figure 6.1, yielding 1944 different simulation scenarios, each of which was randomly replicated 100 times.

6.2.2 Survival predictive functions

In this study, we considered a response variable which is possibly a censored survival time and the Cox proportional hazard model [Cox, 1972] is used for inference. Cox models the instantaneous

Algorithm 1

For proportion π in 1%, 3% and 5% of $p = 1000$ as informative{
For log_2 fold change of informative genes Δ in 0.5, 1, 2{
For absolute pairwise correlation of informative genes ρ in 0.00, 0.25, 0.50, 0.75{
For pairwise correlation of other genes (γ) ~$\boldsymbol{N(0,\theta)}$; θ in 0.00, 0.25, 0.50{
For genes' variance (σ^2) ~$\boldsymbol{Exp(\lambda)}$; λ = 0.5, 1, 2{
For sample size n in 100, 200 {
For number of events τ in 25%, 50%, 75% of n {

Let $p_2 = \pi \times p$ be the number of informative genes of which $1, \ldots, p_1 = \frac{p_2}{2}$ are positively and $p_1 + 1, \ldots . p_2$ are negatively associated to survival time.

1. **Construct covariance matrices from σ^2, ρ & γ**

$$\Sigma = \begin{matrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma'_{12} & \Sigma_{22} & \Sigma_{23} \\ \Sigma'_{13} & \Sigma'_{23} & \Sigma_{33} \end{matrix}$$

For iteration B in 1, ..., 100{

2. **Generate $\boldsymbol{log_2}$ gene expression data**

$\mu = \boldsymbol{U(N, 6, 10)}$ $\quad X = \boldsymbol{mvN}(n, \mu, \Sigma)$

3. **Generate survival time and censoring status:**

Survival time, $T = \left(-\frac{\log(U)}{\xi \exp(\beta' X)}\right)^{1/v}$ and censoring status, $C = \boldsymbol{Bin}(n, \tau)$

where $U = \boldsymbol{U}(n, 0, 1)$, $\beta = c(rep(\Delta,\ p_1),\ rep(-\Delta,\ p_2 - p_1),\ rep(0,\ p - p_2))$, $\xi = 0.01$ and $v = 0.1 \times p$ or $v = (\rho + \theta) \times p$

4. **Build and evaluate predictors using 7 predictive functions**

Predictors are built with $D = \{X, T, C\}$ and evaluated using IBS on a test set $D_t = \{X_t, T_t, C_t\}$ generated from a similar model as D

}
}
}
}
}
}
}
}
}

Figure 6.1: Algorithm to simulate data, build and validate survival predictive models; for each value of the seven variables, the covariance matrix was constructed in step 1 and the mean vector in step 2, the learning and test data were simulated at step 3 and predictive models were built and validated in step 4. Steps 2-4 were then repeated 100 times.

risk of an event at time t for an individual i with gene expression values $X_i = (X_{i1}, \ldots, X_{ip})'$ as

$$h(t|X_i) = h_0(t)\, exp(\beta X_i) \tag{6.2}$$

where $\beta = (\beta_1, \ldots, \beta_p)$ is a vector of regression coefficients and $h_0(t)$ is the baseline hazard. In classical settings ($n > p$), the regression parameters are estimated by maximizing partial likelihood:

$$L(\beta) = \prod_{k \in D} \frac{exp(\beta X_k)}{\sum_{m \in R_k} exp(\beta X_m)} \tag{6.3}$$

where D is the set of indices of the failure times, R_k is the set of indices of the individuals at risk at time t_k and we have assumed there are no ties in the survival times. In the case of ties, we use the "Breslow" approximation to the partial likelihood. We also assumed that the censoring is non-

informative, so that the construction of the partial likelihood is justified. This is equivalent to maximizing the log partial likelihood:

$$l(\beta) = \sum_{k \in D} \left[\beta X_k - log \left(\sum_{m \in R_k} exp(\beta X_m) \right) \right] \tag{6.4}$$

Nevertheless, in high-dimensional *($p \gg n$)* settings, this leads to overfitting. As such several dimension reduction techniques have been proposed in the literature to tackle this curse of dimensionality problem. Seven elective choices that are suited for Cox proportional hazard models and which have been utilized for survival predictions with gene expression data were chosen and are briefly described below:

- *Supervised principal components regression (superPC)*: Supervised principal components regression was first proposed by Bair & Tibshirani, (2004). This procedure ranks genes based on their correlation to the survival time (Cox's score) using the actual gene expression values then picks out a subset of the genes with the highest scores and applies a principal component analysis (PCA) to this subset. After which a few principal components are plugged in eq. (6.2). In our analysis, we determine the optimal subset using cross-validation [Tibshirani & Efron, 2002] and utilized the first principal component (PC1) in a multivariate Cox model. This algorithm is implemented in the R package `superpc` [Bair & Tibshirani, 2004].

- *Cox univariate shrinkage (uniCox):* This is a method for prediction in Cox's proportional hazards model for high-dimensional data. The method assumes that the features are independent in each risk set, so that the partial likelihood in eq. (6.3) factors into a product and the log partial likelihood in eq. (6.4) reduces to a sum of the form

$$J(\beta) = \sum_{j=1}^{p} g_j \beta_j - \eta \sum |\beta_j| \tag{6.5}$$

  where $\eta \geq 0$ is a shrinkage parameter determined through cross-validation by drop in deviance [Tibshirani, 2009]. As such, it is analogous to univariate thresholding in linear regression and nearest shrunken centroids in classification. The method has the attractive property of being essentially univariate in that the features are entered into the model based on the size of their Cox's score statistics [Tibshirani, 2009]. The algorithm is implemented in the R package `uniCox` [Tibshirani, 2009].

- L_2- *penalized Cox regression (PLR2):* Ridge (L_2 penalized) regression shrinks the regression coefficients by imposing a penalty on their squared values and the regression coefficients are estimated by maximizing the penalized log partial likelihood

$$l(\beta) - \eta \sum_{j=1}^{p} \beta_j^2 \tag{6.6}$$

where $l(\beta)$ is the log partial likelihood given in eq. (6.4) and the second term is a penalty determining the amount of shrinkage [Verweij & van Houwelingen, 1994; van Houwelingen et al., 2006]. Many coefficients shrunk toward zero to reduce overfitting but are never zero. The tuning parameter η is determined by cross-validation as implemented in the R package `penalized` [Goeman, 2010].

- L_1- *penalized Cox regression (PLR1):* Lasso (L_1 penalized) regression shrinks the regression coefficients by imposing a penalty on their absolute values and the regression coefficients are estimated by maximizing the penalized log partial likelihood

$$l(\beta) - \eta \sum_{j=1}^{p} |\beta_j| \tag{6.7}$$

where $l(\beta)$ is the log partial likelihood given in eq. (6.4) and the second term is a penalty determining the amount of shrinkage [Goeman, 2010]. Lasso has an advantage over ridge in that many coefficients shrunk exactly to zero thus, performing feature selection. The tuning parameter η is determined by cross-validation as implemented in the R package `penalized` [Goeman, 2010]

- L_1 & L_2 - *penalized Cox regression (PLR12):* Elastic net (L_1 & L_2)- penalized regression was introduced by Zou & Hastie, (2005) and is a combination of the ridge and lasso penalties. L_2 penalty tends to result in all small but non-zero regression coefficients, whereas applying an L_1 penalty tends to result in many regression coefficients shrunk exactly to zero and a few other regression coefficients with comparatively little shrinkage. Combining L_1 and L_2 penalties tends to give a result in between, with fewer regression coefficients set to zero than in a pure L_1 setting and more shrinkage of the other coefficients [Goeman, 2010]. A typical example is that L_1 eliminates correlated variables in its model but in some applications, these variables could be of interest. As such combining both lasso and ridge penalties allows for such

variables to be part of a model. This algorithm is also implemented in the R package `penalized` [Goeman, 2010]

- *Cox model by likelihood based boosting (coxBoost):* This algorithm fits a Cox proportional hazards model by componentwise partial likelihood based boosting. In contrast to gradient boosting, CoxBoost is not based on gradients of loss functions but adapts the offset-based boosting approach from Tutz & Binder, (2007) for estimating Cox proportional hazards models. In each boosting step the previous boosting steps are incorporated as an offset in penalized partial likelihood estimation, which is employed to obtain an update for a single covariate, in every boosting step. This results in sparse fits similar to Lasso-like approaches, with many estimated coefficients being zero. The main model complexity parameter, which has to be selected by cross-validation, is the number of boosting steps [Binder & Schumacher, 2008]. This algorithm is implemented in the R package `coxBoost` [Binder & Schumacher, 2008].

- *Random survival Forest (RF):* Random survival forest algorithm utilizes the approach of Brieman's Random Forest [Breiman, 2001] to right-censored data. Similar to Brieman's algorithm, trees in a survival forest are grown in a two-step randomization process. (i) A tree is grown with a bootstrap sample of two-third the original sample size. (ii) At each node, a random sample of predictors m sampled from all predictors is selected and the predictor with the split that maximizes survival difference between daughter nodes is used. Step (ii) above is repeated until the tree is grown to terminal nodes with minimum size k. Finally, it calculates cumulative hazards function (CHF) for each tree and average to obtain the ensemble CHF and then using out-of-bag (OOB) data, it calculates prediction error for the ensemble CHF [Ishwaran et al., 2008]. The parameters m and k are often optimized using cross-validation. This algorithm is implemented in the R package `randomForestSRC`.

  Though Ishwaran et al., (2008) used OOB CHF to compute C-index as a prediction error which could be used to determine optimal values for m and k, we utilized the integrated Brier score (IBS) defined below, in determining our optimal values for m and k. To do this, we divided our learning data using 5-fold cross-validation and each fold is left out while the remaining folds are used without resampling in step (i) above and at the end, the predicted survival probabilities of the left-out samples at the event time points were used to compute the IBS.

6.2.3 Building and evaluating Cox's survival models

To assess the dependency of the chosen functions on the characteristics of the simulated gene expression data, we built on each simulated dataset, Cox's survival prediction models with all the functions listed above. The simulated dataset was considered as a learning set which was split into an inner-test set and a cross-validation set using 5-fold cross-validation and the parameters of the functions were optimized by maximizing the log partial likelihood following Verweij & van Houwelingen, (1993) procedure unless stated otherwise. With the optimal parameter(s) for each function, the prediction models were built using the entire learning set. The resulting models were evaluated using IBS on a test set consisting of 100 samples generated from the same model as the learning set (see Figure 6.1). The IBS of the functions on this test set were recorded and the process was repeated 100 times (sampling both learning and test sets) for each simulation scenario and the resulting IBS over the 100 replications were used for further downstream analyses.

We utilized the IBS because it is also suitable for non-Cox's models and has thus become a standard evaluation measure for survival prediction methods [Hothorn et al., 2006; Schumacher et al., 2007] and is defined as:

$$IBS = \frac{1}{max(t_i)} \int_0^{max(t_i)} BS(t)dt \tag{6.8}$$

$$BS(t) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\hat{S}(t|X_i)^2 I(t_i \leq t \wedge C_i = 1)}{\hat{G}(t_i)} + \frac{\left[1 - \hat{S}(t|X_i)\right]^2 I(t_i > t)}{\hat{G}(t_i)} \right]$$

where $BS(t)$ is a time dependent Brier score, $\hat{G}(.)$ denotes the Kaplan-Meier estimate of the censoring distribution which is based on the observations $(t_i, 1 - C_i)$ and I is the indicator function. The numerator of the first summand is the squared predicted probability that individual i survives until time t if he actually died (uncensored) before t, or zero otherwise. The better the survival function is estimated, the smaller is this probability. Analogously, the numerator of the second summand is the squared probability that individual i dies before time t if he was observed at least until t, or zero otherwise. Censored observations with survival times smaller than t are weighted with 0 [van Wieringen et al., 2009]. The values of the Brier Score range between 0 and 1 with good predictions at time t resulting in small Brier Scores [van Wieringen et al., 2009]. Time dependent Brier scores and IBS were computed with the R package `ipred` [Peters & Hothorn, 2015].

6.2.4 Random effects linear regression

An average of the IBS of each and every function over 100 replications for each simulated scenario was computed yielding 13608 data points resulting from the 1944 different simulation scenarios by 7 Cox's predictive functions. The IBS were then transformed to an accuracy like measure $\text{trIBS} = 1 - \text{IBS}$ and these accuracies were modeled using a linear random effects regression model with the predictive function as the random effects clustering variable, by transforming the accuracies to an unbounded range using the logit function. For the ℓ^{th} standardized study factor, the random effects model is written as:

$$\varphi = log\left(\frac{\tilde{\pi}(x_{ij})}{1-\tilde{\pi}(x_{ij})}\right) = Y_{ij} = \beta_0 + \vartheta_{0j} + (\beta_1 + \vartheta_{1j})X_{ij}^{\ell} + \varepsilon_{ij} \tag{6.9}$$

where $\beta = (\beta_0, \beta_1)'$ are respectively the fixed intercept and slope, $0 < \tilde{\pi}(x_{ij}) < 1$ is the expected trIBS of function j in scenario i, $\vartheta_j = (\vartheta_{0j}, \vartheta_{1j})' \sim N(0, D)$ are respectively the random intercepts and slopes of the functions while $\varepsilon_{ij} \sim N(0, \sigma^2)$ are the independent and identically distributed residuals, also independent from the random effects ϑ_{j}. D is a 2×2 covariance matrix of the random effects. All the aforementioned study factors were evaluated by univariate and multivariate linear random effects regression models. Multivariate regression evaluation was done by a backward selection whereby at each step, two nested models, with and without a particular study factor, were compared by log-likelihood ratio test at 5% significance level. Each factor ℓ was also evaluated by its explained null variance defined as:

$$\text{Var}_{\ell} = \frac{\text{MSE}_{\text{null}} - \text{MSE}_{\ell}}{\text{MSE}_{\text{null}}} \tag{6.10}$$

where MSE_{null} and MSE_{ℓ} are the mean square errors of the null (random intercept only) and the ℓ^{th} standardized study factor models respectively. The explained null variance of the selected multivariate model was also evaluated.

6.2.5 Software

All statistical analyses were performed in `R software version 3.3.1`, and `Bioconductor` [Gentleman et al., 2004] using the following packages: `mvtnorm` [Genz and Bretz, 2009] for simulating data, the packages listed above in Sections 6.2.2 and 6.2.3 for predictive modeling and evaluation, `lattice` [Sarkar, 2008] for visualization, `lme4` [Bates et al., 2015] for linear random effects modeling and `survival` [Therneau & Grambsch, 2000] to determine informative genes based on their univariate Cox's coefficeints. Additionally, we have

incorporated the codes to replicate and apply these results in the R package `SPreFuGED` [Jong et al., 2016].

6.3 Results

Figure 6.2 shows the average IBS over the 100 random replicates (y-axis) of the functions (lines) for different values of absolute pairwise correlations of informative genes (x-axis), pairwise correlations of non-informative genes (rows) and genes' variances (columns) at a fixed proportion of informative genes ($\pi = 1\%$), effect size ($\beta = 2$), proportion of events ($\tau = 75\%$) and sample size ($n = 200$). From this figure, one notices that for data with completely independent variables or data with independent non-informative genes, univariate methods (superPC and uniCox) outperform other methods while for data with correlated non-informative genes, Lasso (PLR1) and Elastic net (PLR12) outperform other methods. For data with uncorrelated non-informative genes, variability of methods' performance increases with decrease in genes' variances (column 1, rows 3-1) while for data with correlated non-informative genes, variability of methods' performance decreases with decrease in genes' variances (columns 2&3, rows 3-1) .

On the other hand, if the proportion of informative genes increases, the performance of the methods tends to depend on the correlation values of both the informative and non-formative genes and the genes' variances. Specifically, univariate methods (superPC and uniCox) tend to perform worse when the data has uncorrelated non-informative genes and highly correlated informative genes (supplementary Figure S1A column 1, rows 2&3) while IBS of tree-based method RF decreases with increase correlation values of informative genes; this method is among the worse methods for uncorrelated data and best for highly correlated informative genes (supplementary Figure S1A column 1). Other scenarios show: (i) for small log_2 effect size (supplementary Figure S1B) univariate methods again outperform other methods when the non-informative genes are uncorrelated while Lasso and Elastic net outperform others if the non-informative genes are correlated, (ii) a negative association of proportion of events with IBS for most of the functions except for boosting (supplementary Figure S1C) and less variability between the functions on data with correlated non-informative genes and (iii) a decrease in sample size generally leads to an increase in the IBS of all the functions but the per function effect depends on the interaction of this variable with other variables. The random intercepts and slopes of the functions across scenarios indicate the interactions between these functions and data characteristics which thus, indicate that the optimality of a function depends on these data characteristics.

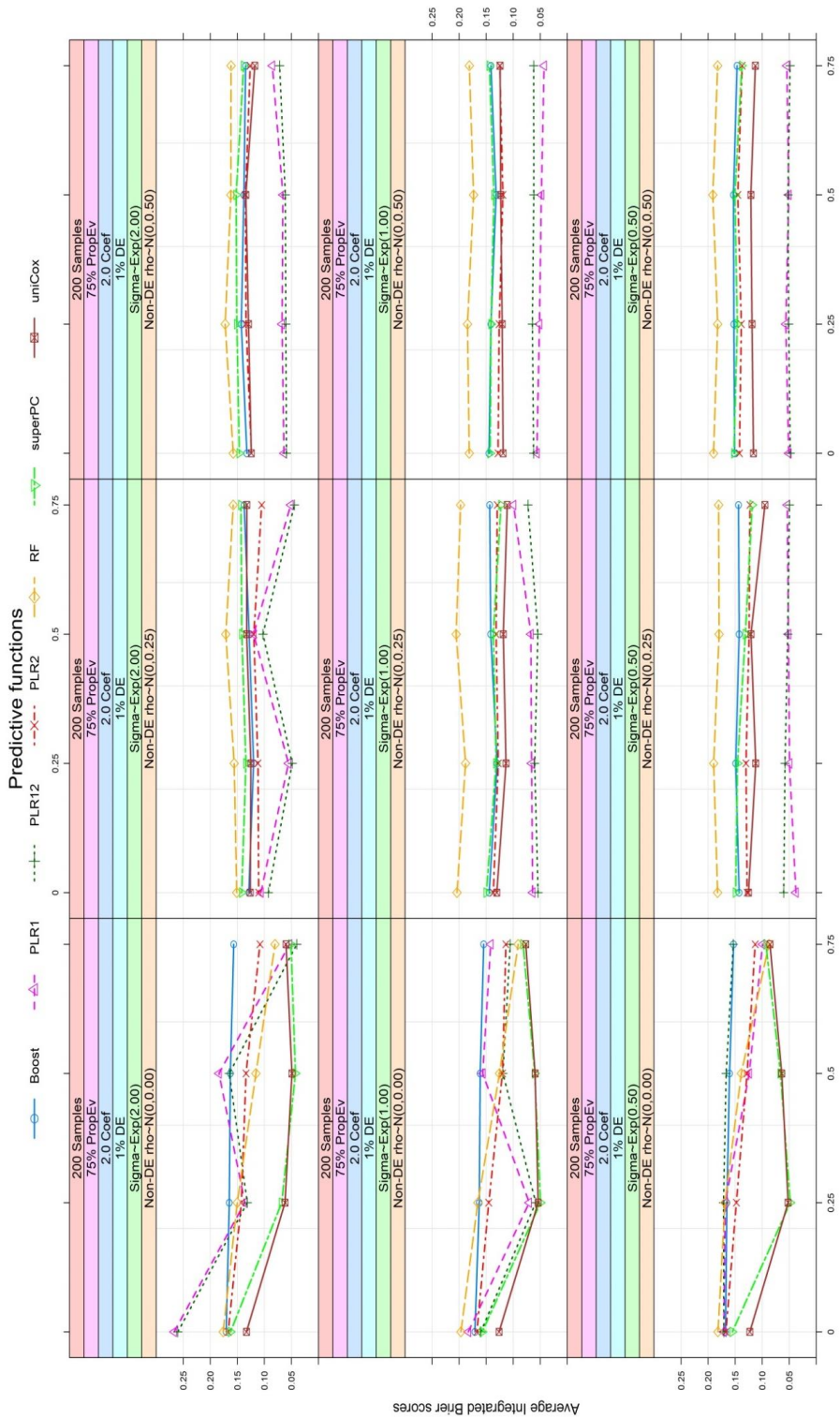

Figure 6.2: Average Integrated Brier Scores (y-axis) of the seven predictive functions (lines) for different vales of absolute pairwise correlations of informative genes (x-axis), pairwise correlations of non-informative genes (rows) and genes' variances (columns) at a fixed proportion of informative genes ($\pi = 1\%$), log_2 effect size ($\beta = 2$), proportion of events ($\tau = 75\%$) and sample size ($n = 200$).

Table 6.2: Structure of the performance data generated from evaluating the predictive functions on the simulated data

ID	Function	sampSize	propEV	variance	otherCor	propDE	log2FC	deCorr	IBS
1	Boost	200	0.75	0.5	0	5	2.0	0.00	0.1694
2	Boost	200	0.75	0.5	0	5	2.0	0.25	0.0795
3	Boost	200	0.75	0.5	0	5	2.0	0.50	0.0533
4	Boost	200	0.75	0.5	0	5	2.0	0.75	0.0417
5	PLR1	200	0.75	0.5	0	5	2.0	0.00	0.1479
6	PLR1	200	0.75	0.5	0	5	2.0	0.25	0.1786
⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋮	⋮
13603	superPC	100	0.25	2	0.5	1	0.5	0.50	0.1384
13604	superPC	100	0.25	2	0.5	1	0.5	0.75	0.1354
13605	uniCox	100	0.25	2	0.5	1	0.5	0.00	0.1380
13606	uniCox	100	0.25	2	0.5	1	0.5	0.25	0.1354
13607	uniCox	100	0.25	2	0.5	1	0.5	0.50	0.1403
13608	uniCox	100	0.25	2	0.5	1	0.5	0.75	0.1398

In order to determine which of the data characteristics associate to the performance of the functions, the simulation results were summarized to a data matrix as presented in Table 6.2 containing 13608 data points stemming from the 1944 simulation scenarios multiplied by the seven functions. Of the 1944 scenarios, univariate functions could not converged in three scenarios due to numerical problems. Thus, all the functions were not evaluated in these scenarios resulting to a total of 21 (3 scenarios by 7 functions) missing values. These scenarios correspond to data with large variances ($\lambda = 0.5$) and highly correlated variables (pairwise correlations within non-informative genes, $\theta = 0.50$ and pairwise correlations within informative genes, $\rho = 0.75$). As such, these missing values were omitted in the random effects analysis.

