Development of biomarkers in non-clinical and clinical studies

Theophile Bigirumurame

Promoter: Prof. dr. Ziv Shkedy

Acknowledgements

I would like to express my special appreciation and thanks to many people, that I have met during my PhD, but this acknowledgement would never cover all people I want to thank. So the list is incomplete.

Firstly, I would like to express my sincere gratitude to my PhD supervisor Prof. Dr Ziv Shkedy, you have been a tremendous mentor for me. I would like to thank you for encouraging my research and for allowing me to grow as a research scientist. Your advice on both research as well as on my career have been priceless. Thank you for your availability to set meeting whenever I felt it was needed. I will never forget ”Theophile let us meet now”. I would like to thank my jury members for all their efforts, and for the helpful comments, which improved my dissertation. Thanks to my coauthors for their comments, suggestions and helpful discussions.

I would also like to thank my officemate Leandro, Leacky for the very interesting wide ranging discussions we had in the office, and for creating a conducive working environment. Thanks also to all the Censtat colleagues for creating such a friendly environment and for support in various forms. I would certainly not forget the secretarial Martine, Hilde and Chantal for all the assistance. I would like to thank Marc Thoelen, Karin Daniels, Theogene Havugimana, Jean Baptiste Bugingo, Barbara Swennen, Eugene (Musenyeri), Theophile Gapasi (Bazina), Evelyne Louis. To my research circle Marijke, Ewoud, Rudradev and Nolen and my travel buddies to Janssen during my first year, Sammy, Doreen, Leacky, Martin, Tanya and Nolen, you have been a source of intellectual discussions, conversations and lasting friendship, thanks!

A special thanks to my family. Words cannot express how grateful I am to my mother, and father, sisters and brothers for all of the sacrifices that you have made on my behalf. Reka mbyandike mu rurimi mushobora kubyumvamo neza. Ndashimira byimazeyo mbere na mbere ababyeyi banyibarutse. Ndabashimira kubwimpano y'ubuzima mwampaye, mukandera, mukankuza kugeza munyeretse inzira igana ishuri. Mwakoze ibishoboka byose sinabura ibyo nakeneraga byose. Ndabashimira kuba mwaranyeretse umubyeyi uruta abandi bose ari we Nyagasina Imana itanga byose. Ndashimira abavandimwe banjye bose. Sinarondora amazina yose ngo mbivemo, ariko ndabashimira ibyo mumfasha byose.

At the end I would like express appreciation to my beloved wife Germaine Uwimpuhwe who was always my support in the moments when there was no one to answer my queries. Ndakeka ibyo nakoraga byose utarabyumvaga neza, ariko kukugira mu buzima bwanjye ntacyo nabinganya. Wambereye, kandi uracyari, inshuti nziza, umufasha, umujyanama, unyitaho nkuko mama wanjye yabigenza. Iyo kode cg modeli zanganga gukonverija, niwowe nabibwiraga kandi ukamfasha kubyakira. Ndagushimira kubwo kuntega amatwi. Ndagushimira urukundo, ibyishimo ndeste n'umunezero wazanye mu buzima bwanjye. Tutarabana ntago nari nakamenye icyo ijambo ry'Imana rivuga ngo: "si byiza ko muntu aba wenyine, reka tumuremere umufasha (mbyanditse ntarebye muri Bibiliya nizereko ntinyuzemo)". Hashimwe Imana Rurema yaremye urukundo ikarubiba mu mitima yacu. Kandi ishimirwe kuba ikiduhaye umwuka w'abazima. Nshuti nziza, ndagukunda, unyihanganire imitoma iranshiranye.

Theophile Bigirumurame
Diepenbeek, 23 september 2016

List of Publications

The materials presented in the dissertation are based on the following publications:

Manuscripts

Bigirumurame ,T. ; Pushpike T. J., Louis, E. ; Liene, B.; Karolien, V. ; de Jonge, E. , Thomeer, M. ; Mesotten, L. ; Stinissen, P.; Vanderzande, D., Shkedy, Z., Adetayo, K. ; Adriaensens, P. Analysis and statistical validation of ^{1}H-NMR Metabolite Profiles for Early Detection of Breast Cancer. *Journal of Data Mining and Bioinformatics.* (submitted).

Bigirumurame T. ; Louis, E. ; Adriaensens, P. ; Mesotten, L.; Vanhove , K. ; Shkedy, Z. ; Thomeer, M. Risk models for lung cancer screening with low-dose computed tomography: the added predictive value of including NMR metabolic phenotype data. *Cancer Prevention Research.* (submitted).

Bigirumurame, T., Perualila T. N. J., Adetayo, K., Shkedy, Z. Joint modeling of bioassay data and genes expression in drug discovery experiments: A supervised principal component analysis approach (Working paper).

Louis, E., Adriaensens, P., Wanda, G.; **Bigirumurame, T.** Baeten, K.; Vanhove, K.; Vandeurzen, K.; Darquennes, K.; Vansteenkiste, J.; Christophe, C.; Shkedy, Z.; Mesotten, L. ; Thomeer, M. (2016) Detection of lung cancer via metabolic changes measured in blood plasma. *Journal of Thoracic Oncology*, **11** (4), 516–523.

Book Chapters

Bigirumurame, T., Shkedy, Z., Burzykowski, T. (2016) Surrogacy validation using SAS software. In: Molenberghs, G., Alonso, A., Van der Elst, W., **Bigirumurame, T.**, Buyse, M., Burzykowski, T., Shkedy, Z. *Applied Surrogate Endpoint Evaluation Methods with SAS and R. Chapman & Hall/CRC Biostatistics Series.* (To be published).

Bigirumurame, T., Shkedy, Z. (2016) Surrogacy in Cloud computing. In: Molenberghs, G., Alonso, A., Van der Elst, W., **Bigirumurame, T.**, Buyse, M., Burzykowski, T., Shkedy, Z. *Applied Surrogate Endpoint Evaluation Methods with SAS and R. Chapman & Hall/CRC Biostatistics Series.* (To be published).

Perualila, J.N., Shkedy, Z., Sengupta, R., **Bigirumurame, T.**, Bijnens, L., Talloen, W., Verbist, B., Göhlmann, H.W.H, QSATR consortium, and Adetayo Kasim (2016) High-dimensional biomarker in drug discovery: QSTAR framework. In: Molenberghs, G., Alonso, A., Van der Elst, W., **Bigirumurame, T.**, Buyse, M., Burzykowski, T., Shkedy, Z. *Applied Surrogate Endpoint Evaluation Methods with SAS and R. Chapman & Hall/CRC Biostatistics Series.* (To be published).

Conference Proceedings

Bigirumurame, T., Perualila, T. N. J., Shkedy, Z., Adetayo, K. (2015) Integrated analysis of multi-source data in drug discovery experiments using structural equation models. In: *Kepler, Johannes (Ed.) Proceedings of the 30th International Workshop on Statistical Modelling, 39–42.*

Contents

Chapter 1

Introduction

The work presented in this dissertation is focused on the development of biomarkers in non-clinical and clinical studies. A biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes or pharmacological responses to therapeutic or other interventions (Biomarkers Definitions Working Group, 2001). Recent developments in biotechnology make it possible to use molecular biomarkers of exposure, toxicity, disease risk, disease status and response to therapy (Clarke *et al.*, 2004; Baek *et al.*, 2009; Amaratunga *et al.*, 2014; Göhlmann and Talloen, 2009). The high-throughput genomic, proteomic and metabolic data are characterized by a large number of variables with relatively small number of samples. In the drug discovery and development process, this technology is used to monitor simultaneously the activity of thousands genes and their response to certain experimental conditions. The identified biomarkers could then be used as diagnostic tests or decision making tools (Clarke *et al.*, 2004; Amaratunga *et al.*, 2014; Baek *et al.*, 2009; Göhlmann and Talloen, 2009; Ambroise and McLachlan, 2002).

The research presented in this dissertation is mainly based on the use of existing methods and models and the novel aspects are related to the introduction of advanced, up-to-date and sophisticated methods in metabolomics and transcriptomics in drug discovery.

The first part of the dissertation is focused on metabolic biomarkers that are used to improve existing diagnostic procedures, screening tools and risk models to identify patients with high risk of developing lung cancer and breast cancer.

In Chapter 2, a brief introduction about the usage of metabolic data for the detection of cancer is given and two case studies in lung and breast cancer are introduced.

In Chapter 3, we present an analysis for the validation of proton nuclear magnetic resonance (^{1}H-NMR) metabolite profiles for early detection of breast cancer. The uniqueness of integration regions (IR) signature for classification is presented as well. Chapter 4 is focused on lung cancer. We present a screening tool based on metabolic changes measured in blood plasma for the detection of lung cancer. In Chapter 5 we focus on the question how to test additive value of metabolic data in risk models for lung cancer that include clinical covariates. In other words, we focus on the question whether use of metabolic data in addition to clinical covariates does improve the accuracy of prediction of the risk for cancer.

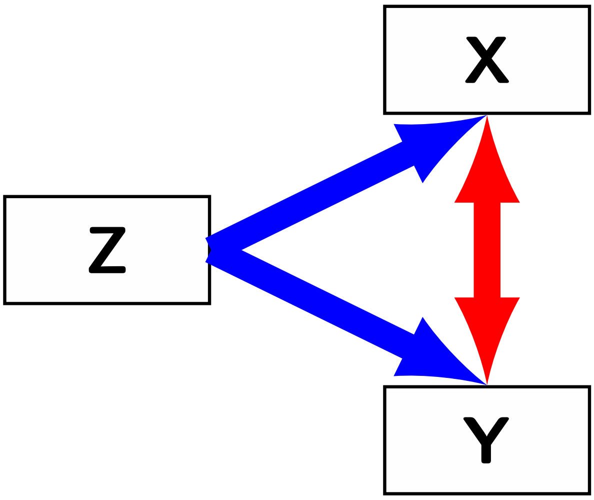


Figure 1.1: *The relationship between a condition (Z), a biomarker (X) and a primary endpoint (Y). The triplet (Y,X,Z) is the basic data structure used in the second and the third part of the dissertation.*

Figure 1.1 illustrates the data structure in the second and the third part of the dissertation. The variable Z represents the treatment (or condition) under study while the variable X is a potential biomarker for the primary endpoint Y. The blue arrows represent the condition effects on both biomarker and primary endpoint whereas the red arrow represents the association between the two endpoints after correcting for the condition effects.

The second part of the dissertation focuses on integrated analysis of multi-sources data in the drug discovery experiments. Chapter 6 introduces the use of transcriptomic data in the drug discovery studies and presents the datasets used for illustration. Joint modeling approach of the bioassay data and gene expression data using supervised principal component analysis, lasso and elastic net approaches are presented in Chapter 7. In Chapter 8, we discuss the use of structural equations models for the identification of causal structures in high dimensional data in the drug discovery.

While the first two parts of the dissertation are focused on the development of high dimensional biomarkers, the third part is focused on software development that can be used for the validation of surrogate endpoints in randomized clinical trials.

In Chapter 9, we give a short introduction about the use of surrogate endpoints in randomized clinical trials and present the datasets used for illustration. In Chapter 10, we present a set of SAS macros which can be used to evaluate surrogate endpoints in different settings. R Shiny application is presented in Chapter 11. Chapter 12 offers concluding remarks and a perspective for future research.

Part I

Development of Metabolic Biomarkers for Cancer

Chapter 2

Metabolic Cancer Studies

2.1 Metabolic Studies for Cancer Diagnostic

In this chapter, we introduce the use of metabolic data in the diagnostic of cancer. Especially, lung and breast cancer are considered. The datasets used, for illustration in Chapter 3 and 4, are introduced as well.

2.1.1 Lung Cancer

Lung cancer is one of the most common malignancies worldwide. It is the leading cause of cancer death in North America and worldwide (Tammemagi *et al.*, 2011). Ferlay *et al.* (2015) reported that 1.82 million new lung cancer (LC) cases were diagnosed in 2012 and 1.6 million LC-related deaths were recorded. Parkin *et al.* (2005) estimated that 1.35 million new lung cancer (LC) cases and 1.18 million LC-related deaths occur every year. Lung cancer is mostly diagnosed at an advanced disease stage, when curative treatments are limited, and this is attributed to the lack of symptoms during the early phases. As a result, the relative five-years survival rate is then very poor, ranging between 5 to 10% worldwide (Boyle *et al.*, 2008).

Screening for lung cancer at an early stage before a patient develops clinical symptoms and when the treatment is most effective should benefit the patient by increasing his/her quality of life and life expectancy (Bourzac, 2014; Shlomi *et al.*, 2014; Wood *et al.*, 2012). An appropriate screening test should be cost-effective. According to Field *et al.* (2013a,b) the benefit-risk balance is maximized when high-risk target population is selected for screening. Robust methods for risk prediction are essential to accurately select individuals with high risk of developing lung cancer for screening.

Currently, risk prediction models include mainly epidemiological and clinical risk factors such as gender, age, and smoking history (Bach *et al.*, 2003; Spitz *et al.*, 2007; Cassidy *et al.*, 2007). Eighty-five percent of the lung cancers are non-small cell lung cancer (NSCLC) and fifteen percent are small cell lung cancer (SCLC). The latter are aggressive malignancy, fast-growing and spread much more quickly (Cuffe *et al.*, 2011; Wood *et al.*, 2012).

Various techniques are available to screen lung cancer. They include chest radiography (CXR), sputum cytology and low-dose computed tomography (LDCT, Brett, 1968; Melamed *et al.*, 1984; Larke *et al.*, 2011). Brett (1968) failed to demonstrate the beneficial effect of CXR screening alone or in combination with sputum cytology on lung cancer mortality. Additionally, CXR does not allow to detect lung tumors smaller than 2 cm (Sone *et al.*, 2000). The LDCT allows to detect lung cancers at smaller tumors and at earlier stage compared to conventional CXR (Henschke *et al.*, 1999; Sone *et al.*, 1998; National Lung Screening Trial Research Team, 2011b). However, these screening methods are characterized by high false positive rates. This results in emotional stress, needless financial cost, and increased risk for healthy people (Humphrey *et al.*, 2004). Healthy people might be exposed to unnecessary radiation, biopsies and surgical procedures which are associated with higher morbidity and mortality rates (Shlomi *et al.*, 2014; Tammemagi and Lam, 2014).

Because of high false positive rate of LDCT, there is a growing interest in improving the accuracy of current risk models by incorporating lung cancer related biomarker for the selection of high-risk individuals eligible for LDCT screening (Raji *et al.*, 2010; Spitz *et al.*, 2008). Blood samples can be obtained non-invasively and without risks for patients (Smolinska *et al.*, 2012; Tsay *et al.*, 2014; Mamas *et al.*, 2011). In some studies DNA repairs markers (Spitz *et al.*, 2008) and genetic factors (Raji *et al.*, 2010) were added to models containing clinical risk factors. This resulted in prediction improvement.

2.1.2 Breast Cancer

Breast cancer is the most commonly diagnosed cancer in women and the leading cause of cancer deaths in women worldwide. In 2008, 1.38 million women were diagnosed with breast cancer worldwide, accounting for approximately 23% of all cancers diagnosed in women (Peter and Bernard, 2008). The incidence of breast cancer is generally higher in developed countries as compared to developing countries, but due to differences in population size, the number of cases becomes roughly equal (690.000 for both developed and developing regions). Gender and age are the most important risk factors for the disease. Breast cancer is diagnosed 100 times more in women than in men and the majority of advanced breast cancers are diagnosed in women older than 50 years. Nevertheless, 89% of women survive five years after the diagnosis in western countries. This is mainly due to the development of more efficient detection and treatment methods (Parkin *et al.*, 1999; Howlader *et al.*, 2012). Currently, several complementary techniques are available for diagnosis and surveillance of breast cancer. They include mammography, physical examination, ultrasonography, MRI and blood-based biomarker tests. Among all these techniques, mammography is still the gold standard (Sickles, 1991; Gartlehner *et al.*, 2013).

Several new approaches are developed in view of early detection, progression follow-up and therapy monitoring of breast cancer. They are primarily based on the detection of blood-based tumor markers and/or genetic profiling (Ebeling *et al.*, 2002; Asiago *et al.*, 2010; Duffy, 2006).

In this part of the dissertation, we focus on metabolomic biomarkers obtained from blood samples. In particular, ^{1}H-NMR based metabolomics data is used. Metabolic phenotype is the end result of genetic and environmental (diet, physical activity) influences and provides a readout of the metabolic state of an individual (Holmes *et al.*, 2008). Note that metabolic phenotype is not only affected by disease processes, but also by confounding factors, such as age, gender, ethnicity, diet, drug administration and lifestyle (Holmes *et al.*, 2008; Kochhar *et al.*, 2006; Lenz *et al.*, 2004).

Cancer cells have to reorganize their metabolism in order to meet their abnormal nutrients demand to support growth, proliferation and survival under suboptimal conditions (Kroemer and Pouyssegur, 2008; Cantor and Sabatini, 2012). ^{1}H-NMR based metabolomics have great potential in cancer diagnosis, prediction of therapy response and development of new therapies since cancer cells are characterized by profound metabolic alterations (Kroemer and Pouyssegur, 2008; Sciacovelli *et al.*, 2014).

2.2 Case Studies

2.2.1 Breast Cancer Dataset

The Breast cancer dataset contains information on about 139 subjects who are clinically free of breast cancer (termed as healthy controls; HC) and 161 patients who are newly diagnosed with breast cancer (BC). Since age is a major risk factor for breast cancer (McPherson *et al.*, 2000; Pike *et al.*, 1993), the subject population was defined to reflect this aspect and indeed the median age of the HC and BC groups are 56 and 61 years, respectively. Their metabolite patterns are quantitatively profiled by proton nuclear magnetic resonance (^{1}H-NMR) spectroscopy. The dataset contains two types of variables. A vector $\mathbf{Y}_{n\times 1}$ containing information about the individual labeling (BC or HC) and a matrix $\mathbf{X}_{n\times m}$ containing the individual metabolic profiles.

$$\mathbf{Y}_{n\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}, \quad \mathbf{X}_{n\times m} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}.$$

Here, Y_i=1 if the subject is a breast cancer patient and zero otherwise, x_{ij} is the metabolite readout of the jth metabolite of the ith subject. Note that we use the terminology metabolite for an integration region readout. This will be clarified later in this section.

Blood Sampling and Processing

Venous blood samples (10 ml) were collected in lithium coated tubes and stored at 4°C within 5 to 10 minutes after collection. Within 6 hours after collection, blood samples were transported on crushed ice (4°C) and centrifuged at room temperature (1600 g, 15 minutes). Subsequently, plasma aliquots of 500μl were transferred into sterile cryovials and stored at -80°C until NMR examination. All samples were analyzed within 3 months. At the time of the ^{1}H-NMR analysis, plasma samples were thawed and homogenized using a vortex mixer. After centrifugation at 4°C (13000 g, 4 minutes), the samples were further diluted to 800μl with deuterium oxide (D_2O, 99.9%, Cambridge Isotope Laboratories Inc, Andover, USA) containing trimethylsilyl-2,2,3,3-tetradeuteropropionic acid (TSP, 3.6mg/12ml, Cambridge Isotope Laboratories Inc, Andover, USA) as a chemical shift reference (0.015 ppm) (Beckonert *et al.*, 2007). Finally, the samples were transferred into 5 mm NMR tubes and analyzed.

^{1}H-NMR Spectroscopic Analyses

Proton NMR spectra were recorded at 21.2°C on a 400 MHz NMR spectrometer (Varian/ Agilent, Nuclear Magnetic Resonance Instruments, USA) with a magnetic field strength of 9.4 Tesla. Slightly T2-weighted spectra were acquired using the Carr-Purcell-Meiboom-Gill (CPMG) pulse sequence to attenuate signals of macromolecules, such as proteins and polysaccharides (Meiboom and Gill, 1958; Hürlimann and Griffin, 2000). Additional water suppression allows optimal detection and quantification (integration) of the resonance signals of low molecular weight metabolites.

The proton NMR spectra were phased properly and divided into 98 fixed integration regions (IRs) based on the chemical shifts of the metabolites. After a baseline correction, the integrals of 96 IRs (all except this of remaining water and TSP) were normalized to the total integration value of all signals except these arising from water, TSP and the (huge) signals of glucose and fructose. These normalized values of the 96 IRs are referred to as var-1,..var-96. Note that the 96 IRs do represent less than 96 metabolites. This is due to the fact that most metabolites have protons in different chemical environments and so give rise to more than one signal in the ^{1}H-NMR spectrum. Moreover, proton spins are often J-coupled with other proton spins, giving rise to J-coupling patterns (doublet, triplet,...).

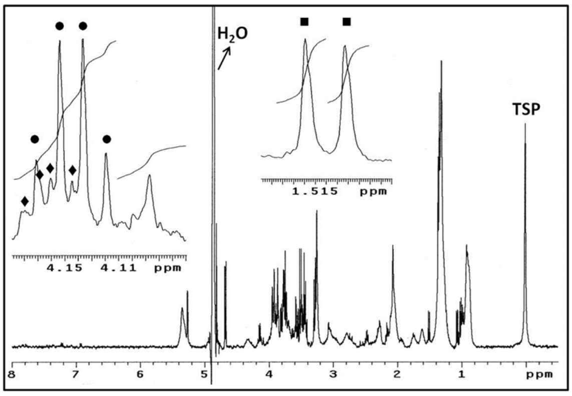


Figure 2.1: *A typical proton NMR spectrum of human plasma. Left inset: integration regions (IRs) between 4.188-4.112 ppm (Var-16) and 4.112-4.060 ppm (Var-17). The Var-16 IR shows the quadriplet (●) of lactate (CH3CHOHCOOH; J=6.9 Hz) superimposed on the double doublet (♦) of the H_α proton of proline (J=8.7 and 6.0 Hz). The Var-17 IR is composed of signals arising from protons of creatinine, fructose, inositol and tryptophan. Right inset: integration regions between 1.533-1.510 ppm (Var-82) and 1.510-1.491 ppm (Var-83), arising from the doublet (■; J=7.2 Hz) of alanine. Adapted from Bigirumurame et al. (2016).*

.

Figure 2.1 illustrates the quantification of the integration regions (Friebolin and Becconsall, 1993; Bovey *et al.*, 1988). The amino acid alanine, for example, shows a doublet around 1.51 ppm arising from the methyl protons (J-coupled to the methine proton with a coupling constant of J=7.2 Hz) and a quadriplet around 3.81 ppm arising from the methine protons (J-coupled to the methyl protons with the same coupling constant of J=7.2 Hz). The inset at the right shows the integration regions of the alanine doublet lines (■) between 1.533-1.510 ppm (Var-82) and 1.510-1.491 ppm (Var-83). Lactate on the other hand gives rise to a quadriplet (CH3CHOHCOOH; J=6.9 Hz) around 4.145 ppm and a doublet (CH3CHOHCOOH; J=6.9 Hz) around 1.350 ppm. The inset at the left shows the integration region of the lactate quadriplet (●) between 4.188-4.112 ppm (Var-16). Remark that the latter IR is superimposed on the double doublet signal (⧫) of the H_α proton of proline (J=8.7 and 6.0 Hz). The other proton J-patterns of proline appear around 2.38 ppm (H_β), 2.06 ppm (H_γ) and 3.38 ppm (H_δ). The experimental, normalized integration values of the 96 IRs are shown in Figure 2.2 for all (300) subjects included.

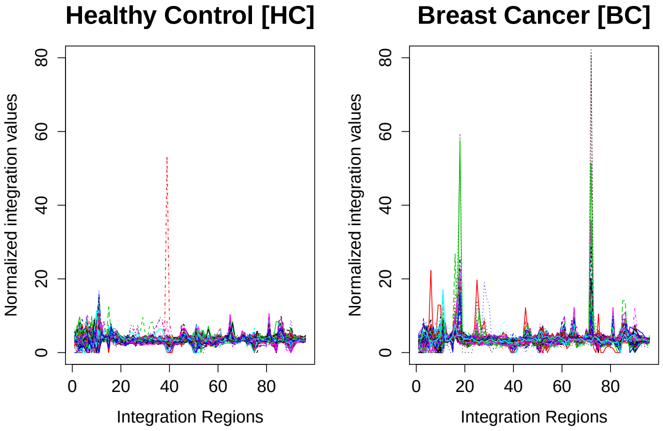


Figure 2.2: *Patient-specific normalized NMR integration values of the 96 integration regions (IRs). Left panel: 139 healthy controls. Right Panel: 161 breast cancer patients.*

2.2.2 Lung Cancer Dataset

For the lung cancer study, 357 patients were included in the Limburg Positron Emission Tomography center (273 patients from Hasselt, Belgium) and at the Department of Respiratory Medicine of University Hospital Leuven (84 patients from Leuven, Belgium) from March 2011 to June 2014. The diagnosis of lung cancer was confirmed by a pathological biopsy or by a clinician specialized in interpreting radiological and clinical lung cancer data. Clinical staging of the tumors was performed according to the *7th* edition of the tumor, node, metastasis (TNM) classification of malignant tumors (Goldstraw *et al.*, 2007). All controls (n=347) were patients with non-cancer diseases who were included at Ziekenhuis Oost-Limburg (ZOL, Genk, Belgium) between March 2012 and June 2014. Exclusion criteria were: (1) not fasted for at least

6 hours; (2) fasting blood glucose concentration ≥ 200 mg/dl; (3) medication intake on the morning of blood sampling and (4) treatment or history of cancer in the past 5 years. The study was conducted in accordance with the ethical rules of the Helsinki Declaration and Good Clinical Practice and was approved by the ethical committees of ZOL, Hasselt University (Hasselt, Belgium) and University Hospital Leuven. An elaborated discussion about the Blood sampling, sample preparation and NMR analysis is given in Louis *et al.* (2015a).

2.2.3 Evaluation of a Given Classifier

In Chapters 3, 4, and 5 we present several methods to classify patients according to their disease status (cancer/ healthy control). We evaluate the performance of the classification methods using different statistics presented in this section.
Consider a binary classification problem and let Y_i and $\hat{Y}_i$ be the true and the predicted status (class) of a subject, respectively

$$Y_i = \begin{cases} 1 & \text{observed class is cancer} \\ 0 & \text{observed class is control} \end{cases} \quad \text{and} \quad \hat{Y}_i = \begin{cases} 1 & \text{predicted class is cancer} \\ 0 & \text{predicted class is control} \end{cases}$$

The observed and the predicted classes are used to form a confusion matrix from which performance measures are computed. Table 2.1 shows such a matrix.

Table 2.1: *Confusion matrix.*

		Predicted status	
		1	**0**
Observed Status	**1**	True Positive (TP)	False Negative (FN)
	0	False Positive (FP)	True Negative (TN)

The following statistics can be computed and used to assess the performance of a given classifier.

- The misclassification error is the total number of mistakes committed using the classification procedure, that is:

$$MCE = \frac{FN + FP}{(FN + FP + TP + FP)}.$$

- The specificity of a classifier measures the proportion of negative cases (controls) that are correctly classified,

$$SPE = \frac{TN}{(TN + FP)}.$$

- The sensitivity of a classifier measures the proportion of positive cases (cancer for instance) that are correctly classified,

$$SEN = \frac{TP}{(TP + FN)}.$$

- The positive predictive value of a classifier measures the proportions of predicted positive cases (cancer subjects) that are true positive cases,

$$PPV = \frac{TP}{(TP + FP)}.$$

- The negative predictive value of a classifier measures the proportions of predicted negative cases (control subjects) that are true negative cases,

$$NPV = \frac{TN}{(TN + FN)}.$$

Chapter 3

Analysis and Statistical Validation of ^{1}H-NMR Metabolite Profiles for Early Detection of Breast Cancer

3.1 Introduction

Metabolomics is a rather recent methodology which encompasses the comprehensive and simultaneous quantitative analysis of small molecules within a given biological system, the so-called metabolites (Nicholson *et al.*, 1999; Pan and Raftery, 2007; Sitter *et al.*, 2010; Barderas *et al.*, 2011; Bain *et al.*, 2009). These metabolites constitute the end products of cellular metabolism and therefore changes in their concentrations may be regarded as functional signatures of the actual state of metabolism, namely the metabolic phenotype (O'Connell, 2012).
The metabolite profiles are most often derived by means of high-resolution and high-throughput analytical methods such as proton nuclear magnetic resonance spectroscopy (^{1}H-NMR) and mass spectrometry (Oakman *et al.*, 2011). The applications of metabolic phenotyping are very diverse and include biomarker identification, disease diagnosis and follow-up, improved insights in biochemical pathways etc. This explains the broad interest of biomedical, toxicological, nutritional and pharmaceutical research fields (Eliassen *et al.*, 2012; Cheng *et al.*, 2005).

In this chapter, several methods are explored to rank and select spectral integration regions and to construct a robust classifier to discriminate between breast cancer patients (BC) and healthy controls (HC). The statistical analysis and the conclusions are based on a data set of 300 subjects (161 BC and 139 HC) described in Section 2.2.1. Last but not least, possibility to define a unique but limited set of integration regions that can be used in the classifier is discussed in detail. This chapter is organized as follows: statistical methodology is given in Section 3.2. The ranking of the IR is given in Section 3.3. Classifiers validation of fixed sets of metabolites are described in Section 3.4. Uniqueness of the signature is studied in Section 3.5, followed by a discussion in Section 3.6.

3.2 Statistical Methodology

Several statistical testing procedures were considered to identify differentiating integration regions (IRs) between both groups. Furthermore, several classification methods were compared, including partial least squares discriminant analysis (Jansson *et al.*, 2009; Giskeødegård *et al.*, 2010; Bryan *et al.*, 2008). All statistical analyses were performed using the R statistical software package (version 3.2.1, R Development Core Team, 2015). In particular, the entire analysis for classification was performed using the R Bioconductor package CMA which allows to perform a wide variety of cross validations and classification methods (Slawski *et al.*, 2008). The scheme of the work flow is presented in Figure 3.1.

Firstly, a feature selection is done within the classification loop (Analysis 1 in Figure 3.1). Secondly, classifiers are built from a fixed list of IRs (Analysis 2 in Figure 3.1), and lastly, the uniqueness of the metabolic signature in the second analysis is investigated by building classifiers on a subset of the IRs which are not included in the second analysis (Analysis 3 in Figure 3.1).

3.2.1 Features Selection

Initially, several statistical tests were conducted (Wilcoxon signed-rank test, t-test, Lasso test, Elastic Net test and the Limma t-test) in order to rank the 96 IRs according to their test statistic (Efron and Tibshirani, 1997; Kohavi, 1995).