For each of the predictive variable, a linear random effects regression model was fitted on the transformed IBS as described in the method section. The individually explained null variance of the study factors are depicted on Figure 6.3. This figure shows that pairwise correlations within informative genes, the number (proportion) of informative genes and the pairwise correlations within non-informative are the leading factors explaining respectively 14.0%, 10.3% and 9.1% of the null variance. While sample size and the number (proportion) of events respectively explain 7.1% and 4.5% respectively. Genes' variances and log_2 effect sizes account for less than 1% each. For the multivariate linear random effects regression model, we started with a complex model of random intercepts and slopes and three ways interactions of the predictive factors. Starting with pairwise log_2 effect size because of its low individually explained null variance, we eliminated

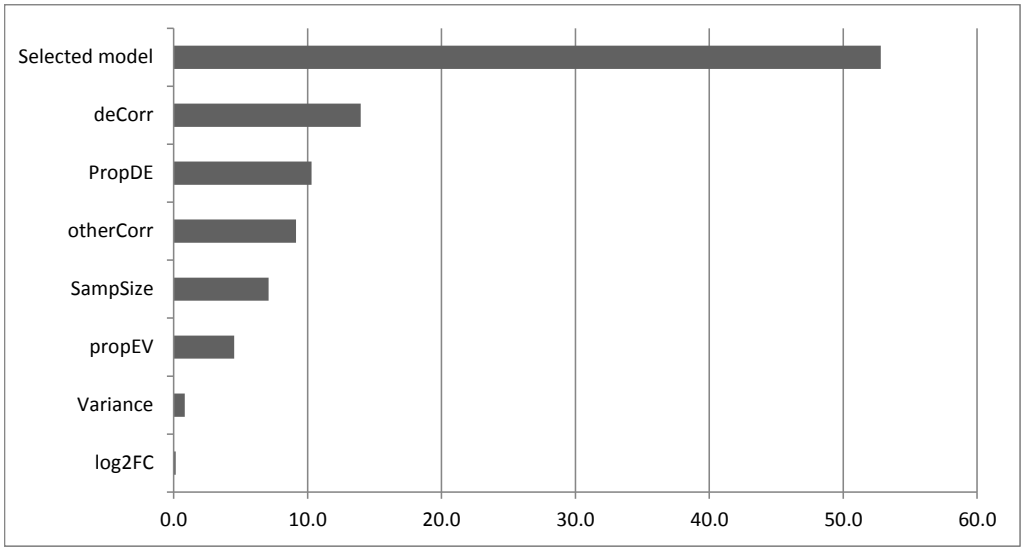

Figure 6.3: Proportion of the null variance explained by each and every studied factor. *The selected model refers to the predictive model presented on Table 6.3.*

variables using the log-likelihood ratio test. We ended up with the model presented on Table 6.3 consisting of the fixed effects of six predictive factors, random intercepts and slopes of the main effects of these factors. As observed univariately, log_2 effect size was eliminated from the multivariate model. This model explains approximately 53% of the null variance as illustrated on Figure 6.3. The left panel of Table 6.3 presents the estimates of fixed effects, the standard errors and the t statistics while the top-right panel presents the net effect of a standard deviation (SD) unit increase of a given factor conditional on common values of other factors. Finally, the bottom-right panel presents the performances of the predictive functions at different values of the data characteristics.

From the top-right panel of this table, one notices that a 1 SD unit increase in pairwise correlations of informative genes, corresponding to $\rho = 0.56$, will lead to an increase in the transformed integrated Brier score (trIBS) with the highest increase observed when other variables are at their highest values but might have a negative effect on the trIBS when other values are extremely low. A similar effect is observed for a 1 SD unit increase in the proportion of DE genes π. A SD unit increase in sample size on one hand will lead to an increase in the trIBS with maximum effect observed when other values are extremely large. On the other hand, a unit increase in the pairwise correlation of non-informative genes leads to a decrease in the trIBS with severe effects when other variables take extremely small or large values. The effects of the proportion of events and genes' variances depend on the values of other variables and could lead to an increase in the

Table 6.3: Fixed effects estimates (left panel) and their conditional on other factors net effects (right panel)

Fixed effects

Parameter	Estimate	Std. Error	t value
(Intercept)	2.031	0.048	42.476
StdDECorr ($\tilde{\rho}$)	0.127	0.011	11.896
StdPropDE ($\tilde{\pi}$)	0.085	0.030	2.818
StdOtherCorr ($\tilde{\theta}$)	0.012	0.039	0.306
StdSampSize ($\tilde{n}$)	0.081	0.019	4.303
StdPropEV ($\tilde{\tau}$)	0.018	0.024	0.734
StdVariance ($\tilde{\sigma}^2$)	0.010	0.013	0.764
StdDECorr2 ($\tilde{\rho}^2$)	-0.026	0.003	-10.174
StdOtherCorr2 ($\tilde{\theta}^2$)	-0.058	0.003	-19.744
StdPropEV2 ($\tilde{\tau}^2$)	0.059	0.003	20.146
StdDECorr*StdOtherCorr	-0.022	0.002	-10.803
StdDECorr*StdPropDE	0.056	0.002	27.194
StdOtherCorr*StdPropDE	0.008	0.002	3.700
StdDECorr*StdPropEV	0.043	0.002	20.736
StdOtherCorr*StdPropEV	-0.003	0.002	-1.282
StdPropDE*StdPropEV	0.020	0.002	9.863
StdDECorr*StdVariance	0.018	0.002	8.656
StdOtherCorr*StdSampSize	0.021	0.002	10.174
StdOtherCorr*StdVariance	-0.007	0.002	-3.418
StdSampSize*StdPropEV	0.017	0.002	8.374
StdDECorr*StdOtherCorr*StdPropDE	0.020	0.002	9.558
StdDECorr*StdOtherCorr*StdPropEV	-0.020	0.002	-9.729
StdDECorr*StdPropDE*StdPropEV	0.016	0.002	7.862
StdOtherCorr*StdPropDE*StdPropEV	-0.017	0.002	-8.342

Conditional net effects of study factors

1 SD increase net effect

Other Variables	$\varphi = \log\left(\frac{\tilde{\pi}(x_{[i]})}{(1-\tilde{\pi}(x_{[i]}))}\right)$	$\tilde{\rho}$	$\tilde{\pi}$	$\tilde{\theta}$	$\tilde{n}$	$\tilde{\tau}$	$\tilde{\sigma}^2$
2 SD		0.355	0.329	-0.120	0.157	0.147	0.032
1 SD		0.212	0.188	-0.066	0.119	0.133	0.021
0 SD		0.101	0.085	-0.046	0.081	0.077	0.010
-1 SD		0.022	0.020	-0.060	0.043	-0.021	-0.001
-2 SD		-0.025	-0.007	-0.108	0.005	-0.161	-0.012

1 SD corresponds to $\rho = 0.56$ $\theta = 0.41$ $\pi = 3.26$ $n = 100$ $\tau = 0.41$ $\sigma^2 = 1.25$

Other Variables	$\varphi = \log\left(\frac{\tilde{\pi}(x_{[i]})}{(1-\tilde{\pi}(x_{[i]}))}\right)$	Boost.	Lasso	Elas. Net	Ridge	RF	superPC	uniCox
2 SD		3.307	3.740	4.020	3.077	2.395	2.778	3.007
1 SD		2.573	2.810	2.957	2.427	1.958	2.286	2.405
0 SD		2.082	2.125	2.138	2.022	1.764	2.038	2.047
-1 SD		1.844	1.691	1.571	1.868	1.822	2.041	1.941
-2 SD		1.865	1.517	1.263	1.974	2.140	2.305	2.095

trIBS when other variables take large values or a decrease in trIBS otherwise. These results together with the individually explained null variance show that (i) the number of informative genes (ii) how correlated both informative and non-informative genes are and (iii) the overall sample size and proportion of events associate to the performance of Cox's predictive functions.

Lastly, the bottom-right panel of the table shows that all predictive functions will perform reasonably well if all data characteristics have large values; with elastic net, Lasso and Boosting being the top three. Nevertheless, if the data characteristics are extremely small, these functions turn to be among the least performing while supervised PC (superPC), Random forest (RF) and univariate Cox (uniCox) become the best performing functions. This indicates that the optimality of a function depends on the data characteristics. It should be noted however that the combination of all other variables simultaneously being at -2 or at +2, is highly unlikely.

6.4 Application

To evaluate the predictive ability of our presented random effects regression model on real-life data, two Affymetrix gene expression datasets were downloaded from ArrayExpress database [www.ebi.ac.uk/arrayexpress/, 12/08/2016] and preprocessed without filtering as previously described by Jong et al., (2014). These datasets are briefly described below:

- *E-GEOD-14814:* It has been shown that there is a significant survival benefit from adjuvant cisplatin/vinorelbine (ACT) in stage IB-II NSCLC, but stage IB patients did not derive significant benefit. In this study, microarray gene expression was used to identify a prognostic gene signature that is stage and histology independent and that may select early stage NSCLC patients who are most likely to have a survival benefit from adjuvant chemotherapy with ACT [Zhu et al., 2009]. All 90 samples were utilized with all-cause-mortality considered in this study.
- *E-GEOD-26549:* Patients with oral preneoplastic lesion (OPL) have high risk of developing oral cancer. Although certain risk factors such as smoking status and histology are known, the ability to predict oral cancer risk remains poor. The study objective was to determine the value of gene expression profiling in predicting oral cancer development. Gene expression profile was measured in 86 of 162 OPL patients who were enrolled in a clinical chemoprevention trial that used the incidence of oral cancer development as a prespecified endpoint. The median follow-up time was 6.08 years and 35 of the 86 patients developed oral cancer over the course [Saintigny et al., 2011].

Table 6.4: Characteristics of the 5 datasets used for evaluating the predictive model.

studyID	#genes	sampSize	propEV	propDE	deCorr	otherCorr	variance
E-GEOD-14814	22283	90	0.322	0.978	0.178	0.273	0.228
E-GEOD-26549	33297	86	0.407	7.037	0.158	0.212	0.140

studyID: ArrayExpress accessing ID and **#genes:** total number of probesets on the array
The third to the eighth columns correspond to the variables under study.

We quantified the data characteristics studied and presented on Table 6.4 as follows: (i) sampSize was computed by counting the samples in the study. (ii) propEV by counting the number of events divided by the sampSize. (iii) propDE was computed by first ranking the probesets based on their absolute coefficients from univariate Cox models [Cox, 1972] and then applying a cutoff on the absolute coefficients, the number of informative probesets was determined and the ratio to the total number of probesets was used as propDE. Different absolute log_2 effect size (coefficient) cutoffs were considered, starting from 2, 1.5, 1, 0.5 to 0.25 with the next value only considered if the total number of probesets retained after a preceding cutoff value is less than 5. In a situation where the last cutoff value does not yield up to five probesets, the 95% quantile of the absolute coefficients was considered as a cutoff value. (iv) variance was determined as the mean of the variances of all the probesets. (v) deCorr as the mean of absolute values of the elements of the upper- (lower-) triangular of the correlation matrix of the informative genes retained in (iii) above. (vi) otherCorr was computed as the standard deviation (SD) of the elements of the upper- (lower-) triangular of the correlation matrix of non-informative probesets. This matrix was computed from all non-informative probesets if they were less than 10000 or a sample of 10000 from these probesets otherwise.

These data characteristics were standardized using the mean and SD of the respective variables from the simulated data. The random effects model was used to predict the transformed integrated Brier scores (trIBS) for all predictive functions in each dataset (supplementary Figure S2). We then built and evaluated Cox's predictive models using the predictive functions using 10-folds cross-validation and a 5-fold inner cross-validation on the learning set to optimize parameter(s) of each function. The average trIBS over the 10-folds were computed and considered as the observed trIBS. To compare the predicted to observed trIBS, we used the ranked base Spearman correlation between the predicted and observed trIBS. The results of this comparison for each dataset are presented on Figure 6.4. Under normal circumstances, small positive and negative correlations around zero can be observed between the predicted and observed trIBS by chance. Nevertheless, the high positive correlation values on this figure indicate agreement

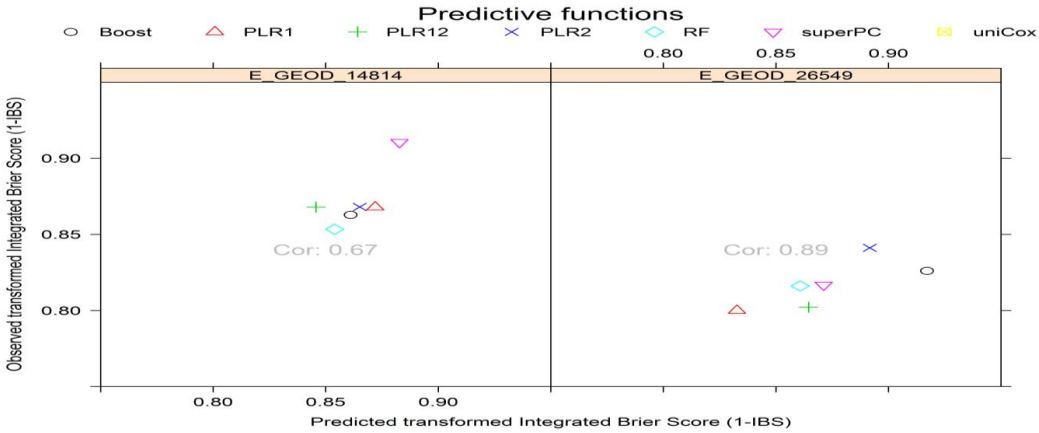

Figure 6.4: Predicted vs Expected (Observed) transformed integrated Brier scores (trIBS). *Cor represents Spearman correlations between the predicted and observed trIBS. Note that "_" corresponds to "–" in the studyID*

between our predicted and observed trIBS and hence the reliability of our model in predicting optimal functions(s) for a given real-life data.

Although the correlations are positive; confirming the ordering of all the functions, we are in general interested in just the ordering of the first or first and second functions determined as optimal on a given dataset. As shown on E-GEOD-14814, the predicted optimal function (superPC) was also observed as the optimal function on this dataset. Boost and Ridge (PLR2) which were respectively predicted optimal and sub-optimal functions on E-GEOD-26549 were indeed observed as both optimal and sub-optimal but with an interchange in ranks. Note that though uniCox had predicted value for these datasets as illustrated on supplementary Figure S2, this function did converge on neither dataset. As was observed in some of the simulation scenarios, univariate functions (uniCox and superPC) might not converge under certain conditions in real-life data. As such, we recommend that if any of these functions is predicted as an optimal function, it should be considered only when it converges, otherwise a sub-optimal function could be used. That notwithstanding, these functions could indeed be the true optimal function when it converges as was the case for superPC on E-GEOD-14814.

6.5 Discussion

We hypothesized that the performance of Cox's predictive functions on gene expression data depends on overall sample size, proportion of events, proportion of informative genes, genes' variances, log_2 effect sizes of informative genes and magnitude of the pairwise correlation within and between informative and non-informative genes. We empirically showed that except for the log_2 effect sizes, all hypothesized data characteristics associate to the integrated Brier scores of

seven often used and clinically relevant Cox's predictive functions. Additionally, we have provided a predictive model to determine an optimal Cox's predictive function for a given gene expression data using the associated data characteristics. An application of the predictive model on two real-life gene expression datasets showed high correlations between predicted and observed performance of optimal function(s) and was able to rule out least optimal function(s) on these datasets.

Microarray gene expression profiling in particular has become a widely used tool for survival predictions [Bøvelstad et al., 2007]. Cox's predictive functions have been shown to perform differently across gene expression datasets [Bøvelstad et al., 2007; van Wieringen et la., 2009] and data characteristics have been shown to differ across gene expression datasets [Jong et al., 2014; Novianti et al., 2015]. While sufficient knowledge is available on the properties of most predictive functions, little is known about data characteristics that associate to the performance of Cox's predictive functions and how to use these characteristics to choose an optimal predictive function for a specific dataset. Since gene expression data often suffers from the curse of dimensionality, prediction with such data utilizes cross-validation (CV) to estimate the error rate that is generalizable to independent data. Nevertheless, since there is no universally best machine learning algorithm, it is common practice to compare several algorithms and report the algorithm that produces the smallest cross-validation error. This approach leads to selection bias because there is a high probability of a function to have the smallest error simply by chance. This probability increases with both the number of algorithms compared and a decrease in sample size. Thus, selection bias is often high in data with small sample sizes and when several algorithms are compared [Ding et al. 2014].

Ding et al., (2014) stated as an example that one uses a small pilot data, compares several machine learning methods and selects the minimum error predictor (MEP) with a falsely small error because of the selection bias. When the model proceeds to a large cohort validation for translational research, it will likely fail. As such, several bias correction methods have been proposed in the literature of class prediction with gene expression data which could easily be adapted for survival prediction. Nevertheless, no such method is a hundred percent effective when several least optimal algorithms are compared on a dataset with a small sample size, as observed by these authors. Another school of thought prefers super-learners which combine the scores of several predictive functions to improve overall prediction. Nevertheless, super-learners are less acceptable in medical applications because of their less interpretability [Simon, 2014].

Additionally, super-learners often utilize the entire genome instead of a selected profile, making practical application time consuming and costly. Thus, traditional predictive functions that can yield interpretable models are still preferred over super-learners.

In this study, we outlined data characteristics together with clinically relevant and often used Cox's predictive functions and investigated their effects on the predictive performance of these functions using simulation studies. In simulating gene expression data, we utilized the approach proposed by Jong et al., (2016). Based on the simulation results, we have provided a guide for choosing an optimal Cox's predictive function for a specific dataset using the associated data's characteristics and the studied functions through a linear random effects predictive model. As a meta-model one would expect it to explain close to 100% of the variance in the simulated data but our predictive model accounts for approximately 53% of the variability in the simulated data. The remaining unexplained variance may be associated to sampling variability stemming from the several (108) random covariance matrices used to generate both learning and test sets as well as the different learning and test sets generated at each iteration. Additionally, a linear model might not be a best fit for the data and non-linear models could be worth considering.

Although we used different Cox's predictive functions and evaluated these functions using integrated Brier scores, our simulation results show that if there are highly correlated genes not associated to the survival time (non-formative genes), most predictive functions will perform poorly. This can be expected in a disease where by many genes in the same pathway(s) are not associated to the survival time but are expressed in similar manner. On the other hand, predictive functions will have better performance with increase in the overall sample size. This is contrary to the believe that the effective sample size in classical Cox's regression is the number of events. For other variables, their effects depends on the interactions between these variables and the predictive functions which was modeled via the random effects linear regression. Our simulation results also confirm the results of Bøvelstad et al., (2007) and van Wieringen et la., (2009) that the performance of a predictive function depends on the dataset in question as shown by the fact that the optimality of a given function is scenario dependent.

Additionally, we have provided a predictive model that can serve as a guide to choose a Cox's predictive function for a given gene expression dataset and its application on two real-life datasets showed high correlations between the predicted and the observed ranks of these functions. Though the correlations are not significantly high, the ranks of the first two optimal functions were maintained, thereby satisfying the objective of our model in determining the optimal function(s)

on a given data. We think that the interchange between the top two functions on the E-GEOD-26594 could be due to the small number of cross-validations (10-folds cross-validation). A large number of Bootstrap samples could yield expected trIBS similar to predicted trIBS. More so, the predicted trIBS were computed based on the entire datasets but the training of the predictive functions was based on a subset of the data. Considering the small sample sizes, there could be huge differences between the estimated characteristics of the entire data and the training data thus resulting to the variability in the predicted and observed performance. That stated, we must note that our predictive model predicts the expected performance on an independent dataset and not the performance on a cross-validated dataset, which could be slightly different from the actual expected value. The drawback of our model is that irrespective of the dataset at hand, it will predict trIBS for all the seven functions but some of these functions might not converge on such data. As such, on datasets where non-convergence is a problem for a predicted optimal function, the sub-optimal function in line could be considered.

In this study, we developed an approach that is aiming to be free of biases, to choose an algorithm that is optimal for a given dataset. For a given dataset, our approach quantifies data characteristics, uses a model to predict which algorithm will be optimal for this data and then uses that algorithm for prediction. One will immediately think of two sources of biases in our approach namely, selection bias and optimistic bias.

(i) Selection bias, because through our prediction model we are implicitly comparing algorithms using their predicted errors and choosing the one with the smallest error. Nevertheless, our predicted errors are expected values based on large independent datasets. Hence, cannot occur by chance as CV errors. Theoretical properties of bias show that for a fixed number of algorithms, bias approaches zero as sample size grows large enough [Ding et al., 2014]. Hence, our approach is free of selection bias since the test sets sample sizes in our simulations were large (100 samples by 100 iterations).

(ii) Optimistic bias, because using the data to determine an optimal algorithm introduces bias in the assessment of predictive performance resulting from overfitting. Nevertheless, we believe our approach is less prone to optimistic bias because: (a) most important variables (e.g. correlations) in our model do not utilize class labels and those utilizing the class labels for their estimations are summarized single values; an approach which is completely different from resubstitution that account for most if not all of optimistic bias. (b) though used for algorithm selection, the samples have not been seen by any such algorithm.

Hence, the hypothesis chosen from the hypotheses class by the selected algorithm is entirely based on the training set making the test set an independent set to such a hypothesis.

That notwithstanding, should there be any optimistic bias from our approach, it happens only on the experimental data and not on the mandatory independent validation data often required nowadays. More so, such optimistic bias on a generalizable algorithm is preferable to a less generalizable algorithm falsely chosen due to selection bias. Generalizability is plausible because the generalized performance of an algorithm (hypotheses class) depends on its interaction with the data characteristics (assumed data distribution) and since the pilot study is often a random sample from the population, the estimates of the data characteristics should be close to expected values in the population. Hence, our approach comes handy in that it allows one to use exclusively (without selection bias) the data characteristics from the pilot study to select an algorithm that can be generalized to larger cohorts.

Acknowledgements

This work has been supported by the VIRGO consortium, which is funded by the Netherlands Genomics Initiative and by the Dutch Government (FES0908). The funding agencies in no way influenced the outcome or conclusions of the study. The authors would like to thank the VIRGO consortium and its sponsors, the Netherlands Genomics Initiative and the Dutch Government for the financial support. Next to this, we thank Martin Marinus and the HPC-team at UMC Utrecht for the high performing computing facilities.

Supplementary Material

Supplementary material for this manuscript is available upon request to the authors and include:

Figure S1: Average integrated Brier scores of the seven predictive functions for: **(A)** sample size of 200, 75% proportion of events, 3% of informative genes and log_2 effect size of 2; **(B)** sample size of 200, 75% proportion of events, 3% of informative genes and log_2 effect size of 0.5; **(C)** sample size of 200, 25% proportion of events, 3% of informative genes and log_2 effect size of 2; **(D)** sample size of 100, 75% proportion of events, 3% of informative genes and effect size of 2;

Figure S2: Predicted transformed integrated Brier scores (trIBS) on the real-life dataset by our linear random effects regression model using the characteristics shown on Table 6.4.

Source code has been incorporated in the R package `SPreFuGED`

PART III

IDENTIFICATION AND VALIDATION OF PREDICTIVE BIOMARKERS.

Chapter 7

Transcriptome assists prognosis of disease severity in respiratory syncytial virus infected infants.

Victor L. Jong, Inge M. L. Ahout, Henk-Jan van den Ham, Jop Jans, Fatiha Zaaraoui-Boutahar, Aldert Zomer, Elles Simonetti, Maarten A. Bijl, H. Kim Brand, Wilfred F.J. van IJcken, Marien I. de Jonge, Pieter L. Fraaij, Ronald de Groot, Albert D. M. E. Osterhaus, Marinus J. C. Eijkemans, Gerben Ferwerda & Arno C. Andeweg

Nature Scientific Reports 2016; **6**:36603

Abstract

Motivation: Respiratory syncytial virus (RSV) causes infections that range from common cold to severe lower respiratory tract infection requiring high-level medical care. Prediction of the course of disease in individual patients remains challenging at the first visit to the pediatric wards and RSV infections may rapidly progress to severe disease. In this study we investigate whether there exists a genomic signature that can accurately predict the course of RSV.

Results: We used early blood microarray transcriptome profiles from 39 hospitalized infants that were followed until recovery and of which the level of disease severity was determined retrospectively. Applying support vector machine learning on age by sex standardized transcriptomic data, an 84 gene signature was identified that discriminated hospitalized infants with eventually less severe RSV infection from infants that suffered from most severe RSV disease. This signature yielded an area under the receiver operating characteristic curve (AUC) of 0.966 using leave-one-out cross-validation on the experimental data and an AUC of 0.858 on an independent validation cohort consisting of 53 infants. A combination of the gene signature with age and sex yielded an AUC of 0.971.

Conclusion: The presented signature may serve as the basis to develop a prognostic test to support clinical management of RSV patients.

7.1 Introduction

Respiratory syncytial virus (RSV) causes infections that range from common cold to severe lower respiratory tract infection that in some instances may have a fatal outcome. Especially infants, elderly and patients with underlying chronic disorders suffer from severe RSV infections [Simoes, 1999; Falsey et al, 2014]. In infants, RSV is the leading cause of lower respiratory tract infections (LRTI) and is responsible for 80% of acute bronchiolitis cases [Bush & Thomson, 2007]. RSV infections pose a huge burden on society in terms of disease, logistics and socio-economic sequelae. There is an unmet need for an RSV vaccine, despite considerable research efforts no licensed vaccine has been developed.

In industrialized countries, 1-5% of infants with RSV infection are hospitalized [Nair et al., 2013; Hall et al., 2009; Stockman et al., 2012; Jain et al., 2015]. Some of these infants yet suffer from severe disease upon admittance, while others are admitted without severe symptoms since the course of bronchiolitis is highly variable and the need for supportive care cannot be predicted [Adams & Doull, 2009; Meissner, 2016]. Several risk factors for developing severe RSV disease in infants have been identified, including preterm birth, young age, sex and environmental factors like in-house smoking [Tregoning & Schwarze, 2010]. Notwithstanding these known risk factors, current medical practice does not allow accurate prediction of whether an infant will further progress to severe RSV disease or not and could even be sent home safely. Genomic technologies have contributed to study the virus-host interaction, including virus discovery, pathogenesis studies, the design of antiviral strategies and identification of biomarkers to support clinical management of infectious diseases [Pulendran et al., 2013; Sekaly & Pulendran, 2011; van de Weg et al., 2015; Zhai et al., 2015]. For RSV infections, this has supported the characterization of vaccine-induced skewed host responses upon infection [Schuurhof et al., 2010; van Diepen et al., 2015]. Meijas et al., (2013) recently used blood transcriptome profiles obtained within 3 days of hospitalization to characterize the host response to RSV infection in infants compared with rhinovirus or influenza infections and identified transcriptional profiles that associate with RSV disease severity. However, a prognostic model for RSV severity based on gene expression profiles collected at admittance to the hospital has not been developed.

In this study we aim to identify and validate a gene signature that discriminates severe from less severe RSV LRTI that do not require advanced support. Such a signature together with other clinical parameters may improve the prognosis of less severe patients that could be safely sent home.

7.2 Material and Methods

7.2.1 Study design

Study subjects were recruited at Canisius Wilhelmina Hospital, Radboud University Medical Center, Nijmegen, and Erasmus Medical Center, Rotterdam, The Netherlands. Nasopharyngeal wash and blood samples were prospectively obtained from patients less than 2 years of age with a viral bronchiolitis. Patient enrolment occurred 7 days a week and samples were taken within 24 hours after first contact with the hospital. Seventy-three percent of all eligible bronchiolitis patients agreed to participate in the study. The major reasons for non-inclusion were parental availability to sign consent and the hesitancy for the venipuncture. Only patients with an RSV infection, as retrospectively determined by PCR were included in the study. Exclusion criteria were: immunodeficiency, systemic steroid treatment in the previous 2 weeks, blood transfusion, congenital heart and chronic lung disease. A Tempus tube (Tempus^{TM}, Applied Biosystems, Austria) and sodium heparin tube were filled with 3ml of blood. Medical history, demographic and clinical data were collected from medical records and questionnaires. The (hospitalized) patients were followed until recovery and were retrospectively classified as: mild for children without hypoxia, moderate for patients requiring supplemental oxygen (oxygen saturations <90%, ≥10 minutes) and severe for children requiring mechanical ventilation due to apnea, exhaustion and/or respiratory failure. Recovery samples were obtained after 4-6 weeks, during home visits. Blood samples were obtained from healthy controls (HC) without underlying diseases or medication subjected to elective surgery.

7.2.2 Study approval

The study protocol was approved by the Regional Committees on Research involving Human Subjects of Arnhem-Nijmegen and Rotterdam and were conducted in accordance with the principles of the Declaration of Helsinki. Written informed consent was obtained from the parents of all children prior to inclusion in the study.

7.2.3 Sample processing and blood transcriptome profiling

Inclusion of patients and sample collection was performed by a single MD at the hospitals. Multiplex RT-PCR was performed to test the nasopharyngeal washes on 15 different viral pathogens, as previously described [Templeton et al., 2004]. Blood was collected in Tempus tubes for immediate stabilization of RNA and subsequently stored at -80°C. Total RNA was isolated from each blood sample, processed, assessed, labelled and hybridized to a single Affymetrix Human

Genome U133 plus 2 gene chips; and image analysis was performed in the same lab and by one technician as described in supplementary material. The raw data has been deposited in the ArrayExpress database under access number E-MTAB-5195.

7.2.4 Data preprocessing

Microarray data was preprocessed using `R 3.1.2` [R core Team, 2014] and `Bioconductor` [Gentleman et al, 2004]. Upon initial quality control and VSN normalization (to render samples comparable), probeset (a combination of multiple probes) summarization was performed by median polish [Gautier et al., 2004; Huber et al., 2002]. Unless otherwise stated, all probesets/genes present on the Affymetrix GeneChip were used for data analysis. Samples were labelled and hybridized in two batches which did not correspond to any biological variable as samples were randomly assigned to the batches. The normalized expression values were adjusted for a batch effect (see supplementary Figure S1) using "ComBat" [Johnson et al., 2007]. Additionally, we assessed confounding effects of clinical parameters age and sex on gene expression–severity relationship using "biasograms" [Krzystanek et al., 2013].

7.2.5 Differential expression analysis

To obtain a global view of the blood transcriptome changes in response to RSV infection (i.e. to evaluate whether whole transcriptome changes associate with severity), a principal component analysis (PCA) as an exploratory analysis was performed on the age by sex standardized data. Next, a differential expression (DE) analysis was performed on the normalized-batch-adjusted data controlling for an age by sex effect using empirical Bayes linear models [Symth, 2004] implemented in the R package `limma` [Ritchie et al., 2015]. Details of the models are found in supplementary material. We controlled for multiple testing via false discovery rate (FDR) using a Benjamini and Hochberg procedure [Benjamini & Hochberg, 1995]. Gene set enrichment analysis was performed using Ingenuity pathway analysis (IPA, www.qiagen.com/ingenuity).

7.2.6 Identification and evaluation of prognostic biomarkers

Since we are interested in identifying RSV-infected infants that will progress to severe stage upon presentation to the hospital, we grouped mild and moderate samples and aimed to separate these samples from infants that were presented with or progressed to severe disease after hospitalization. We chose to utilized probabilistic predictors (to predict the chance of an RSV-infected infant to be severe) because in clinical applications, probabilities are more informative than absolute yes or no predictions [Pepe, 2005]. Several probabilistic predictors exist in the

literature and their performance depends on the type of the data they are being applied on [Kim & Simon, 2011]. Using results of Jong et al., (2014); Kim & Simon, (2011) and observed correlations in the data, three probabilistic classification functions that could be optimal for this data were chosen as described in supplementary material. These functions were support vector machines (SVM) [Schölkopf & Smola, 2001], shrunken centroids discriminant analysis (SCDA) [Tibshirani et al., 2002] and random forest (RF) [Breiman, 2001].

For each classification function, the experimental data was split into a learning set and a test set using leave-one-out cross-validation (LOOCV). Cross-validation reduces optimistic bias by ensuring that our models are evaluated on an independent dataset that was not used to constructed these models. Most probabilistic classification functions require hyper-parameters to perform variable selection among the huge number of variables (probesets). Usually, the best values for these hyper-parameters are also determined by cross-validation. Thus, the parameter(s) of the function were optimized using an inner loop of five-fold cross-validation on the learning set. Next, a prognostic model was built with the optimal parameter(s) on the entire learning set and evaluated with the test set, as described in supplementary material. The following R packages; `CMA` [Slawski et al., 2008], `e1071` [Meye et al., 2014], `pamr` [Hastie et al., 2014] and `randomForest` [Liaw & Wiener, 2002] were utilized for class prediction. The best calibrated and refined function amongst the three functions was selected and its performance evaluated using the area under the receiver operating characteristic (ROC) curve (AUC). Finally, the transcripts that maximized the binomial log-likelihood function, with the leave-one-out cross-validated data were retained as a gene signature from the selected function as described in supplementary material.

7.2.7 Comparison of biomarkers to clinical parameters.

Age and sex are readily available clinical parameters that have been determined to be associated to RSV disease severity [Berger et al., 2009]. To assess the gain attained with a genomic model over a model with these clinical parameters, and the effect of standardization, the leave-one-out cross-validated predicted probabilities of progressing to severe for all samples were transformed to genomic scores (a genomic score is single measure of the genome of a sample as predicted by a model) for models with unstandardized and standardized data. Logistic regression models (see supplementary material) were then fitted with the genomic scores and/or clinical parameters and their AUCs compared.

7.2.8 Validation of biomarkers

For an independent validation, a subset of the Illumina RSV data of Meijas et al., (2013) was used. Since the experimental data and validation data were obtained using different platforms, we linked the data using gene symbols and applied cross-platform transformation (to render gene expression comparable across datasets) as described in supplementary material. The transformed data was supplied to our prognostic model for predictions of probabilities of severity. For a confirmatory analysis of how well our prognostic model performs, we built and evaluated a prediction model with the chosen function (same function used to build our prognostic model) on the entire Illumina data and compared our validation performance to the performance from this (unrestricted) data.

7.3 Results

7.3.1 Study subjects and sampling

Thirty-nine infants hospitalized with acute RSV bronchiolitis were included in the study. Nasopharyngeal wash and whole blood samples for mRNA profiling were collected within 24 hours upon hospital admittance. Table 7.1 presents the characteristics of the study subjects. As expected, patients with the most severe course of RSV bronchiolitis were significantly younger than those with a relative mild or moderate course of this disease. The variables related to disease severity; duration of oxygen, and length of stay in the hospital were highest in the severe group, with ventilation indicating the method by which oxygen was supplied. The proportion of co-infections was lower in severe patients as compared to the other severity categories. There were

Table 7.1: Patient characteristics (n represents the number of samples per group)

Parameters	Mild (n=7)	Moderate (n=14)	Severe (n=18)
Age (days)	153 [84, 291]	185 [60, 333]	31 [17, 76]
Gestational age (weeks)	40 [29, 41]	40 [37, 41]	39 [37, 40]
Birth weight (kg)	3.5 [3.0, 4.2]	3.4 [3.1, 3.9]	3.3 [2.5, 4.0]
Symptomatic days	4 [2, 6]	4 [3, 6]	3 [2, 4]
Duration on O_2 (days)	0	3 [2, 5]	8 [7, 11]
Ventilation	None	Supplemental	Mechanical
Length of stay (days)	4 [2, 6]	5 [3, 8]	11 [9, 13]
Breastfeeding	4 (57)	11 (79)	12 (67)
Male gender	5 (71)	10 (71)	12 (67)
RSV + other virus(es)	4 (57)	8 (57)	3 (17)

Data are presented as median and interquartile range (IQR) in square brackets [,] or number and percentage in brackets ().The median age of the healthy controls was 536 days (IQR [472, 602]).

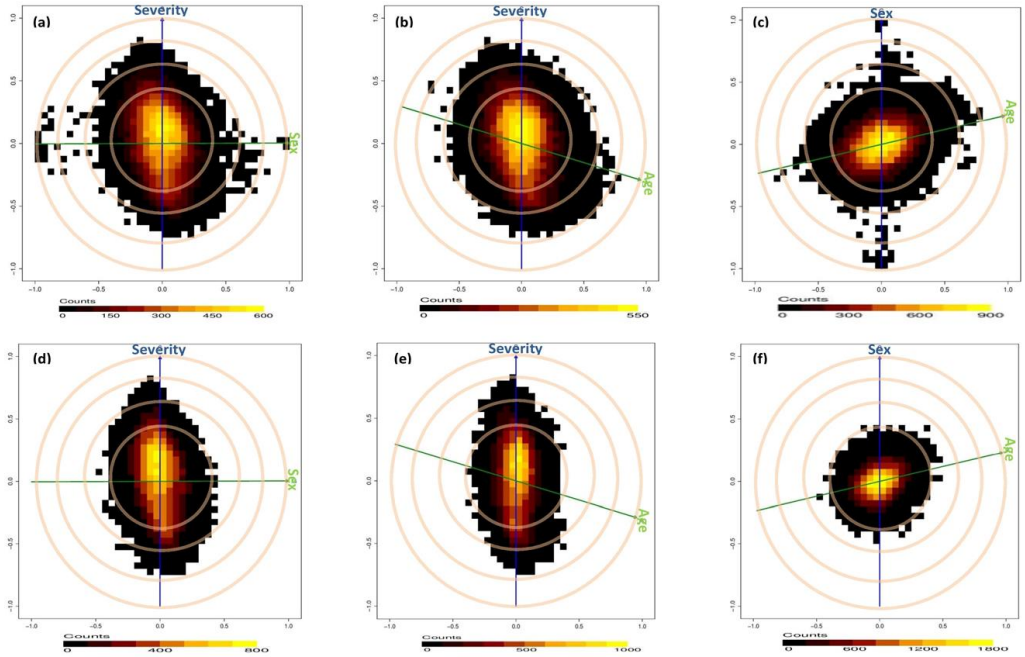

Figure 7.1: Confounding effect of Sex, Age and Age by Sex on gene expression–severity relationship, before: (a, b , c) and after: (d, e , f) an age by sex standardization. *The blue and green lines represent the clinical variables, the cosine of the angle between the lines represents its correlation to the blue line (Sex is not correlated to Severity, Age is negatively correlated to Severity i.e. younger kids become severe and Age is positively correlated to sex i.e. females are older). The cloud of points represent the transcripts and their correlations to both variables with most transcripts uncorrelated to the variables (yellow cloud) while a considerable number (black cloud) are correlated to Severity, Sex, Age and Age*Sex. The associations between the transcripts and Sex, Age or Age*Sex are significantly eliminated after standardization while retaining that of Severity*

no differences in the occurrence of other known risk factors.

7.3.2 Age and sex as confounders of gene expression–severity relationship.

Figure 7.1a and 1b respectively illustrate the confounding effects of sex and age on the gene expression-disease severity relationship. These figures show that whereas age is negatively correlated to severity, sex is uncorrelated to severity. Nevertheless, the high positive/negative correlations of a considerable number of transcripts to sex and age, as well as severity, indicate a confounding effect of these variables on the expression-severity relationship of these transcripts, thus warrant adjustment. Figure 7.1d and 1e illustrate the “biasograms” after an age by sex standardization. These figures show that standardization has no effect on severity correlated transcripts but as expected, transcripts that were originally correlated to age and sex become uncorrelated. A positive correlation of age to sex which signifies an age by sex interaction as a

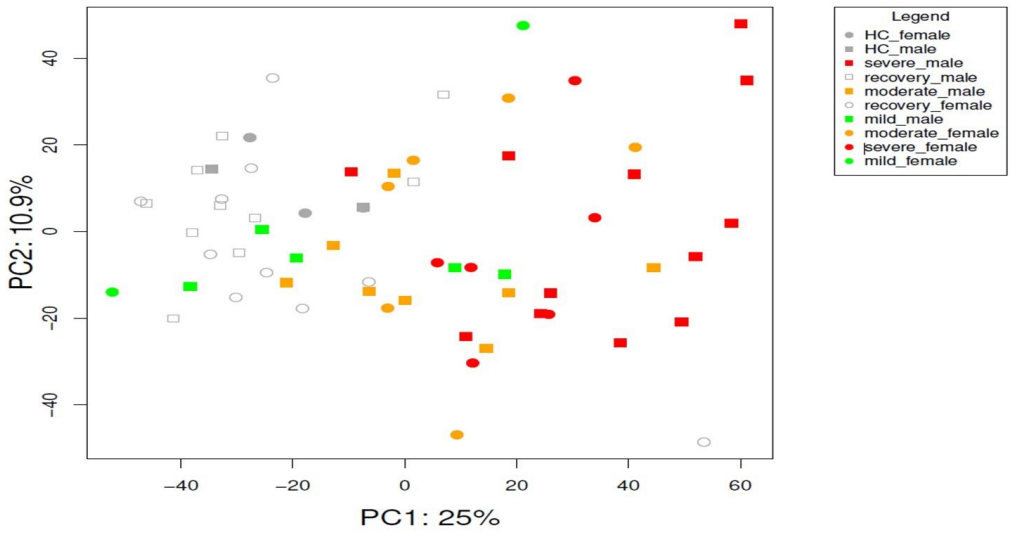

Figure 7.2: Global blood transcriptome profiling with principal component analysis: the first principal component (PC1) accounts for 25% of the variance in the dataset and associates with disease severity. *This can be observed as a shift from healthy controls and recovery cases (left) through mild and moderate to severe cases (right), with considerable overlap.*

potential confounder on the gene expression-severity relationship was also observed (Figure 7.1c) and eliminated after standardization (Figure 7.1f).

7.3.3 Global blood transcriptome profiles associate with RSV disease severity

Figure 7.2 illustrate a PCA on the whole transcriptome and the first principal component accounts for 25% of the variance in the transcriptomes and associates with disease severity. Transcriptome profiles of HC and recovery samples group together on the first principal component and are located opposite to profiles of severe infants. The distinct groups do not form discrete clusters in the PCA but gradually shift from mild through moderate to severe, with considerable overlap. This shows that the blood mRNA profiles substantially capture the severity of lower respiratory tract RSV infection.

7.3.4 Number of differential gene expression relates to RSV disease severity

Table 7.2 presents results of differential gene expression analysis and reveals that the number of DE transcripts increases with disease severity. No DE transcript was identified between mild versus HC samples when applying a FDR of 5% and absolute fold change (FC) threshold of 2. However, 17 and 221 transcripts were DE between moderate and severe versus HC respectively. Interestingly, all transcripts that are DE in moderate class are also DE in severe class with larger FC. About 90%

Table 7.2: Number of differentially expressed transcripts for each contrast at FDR of 5% and absolute fold change cutoff of 2

	Mil-HC	Mod-HC	Sev-HC	Mod-Mil	Sev-Mil	Sev-Mod	RC-HC	Sev-(Mil+Mod)/2
UP	0	15	194	0	164	42	0	82
Down	0	2	27	0	14	7	1	13
Total	0	17	221	0	178	49	1	95

Where Mil: Mild, Mod: Moderate, Sev: Severe, HC: Healthy controls (<2years) and RC: Recovery samples

of these DE transcripts are up-regulated. Comparison of HC with recovery samples revealed a single down-regulated transcript while moderate versus mild yielded no DE transcript, severe versus mild or moderate yielded 178 and 49 DE transcripts respectively. Lastly, 95 transcripts were DE between severe versus combined mild/moderate samples.