3.2.2 Cross Validation

In order to select the IRs features for the classifier, a 3-fold cross validation procedure was set up consisting of 1000 iterations. For each iteration of the 3-fold cross valida-

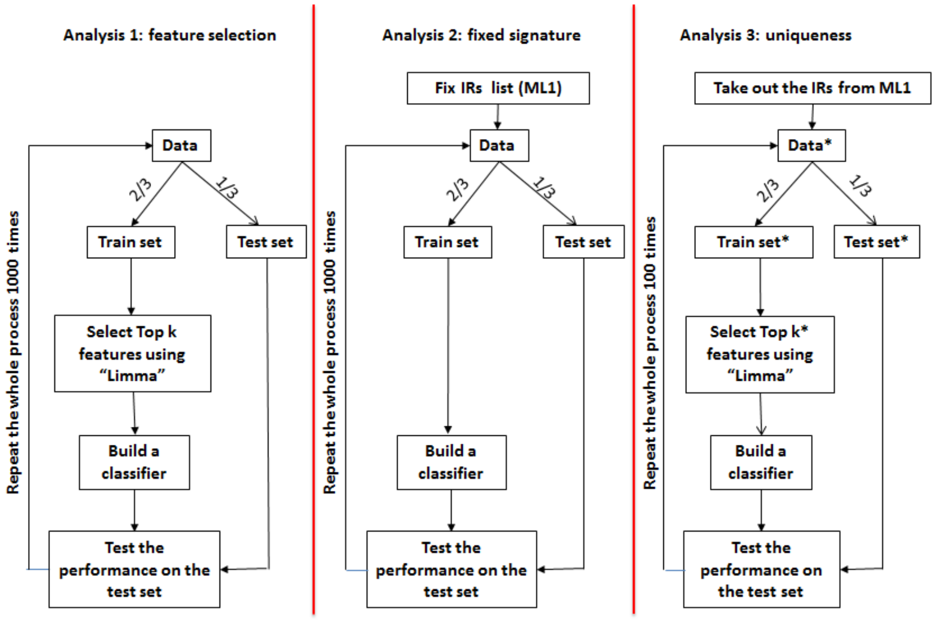

Analysis 1: feature selection
Data
2/3
1/3
Train set
Test set
Select Top k features using "Limma"
Build a classifier
Test the performance on the test set
Repeat the whole process 1000 times
Analysis 2: fixed signature
Fix IRs list (ML1)
Data
2/3
1/3
Train set
Test set
Build a classifier
Test the performance on the test set
Repeat the whole process 1000 times
Analysis 3: uniqueness
Take out the IRs from ML1
Data*
2/3
1/3
Train set*
Test set*
Select Top k* features using "Limma"
Build a classifier
Test the performance on the test set
Repeat the whole process 100 times

Figure 3.1: *Work flow used to build classifiers based on different IRs lists.*

tion, the total number of patients (300) is randomly split into a training set consisting of two thirds of the patients (200) and a test set consisting of the remaining one third of the patients (100). The ratio of BC/HC patients in these two groups is always equal to the BC/HC ratio of the total data set. Schematically, the cross validation scheme is shown in Figure A.1.

At each iteration of the cross validation loop, a top-k ($k = 2, 3, 4, \dots, 43$) list of IRs is selected based on the feature selection procedure. The selected IRs are then used to build classifiers by different classification methods. Finally, these trained classifiers are validated on the test set on the basis of estimated misclassification error (MCE), sensitivity and specificity.

3.2.3 Classification Methods

Several methods were considered (Statnikov *et al.*, 2005; Ambroise and McLachlan, 2002; Golub *et al.*, 1999; Jansson *et al.*, 2009), including linear discriminant analysis (LDA), diagonal discriminant analysis (DLDA, Guo *et al.*, 2007), partial least squares linear discriminant analysis (PLS-LDA, Boulesteix and Strimmer, 2007; Park and Hastie, 2007; Boulesteix, 2004), support vector machine (SVM, Guyon *et al.*, 2002), random forest (RF, Breiman, 2001), Fisher's discriminant analysis (FDA, Ripley, 1996) and quadratic discriminant analysis (QDA, MacLachlan, 1992). From a statistical point of view, these methods differ from one another by some assumptions and the way they construct the class prediction rule.

3.2.4 A Fixed Metabolic Signature

In a further step, we considered to fix two top-k IRs sets (ML1 and ML2) during the cross validation process. To that end, top-40 IRs were selected in two different ways, (1) the most frequently selected IRs during the cross validation by the Limma t-test (because this test is commonly used for feature selection; Smyth, 2005) and (2) the combination of IRs yielding the best overall performance with respect to lowest mean classification error (MCE) and highest specificity and sensitivity during the cross validation. This was done in order to remove the noisy IRs and to improve the classification (Slawski *et al.*, 2008; Amaratunga *et al.*, 2014).

Finally, we investigated the uniqueness of the best performing ML1 set of IRs by the 'leave one IR out' cross validation (Friedman *et al.*, 2001) and analyzed the classification results obtained by using only the remaining IRs.

3.3 Ranking of IRs and Classification (Analysis 1)

Initially, the 96 integration regions (IRs) of the ^{1}H-NMR spectra were ranked based on a Limma t-test (Smyth, 2004, 2005). The result for the entire dataset (without cross validations) is presented in Figure 3.2. It was found that 64 of the 96 IRs were significantly different between the HC and BC groups and that IR-16 (Var-16) was the top IR. Figure 3.3 depicts the volcano plot, the unadjusted and adjusted p-value (FDR, Benjamini and Hochberg, 1995) for all IRs. Note that the differential expression analysis is presented only to visualize the strength of the signal in the data. Feature selection for the classifier, as described in the next section, is not based on the results presented in this section.

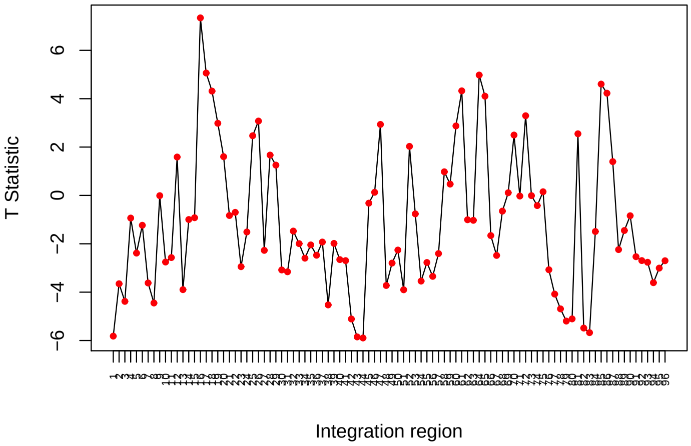


Figure 3.2: Integration regions (IRs) and their estimated Limma t-test statistics.

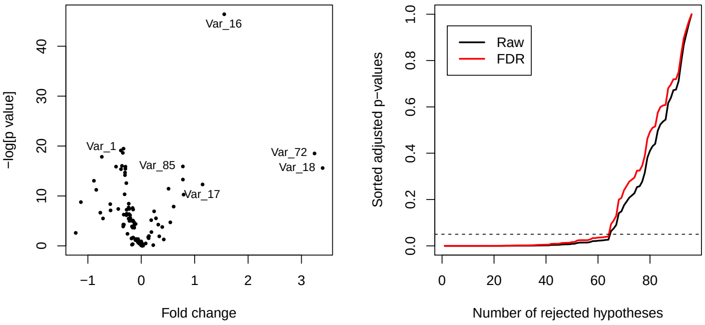


Figure 3.3: Left Panel: Volcano plot. Right Panel: Unadjusted and adjusted p-values for all IRs.

3.3.1 Ranking and Selection of Top IRs

Our first goal was to rank the IRs based on their frequency of selection in one of the following feature selection tests: the Wilcoxon signed-rank test, t-test, Lasso test, Elastic Net test and Limma t-test. Next and for each iteration, the (variable) training group is used to rank the IRs and to select top-k IRs (e.g. a top-3, a top-8, a top-43,. . .) according to their p-values obtained by one of the statistical tests mentioned above. At the end of the iteration procedure, the frequency of selection in top-k lists of significance was obtained for each IR. For example, Table 3.1 shows the lists of IRs most frequently selected in a top-15 based signature on the statistical tests mentioned above. Note that these lists have several IRs in common. For instance, 12 out of 15 IRs have been selected in the top-15 list by the Wilcoxon signed-rank test as well as by the t-test and the Limma t-test, namely: 1, 8, 16, 42, 43, 44, 78, 79, 80, 82, 83 and 85. The t-test and Limma t-test both select the same IRs as top-15.

Table 3.1: *Lists of IRs most frequently selected in a top-15 by applying the Wilcoxon signed-rank test, t-test, lasso test, elastic net test and the Limma t-test in the 3-fold cross validation procedure. Note that for each iteration, the HC and BC groups are compared by the statistical test and that the IRs are ranked based on their test statistics.*

	wilcox.test	t.test	lasso	elasticnet	limma t-test
1	Var_16	Var_16	Var_58	Var_58	Var_16
2	Var_44	Var_43	Var_47	Var_64	Var_43
3	Var_43	Var_44	Var_64	Var_47	Var_44
4	Var_1	Var_83	Var_16	Var_94	Var_1
5	Var_72	Var_1	Var_83	Var_79	Var_83
6	Var_79	Var_82	Var_14	Var_14	Var_82
7	Var_83	Var_79	Var_94	Var_83	Var_79
8	Var_82	Var_42	Var_79	Var_26	Var_42
9	Var_85	Var_17	Var_19	Var_85	Var_17
10	Var_80	Var_80	Var_92	Var_92	Var_80
11	Var_18	Var_64	Var_26	Var_48	Var_64
12	Var_78	Var_8	Var_74	Var_16	Var_8
13	Var_13	Var_78	Var_85	Var_54	Var_78
14	Var_8	Var_61	Var_44	Var_19	Var_85
15	Var_42	Var_85	Var_54	Var_91	Var_61

For the remainder of this chapter, features selection is based on the Limma t-test. Figure 3.4 shows, as an example, the frequency of selection of the IRs in a top-3, top-8, top-15 and top-43, obtained by applying the Limma t-test in the 3-fold cross-validation procedure. An IR selected in a lower top list is also selected in all higher top lists. It can be noticed that the more extended is the top-k considered, the more different IRs become frequently selected. However, it is clear that certain IRs which have a higher tendency to be selected might differentiate better between the BC and HC groups.

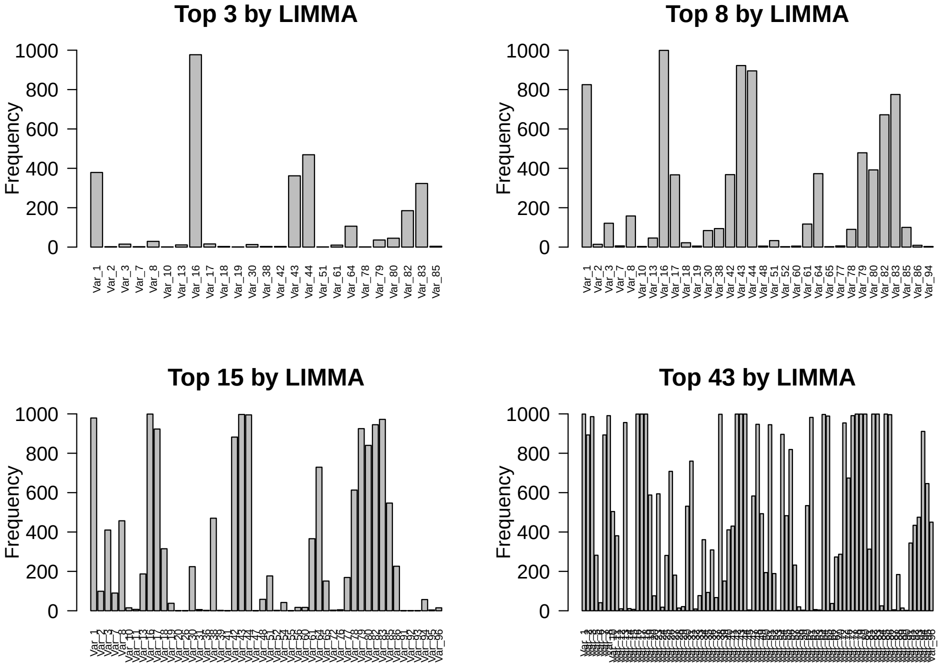


Figure 3.4: *Frequency of selection of IRs in a top-k list, obtained by applying the Limma t-test for the 1000 iterations of the 3-fold cross-validation procedure. For each of the 1000 iterations, IRs are ranked based on their Limma t-test significance and the frequency of selection in a top-k is determined. Remark that the top-k IRs can be different from iteration to iteration. Top left panel: k=3; Top right panel: k=8; Lower left panel: k=15 and Lower right panel: k=43.*

3.3.2 Evaluation of Different Classification Methods

During the 3-fold cross validation procedure, several classification methods were evaluated (Dudoit *et al.*, 2002; Statnikov *et al.*, 2005). The top-k IRs, selected for each iteration of the 3-fold cross-validation on the training group by the Limma t-test, were used to build 'top-k-based' classifiers by means of different classification methods. For each iteration, these trained classifiers were evaluated on the basis of misclassification error, sensitivity and specificity in the classifier test group (remaining 1/3 of the subjects). Note that, since the training and test groups are variable for each iteration of the 3-fold cross validation, so will be the selected top-k IRs and resulting 'top-k-based' classifiers.

In this study, the seven classification methods mentioned in Section 3.2.3 were performed and compared. Table 3.2 and Figure 3.5 present an overview of the median overall misclassification error (MCE) obtained by the seven different classification methods for six 'top-k-based' classifiers (k=3, 12, 20, 30, 40 and 43) after 1000 iterations. Figure 3.5 shows the median overall misclassification error with the corresponding 95% confidence intervals for different top k IRs.

Table 3.2: *Median overall misclassification error obtained by the seven different classification methods for six different 'top-k-based' classifiers (k=3, 12, 20, 30, 40 and 43; top-k IRs selected by the Limma t-test). The median misclassification error of BC as HC, sensitivity and specificity are presented in Table A.1, Table A.2 and Table A.3 in the appendix.*

Method	3	12	20	30	40	43
LDA	0.30	0.28	0.27	0.26	0.25	0.25
DLDA	0.31	0.32	0.31	0.31	0.32	0.32
FDA	0.31	0.28	0.27	0.26	0.25	0.25
PLSLDA	0.29	0.28	0.30	0.29	0.29	0.29
SVM	0.26	0.25	0.25	0.24	0.24	0.24
RF	0.29	0.25	0.23	0.23	0.22	0.22
QDA	0.33	0.31	0.28	0.27	0.27	0.28

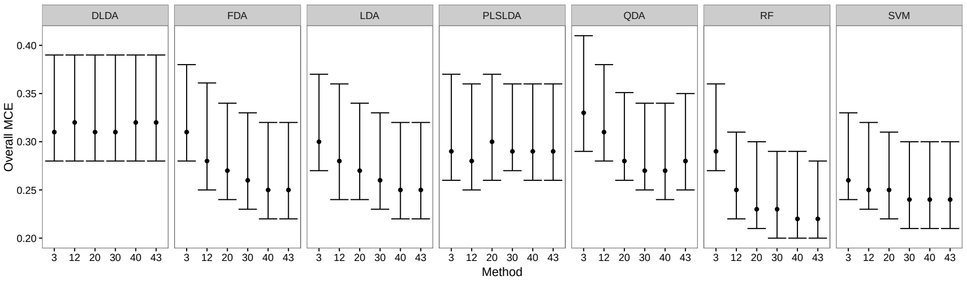


Figure 3.5: *The median overall misclassification error with corresponding confidence intervals per classification method and top k list .*

Figure 3.6 presents the median of overall MCE and MCE of BC as HC, as well as the median sensitivity and specificity for the different classification methods and several top-k IRs. Concerning the classification methods, it can be noticed that although most of them perform more or less similarly, the SVM, RF and FDA methods perform slightly better for this dataset. Based on the overall performance, i.e. yielding the highest sensitivity and specificity and the lowest misclassification errors, it was decided to proceed with the SVM classification method. The distribution of the misclassification errors, sensitivity and specificity, obtained by applying the SVM classification method during the cross validation, are further visualized in Figure 3.7 for several top-k IRs. For the (variable) top-40 IRs, the median overall MCE, MCE of BC as HC, sensitivity and specificity obtained by the SVM classification method

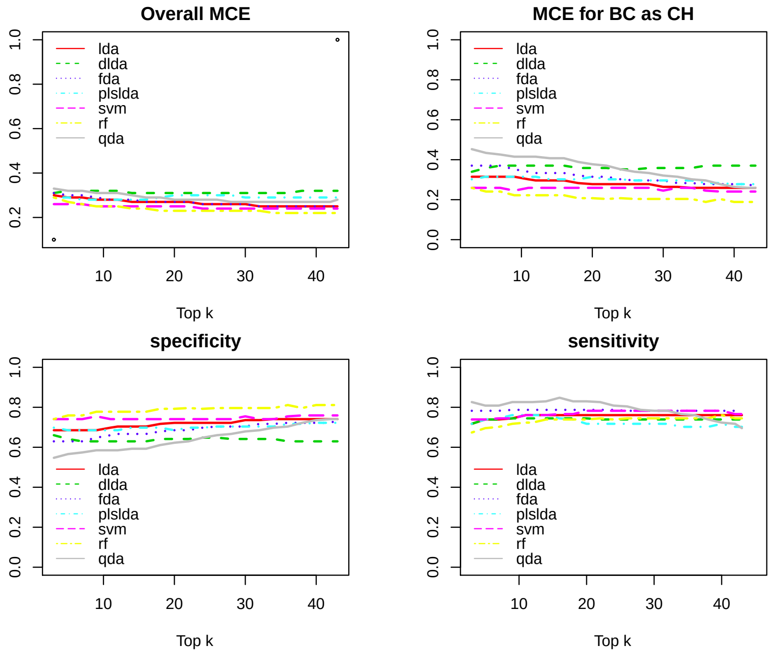


__Figure 3.6:__ Overview of the median overall misclassification error, median misclassification error of BC as HC, median specificity and median sensitivity obtained for several 'top-k-based' classifiers (selection method Limma t-test; k = 3 5 7 9 11 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 43) by the different classification methods.

is 0.24, 0.22, 0.78 and 0.76, respectively.

3.4 Classifier Validation for Fixed Sets of IRs (Analysis 2)

In a next step, we first fixed selected sets of IRs and applied the classification method using these fixed IRs lists. The idea here is to take out the noisy IRs in order to reduce the misclassification errors and to improve the sensitivity and specificity. Therefore, we considered two fixed sets of top-40 IRs, the ML1 and ML2 lists. For the ML1 list, we considered the top-40 IRs with the highest frequency of selection in the 3-fold cross validations (see Section 3.2.2). This set of top-40 IRs is referred to as the ML1 list. Figure 3.8 shows the Limma t-test statistics (on the total dataset without cross validation) versus the frequency of selection in the cross validations. This figure clarifies that IRs with better Limma t-test statistics have a higher chance to be selected as top during the 3-fold cross validations.

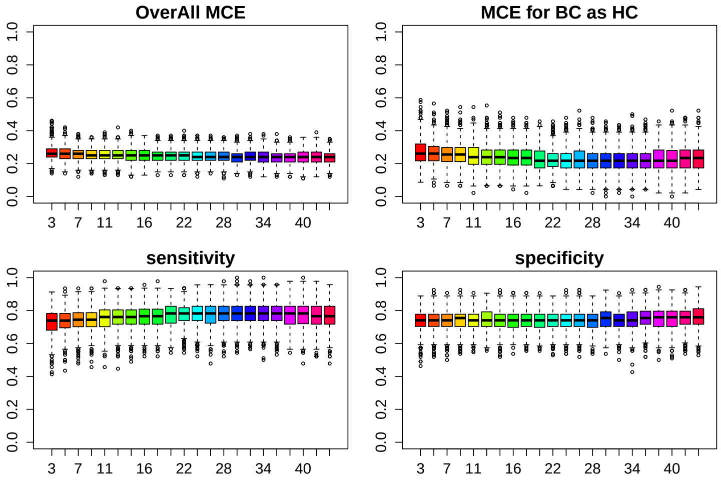


Figure 3.7: *Distribution of the 3-fold cross validation results for several 'top-k-based' classifiers (selection method Limma t-test; k = 3 5 7 9 11 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 43) built by the SVM classification method. The bars indicate the median overall misclassification error, median misclassification error of BC as HC, median specificity and median sensitivity. Remark that the top-k IRs are not fixed but variable from iteration to iteration.*

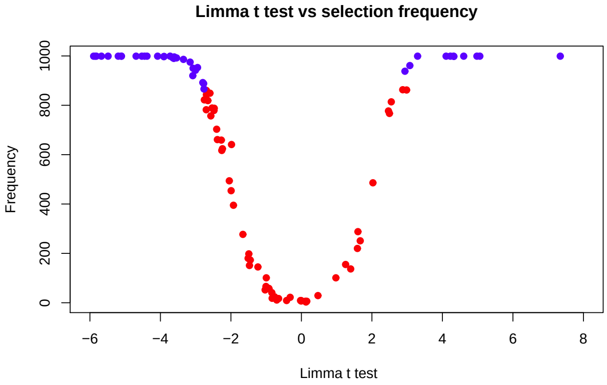


Figure 3.8: *Limma t-test statistics versus the frequency of selection in the 3-fold cross validations. The top-40 IRs are indicated in blue and their corresponding indices are: 1 2 3 7 8 13 16 17 18 23 26 30 31 38 42 43 44 47 48 49 51 54 55 56 61 64 65 72 76 77 78 79 80 82 83 85 86 93 94 95. This set of top-40 IRs is referred to as the ML1 list.*

For the ML2 list, we considered the 'top-40-based' SVM classifier yielding the best overall performance on the criteria considered, being the overall misclassification error, misclassification error of BC as HC, specificity and sensitivity during the 3-fold cross validations (see Section 3.3.2). The classification results obtained by this classifier are indicated by the red bullets in Figure 3.9. This set of top-40 IRs is referred to as the ML2 list. It should be noticed that the ML1 and ML2 lists have 32 IRs out of the 40 IRs in common.

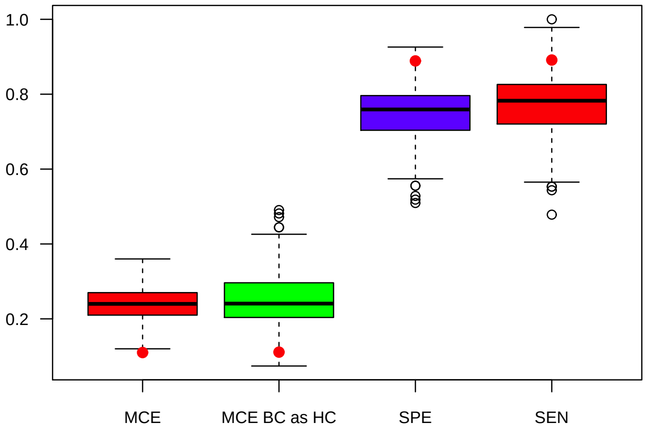


Figure 3.9: *Distribution of the 3-fold cross validation results for the 'top-40-based' classifiers (selection method Limma t-test) built by the SVM classification method. The bars indicate the median overall MCE, median MCE of BC as HC, median specificity and median sensitivity. The red bullets indicate the classifier with the best overall performance (Overall MCE=0.11, MCE of BC as HC=0.11, SEN=0.89, SPE=0.89) and consists of IRs with indices: 1,2,3,5,7,8,13,16,17,18,19,20,23,25,26,38,41,42,43,44,48,49,51,54,56,60,61,64,65, 70,72,77,78,79,80,82,83,85,86,96. This set of top-40 IRs is referred to as the ML2 list.*

3.4.1 Fixed Set of IRs Based on Highest Frequency of Selection: ML1

First, the top-40 IRs with the highest frequency of selection in the 3-fold cross validations (see Section 3.4) were selected. This set of top-40 IRs was fixed and referred to as ML1 (IRs: 1, 2, 3, 7, 8, 13, 16, 17, 18, 23, 26, 30, 31, 38, 42, 43, 44, 47, 48, 49, 51, 54, 55, 56, 61, 64, 65, 72, 76, 77, 78, 79, 80, 82, 83, 85, 86, 93, 94, 95). In a next step, a 3-fold cross validation was performed to evaluate the performance of ’top-ML1-based’ classifiers built by the different classification methods mentioned before. The distribution and median overall misclassification error, misclassification error of BC as HC, sensitivity and specificity obtained by the different classification methods are shown in Figure 3.10.

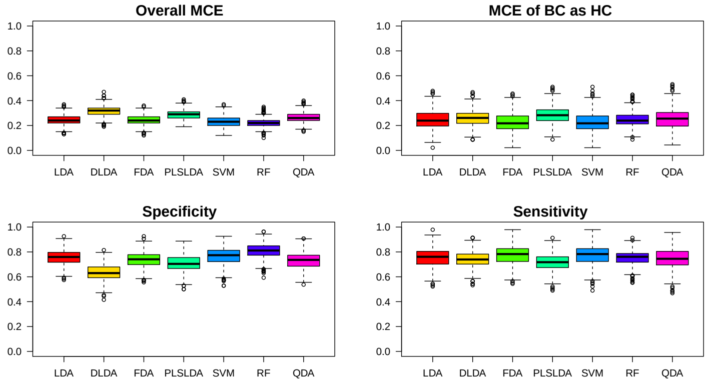


Figure 3.10: *Distribution of the 3-fold cross validation results for the ’top-ML1-based’ classifiers built by the different classification methods. The bars indicate the median overall misclassification error, median misclassification error of BC as HC, median specificity and median sensitivity.*

Although most of the classification methods perform more or less the same, the SVM and RF methods result in a slightly better overall performance for this dataset. The left panel of Table 3.3 presents the median misclassification errors, sensitivity and specificity of the ’top-ML1-based’ classifiers built by the SVM classification method. By comparing the performance results shown in Table 3.3 with those in Tables 3.2, A.1, A.2 and A.3, it is clear that removal of the noisy IRs results in a slightly further improvement. Receiver Operating Characteristic (ROC) curves are presented in the left panel of Figure 3.12 to visualize the sensitivity versus specificity performance of this SVM classifier during the iterations of the 3-fold cross validation procedure.

3.4.2 Fixed Set of IRs Based on Best Overall Performance: ML2

In Section 3.3.2 ’top-40-based’ SVM classifiers were constructed and evaluated during the 3-fold cross validation (see Figure 3.7; k = 40). The results of the classifier yielding the best overall performance (lowest MCE, 0.11 and highest specificity, 0.89 and sensitivity, 0.89) was shown in Figure 3.9 (red bullets). This set of top-40 IRs was fixed and referred to as ML2 (IRs: 1, 2, 3, 5, 7, 8, 13, 16, 17, 18, 19, 20, 23, 25, 26, 38, 41, 42, 43, 44, 48, 49, 51, 54, 56, 60, 61, 64, 65, 70, 72, 77, 78, 79, 80, 82, 83, 85, 86, 96). In this section, we further evaluate the performance of classifiers built with this fixed set of IRs and different classification methods in a 3-fold cross validation. The median overall misclassification error, misclassification error of BC as HC and sensitivity and specificity obtained by the different classification methods are shown in Figure 3.11.

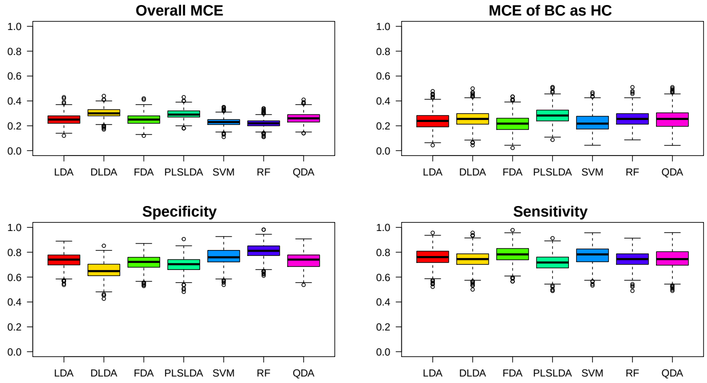


Figure 3.11: *Distribution of the 3-fold cross validation results for the 'top-ML2-based' classifiers built by the different classification methods. The bars indicate the median overall misclassification error, median misclassification error of BC as HC, median specificity and median sensitivity.*

Although most of the classification methods perform more or less the same, the SVM and RF methods result in a slightly better overall performance for this dataset. The right panel of Table 3.3 presents the median misclassification errors, sensitivity and specificity of the "top-ML2-based" classifiers built by the SVM classification method. Receiver operating characteristic curves are presented in the right panel of Figure 3.12 to visualize the sensitivity versus specificity performance of the SVM classifier during the iterations of the 3-fold cross validation procedure.

The results presented here demonstrate that the two "top-ML-based" SVM classifier results in comparable performances. As a conclusion, we can state that both top ML1 and top ML2 lists of integration regions allow an efficient classification of patients between the BC and HC groups, and this by using classification methods different from the most often used OPLS-DA method.

Table 3.3: *Median misclassification errors, sensitivity and specificity for the 'top-ML1-based' (left panel) and 'top-ML2-based' (right panel) SVM classifiers.*

	ML1			ML2		
	Lower 95% C.I	Median	Upper 95% C.I	Lower 95% C.I	Median	Upper 95% C.I
Overall MCE	0.20	0.23	0.31	0.21	0.23	0.30
MCE BC as HC	0.17	0.22	0.37	0.17	0.22	0.37
Specificity	0.72	0.77	0.89	0.72	0.76	0.87
Sensitivity	0.72	0.78	0.91	0.72	0.78	0.91

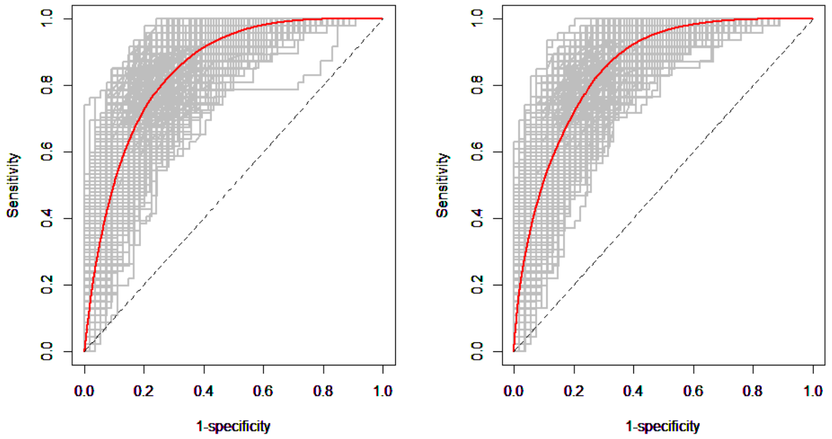


Figure 3.12: *ROC curves showing the performance of the 'top-ML1-based' (left panel) and 'top-ML2-based' (right panel) SVM classifier at each iteration of the 3-fold cross validation (gray colored curves). The median ROC curves are indicated by the red colored curves.*

3.5 Uniqueness of a Fixed IR Signature for Group Classification (Analysis 3)

Based on the results of Section 3.4.1, the ML1 set of IRs was chosen for further evaluation toward its uniqueness in classifying the BC and HC patients. Our first goal was to perform 3-fold cross validations while leaving out one of the IRs of the ML1 list. This is presented for the SVM classification method but other methods were considered as well. The underlying goal of this particular analysis is to check whether a certain IR, if left out, exerts a substantial impact on the overall classification results. For instance, if a substantial decrease in sensitivity and specificity and a substantial increase in misclassification errors is observed upon leaving out a certain IR, it highlights the importance of this particular IR in the list.