7.3.5 RSV induced blood transcriptome profiles reveal an inflammatory response

Figure 7.3 shows that multiple relevant categories of molecular and cellular functions are significantly enriched when comparing severe to HC samples. With "Cell-to-Cell Signaling and Interaction" top category, gene sets related to activation of several types of immune cells

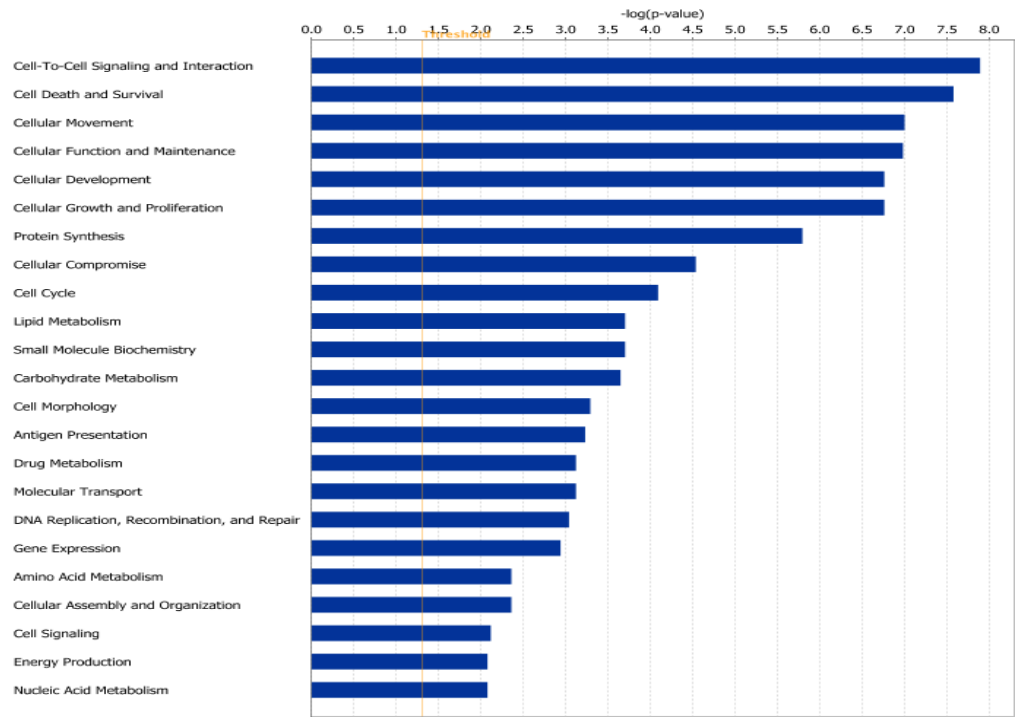

Figure 7.3: Ingenuity pathway analysis (IPA) Molecular and Cellular functions gene set analysis for severe vs healthy control contrast

including lymphocytes, granulocytes and specifically neutrophils are most significantly enriched. In addition, gene sets that are involved in migration and tissue infiltration of these same activated cell types are most significantly enriched within the category “Cellular Movement” that ranks third on this figure. Finally, several high ranking molecular and cellular function categories and their underlying gene sets indicate the immune cells involved are strongly proliferating. A list of genes involved in each of these pathways is presented in supplementary Table S1. Taken together, blood transcriptome changes in RSV disease reveal a typical inflammatory response to a viral infection.

7.3.6 Early blood transcriptome changes to predict a severe outcome of RSV infection.

To construct a predictive model, we combined mild and moderate cases as a single group and three probabilistic classification functions were chosen based on supplementary Figure S2 and results of Jong et al., (2014); Kim & Simon, (2011). Using these functions, classifiers were built and evaluated using LOOCV on the experimental data. SVM was chosen as the best calibrated and refined as shown on supplementary Figure S3 and henceforth considered for all analyses. The

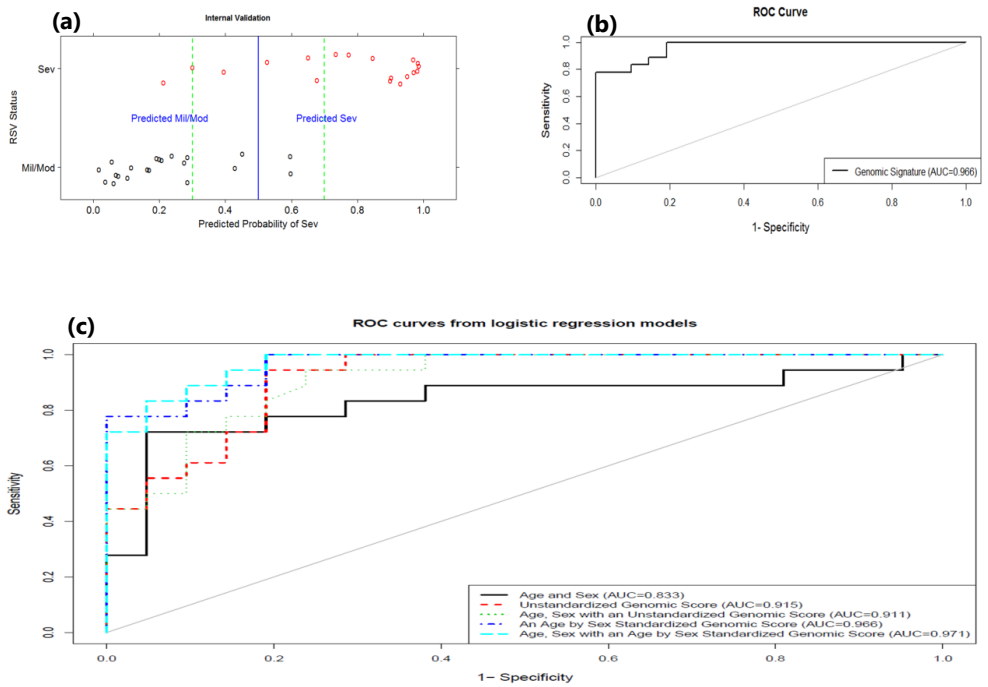

Figure 7.4: Internal validation of gene signature: ***(a)*** *samples’ predicted probabilities of being severe.* ***(b)*** *shows the ROC curve and the AUC for predicted probabilities. The AUC value of approximately 1 indicates how accurate our signature performs on this internal validation set.* ***(c)*** *shows that a genomic model from the age by sex standardized data out performs that from the unstandardized data. In addition, there is a significant difference between a model with clinical parameters and that with a genomic score and just a slight improvement when both parameters are included in a model.*

LOOCV predicted probabilities from SVM were used to evaluate its performance and are plotted on Figure 7.4a against the true RSV status as retrospectively determined. This figure shows that 5 samples out of 39 were misclassified at a 50% cutoff and when applying a proposed uncertainty band of 30-70% just one false negative is witnessed. Evaluation of the clinical characteristics of the single false negative patient as well as those patients with uncertain predictions (plotted within the proposed uncertainty band) did not reveal any recognizable pattern. The false negative patient had uniquely RSV and only a single patient plotted within the uncertainty band had RSV + other virus(es). Figure 7.4b presents the corresponding ROC curve from the LOOCV predicted probabilities and AUC of 0.966 demonstrates the high discriminative power of our prognostic model.

7.3.7 Genomic biomarkers outperform clinical parameters age and sex

Figure 7.4c presents the ROC curves from logistic models of clinical parameters and/or genomic scores. The genomic score model from the standardized data (AUC=0.966) outperforms that from the unstandardized data (AUC=0.915). In addition, there is a significant difference between a model with age and sex only (AUC=0.833) and that with a standardized (AUC=0.966) or unstandardized genomic score (AUC=0.915). Whereas there is a slight improvement from clinical parameters and standardized genomic score model (AUC=0.971), there is no improvement from the clinical parameters and unstandardized genomic score (AUC=0.911), indicating that indeed standardization completely removed an age-by-sex effect on the gene expression data.

7.3.8 An 84 gene expression signature predicts absence of disease progression

To extract the prognostic signature, we selected top transcripts maximizing the binomial log-likelihood function using LOOCV predicted probabilities as illustrated on supplementary Figure S4. This figure depicts a 1-SE maximum of 95 transcripts corresponding to 84 unique genes, which are displayed on Table 7.3. Of the 95 transcripts constituting the prognostic signature, 81 (85.26%) were found to be significantly DE between severe and non-severe patients (FDR cutoff of 5%, supplementary Table S2). The inclusion of non-DE transcripts in the classification model is expected, since not only DE genes are instrumental in class discrimination as illustrated by the two-dimensional scenario in supplementary Figure S4.

Table 7.3: Gene signature of 95 transcripts (probesets) corresponding to 84 genes.

Nº	Affymetrix IDs	Gene Symbol	Illumina IDs
1	1553605_a_at	ABCA13	0pf.QRwb4sWsn7jQKQ
2	238439_at 239196_at	ANKRD22	oXed_glbkhVEMCDCTo xnW5awXa6P_EblCH0s HIf_t7qU.BejlFP86o
3	209369_at	ANXA3	orojh4FCCMN1ArUY6k
4	205678_at	AP3B2	9SneFS4fUDnuF7_nVM
5	206632_s_at	APOBEC3B	9Ff619hpIlXH7nOK6I
6	206177_s_at	ARG1	0lVMTIuod_g0ohAnFE
7	232197_x_at	ARSB	3IDh9IVCveAqVdCl2o cMniVI7195NdYNUqeo BJIKm4IFqJ01E3eHp4 0u0lM7OqfU.quOIC94
8	201242_s_at	ATP1B1	unu3iN6N5U0f6cuEqc 9pU2OJP.BMdTLf9e64 fvOfs_Dnt_E3fl_11o
9	214575_s_at	AZU1	QBuHuFVihS7ZVBBu5E
10	221530_s_at	BHLHE41	EUN35CXUp3p.txedvU
11	205557_at	BPI	EqRaeX9VKg13UjfBQg
12	210244_at	CAMP	rItcu7lcV6dKfep3iA
13	209498_at 211889_x_at	CEACAM1	KpCYUm43RIkQFdLdQU fr.cFeEaHuTH9XTU54 xKXCKf3Uoo8n6L3cx0
14	203757_s_at 211657_at	CEACAM6	9criZPiUuucDnEx1B4
15	206676_at	CEACAM8	ubtBVA.8bOI1.LUVDo
16	209395_at 209396_s_at	CHI3L1	Zn_tzf4mVHIVHodF1U
17	208168_s_at	CHIT1	NA
18	220496_at	CLEC1B	uSC2Dit6C6ijp2vN0g
19	208146_s_at	CPVL	cdfjqToegKNUoMQgpI ZeEuC6evE1pCauH1NI
20	210262_at	CRISP2	ukOTfdVSNNEgNH9KQ4
21	207802_at	CRISP3	uqDXiAQ0SFE4hwohO4
22	205653_at	CTSG	0RKUnUS75SJfDAbUSI
23	202859_x_at	CXCL8	3Vy3nJSjUQtfvUe5fo EVLftPlJ7vrLnu_Dxo
24	207269_at	DEFA4	6530r9KdBCqTred_SM
25	206871_at	ELANE	NtF0nlRFRHdSTekQwE
26	224225_s_at	ETV7	9KCUpFJlkVJVFXsDoM
27	213506_at	F2RL1	BSHqBKlSg3u6xAXk64 ZAl9Iq.oUTd.oAy.kE
28	211734_s_at	FCER1A	WgUoCn0h94FRNcJQFU KV7kDSLO4uggquLXB4
29	231093_at	FCRL3	ZUclXxUi6VEoBJeRT8
30	205110_s_at	FGF13	67unrLpPnjv_uzOezU Hd7t51A7ISQ_3qyhSc Hvun37sCEoHuge477c TmCOYpTxe.jV7aAXpE
31	223836_at	FGFBP2	xmX900n31irhK4CFXE
32	223767_at	GPR84	ok0Fe53OSU4FlD1lCQ
33	210321_at	GZMH	rgmXcEkpOlcF9d6l4U
34	206647_at	HBZ	lteivVSVR8WbDmUMBQ
35	213537_at	HLA-DPA1	NFtdNMC3eb3pThValQ
36	203290_at	HLA-DQA1	NA
37	209480_at	HLA-DQB1	TfXanqXzTU0Si0gKnU
38	206697_s_at	HP	fpPOCkkS1WAIIRYIlc
39	215118_s_at	IGH	NA
40	212592_at	IGJ	HhecR84SAQucRJ7rVE
41	217148_x_at	IGLC1	NA
42	234764_x_at	IGLV1-44	NA
43	227140_at	INHBA	cp0iOCS4CISg_oqAqI
44	206494_s_at 206493_at 216956_s_at	ITGA2B	H3o0v7nktkZOvQbtdU
45	212531_at	LCN2	WUTbfV7VDYUuzYaeLk
46	208450_at	LGALS2	WT6GkkHqpCBBqCNF7k
47	222196_at	LOC389906	NA
48	238717_at	LOC441528	NA
49	1569110_x_at	LOC728613	NA
50	202018_s_at	LTF	cQO5fddUoFsfScd65c
51	244523_at	MMD	6p_X8jaueM_Xv1yw6k
52	207329_at 231688_at	MMP8	fgvcFUrFTg7ifFKt.k BelT8r_h3KIiQoopKI
53	203936_s_at	MMP9	fn3Hpm4vVouL4FK6FA
54	203949_at 203948_s_at	MPO	KXq9Q16d7p_cnO4X4k
55	229510_at	MS4A14	QqQF0qheQ0ghuCkj.Q 0JYKEkJ0rogLlAHvOU
56	210254_at 1554892_a_at	MS4A3	lKdKj.P0uB56B0ROV4 Qen_rt.er1_QxUxR9Q B3q9A94B7R88Xs4E6k
57	201058_s_at	MYL9	Tbs1URA5CdFCtV3S1U 3CBVEhgxeipOOJilWo
58	209290_s_at	NFIB	BNHPXrzdFBgpyimm0g QUlzHnJJ50x5LQ8egk
59	221690_s_at	NLRP2	EFIl20d6F6tINX6rcs
60	206343_s_at	NRG1	ruNF5QODCEECIJIKkI EX0VFaBXiuIGVC0kHQ rcTUUeoQlN_wDuaIKI 9QpESNQOTkjUnNJB00 uX15cu4f_VUIuXoST0 ZZlmuuNioCuRWSaaSQ
61	212768_s_at	OLFM4	rgjyBSHxwUPUVVFBVc c70LXLcyj6S.A5.HVU
62	210004_at	OLR1	9V0C7RDnXXjh11UgBQ
63	205040_at	ORM1	fG6ilCJP2dH5410qEU
64	227474_at	PAX8-AS1	NA
65	225207_at	PDK4	ESC1yEuE.fDrqLUnTk
66	207384_at	PGLYRP1	3mxHd1KQUncUXTUg_k
67	41469_at 203691_at	PI3	NeHlSg0lLnuCnfmo6U
68	207341_at	PRTN3	EX243qjuUVl.1eH30Y
69	211748_x_at 212187_x_at	PTGDS	HXlUBYwuThoEModVKk
70	220570_at	RETN	Qz8qQwkk_l6LlIm0XU
71	206851_at	RNASE3	9aLI9auX8dR7hdMDUo
72	213566_at	RNASE6	Te4VV0giY1VcQvr17E
73	214539_at	SERPINB10	f0LQV4ks6o4uOX0lAk
74	1553177_at	SH2D1B	0TlE1O9_rTf.5X9Oi0
75	219519_s_at	SIGLEC1	iiFGSdPjXjklE7iF0E
76	220000_at	SIGLEC5	Q13VUdxtx69ahqiNUw oJC4o.lSL0tenoolEk
77	203021_at	SLPI	No174RVAVBCigl6guU
78	202286_s_at	TACSTD2	WihIT5WgHS3Pz30n5U
79	205513_at	TCN1	i4uimBR4lCiesvG_1k
80	209651_at	TGFB1I1	EJXHlJXn14J32nhJWc iJVfHqoq1_Q9SYNopQ
81	219410_at	TMEM45A	6qievf0j6P7Xs1VS6I r_iBJ6cKOHcLse.k.U
82	206641_at	TNFRSF17	iWCth3hT5.UdUnOigo
83	218876_at	TPPP3	ENZUufqUJFe6TUJ1Xo QLe5eyXThVBUlUpOnA
84	231122_x_at	ZDHHC19	69eJXi6CX97l_V.lR0

7.3.9 Performance of the genomic signature retained on an independent dataset.

For an independent validation, a subset of the Illumina RSV data of Meijas et al., (2013) was used. Since the experimental data and validation data were obtained using different platforms, we linked the data using gene symbols and applied cross-platform transformation (to render gene expression comparable across datasets) as shown on supplementary Figure S6 and extensive described in the supplementary material. Figure 7.5a presents predicted probabilities of severe on the validation data using 75 of our 84 prognostic gene signature that were common in both experimental and validation datasets, while Figure 7.5b presents the LOOCV predicted probabilities from SVM on the entire Illumina data. Both figures show that using the unrestricted data leads to more certain probabilities and slightly improves specificity compared to our signature. Nonetheless, Figure 7.5c illustrates a large agreement between the predicted probabilities of the two models, while Figure 7.5d clearly reveals that both models are alike as demonstrated by the AUCs of 0.858 and 0.856 for our signature and the unrestricted model respectively. To assess the concordance of the expression pattern of our signature on both datasets (Affymetrix and Illumina), we plotted the log_2fold changes of the common 75 genes as

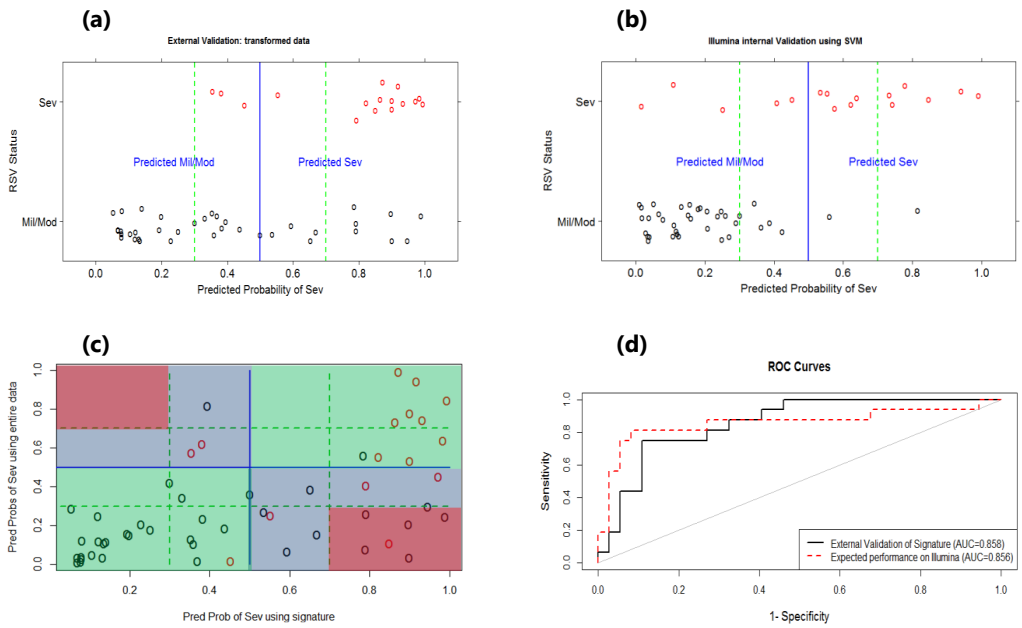

Figure 7.5: Predicted probabilities of being severe from the validation data against true RSV status using; *our diagnostic signature **(a)** and LOOCV on unrestricted data **(b)**. **(c)** illustrates the agreement of the predictions from both models, green regions are perfect agreement, blue are disagreements at a 50% cutoff and red are disagreements at a 30% -70% uncertainty band. Finally, **(d)** presents the ROC curves and AUC from both models illustrating similar AUC values.*

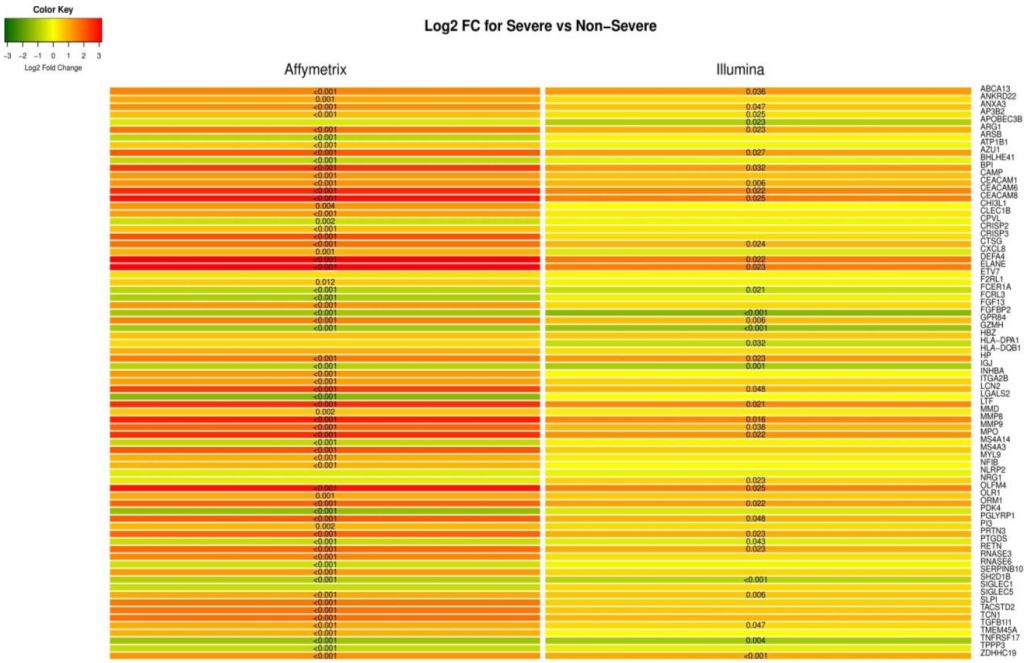

Figure 7.6: log_2 Fold change between Severe vs Non-Severe infants for 75 common genes in Affymetrix and Illumina datasets. *Red represent up-regulation while green represents a down-regulation and the significant FDR adjusted p-values are placed in the cells. As one can clearly see, there is a huge overlap in the direction expressions across datasets. Where there are slight differences, these differences are not significant as shown by a non-significant p-value in at least one of the datasets.*

shown on Figure 7.6. From this figure, one clearly sees that there is a huge overlap in the direction of expressions across datasets. Where there are slight differences, these differences are not significant as shown by a non-significant p-value in at least one of the datasets.

7.4 Discussion

RSV infection in infants may cause life-threatening disease. No vaccine is yet available and triage of patients is challenging since RSV infections may rapidly progress to severe disease. No reliable prognostic model to predict which RSV patient will not progress to severe disease and could be safely send home is available either. Thus, clinical care is symptom-based and a significant proportion of RSV infected infants is hospitalized for observation purposes. We have provided an 84 gene signature that discriminates hospitalized infants with less severe RSV infection from those infants with severe RSV disease. The identified signature yielded a LOOCV AUC of 0.966 on the experimental data and was independently validated with an AUC of 0.858 and might serve as a basis to develop a prognostic test for clinical management of RSV disease.

In line with epidemiological observations Berger et al., (2009) and observations of Mejias et al., (2013), we showed the confounding effects of age and sex on gene expression-severity relationship for RSV disease. Studies in any RSV patient cohort with a naturally occurring "skewed" distribution of age and sex can be standardized for these parameters. By adjusting for an age-by-sex effect in our analyses, we obtained age-by-sex independent results which can be effectively applied to any patient(s). The high performance of our signature on the age and sex matched validation data signifies age-by-sex independence and robustness of this signature. Fewer co-infections were observed in severe patients (Table 7.1). A similar trend has been described previously [Brand et al., 2012]. In our cohort study we did not take into account co-infections since no consistent association between the occurrence or absence of co-infections with RSV disease severity have been reported [Brand et al., 2012; Hasegawa et al., 2014; Mansbach et al., 2008; Chorazy et al., 2013; Martin et al., 2012]. Furthermore, we aimed at the identification of a gene signature in a natural "real-life" cohort of patients not stratified according to age or occurrence of co-infections.

We hypothesized that changes in blood cell type distribution and/or mRNA expression changes of the circulating cells collected from peripheral blood reflect local lung host response characteristics that associate with disease severity. PCA and DE analysis indeed revealed significant changes in the transcriptome profile of whole blood. Gene set analysis further shows that relevant processes are monitored including the activation, migration and tissue infiltration of lymphocytes, granulocytes and neutrophils. Individual DE genes in severe RSV disease revealed overexpression of the neutrophil associated genes MMP8 and MMP9, which have previously been related to severe RSV disease [Brand et al., 2012]. ARG1 and CHI3L1 that have been linked to alternatively activated macrophages in a mouse model for vaccine enhanced RSV disease van Diepen et al., (2015) were also found to be strongly up-regulated. This suggests that the collected blood transcriptome profiles indeed reflect local lung host response.

In our class prediction analysis, three functions were evaluated and the best was chosen. While it has been pointed by Varma et al., (2006); Tibshirani & Tibshirani, (2009); Bernau et al., (2013); Ding et al., (2014) that selecting a minimal-error classifier leads to selection bias that should be corrected, the literature does not stipulate a selection bias when using calibration and refinement scores as evaluation measures. Nevertheless, we employed the nested cross-validation correction of selection bias Tibshirani & Tibshirani, (2009) in our model building procedure by splitting our experimental data into learning and test sets with an inner loop split on the learning set for

parameter(s) optimization. Though found to contain high variance, we utilized leave-one-out cross-validation for the test set because it yields approximately an unbiased estimate of the true (expected) prediction error [Hastie et al., 2009] and because we were interested in the individual sample predicted probability of severe and not entirely on the expected predicted error. Nevertheless, where we were interested in the expected predicted error, as in the optimization of parameters, we utilized five-fold cross-validation as recommended by Breiman & Spector, (1992). To validate the identified signature, an independent dataset generated on a different platform was used. Despite (i) the several sources of variability between our experimental data and the validation data that stem from - but not limited to - array platforms and different clinical cutoffs of RSV severity statuses, (ii) different time of profiling, 1-3days after hospitalization and (iii) loss of information due to a reduction in signature because of no corresponding transcripts on Illumina platform and the aggregation of multiple transcripts to genes, our signature yielded an AUC of 0.858 that was comparable to accuracy (AUC of 0.856) when using the Illumina data (validation set) as experimental set. Cross-platform validation is rare due to lack of guidance on how this can be done reliably. We presented a cross-platform validation procedure.

The RSV patients enrolled in the study displayed varying disease severities but were all hospitalized thus representing a severe disease enriched subset of RSV infected infants. The patients enrolled however also represent a natural cohort of patients including a significant number of patients that eventually did not require extensive medical care and could have been discharged home. Since the blood samples were collected soon after hospital admission, the generated blood transcriptomes and the derived gene signature may serve as a basis for the development of a novel genomic tool to support clinical management of RSV disease including triage of patients presenting at the hospital provided that a rapid (real time) gene test can be developed. Larger transcriptome data sets are however required to construct predictive models that may also allow for discriminating mild from moderate and moderate from severe cases. Ultimately, one would like to extend the RSV biomarker program to earlier time point samples (e.g. obtained when visiting a general practitioner) and to samples collected from patients infected by other (respiratory) infectious agents or pathological conditions (comorbidities) in order to identify specific respiratory viral prognostic biomarkers. To this end a novel gene signature have to be developed using a much larger early blood sample cohort. The current results support the development of diagnostic tests for personalized medicine that not only provide information on the causative infectious agent, but also about the disease severity that may be expected.

Acknowledgements

This study was financially supported by the VIRGO consortium, which is funded by the Netherlands Genomics Initiative and by the Dutch Government (FES0908). We are grateful for the parents and children that participated in this study, the physicians and nurses from the participating hospitals and Mariëtte Las from Pediatric Drug Research Centre, Nijmegen. We wish to thank Mediq Tefa for providing the Cheiron Dynamic II apparatus during the study.

Supplementary Material

Supplementary information accompanies this paper at http://www.nature.com/srep (DOI: 10.1038/srep36603) and includes:

Material and methods: A detail description of the study design, sample processing and data analyses methods.

Figure S1: Principal component analysis (PCA) of all probesets before (a) and after (b) batch adjustment.

Figure S2: Distribution of transcripts' pair-wise correlation values for the RSV experimental data

Figure S3: A plot of the predicted probabilities and the relative frequencies of severe for the initial three classification functions.

Figure S4: Illustration of how the gene signature was extracted by maximizing the binomial log-likelihood function.

Figure S5: A simulated 2-dimensional plot of two genes which individually do not separate the groups but together separates the group almost perfectly. This illustrates the use of non-differentially expressed genes in class discrimination.

Figure S6: Algorithm for building, evaluating and validating the class prediction model.

Table S1: List of genes involved in each and every pathway identified by IPA.

Table S2: Differentially expressed transcripts for the three contrasts and those that were included in the prognostic signature.

Chapter 8

General discussion.

8.1 Research motivation

Recent developments of high-throughput (array, sequencing, mass spectrometric etc.) technologies have given enormous hope for the advancement of personalized medicine. These technologies are capable of quantifying patients' information at different molecular levels and have recently become regular tools for medical research. The amount of data often generated by such technologies is quite huge, producing thousands of variables on just a few hundreds of samples. Gene expression for instance, measures the expression of tens of thousands of genes and has become a widely used tool to identify particular disease subpopulations and to perform diagnostic and prognostic predictions in medical applications [van 't Veer et al., 2002; Huang et al., 2010]. Prediction analyses with such data is often challenging as regular statistical prediction approaches that often require a large number of samples relative to the number parameters, come short. As such, several machine learning and advanced statistical algorithms (functions) have been proposed in the literature to overcome this problem of small sample relative to the number of parameters, often referred to as the curse of dimensionality.

Despite the huge number of such functions outlined in the literature, none consistently outperform others on all datasets as observed by Lee et al., (2005); van Wieringen et al., (2009); Kim & Simon, (2011) that functions perform differently across datasets. It is therefore a challenge to choose an optimal (best fit) function for a given dataset. A common practice is to compare functions and select the best based on their cross-validated errors, but this is often computationally intensive and even when feasible leads to selection bias [Varma et al., 2006; Tibshirani & Tibshirani, 2009; Bernau et al., 2013; Ding et al., 2014] because there is a high probability that a function might have the smallest cross-validation error simply by chance. This probability increases with both the number of algorithms compared and a decrease in sample size. Thus, selection bias is often high in data with small sample sizes and when several algorithms are compared [Ding et al. 2014]. Hence, several bias correction methods have been proposed in the literature of class prediction with gene expression data. Nevertheless, no such method is 100% effective when several least optimal algorithms are compared on a dataset with a small sample size, as observed by Ding et al., (2014).

As such, making a choice of a function to utilize on a given dataset is either at random or by familiarity. An approach we think might lead to selecting the least optimal (worse fit) function for a given dataset. While the characteristics of the functions are known, the literature on why their

performance vary across datasets is quite sparse. In this thesis we hypothesized that gene expression data characteristics affect the performance of the functions. And we seek to identify such data characteristics, utilize these characteristics in evaluating and selecting (with little or no bias) an optimal function for diagnostic and prognostic research with every given gene expression dataset.

8.2 Research outcome

Part I of the thesis is dedicated on identifying gene expression data characteristics that affect the performance of predictive functions. In Chapter 2, we assessed whether correlation structures differ across real-life datasets. Correlation which was observed by Kim & Simon, (2011) via simulations to be associated to the performance of predictive functions was assessed in real-life data. We evaluated the homogeneity of correlation structures within and between 12 downloaded datasets of six etiological disease categories: inflammatory, immune, infectious, degenerative, hereditary and acute myeloid leukemia (AML). The effect of filtering (as one of a preprocessing step for gene expression data) on the correlation structures was also assessed. The datasets were preprocessed by a common procedure incorporating platform specific recommendations and two filtering methods (detection call and variance filtering). Homogeneity of correlation matrices between and within datasets of etiological diseases was assessed using the Box's M statistic on permuted samples. We found that correlation structures significantly differ between datasets of the same disease category and of different etiological disease categories and that variance filtering eliminates more uncorrelated variables than detection call filtering and thus renders the data highly correlated. This therefore signifies that functions that are by design sensitive to correlations will have varying performance across datasets of the same disease category and across disease categories.

In chapter 3, we identified data characteristics that affect the predictive accuracy of classification functions. To achieve this, datasets from twenty-five studies meeting predefined inclusion and exclusion criteria were downloaded. Nine classification functions within the categories: discriminant analyses, tree based, regularization and shrinkage and nearest neighbor methods, that are often utilized in biomedical applications were considered. Consequently, nine classifiers were built for each dataset using the same procedure and their performances were evaluated by calculating their accuracies. The characteristics of each experiment were recorded, (i.e., observed disease, medical question and tissue types) together with the characteristics of the gene

expression data (sample size, number of differentially expressed genes, fold changes and within-class correlations). Their effects on the accuracy of classification functions were statistically assessed by random effects logistic regression models considering both the studies and classification functions as random variables.

The number of differentially expressed genes and the average fold change had significant impact on the accuracy of classification functions and respectively accounted for 72% and 57% of the between study variation. Multiple random effects logistic regression with forward selection yielded the two aforementioned study factors and within class correlation as factors affecting the accuracy of classification functions, explaining 91.5% of the between study variation. Our results showed that the number of differentially expressed genes, the fold change, and the correlation in gene expression data significantly affect the accuracy of class prediction models.

Although we identified 3 data characteristics significantly associated to the accuracy of classification functions in Chapter 3, the net effects of these factors and the non-significant factors could have been confounded by unobserved study factors. Thus, in Part II of this thesis we focused on empirically quantifying the effects of the gene expression data characteristics on the accuracy, Brier score and integrated Brier score of direct classification, probabilistic classification and Cox's prediction functions respectively. And provided predictive models for identifying an optimal function for any given dataset using our empirical results.

In Chapter 4, gene expression data were simulated for different values of gene-pairs correlations, sample size, genes' variances, deferentially expressed genes and fold changes. To simulate gene expression data, we assumed from observations in Chapter 2 that; (i) genes' variances follow an exponential distribution and (ii) the correlation structure is composed of three clusters of up-regulated, down-regulated (referred to as DE-genes) and noisy genes (non-DE genes). We further assumed that the expression values are normally distributed as often expected in real-life log_2 transformed gene expression data. For each simulation scenario, ten direct classifiers were constructed with simulated training data and evaluated with independent test data, using ten classification functions frequently utilized in the literature. The resulting accuracies from different simulation scenarios and classification functions were then modeled using linear mixed effects regression models on the studied data characteristics.

Our models showed that sample size, pairwise correlations of non-DE genes and the proportion of DE genes are the leading factors affecting the accuracy of direct classifiers and respectively accounting for approximately 17%, 14% and 13% of the null variance. While genes' variances and

fold change respectively account for 8% and 7% of the null variance, pairwise correlations between DE genes accounts for simply 1%. Our multiple regression model consisting of the main effects and two-way interactions between these variables explained approximately 70% of the null variance and is used to predict the accuracy of the functions on a given dataset. This model was shown to have high correlations between predicted and expected accuracies of the classification functions on eight real-life datasets. Thus, it serves as a guide for determining an optimal direct classification function for any given gene expression dataset.

Although direct classification is a common practice, it is considered in medical applications to be of less importance relative to probabilistic classification. Medical decision making is complex and misclassification costs are often high [Kim & Simon, 2011]. Thus, probabilistic classifiers that provide an estimate of the probability of class membership for new cases are considered to be more useful than classification rules that simply assign cases to a class [Pepe, 2005; Malley et al., 2012]. Additionally, it has been shown by Kim & Simon, (2011) that an optimal direct classifier on a given dataset is not necessarily an optimal probabilistic classifier for that dataset. Thus, it is worth providing a guide for selecting an optimal probabilistic classification function for any given dataset similar to the guide provided in Chapter 4 for direct classifiers. In Chapter 5, we compared probabilistic and direct classifiers and devised a predictive model for determining an optimal probabilistic function for class prediction with a given dataset.

In a similar manner as in Chapter 4, Gene expression data were simulated for different values of gene-pairs correlations, sample size, genes' variances, number of deferentially expressed genes and fold changes. Nine out of the ten functions considered in Chapter 4 that by design can produce or could self-process to produce probabilities were considered. For each simulated dataset, nine probabilistic classifiers were built and evaluated using Brier score. The resulting Brier scores from the functions on independent test sets at different simulation scenarios were then modeled using linear random effects regression models on the studied data characteristics. Pairwise correlations of non-DE genes and sample size were the leading factors respectively accounting for approximately 18% and 17% of the null variance while the proportion of differentially expressed genes, genes' variances and log_2 fold change respectively account for approximately 9%, 8% and 5% of the null variance, pairwise correlations between DE genes accounts for simply 1%. Unlike sample size as leading factor in explaining the null variance for direct classifiers, pairwise correlations within non-DE genes was the leading factor explaining the null variance for probabilistic classifiers.

Nevertheless, the multiple regression analysis showed an association of the main effects and two-way interactions of all the variables to the Brier scores and yielded a model that explained approximately 62% of the null variance and predicts the Brier score of the functions on any given dataset. An application of our model on twelve real-life datasets also showed high positive correlations between the predicted and expected Brier scores. Hence, we have presented a predictive model that might serve as a guide for determining an optimal probabilistic function for any given gene expression data. Additionally, have we confirmed results of Kim & Simon, (2011) that the optimality of a function on a given dataset depends on whether it is trained as a direct or probabilistic classifier.

Just like for class prediction analysis, several (but relatively few) functions have been developed in the literature of survival prediction with high-throughput data. And these functions have also been shown to perform differently across gene expression datasets irrespective of the measure of evaluation used [Bøvelstad et al., 2007; van Wieringen et la., 2009]. In Chapter 6, we focused on empirically identifying gene expression data characteristics associated to the performance of Cox's predictive functions. Additionally, we seek to provide a guide for determining an optimal Cox's predictive function for any given gene expression data in this chapter.

Gene expression data were simulated for different values of sample size, proportion of events, proportion of informative genes, genes' variances, log_2 effect sizes of informative genes and absolute values of the pairwise correlation within and between informative and non-informative genes. For each simulated dataset, seven Cox's predictive functions were trained and evaluated using integrated Brier score (IBS). The resulting IBS from different simulation scenarios and predictive functions were then modeled using linear random effects regression models on the studied data characteristics. These models showed that pairwise correlations within informative genes, the number (proportion) of informative genes and the pairwise correlations within non-informative genes are the leading factors explaining respectively 14.0%, 10.3% and 9.1% of the null variance. While sample size and the number (proportion) of events respectively explain 7.1% and 4.5%. Genes' variances and log_2 effect sizes account for less than 1% each.

A multiple regression analysis showed no association of the effect size to the IBS of the functions but an association of the main effects, two-way and three-way interactions of other variables. This model explained barely 53% of the null variance. An application of our model on two real-life datasets showed high agreement between the predicted and expected IBS. Thus we've shown in this chapter that sample size, proportion of events, proportion of informative genes, genes'

variances, absolute values of the pairwise correlation within and between informative and non-informative genes associate to the performance of Cox's predictive functions. And we have presented a predictive model that might serve as a guide for determining an optimal Cox's predictive function for any given gene expression data.

For easy application of our proposed models in Part II, an R package titled ""`SPreFuGED`: **S**electing a **Pre**dictive **Fu**nction for a given **G**ene **E**xpression **D**ata" has been developed to allow researchers determine an optimal predictive function for any given gene expression dataset. In this package, one simply supplies the gene expression dataset as a matrix, the response variable as a binary vector or a survival time & status matrix and the type of (direct, probabilistic or survival) predictions to be made. The package returns either the predicted accuracies, Brier scores or integrated Brier scores (respectively for direct, probabilistic or survival predictions) for all the functions from which the optimal one can be chosen for analysis of such dataset.

In Part III (Chapter 7), we focused on the application of our preliminary methodological results to identify a predictive signature for Respiratory syncytial virus (RSV) disease severity in infants. RSV causes infections that range from common cold to severe lower respiratory tract infection requiring high-level medical care. Prediction of the course of disease in individual patients remains challenging at the first visit to the pediatric wards and RSV infections may rapidly progress to severe disease. As such, infants are often hospitalized for observational purpose and in peak seasons, this often leads to congestion at the pediatric wards posing a huge socio-economic burden. There is an unmet need for an RSV vaccine but despite considerable research efforts no licensed vaccine has been developed. Additionally, there is no reliable tool to identify which RSV patient will progress to severe disease stage.

In this chapter, we used early blood gene expression profiles from 39 hospitalized infants that were followed until recovery and of which the level of disease severity was determined retrospectively. Applying our approaches for selecting an optimal probabilistic function for this gene expression data, support vector machine was selected. This function applied on age by sex standardized gene expression data yielded an 84 gene signature that discriminated hospitalized infants with eventually less severe RSV infection from infants that suffered from most severe RSV disease. This signature was shown to be highly accurate in its predictions on the experimental data via cross-validation and on an independent validation cohort. We also demonstrated the added value of gene expression data when combined with phenotypic factors like age and sex in predicting RSV disease severity. The current results support the development of diagnostic tests

for personalized medicine that not only provide information on the causative infectious agent, but also about the disease severity that may be expected.

8.3 Perspectives for future research

- In Part I, we presented correlation matrices as histograms and used basic statistical methods to summarize a within group correlation matrix into a single value that was utilized in our random effects regression analysis in Chapter 3. This might not be a proper way to summarize a matrix and might have led to loss of information. As such, the topic on summarizing a matrix to a single value remains open for further research.
- In Part II chapters 4 & 5 concerning binary classification, we assumed equal proportions of samples in both classes thereby ignoring class-imbalance. Nevertheless, in highly imbalanced datasets, accuracy may yield overoptimistic results, because a classification model might easily send all samples to the majority class. A meaningful classification model necessarily should have higher accuracy than the proportion of the majority class. This factor was not considered in our simulations and its effect on the performance of the classification functions remains a topic to be investigated.
- Still in Part II we assumed that there is little or no optimistic bias using our approach but this was not systematically investigated and remains a topic to pursue.
- In the entire Part II, we chose the values of the gene expression variables based on observed data and what is often assumed in the field. These assumptions might be violated and values may differ over time; for instance, with the decrease in the cost and time required to profile a sample, the total sample size might increase from hundreds to thousands. As such, extending these simulations results for different values of the variables remains a topic of research that can be pursued to meet up with recent developments.
- Also in Part II, we observed that our random effects linear models do not explain the entire null variance as would have been expected. This could be an indication that our models do not fit our simulation results perfectly. As such, other models like non-linear models could be tested on our simulations results whether they best explain the data and possibly improve the predictions assign to the predictive functions.
- Although we focused on microarray technologies for measuring gene expression data, we think that the biological measurements of a particular gene expression data should be preserved across technologies e.g. from microarray to RNASeq, and our models could easily be

applied across technologies. Nevertheless, this topic was not address in this thesis and could be assessed.