Our second goal was to judge whether or not we could achieve the same overall performance by means of a classifier built with the remaining 56 IRs (96-40), so by ignoring the IRs of the ML1 list. All the above was investigated in order to find out whether or not a unique, well performing classifier could be defined based on only a limited number of IRs. In principle, this means that it should not be possible to find another combination of IRs that performs as good as the ML1 set.

3.5.1 Leave-one-out IR Analysis

Classification results based on the 'leave-one-out IR' cross validations (LOOCV) are shown in Figure 3.13 and 3.14. These analyses were performed on the complete top-40 ML1 set, as well as on fixed ML1 subsets of 3, 12, 30 and 40 IRs (based on the frequency of selection as described in Section 3.4). For the top-3 IR set for example, this means that we left out one IR at a time and performed the 3-fold cross validation with the remaining two IRs. Figure 3.13 and 3.14 show that the results improve when more IRs are used in the classifiers.

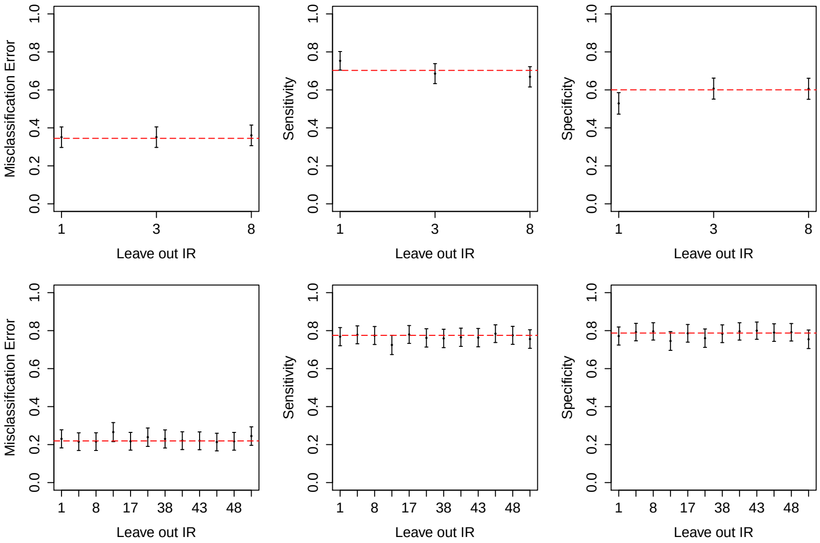


__Figure 3.13:__ SVM-based classification results obtained by the LOOCV (leave-one-out IR cross validations) for the top-3 IRs (top row) and the top-12 IRs (bottom row). The red dashed horizontal lines indicate the median 3-fold cross validation results obtained by using all 3 (top row) and 12 (bottom row) fixed IRs. Top-3 IRs list include IRs with indices:1, 3, 8. Top-12 IRs list include IRs with indices: 1, 3, 8, 16, 17, 18, 38, 42, 43, 44, 48, 64.

This holds for the misclassification error as well as for the sensitivity and specificity. Results for the top-30 subset and complete top-40 set are shown in Figure 3.14. It is further observed that the 95% confidence intervals are wider for the top-3 and top-12 sets as compared to the top-30 and top-40 sets.

For the top-40 ML1 set, all LOOCV's perform more or less the same. In conclusion, the impact of "leaving-out a specific IR" on the classification results becomes substantial for small subsets but is almost negligible for top-30 and the ML1 set of 40 IRs.

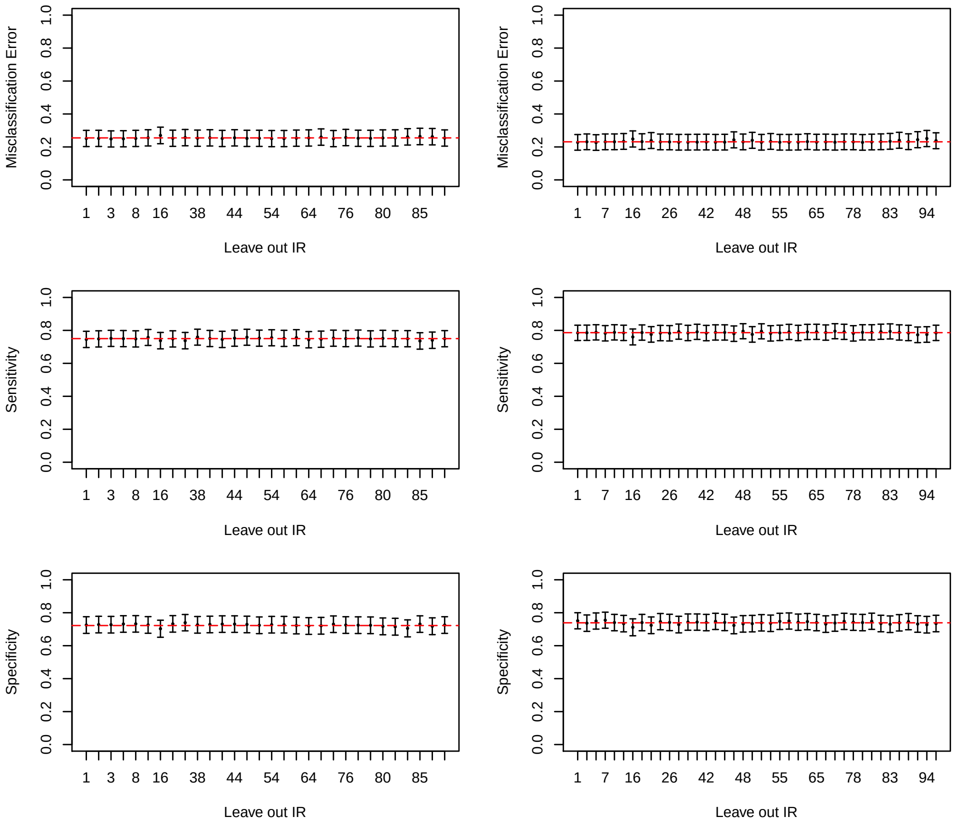


Figure 3.14: *SVM-based MCE (first row), sensitivity (second row), specificity (third row) results obtained by the LOOCV for the top-30 (first column) and top-40 (second column) IRs of the ML1 set. The red dashed horizontal lines indicate the median 3-fold cross validation results obtained by using all 30 (first column) or all 40 (second column) fixed IRs.*

3.5.2 Metabolic Signature Excluding the IRs in ML1

Our second goal was to evaluate whether the same classification results could be achieved by only using the remaining 56 IRs. Therefore, the same analysis proce-

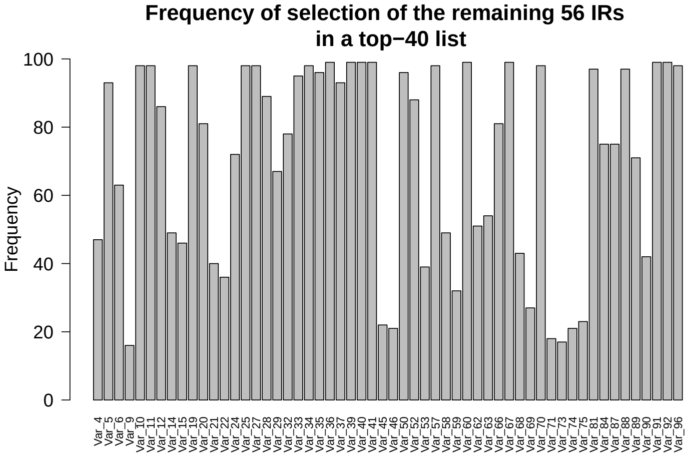


Figure 3.15: *Frequency of selection of the remaining 56 IRs in a top-40, obtained by applying the Limma t-test for the 100 iterations of the 3-fold cross-validation. For each iteration, the 56 IRs were ranked according to their Limma t-test significance and the frequency of selection in a top-40 was determined. Note that the top-40 IRs can be different from iteration to iteration.*

dure was followed as described in Section 3.3.1 and 3.3.2, i.e. for each iteration of the 3-fold cross validation, top-k IRs (k=3, 5,..., 43) were selected and 'top-k-based' classifiers were validated by means of the different classification methods. Note that the composition of the top-k lists can be different from iteration to iteration.

Figure 3.15 presents the frequency of selection of the remaining 56 IRs in a top-40 by the Limma t-test in the 3-fold cross validation of 100 iterations.

Figure 3.16 presents the median classification results obtained by the different classification methods for several top-k IRs selected out of the remaining 56 IRs while Table 3.4 summarizes the median overall misclassification errors.

The median misclassification errors of BC as HC as well as the median sensitivities and specificities are presented in Table A.4 in the appendix. These results demonstrate that also with a top-40 list selected out of the 56 remaining IRs, some of the classification methods still perform reasonably well. For the SVM method for instance, the resulting median overall MCE, MCE of BC as HC, sensitivity and specificity are respectively 0.29, 0.25, 0.75 and 0.65 as compared to 0.23, 0.22, 0.78 and 0.77 obtained by means of the fixed top-40 ML1 list.

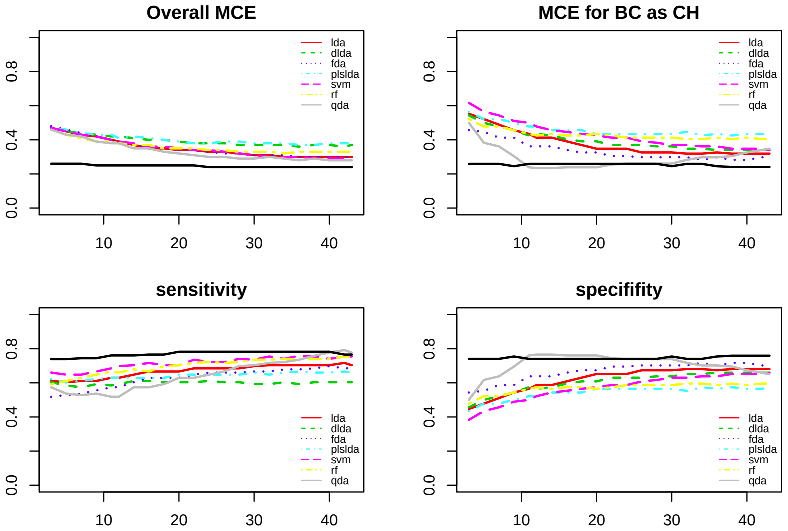


Figure 3.16: *Overview of the median classification results obtained by the different classification methods for several top-k IRs selected out of the remaining 56 IRs (96 IRs - the 40 IRs of the ML1 list) by the Limma t-test. The black line show the SVM results from all 96 IRs.*

Table 3.4: *Median overall misclassification error obtained by the different classification methods for several 'top-k-based' classifiers (k = 3, 12, 20, 30, 40, and 43). Left Panel: the IRs were selected out of the remaining 56 IRs. Right Panel: Most selected IRs from all 96 IRs. The median misclassification error of BC as HC as well as median sensitivity and specificity are presented in Table A.4.*

Method	3	12	20	30	40	43
LDA	0.47	0.39	0.34	0.31	0.30	0.30
DLDA	0.47	0.42	0.39	0.37	0.37	0.37
FDA	0.48	0.39	0.34	0.31	0.30	0.30
PLSLDA	0.47	0.41	0.39	0.38	0.38	0.38
SVM	0.47	0.39	0.34	0.31	0.29	0.29
RF	0.46	0.38	0.35	0.33	0.33	0.33
QDA	0.46	0.38	0.32	0.29	0.28	0.28

Method	3	12	20	30	40	43
LDA	0.30	0.28	0.27	0.26	0.25	0.25
DLDA	0.31	0.32	0.31	0.31	0.32	0.32
FDA	0.31	0.28	0.27	0.26	0.25	0.25
PLSLDA	0.29	0.28	0.30	0.29	0.29	0.29
SVM	0.26	0.25	0.25	0.24	0.24	0.24
RF	0.29	0.25	0.23	0.23	0.22	0.22
QDA	0.33	0.31	0.28	0.27	0.27	0.28

This strongly indicates that, although the most efficient classifiers to differentiate between breast cancer patients and healthy controls are based on the IRs of the ML1 list, the ML1 list seems not to be fully unique for the dataset. As a consequence, it seems not possible to define a limited set of unique metabolites to differentiate between the two groups. Note that for a ML1 list for which less metabolites are included (for example top 3, top 8 etc), the classification results using the metabolite which are not in the ML1 list is expected to improve. For example, if we exclude the top 3 metabolites from the data, the results from the classification based on

93 (96–3) remaining metabolites are expected to be similar to the results for the classification based on the top 3 metabolites. The main reason might be sought in the interconnection of metabolites in the biochemical pathways, making the signal in the data very strong.

3.6 Discussion

In this chapter, ^{1}H-NMR metabolomics is statistically evaluated as a complementary methodology with the potential to early diagnosis of breast cancer, even before it becomes clinically or radiologically detectable. The study is carried out on the basis of a blood plasma dataset of 300 subjects, including 161 breast cancer patients and 139 healthy controls, and uses the values of 96 well defined integration regions (IRs) of the proton spectra which represent the concentration of the low-molecular-weight plasma metabolites.

By means of a 3-fold cross validation procedure consisting of 1000 iterations, several feature selection methods were used to rank and select top-k IRs lists and to build classifiers by different classification methods. The resulting classifiers were validated in the test group (1/3 of the subjects) on the basis of misclassification errors (MCE), sensitivity and specificity. It was observed that although most of the classification methods performed more or less the same, the overall performances obtained by the SVM and RF methods were slightly better for the dataset examined. Based on variable sets of top-40 IRs, selected by the Limma t-test in the iteration procedure, the SVM classification method resulted in a median overall MCE, MCE of BC as HC, sensitivity and specificity of 0.24, 0.22, 0.78 and 0.76, respectively.

In the next step, two fixed sets of top-40 IRs were defined, namely the ML1 and ML2 lists. The goal here was to remove the IR selection bias and the noisy IRs in order to reduce the misclassification errors and to improve the sensitivity and specificity. The ML1 list was composed of the top-40 IRs with the highest frequency of selection by the Limma t-test in the 3-fold cross validation, while the ML2 list was composed of the top-40 IRs of the SVM classifier yielding the best overall performance. Based on these two fixed sets of top-40 IRs, classifiers were built by different classification methods in a 3-fold cross validation and were validated in the test group. It was found that although most of the classification methods performed more or less the same, the overall performance obtained by the SVM and RF methods was slightly better for the dataset. The SVM classifier based on the ML1 list resulted in quite similar results as the ML2 list. The median overall MCE, and specificity were equal to 0.23 and 0.77, respectively. Hence it can be concluded that the two list of IRs are robust sets to construct a promising classifier for the early detection of breast cancer.

In the final step, the top-ML1 list was evaluated towards its uniqueness for classification. The goal was to investigate whether or not the ML1 list was unique with respect to the construction of a well performing classifier. Here, the same cross validation procedure was followed to rank and select variable sets of top-40 IRs, but in this case by selecting IRs only out of the remaining 56 IRs. Based on the resulting SVM classifiers, a median overall MCE, MCE of BC as HC, sensitivity and specificity of respectively 0.29, 0.25, 0.75 and 0.65 were obtained. Although the selection out of the remaining 56 IRs results in a weakening of the classification performance, the outcome strongly indicates that it will be difficult to define a classifier based on a limited set of IRs, and thus a limited number of metabolites. The main reason probably has to be sought in the interconnection of metabolites in the biochemical pathways.

Chapter 4

Development of a Metabolic Signature for Lung Cancer

Lung cancer is the leading cause of cancer death worldwide with a five-year survival of only ±15% (Ferlay *et al.*, 2015; Mulshine and Sullivan, 2005). A promising screening tool for lung cancer is low-dose computed tomography (LDCT), which has been shown to reduce lung cancer mortality by 20% as compared to chest radiography screening (National Lung Screening Trial Research Team, 2011b). However, LDCT screening has some disadvantages such as the high cost associated with screening all patients at risk according to current risk models, radiation exposure and the low positive predictive value (high rate of false positive results, Bach *et al.*, 2012). Because of these limitations, other detection platforms are being evaluated, all with their advantages and shortcomings (Hasan *et al.*, 2014).

Over the past decade, accumulating evidence has shown that cancer cell metabolism differs from that of normal cells (Cantor and Sabatini, 2012; Munoz-Pinedo *et al.*, 2012; Sciacovelli *et al.*, 2014). More specifically, it is reprogrammed to promote cell proliferation and survival and is driven by aberrant signaling pathways induced by the activation of oncogenes/inactivation of tumor suppressor genes (Iurlaro *et al.*, 2014). One of the main adaptations of cancer cells is that, even in the presence of normal oxygen levels, they rely on anaerobic energy production through glycolysis, a hallmark known as the Warburg effect (Upadhyay *et al.*, 2013).

As metabolites are the end products of cellular processes, changes in their concentration reflect alterations in the metabolic phenotype (Holmes *et al.*, 2008). Proton nuclear magnetic resonance (^{1}H-NMR) based metabolomics allows a fast, non-invasive identification and quantification of complex mixtures of metabolites, as in plasma (Bervoets *et al.*, 2015; Lindon and Nicholson, 2008; Louis *et al.*, 2015b).

The aim of the analysis presented in this chapter is to establish a metabolic signature for lung cancer diagnostic. Our gaol is to develop a metabolic classifier that can be used for lung cancer screening.

The remainder of this chapter is organized as follows: the data is described in Section 4.1. Statistical methods are introduced in Section 4.2. Application to data is given in Section 4.3, followed by a discussion in Section 4.4.

4.1 Data Structure

In total, 233 out of the 357 lung cancer patients and 226 out of the 347 controls (for an elaborated description see Section 2.2.2) were randomly assigned to the training cohort, leaving a validation cohort of 98 lung cancer patients and 89 controls. Similar to Chapter 3, two parts of the dataset are used, the vector $\mathbf{Y}_{n\times 1}$ containing the individual labeling (LC or C) and the matrix $\mathbf{X}_{n\times m}$ containing the individual metabolic profiles.

4.2 Statistical Analysis

4.2.1 Classification and Cross Validation

The analysis presented in this chapter is an initial analysis in which the partial least squares discriminant analysis (PLS-DA) method was used as a classification method (Kramer, 1998; Bayne, 1999; Szymańska *et al.*, 2012; Barton *et al.*, 2008). Recall that PLS-DA is a variant of the PLS regression which constructs a set of orthogonal X-components $t_h = Xw_h^*$ and Y-components $u_h = Yc_h$ maximizing the covariance between the response u_h and the linear combination of the predictor variables t_h (Höskuldsson, 1988; Pérez-Enciso and Tenenhaus, 2003; Nguyen and Rocke, 2002). Here, w_h^* is a vector containing the weights given to each original variable in the kth component and c_h is the regression coefficient of y_k on hth X-component variable.

PLS-DA consists of two steps. The first step is a dimension reduction, which finds m appropriate linear transformations $t_1, ..., t_m$ of the vector of predictors X, where m is a tuning parameter. The second step is the linear discriminant analysis using the new components $t_1, ..., t_m$ as predictor variables. To find the optimal components number m, a cross validation method is applied. The training cohort is divided into an internal training (2/3 of the observations) and internal test set (1/3 of the observations).

For each number of the components (m ranging from 1 to 15), the cross-validation step was repeated 1000 times. The value of m, minimizing the misclassification error, is then used to predict the class of the observations from the validation cohort.
The PLS-DA discussed here is a standard classification procedure within the metabolomic studies (Gromski *et al.*, 2015; Barton *et al.*, 2008). In addition to the analysis using the PLS-DA as a classification method, secondary analysis was performed in which different classification methods were used. The second analysis was conducted in order to investigate whether the results obtained using the PLS-DA method and other classification methods described in Chapter 3 are comparable. The workflow of the secondary analysis is shown in Figure 4.1

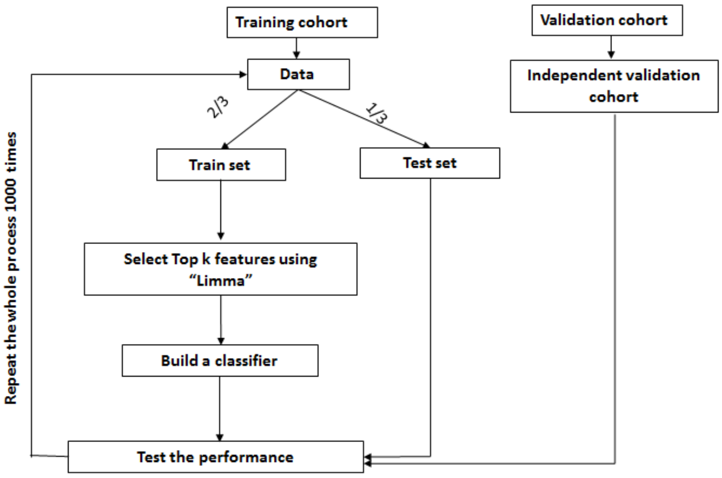

Training cohort
Validation cohort
Data
Independent validation cohort
2/3
1/3
Train set
Test set
Select Top k features using "Limma"
Build a classifier
Test the performance
Repeat the whole process 1000 times

Figure 4.1: *Workflow to build classifiers for the lung cancer data.*

Besides the test set obtained via 3-fold CV, an independent validation cohort was used as well. For the training data, for each step in the CV loop, the ratio of LC/C patients in the two cross-validated datasets (i.e. the training and test sets) was always equal to the LC/C ratio of the training cohort data. Lasso (Tibshirani, 1996), random forest (Breiman, 2001), support vector machine (Guyon *et al.*, 2002), linear discriminant analysis (Ripley, 1996), quadratic discriminant analysis (MacLachlan, 1992) methods were used to construct classifiers on both training and validation cohorts.

4.3 Results

4.3.1 Partial Least Square Discriminant Analysis

Partial Least Square Discriminant Analysis (PLS-DA) was used to train a classification model (classifier) in discriminating between lung cancer patients and controls based on data input from their metabolic phenotype. The resulting model was validated on an independent cohort. Table 4.1 shows the characteristics of the training and validation cohorts.

Table 4.1: *Characteristics of the subjects included in the study. Data are presented as mean ± standard deviation and range, unless otherwise indicated. Abbreviations: BMI: body mass index, C: controls, COPD: chronic obstructive pulmonary disease, LC: lung cancer patients.*

	Training cohort		**Validation cohort**	
	C	**LC**	**C**	**LC**
Number of subjects, N	226	233	89	98
Gender, N (%)				
Male	119 (53)	160 (69)	44 (49)	66 (67)
Female	107 (47)	73 (31)	45 (51)	32 (33)
Age, yrs	67 ± 11	68 ± 10	69 ± 10	64 ± 9
(range)	(38 - 88)	(36 - 88)	(47 - 89)	(45 - 83)
BMI, kg/m2	28.3 ± 5.0	25.8 ± 4.5	28.4 ± 5.7	26.2 ± 4.7
(range)	(18.7-46.7)	(17.5- 41.8)	(16.2-52.0)	(16.8-38.5)
COPD, N (%)	39 (17)	119 (51)	9 (10)	35 (36)
Taking lipid-lowering medication, N (%)	124 (55)	122 (52)	56 (63)	39 (40)
Diabetes, N (%)	23 (10)	40 (17)	20 (22)	12 (12)
Smoking habits				
Smoker, N (%)	47 (21)	113 (49)	15 (17)	48 (49)
Ex-smoker, N (%)	102 (45)	110 (47)	36 (40)	46 (47)
Non-smoker, N (%)	77 (34)	10 (4)	38 (43)	4 (4)
Pack years	16 ± 24	33 ± 21	13 ± 18	38 ± 21
(range)	(0-175)	(0-125)	(0-60)	(0-150)

The cross validation results are shown in Figure 4.2. The minimum misclassification error is obtained when the PLS-DA is built with 6 components.

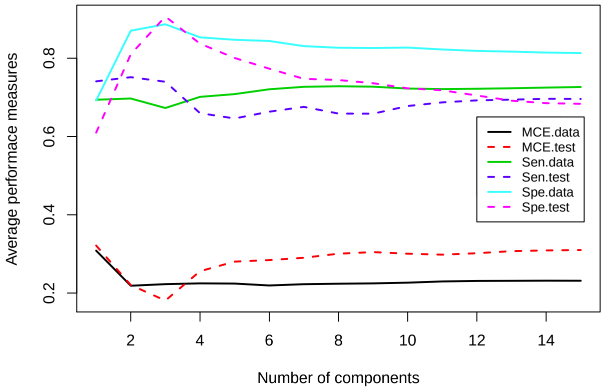


Figure 4.2: *Average performance measures for different components number. Solid lines: statistics calculated on the test set of the training cohort. Dashed lines: statistics calculated on the fixed validation cohort.*

Applying a PLS-DA with six components on the training set resulted in a model that allows to classify 82% of the 233 lung cancer patients and 89% of the 226 controls correctly.

The predictive accuracy of the model was assessed by applying it to the independent cohort of 98 lung cancer patients and 89 controls, resulting in a sensitivity of 75% and a specificity of 82%. The sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) of this model are shown in Table 4.2.

Table 4.2: *Characteristics of the trained PLS-DA classification models. Abbreviations: NPV: negative predictive value, PLS-DA: partial least squares discriminant analysis, PPV: positive predictive value, SEN: sensitivity: SPE: specificity.*

	SEN (%)	**SPE (%)**	**PPV (%)**	**NPV (%)**
Training cohort	82	89	89	82
Validation cohort	75	82	82	75

Table 4.3 shows the results obtained when the PLS-DA analysis is repeated 1000 times on both cohorts data with six components.

Table 4.3: *Cross validated PLS-DA. Left panel: results from the test set. Right panel: results from the independent validation cohort. Lower: 25% quantile, Upper: 97.5% quantile.*

	SEN	SPE	PPV	NPV	MCE
Lower	0.68	0.81	0.79	0.72	0.20
Median	0.72	0.85	0.83	0.75	0.22
Upper	0.82	0.93	0.92	0.83	0.28

	SEN	SPE	PPV	NPV	MCE
Lower	0.63	0.75	0.74	0.66	0.27
Median	0.66	0.78	0.76	0.68	0.28
Upper	0.74	0.84	0.83	0.73	0.34

4.3.2 Classification Using Other Methods

In this section, five classification methods mentioned in Section 4.2 were also used. Their performances are compared to PLS-DA performance. Table 4.4 presents an overview of the median specificity, sensitivity, misclassification error, positive and negative predictive values (PPV and NPV) obtained by the five classifiers for six 'top-k-based' classifiers (k=3, 12, 20, 30, 40, 43). Similar to Chapter 3, the top-k metabolites were selected using Limma t-test Smyth (2005). Table 4.4 shows that for top40 and top43, LDA, SVM, LASSO methods give comparable results to the results obtained from PLS-DA method with six components in Table 4.3 (left panel). These results were obtained on the test set in the cross-validation steps.

Table 4.4: *Classification results on the test set. Median specificity, sensitivity, misclassification error (MCE), negative predictive value (NPV), positive predictive value (PPV) obtained by different classification methods based on six 'top-k' metabolites. RF: Random forest, SVM: Support vector machine, LDA: linear discriminant analysis, QDA: quadratic discriminant analysis.*

Method	Top3	Top12	Top20	Top30	Top40	Top43
	SPECIFICITY					
LASSO	0.65	0.78	0.80	0.83	0.84	0.84
RF	0.71	0.76	0.79	0.79	0.79	0.79
SVM	0.74	0.80	0.82	0.84	0.84	0.84
QDA	0.76	0.73	0.79	0.80	0.76	0.75
LDA	0.76	0.83	0.84	0.85	0.85	0.84
	SENSITIVITY					
LASSO	0.77	0.71	0.70	0.73	0.76	0.76
RF	0.68	0.69	0.71	0.73	0.73	0.73
SVM	0.72	0.68	0.72	0.75	0.76	0.75
QDA	0.71	0.68	0.64	0.65	0.67	0.68
LDA	0.71	0.66	0.69	0.73	0.74	0.74
	MCE					
LASSO	0.29	0.25	0.25	0.22	0.21	0.20
RF	0.31	0.27	0.25	0.25	0.24	0.24
SVM	0.27	0.27	0.24	0.21	0.20	0.21
QDA	0.27	0.29	0.29	0.28	0.29	0.29
LDA	0.27	0.26	0.24	0.22	0.21	0.21
	NPV					
LASSO	0.73	0.72	0.72	0.75	0.77	0.77
RF	0.68	0.71	0.72	0.73	0.74	0.74
SVM	0.72	0.70	0.74	0.76	0.77	0.77
QDA	0.72	0.69	0.68	0.68	0.69	0.69
LDA	0.72	0.70	0.73	0.75	0.76	0.76
	PPV					
LASSO	0.70	0.77	0.79	0.81	0.82	0.83
RF	0.70	0.75	0.77	0.78	0.78	0.78
SVM	0.74	0.77	0.80	0.82	0.83	0.82
QDA	0.75	0.73	0.76	0.76	0.74	0.73
LDA	0.75	0.79	0.82	0.83	0.83	0.83

Table 4.5 shows the results obtained on the independent validation cohort data. The estimated performance measures are very close to the results obtained on the test set. Among the five classification methods, QDA was giving poor performance.

Table 4.5: *Validation cohort data. Median specificity, sensitivity, misclassification error (MCE), negative predictive value (NPV), positive predictive value (PPV) obtained by different classification methods based on six 'top-k' metabolites.*

Method	Top3	Top12	Top20	Top30	Top40	Top43
	SPECIFICITY					
LASSO	0.65	0.78	0.80	0.83	0.84	0.84
RF	0.71	0.76	0.79	0.79	0.79	0.79
SVM	0.74	0.80	0.82	0.84	0.84	0.84
QDA	0.76	0.73	0.79	0.80	0.76	0.75
LDA	0.76	0.83	0.84	0.85	0.85	0.84
	SENSITIVITY					
LASSO	0.81	0.78	0.80	0.76	0.71	0.71
RF	0.77	0.80	0.82	0.82	0.82	0.82
SVM	0.76	0.77	0.78	0.72	0.71	0.70
QDA	0.74	0.72	0.73	0.74	0.80	0.81
LDA	0.74	0.73	0.73	0.71	0.72	0.72
	MCE					
LASSO	0.30	0.27	0.25	0.24	0.21	0.20
RF	0.29	0.25	0.23	0.22	0.24	0.24
SVM	0.29	0.25	0.22	0.24	0.20	0.21
QDA	0.28	0.31	0.26	0.26	0.29	0.29
LDA	0.28	0.25	0.24	0.25	0.21	0.21
	NPV					
LASSO	0.72	0.73	0.76	0.74	0.72	0.72
RF	0.72	0.76	0.78	0.79	0.78	0.79
SVM	0.71	0.74	0.76	0.73	0.72	0.72
QDA	0.71	0.68	0.72	0.73	0.75	0.76
LDA	0.71	0.72	0.73	0.71	0.72	0.72
	PPV					
LASSO	0.68	0.73	0.75	0.79	0.80	0.81
RF	0.70	0.75	0.76	0.76	0.76	0.77
SVM	0.71	0.76	0.79	0.81	0.80	0.80
QDA	0.72	0.70	0.77	0.77	0.73	0.73
LDA	0.72	0.78	0.79	0.79	0.78	0.78

4.4 Discussion

The results presented in this chapter demonstrate that (1) the metabolic classifier allows to classify 82% of the lung cancer patients and 89% of the controls correctly and (2) the metabolic classifier discriminates between lung cancer patients and controls of the independent cohort with a sensitivity of 75%, a specificity of 82% .