- In the RSV cohort presented in Part III (Chapter 7), blood samples were collected soon after hospital admission and disease progress might have reach an advanced stage in most patients. It will be nice to extend the RSV biomarker program to earlier time point samples (e.g. obtained when visiting a general practitioner) and to samples collected from patients infected by other (respiratory) infectious agents or pathological conditions (comorbidities) in order to identify specific respiratory viral prognostic biomarkers. To this end a novel gene signature have to be developed using a much larger early blood sample cohort.

8.4 Conclusion

In thesis, we have identified gene expression data characteristics that affect the performance of most predictive functions (machine learning algorithms) and we have presented models for selecting an optimal predictive function for a given gene expression data with little or no bias. Our models are flexible in that our simulations results are publicly available and can be easily extended with new predictive functions or simulation scenarios and the models updated. Additionally, these models have been assembled into an R package `SPreFuGED` and can be easily applied by both statisticians and non-statisticians.

Chapter 8

Bibliography

Bibliography

Adams, M., & Doull, I. (2009). Management of bronchiolitis. *Paediatrics and Child Health*, *19*(6), 266–270.

Arijs, I., Li, K., Toedter, G., Quintens, R., Van Lommel, L., Van Steen, K., … Rutgeerts, P. (2009). Mucosal gene signatures to predict response to infliximab in patients with ulcerative colitis. *Gut*, *58*(12), 1612–1619.

Bacher, U., Schnittger, S., Macijewski, K., Grossmann, V., Kohlmann, A., Alpermann, T., … Haferlach, T. (2012). Multilineage dysplasia does not influence prognosis in CEBPA-mutated AML, supporting the WHO proposal to classify these patients as a unique entity. *Blood*, *119*(20), 4719–4722.

Bair, E., & Tibshirani, R. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biology*, *2*(4).

Bansard, C., Lequerre, T., Derambure, C., Vittecoq, O., Hiron, M., Daragon, A., … Salier, J.-P. (2011). Gene profiling predicts rheumatoid arthritis responsiveness to IL-1Ra (anakinra). *Rheumatology*, *50*(2), 283–292.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). lme4: Linear mixed-effects models using Eigen and S4. R package.

Becker, R. A., Chambers, J. M., & Wilks, A. R. (1988). *The New S Language*. Wadsworth & Brooks/Cole.

Beghini, A., Corlazzoli, F., Del Giacco, L., Re, M., Lazzaroni, F., Brioschi, M., … Cairoli, R. (2012). Regeneration-associated WNT signaling is activated in long-term reconstituting AC133bright acute myeloid leukemia cells. *Neoplasia* , *14*(12), 1236–48.

Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, *34*, 1713–1723.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289 – 300.

Berger, T. M., Aebi, C., Duppenthaler, A., & Stocker, M. (2009). Prospective Population-based Study of RSV-related Intermediate Care and Intensive Care Unit Admissions in Switzerland over a 4-Year Period (2001–2005). *Infection*, *37*(2), 109–116.

Bernau, C., Augustin, T., & Boulesteix, A.-L. (2013). Correcting the Optimal Resampling-Based Error Rate by Estimating the Error Rate of Wrapper Algorithms. *Biometrics*, *69*(3), 693–702.

Binder, H., & Schumacher, M. (2008). Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics*, *9*(14).

Blagus, R., & Lusa, L. (2010). Class prediction for high-dimensional class-imbalanced data. *BMC Bioinformatics*, *11*(1), 523.

Blalock, E. M., Geddes, J. W., Chen, K. C., Porter, N. M., Markesbery, W. R., & Landfield, P. W. (2004). Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proceedings of the National Academy of Sciences*, *101*(7), 2173–2178.

Bochukova, E. G., Soneji, S., Wall, S. A., & Wilkie, A. O. M. (2010). Scalp fibroblasts have a shared expression profile in monogenic craniosynostosis. *Journal of Medical Genetics*, *47*(12), 803–808.

Boulesteix, A.-L. (2013). On representative and illustrative comparisons with real data in bioinformatics: response to the letter to the editor by Smith et al. *Bioinformatics*, *29*(20), 2664–2666.

Bøvelstad, H. M., Nygård, S., Størvold, H. L., Aldrin, M., Borgan, Frigessi, A., & Lingjærde, O. C. (2007). Predicting survival from microarray data - A comparative study. *Bioinformatics*, *23*(16), 2080–2087.

Brand, H. K., de Groot, R., Galama, J. M. D., Brouwer, M. L., Teuwen, K., Hermans, P. W. M., … Warris, A. (2012). Infection with multiple viruses is not associated with increased disease severity in children with bronchiolitis. *Pediatric Pulmonology*, *47*(4), 393–400.

Brand, K. H., Ahout, I. M. L., de Groot, R., Warris, A., Ferwerda, G., & Hermans, P. W. M. (2012). Use of MMP-8 and MMP-9 to assess disease severity in children with viral lower respiratory tract infections. *Journal of Medical Virology*, *84*(9), 1471–1480.

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

Breiman, L., & Spector, P. (1992). Submodel selection and evaluation in regression . The X-random case. *International Statistical Review*, *60*(3), 291–319.

Bronner, I. F., Bochdanovits, Z., Rizzu, P., Kamphorst, W., Ravid, R., van Swieten, J. C., & Heutink, P. (2009). Comprehensive mRNA Expression Profiling Distinguishes Tauopathies and Identifies Shared Molecular Pathways. *PLoS ONE*, *4*(8), e6826.

Bush, A., & Thomson, A. H. (2007). Acute bronchiolitis. *BMJ*, *335*(7628), 1037–1041.

Chorazy, M. L., Lebeck, M. G., McCarthy, T. A., Richter, S. S., Torner, J. C., & Gray, G. C. (2013). Polymicrobial Acute Respiratory Infections in a Hospital-based Pediatric Population. *The Pediatric Infectious Disease Journal*, *32*(5), 460–466.

Cox, D. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 187–220.

Ding, Y., Tang, S., Liao, S. G., Jia, J., Oesterreich, S., Lin, Y., & Tseng, G. C. (2014). Bias correction for selecting the minimal-error classifier from many machine learning models. *Bioinformatics*, *30*(22), 3152–3158.

Bibliography

Dudoit, S., Fridlyans, J., & Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, *97*(457), 77–87.

Dupuy, A., & Simon, R. M. (2007). Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting. *JNCI Journal of the National Cancer Institute*, *99*(2), 147–157.

Falsey, A. R., McElhaney, J. E., Beran, J., van Essen, G. A., Duval, X., Esen, M., … Taylor, S. (2014). Respiratory Syncytial Virus and Other Respiratory Viral Infections in Older Adults With Moderate to Severe Influenza-like Illness. *Journal of Infectious Diseases*, *209*(12), 1873–1881.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–22.

Gautier, L., Cope, L., Bolstad, B. M., & Irizarry, R. A. (2004). affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, *20*(3), 307–315.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., … Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, *5*(10), R80.

Genz, A., & Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities*. Heidelberg: Springer-Verlag.

Goeman, J. J. (2010). L1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal*, *52*(1), 70–84.

Golub, G., & van Loan, C. (1996). *Matrix Computations* (3rd ed.). Baltimore: Johns Hopkins.

Greco, S., Fasanaro, P., Castelvecchio, S., D'Alessandra, Y., Arcelli, D., Di Donato, M., … Martelli, F. (2012). MicroRNA Dysregulation in Diabetic Ischemic Heart Failure Patients. *Diabetes*, *61*(6), 1633–1641.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, *46*(1-3), 389–422.

Hall CB, Weinberg GA, Iwane MK, & et.al. (2009). The burden of respiratory syncytial virus infection in young children. *N Engl J Med*, *360*(6), 588–98.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques* (2nd ed.). CA: Morgan Kaufmann.

Hasegawa, K., Mansbach, J. M., Teach, S. J., Fisher, E. S., Hershey, D., Koh, J. Y., … Camargo, C. A. (2014). Multicenter Study of Viral Etiology and Relapse in Hospitalized Children With Bronchiolitis. *The Pediatric Infectious Disease Journal*, *33*(8), 809–813.

Hastie, T., Tibshirani, R., & Friedman, J. (2003). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. NY: Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. NY: Springer.

Hastie, T., Tibshirani, R., Narasimhan, B., & Chu, G. (2014). pamr: Pam: prediction analysis for microarrays. R Package.

Hothorn, T., Buhlmann, P., Dudoit, S., Molinaro, A., & Van Der Laan, M. J. (2006). Survival ensembles. *Biostatistics*, *7*(3), 355–373.

Huang, J., Shi, W., Zhang, J., Chou, J. W., Paules, R. S., Gerrish, K., … Bushel, P. R. (2010). Genomic indicators in the blood predict drug-induced liver injury. *The Pharmacogenomics Journal*, *10*(4), 267–277.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A., & Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, *18*(Suppl.1), S96–S104.

Hyrcza, M. D., Kovacs, C., Loutfy, M., Halpenny, R., Heisler, L., Yang, S., … Der, S. D. (2007). Distinct Transcriptional Profiles in Ex Vivo CD4+ and CD8+ T Cells Are Established Early in Human Immunodeficiency Virus Type 1 Infection and Are Characterized by a Chronic Interferon Response as Well as Extensive Transcriptional Changes in CD8+ T Cells. *Journal of Virology*, *81*(7), 3477–3486.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *Annals of Applied Statistics*, *2*(3), 841–860.

Jain, S., Self, W. H., Wunderink, R. G., Fakhran, S., Balk, R., Bramley, A. M., … Finelli, L. (2015). Community-Acquired Pneumonia Requiring Hospitalization among U.S. Adults. *New England Journal of Medicine*, *372*(9), 835-45.

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

Jong, V. L., Novianti, P. W., Roes, K. C. B., & Eijkemans, M. J. C. (2014). Exploring homogeneity of correlation structures of gene expression datasets within and between etiological disease categories. *Statistical Applications in Genetics and Molecular Biology*, *13*(6), 717–732.

Jong, V. L., Novianti, P. W., Roes, K. C. B., & Eijkemans, M. J. C. (2016). Selecting a classification function for class prediction with gene expression data. *Bioinformatics*, *32*(12), 1814–1822.

Kabakchiev, B., Turner, D., Hyams, J., Mack, D., Leleiko, N., Crandall, W., … Silverberg, M. S. (2010). Gene Expression Changes Associated with Resistance to Intravenous Corticosteroid Therapy in Children with Severe Ulcerative Colitis. *PLoS ONE*, *5*(9), e13085.

Bibliography

Kattan, M. W., Eastham, J. A., Stapleton, A. M., Wheeler, T. M., & Scardino, P. T. (1998). A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer. *Journal of the National Cancer Institute*, *90*(10), 766–71.

Kaufman, L., & Rousseeuw, P. J. (Eds.). (2005). *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Kim, K. I., & Simon, R. (2011). Probabilistic classifiers with high-dimensional data. *Biostatistics*, *12*(3), 399–412.

Kruppa, J., Liu, Y., Biau, G., Kohler, M., König, I. R., Malley, J. D., & Ziegler, A. (2014a). Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory. *Biometrical Journal*, *56*(4), 534–563.

Kruppa, J., Liu, Y., Biau, G., Kohler, M., König, I. R., Malley, J. D., & Ziegler, A. (2014b). Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory. *Biometrical Journal*, *56*(4), 534–563.

Krzystanek, M., Szallasi, Z., & Eklund, A. C. (2013). Biasogram: Visualization of Confounding Technical Bias in Gene Expression Data. *PLoS ONE*, *8*(4), e61872.

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*(559).

Langfelder, P., Zhang, B., & Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, *24*(5), 719–720.

Le Dieu, R., Taussig, D. C., Ramsay, A. G., Mitter, R., Miraki-Moud, F., Fatah, R., … Gribben, J. G. (2009). Peripheral blood T cells in acute myeloid leukemia (AML) patients at diagnosis have abnormal phenotype and genotype and form defective immune synapses with AML blasts. *Blood*, *114*(18), 3909–3916.

Lee, J. C., Lyons, P. A., McKinney, E. F., Sowerby, J. M., Carr, E. J., Bredin, F., … Smith, K. G. C. (2011). Gene expression profiling of CD8+ T cells predicts prognosis in patients with Crohn disease and ulcerative colitis. *Journal of Clinical Investigation*, *121*(10), 4170–4179.

Lee, J. W., Lee, J. B., Park, M., & Song, S. H. (2005). An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics and Data Analysis*, *48*(4), 869–885.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*, 18–22.

Lunnon, K., Sattlecker, M., Furney, S. J., Coppola, G., Simmons, A., Proitsi, P., … Hodges, A. (2013). A blood gene expression marker of early Alzheimer's disease. *Journal of Alzheimer's Disease*, *33*(3), 737–753.

Majeti, R., Becker, M. W., Tian, Q., Lee, T.-L. M., Yan, X., Liu, R., … Weissman, I. L. (2009). Dysregulated gene expression networks in human acute myelogenous leukemia stem cells. *Proceedings of the National Academy of Sciences*, *106*(9), 3396–3401.

Malley, J. D., Kruppa, J., Dasgupta, A., Malley, K. G., & Ziegler, A. (2012). Probability machines: consistent probability estimation using nonparametric learning machines. *Methods of Information in Medicine*, *51*(1), 74–81.

Mansbach, J. M., McAdam, A. J., Clark, S., Hain, P. D., Flood, R. G., Acholonu, U., & Camargo, C. A. (2008). Prospective Multicenter Study of the Viral Etiology of Bronchiolitis in the Emergency Department. *Academic Emergency Medicine*, *15*(2), 111–118.

Marczyk, M., Jaksik, R., Polanski, A., & Polanska, J. (2013). Adaptive filtering of microarray gene expression data based on Gaussian mixture decomposition. *BMC Bioinformatics*, *14*(1101).

McLachlan, G. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. NY: Wiley.

Meissner, H. C. (2016). Viral Bronchiolitis in Children. *New England Journal of Medicine*, *374*(1), 62–72.

Mejias, A., Dimo, B., Suarez, N. M., Garcia, C., Suarez-Arrabal, M. C., Jartti, T., … Ramilo, O. (2013). Whole Blood Gene Expression Profiles to Assess Pathogenesis and Disease Severity in Infants with Respiratory Syncytial Virus Infection. *PLoS Medicine*, *10*(11), e1001549.

Menke, A., Arloth, J., Pütz, B., Weber, P., Klengel, T., Mehta, D., … Binder, E. B. (2012). Dexamethasone Stimulated Gene Expression in Peripheral Blood is a Sensitive Marker for Glucocorticoid Receptor Resistance in Depressed Patients. *Neuropsychopharmacology*, *37*(6), 1455–1464.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2014). Misc functions of the Department of Statistics (e1071), TU Wien.

Mogensen, U. B., Ishwaran, H., & Gerds, T. A. (2012). Evaluating Random Forests for Survival Analysis using Prediction Error Curves. *Journal of Statistical Software*, *50*(11), 1–23.

Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., … Groop, L. C. (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, *34*(3), 267–273.

Nair, H., Simões, E. A., Rudan, I., Gessner, B. D., Azziz-Baumgartner, E., Zhang, J. S. F., … Campbell, H. (2013). Global and regional burden of hospital admissions for severe acute lower respiratory infections in young children in 2010: a systematic analysis. *The Lancet*, *381*(9875), 1380–1390.

Novianti, P. W., Jong, V. L., Roes, K. C. B., & Eijkemans, M. J. C. (2015). Factors affecting the accuracy of a class prediction model in gene expression data. *BMC Bioinformatics*, *16*(199).

Bibliography

Novianti, P. W., Roes, K. C. B., & Eijkemans, M. J. C. (2014). Evaluation of Gene Expression Classification Studies: Factors Associated with Classification Performance. *PLoS ONE*, *9*(4), e96063.

Ntzani, E. E., & Ioannidis, J. P. (2003). Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *The Lancet*, *362*(9394), 1439–1444.

Ogata, S., Ogihara, Y., Nomoto, K., Akiyama, K., Nakahata, Y., Sato, K., … Ishii, M. (2009). Clinical Score and Transcript Abundance Patterns Identify Kawasaki Disease Patients Who May Benefit From Addition of Methylprednisolone. *Pediatric Research*, *66*(5), 577–584.

Olsen, J., Gerds, T. A., Seidelin, J. B., Csillag, C., Bjerrum, J. T., Troelsen, J. T., & Nielsen, O. H. (2009). Diagnosis of ulcerative colitis before onset of inflammation by multivariate modeling of genome-wide gene expression data. *Inflammatory Bowel Diseases*, *15*(7), 1032–1038.

Payton, J. E., Grieselhuber, N. R., Chang, L.-W., Murakami, M., Geiss, G. K., Link, D. C., … Ley, T. J. (2009). High throughput digital quantification of mRNA abundance in primary human acute myeloid leukemia samples. *Journal of Clinical Investigation*, *119*(6), 1714–1726.

Pepe, M. S. (2005). Evaluating technologies for classification and prediction in medicine. *Statistics in Medicine*, *24*(24), 3687–3696.

Peters, A., & Hothorn, T. (2015). ipred: Improved Predictors. R package.

Pulendran, B., Oh, J. Z., Nakaya, H. I., Ravindran, R., & Kazmin, D. A. (2013). Immunity to viruses: learning from successful human vaccines. *Immunological Reviews*, *255*(1), 243–255.

R Development Core Team. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rasimas, J., Katsounas, A., Raza, H., Murphy, A. A., Yang, J., Lempicki, R. A., … Rosenstein, D. (2012). Gene Expression Profiles Predict Emergence of Psychiatric Adverse Events in HIV/HCV-Coinfected Patients on Interferon-Based HCV Therapy. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, *60*(3), 273–281.

Ripley, B. (1996). *Pattern Recognition and Neural Networks*. MA: Cambridge University Press.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*.

Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507–2517.

Saintigny, P., Zhang, L., Fan, Y.-H., El-Naggar, A. K., Papadimitrakopoulou, V. A., Feng, L., … Mao, L. (2011). Gene Expression Profiling Predicts the Development of Oral Cancer. *Cancer Prevention Research*, *4*(2), 218–229.

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R*. NY: Springer.

Schäfer, J., & Strimmer, K. (2005). A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. *Statistical Applications in Genetics and Molecular Biology*, *4*(1).

Scherzer, C. R., Eklund, A. C., Morse, L. J., Liao, Z., Locascio, J. J., Fefer, D., … Gullans, S. R. (2007). Molecular markers of early Parkinson's disease based on gene expression in blood. *Proceedings of the National Academy of Sciences*, *104*(3), 955–960.

Schölkopf, B., & Smola, A. (2002). *Learning with Kernels*. MA: MIT Press.

Schumacher, M., Binder, H., & Gerds, T. (2007). Assessment of survival prediction models based on microarray data. *Bioinformatics*, *23*(14), 1768–1774.

Schuurhof, A., Bont, L., Pennings, J. L. A., Hodemaekers, H. M., Wester, P. W., Buisman, A., … Janssen, R. (2010). Gene Expression Differences in Lungs of Mice during Secondary Immune Responses to Respiratory Syncytial Virus Infection. *Journal of Virology*, *84*(18), 9584–9594.

Scian, M. J., Maluf, D. G., Archer, K. J., Suh, J. L., Massey, D., Fassnacht, R. C., … Mas, V. (2011). Gene Expression Changes Are Associated With Loss of Kidney Graft Function and Interstitial Fibrosis and Tubular Atrophy: Diagnosis Versus Prediction. *Transplantation*, *91*(6), 657–665.

Sekaly, R., & Pulendran, B. (2012). Systems biology in understanding HIV pathogenesis and guiding vaccine development. *Current Opinion in HIV and AIDS*, *7*(1), 1–3.

Shi, L., Campbell, G., Jones, W. D., Campagne, F., Wen, Z., Walker, S. J., … Wolfinger, R. D. (2010). The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology*, *28*(8), 827–838.

Simon, R. (2014). Class probability estimation for medical studies. *Biometrical Journal*, *56*(4), 597–600.

Simon, R. M., Subramanian, J., Li, M. C., & Menezes, S. (2011). Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data. *Briefings in Bioinformatics*, *12*(3), 203–214.

Slawski, M., Daumer, M., & Boulesteix, A.-L. (2008). CMA: a comprehensive Bioconductor package for supervised classification with high dimensional data. *BMC Bioinformatics*.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, *3*, Article3.

Statnikov, A., Aliferis, C. F., Tsamardinos, I., Hardin, D., & Levy, S. (2005). A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, *21*(5), 631–643.

Bibliography

Stijnen, T., Hamza, T. H., & Özdemir, P. (2010). Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Statistics in Medicine*, *29*(29), 3046–3067.

Stirewalt, D. L., Pogosova-Agadjanyan, E. L., & Ochsenreither, S. (2012). E-GEOD-37307 - Aberrant expressed genes in AML. [www.ebi.ac.uk/arrayexpress/, 15/03/2013].

Stockman, L. J., Curns, A. T., Anderson, L. J., & Fischer-Langley, G. (2012). Respiratory Syncytial Virus-associated Hospitalizations Among Infants and Young Children in the United States, 1997–2006. *The Pediatric Infectious Disease Journal*, *31*(1), 5–9.

Stojanov, S., Lapidus, S., Chitkara, P., Feder, H., Salazar, J. C., Fleisher, T. A., … Kastner, D. L. (2011). Periodic fever, aphthous stomatitis, pharyngitis, and adenitis (PFAPA) is a disorder of innate immunity and Th1 activation responsive to IL-1 blockade. *Proceedings of the National Academy of Sciences*, *108*(17), 7148–7153.

Suárez-Fariñas, M., Shah, K. R., Haider, A. S., Krueger, J. G., & Lowes, M. A. (2010). Personalized medicine in psoriasis: developing a genomic classifier to predict histological response to Alefacept. *BMC Dermatology*, *10*(1).

Tan, F. K., Hildebrand, B. A., Lester, M. S., Stivers, D. N., Pounds, S., Zhou, X., … Arnett, F. C. (2005). Classification analysis of the transcriptosome of nonlesional cultured dermal fibroblasts from systemic sclerosis patients with early disease. *Arthritis & Rheumatism*, *52*(3), 865–876.