The metabolic phenotype, which is represented by the relative abundance of plasma metabolites, has to be seen as a single biomarker that cannot be defined based on a cut-off value. It is demonstrated that the combination of a series of subtle metabolic alterations (metabolites of which the plasma concentration is increased/decreased in lung cancer patients compared to controls), detected by ^{1}H-NMR spectroscopy and presented by PLS-DA, enables to diagnose lung cancer.

Recently, many studies have explored lung cancer metabolism, but mostly by mass spectrometry (MS) techniques rather than by ^{1}H-NMR spectroscopy (Wen *et al.*, 2013; Hori *et al.*, 2011; Chen *et al.*, 2015). Although MS is without doubt more sensitive, ^{1}H-NMR spectroscopy requires no invasive extraction procedures, and so minimal sample preparation (Lindon and Nicholson, 2008). Both techniques are therefore complementary and of importance in the field of metabolomics. Furthermore, most published NMR studies focused on the metabolic composition of the lung cancer tissue despite the fact that metabolic phenotyping of blood plasma has the advantage to assess more directly complex interaction between tumor and host (Chen *et al.*, 2011; Duarte *et al.*, 2010). Moreover, blood samples can be obtained non-invasively and with minimal risk for the patient (Mamas *et al.*, 2011). According to a review of Duarte *et al.* (2013), only Rocha *et al.* (2011) investigated lung cancer-induced metabolic alterations in plasma by ^{1}H-NMR spectroscopy, demonstrating a discrimination between 85 lung cancer patients and 78 controls with a sensitivity and specificity of ± 90% but unfortunately without validation in an independent cohort.

Currently, low-dose computed tomography (LDCT) is the most studied tool to screen for lung cancer. The NELSON trial demonstrates that LDCT screening has a sensitivity of 85% and a specificity of 99% in comparison to no screening (Horeweg *et al.*, 2014). However, a major limitation of LDCT is the low PPV ranging from 4% in the National Lung Screening Trial to 40% in the NELSON trial. This means that more than half of the study participants were referred for further investigations, being not without cost and risk, on the basis of false-positive results (National Lung Screening Trial Research Team, 2011b; Horeweg *et al.*, 2014). Strengthening of current risk models by incorporating metabolic phenotype information might be the way to better identify patients eligible for LDCT screening. This will be the main issue in Chapter 5, where the added predictive value of metabolic data is studied.

In this respect, ^{1}H-NMR based metabolomics seems to be reasonably able to discriminate between early stage patients and a randomly selected equally populated group of controls. This indicates that metabolic alterations present in the initial phase of cancer development can already be detected by ^{1}H-NMR based-metabolomics. Although these results look promising, the number of early stage patients needs to be increased to confirm.

Next to PLS-DA, different classification methods were used to classify LC and C patients in a cross validation way. Results showed that about 40 metabolites were enough to have comparable performance. The difference between PLS-DA and those classifiers was the metabolites selection. PLS-DA constructs components without any selection on the metabolites.

In conclusion, we validated ^{1}H-NMR metabolic phenotype of blood plasma as a complementary tool to discriminate between lung cancer patients and controls.

In the next chapter we focus on a slightly different problem. Our intent is not to use the metabolome as a separate screening tool but to complement current risk models with additional parameters to better select high-risk individuals eligible for LDCT screening. Therefore, the question we discuss in Chapter 5 is related to the benefit of using metabolic data in addition to epidemiological and clinical variables in a risk model for lung cancer.

Chapter 5

Risk Models to Select Individuals Eligible for Lung Cancer Screening with Low-Dose Computed Tomography: Adding the Metabolic Phenotype

5.1 Introduction

One of the main criteria for a screening test is the cost-effectiveness, meaning that the number of false positive results should be low to prevent unnecessary surgical interventions (Tammemagi and Lam, 2014; Wood *et al.*, 2012). To maximize the benefit-risk balance, accurate selection of high-risk target population for lung cancer screening programs necessitates robust methods for risk prediction (Field and Duffy, 2008; Field *et al.*, 2013b).

Current risk models for prediction of lung cancer have tended to concentrate on clinical risk factors, including age, smoking behavior, previous history of cancer and family history of lung cancer (Cassidy *et al.*, 2008; Spitz *et al.*, 2007; Hoggart *et al.*, 2012; Tammemagi *et al.*, 2013; Bach *et al.*, 2003). Since lung cancer predominantly occurs in elderly people and smoking is an important risk factor, high-risk individuals in the two largest randomized controlled trials designed to evaluate the impact of low-dose computed tomography (LDCT) screening on lung cancer mortality were selected on the basis of age and smoking behavior (National Lung Screening Trial Research Team, 2011a; Zhaoa *et al.*, 2011). More specifically, eligible participants for the North-American National Lung Screening Trial (NLST) were aged between 55 and 74 years and had a smoking history of at least 30 pack years. Former smokers were only included in the study if they had quit smoking within the past 15 years (National Lung Screening Trial Research Team, 2011a,b). Furthermore, the Dutch-Belgian lung cancer screening trial (Dutch acronym - NELSON) recruited subjects aged between 50 and 75 years, who smoked 15 or more cigarettes a day for more than 25 years ($\geq$18.75 pack years) or 10 or more cigarettes a day for more than 30 years ($\geq$15 pack years). Former smokers were only included in the study if they had quit smoking for less than 10 years (Zhaoa *et al.*, 2011; van den Bergh *et al.*, 2008). The major drawback of both LDCT screening studies is the low positive predictive value (PPV), ranging from 3.8% in the NLST study to 40.4% in the NELSON study. This indicates that more than half of the study participants were referred for further investigations, being not without cost and risk, on the basis of false positive results (National Lung Screening Trial Research Team, 2011b, 2013; Horeweg *et al.*, 2014; Bach *et al.*, 2012). Consequently, there is an increasing interest to improve the accuracy of risk models by adding lung cancer risk-related biomarkers. This is done in order to better select high-risk individuals eligible for lung cancer screening with LDCT and so to lower false positive rate and corresponding financial burden.

In the previous chapter, we have shown that metabolic signature can be used to predict the disease status of a subject. In this chapter, we focus on a different question: what can a metabolic signature add to a risk model in addition to clinical variables? In other words, can we improve the accuracy when the metabolic data is added to the model? This can be done by considering the clinical and the metabolic data in different ways. A first approach is to treat both data types in the same way (naive approach). A combined prediction model is built by treating clinical and metabolic predictors in the same way. A second approach is to fit a model on the clinical data, and use the predicted values from this model as an offset in a model comprising metabolic data (clinical offset). A third approach is to fit a model using both data types, but favoring the clinical data. We adopted the third approach in our analyses. This was achieved by not applying penalty on the clinical risk factors in penalized regression models using lasso and elastic net penalties. A forth approach that can be used is to summarize the metabolic data in a form of new component (or score) and add this component in a prediction model comprising clinical variables (Boulesteix and Sauerbrei, 2011; De Bin *et al.*, 2014).

A blood-based diagnostic biomarker signifies an attractive option to complement risk models used to select high-risk individuals eligible for lung cancer screening with LDCT since blood samples can be obtained in a non-invasive way and with minimal risk for the patient (Mamas *et al.*, 2011; Tsay *et al.*, 2014). Louis *et al.* (2015a,b, 2016) have demonstrated that the metabolic phenotype of blood plasma, determined by ^{1}H-NMR spectroscopy, not only enables to discriminate between cancer patients and controls but also between different cancer types such as lung and breast cancer. ^{1}H-NMR spectroscopy, one of the main analytical platforms used in metabolomics studies, is a very reproducible technique which permits a fast and non-invasive identification and special quantification of complex mixtures of metabolites, as in plasma, with minimal sample preparation and relatively low cost on a per sample basis (Emwas *et al.*, 2013; Lindon and Nicholson, 2008). Hence, ^{1}H-NMR-based metabolomics of blood plasma seems to provide an attractive blood-based diagnostic biomarkers to add to risk models used for the selection of high-risk individuals eligible for lung cancer screening with LDCT.

This chapter is organized as follows: in Section 5.2 data and analysis setting are described. An application to the data is given in Section 5.3, followed by a discussion in Section 5.4.

5.2 Data and Analysis Setting

5.2.1 Cross Validation Procedure

Blood sampling, samples preparation and NMR analysis protocols used have been described in detail by Louis *et al.* (2016, 2015a). For the analysis presented in Section 5.3.4, a cross validation loop is used to estimate the performance statistics. The study population described in Section 2.2.2 was randomly split into two cohorts. A training cohort (two thirds of the subjects) and a validation cohort (one third of the subjects, referred to as test cohort 1). The second cohort is a fixed validation cohort (referred to as test cohort 2). Figure B.1 shows the data splitting scheme.

The subject characteristics of the first cohort of 536 subjects (273 lung cancer patients and 263 controls) are presented in Table 5.1.

Table 5.1: *Subject characteristics of cohort 1. Data are presented as mean ± standard deviation, unless otherwise indicated. Univariate logistic regression models were used to calculate p-values for continuous variables, while a Chi-square test was used to compute p-values for categorical variables.*

	Controls (n=263)	**Patients (n=273)**	**p-value**
Gender, n (%)			
Male	139 (53)	186 (68)	0.0003
Female	124 (47)	87 (32)	
Age, yrs	66 ± 11	68 ± 10	0.056
BMI, kg/m^2	28.0 ± 5.1	25.8 ± 4.5	<0.0001
Smoking habits, n (%)			
Smoker	60 (23)	131 (48)	<0.0001
Ex-smoker	111 (42)	131 (48)	
Non-smoker	92 (35)	11 (4)	
Smoking pack years	16 ± 23	33 ± 21	<0.0001
Previous mine-worker, n (%)			
Yes	15 (6)	24 (9)	0.001
No	124 (47)	162 (59)	
Not applicable	124 (47)	87 (32)	
Prior diagnosis of malignant tumor, n (%)			
Yes	9 (3)	24 (9)	0.016
No	254 (97)	249 (91)	
COPD, n (%)			
Yes	30 (11)	139 (51)	<0.0001
No	233 (89)	134 (49)	
Diabetes, n (%)			
Yes	47 (18)	47 (17)	0.932
No	216 (82)	226 (83)	
Taking lipid-lowering medication, n (%)			
Yes	149 (57)	142 (52)	0.322
No	114 (43)	131 (48)	
Taking malfunctioning thyroid medication, n (%)			
Yes	18 (7)	9 (3)	0.093
No	245 (93)	264 (97)	
Taking anti-arrhythmic medication, n (%)			
Yes	13 (5)	33 (12)	0.005
No	250 (95)	240 (88)	
Taking blood pressure-lowering medication, n (%)			
Yes	193 (73)	168 (62)	0.005
No	70 (27)	105 (38)	
Taking anti-coagulants medication, n (%)			
Yes	163 (62)	155 (57)	0.255
No	100 (38)	118 (43)	

5.2.2 Data Structure

Let $\mathbf{X}$ be a $n \times m$ data matrix, where n is the sample size and m the number of the molecular variables (metabolites in our case). The $\mathbf{X}$ matrix contains information about 102 metabolites on 536 samples. Let $\mathbf{Z}$ be a $n \times p$ clinical variables matrix. It is assumed that $p < n$, but no such restriction is put on m. Let $\mathbf{Y}$ $n \times 1$ be a vector in which the ith entry is an indicator variable which is equal to 1 if the patient has lung cancer (LC) and 0 otherwise. We further assume that:

$$Y_i \sim Binomial(1, \pi_i), \quad \text{with} \quad Y_i = \begin{cases} 1 & : \text{Lung cancer } (\pi_i), \\ 0 & : \text{Otherwise } (1\text{-}\pi_i). \end{cases}$$

The three data sources $\mathbf{Y}$, $\mathbf{Z}$ and $\mathbf{X}$ can be represented as matrices given by:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}.$$

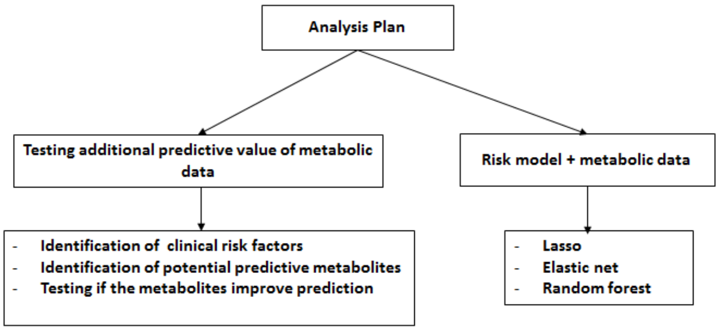

Analysis Plan
Testing additional predictive value of metabolic data
Risk model + metabolic data
- Identification of clinical risk factors
- Identification of potential predictive metabolites
- Testing if the metabolites improve prediction
- Lasso
- Elastic net
- Random forest

Figure 5.1: *Analysis plan. Left panel: identification and testing of metabolic variables, Right Panel: Development of a risk models taking metabolic data into account*
.

In Section 5.3, we follow the analysis plan depicted in Figure 5.1 in order to construct risk models taking into account clinical and metabolic data. Firstly, we identify significant clinical risk factors. Secondly, we investigate if we can find significant metabolites given clinical risk factors. Next to identification, three different testing procedures are also used to test the additive predictive value for metabolic data. Lastly, three different modeling approaches are used to construct risk models including clinical variables and metabolic data. These models are used to evaluate the gain, in terms of accuracy in prediction, when the metabolic data are included in the predictive model.

5.3 Application to the Data

5.3.1 Risk Model with Clinical Covariates

5.3.1.1 Model Formulation

Following Cassidy *et al.* (2008) (Liverpool Lung Project model, LLP), we first consider a model comprising of only clinical covariates. We refer to this model as the baseline risk model and its linear predictor is given by:

$$\begin{aligned} logit(\pi_i) \quad &= \mathbf{Z}\gamma, \\ &= \gamma_0 + \gamma_1 \mathrm{AGE}_i + \gamma_2 \mathrm{PACK}_i + \gamma_3 \mathrm{BMI}_i + \gamma_4 \mathrm{COPD}_i + \gamma_5 \mathrm{ARR}_i \\ &+ \gamma_6 \mathrm{SMOKING}_i + \gamma_7 \mathrm{COA}_i + \gamma_8 \mathrm{SEX}_i + \gamma_9 \mathrm{PMW}. \end{aligned} \tag{5.1}$$

Here, AGE is the patient's age , PACK : smoking pack years, BMI : body mass index, COPD : chronic obstructive pulmonary disease, ARR : taking anti-arrhythmic medication, SMOKING : smoking status (active, quit, never), COA : taking anti-coagulant medication, PMW : previous mine worker.

5.3.1.2 Results

Univariate testing, based on a univariate logistic regression model, indicated that 9 clinical risk factors were significantly associated to lung cancer, i.e. gender, body mass index (BMI), smoking habits, number of smoking pack years, previous mine-worker, prior diagnosis of a malignant tumor, presence of chronic obstructive pulmonary disease (COPD), intake of anti-arrhythmic medication and intake of medication against high blood pressure. All clinical risk factors were then included in a multiple logistic regression model using a stepwise model selection procedure. Table 5.2 shows that age, smoking habits, number of smoking pack years, BMI, presence of COPD, intake of anti-arrhythmic medication and intake of anti-coagulants medication were statistically significant.

Table 5.2: *Odds ratio estimates from a multiple logistic regression model comprising statistically significant clinical risk factors. Abbreviations: CI: confidence interval, OR: odds ratio.*

	OR (95% CI)	**p-value**
(Intercept)	0.103 (0.012-0.835)	0.035
Age	1.035 (1.013-1.059)	0.002
Smoking pack years	1.016 (1.005-1.028)	0.0057
BMI	0.922 (0.879-0.965)	0.0006
COPD	5.466 (3.316-9.248)	<0.0001
Taking anti-arrhythmic medication	4.293 (1.932-10.103)	0.0005
Smoker	9.354 (4.031-23.156)	<0.0001
Ex-smoker	6.525 (3.052-14.996)	<0.0001
Taking anti-coagulants medication	0.406 (0.253-0.642)	0.0001

5.3.2 Feature-by-feature Analysis

5.3.2.1 Model Formulation

In order to test for statistical significance of a single feature, given a set of clinical risk factors, a feature-by-feature model was formulated and compared to the baseline risk model in equation (5.1). That is:

$$\begin{aligned} M_0 &: logit(\pi_i) = \gamma_0 + \textstyle\sum_{l=1}^{p} \gamma_l Z_{il}, \\ M_1 &: logit(\pi_i) = \alpha_0 + \textstyle\sum_{l=1}^{p} \alpha_l Z_{il} + \beta_j X_{ij}. \end{aligned} \tag{5.2}$$

The parameter β_j is statistically significant whenever the following null hypothesis is rejected:

$$H_0 : \beta_j = 0, \qquad j = 1, 2, ..., m. \tag{5.3}$$

Since many tests are conducted, the false discovery rate (FDR, Benjamini and Hochberg, 1995) adjustment procedure is used to correct for multiple testing .

5.3.2.2 Results

Multiple logistic regression models comprising statistically significant clinical risk factors and one metabolite at a time were fitted to find which metabolites have statistically significant effects on disease status prediction. After adjusting for multiple testing, 53 out of the 102 metabolites were found to have significant effects on disease status prediction. Figure 5.2 shows a volcano plot, adjusted and unadjusted p-values obtained from multiple logistic regression models containing significant clinical risk factors and one integration value at a time.

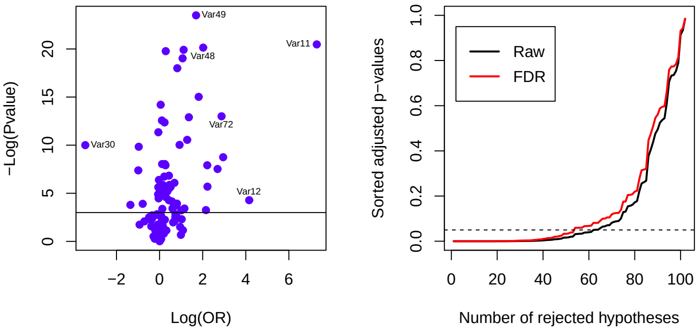


Figure 5.2: *Unadjusted and adjusted p-values obtained for the tests of the NMR integration values. Left panel: volcano plot. Right panel: p-values obtained from a model containing significant clinical risk factors and one NMR integration region at a time.*

5.3.3 Testing the Added Predictive Value of the Metabolic Signature

5.3.3.1 The Likelihood Ratio Test

In Section 5.3.2.1, we presented a testing procedure based on models comprising of significant clinical risk factors and one metabolite at a time. If the sample size is greater than the total number of all variables ($p + m < n$), one can test statistical significance of many metabolites given a set of clinical values using likelihood ratio test. The linear predictor of the alternative model in equation (5.2) becomes:

$$logit(\pi_i) = \alpha_0 + \sum_{l=1}^{p} \alpha_l Z_{il} + \sum_{j=1}^{m} \beta_j X_{ij}. \tag{5.4}$$

The corresponding null hypothesis is given by :

$$H_0 : \beta_1 = \beta_2 = ... = \beta_m = 0. \tag{5.5}$$

This hypothesis is tested by comparing the baseline model (M_0) in equation (5.2) with the model in equation (5.4) using the likelihood ratio test. The corresponding test statistic is a χ^2_m. A rejection of H_0 implies that at least one of the β_j's is statistically different from zero.

5.3.3.2 Global Test I (Goeman *et al.*, 2005)

Although for the case study considered in this chapter, the likelihood ratio test can be used to test the null hypothesis formulated in equation (5.5) without a problem ($n = 536$ and $m + p = 109$), we use additional tests developed in the context of high dimensional data. If $m + p > n$, the model in equation (5.4) cannot be fitted and the likelihood ratio test is not applicable. Goeman *et al.* (2004) developed a global test for a group of high dimension variables to test association with a clinical outcome. Goeman *et al.* (2005) extended their method and made it possible to adjust for presence of covariates. It allows to shift the hypothesis testing from a single variable test (i.e. feature-by-feature analysis), to a group of variables (genes or metabolites pathway). It can be applied to high dimensional data such genomic, metabolomic, proteomic data. Similar to the previous section, Goeman *et al.* (2005) considered the null hypothesis formulated in (5.5). To obtain a test that is applicable for any value of m, they assumed that the regressions coefficients $\beta_1, ..., \beta_m$ are random variables and a priori independent with mean zero and common variance τ^2. A single unknown parameter τ^2 determines how much the regression coefficients can deviate from zero. The null hypothesis becomes :

$$H_0 : \tau^2 = 0. \tag{5.6}$$

Goeman *et al.* (2004) showed that the global test statistic is given by:

$$Q = \frac{1}{m} \sum_{j=1}^{m} Q_j,$$

where, Q_j is the test statistic that would have been calculated if a single feature is included in the model. The test Q is derived in stages. First, it is assumed that all parameter except τ^2 are known, i.e. the regression coefficients for the clinical covariates. Secondly, the score test for τ^2 is derived, and can be generalized to the situation with unknown parameters. In this latter case, the parameter values of $\gamma_1, \ldots, \gamma_p$ are replaced by their estimates. As pointed out by Goeman *et al.* (2004), the global test Q for a group of m features is the average of the m statistics calculated for m features (for example if the score test is used to test the null hypothesis $H_0 : \beta_j = 0$ for the models formulated in equation (5.3)).

5.3.3.3 Global Test II (Boulesteix and Hothorn, 2010)

In this section, we considered the testing procedure proposed by Boulesteix and Hothorn (2010) to test the added predictive value for the high dimensional data. It is based on boosting algorithm in which the high dimension data are selected given that the clinical variables are set as an offset (clinical offset). Boulesteix and Hothorn (2010) proposed a two-stage approach, in which in the first stage a logistic regression is fitted on the clinical variables (for a binary response). At the second-stage, component-wise fitting is performed using boosting algorithm (Bühlmann and Hothorn, 2007).

At each boosting iteration, a variable X_j minimizing the log-likelihood loss function given by:

$$\rho_{log-lik}(\tilde{y}, f) = log_2(1 + exp(-2\tilde{y}f)),$$

where, $\tilde{y} = (2y - 1)$ and $f = log(p/(1 - p))/2$ enters the model (Bühlmann and Hothorn, 2007). A permutation based testing procedure is performed to test the additional predictive value of the X_j's variables given the clinical variables. The model under alternative hypothesis is given by:

$$logit(\pi_i) = \alpha + \sum_{j=i}^{p} \gamma_l Z_{il} + \sum_{j=1}^{m} \beta_j^* X_j,$$

here, β_j^* represents the estimated parameter from a permuted data. The null hypothesis of no added predictive value is formally stated as:

$$H_0 : \beta_1^* = ... = \beta_m^* = 0.$$

Note that only columns of $\mathbf{X}$ are permuted. The two-stage procedure is applied and the negative binomial log-likelihood ℓ is computed for the permuted data set. The whole procedure is repeated B times, yielding $\ell_1, ..., \ell_B$. The permutation p-value is then obtained as

$$p - value = \frac{1}{B} \sum_{b=1}^{B} \mathbf{I}(\ell_b \leq \ell_{obs}).$$

Here, $\mathbf{I(.)}$ denotes the indicator function which takes the value of 1 if $\ell_b \leq \ell_{obs}$ and zero otherwise. The main difference between the tests presented in Section 5.3.3.2 and 5.3.3.3 is that the test proposed by Boulesteix and Hothorn (2010) includes a variable selection step (in the second-stage), whereas there is not variable selection in the test proposed by Goeman *et al.* (2005).

5.3.3.4 Results

The three tests procedures were applied to the ^{1}H-NMR metabolic phenotype data to test whether they do have an added predictive value for lung cancer status prediction. All tests indicate that NMR metabolic phenotype data have significant impact on the disease status prediction. Table 5.3 presents p-values obtained for the three tests.

Table 5.3: *P-values obtained from the three testing procedures.*

Goeman et al.	**Boulesteix and Hothorn**	**LRT**
p-value	p-value	p-value
8.1×10^{-5}	0	2.2×10^{-16}

5.3.4 Metabolic Based Risk Model for Lung Cancer

In the previous section, we showed that metabolic data has a statistical significant effect if added to clinical covariates in the risk model. However, up to this point we did not quantify the gain in using the metabolic data in the risk model (in addition to the clinical covariates). In this section, we investigate what is the gain in terms of reduction in misclassification error, when the metabolic data is added to the risk model. Our aim is to use the metabolic data, in addition to the clinical covariate, in a risk model to predict the disease status. In Section 5.3.4.1, we use predictive models based on penalized logistic regression and random forest (Breiman, 2001). Note that since we use models to predict the disease status, a cross validation loop is used to estimate the performance statistics. For the penalized logistic regression model, the cross validation procedure was presented in Section 5.2.1. The cross validation for the random forest procedure is discussed in Section 5.3.4.1.

5.3.4.1 Model Formulation: Lasso and Elastic Net

Multiple logistic regression models comprising significant clinical risk factors defined in equation (5.1) is used as a baseline model. Predictive models were developed on a training cohort and evaluated on two independent validation cohorts (test cohort 1 and 2 in Figure B.1). Three different predictive models were used: random forest (RF, Breiman, 2001), penalized logistic regression models using lasso (Tibshirani, 1996; Friedman *et al.*, 2001) and elastic net penalties (Friedman *et al.*, 2001; Zou and Hastie, 2005). Given the logistic regression in equation (5.4), the negative log likelihood with penalty takes the following form:

$$-\frac{1}{N}\sum_{i=1}^{N}\{y_i \log \Pr(Y=1|x_i,z_i) + (1-y_i) \log \Pr(\mathrm{Y=0|x_i,z_i})\} + \lambda P_\alpha(\beta_j),$$

$$= -\frac{1}{N}\sum_{i=1}^{N}\{y_i(\beta_0 + x_i'\beta + z_i'\gamma) - \log(1+e^{\beta_0+x_i'\beta+z_i'\gamma})\} + \lambda P_\alpha(\beta_j),$$

where, $P_\alpha(\beta_j) = \sum_{j=1}^{m}\left[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|\right]$ is the elastic net penalty (Zou and Hastie, 2005). P_α is a compromise between the ridge-regression penalty ($\alpha = 0$) and the lasso penalty ($\alpha = 1$). The penalty parameter λ introduces shrinkage in β_j coefficients. The above notation suggests that only coefficients related to **X** matrix (metabolic data) are penalized. This is the 'favoring' approach (Boulesteix and Sauerbrei, 2011; De Bin *et al.*, 2014).

Random Forest (RF, Breiman, 2001) is an ensemble method for classification that builds many decision trees based on bootstrap samples from the original data and aggregates their predictions to improve the predictive accuracy and to control for over-fitting. The re-sampling scheme used in RF is different from the k-fold cross validation used so far for other classification methods. RF does not make distinction between the clinical and metabolic variables, as stated before it uses a 'naive' approach to test for added predictive value (De Bin *et al.*, 2014). Figure 5.3 depicts a schematic representation of RF resampling.

RF has two parameters which have to be fixed before it is run, namely the number of the trees to grow and the number of features to select at each split.

- Fix the number of trees to grow,
  - The dataset is split into a train set and out-of-bag (OOB, test set) using bootstrap draws (sampling with replacement),
  - Randomly select a fixed number of features at each split (on the train set),
  - Grow a tree without pruning. Some stopping criteria have to be fulfilled in order to stop the tree growth,
  - Predict class membership of the observations in the OOB,
- At the end of the run, by majority vote, get the final class membership of each observation when it was in the out-of-bag,
- Compute the OOB error (MCE) and other performance measures such as sensitivity, specificity, positive and negative predictive values.

As pointed out by Breiman (2001), each tree is grown using a different bootstrap sample from the original data.

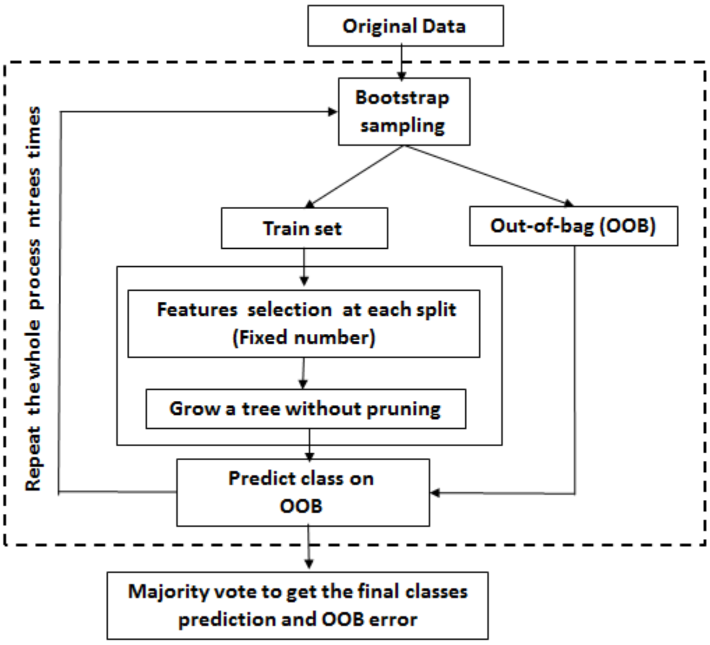

Original Data
Bootstrap sampling
Train set
Out-of-bag (OOB)
Features selection at each split (Fixed number)
Grow a tree without pruning
Predict class on OOB
Repeat the whole process ntrees times
Majority vote to get the final classes prediction and OOB error

Figure 5.3: *Random forest algorithm. Adapted from Boulesteix et al. (2012).*

5.3.4.2 Results

Predictive models with and without ^{1}H-NMR metabolic phenotype data were developed using lasso and elastic net penalties in order to evaluate the added predictive value of the ^{1}H-NMR metabolic phenotype data. Recall that we used the favoring approach when the metabolic data was added to the models with penalties (clinical data are not penalized). For the lasso method, the average misclassification error (MCE) of test cohort 1 dropped from 24.6% to 18.6% when the ^{1}H-NMR metabolic phenotype data were included in the model. The average MCE of test cohort 2 dropped from 23.0% to 22.6%. Similar patterns were observed for the models based on the elastic net penalty. The average MCE of test cohort 1 declined from 24.6% to 17.8% and that of test cohort 2 dropped from 23.0% to 21.4% when the ^{1}H-NMR metabolic phenotype data were included in the model (Table 5.4).

Density estimates for the distribution of the MCE are shown in Figure 5.4. We notice, for both test sets, that the distribution of the MCE when the metabolic data is added to the model is shifted to the left (compared with the distribution of the MCE when only clinical covariates are used in the risk model). For both test cohorts, the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) of the penalized lasso and elastic net models remain stable when ^{1}H-NMR metabolic phenotype data were added to the logistic regression model.

For the RF, the average MCE with only clinical risk factors was 22.7% for test cohort 1 and dropped to 18.2% when the ^{1}H-NMR metabolic phenotype data were added to the model. The average MCE of test cohort 2 declines from 28.6 to 12.3% when both clinical and metabolic phenotype data were included in the model. Moreover, the PPV and the specificity increase for test cohort 1 when the metabolic phenotype data were included in the model, while the NPV and sensitivity remained quasi stable. Furthermore, PPV, NPV, sensitivity and specificity increased for test cohort 2 when the metabolic phenotype data were included in the model (Table 5.4).
Figure 5.5 shows the variable importance plot which indicates that 24 out of the top 28 most discriminating variables constitute ^{1}H-NMR integration values.
These results further confirm that ^{1}H-NMR metabolic phenotype information has added value for disease prediction. Besides the number of smoking pack years, the presence of COPD and the smoking habits, the plasma concentration of threonine (VAR49 and 50), glycerol (VAR45 and 46) and valine (VAR48) contribute the most to the discriminative power of the model. Thus, in addition to known clinical risk factors for lung cancer (smoking and the presence of COPD), altered plasma levels of these metabolites seem to be risk factors for lung cancer as well.

Table 5.4: *Prediction performance parameters resulting from the penalized logistic regression models using lasso and elastic net penalties and the random forests analysis. Data are presented as percentage ± standard deviation. Abbreviations: MCE: misclassification error, PPV: positive predictive value, NPV: negative predictive value, SEN: sensitivity, SPE: specificity.*

		MCE (%)	PPV (%)	NPV (%)	SENS (%)	SPEC (%)
Logistic model	Test 1	24.6 ± 2.7	75.7 ± 4.2	75.4 ± 4.5	76.6 ± 4. 7	74.4 ± 4.8
(clinical)	Test 2	23.0 ± 1.7	80.7 ± 1.2	74.2 ± 2.2	71.0 ± 3.5	83.0 ± 2.1
Lasso	Test 1	18.6 ± 3.4	76.8 ± 5. 7	74.8 ± 5.4	75.4 ± 5.3	76.2 ± 6.1
(Clinical + metabolic phenotype)	Test 2	22.6 ± 3.4	81.5 ± 2.9	74.5 ± 4.0	70.9 ± 5.8	84.0 ± 2.6
Elastic net	Test 1	17.8 ± 3.5	77.3 ± 5.9	75.2 ± 5.6	75.7 ± 5.5	76,7 ± 6.3
(Clinical + metabolic phenotype)	Test 2	21.4 ± 3.4	81.9 ± 2.7	76.1 ± 4.1	73.3 ± 5.8	83.9 ± 2.2
Random Forests	Test 1	22.7 ± 0.9	76.2 ± 0.6	78.6 ± 1.5	80.6 ± 1.8	73.9 ± 0.8
(Clinical)	Test 2	28.6 ± 2.3	73.9 ± 2.2	69.5 ± 2.4	66.3 ± 3.2	76.6 ± 1.9
Random Forests	Test 1	18.2 ± 0.6	84.6 ± 0.7	79.3 ± 0.7	78.6 ± 0.8	85.1 ± 0.8
(clinical + metabolic phenotype)	Test 2	12.3 ± 0.5	88.3 ± 0.8	87.2 ± 0.4	87.1 ± 0.4	88.5 ± 0.9

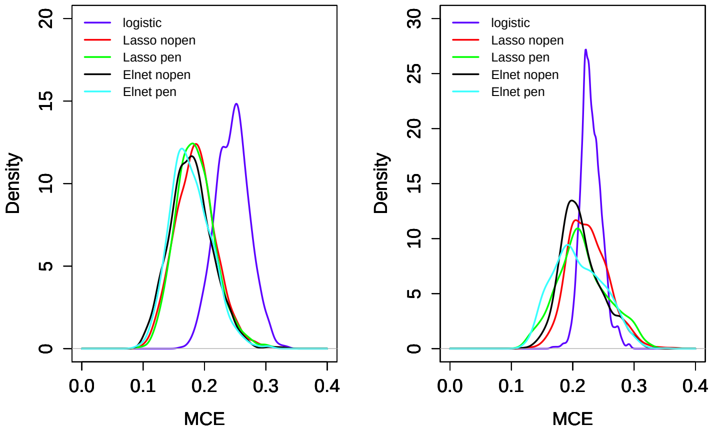


Figure 5.4: *Density estimates for the distribution of the misclassification error. Left panel: density plot for test cohort 1. Right panel: density plot for test cohort 2. Abbreviations: Elnet: elastic net, nopen: not penalized, pen: penalized.*

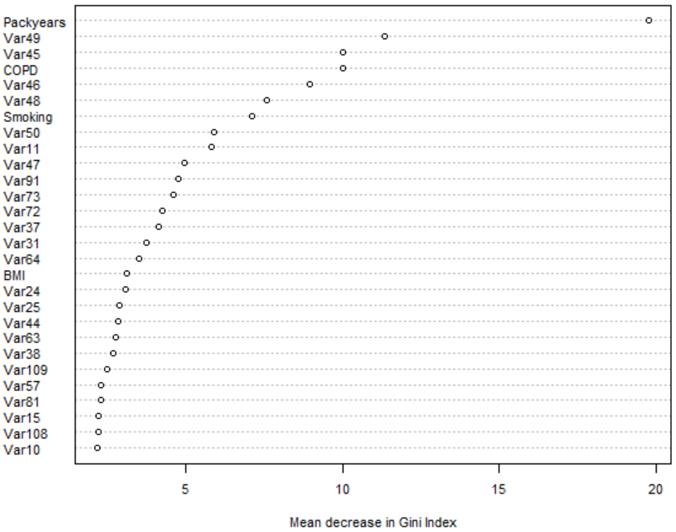


Figure 5.5: *Variable importance plot obtained from a single random forests analysis including both clinical risk factors and ^{1}H-NMR metabolic phenotype data (top 28 important variables). Abbreviations: BMI: body mass index, COPD: chronic obstructive pulmonary disease.*

The left panel in Figure 5.6 shows that the 28 most discriminating variables were almost always selected (in the top 30 important variables) in all 1000 runs of the RF analysis.

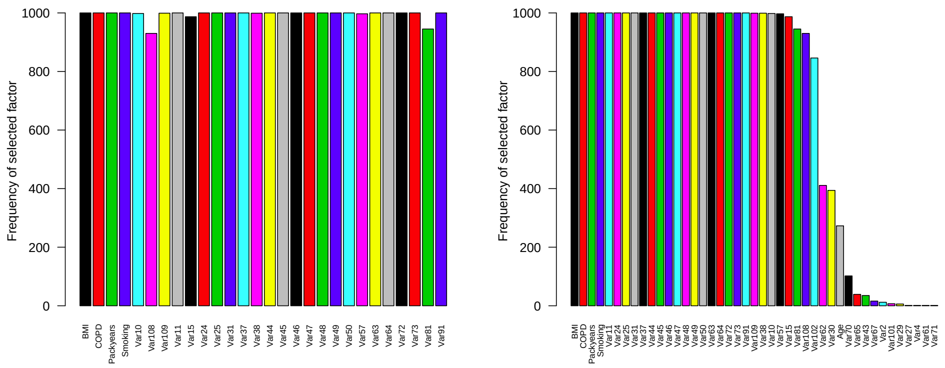


Figure 5.6: *Frequency of the selection of variables in 1000 runs of random forests. Left panel: variables with the highest importance score in the single run presented in Figure 5.5; Right panel: All variables selected at least once (in the top 30 important variables in each run).*

5.4 Discussion

In this chapter, we have shown that the addition of ^{1}H-NMR metabolic phenotype data improves the MCE of classical risk models that only take clinical risk factors into account. Both penalized logistic regression models using lasso and elastic net penalties and the RF indicate that the addition of ^{1}H-NMR metabolic phenotype data to classical risk models that only take clinical risk factors into account reduces the MCE. These findings are comparable to those found for studies in which the addition of genetic risk markers and mutagen sensitivity data improved the performance of risk models including only clinical risk factors (Raji *et al.*, 2010; Spitz *et al.*, 2008). For the random forest analysis, the improvement in MCE is shown to reach 16%. It is demonstrated that the inclusion of NMR metabolic fingerprint data has potential to improve the selection of high-risk individuals eligible for lung cancer screening with LDCT. The proposed methodology paves the way to a reduction of the false positive rate and corresponding financial burden.

Part II

High Dimensional Biomarkers in Drug Discovery

Chapter 6

Development of High Dimensional Biomarkers Within the QSTAR Framework

6.1 The QSTAR Framework

The drug discovery and development process are costly and time consuming. The development of a new drug costs in the order of billion US dollar (depending on the drug type) and takes about a decade (Adams and Brantner, 2006; DiMasi *et al.*, 2003). It is crucial to identify early failure and thereby save time and investment. The decision to continue/stop a development process in drug discovery must be made during all phases (Cowlrick *et al.*, 2011; Fischer and Heyse, 2005). The decision should ideally be based on scientific parameters that are predictive of later outcomes, and which can be determined quickly and at relatively low cost. A major problem during early drug discovery is the time gap between the compounds selection and the identification of potential undesirable effects (off target effects) such as toxicity. As pointed out by Verbist *et al.* (2015), in many cases the off target effects are discovered in a late stage of the development which means that a lot of resources could be saved if the off target effects associated with a new compound could be identified earlier.

Currently, microarray technology is used to monitor simultaneously the activity of thousands genes and their response under certain experimental condition(s) (Amaratunga *et al.*, 2014). Therefore, microarray data can be used in order to identify many new additional targets for drug discovery. The detection of biochemical pathways, genes or proteins that are disturbed by a disease, or linked to a certain treatment is made easy by the use of technologies such as chemoinformatics, genomics, proteomics.

In the second part of the dissertation, we focus on the Quantitative Structure-Transcriptional-Assay Relationships (QSTAR, Verbist *et al.*, 2015; Perualila *et al.*, 2016) modeling framework. Early drug discovery research and development process generates multiple sources of high dimensional data such as high-throughput screening (HTS), chemical structures, gene expression, image-based high-content screening (HCS, Grepin and Pernelle, 2000; Mayr and Bojanic, 2009). This type of high-dimensional data is characterized by a high number of features (variables) and relatively small number of samples. Within the QSTAR framework, transcriptomics data are integrated with structural compound information as well as bioactivity data in order to analyze compound effects in biological systems (Verbist *et al.*, 2015).

Perualila *et al.* (2016) proposed a joint modeling (JM) approach to integrate the three data types. Their modeling approach allows to (1) identify gene signatures associated with bioactivity of chemical structures, (2) determine chemical substructures (fingerprint features, FF) of compounds that are related with effects to the bioassay data for target(s) of interest and (3) to determine whether this effect can also be confirmed by the gene expression changes. In this part of the dissertation, we propose the use of Structural Equation Modeling (SEM, Bollen, 1989; Danner *et al.*, 2015; Loehlin, 1998) in combination with model averaging techniques (Burnham and Anderson, 2003; Kuiper *et al.*, 2014; Lin *et al.*, 2012; Claeskens and Hjort, 2008) to compare and select the causal structure that best fits the data. SEM allows investigating several causal models that could explain the relationship between the chemical structure and biological activity with the gene expression. The JM and SEM modeling approaches are performed using a gene-by-gene analysis. In case a joint analysis is of interest, we propose to use supervised principal component analysis (SPCA, Bair *et al.*, 2006), lasso (Tibshirani, 1996) and elastic net (Zou and Hastie, 2005) methods to combine information from the three data sources.

6.2 Case Studies

6.2.1 EGFR Data

The dataset used in this part of the dissertation was obtained from an oncology discovery project (EGFR project), which was focused on the inhibition of the fibroblast growth factor receptor (FGFR). In total, 35 compounds and one DMSO control compound were profiled on SKOV3 ovarian carcinoma cells and all compounds showed strong FGFR inhibition in a cellular assay experiment. Different concentration of the compounds were used. At the end of incubation, cell growth was determined. The concentration at which the cell growth is reduced by 50% was retained. In all analyses, the pIC50 scale (-log IC50) is used.

For the microarray experiment, human ovarian carcinoma cells were seeded in flasks and then compound was added. After the pre-processing steps, the gene expression data contain 3595 genes. This dataset was used in the analysis of Verbist *et al.* (2015) and Perualila *et al.*, (2016). Both publications highlighted a particular chemical feature that was linked to both cell growth activity and gene down regulation which was detrimental to biological activity and, in turn, very likely also to the efficacy of the compound. The identified chemical structure (also referred as fingerprint feature) was used to demonstrate the gene-by-gene structural equation modeling approach.

6.2.2 ROS Data

The ROS project sought to develop compounds that inhibits ROS1 (reactive oxygen species). ROS1 is highly-expressed in a variety of tumor cell lines and belongs to the sevenless subfamily of tyrosine kinase insulin receptor genes. The dataset consists of 89 compounds tested for target inhibition. A total of 7100 genes were used in the analysis.

6.2.3 Data Structure

As mentioned above, these datasets contained information about: (1) chemical structure of the compounds (measured by fingerprint features that represent different substructures in the compounds), (2) gene expression data, and (3) biological activity of the compounds (measured by pIC50 for different bioassays). Our aim is to model the association between the gene expression and the bioactivity variable(s) taking into account the chemical structure.

Let x_{ji} be the jth gene expression for the ith compound, $j = 1, 2, \ldots, m$ and $i = 1, 2, \ldots, n$; y_{bi}, $b = 1, \ldots, B$, be the bth the bioactivity variable of the ith compound measured on the bth assay; and z_{ki} , $k = 1, \ldots, K$, be the kth fingerprint feature in the ith compound. Note that z_{ki} is a binary indicator, $z_{ki} = 0$ represents the absence of the kth chemical structure in the ith compound while $z_{ki} = 1$ denotes the presence of the kth chemical structure.

The three data sources can be represented as matrices given by

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdot & y_{1n} \\ y_{21} & y_{22} & \cdot & y_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ y_{B1} & y_{B2} & \cdot & y_{Bn} \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} z_{11} & z_{12} & \cdot & z_{1n} \\ z_{21} & z_{22} & \cdot & z_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ z_{K1} & z_{K2} & \cdot & z_{Kn} \end{bmatrix},$$

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdot & x_{1n} \\ x_{21} & x_{22} & \cdot & x_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ x_{m1} & x_{m2} & \cdot & x_{mn} \end{bmatrix}.$$

Note that for the analysis presented in this part of the dissertation, we used one fingerprint feature and one bioactivity variable.

Chapter 7

Joint Modeling of Bioassay Data and Genes Expression in Drug Discovery Experiments: A Supervised Principal Component Analysis, Lasso and Elastic Net Approaches

7.1 Introduction

Microarray technology is extensively used in biological and medical studies to monitor simultaneously the activity of thousands of genes and their response under certain condition(s) (treatment, disease status, time, etc.). In some experiments, in addition to gene expression, other variables are available and the question of interest is to identify whether or not the gene expressions can serve as biomarkers for a response of interest (Perualila *et al.*, 2016; Tilahun *et al.*, 2010; Chen *et al.*, 2008). This results in high dimensional data characterized by a high number of variables and few samples. Furthermore, although the number of genes is large, it is expected that only a small number of genes will be associated with the conditions under investigation or

with a response variable of interest. In this chapter we aim at developing predictive models to integrate the different data sources. We first review the joint modeling approach by Perualila *et al.* (2016) and the conditional model which are implemented as gene-specific models. Both modeling approaches allow to identify potential genetic biomarker for compound efficacy as measured by pIC50. Next to the gene-specific approach, we study how to combine information from many genes to form a joint biomarker using methods such as supervised principal component analysis (SPCA, Bair *et al.*, 2006), lasso (Tibshirani, 1996) and elastic net (Zou and Hastie, 2005).
This chapter is organized as follows: in Section 7.2, we introduce the joint and conditional models. In Section 7.3, the supervised principal component analysis (SPCA) method is presented. Section 7.4 describes the cross validation procedure used in the analysis. Predictive models using lasso and elastic net penalties are presented in Section 7.5. An application to the data is given in Section 7.6, followed by a discussion in Section 7.7.

7.2 Joint and Conditional Models for Gene Expression and Bioassay Data

7.2.1 Gene Specific Approach

Let X_{ji} be the jth gene expression ($j = 1, 2, ..., m$), of the ith compound ($i = 1, 2, ..., n$). The measurement for the bioactivity (pIC50 in our setting) is denoted by Y_i. Let Z_i be an indicator variable, which takes the value 1 if the fingerprint feature is present in the ith compound, and 0 otherwise. Perualila *et al.* (2016) consider the following gene-specific joint model:

$$\begin{pmatrix} X_{ji} \\ Y_i \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_{X_j} + \alpha_j Z_i \\ \mu_Y + \beta Z_i \end{pmatrix}, \Sigma_j\right]. \tag{7.1}$$

The gene-specific covariance matrix Σ_j is given by:

$$\Sigma_j = \begin{pmatrix} \sigma_{jj} & \sigma_{jY} \\ \sigma_{jY} & \sigma_{YY} \end{pmatrix}. \tag{7.2}$$

The parameters α_j and β are the fingerprint feature effects upon the jth gene and bioactivity variable, respectively. The parameters μ_{X_j} and μ_Y are the gene-specific and bioactivity intercepts, respectively.
The covariance matrix (7.2) can be used to quantify the association between the gene expression and the bioactivity after correcting for the fingerprint feature effects using

the adjusted correlation given by:

$$\rho_j = \frac{\sigma_{jY}}{\sqrt{\sigma_{jj}\sigma_{YY}}}. \tag{7.3}$$

Note that ρ_j is the gene-specific correlation coefficient between gene expression and the bioactivity variable after adjusting for fingerprint feature effect. The joint model (7.1) implies the following conditional model (Burzykowski *et al.*, 2005):

$$Y_i = \tilde{\mu} + \tilde{\beta}_j Z_i + \tilde{\alpha_j} X_{ji} + \tilde{\varepsilon}_{ij}. \tag{7.4}$$

Here, $\tilde{\mu} = \mu_Y - \sigma_{jY}\sigma_{jj}^{-1}\mu_j$, $\tilde{\beta}_j = \beta - \sigma_{jY}\sigma_{jj}^{-1}\alpha_j$, $\quad \tilde{\alpha_j} = \sigma_{jY}\sigma_{jj}^{-1}$, and $\tilde{\varepsilon}_{ij} \sim N(0, \sigma_{YY} - \sigma_{jY}^2\sigma_{jj}^{-1})$.

The above models are fitted gene-by-gene and it is expected that only a small number of genes is associated with the bioactivity variable. In the next section we describe a method which allows to combine information from many genes to form a joint biomarker (gene profile) for prediction.

7.3 Supervised Principal Component Analysis Approach

The gene-specific joint modeling approach in the previous section allows to identify individual genes associated with pIC50 (the bioactivity variable). The next section focuses on the question of how to combine information about the expression level from many genes in order to form a ”gene profile”?

In the microarray setting, the number of predictors (m) is larger than the sample size (n) and the design matrix $\mathbf{X}$ is likely to be singular which makes linear regression model comprising all predictors (genes) no longer feasible.

One way to overcome this problem is to reduce the dimension of the design matrix using principal component analysis (PCA) for example. However, a drawback of PCA is that there is no guarantee that the principal components are associated with the response of primary interest (Bøvelstad *et al.*, 2007; Alter *et al.*, 2000; Tilahun *et al.*, 2010).

Bair *et al.* (2006) proposed the supervised principal components (SPCA) method if $m >> n$. SPCA is similar to conventional principal components analysis except that one only uses predictors with the strongest estimated association with the response. The SPCA relies on the underlying assumption that there is a latent variable $U(X)$ (the gene profile), which is associated with the response variable Y.

SPCA consists of the following steps done using a cross validation:

- The first step consists of fitting a gene-specific conditional model (7.4) to all genes.

- At the second step, select k genes ($k << m$) having the strongest estimated association with the response. Form a reduced gene expression matrix $\mathbf{X}_k$.

- At step three, the first principal component is constructed using the reduced gene expression matrix ($\mathbf{X}_k$) from second step. To choose the optimal number of genes to use in the first principal component, step three is an iterative procedure: the first iteration consists of the top 2 genes selected in the first step, and at each iteration a new gene is added until all genes in the reduced matrix are used ($2 \leq i \leq k$).

Let $\hat{U}(X)$ be the first principal component of the reduced matrix. We consider the following linear regression model:

$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2 \hat{U}(X)_i + \varepsilon_i. \tag{7.5}$$

Note that the gene profile $\hat{U}(X)$ is the estimate for the latent variable $U(X)$. Since our aim is to construct a predictive model, in step four, the model (7.5) is fitted to find top k genes from which the first PC maximizes the goodness of fit measure that will be discussed in Section 7.4.1.

7.4 Cross Validation

The above four steps are run within a 3-fold cross validation repeated 1000 times. At the end, the genes are ranked based on their selection frequencies. The procedure is represented schematically in Figure 7.1.

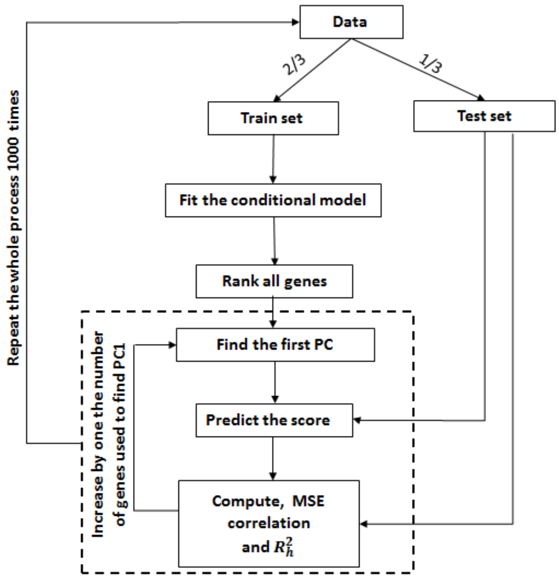

Data
2/3
1/3
Train set
Test set
Fit the conditional model
Rank all genes
Find the first PC
Predict the score
Compute, MSE correlation and R_h^2
Repeat the whole process 1000 times
Increase by one the number of genes used to find PC1

Figure 7.1: *SPCA. Flowchart of the steps involved in the SPCA method.*

Features selection and gene profile construction are done using the train set and evaluation on the test set. The square (broken line) in Figure 7.1 shows the iterative step (the internal loop), in which at each iteration a new gene is added to build a new PC. For the analysis presented below, a second cross validation loop is conducted using fixed list of genes for the signature.

7.4.1 Information-theoretic Approach

Selection of the top k genes and evaluating the quality of the gene profile as a biomarker to bioactivity variable can be done using a surrogacy measure that was developed within the information theoretic approach (Alonso and Molenberghs, 2007) in the context of randomized clinical trials. Following this approach, a gene profile is called a good biomarker for the bioactivity variable if a ”large” amount of uncertainty about the bioactivity variable is reduced when the gene profile is known. Consider the following models:

$$\begin{cases} M_0 : Y_i = \mu_Y + \beta_1 Z_i + \varepsilon_{0i}, \\ M_1 : Y_i = \theta_0 + \theta_1 Z_i + \theta_2 \hat{U}(X)_i + \varepsilon_{1i}. \end{cases} \tag{7.6}$$

Note that the first equation in (7.6) relates the expected value of the bioactivity variable only to the fingerprint, while the second connects both the gene profile and the fingerprint to the expected value of the bioactivity variable. As shown by Alonso and

Molenberghs (2007), the degree of association between the gene profile and the bioactivity variable can be measured by the squared informational coefficient of correlation (SICC) given by:

$$R_h^2 = 1 - exp\left(\frac{-G^2}{n}\right). \tag{7.7}$$

Here, G^2 denotes the likelihood ratio statistic to compare the two models in equation (7.6), and n is the sample size. R_h^2 satisfies a number of useful properties: it ranges in the unit interval and is equal to zero when $\hat{U}(X)$ and Y are independent. The first PC is constructed in order to maximize R_h^2, i.e., using surrogacy terminology we can say that the gene profile is constructed in order to maximize a surrogacy measure. From that point of view, given the expression matrix, $\hat{U}(X)$ is the "best" biomarker for Y. Note that although in our setting Y is assumed to be a normally distributed variable, R_h^2 is still valid for other distributions of Y.

Since $U(X)$ is constructed from the high dimensional expression matrix, as shown in Figure 7.1, R_h^2 is calculated on the test set and, in addition, we evaluate the gene profile as a predictor for Y using the mean squared error given by

$$MSE = mean((Y_{test} - \hat{Y}_{test})^2).$$

7.5 Predictive Models Using the Lasso and the Elastic Net

7.5.1 Lasso

In this section, we discuss two different approaches for gene selection (while taking into account the fingerprint feature), namely lasso (Tibshirani, 1996) and elastic net (Zou and Hastie, 2005). Tibshirani (1996) proposed to minimize the residuals sum of squares subject to the constraint $\sum_{j=1}^{m} |\beta_j| \leq t$. The regression coefficients are estimated in order to minimize the following expression (Hastie *et al.*, 2015):

$$\underset{(\gamma,\beta)}{\text{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \gamma z_i - \sum_{j=1}^{m} x_{ji}\beta_j)^2 + \lambda \sum_{j=1}^{m} |\beta_j| \right\}. \tag{7.8}$$

Note that the fingerprint feature's coefficient γ is not penalized. It implies that this variable will always be included in the model. Due to the penalty term $\lambda \sum_{j=1}^{m} |\beta_j|$ some of the coefficients are shrunk toward zero resulting in variables selection. The lasso penalty causes the estimates of the non-zero coefficients to be biased toward zero

(Hastie *et al.*, 2015). One way to reduce this bias is to use the lasso to identify non-zero coefficients, and then fit an unrestricted linear model to the selected variables. Alternatively, one can apply lasso again on the selected variable (i.e., lasso after lasso). This is known as the relaxed lasso (Meinshausen, 2007; Hastie *et al.*, 2015). The penalty parameter, λ, is chosen by cross validation and the optimal value of λ is chosen to be the minimizer of the mean squared error (Hastie *et al.*, 2015). The lasso method has some limitations. It can only select a number of features (genes) at most equal to the number of samples. When a group of highly correlated variables exists, lasso method will only select one of the variables from the group and does not care which one is selected (Zou and Hastie, 2005).

7.5.2 Elastic Net

To overcome the limitations of the lasso penalty, Zou and Hastie (2005) proposed the elastic net penalty. This penalty has the following form: $\lambda \sum_{j=1}^{m}[(1-\alpha)\beta_j^2+\alpha|\beta_j|]$. It is a mixture of the lasso and the ridge regression penalties. The elastic net coefficients are the minimizer of (Hastie *et al.*, 2015)

$$\underset{(\gamma,\beta)}{\operatorname{argmin}} \left\{ \frac{1}{2N}\sum_{i=1}^{N}(y_i-\beta_0-\gamma z_i-\sum_{j=1}^{m}x_{ji}\beta_j)^2+\lambda\sum_{j=1}^{m}[\frac{1}{2}(1-\alpha)\beta_j^2+\alpha|\beta_j|] \right\}. \quad (7.9)$$

Similar to lasso, the elastic net provides an automatic variable selection procedure due to the shrinkage of the parameters toward zero and it can select groups of correlated variables. Again, the parameter γ corresponding to the fingerprint feature variable in equations (7.9) is not penalized. The shrinkage and the mixing parameters , λ and α respectively, are determined by an internal cross validation. Note that for $\alpha = 1$ the elastic net solution is identical to the lasso solution while for $\alpha = 0$ it is identical to ridge regression. For the predictive model, a cross validation loop of B iterations is performed (see an illustration in Figure 7.2). At each iteration, the predictive model consists of k genes with non-zero coefficients selected using lasso or elastic net. These genes are kept for further analyses.

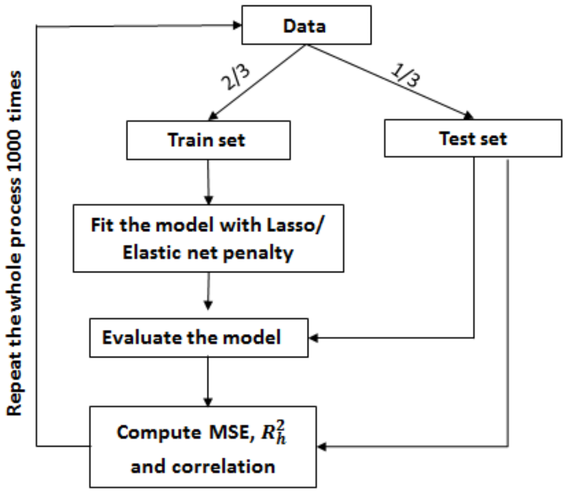

Data
2/3
1/3
Train set
Test set
Fit the model with Lasso/ Elastic net penalty
Evaluate the model
Compute MSE, R_h^2 and correlation
Repeat the whole process 1000 times

Figure 7.2: *Penalized regression models. Flowchart of the steps involved in the development of the predictive model.*

7.5.3 Fixed Gene List

At the end of the 3-fold cross validation repeated 1000 times, two lists of genes with the highest selection frequencies are produced (one list per each penalty). Extra 1000 iterations, with a 3-fold cross validation, are run on each fixed list to assess the predictive model.

Within the loop, the optimal λ value is determined on the train for fixed k genes ($k = 2, 3, ...11$, for lasso and $k = 2, 3, ...20$, for elastic net). For each value of k, the process is repeated 1000 times. A grid of the elastic net mixing parameter is also provided ($\alpha = 0, 0.1, ..., 1$). The squared correlation (ρ^2) between the observed and predicted values, MSE and the squared informational coefficient of correlation (SICC, in equation (7.7)) from information theory (Alonso and Molenberghs, 2007) are used to assess the model fitted on the test set.

7.6 Application to the Data

7.6.1 Supervised Principal Component Analysis

7.6.1.1 Cross Validation Results

The EGFR data described in Section 6.2.1 was used in all analyses performed in this chapter. For each cross validated sample, the gene-specific models were fitted to the train dataset to rank genes (based on parameter test statistics). The iterative procedure starts with top 2 genes, and subsequently the number of the top genes was increased by one. In each split, the combination of genes giving the highest R_h^2 on the test set were recorded. Figure 7.3 (top left panel) shows the average R_h^2 with 95% confidence intervals for each top k ($k = 2, 3, ...100$). The top right panel depicts selection frequencies for the top 22 genes in all 1000 data splits. MSE and ρ^2 measures are shown in the bottom row.