Templeton, K. E., Scheltinga, S. A., Beersma, M. F. C., Kroes, A. C. M., & Claas, E. C. J. (2004). Rapid and Sensitive Method Using Multiplex Real-Time PCR for Diagnosis of Infections by Influenza A and Influenza B Viruses, Respiratory Syncytial Virus, and Parainfluenza Viruses 1, 2, 3, and 4. *Journal of Clinical Microbiology*, *42*(4), 1564–1569.

Therneau, T., & Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. NY: Springer.

Tibshirani, R. (1996). Regression Selection and Shrinkage via the Lasso. *Journal of the Royal Statistical Society B*, 58(1), 267-288

Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, *99*(10), 6567–6572.

Tibshirani, R. J. (2009). Univariate shrinkage in the Cox model for high dimensional data. *Statistical Applications in Genetics and Molecular Biology*, *8*(1), Article21.

Tibshirani, R. J., & Efron, B. (2002). Pre-validation and inference in microarrays. *Statistical Applications in Genetics and Molecular Biology*, *1*(1).

Tibshirani, R. J., & Tibshirani, R. (2009). A bias correction for the minimum error rate in cross-validation. *The Annals of Applied Statistics*, *3*(2), 822–829.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B*, *63*(2), 411–423.

Toedter, G., Li, K., Marano, C., Ma, K., Sague, S., Huang, C. C., … Baribaud, F. (2011). Gene Expression Profiling and Response Signatures Associated With Differential Responses to Infliximab Treatment in Ulcerative Colitis. *The American Journal of Gastroenterology*, *106*(7), 1272–1280.

Tregoning, J. S., & Schwarze, J. (2010). Respiratory Viral Infections in Infants: Causes, Clinical Symptoms, Virology, and Immunology. *Clinical Microbiology Reviews*, *23*(1), 74–98.

Tutz, G., & Binder, H. (2007). Boosting ridge regression. *Computational Statistics & Data Analysis*, *51*(12), 6044–6059.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., … Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, *415*(6871), 530–536.

van de Weg, C. A. M., van den Ham, H.-J., Bijl, M. A., Anfasa, F., Zaaraoui-Boutahar, F., Dewi, B. E., … Andeweg, A. C. (2015). Time since Onset of Disease and Individual Clinical Markers Associate with Transcriptional Changes in Uncomplicated Dengue. *PLOS Neglected Tropical Diseases*, *9*(3), e0003522.

van Diepen, A., Brand, H. K., de Waal, L., Bijl, M., Jong, V. L., Kuiken, T., … Andeweg, A. C. (2015). Host Proteome Correlates of Vaccine-Mediated Enhanced Disease in a Mouse Model of Respiratory Syncytial Virus Infection. *Journal of Virology*, *89*(9), 5022–5031.

van Houwelingen, H. C., Bruinsma, T., Hart, A. A. M., van't Veer, L. J., & Wessels, L. F. A. (2006). Cross-validated Cox regression on microarray gene expression data. *Statistics in Medicine*, *25*(18), 3201–3216.

van Wieringen, W. N., Kun, D., Hampel, R., & Boulesteix, A.-L. (2009). Survival prediction using gene expression data: A review and comparison. *Computational Statistics & Data Analysis*, *53*(5), 1590–1603.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, *7*(1), 91.

Vedin, I., Cederholm, T., Freund-Levi, Y., Basun, H., Garlind, A., Irving, G. F., … Palmblad, J. (2012). Effects of DHA- Rich n-3 Fatty Acid Supplementation on Gene Expression in Blood Mononuclear Leukocytes: The OmegAD Study. *PLoS ONE*, *7*(4), e35425.

Verweij, P. J. M., & Van Houwelingen, H. C. (1993). Cross-validation in survival analysis. *Statistics in Medicine*, *12*(24), 2305–2314.

Verweij, P. J., & van Houwelingen, H. C. (1994). Penalized likelihood in Cox regression. *Statistics in Medicine*, *13*(23-24), 2427–36.

Bibliography

Walter, M., Bonin, M., Pullman, R. S., Valente, E. M., Loi, M., Gambarin, M., … Grundmann, K. (2010). Expression profiling in peripheral blood reveals signature for penetrance in DYT1 dystonia. *Neurobiology of Disease*, *38*(2), 192–200.

Wang, L., Zhu, J., & Zou, H. (2006). The doubly regularized support vector machine. *Statistica Sinica*, *16*(2), 589.

Wessels, L. F. A., Reinders, M. J. T., Hart, A. A. M., Veenman, C. J., Dai, H., He, Y. D., & 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.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer Publishing Company, Incorporated.

Willenbrock, H., Juncker, A. S., Schmiegelow, K., Knudsen, S., & Ryder, L. P. (2004). Prediction of immunophenotype, treatment response, and relapse in childhood acute lymphoblastic leukemia using DNA microarrays. *Leukemia*, *18*(7), 1270–1277.

Wu, F., Dassopoulos, T., Cope, L., Maitra, A., Brant, S. R., Harris, M. L., … Chakravarti, S. (2007). Genome-wide gene expression differences in Crohn's disease and ulcerative colitis from endoscopic pinch biopsies: Insights into distinctive pathogenesis. *Inflammatory Bowel Diseases*, *13*(7), 807–821.

Yang, K., Shan, G., & Zhao, L. (2006). Correlation Coefficient Method for Support Vector Machine Input Samples. In *2006 International Conference on Machine Learning and Cybernetics* (pp. 2857–2861).

Ye, G., & Chen, Y. (2011). Efficient variable selection in support vector machines via the alternating direction method of multipliers. *Of the International Conference on Artificial*, *15*, 832–840.

Zhai, Y., Franco, L. M., Atmar, R. L., Quarles, J. M., Arden, N., Bucasas, K. L., … Couch, R. B. (2015). Host Transcriptional Response to Influenza and Other Acute Respiratory Viral Infections – A Prospective Cohort Study. *PLOS Pathogens*, *11*(6), e1004869.

Zhang, B., & Horvath, S. (2005). A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology*, *4*(1).

Zhang, J., & Boos, D. D. (1992). Bootstrap Critical Values for Testing Homogeneity of Covariance Matrices. *Journal of the American Statistical Association*, *87*, 425–429.

Zhu, C.-Q., Ding, K., Strumpf, D., Weir, B. A., Meyerson, M., Pennell, N., … Tsao, M.-S. (2010). Prognostic and Predictive Gene Signature for Adjuvant Chemotherapy in Resected Non-Small-Cell Lung Cancer. *Journal of Clinical Oncology*, *28*(29), 4417–4424.

Zhu, J. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics*, *5*(3), 427–443.

Zhu, J., Rosset, S., Hastie, T., & Tibshirani, R. (2004). L1-norm support vector machines. *Advances in Neural Information Processing Systems*, *16*, 49–56.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic-net. *Journal of the Royal Statistical Society*, *67*(2), 301–320.

Summary

Summary

In standard medicine, a patient visits his doctor with some symptoms and he gets an evaluation possibly accompanied with some tests. If the patient is fortunate, he is diagnosed of a particular disease and placed on a certain treatment plan. Such plan almost has nothing to do with the specificities of this patient but it is simply based on evidence that the plan works for most people with similar symptoms. That is because medicine as we know, revolves around “standards of care”; the best courses of prevention or treatment for the general population or the average person on the street. When a first plan does not work, the doctor and patient move to a second plan and then to the next; rending standard medicine a trial and error process, with life on the line.

The completion of the human genome project in 2003 and the rapid advancement of high-throughput technologies have called for more patient specific diagnoses and treatments. Technology has now moved from establishing genetic background to measuring the activity of genes (‘gene expression’). These developments have resulted to a medical procedure that separates patients into different groups with medical decisions, practices, interventions and treatments being tailored to the individual patient based on their predicted response or risk of disease and is often referred to as personalized medicine or precision medicine. Personalized medicine emphasizes how unique a patient’s disease risk is; based on his genomic makeup, environment and lifestyle. The genomic makeup particularly, has led to the identification of disease subtypes, response to treatment categories and patients’ prognoses. This has led to a dramatic increase in genomic research over the years. By incorporating gene activity data, prognosis and diagnosis of a patient can be further refined.

Although genomic analyses have led to accurate predictions of patient specific diagnosis and prognoses, the challenge with genomic data is that the amount of data often generated by genomic technologies is quite huge; producing thousands of variables on just a few hundreds of samples. Thus, prediction analyses with genomic data is often problematic as regular statistical prediction methods that often require a large number of samples relative to the number parameters, fall short. As such, several machine learning and advanced statistical functions have been proposed in the literature to overcome this problem of small samples relative to the number of parameters. Despite the large number of such predictive functions outlined in the literature, there is no clear winner, as the functions perform differently across genomic datasets. It is therefore a challenge to choose an optimal predictive function for a given genomic dataset.

A common practice is to compare functions and select the best, but this is often computationally intensive and even when feasible leads to selection bias. Thus, making a choice of a function to utilize on a given dataset is either random, by familiarity or by another trial and error type process. An approach we believe might lead to selecting the least optimal predictive function for a given dataset thereby resulting to less accurate predictions. While the characteristics of most of the predictive functions are known, the literature on the reasons for the variability of the performance of these functions across datasets is quite sparse. In this thesis, we hypothesized that the characteristics of gene expression (genomic) data affect the performance of the functions. Our goal was to identify such data characteristics and propose a guide to utilize these characteristics to select an optimal predictive function for every given gene expression dataset.

In Chapter 2 we assessed the homogeneity of correlation structures between gene expression datasets. The assessment of the homogeneity of correlation structures was performed because correlation had previously been reported in the literature as a data characteristic associated with the performance of predictive functions but whether it is significantly different across datasets was not ascertained. The key finding of this chapter was that correlation structures significantly differ between gene expression datasets of the same disease category and between different etiological disease categories.

In chapter 3, we identified data characteristics that affect the predictive accuracy of classification functions. To achieve this, we downloaded several gene expression datasets from public repositories and nine classification functions that are often utilized in biomedical applications were used to build and evaluate class prediction models on these datasets. The characteristics of the datasets were quantified and their effects on the accuracy of the classification functions were statistically assessed through random effects logistic regression models considering both the studies and classification functions as random variables. We found that the number of differentially expressed genes, the fold change, and the correlation within gene expression data significantly affect the accuracy of class prediction models.

Although we identified three data characteristics significantly associated to the accuracy of classification functions in Chapter 3, the net effects of these factors and the non-significant factors could have been confounded by unobserved study factors. Thus, in Part II (Chapters 4-6) of this thesis we focused on empirically quantifying the effects of the gene expression data characteristics on the accuracy, Brier score and integrated Brier score of direct classification, probabilistic classification and Cox's predictive functions respectively.

Summary

In Chapter 4, gene expression data were simulated for different values of gene-pairs correlations, sample size, genes' variances, deferentially expressed genes and fold changes following our observations in Chapters 1 & 2. For each simulation scenario, ten direct classifiers were constructed and evaluated with simulated training and test data, using ten classification functions frequently utilized in the literature. The resulting accuracies from different simulation scenarios and classification functions were then modeled using linear mixed effects regression models on the studied data characteristics. Our modeling process resulted to a random effects model consisting of the main effects and two-way interactions of these variables. This model showed high correlations between predicted and expected accuracies of the classification functions via its application on several real-life datasets. Thus, it serves as a guide for determining an optimal direct classification function for any given gene expression dataset.

Even though direct classification is a common practice, it is considered in medical applications to be of less importance relative to probabilistic classification because medical decision making is complex and misclassification costs are often high. Thus, probabilistic classifiers that provide an estimate of the probability of class membership for new cases are considered to be more useful than classification rules that simply assign cases to a class. Additionally, the literature reveals that an optimal direct classification function is not necessarily an optimal probabilistic classification function on a given data. As such, in Chapter 5, we simulated data as in Chapter 4 and nine out of the ten functions considered in Chapter 4 that by design can produce or could self-process to produce probabilities were utilized to build and evaluate (using Brier score) probabilistic classifiers on each scenario. The resulting Brier scores were then modeled using linear random effects regression models on the studied data characteristics yielding a model for predicting an optimal probabilistic classification function for a given gene expression data.

Just like for class prediction analysis, several functions have been developed in the literature of survival prediction with high-throughput data. And these functions have also been shown to perform differently across gene expression datasets irrespective of the measure of evaluation used. In Chapter 6 we identified data characteristics associated to the performance of Cox's predictive functions and we provided a guide for determining an optimal Cox's predictive function for any given gene expression data. The data characteristics considered in Chapters 4 & 5 and the number of events were systematically varied in our simulated gene expression data. For each simulated dataset, seven Cox's predictive functions were trained and evaluated using integrated Brier score (IBS). The resulting IBS from different simulation scenarios and predictive functions

were then modeled using linear random effects regression models on the studied data characteristics. This yielded a model for predicting an optimal Cox's predictive function for a given gene expression data.

In Part III (Chapter 7), we utilized our results on predicting an optimal function for a given gene expression data to choose a function for the prognoses of Respiratory syncytial virus (RSV) disease severity in infants. There is an unmet need for an RSV vaccine but despite considerable research efforts no licensed vaccine has been developed. Additionally, there is no reliable tool to identify which RSV patient will progress to severe disease stage posing a challenge in RSV management. In this chapter we utilized our predicted optimal probabilistic classification function to construct and validate a probabilistic classifier for the management of RSV. We provided an 84 gene signature that discriminated hospitalized infants with eventually less severe RSV infection from infants that suffered from most severe RSV disease.

In general, we have identified gene expression data characteristics that affect the performance of most predictive functions and have provided predictive models that can determine an optimal predictive function on any given dataset. For applicability, we have assembled our proposed models in an R package titled "`SPreFuGED`: **S**electing a **Pre**dictive **Fu**nction for a given **G**ene **E**xpression **D**ata".

Summary

Nederlands samenvatting

Nederlands samenvatting

Wanneer iemand bij de dokter komt met klachten, stelt de dokter de symptomen vast en worden er eventueel een aantal testen uitgevoerd. In veel gevallen kan er dan een diagnose gesteld worden en krijgt de patiënt een specifiek behandelplan. Zo'n plan heeft doorgaans weinig te maken met de kenmerken van de patiënt zelf, maar is gebaseerd op bewijs dat deze behandeling werkt bij mensen met dit soort symptomen. Dat komt doordat de huidige geneeskunde is gebaseerd op 'standards of care'; het beste behandelpatroon voor de bevolking als geheel, een soort medische versie van Jan Modaal (m/v). Wanneer dit behandelplan niet werkt zullen de dokter en patiënt uitwijken naar een tweede behandelplan en zo verder. Zo probeert de arts dus met vallen en opstaan de patiënt te genezen van zijn of haar (levensbedreigende) kwaal.

Het voltooien van het Humaan genoom project en de snelle ontwikkelingen in nieuwe technologie geven de mogelijkheid om diagnose en behandeling meer toe te spitsen op de patient. Inmiddels is de techniek zover dat men de genetische achtergrond van een individu kan vaststellen, maar ook de activiteit van genen ('gen-expressie') kan meten. Met behulp van deze nieuwe technologie kunnen patiënten ingedeeld worden in groepen op basis van achtergrond en mogelijke reactie op een behandeling. Hierdoor kunnen diagnoses en behandelingen beter zijn afgestemd op wat deze specifieke patiënt nodig heeft – dit wordt 'gepersonaliseerde' of 'precisie-geneeskunde' genoemd. Gepersonaliseerde geneeskunde legt de nadruk op de unieke eigenschappen van een patiënt door het meenemen van genetische achtergrond, de omgeving en gewoonten van het individu tijdens het behandelen van ziekte. Met name op basis van genetische achtergrond is er de laatste jaren veel vooruitgang geboekt in het indeling van patiënten en het gebruik van deze nieuwe technologie. Door ook gen-activiteit in ogenschouw te nemen kan dit beeld verder verfijnd worden.

Hoewel het gebruik van genoom-data hebben geleid tot nauwkeurige voorspellingen over prognose en diagnose, blijft het een uitdaging om van de grote hoeveelheiden data die beschikbaar zijn effectief te gebruiken – niet elk van de 3 miljard letters in onze genetische code is even informatief. Hetzelfde geldt voor gen expressie: niet alle genactiviteit is van belang. Doorgaans zijn er duizenden of meer meetwaarden voor slechts een paar honderd patiënten. Reguliere statistiek heeft meer observaties dan variabelen nodig heeft om het effect van deze variabele vast te stellen en kan dus niet toegepast worden in deze situatie. In plaats daarvan zijn er de laatste jaren nieuwe statische en computationele technieken ontwikkeld die toch conclusies kunnen trekken uit weinig observaties ten opzichte van het aantal variabelen. Er zijn inmiddels veel van dit soort technieken beschreven, maar het is niet duidelijk wat de beste aanpak is voor

een specifieke vraag of patient. De uitdaging is dus om de goede statistische of computationele methode te kiezen voor het verfijnen van diagnose en prognose bij patiënten.

Voordat een bepaalde statistische of computationele method wordt toegepast wordt er doorgaans eerst een aantal methoden met elkaar vergeleken. De specifieke kenmerken van de verschillende methoden zijn doorgaans goed beschreven, maar de (on)bruikbaarheid van deze methoden in verschillende situaties doorgaans niet. Methoden worden vaak vergeleken door middel van simulaties die veel rekenkracht vergen en vaak geen goed onafhankelijk beeld geven. Daardoor is het kiezen van een methode nu vaak geen precisiewerk, maar een kwestie van vallen en opstaan totdat het gewenste resultaat is bereikt.

In dit proefschrift is de hypothese dat de meetwaarden specifieke kenmerken bevatten die kunnen helpen bij het vaststellen van de meest bruikbare methode. Ons doel is om deze kenmerken te vinden en effectief toe te passen om zo de meest bruikbare statistische of computationele methode te kiezen voor een specifiek probleem.

Résumé en français

Résumé en français

En médecine classique, un patient visite son médecin avec quelques symptômes et il obtient un examen éventuellement accompagnée de quelques tests. Si le patient est chanceux, il est diagnostiqué d'une maladie particulière et placé sous un certain traitement. Ce plan n'a presque rien à voir avec les spécificités de ce patient, mais il est simplement basé sur le fait que le traitement fonctionne pour la plupart des personnes avec des symptômes similaires. C'est parce que la médecine que nous connaissons, tourne autour de «normes de soins»; Les meilleurs plans de prévention ou de traitement pour la population en général ou la personne moyenne dans la rue. Quand un premier plan ne fonctionne pas, le médecin et le patient passent à un deuxième plan, puis à l'autre; rendant la médecine standard a un processus d'essai et d'erreur, avec la vie en jeu.

L'achèvement du projet de génome humain en 2003 et l'avancement rapide des nouvelles technologies ont nécessité plus de diagnostics personnalisés et de traitements spécifiques. Ces développements ont abouti à une procédure médicale qui sépare les patients en différents groupes avec des décisions médicales, des pratiques, des interventions et des traitements adaptés au profil individuel des patients en fonction de leur réponse prédite ou risque de maladie et est souvent appelée médecine personnalisée ou médecine de précision. La médecine personnalisée met l'accent sur l'unique risque de maladie d'un patient est; Basé sur son sillage génomique, son environnement et son style de vie. Le sillage génomique en particulier, a conduit à l'identification des sous-types de la maladie, la réponse aux catégories de traitement et le pronostic des patients. Cela a conduit à une augmentation spectaculaire de la recherche génomique au fil des ans.

Bien que les analyses génomiques aient conduit à des prédictions précises du diagnostic et des pronostics propres aux patients, le défi posé par les données génomiques est que la quantité de données souvent générées par les technologies génomiques est assez importante; Produisant des milliers de variables sur quelques centaines d'échantillons. Ainsi, les analyses de prédiction avec des données d'expression génomiques sont souvent problématiques et ne corroborent pas les méthodes de prédiction statistique classiques qui nécessitent souvent un grand nombre d'échantillons par rapport aux paramètres numériques. En tant que tel, plusieurs fonctions d'apprentissage et fonctions statistiques avancées ont été proposées dans la littérature pour surmonter ce problème de petits échantillons par rapport au nombre de paramètres. Malgré le grand nombre de ces fonctions prédictives décrites dans la littérature, il n'y a pas de lauréat clair, car les fonctions fonctionnent différemment dans les ensembles de données génomiques. Il est

donc difficile de choisir une fonction prédictive optimale pour un ensemble de données génomiques donné.

Une pratique courante est de comparer les fonctions et de sélectionner les meilleures, mais cela est souvent intensif sur le plan des simulations et calculs puis même lorsque cela est possible entraîne un biais de sélection. Ainsi, le choix d'une fonction à utiliser sur un ensemble de données donné est soit aléatoire, soit par familiarité, soit par un autre processus de type essai et erreur. Une approche que nous croyons pourrait conduire à sélectionner la fonction de prévision la moins optimale pour un ensemble de données donné résultant ainsi de prédictions moins précises. Alors que les caractéristiques de la plupart des fonctions prédictives sont connues, la littérature sur les raisons de la variabilité de la performance de ces fonctions à travers les ensembles de données est assez rare. Dans cette thèse, nous avons émis l'hypothèse que les caractéristiques des données d'expression génétique (génomique) affectent la performance des fonctions. Notre objectif était d'identifier ces caractéristiques de données et de proposer un guide pour utiliser ces caractéristiques afin de sélectionner une fonction prédictive optimale pour chaque ensemble d'expression génétique donnée.

Dans le chapitre 2, nous avons évalué l'homogénéité des structures de corrélation entre les ensembles de données d'expression génique. L'évaluation de l'homogénéité des structures de corrélation a été réalisée parce que la corrélation avait été précédemment rapportée dans la littérature comme une caractéristique des données associée à la performance des fonctions prédictives mais si elle était significativement différente dans les ensembles de données n'a pas été déterminée. La conclusion clée de ce chapitre était que les structures de corrélation différaient significativement entre les ensembles de données d'expression génique de la même catégorie de maladie et entre différentes catégories de maladies étiologiques.