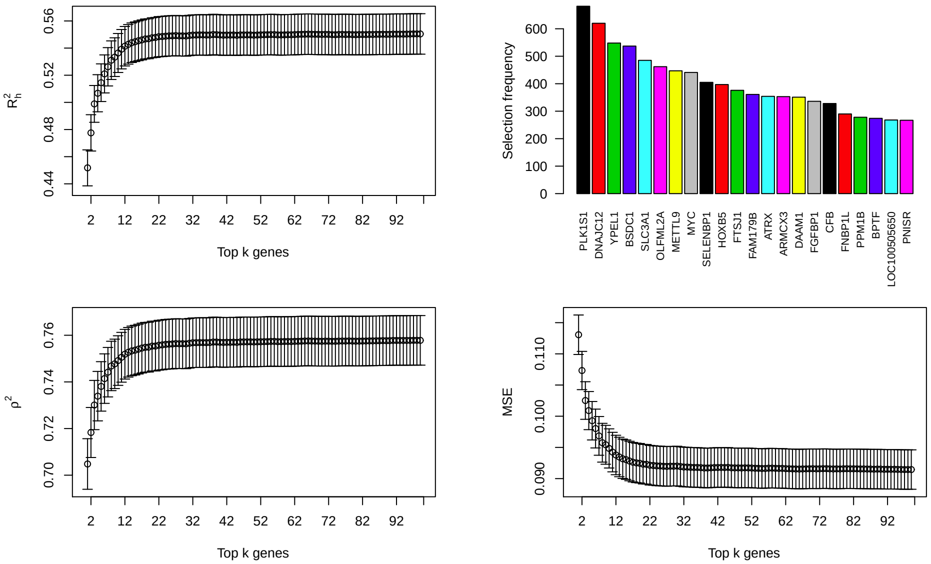


Figure 7.3: *SPCA: performance measures obtained for models using top k genes in the gene profile (on the test set); Top left panel: R_h^2 measures, Top right panel: Frequency of mostly selected genes, Bottom left:ρ^2 measures, Bottom right: MSE. All measures are obtained from 1000 cross-validated datasets.*

From the left panel of Figure 7.3, it can be seen that when 22 to 25 genes are included in the gene profile, there was not much improvement in R_h^2. Next, a list with the highest selection frequency was identified. The top right panel of Figure 7.3 shows the top 22 genes. Figure 7.4 shows the density plot for the R_h^2 and the ρ^2.

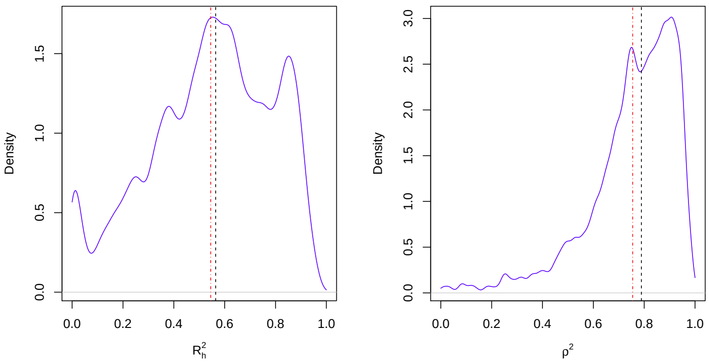


Figure 7.4: *SPCA. Density plots of R_h^2 and ρ^2 for 1000 cross validated samples. The red dotted lines present the mean, while the black dotted lines present the median.*

7.6.1.2 Fixed Features in a Gene Profile: Leave-one-out Cross Validation

In this Section, we present the results with a fixed list of 22 genes using leave-one-out cross-validation. Following Tilahun *et al.* (2010), two approaches were used to construct the gene profile.

- The first approach consists of taking the first top k genes and constructing the first principal component. Figure 7.5 (left panel) and Table 7.1 show that $R^2_{h_{cv}}$ is ranging between 0.68 to 0.744.

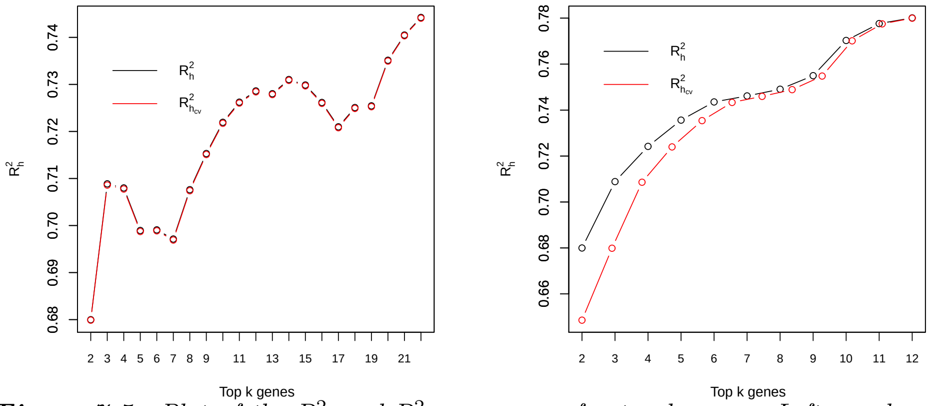


Figure 7.5: *Plot of the R_h^2 and $R^2_{h_{cv}}$ measures for top k genes. Left panel: genes selected based on the first principal component. Right panel: a given gene is added if it results in increase of R_h^2 and $R^2_{h_{cv}}$.*

Table 7.1: *SPCA results based on top k genes,R_h^2, $R_{h_{cv}}^2$: The measure of association without and with leave one cross validation.*

Top	R_h^2	$R_{h_{cv}}^2$
2	0.6800	0.6799
3	0.7089	0.7086
4	0.7080	0.7078
5	0.6990	0.6987
6	0.6991	0.6989
7	0.6971	0.6969
8	0.7076	0.7074
9	0.7153	0.7151
10	0.7219	0.7217
11	0.7262	0.7260
12	0.7286	0.7284
13	0.7280	0.7278
14	0.7310	0.7309
15	0.7299	0.7297
16	0.7262	0.7260
17	0.7210	0.7208
18	0.7251	0.7249
19	0.7254	0.7253
20	0.7351	0.7350
21	0.7405	0.7403
22	0.7442	0.7441

- With the second approach, a gene is included as a part of the joint biomarker if it results in an increase of the R_h^2. For this approach, 12 genes were added in the construction of the biomarker, giving an $R_{h_{cv}}^2$ value of 0.78, which is higher than taking, for example, the top 22 genes at once. Figure 7.5 (right panel) and Table 7.2 show the results obtained with and without leave-one-out cross validation.

Table 7.2: *SPCA results based on top k genes,R_h^2, $R_{h_{cv}}^2$. Only genes which increase R_h^2 are used to build the first PC.*

Top	R_h^2	$R_{h_{cv}}^2$
2	0.6800	0.6799
3	0.7089	0.7086
4	0.7242	0.7240
5	0.7356	0.7354
6	0.7435	0.7433
7	0.7461	0.7460
8	0.7491	0.7489
9	0.7550	0.7548
10	0.7703	0.7701
11	0.7776	0.7775
12	0.7801	0.7800

7.6.1.3 Fixed Features in a Gene Profile: 3-fold Cross Validation.

In this section, the analysis was performed using fixed lists of features in gene profile. For each top k value, a loop of 3-fold cross validation with 1000 iterations was run.

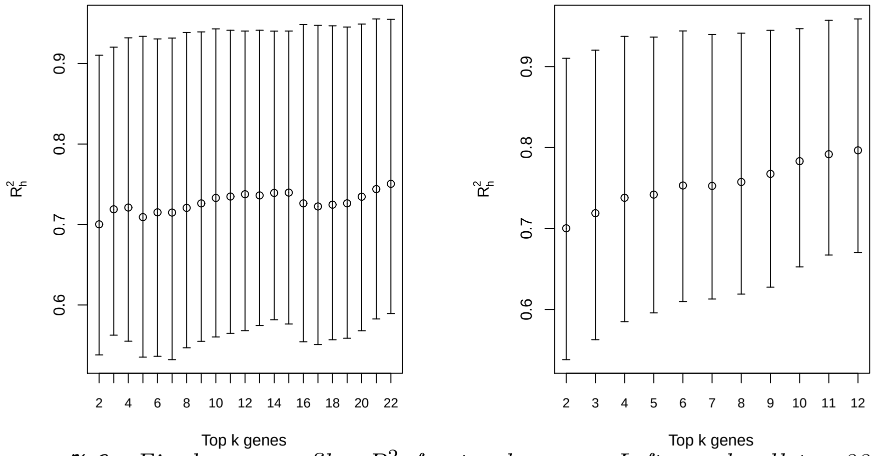


Figure 7.6: *Fixed gene profile. R_h^2 for top k genes. Left panel: all top 22 genes; Right panel: genes from the list with increasing R_h^2.*

Figure 7.6 shows the quantiles ($25th$, median, $97.5th$) of R_h^2 measures for different values of top k. The left panel shows the quantiles for the top 22 genes, while the right panel shows the quantiles for the top 12 genes used in the second approach. As expected, for the short list, the association measures increase as the number of features increases in the first principal component.

7.6.2 Lasso and Elastic Net

7.6.2.1 Cross Validation Results

For each iteration, features having non-zero coefficients were kept. Figure 7.7 (top left panel) shows the selection frequencies for top 11 genes. The top left panel and lower panel depict the density plots for both MSE and the squared correlation between the observed and the predicted values on the test set .

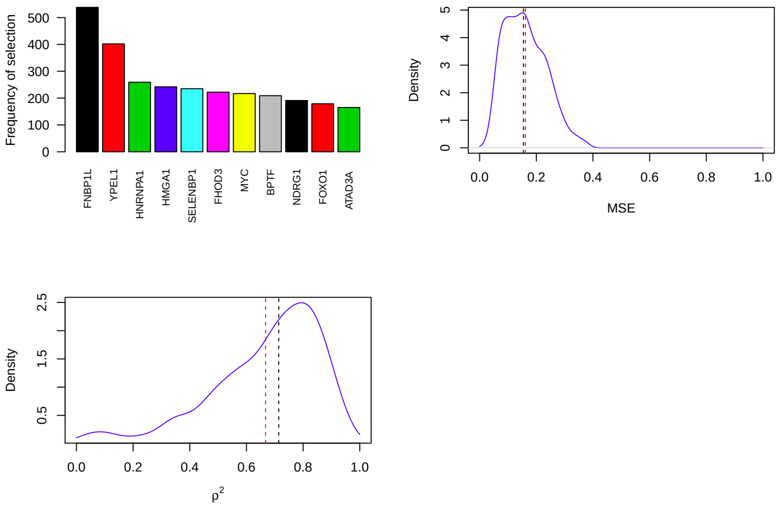


__Figure 7.7:__ Lasso. Upper left panel: selection frequency for top genes; Upper right panel: MSE density plot; Bottom left panel: ρ^2 density plot (red dotted lines present the mean, the black dotted lines present the median). Results from 3-fold cross-validation repeated for 1000 times with lasso.

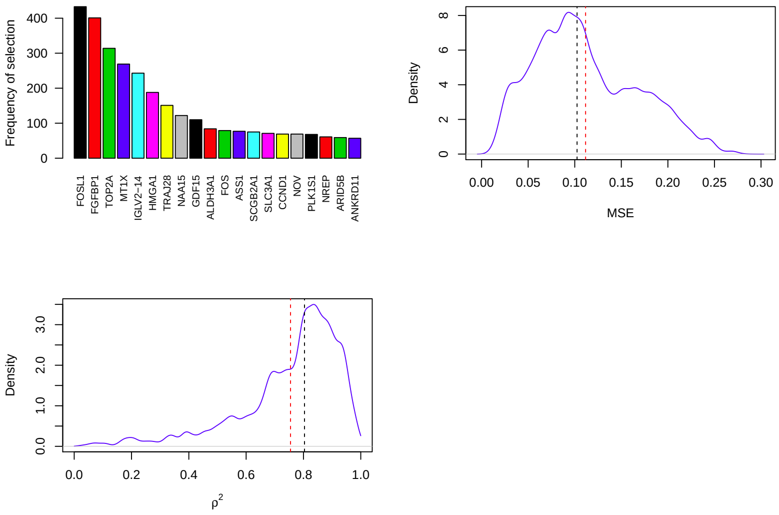


__Figure 7.8:__ Elastic net. Upper left panel: selection frequency for top genes; Upper right panel: MSE density plot; Bottom left panel: ρ^2 density plot (red dotted lines present the mean, the black dotted lines present the median). Results from 3-fold cross-validation repeated for 500 times with elastic net.

Similar to lasso, the elastic net penalty parameters were tuned in 3-fold cross validation procedure. A range of values for the mixing parameter α was defined and, for each value, the 3-fold cross validation was repeated 500 times. For each optimal α value, genes with non zero coefficients were kept in order to find the most selected features. Figure 7.8 (top left panel) shows the selection frequencies for the top 20 genes for an elastic net model with $\alpha = 0.89$ (Figures for other α values are presented in the appendix C). The top right panel depicts the density plot for MSE. The left bottom panel shows the density plot for the squared correlation between the observed and the predicted on the test set.
As expected, the elastic net penalty, has a tendency to select more genes compared to lasso.

7.6.2.2 Fixed List of Genes

Final prediction models were constructed using top 11 and top 20 genes, for lasso and elastic net, respectively. For each top k value, the data were cross-validated 1000 times. The squared correlation (ρ^2) between the observed and predicted values and the squared informational coefficient of correlation (SICC) were used to assess the model fitted on the test set. Figure 7.9 shows the distribution of R^2_h, ρ^2 and MSE for lasso.

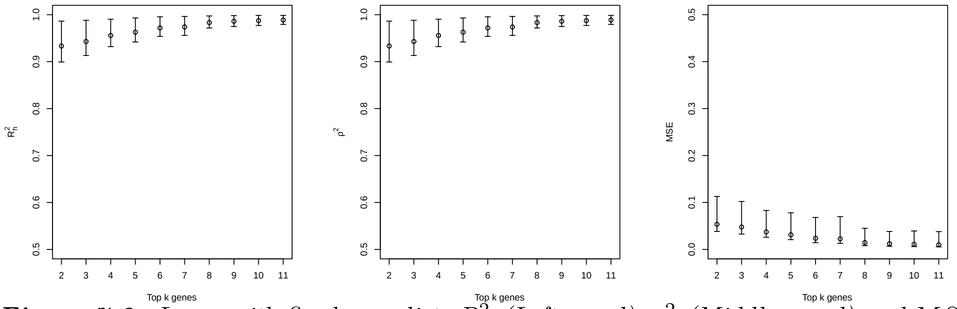


Figure 7.9: *Lasso with fixed gene list. R^2_h (Left panel), ρ^2 (Middle panel) and MSE (right panel) distributions with top k genes ($k = 2, 3, ..., 11$).*

We notice that the variability in R^2_h, ρ^2 and MCE are relatively high when the number of genes used to build the predictive model is relatively small.

Similar analyses were performed for models with elastic net penalties. A grid of α values was provided ($\alpha = 0, 0.1, 0.2, ..., 1$). As before, for a given values of α and k, the 3-fold cross validation was repeated 1000 times. Figure 7.10 presents the distributions of R_h^2, ρ^2 and MSE for the mixing parameter $\alpha = 0.5$. Results for other α values are presented in the appendix C.

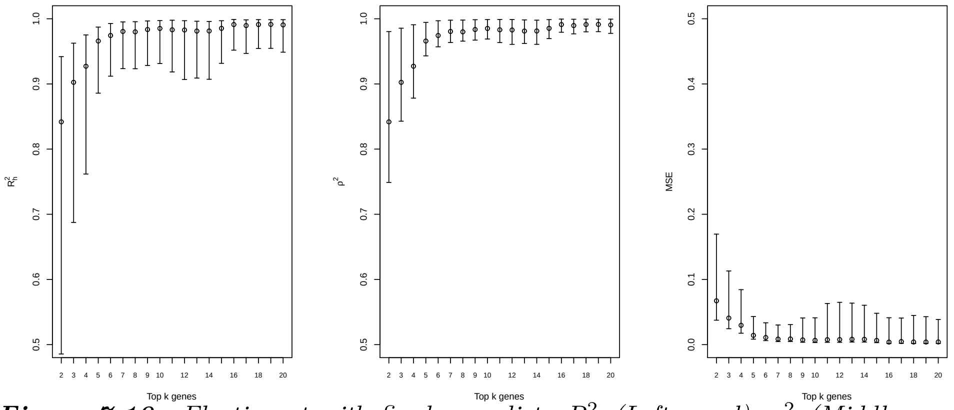


Figure 7.10: *Elastic net with fixed gene list. R_h^2 (Left panel), ρ^2 (Middle panel) and MSE (right panel) distribution for top k genes ($k = 2, 3, ..., 20$). The mixing parameter $\alpha = 0.5$.*

7.7 Discussion

The starting point of this chapter was a gene-specific model, either a joint model or a conditional model was used to identify genes which are associated with the bioactivity of the compound (measured by pIC50). Our aim in this chapter was to construct a biomarker that includes information from multiple genes. Of course, when such a biomarker is constructed the next question is how to evaluate it.

Alonso and Molenberghs (2007) argue that ”a variable is a valid surrogate for a true endpoint at the individual level, if uncertainty about true endpoint is reduced by a ”large” amount when the surrogate endpoint is known”.

In this chapter, we proposed to use a measure which is often used for the evaluation of surrogate endpoint in randomized clinical trials, namely the R_h^2. We have shown that, using data reduction methods such as SPCA, or penalized regression models, we can construct biomarkers which maximize R_h^2.

An issue that was not addressed in this chapter is inference. The question is not just how good is the biomarker but if the value of the R_h^2 is unlikely to be observed in random data. For this, a resampling based inference procedure should be developed. This is a topic for future investigation.

Chapter 8

Identification of Causal Structures in High-Dimensional Data Using Structural Equation Models

8.1 Introduction

The drug discovery and development process is typically costly and time consuming. This is largely attributed to the time gap between the lead compounds selection and the identification of their off target effects in later toxicity studies (Hughes *et al.*, 2011; DiMasi *et al.*, 2003; Adams and Brantner, 2006). As a consequence, a compound development project could be terminated at a time when substantial resources have already been spent. Therefore, it is crucial to identify early failure of candidate compounds in order to save time and investment in a later stage. In this regard, high-dimensional biological data, which can be determined quickly and at relatively low cost, could be useful to speed up the understanding of the molecular basis of disease and to examine efficacy and toxicity response of candidate drugs (Lennon, 2000; Kraljevic *et al.*, 2004; Starkuviene and Pepperkok, 2007). Recently, Verbist *et al.* (2015) demonstrated the use of transcriptomic biomarkers for compound's activity providing an early insight to the mechanism of action of a specific (or a specific set of) compound(s). Perualila *et al.* (2016) presented a joint modeling approach to identify genes that are associated with the efficacy data accounting for the chemical structure of the compounds.

In this chapter, we further exploit the relationship of the three data sources in drug discovery studies presented by Perualila *et al.* (2016) via structural equation modeling (SEM). Similar to the joint modeling approach, SEM allows for the simultaneous estimation and testing of the effect of chemical structure on the bioactivity data and the gene expression data. The SEM allows to investigate several causal models that could explain the relationship between the chemical structure and biological activity with the gene expression as the proposed mediator (Danner *et al.*, 2015; Li *et al.*, 2006). Specifically, SEM decomposes the total effect of the chemical structure on biological activity variable into direct and indirect effects. An indirect effect is the effect of the chemical structure on the bioactivity variable via the gene expression. Danner *et al.* (2015) showed that statistical test of indirect effect (or mediation effect) can be improved by assessing the fit of all possible structural models in order to avoid spurious mediation effects and to gain more scientific insights. Note that given the experimental variable in our setting, the fingerprint feature is treated only as an explanatory variable and some causal structures originally proposed by Danner *et al.* (2015) were not considered. This point is further discussed in Section 8.2.2. In this chapter, we extend the approach presented in Danner *et al.* (2015) and propose the use of SEM in combination with model averaging techniques (Burnham and Anderson, 2003, 2004; Claeskens and Hjort, 2008) to compare and select causal structures that best fit the data. A gene specific SEM is fitted with an aim to classify genes according to the causal structure between the bioactivity variable, chemical structure and gene expression.

This chapter is organized as follows: in Section 8.2.1 we present the path analysis model and describe how the chemical structure effect can be decomposed into the direct and indirect effects. The model averaging technique to compare between different causal structures and select the most plausible is presented in Section 8.2.3. The proposed method is applied to the data in Section 8.3, followed by a discussion in Section 8.4.

8.2 Methodology

8.2.1 Structural Equation Modeling

Structural equation modeling (SEM) approach has been widely used in many fields, such as economics, sociology and psychology (Bollen, 1989; Loehlin, 1998; Jöreskog, 1993). The key idea behind SEM is that the causal relationships among the variables determine the expected pattern of correlation (Li *et al.*, 2006).

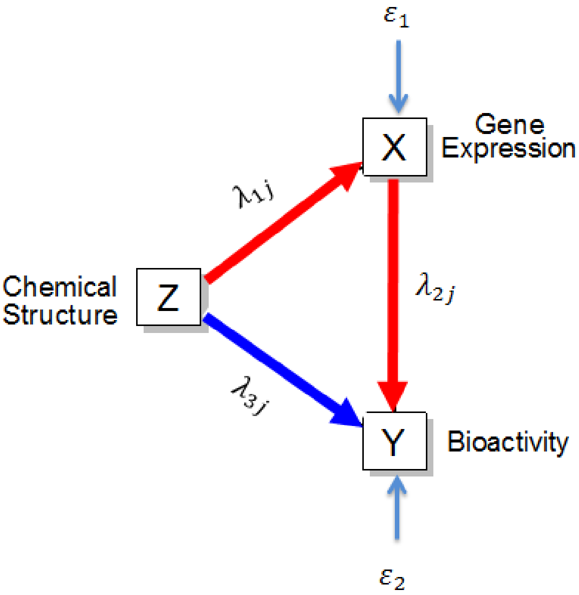


Figure 8.1: *Path diagram. Fingerprint feature indirect effect on the pIC50 through the gene expression (red arrows). Direct effect is represented by the blue arrow.*

In this chapter, a path analysis model is considered, i.e., a SEM with three observed variables: (1) the jth gene expression of the ith compound, X_{ji}, $j = 1, 2, \ldots, m$, and $i = 1, 2, \ldots, n$; (2) the bioactivity variable (pIC50) denoted by Y_i, and (3) the chemical structure or fingerprint (FP) feature denoted by Z_i, an indicator variable, which takes value 1 if the fingerprint feature is present in the ith compound, and 0 otherwise. Figure 8.1 presents a path diagram, an example of a directed acyclic graph (DAG), displaying the causal relation between the three variables (Greenland *et al.*, 1999). The path diagram shown in Figure 8.1 corresponds to the two models given by:

$$\begin{aligned} X_{ji} &= \lambda_{1j} Z_i + \varepsilon_{1i}, \\ Y_i &= \lambda_{3j} Z_i + \lambda_{2j} X_{ji} + \varepsilon_{2i}. \end{aligned} \tag{8.1}$$

Here, λ_{1j} and λ_{3j} are the fingerprint effects on the jth gene expression and the pIC50, respectively. The parameter λ_{2j} is the gene-specific effects on the pIC50. It is further assumed that (Bollen, 1989):

$$\begin{aligned} &\varepsilon_{ki} \sim N(0, \delta_k^2), \quad \text{k=1,2}, \\ &cov(\varepsilon_{1i}, \varepsilon_{2i}) = 0, \\ &var(Z) = \phi. \end{aligned} \tag{8.2}$$

The path analysis model specified in equation (8.1) allows to distinguish direct, indirect, and total effects of the fingerprint feature Z on the pIC50. Three possible direct effects can be seen from Figure 8.1: λ_{3j}, the direct effect of the FP feature on the pIC50: λ_{1j}, a gene-specific direct effect of the FP feature on the expression level of jth gene, and λ_{2j}, a gene-specific direct effect of the jth gene on the pIC50. A variable Z is said to have an indirect effect on the variable Y if the effect of Z on Y is mediated by at least one intervening variable (Bollen, 1989). In the path analysis model formulated in equation (8.1), the fingerprint feature has an indirect effect on the bioactivity variable through the jth gene expression. This indirect effect is marked by red arrows in Figure 8.1, and equals to $\lambda_{1j} \times \lambda_{2j}$. The sum of the direct and indirect effects is equal to the total effect of the variable Z on the variable Y (Bollen, 1989).

8.2.2 Causal Structure

Hypothesis tests for coefficients in equation (8.1) allow to investigate whether the FP feature has a significant direct and indirect effect on pIC50. However such statistical tests are based on the structural model specified in equation (8.1) while other causal structures are not considered (Danner *et al.*, 2015; Fiedler *et al.*, 2011). For some genes, a path analysis model containing indirect effect might not be the best model and there is no reason to exclude alternative models which represent different causal structures.

Note that there is a fundamental difference between the approach presented by Danner *et al.* (2015) and the approach we present in this chapter. With three-variate structures (gene expression, FP feature and pIC50), there are 27 possible causal structures for the triplet (X_{ij}, Y_i, Z_i). In the drug discovery setting, the FP feature variable is a design variable and it is not affected by neither the gene expression nor the bioactivity variable. These constraints reduce the number possible causal structures to 12. Figure 8.2 left panel shows one of the discarded structures.

In addition, since the primary interest is placed on the effect of gene expression and fingerprint feature on the bioactivity variable (i.e. QSAR and QSTAR), four extra causal structures are excluded. Figure 8.2 right panel shows a typical structure which is not consider (arrow labeled a). If both directions (a and b) are of interest the joint model formulated in equation (7.1) can be used.

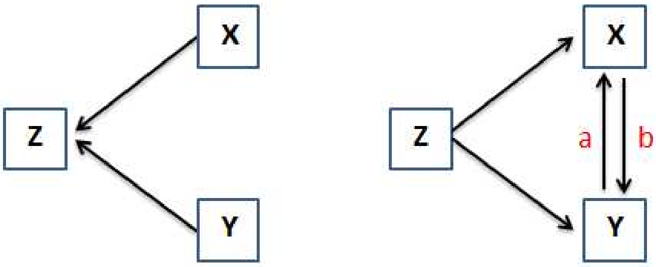


Figure 8.2: *Some of the causal structures ruled out by the experiment design*

Figure 8.3 shows a set of 8 causal models considered in our analysis.

- The causal structure represented by Model 1, the independent model, assumes that none of the variables affects each other. Given that the FP feature was chosen in a such way to maximize the difference on pIC50, this structure is less probable to be found.

- For Model 2, the causal structure corresponds to genes for which the FP feature affects the pIC50, but this effect is not mediated via the gene expression. The only effect in this causal structure is a direct effect from Z to Y.

- In Model 3, the FP feature only affects the gene expression but not the pIC50.

- The causal structure in Model 4 corresponds to genes which only affect the pIC50. The FP feature does not effect neither the gene expression nor the pIC50.

- The causal structure in Model 5 corresponds to a complete mediation which means that the FP feature affects the gene expression, which in turn affects the pIC50. In the drug discovery setting, genes having this causal structure are of interest since they could provide an understanding about the mechanism of action of a given class of compounds having the specific FP feature.

- The causal structure in Model 6, the conditional independence model, corresponds to genes for which the FP feature affects both the expression level and the pIC50 but given the FP level, gene expression and pIC50 are not correlated. Hence, gene expression level and the pIC50 are mutually caused by the presence or absence of the FP feature (Morgan and Winship, 2007; Pearl, 2009).

- The causal structure in Model 7 corresponds to genes for which the expression level as well as the FP feature affect the pIC50. However, FP feature does

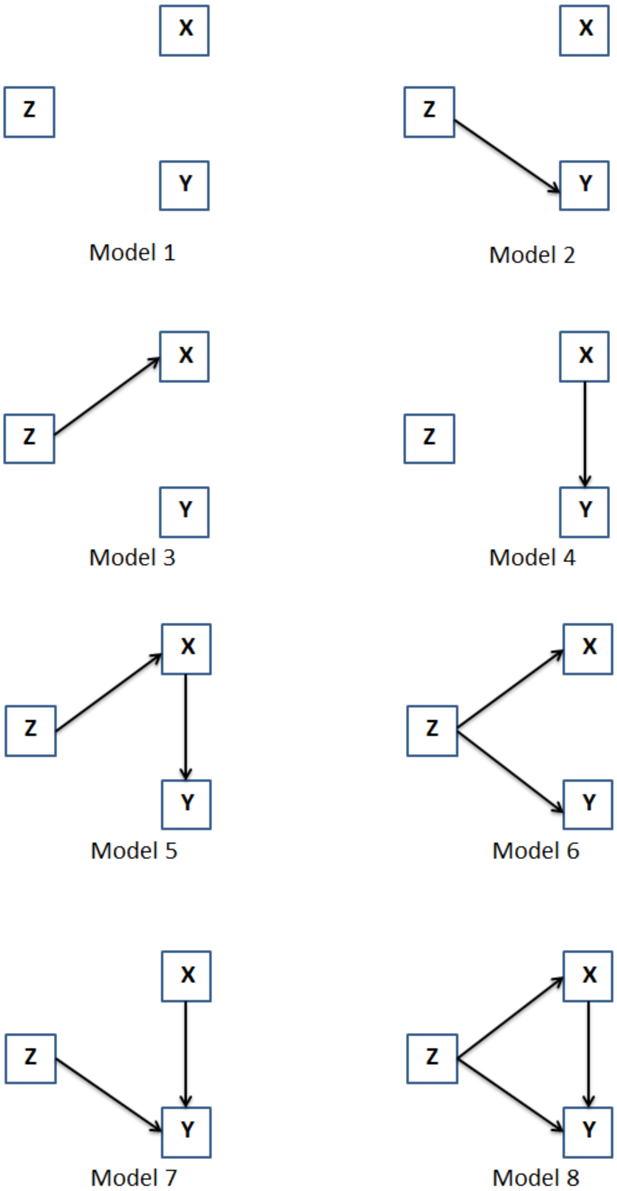


Figure 8.3: *Illustration of possible causal structure for the triplet (X_{ji}, Y_i, Z_i). Model 1: independent model, Model 2— Model 4: single effect models, Model 5: complete mediation model, Model 6: common cause model, Model 7: common effect on Y, Model 8: Partial mediation.*

not affect the gene expression. This structure corresponds to genes for which the gene expression and the FP feature would collide at the pIC50. According to Greenland *et al.* (1999), a collider is a variable that blocks the association between the variables that influence it.

- This causal structure presented in Model 8 corresponds to genes for which the expression level are affected by the presence or absence of the FP feature. The expression level in turn affects the pIC50. In addition, the FP feature affects the pIC50. This model is formulated in equation (8.1).

The eight causal structures represented by the 8 path models are fitted for each gene separately and can be used to select the best causal structure for each gene and to classify genes according to their causal structures.

8.2.3 Model Averaging Techniques

The modeling approach of Danner *et al.* (2015) requires to fit all possible structural models (8 in our setting) and to choose the structural model which best fits the data using model selection criteria such as Akaike's information criterion (AIC, Akaike, 1974) or the Bayesian information criterion (BIC, Schwarz, 1978). Following this approach, estimation and inference are based on the selected model (a procedure called post selection inference and estimation). Such a procedure does not take into account model uncertainty (i.e., the fact that more than one model was fitted to the data). In this chapter, we propose to use the model averaging technique for this purpose (Lin *et al.*, 2012; Kuiper *et al.*, 2014; Claeskens and Hjort, 2008; Kuiper *et al.*, 2011). All candidate models are fitted and their corresponding information criterion (IC) are computed. The IC, which takes into account both the goodness-of-fit and the model complexity, is used to calculate the posterior probability of the model. Let $M_1, \ldots, M_R$, be a set of R candidate models fitted to the data. The posterior model probability $P(M_r|data), r = 1, \ldots, R$, derived from the IC can be used to select causal structure (model) which better fits the observed data for a given gene. The posterior probability of the model M_r given the data (Burnham and Anderson, 2003) is given by

$$P(M_r|data) = \frac{P(data|M_r)P(M_r)}{\sum_{r=1}^{R} P(data|M_r)P(M_r)}, \quad r = 1, \ldots, R. \tag{8.3}$$

Here, $P(data|M_r)$ and $P(M_r)$ are the model likelihood and the prior probabilities of the rth model, respectively. Following Lin *et al.* (2012) non informative prior are used, i.e. $P(M_r) = 1/R$ for all models. The model likelihood $P(data|M_r)$ is approximated by (Burnham and Anderson, 2003)

$$P_{IC}(data|M_r) = \exp(-\frac{1}{2}\Delta IC_r), \tag{8.4}$$

where, $\Delta IC_r = IC_r - IC_{min}$, with $IC_{min} = min(IC_1, \ldots, IC_R)$. Hence, combining equations (8.3) and (8.4) together and assuming equal prior probabilities, we get

$$w_r = P_{IC}(M_r|data) = \frac{\exp(-\frac{1}{2}\Delta IC_r)}{\sum_{r=1}^{R}\exp(-\frac{1}{2}\Delta IC_r)}. \tag{8.5}$$

The weight, w_r, can be considered as an approximation of the posterior probability of a given model to be the best model among all fitted models given the data. AIC and BIC can be used in equation (8.5). The BIC penalty is higher than the AIC when there are more than seven observations (i.e., it favors simpler models as the sample size increases).

8.3 Application to the Data

Two case studies, EGFR and ROS, presented in Chapter 6, are used for illustration of the methodology presented above.

8.3.1 EGFR Data Overall Results

The model averaging technique described in the previous section was applied to the EGFR dataset assigning each gene to one of the causal structures based on the highest weight defined in equation (8.5).

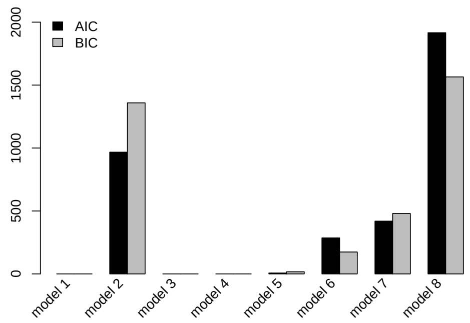


Figure 8.4: *Classification of genes to different casual structures based on AIC and BIC. Genes are classified based on the maximum posterior model probability.*

Table 8.1: *Classification of genes into different casual structures according to the maximum posterior probability.*

Structure	AIC	BIC
Model 1	0	0
Model 2	967	1359
Model 3	0	0
Model 4	0	0
Model 5	7	17
Model 6	286	174
Model 7	419	480
Model 8	1916	1565

It is evident from Figure 8.4 and Table 8.1 that, compared to the AIC, BIC assigns more genes to Model 2 and less genes to Model 8. This is because BIC favors simpler structures as pointed out in the previous section. Moreover, for this case study, some causal structures are less likely exhibited such as Models 1, 3 and 4. This is expected since the FP feature used has been previously shown to be linked to the pIC50 (Model 2 as the simplest). The AIC and BIC criteria identified most of the genes as having the causal structures represented by Model 2 (27% and 38% using AIC or BIC, respectively) and Model 8 (53% and 44% using AIC or BIC, respectively). Recall that genes having the causal structure represented by Model 2 are not of interest in the drug discovery process. The FP feature has an effect on the pIC50, but this effect is not mediated via the gene expression.

8.3.2 Example of Specific Genes

8.3.2.1 CAP1

The left panel in Figure 8.5 shows the posterior model probability of the 8 causal models for the gene CAP1. The highest posterior model probability is obtained for the causal structure in Model 2 ($w_2 = 0.73$). The expression level of this gene is plotted against the pIC50 as shown in the right panel of Figure 8.5. Note how the pIC50 level is different across the FP groups but the gene expression is not affected by the FP.

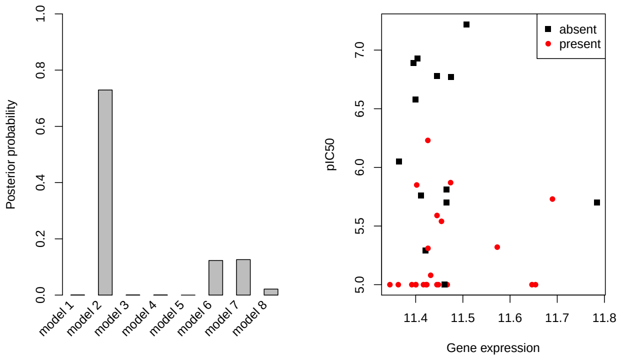


Figure 8.5: *Structural models for gene CAP1. Left panel: Posterior model probabilities; Right Panel: gene expression versus pIC50.*

8.3.2.2 FOSL1

Genes having the causal structure represented by Model 8 are of interest in the drug discovery process since the FP feature affects the gene expression level, which in turn affects the pIC50. In addition, the FP feature affects the pIC50.

Figure 8.6 shows the posterior model probabilities (left panel) and the scatter plot showing the relationship pattern of the three variables (right panel) for the gene FOSL1. For this gene, the highest posterior probability is obtained for Model 8 ($w_8 = 0.89$).

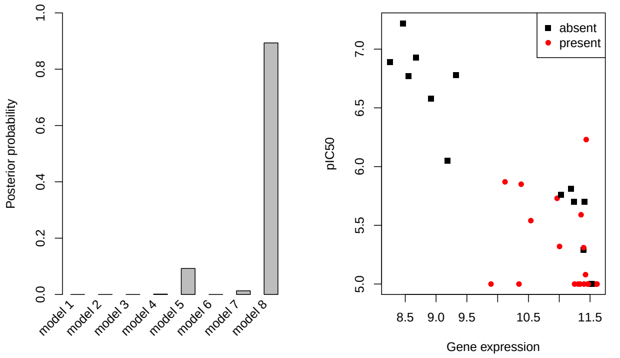


Figure 8.6: *Structural models for gene FOSL1. Left panel: Posterior model probabilities; Right Panel: gene expression versus pIC50.*

8.3.2.3 PNISR

In total, 17 genes were identified as having the causal structure represented by Model 5. For this structure, a complete mediation is observed. The FP feature affects the gene expression level, which in turn affects the pIC50. Figure 8.7 shows the posterior model probabilities (left panel) for the gene PNISR.

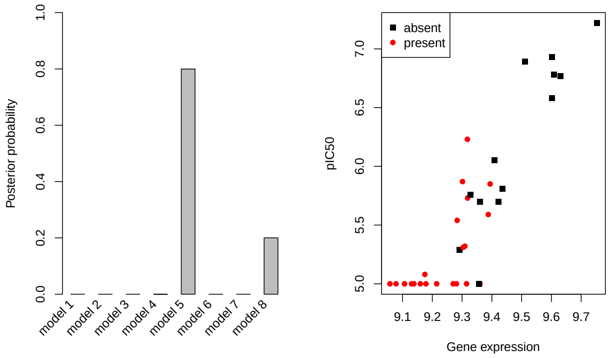


Figure 8.7: *Structural models for gene PNISR. Left panel: Posterior model probabilities; Right Panel: gene expression versus pIC50.*

In the drug discovery process, this type of genes is of interest since total effect of the FP feature on the bioactivity is absorbed by the indirect effect via the gene expression.

8.3.2.4 YY1

The causal structure represented by Model 6 was observed for 8% of the genes using AIC (5% of the genes using BIC). It is an interesting causal structure with respect to drug discovery since the FP feature has an effect on both gene expression and pIC50, but the gene expression level does not affect the pIC50. Figure 8.8 shows one example, the gene YY1, that follows this causal structure.

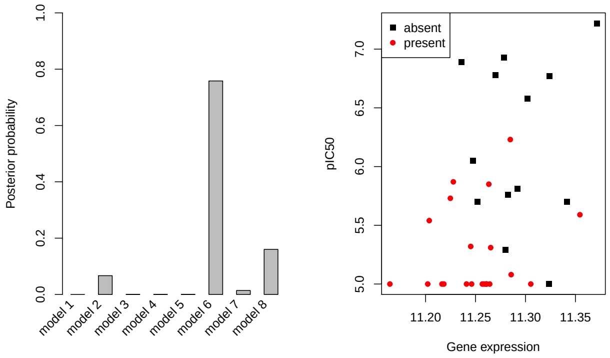


Figure 8.8: *Structural models for gene YY1. Left panel: Posterior model probabilities; Right Panel: gene expression versus pIC50.*

Figure 8.9 summarizes the causal structures for the above-mentioned genes with their corresponding parameter estimates.

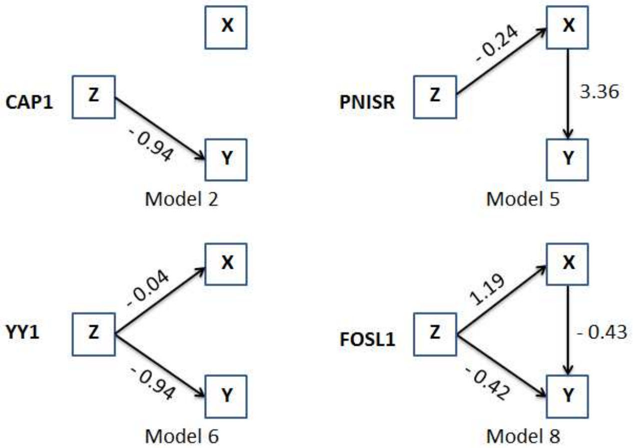


Figure 8.9: *The EGFR case study, SEM parameter estimates of the four genes exhibiting different causal structures.*

8.3.3 Application to ROS Data

The 8 SEMs were fitted to the ROS data and the classification to the different causal structures is presented in Table 8.2 and Figure 8.10.

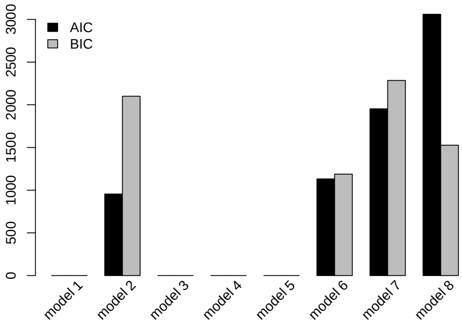


Figure 8.10: *Classification of gene to different casual structures based on AIC and BIC. Genes are classified based on the maximum posterior model probability.*

Table 8.2: *Classification of genes into different casual structures according to the maximum posterior model probability.*

Structure	AIC	BIC
Model 1	0	0
Model 2	955	2100
Model 3	0	0
Model 4	0	0
Model 5	0	0
Model 6	1132	1188
Model 7	1953	2285
Model 8	3060	1527

For gene LRRTM2, the highest model probability (w_2=0.82) was obtained for Model

2 which implies that for this gene the most probable causal structure assumes only an effect of the chemical structure on pIC50 as it can be clearly seen in Figure 8.11.

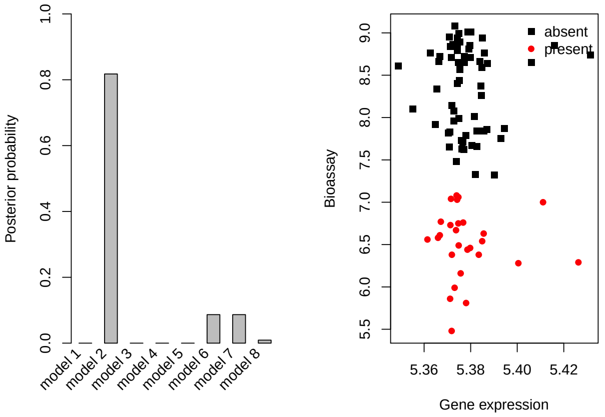


Figure 8.11: *Structural models for gene LRRTM2. Left panel: Posterior model probabilities; Right Panel: gene expression versus pIC50.*

As mentioned in the previous section, the conditional independence model (Model 6), which has the highest model probability for gene PFKFB3 (w_6=0.95, see Figure 8.12) is of primary interest in drug discovery. This causal structure implies that the association between gene expression and pIC50 is derived by the chemical structure (and conditional on the chemical structure the two variables are independent).

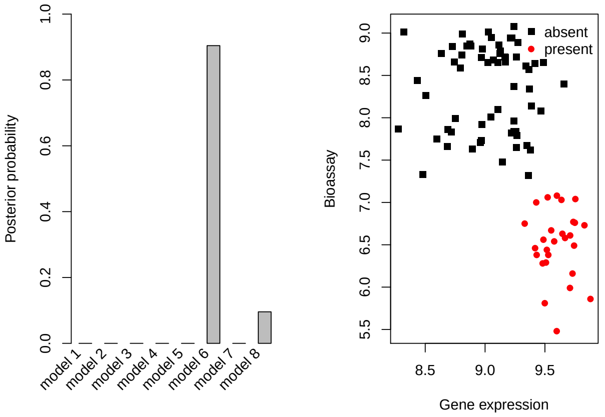


Figure 8.12: *Structural models for gene PFKFB3. Left panel: Posterior model probabilities; Right Panel: gene expression versus pIC50.*

For gene EXPH5, shown in Figure 8.13, the highest model probability was observed for model 7 (w_7=0.90) for which the causal structure includes only direct effects.

Finally, model 8, with w_8=0.97, is the most probable model for gene MPPL17 (see Figure 8.14) implying that in addition to the chemical structure effect, the expression level and pIC50 are correlated.

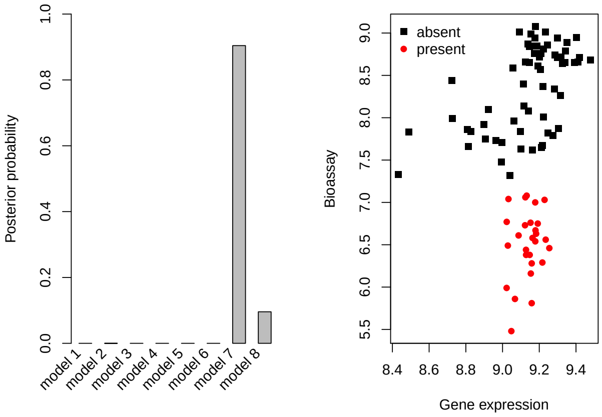


Figure 8.13: *Structural models for gene EXPH5. Left panel: Posterior model probabilities; Right Panel: gene expression versus pIC50.*

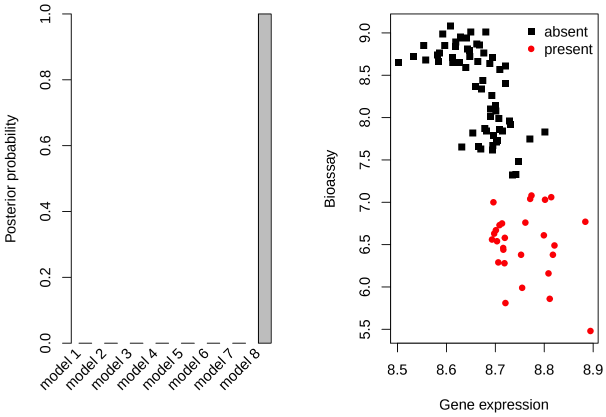


Figure 8.14: *Structural models for gene MRPL17. Left panel: Posterior model probabilities; Right Panel: gene expression versus pIC50.*

Figure 8.15 displays parameter estimates for the above genes and their corresponding causal structure.

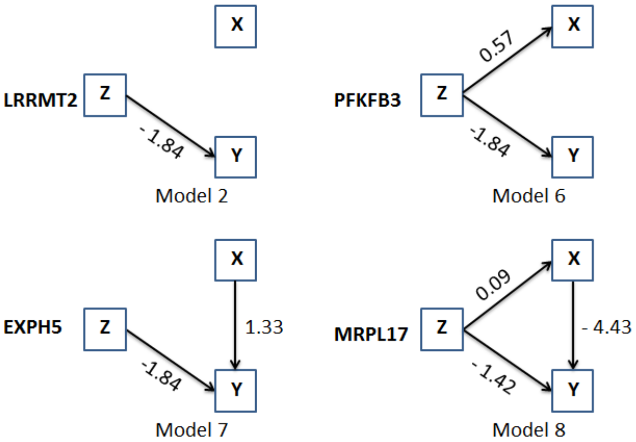


Figure 8.15: *The ROS project, SEM parameter estimates of the four genes exhibiting different causal structures.*

8.4 Discussion

In this chapter, we discussed how structural equation models can be used for the integration of multi-source high-dimensional data in the drug discovery process. SEM allows to classify genes according to their causal structures with the bioactivity and chemical structure data in the early phase of drug discovery.

The selection of causal structure per gene is done via model comparison. To this end, for each gene, different causal structures are assumed, and the structure giving the highest posterior model probability is retained. Recall that some of the causal structures were not considered due to the experiment constraints. Genes are then grouped depending on the causal structure that gives the best fit.

The genes FOSL1, PNISR, YY1 (EGFR data) are of interest in the drug discovery studies since they can be used to establish the relationship between chemical structure and bioactivity. Selection and evaluation of genes which can be used as potential biomarkers in drug discovery can help the development team to better understand the mechanism of action of a new set of compounds and therefore substantially shorten development time or the time to reach a critical decision point, such as candidate selection, in drug development.

The analysis presented in this chapter was done gene by gene, and genes identified by this approach could be combined in further analyses to construct joint biomarkers using different methods. Methods such as supervised principal component analysis (SPCA), lasso and elastic net can also be used if the interest is to identify a set of multiple genes that could be used to predict the response.

Part III

Evaluation of Surrogate Endpoints in Clinical Trials: Software Development

Chapter 9

Surrogacy in Randomized Clinical Trials: An Introduction and Case Studies

In the first and the second part of this dissertation, we discussed different methods that can be used for the identification of metabolomics and genomic biomarkes. In this part of the dissertation we focus on the evaluation of surrogate endpoints in the clinical trial setting.

Clinical trials are designed to evaluate therapeutic efficacy of new drugs using clinical endpoints that reflect concrete benefits for patients. Such endpoints include disease outcome, survival time, death, etc. However, in many cases these trials require large number of patients and long time to complete. Surrogate endpoints, which would shorten the duration of assessment and allow to evaluate the effectiveness of new drugs, are of interest for both researchers and patients (Burzykowski *et al.*, 2005). For example, considerable reduction in sample size and trial duration can be achieved by replacing rare or late-occurring clinical endpoints with frequent or short-term surrogate endpoints (Lin *et al.*, 1997).

As discussed by Ellenberg and Hamilton (1989), in some cases the investigator must rely on a surrogate endpoint since the clinical endpoint is not available, is difficult to measure, requires expensive, invasive or uncomfortable procedure. Ellenberg and

Hamilton (1989) define a surrogate endpoints as: ”an endpoint that can be used in lieu of other endpoints in the evaluation of experimental treatments or other interventions”. This endpoint is useful when it can be measured earlier, more conveniently or more frequently than the endpoints of interest (clinical endpoint). It should also be clinically relevant and biologically plausible (De Gruttola *et al.*, 1997; Lonn, 2001; Weir and Walley, 2006). Before a surrogate endpoint can replace a final endpoint in the evaluation of an experimental treatment, it must be formally ’validated’, a process that has caused much controversy and has not been fully elucidated so far (Burzykowski *et al.*, 2005).

As pointed out by Buyse *et al.* (2000), the main reason to validate a surrogate endpoint is to be able to predict the effect of treatment on the true endpoint, based on the observed effect of the treatment on the surrogate endpoint, with sufficient precision to distinguish safely between effects that are clinically worthwhile and effects that are not.

In this part of the dissertation, we focus on software development and introduce two softwares products, in SAS and R, for the analysis of surrogate endpoint in randomized clinical trials. The development of both products was done keeping in mind that the users will not be necessary statisticians. Therefore, both products provide a user-friendly and easy to interpret standard output which contains only the main results of the analysis. The SAS product requires a SAS license, basic knowledge about the SAS software but no knowledge in statistics. The R shiny App, an online application, does not require from the user anything except a dataset ready for the analysis. The user can run the analysis from his/her smartphone (or any other mobile device) and does not need to install or know any statistical software.

9.1 Case Studies

Several datasets are used for illustration in this part of the dissertation. A short description of each case study is given below.

9.1.1 ARMD Study

The ARMD trial is a clinical trial for patients with age-related macular degeneration (ARMD), a condition in which patients progressively lose vision (Pharmacological Therapy for Macular Degeneration Study Group, 1997). In the ARMD trial, visual acuity was examined using standardized vision charts that display lines with five letters of decreasing sizes. The patients had to read these letters from top (largest letters) to bottom (smallest letters). Visual acuity was quantified as the total number of letters that were correctly read by a patient.

A total of 181 patients from $N = 36$ centers participated in the ARMD trial. There were two treatment conditions: interferon-α and placebo (coded as 1 = interferon-α and -1 = placebo). The true endpoint (T) is the change in visual acuity 52 weeks after the start of the treatment. The candidate surrogate endpoint (S) is the change in visual acuity 24 weeks after starting the treatment. A total of 84 and 97 patients were enrolled in the placebo and interferon-α treatment conditions, respectively. The aim of the ARMD trial was to show that interferon-α is superior to the placebo treatment (using visual acuity as the primary endpoint).

9.1.2 Ovarian Cancer Study

The method was applied to a data from a meta-analysis for four multicenter trials in advanced ovarian cancer (Omura *et al.*, 1991). The aim of this study was to compare two treatments: cyclophosphamide plus cisplatin (CP) versus cyclophosphamide plus adriamycin plus cisplatin (CAP). The binary indicator for treatment is equal to 0 for CP and 1 for CAP. The surrogate endpoint was the progression free survival time, defined as the time (in years) from randomization to clinical progression of the disease or death, whereas the true endpoint was the survival time, defined as the time (in years) from randomization to death from any cause.

Technically, the ovarian cancer study is a meta-analysis but it contains only four trials. Thus the center was used as the unit of analysis for larger two trials, and the trial as the unit of analysis for the two small trials. A total of 50 ”units” are then available for the analysis, with a number of individual patients per unit ranging from 2 to 274. The analysis was restricted to centers with at least three patients on each treatment arm due to estimability constraints (Burzykowski *et al.*, 2001). As a result, data for 39 centers were used with a total sample size of 1153 patients (569 in the treatment arm and 584 in the control arm).

Figure 9.1 displays the Kaplan Meier curves of the survival time (true endpoint) and progression free survival time (surrogate endpoint).

9.1.3 Colorectal Cancer Study

The colorectal data was from 28 advanced colorectal cancer trials (Advanced Colorectal Cancer MetaAnalysis Project, 1992, 1994; Meta-Analysis Group in Cancer, 1996, 1998). The individual patients data were collected by the Meta-Analysis Group in Cancer between 1990 and 1996 to obtain an overall quantitative assessment of the

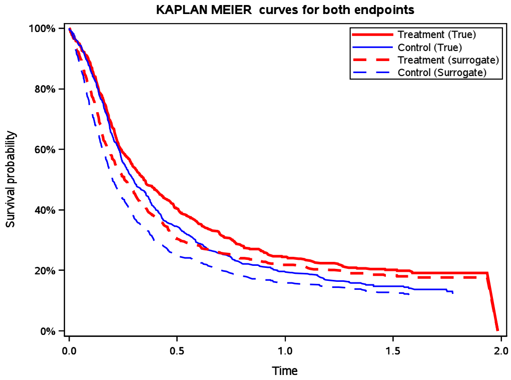


Figure 9.1: *Advanced ovarian data. Kaplan Meier curves for the survival time (true endpoint) and progression free survival (surrogate endpoint) for the two treatment groups CP and CAP.*

value of several experimental treatments in advanced colorectal cancer. In the four meta-analyses, the comparison was between an experimental treatment and a control treatment. The control treatments, referred to hereafter as ’FU bolus’, were similar across the four meta-analyses and consisted of fluoropyrimidines (5FU or FUDR) given as a bolus intravenous injection. The experimental treatments, referred to hereunder as ’experimental FU’, differed across the four meta-analyses and consisted of 5FU modulated by leucovorin (Advanced Colorectal Cancer Meta-Analysis Project, 1992), of 5FU modulated by methotrexate (Advanced Colorectal Cancer Meta-Analysis Project, 1994), of 5FU given in continuous infusion (Meta-Analysis Group in Cancer, 1998) and of hepatic arterial infusion of FUDR for patients with metastases confined to the liver (Meta-Analysis Group in Cancer, 1996). As noted by Daniels and Hughes (1997), the use of an ’experimental’ treatment that varies among the trials can be defended on the grounds of generalizability of the results of the validation process to future clinical trials and treatments. The experimental treatments in our example might be considered as representatives of ’the modifications of the standard fluoropyrimidine-based regimen’ in advanced colorectal cancer. Several of the 28 trials were multiarmed. In total, 33 randomized comparisons were considered in the four meta-analyses. Individual patient data were available for 27 of the comparisons (in 24 studies). Each of the comparisons is considered as a separate ’trial’.

9.1.4 Schizophrenia Study

This dataset combines the data that were collected in five double-blind randomized clinical trials. In these trials, the objective was to examine the efficacy of risperidone to treat schizophrenia. Schizophrenia is a mental disease that is hallmarked by hallucinations and delusions (Association, 2000).

In each trial, the Clinical Global Impression (CGI; Guy (1976)), the Brief Psychiatric Rating Scale (BPRS; Overall and Gorham (1962)), and the Positive and Negative Syndrome Scale (PANSS; Singh and Kay (1975)) were administered. These instruments are clinical rating scales that are routinely used to assess symptom severity in patients with schizophrenia (Mortimer (2007)). The patients in the different trials were administered risperidone or an active control (e.g., haloperidol, levomepromazine, or perphenazine) for four to eight weeks. The main endpoints of interest were the change in the CGI score (= CGI score at the end of the treatment - CGI score at the start of the treatment), the change in the PANSS score, and the change in the BPRS score.

A total of $2,128$ patients participated in the five trials ($1,591$ patients received risperidone and 537 patients were given an active control). The patients were treated by a total of $N = 198$ psychiatrists. Each of the psychiatrists treated between $n_i = 1$ and 52 patients. In the subsequent sections, different combinations of the endpoints (CGI and PANSS) in various forms (binary, continuous) was considered as the candidate surrogate and the true endpoint.

9.1.5 Prostate Cancer Study: A Meta-analysis of Two Trials

This dataset comprises two trials that compared oral liarozole, an experimental retinoic acid metabolism-blocking agent ($Z = 1$), with an antiandrogenic drug considered as control ($Z = 0$): cyproterone acetate in the first trial and flutamide in the second (Buyse *et al.*, 2003). In both trials, patients were in relapse after first-line endocrine therapy. The trials accrued 312 and 284 patients, respectively.

Each trial was multinational and multicentric, and the unit of analysis for the surrogacy analysis was chosen to be the country in which the patients were treated. There were 19 countries containing between 4 to 69 patients. The primary endpoint of the trials was overall survival from the start of treatment. In both trials, patients were assessed at baseline (before the start of treatment), at 2 weeks, monthly for 6 months, at 3-month intervals until the second year, and at 6-month intervals until treatment discontinuation or death. The assessments included measurement of the prostate-specific antigen (PSA) level. PSA is a glycoprotein that is found almost exclusively in normal and neoplastic prostate cells. Changes in PSA often antedate changes in bone scan, and they have been used as an indicator of response in patients with androgen-independent prostate cancer. For the surrogacy analyses, overall survival (OS) was considered the true endpoint (T). In the analysis the logarithm of PSA, measured at about 28 days, was used as a surrogate for overall survival (OS). The data were analyzed before by Buyse et al. (2003), but without considering the value of PSA at a particular time point as a candidate surrogate for OS. For the analysis purposes, patients were grouped by trial and by country. Treatment with flutamide or cyproterone acetate was considered as the experimental treatment, while liarozole was regarded as the control treatment.

Among the 596 patients included in the dataset, 421 had a PSA measurement obtained at about 28 days (±6 days). There are 19 trial-by-country groups containing between 2 to 55 patients per group. Two groups (one with two and one with seven patients) have to be eliminated from the analysis, because at least one of the treatment arms within the group does not contain any deaths. Consequently, the analysis included 412 patients spread across 17 groups. The data were provided by the Janssen Research Foundation (see Buyse *et al.*, 2003).

Chapter 10

Surrogacy Validation Using SAS

10.1 Introduction

The ability to conduct any statistical analysis on large scale and by many users depends on the availability of a software product with the capacity to conduct the analysis of interest. In this chapter, we present a SAS product for the analysis of surrogate endpoint in randomized clinical trials. Table 10.1 presents different surrogacy settings and corresponding models implemented in the SAS macros. The usage of each SAS macro is illustrated using a case study for a specific setting. The SAS macros presented in this chapter do not require from the user to formulate a model in SAS but rather to specify a surrogacy model and the macro formulate the model in SAS automatically.