Au chapitre 3, nous avons identifié des caractéristiques de données qui influent sur la précision prédictive des fonctions de classification. Pour ce faire, nous avons téléchargé plusieurs ensembles de données d'expression génique à partir de dépôts publics et neuf fonctions de classification qui sont souvent utilisées dans des applications biomédicales ont été utilisées pour construire et évaluer des modèles de prédiction de classe sur ces ensembles de données. Les caractéristiques des ensembles de données ont été quantifiées et leurs effets sur la précision des fonctions de classification ont été évalués statistiquement au moyen de modèles de régression logistique à effets aléatoires, considérant à la fois les études et les fonctions de classification comme variables aléatoires. Nous avons constaté que le nombre de gènes exprimés différemment, le changement

de pli et la corrélation à l'intérieur des données d'expression génique affectent significativement la précision des modèles de prédiction de classe.

Bien que nous ayons identifié trois caractéristiques de données significativement associées à l'exactitude des fonctions de classification dans le chapitre 3, les effets nets de ces facteurs et les facteurs non significatifs auraient pu être confondus par des facteurs d'étude non observés. Ainsi, dans la partie II (chapitres 4-6) de cette thèse, nous nous sommes concentrés sur la quantification empirique des effets des caractéristiques des données d'expression génique sur la précision, le score de Brier et le score de Brier intégré de classification directe, de classification probabiliste et de fonctions prédictives de Cox respectivement.

Dans le chapitre 4, les données d'expression génique ont été simulées pour différentes valeurs de corrélations de couples de gènes, de taille d'échantillon, de variances de gènes, de gènes exprimés de manière différentielle et de changements de plis selon nos observations dans les chapitres 1 et 2. Pour chaque scénario de simulation, Avec simulation de la formation et des données d'essai, en utilisant dix fonctions de classification fréquemment utilisées dans la littérature. Les exactitudes obtenues à partir des différents scénarios de simulation et des fonctions de classification ont ensuite été modélisées en utilisant des modèles de régression linéaire à effets mixtes sur les caractéristiques des données étudiées. Notre modélisation a donné lieu à un modèle d'effets aléatoires composé des principaux effets et des interactions bidirectionnelles de ces variables. Ce modèle a montré des corrélations élevées entre les exactitudes prédites et attendues des fonctions de classification via son application sur plusieurs ensembles de données de la vie réelle. Ainsi, il sert de guide pour déterminer une fonction de classification directe optimale pour tout un ensemble d'informations d'expression génique donné.

Même si la classification directe est une pratique courante, elle est considérée dans les applications médicales comme moins importante par rapport à la classification probabiliste parce que la prise de décision médicale est complexe et les coûts de classification erronée sont souvent élevés. Ainsi, les classificateurs probabilistes qui fournissent une estimation de la probabilité d'appartenance à la classe pour les nouveaux cas sont considérés comme plus utiles que les règles de classification qui attribuent simplement des cas à une classe. De plus, la littérature révèle qu'une fonction de classification directe optimale n'est pas nécessairement une fonction de classification probabiliste optimale pour une base de données spécifique. Ainsi, au chapitre 5, nous simulons des données comme dans le chapitre 4 et neuf des dix fonctions considérées au chapitre 4 qui, par conception, peuvent produire ou pourraient auto-traiter pour produire des

probabilités ont été utilisées pour construire et évaluer (en utilisant le score de Brier) Classificateurs sur chaque scénario. Les résultats de Brier résultants ont ensuite été modélisés en utilisant des modèles de régression linéaire à effets aléatoires sur les caractéristiques de données étudiées, donnant un modèle pour prédire une fonction de classification probabiliste optimale pour des données d'expression de gêne données.

Tout comme pour l'analyse de prédiction de classe, plusieurs fonctions ont été développées dans la littérature de la prédiction de survie avec des données à haut débit. Et ces fonctions ont également été montrées pour effectuer différemment à travers des jeux de données d'expression de gène indépendamment de la mesure de l'évaluation utilisée. Dans le chapitre 6, nous avons identifié des caractéristiques de données associées à la performance des fonctions prédictives de Cox et nous avons fourni un guide pour déterminer la fonction prédictive optimale de Cox pour toutes les données d'expression génique données. En plus des caractéristiques des données considérées dans les chapitres 4 et 5, le nombre d'événements a été systématiquement modifié dans nos données d'expression génique simulées. Pour chaque ensemble de données simulées, sept fonctions prédictives de Cox ont été formées et évaluées à l'aide du score de Brier intégré (IBS). Les IBS résultant de différents scénarios de simulation et de fonctions prédictives ont ensuite été modélisés en utilisant des modèles linéaires de régression des effets aléatoires sur les caractéristiques des données étudiées. Ceci a donné un modèle pour prédire une fonction prédictive de Cox optimale pour des données d'expression de gène données.

Dans la partie III (chapitre 7), nous avons utilisé nos résultats sur la prédiction d'une fonction optimale pour une donnée d'expression génique pour choisir une fonction pour les pronostics de la gravité de la maladie du ''virus syncytial respiratoire (RSV)' chez les nourrissons. Il existe un besoin non satisfait d'un vaccin contre le VRS, mais malgré des efforts de recherche considérables, aucun vaccin homologué n'a été mis au point. En outre, il n'existe aucun outil fiable pour identifier quel patient atteindra le stade de la maladie grave et posera un défi dans la prise en charge du VRS. Dans ce chapitre, nous avons utilisé notre fonction de classification probabiliste optimale prédite pour construire et valider un classificateur probabiliste pour la gestion du VRS. Nous avons fourni une signature de 84 gènes qui a discriminé les nourrissons hospitalisés avec éventuellement moins sévère infection par le VRS chez les nourrissons qui souffraient de la plus grave maladie du VRS.

En général, nous avons identifié des caractéristiques de données d'expression génique qui affectent la performance de la plupart des fonctions prédictives et ont fourni des modèles

prédictifs qui peuvent déterminer une fonction prédictive optimale sur un ensemble de données spécifique. Pour l'applicabilité, nous avons assemblé nos modèles proposés dans un paquet R intitulé "`SPreFuGED`: **S**electing a **Pre**dictive **Fu**nction for a given **G**ene **E**xpression **D**ata".

List of publications

Publications in this thesis

Jong, V. L., Ahout, I. M. L., van den Ham, H.-J., Jans, J., Zaaraoui-Boutahar, F., Zomer, A., … Andeweg, A. C. (2016). Transcriptome assists prognosis of disease severity in respiratory syncytial virus infected infants. *Scientific Reports*, *6*(36603).

Jong, V. L., Novianti, P. W., Roes, K. C. B., & Eijkemans, M. J. C. (2016). Selecting a classification function for class prediction with gene expression data. *Bioinformatics*, *32*(12), 1814–1822.

Novianti, P. W., **Jong, V. L.,** Roes, K. C. B., & Eijkemans, M. J. C. (2015). Factors affecting the accuracy of a class prediction model in gene expression data. *BMC Bioinformatics*, *16*(199).

Jong, V. L., Novianti, P. W., Roes, K. C. B., & Eijkemans, M. J. C. (2014). Exploring homogeneity of correlation structures of gene expression datasets within and between etiological disease categories. *Statistical Applications in Genetics and Molecular Biology*, *13*(6), 717–732.

Other publications

Novianti, P. W., **Jong, V. L.,** Roes, K. C. B. & Eijkemans, M. J. C. (2017). Meta-analysis approach as a gene selection method in class prediction: Does it improve model performance? "A case study in acute myeloid leukemia". *BMC Bioinformatics*, *18(210)*.

van Ooijen, M., **Jong, V. L.,** Eijkemans, M. J. C., Heck, A. J. R., Andeweg, A. C., Binai, N. A. & van den Ham H.-J. (2017). Identification of differentially expressed peptides in high-throughput proteomics data. *Briefings in Bioinformatics* (**DOI:** https://doi.org/10.1093/bib/bbx031)

Novianti, P. W., Van Der Tweel, I., **Jong, V. L.,** Roes, K. C. B., & Eijkemans, M. J. C. (2015). An application of sequential meta-analysis to gene expression studies. *Cancer Informatics*, *14*, 1–10.

van Diepen, A., Brand, H. K., de Waal, L., Bijl, M., **Jong, V. L.,** Kuiken, T., … Andeweg, A. C. (2015). Host Proteome Correlates of Vaccine-Mediated Enhanced Disease in a Mouse Model of Respiratory Syncytial Virus Infection. *Journal of Virology*, *89*(9), 5022–5031.

Acknowledgment

Acknowledgement

Dedicated to

Linda Tanyi

"Trust in the LORD with all thine heart; and lean not on thine own understanding. In all thy ways acknowledge him, and he shall direct thy paths." ***Prov. 3:5-6 (KJV).*** Father I thank you for not allowing me to lean entirely on my understanding throughout these years and I exalt your name for directing my paths. May I never depart from your presence for I know you have plans to prosper me and not to harm me, plans to give me hope and a future.

Dear Prof. dr. ir. M. J. C. Eijkemans (Rene), when we first met in 2011 during the interview for the position of a PhD/Consultant, I made a wrong impression of who you actually are. The tricky nature and complexity of the questions you posed made me to think you were being unnecessarily hard on me but immediately I was recruited and I started working under you, I got to know exactly who you really are. As a supervisor, rather than giving instructions on what to or not to do, you simply ask questions and allow me to make the decisions on the next steps. This unique approach of yours gave me freedom to develop to who I am today. Apart from being a supervisor, you have being more than a dad in that you go beyond research (during our weekly meetings) to inquire about my social life as a curious leader that would want to know whether or not his employee was happy. And very often you will offer a sincere and genuine advise whenever I openly share my social worries with you. Words alone cannot express the respect I and my family do have for you. We sincerely wish to thank you for the supervision, kindness and love you have shown to me(us) and above all for offering me a position to continue within the department. May the Almighty richly bless you and your family.

Dear Prof. dr. K. C. B. Roes (Kit), from day one during the interview in 2011, I have known you as a person of "few but mighty words". This imposed title literally means you will say too little but the little will be very insightful. The "few but mighty words" came in at different stages during my PhD and have accumulated to the current thesis. I am known as a proactive person but when I started as a PhD student, I was a bit shy and did not want to participate in departmental discussions out of my field of research but you were rather quick to notice my proactive nature. You encouraged me to participate in all discussions and this has transformed me to a better team player and consultant. My sincere appreciations to your kind supervision and leadership and I thank you for maintaining me at the department. May the Lord grant your heart desires.

Dear dr. A. C. Andeweg (Arno), until I met you, I thought I had a working knowledge of functional biology but I might have been a beginner. Working in your team during these five years has been quite a boost in my professional career. Apart from being exposed to several real-life challenges as a statistical consultant, our lengthy discussions gradually transformed me to a "biologist?". I am sorry if I did try at one point to transform you to a statistician. Thank you very much for the fruitful collaborations, guidance and biological insights.

Dear dr. H.-J. van den Ham (Henk-Jan), I have been blessed to have you within the team at Erasmus MC. As a biologist turn bioinformatician, you were the bridge between me (a statistician) and the rest of the biologists. Additionally, you have also been my "sworn translator", translating everything from Dutch to English and vice versa. Additionally, you have offered numerous helps to

ensure I fully integrate in Dutch society during my initial days and when I had to move to Rotterdam. Thank you very much for everything and especially for translating my thesis title and summary to Dutch (as usual). May the God we both serve never forsakes you or any member of your generations.

Rene, Kit, Arno and Henk-Jan thank you all for the trust you had in me in selecting me to fill the PhD/Consultant position five years ago and for ensuring that I rendered my services successfully.

Dear dr. P. W. Novianti (Putri), you were my immediate and lone team member (team high-dimensional PhDs) and the daily discussions let to exchange of ideas that resulted to several research outcomes. Additionally, I learned from your successes and mistakes on several aspects including; Julius Center's, UMC Utrecht's and Belastingdienst's (Taxation) procedures. You will forever remain a significant person in my entire academic and professional career. Thank you very much for being there and may you forever be blessed.

To the members of the reading committee of my thesis: Prof. dr. A. B. J. Prakken, Prof. dr. L. F. A. Wessels, Prof. dr. A. H. Zwinderman, Prof. dr. A.-L. Boulesteix and dr. J. de Ridder. I thank you for your time and effort in reading and assessing my thesis.

A M. Aissami Abdou, je vous remercie pour votre soutien moral et surtout pour votre aide avec le résumé en français.

To current and former members of the department of Biostatistics and Research Support: Bert, Caroline, Cas, Esther, Ingeborg, Julien, Konstantinos, Maarten, Marijn, Paul, Peter, Rebecca, Rutger, Stavros and Willems. It was a great pleasure to work with all of you especially during the consultancy sessions. My sincere appreciations to the feedbacks provided during presentations and on the final version of this thesis. I am very pleased to still be part of this team that I can proudly call my professional family.

To the (former) members of the bioinformatics team at the Viroscience lab: Anita, David, Fatiha and Maarten. Thank you all for the collaborations and teamwork.

To other members of the VIRGO consortium and most especially: Inge Ahout, Jop Jans, Aldert Zomer, Elles Simonetti, Kim Brand, Wilfred van IJcken, Marien de Jonge, Pieter Fraaij, Ronald de Groot, Albert Osterhaus, Gerben Ferwerda, Angela van Diepen, Leon de Waal, Thijs Kuiken, Geert van Amerongen, Peter Hermans, Albert Heck and Nadine Binai. I thank you for your kind collaborations.

To the master students I co-supervised (currently co-supervising) during your internships: Michiel and Erik-Jan, your research input led (will lead) to a manuscript. I sincerely thank you for such input and wish you the best in your future career.

To the international PhD students of the Department of Biostatistics: Julien, Konstantinos, Putri and Stavros. Thank you for automatically changing the language of communication during

departmental meetings to English. Special gratitude for the quality time we spent together in the office, during conferences and social activities.

To all the PhD students I shared a room with: Julien and Stavros; Gerdien, Paulien, Putri, Stan, Suzanne and Thomas; Henk, Cindy, Jaike, Konstantinos, Madelief, Michelle, Sarah and Veerle. Thank you all for the noise (some was quite useful though), the birthday anniversaries and the social gatherings and the time we spent together. I wish you all the best in your respective careers.

To the African PhDs and Postdocs at the Julius Center: Daniel, Henok, Kayode, Mary and Sanni. I enjoyed your warm receptions in every gathering and the affections we shared at each moment we met. May the Lord richly bless you and your respective families.

To all other former and current PhD students at the Julius Center: Anoukh, Floriaan, John, Kim, Manon, Noor, Romin, Sophie, Welling, Willemijn and Wouter. Thank you all for your lovely companion in and out of the UMC Utrecht.

To friends of friends: Maria, Martina, Mathilde and Shona. The love you showered my friends and myself made me proud of you all. May you keep up with the spirit.

If I were able to simulatneously fill the position of a PhD student and a Statistical consultant, it was because of the educational background I acquired at my masters degree. To that, I wish to thank all my lecturers at the Center of Statistics, Universiteit Hasselt and most especially my then head of department Prof. dr. Tomasz BURZYKOWSKI and my master thesis supervisors Prof. dr. Ziv SHKEDY and dr. Dan LIN.

After leaving Cameroon to Europe, it was obvious it wasn't going to be like home but due to the efforts of several persons a few of which I will mention below, I could still make a home away from home.

Mr. & Mrs. Ketunze, Dr. & Mrs. Nji, Dr. & Mrs. Ako, Dr. C. Akwi, Elvis, Nico, Sylva, Raoul Matho, Raoul Njunang, Jean Filbert, Mangi Adeline, Geraldine Agbor, Ese, Veteran, Promise, Derek, Henry, Roland Ayuk, Grand Valy, Victor, Fidelis and Marcellus for their support during my stay in Belgium and/or Netherlands.

Dr. & Dr. Mrs. Musoro, I wish to thank you for being there in time of need. I thank you for providing me shelter in Amsterdam for approximately a month at the start of my PhD. May you be blessed in Jesus' name.

To Mr. & Mrs. Nteleah, Mr. & Mrs. Tambe, Mr. & Mrs. Ayuk, Mr. & Mrs. Tanyi, Mr. & Mrs. Agbor, Mr. & Mrs. Nongni, Mr. & Mrs. Agbortoko, Mr. & Mrs. Tambe-Nkongho, Mr & Mrs. Enow, Chief & Mafor. Enow, Pastor Yerima, Bro Tabe, Ma Eli Baiye, Ma Emilia Kema, Manyang Anastasia, Manyang Emilia Ebai, Mafor Mary, Sis. Hannah, Ma Esther, Sis Jacky, Sis Mado, Sis Rita, Ma Becky and the rest of the members of Manyang-Nfai and MECA Holland. Thank you for being a family. Your warm and wonderful support shows that indeed you are my family in the Netherlands. May God continuously bless you all.

Acknowledgement

Special thanks to my mama and papa; Mr. & Mrs. Egbe and mijn jongere broers; Arbort en Reybort. Words cannot express my joy for having you as a family and the immense support I have received from you throughout this entire period in the Netherlands. I shall forever be indebted to you and I pray our Lord richly bless you.

To all the brethren of Mountain of Refuge and Restoration Ministries Rotterdam; headed by Pastor Funmi and visiting leader Apostle Dr. J. Ngambi, men's group headed by Elder Koroma, women's group headed by Sister Queen and the youths (by fire by thunder). I thank you all for the prayers and all forms of support. May our Savior empower and multiply the ministries.

To my family members and friends in Cameroon and around the world; Mr. & Mrs. Akwanga, Mr. & Mrs. Nso L., Pa Atong, Rose Atong, Eta Atong, Mercy Atong, Kate Atong, Glory Atong, Thomas Atong, Valentine Fru, Gordon Nso, Zacs Forchu, Peter Leku, Clery Eta, John Zah, Terrence Baya, Norbert Achu, Froumsia Dokrom, Valentine Ndah, James Forchu, Dr. & Mrs. Akosa and those I have not mentioned. Thank you all for your supports throughout these five years.

To my lovely parents; Chief Mfoataw D. O. & Martha Enow and my siblings; Michael, Florence, Rose, Gladys, Elson, Misherine and Rene Rector. If I am to be called a doctor today, it is because of your never ceasing support to me from birth through primary school, Colleges, Universities up till date. I thank God for blessing me with such a loving family. It is rather unfortunate none of you could witness my defense in person but I do know you are always with me in spirit. I want to seize this opportunity to thank you all and to express my everlasting love for all of you. May our good Lord continue to bless us all.

To my lovely wife Jenne and my sons Adriel and Othniel Britt, although you would want me home before 22:00 daily, it was not always possible. Even when I am home, I am sometimes locked up in my reading room. Despite all these, you never complained. I want to sincerely apologize for depriving you of my presence and to thank you for all the provisions and sacrifices you have made to ensure my success.

If I were to have a lovely angel as Jenne in my life, it was thanks to the support of several persons that made this possible. These include:

i. my parents in-law Mr. Tanyi & Ma Mary Etarock and other family members. Thank you all for this angel you offered to me and for your supports and prayers all this while.

ii. my special European crew (Rene, Caroline, Stavros and Suzette) that attended our wedding in Cameroon and assured Jenne of her wellbeing in Netherlands. I want to seize this occasion to thank you once more for this support that shall forever be cherished.

Anyone keeping count of the number of occurrences of each name would have noticed that the mode is no one else than my special dude, Stavros. You are a friend indeed and I had no choice than to name you one of my paranymphs. I want to thank you and the other paranymph, Valeri, for the efforts made to ensure that my thesis defense and reception is a success.

Curriculum vitea

Curriculum vitea

Victor L. Jong was born on Oct 25, 1982 in Agong, Cameroon. He studied at the University of Buea, Cameroon (2002-2006) where he obtained a BSc. in Mathematics and Computer Sciences with a cum laude. He worked as an Information Technology (IT) instructor at Bilingual Grammar School (BGS) Buea, Cameroon (2005-2007) during which he and his colleagues established the first multimedia center of that institution that would serve the entire southwest province of Cameroon. He designed and offer pioneer IT courses to secondary and high school students of BGS and IT lessons to civil administrators of the province. In Feb 2007, he moved to work at Presprint Plc. Limbe, Cameroon, as a Programmer. At Presprint Plc. (2007-2009) he headed the application development team that developed and administered the first database driven web application for the computerization of drivers' licenses, as part of a contract between Presprint Plc. and Cameroon's ministry of transport. His joint effort with his team members and members of other teams yielded an efficient information system that led to the start of production of computerized drivers' licenses in Cameroon at the dawn of 2009.

In Sept 2009, he relocated to Belgium to pursue a masters degree at Universiteit Hasselt. At the Center of Statistics (CenStat), Universiteit Hasselt (2009-2011) he earned a cum laude MSc. in Statistics with specialization in Statistical Bioinformatics. His master thesis titled "Classification and class prediction for different chemical structures using gene expression data" led to the identification of a preliminary gene signature for screening compound toxicity in animal models. From 2012 to 2017 he worked on this PhD thesis at the Julius Center of Health Sciences and Primary Care, University Medical Center Utrecht, the Netherlands under the supervision of Prof. dr. ir. M. J. C. Eijkemans and Prof. dr. C. B. Roes. During this period, he also served as a statistical consultant at the Department of Viroscience, Erasmus Medical Center, Rotterdam, the Netherlands, within the Bioinformatics team headed by dr. A. C. Andeweg and dr. H.-J. van den Ham. Throughout his PhD., he has given lectures and supervised computer labs and tutorial secessions of several statistical courses at bachelor and master levels.

At present, Victor L. Jong is working as an assistant professor at the Julius Center of Health Sciences and Primary Care, University Medical Center Utrecht, the Netherlands, with research focus on methodological aspects relating to prediction in high-dimensional (big) data.