10.2 General Structure of the SAS Macros Available for the Analysis of Surrogate Endpoints

The SAS macros, developed for the analysis presented in this chapter, have the same general form. Depending on the surrogacy setting, the macros require from the user to prepare a dataset with a specific structure. Note that, different macros require different dataset structures. A generic call of a surrogacy SAS marco has the following form:

Table 10.1: *SAS macros available for the evaluation of surrogate endpoints in randomized clinical trials. Statistics mentioned in the table, such as R^2_{indiv} etc, are discussed in different sections in the chapter.*

Surrogacy Setting	Macro name	Model	Surrogacy measures	Section in the chapter	Case study
Normal/Normal	`CONTCONTFULL`	Full fixed	R^2_{trial} and R^2_{indiv}	10.3.1	ARMD data
Normal/Normal	`CONTCONTRED`	Reduced fixed	R^2_{trial} and R^2_{indiv}	10.3.2	True endpoint= Diff52
Normal/Normal	`CONTRANFULL`	Full random	R^2_{trial} and R^2_{indiv}	10.3.3	Surrogate endpoint=Diff24
Normal/Normal	`CONTRANRED`	Reduced random	R^2_{trial} and R^2_{indiv}	10.3.4	
Survival/Survival	`TWOSTAGEKM`	Two-stage	R^2_{indiv}	10.4.1	Ovarian data
Survival/Survival	`TWOSTAGECOX`	Two-stage	R^2_{trial}	10.4.2	True endpoint= OS
Survival/Survival	`COPULA`	joint model	R^2_{trial} and R^2_{indiv}	10.4.3	Surrogate endpoint=PFS
Survival/Categorical (Binary)	`SURVCAT`	Joint model	R^2_{trial} and Global odds	10.9.3	Colorectal data True endpoint=OS Surrogate endpoint=Remission
Normal/Binary	`NORMALBIN`	Joint model Normal-binary	R^2_{trial} and R^2_{indiv}	10.7	Schizo data True endpoint=PANSS Surrogate endpoint=CGI
Binary/Binary	`BINBIN`	Bivariate probit	R^2_{trial} and R^2_{indiv}	10.8	Schizo data True endpoint=CGI Surrogate endpoint=PANSS
Survival/Binary	`SURVBINIT`	Information theory	R^2_{ht} and R^2_{h}	10.9	Colorectal data True endpoint=OS Surrogate endpoint=Remission
Normal/Binary	`NORMBINIT`	Information theory	R^2_{ht} and R^2_{h}	10.9	Schizo data True endpoint=PANSS Surrogate endpoint=CGI
Binary/Binary	`BINBINIT`	Information theory	R^2_{ht} and R^2_{h}	10.9	Schizo data True endpoint=CGI Surrogate endpoint=PANSS
Normal/Normal	`NORMNORMIT`	Information theory	R^2_{ht} and R^2_{h}	10.9	ARMD data True endpoint=Diff52 Surrogate endpoint=Diff24

```
% macroname (data=,true=,surrog=,trial=,treatment=,patid=,
             optional-arguments)
```

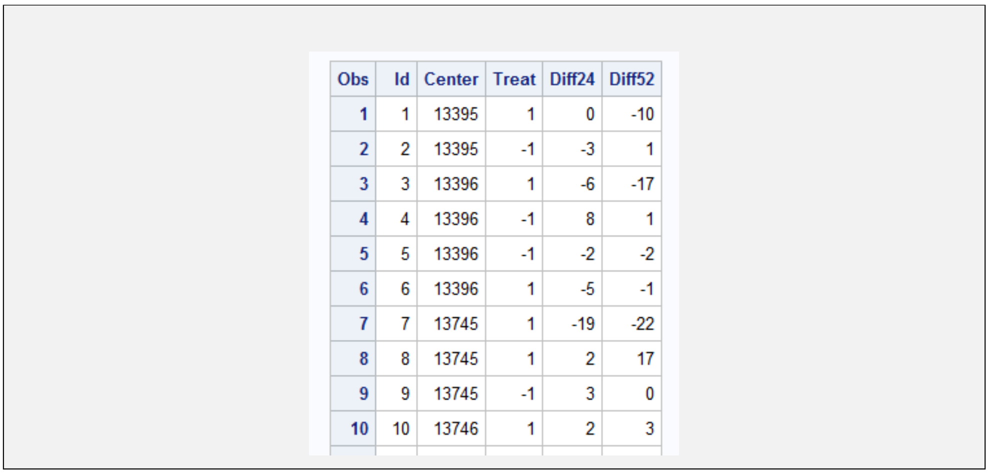
The following arguments are used in all macros:

- `data:` a dataset containing one record per patient with measurements for both the true and the surrogate endpoints.
- `true:` a measurement of the true endpoint.
- `surrog:` a measurement of the surrogate endpoint.
- `treatment:` treatment indicator variable (1= active, -1=control).
- `trial:` the unit of the study for which trial-level surrogacy will be estimated.
- `patid:` patient's identification number.
- `optional-arguments:` optional arguments that should be provided to conduct a specific analysis, for example, a leave-one-out trial analysis discussed in Section 10.3.1.

Graphical and numerical outputs are customized in such a way that users only see relevant information depending on the surrogacy setting. Note that the standard SAS output tables are not produced and it is recommended to check the SAS log window for possible problems.
For the remainder of the chapter, we present different surrogacy settings, their corresponding macros, case studies and outputs. For all surrogacy settings the models are only briefly presented.

10.3 Validation of Surrogacy Using a Joint Modeling Approach for Two Normally Distributed Endpoints

In this section we presented SAS macros for surrogacy setting in which both endpoints are continuous. For all models discussed in this section the ARMD data were used for illustration. Visual acuity at week 52 (`Diff52`) and visual acuity at week 24 (`Diff24`) are the true and the surrogate endpoints, respectively. Each line in the data contains information about one patient. A partial print of the data is shown below.



Obs	Id	Center	Treat	Diff24	Diff52
1	1	13395	1	0	-10
2	2	13395	-1	-3	1
3	3	13396	1	-6	-17
4	4	13396	-1	8	1
5	5	13396	-1	-2	-2
6	6	13396	1	-5	-1
7	7	13745	1	-19	-22
8	8	13745	1	2	17
9	9	13745	-1	3	0
10	10	13746	1	2	3

10.3.1 The Full Fixed-effects Model

Model Formulation

This approach is based on a hierarchical two-stage model used by Buyse *et al.* (2000) to validate a surrogate endpoint, when both endpoints are assumed to be normally distributed. Briefly, the first-stage consists of a joint model for the surrogate and true endpoints given by:

$$\begin{cases} S_{ij} = \mu_{Si} + \alpha_i Z_{ij} + \varepsilon_{Sij}, \\ T_{ij} = \mu_{Ti} + \beta_i Z_{ij} + \varepsilon_{Tij}, \end{cases} \tag{10.1}$$

where, μ_{Si} and μ_{Ti} are trial-specific intercepts for S and T, α_i and β_i are trial-specific treatment effects upon the surrogate and the true endpoint, respectively. The error terms, ε_{Sij} and ε_{Tij}, are bivariate normally distributed with zero mean and covariance matrix given by:

$$\Sigma = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ & \sigma_{TT} \end{pmatrix}. \tag{10.2}$$

The individual-level surrogacy is assessed by the squared correlation between S and T after adjusting for trial-specific treatment effects, that is

$$R^2_{indiv} = \frac{\sigma^2_{ST}}{\sigma_{SS}\sigma_{TT}}. \tag{10.3}$$

For the full fixed-effects model, the trial-level surrogacy is estimated using the coefficient of determination obtained by fitting the following model:

$$\widehat{\beta}_i = \lambda_0 + \lambda_1 \widehat{\mu}_{Si} + \lambda_2 \widehat{\alpha}_i + \varepsilon_i, \qquad (10.4)$$

where, $\widehat{\beta}_i$, $\widehat{\mu}_{Si}$, and $\widehat{\alpha}_i$ are the parameter estimates obtained from the joint model specified in equation (10.1).
This approach ignores the fact that the estimated treatment effects $\hat{\alpha}_i$ and $\hat{\beta}_i$ will typically come from trials with large variations in sizes. One way to address this issue is by weighing the contributions according to trial size, resulting in a weighted linear regression (Tibaldi *et al.*, 2003). Such an approach may account for some but not all of the heterogeneity in information content between trial-specific contributions. We adopt this correction in all two-stage approaches presented in this chapter. Standard errors for R^2_{indiv} and R^2_{trial} can be calculated using the delta method.

Sensitivity Analysis: Leave-one-out Evaluation

The surrogacy measures estimated in the previous section are derived using all units in the study (i.e, all trials, centers, etc.). To check the stability of the estimated measures, we propose to use a ”leave-one-out” evaluation procedure as a sensitivity analysis approach. It is an iterative procedure in which at each iteration one trial is left out and the surrogacy measures are estimated using the remaining trials in the data. The procedure is shown in Figure 10.1.

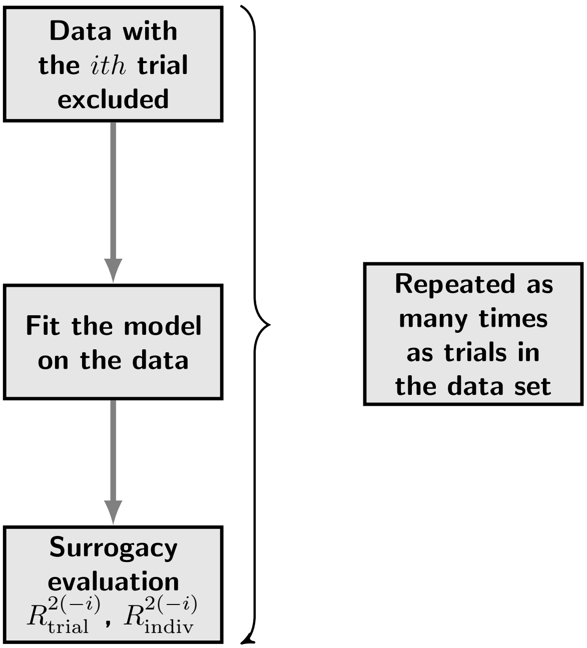

Data with the ith trial excluded
Fit the model on the data
Surrogacy evaluation $R^{2(-i)}_{\text{trial}}$, $R^{2(-i)}_{\text{indiv}}$
Repeated as many times as trials in the data set

Figure 10.1: *A leave-one-out evaluation procedure.*

At the end of the run, parameter estimates for both R^2_{trial} and R^2_{indiv} are obtained for each trial and influential trials can be identified.

The SAS Macro %CONTCONTFULL

The SAS macro %CONTCONTFULL can be used to conduct the analysis for a setting with two normally distributed endpoints using the full fixed-effects model. The following macro call is used:

```
%CONTCONTFULL(data=armd,true=diff52,surrog=diff24,trt=treat,
              trial=center,patid=patientId,weighted=1,
              looa=1)
```

The optional arguments that we use are:

- `weighted`: an option which allows to use weighted regression (weighted=1) in the computation of the trial-level surrogacy. The number of patients in the trial is used as the weight.
- `looa`: an option to perform a leave-one-out trial analysis (looa=1). Both surrogacy measures are computed by leaving out each trial so as to check the influence of the trial on the surrogacy measures.

Data Analysis and Output

The macro %CONTCONTFULL produces default numerical and graphical outputs. The first part of the output, shown in Figure 10.2, consists of two descriptive plots which show the distribution of the patients in the trials by treatment arms (left panel) and a scatter plot between the true and the surrogate endpoints (right panel).

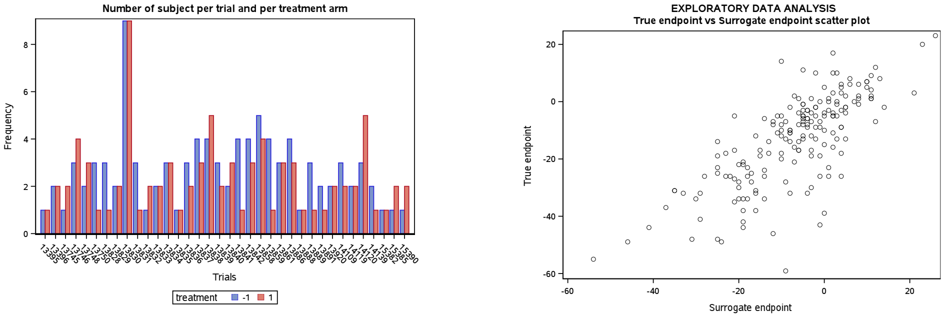
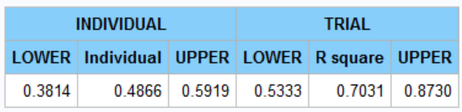


__Figure 10.2:__ ARMD study. Descriptive plots. Left Panel: Patients distribution in the trials by treatment arm. Right Panel: Scatter plot between the true and the surrogate endpoints.

As shown in the panel below, individual- and trial-level surrogacy are equal to $\hat{R}^2_{indiv} = 0.4866$ (0.3814, 0.5919) and $\hat{R}^2_{trial} = 0.7031$ (0.5333, 0.8730), respectively. Both surrogacy measures indicate that the change in visual acuity after 24 weeks is a surrogate of moderate value for the visual acuity at 52 weeks after starting the interferon-α treatment.



INDIVIDUAL			TRIAL		
LOWER	Individual	UPPER	LOWER	R square	UPPER
0.3814	0.4866	0.5919	0.5333	0.7031	0.8730

The results for the leave-one-out sensitivity analysis are presented in Figure 10.3. Note how the value of R^2_{trial} decreased when center ”13830” was left out ($\hat{R}^{2(-13830)}_{trial} = 0.6295$). We can see in Figure 10.2 (left panel) that this trial is the one with highest number of patients in both treatment arms. Leaving out this center has an impact on the estimated trial-level surrogacy, since it has the highest weight.

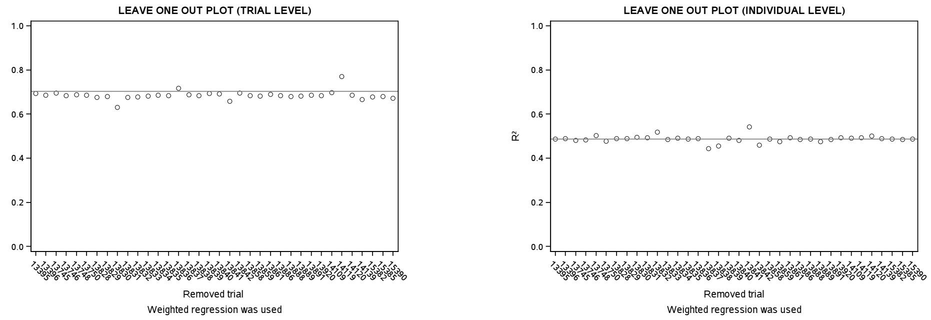
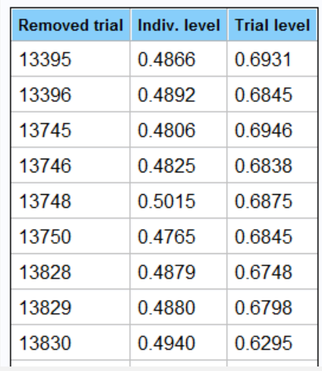


Figure 10.3: *ARMD study. Sensitivity analysis. Left Panel: trial-level surrogacy. Right Panel: individual-level surrogacy.*

A table containing the surrogacy measures estimated using the leave-one-out sensitivity analysis is presented below.



Removed trial	Indiv. level	Trial level
13395	0.4866	0.6931
13396	0.4892	0.6845
13745	0.4806	0.6946
13746	0.4825	0.6838
13748	0.5015	0.6875
13750	0.4765	0.6845
13828	0.4879	0.6748
13829	0.4880	0.6798
13830	0.4940	0.6295

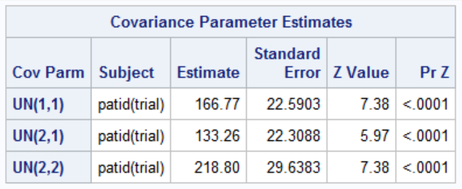
SAS Codes for the First-stage

Although, not visible to the users, the analysis implemented in the %CONTCONTFULL macro is based on procedure MIXED. In this section we discuss in more details the implementation of the joint model specified in equation (10.1) in SAS. Using procedure MIXED, the following code can be used to fit the joint model in equation (10.1).

```
proc mixed data=dataset covtest;
class endp patid trial;
model outcome = endp*trial endp*treat*trial / solution noint ;
repeated endp / type=un subject=patid(trial) ;
ods output solutionF=eb CovParms=covar ;
RUN;
```

The above code presumes that there are two records per subject in the input data set, the first one corresponding to the surrogate endpoint and the second one to the true endpoint. The variable `endp` (endpoint) is an indicator variable for the endpoint (coded -1 for surrogate and 1 for true endpoint), the variable `outcome` contains measurements obtained from each endpoint and the variable `treat` is assumed to be -1/1 coded (Burzykowski *et al.*, 2005).

The `repeated` statement is used for the estimation of the residual covariance matrix Σ in equation (10.2). For the mean structure, the interaction term `endp*trial` in the model statement allows to fit trial-specific intercepts on both endpoints while the three-way interaction `endp*treat*trial` produces the trial-specific treatment effects on the both endpoints. The options `solutionF` and `CovParms` in the `ods output` statement allow to output datasets containing fixed-effects estimates and the errors covariance matrix, respectively, for further analysis. The estimated covariance matrix is shown below.



Covariance Parameter Estimates					
Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
UN(1,1)	patid(trial)	166.77	22.5903	7.38	<.0001
UN(2,1)	patid(trial)	133.26	22.3088	5.97	<.0001
UN(2,2)	patid(trial)	218.80	29.6383	7.38	<.0001

Individual-level surrogacy measure is estimated by:

$$\hat{R}^2_{indiv} = \frac{133.26^2}{166.77 \times 218.80} = 0.4866. \tag{10.5}$$

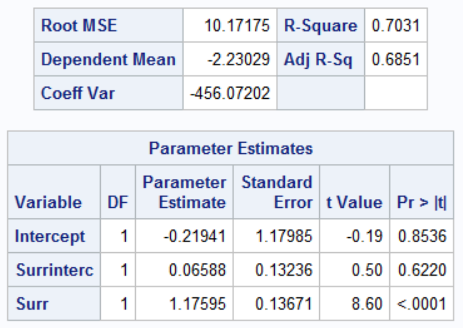
Note that the nesting notation in the `subject=patid(trial)` option is necessary for SAS to recognize the nested structure of the data (subjects are clustered within trials). The hierarchical nature of the data enables SAS to build a block-diagonal covariance matrix, with diagonal blocks corresponding to different trials, which speeds up computations considerably.

SAS Codes for the Second-stage

The second-stage model, from which trial-level surrogacy is estimated, is based on procedure `REG`.

```
proc reg data=secondstage;
model true=surrinterc surrogate;
weight n;
ods output FitStatistics=rsq;
run;
```

Here, `true` is the parameter estimate for the treatment effect on the true endpoint ($\hat{\beta}_i$), `surrinterc`, `surrogate` are the parameter estimates for trial-specific intercept ($\hat{\mu}_{S_i}$) and treatment effects ($\hat{\alpha}_i$) on the surrogate endpoint, respectively. The statement `weight` is used to account for the variability in trials sizes as discussed in Section 10.3.1. SAS output from the second-stage model is shown in the panel below.



Root MSE	10.17175	R-Square	0.7031
Dependent Mean	-2.23029	Adj R-Sq	0.6851
Coeff Var	-456.07202		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.21941	1.17985	-0.19	0.8536
Surrinterc	1	0.06588	0.13236	0.50	0.6220
Surr	1	1.17595	0.13671	8.60	<.0001

The estimated regression line is $\hat{\beta}_i = -0.22 + 0.07\hat{\mu}_{S_i} + 1.18\hat{\alpha}_i$ with the trial-level surrogacy $\hat{R}^2_{trial(f)} = 0.7031$.

10.3.2 The Reduced Fixed-effects Model

Model Formulation

The reduced fixed-effects model assumes common intercepts for S and T in equation (10.1). Hence, trial-specific μ_{Si} and μ_{Ti} are replaced by μ_S and μ_T respectively. The full fixed-effects model in equation (10.1) can be rewritten as:

$$\begin{cases} S_{ij} = \mu_S + \alpha_i Z_{ij} + \varepsilon_{Sij}, \\ T_{ij} = \mu_T + \beta_i Z_{ij} + \varepsilon_{Tij}. \end{cases} \tag{10.6}$$

The term $\hat{\mu}_{Si}$ is dropped at the second-stage. This implies that trial-level surrogacy is assessed using the coefficient of determination obtained from the model:

$$\widehat{\beta}_i = \lambda_0 + \lambda_1 \widehat{\alpha}_i + \varepsilon_i. \tag{10.7}$$

Individual-level surrogacy can be assessed using the adjusted association in equation (10.3).

The SAS Macro %CONTCONTRED

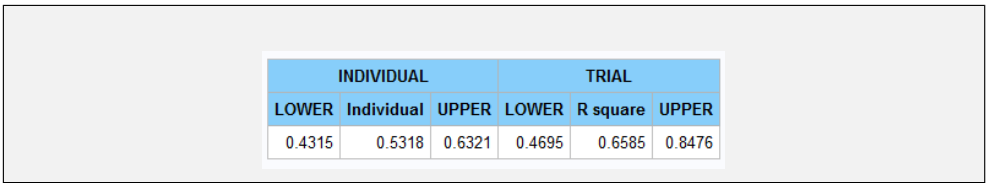
The SAS macro %CONTCONTRED can be used to fit the reduced joint model specified in equation (10.6). For the ARMD data we use:

```
%CONTCONTRED(data=armd,true=diff52,surrog=diff24,trt=treat,
             trial=center,patid=patientId,weighted=1,
             looa=1)
```

The specification of the macro's arguments is the same as the specification presented in Section (10.2).

Data Analysis and Output

Surrogacy measures obtained from the reduced fixed-effects model are shown below. Similar to the results presented in Section 10.3.1, the surrogacy measures $R^2_{indiv} = 0.5318$ (0.4315, 0.6321) and $R^2_{trial(r)} = 0.6585$ (0.4695, 0.8476) indicated that visual acuity 24 weeks after starting the interferon-α treatment is a surrogate of moderate value for the visual acuity at 52 weeks after starting the interferon-α treatment.



INDIVIDUAL			TRIAL		
LOWER	Individual	UPPER	LOWER	R square	UPPER
0.4315	0.5318	0.6321	0.4695	0.6585	0.8476

Trial-specific parameter estimates for treatment effects are shown in Figure 10.4. The regression line fitted at the second-stage is added. The circle sizes in the plot are proportional to the number of patients from each trial.

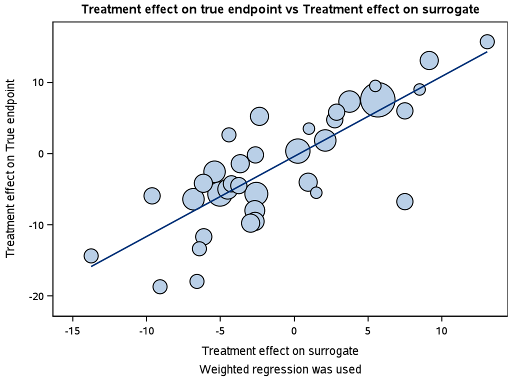


__Figure 10.4:__ ARMD study. Estimation of trial-level surrogacy using a two-stage model. Trial-specific treatment effects obtained from the reduced fixed-effects model. Circle areas are proportional to the trial size.

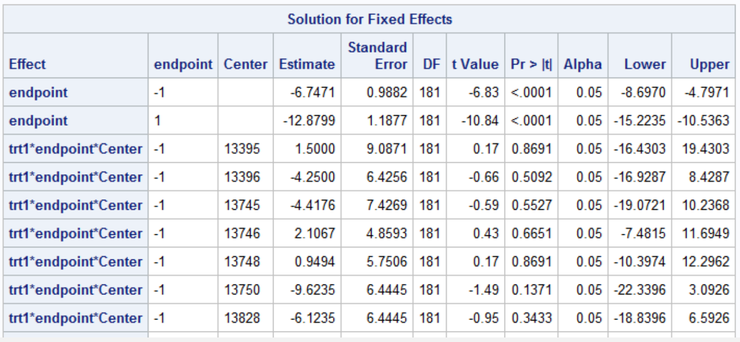
Similar to the analysis presented in the previous section, if the argument `looa=1` is used, a "leave-one-out" analysis is performed .

SAS Codes for the First-stage

The joint model formulated in equation (10.6) can be fitted using SAS procedure `MIXED` in the following way:

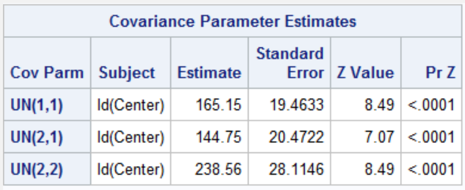
```
proc mixed data=dataset covtest;
class endp patid trial;
model outcome = endp endp*treat*trial / S noint ;
repeated endp / type=un subject=patid(trial) ;
ods output solutionF=eb CovParms=covar ;
run;
```

Note that the two-way interaction term `endp*treat` is dropped from the model statement and instead we use the variable `endp`, as result, a common intercept is fitted to the two endpoints. The panel below shows the parameter estimates for the first-stage model.



Solution for Fixed Effects										
Effect	endpoint	Center	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
endpoint	-1		-6.7471	0.9882	181	-6.83	<.0001	0.05	-8.6970	-4.7971
endpoint	1		-12.8799	1.1877	181	-10.84	<.0001	0.05	-15.2235	-10.5363
trt1*endpoint*Center	-1	13395	1.5000	9.0871	181	0.17	0.8691	0.05	-16.4303	19.4303
trt1*endpoint*Center	-1	13396	-4.2500	6.4256	181	-0.66	0.5092	0.05	-16.9287	8.4287
trt1*endpoint*Center	-1	13745	-4.4176	7.4269	181	-0.59	0.5527	0.05	-19.0721	10.2368
trt1*endpoint*Center	-1	13746	2.1067	4.8593	181	0.43	0.6651	0.05	-7.4815	11.6949
trt1*endpoint*Center	-1	13748	0.9494	5.7506	181	0.17	0.8691	0.05	-10.3974	12.2962
trt1*endpoint*Center	-1	13750	-9.6235	6.4445	181	-1.49	0.1371	0.05	-22.3396	3.0926
trt1*endpoint*Center	-1	13828	-6.1235	6.4445	181	-0.95	0.3433	0.05	-18.8396	6.5926

The individual-level surrogacy measure can be estimated using the residual covariance matrix shown in the panel below.



Covariance Parameter Estimates					
Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
UN(1,1)	Id(Center)	165.15	19.4633	8.49	<.0001
UN(2,1)	Id(Center)	144.75	20.4722	7.07	<.0001
UN(2,2)	Id(Center)	238.56	28.1146	8.49	<.0001

$$\hat{R}^2_{indiv} = \frac{144.75^2}{165.15 \times 238.56} = 0.5318. \tag{10.8}$$

The second-stage model, from which trial-level surrogacy is estimated, can be fitted in the same way as in Section (10.3.1).

10.3.3 The Full Mixed-effects Model

The first-stage proposed in equation (10.1) assumed trial-specific intercepts (μ_{Si} and μ_{Ti}), trial-specific treatment effects of Z on the endpoints (α_i and β_i) and error terms covariance matrix in equation (10.2). At the second-stage, it is assumed that

$$\begin{pmatrix} \mu_{S_i} \\ \mu_{T_i} \\ \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} m_{S_i} \\ m_{T_i} \\ a_i \\ b_i \end{pmatrix}, \tag{10.9}$$

where the second term on the right hand side is assumed to follow a zero-mean normal distribution with covariance matrix:

$$D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ & d_{TT} & d_{Ta} & d_{Tb} \\ & & d_{aa} & d_{ab} \\ & & & d_{bb} \end{pmatrix}. \tag{10.10}$$

The random effects representation is obtained by combining the first-stage in equation (10.1) and the second-stage in equation (10.9):

$$\begin{cases} S_{ij} = \mu_S + m_{Si} + (\alpha + a_i)Z_{ij} + \varepsilon_{Sij}, \\ T_{ij} = \mu_T + m_{Ti} + (\beta + b_i)Z_{ij} + \varepsilon_{Tij}. \end{cases} \tag{10.11}$$

Here, μ_S and μ_T are fixed intercepts, α and β are the fixed treatment effects on the two endpoints, m_{Si} and m_{Ti} are random intercepts, a_i and b_i are trial-specific random treatment effects. The vector of random effects $(m_{Si}, m_{Ti}, a_i, b_i)$ follows a normal distribution with zero-mean and covariance matrix in equation (10.10). The error terms, ε_{Sij} and ε_{Tij}, are assumed to follow a bivariate normal distribution with the covariance matrix in equation (10.2).

The quality of the surrogate at the trial-level may then be calculated as the coefficient of determination for predicting the effect of Z on T, given the effect of Z on S:

$$R^2_{trial(f)} = R^2_{b_i|m_{Si},a_i} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}}. \tag{10.12}$$

Coefficient in equation (10.12) is unitless and ranges in the unit interval if the corresponding variance-covariance matrix is positive definite, two desirable features for its interpretation (Buyse *et al.*, 2000).

At the individual-level, the association between endpoints is the squared correlation coefficient between S and T after adjusting for the trial and treatment effects:

$$R^2_{indiv} = R^2_{\epsilon_{Tij}|\epsilon_{Sij}} = \frac{\sigma^2_{ST}}{\sigma_{SS}\sigma_{TT}}. \tag{10.13}$$

The SAS Macro `%CONTRANFULL`

The `%CONTRANFULL` macro can be used to perform the analysis and it can be invoked using the following call:

```
%CONTRANFULL(data=simdata,true=true,surrog=surr,trt=treat,
trial=trial,patid=patientId,looa=1).
```

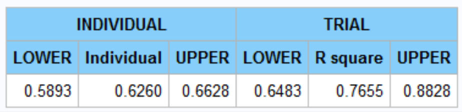
The macro's arguments are the same as those presented in Section 10.2.

Data Analysis and Output

Due to convergence problems with the full random effects, a simulated set of data was used to generate numerical and graphical outputs. The following parameters were used to simulate the data: 1000 observations from 50 trials were generated from a multivariate normal distribution with the mean vector $(\mu_S, \mu_T, \alpha, \beta) = (5, 5, 5, 5)$, and covariance matrices given by:

$$D = \begin{pmatrix} 10 & 8 & 0 & 0 \\ & 10 & 0 & 0 \\ & & 10 & 9 \\ & & & 10 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 10 & 8 \\ & 10 \end{pmatrix}.$$

As shown in the panel below, the surrogacy measures are equal to $\hat{R}^2_{indiv} = 0.6260$ (0.5893, 06628) and $\hat{R}^2_{trial} = 0.7655$ (0.6483, 0.8828).



INDIVIDUAL			TRIAL		
LOWER	Individual	UPPER	LOWER	R square	UPPER
0.5893	0.6260	0.6628	0.6483	0.7655	0.8828

Figure 10.5 shows the empirical Bayes estimates for the trial-specific random treatment effects.

SAS Codes for the Full Mixed-effects Model

The full mixed-effects model can be fitted using SAS procedure `MIXED`. A possible code is given below.

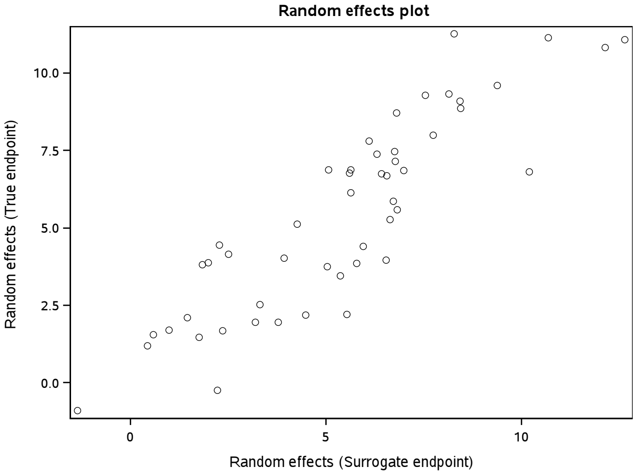
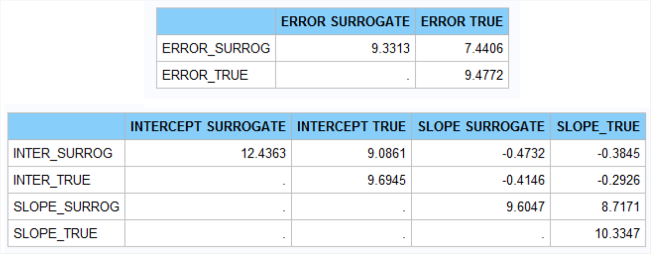


Figure 10.5: *Trial-specific random effects plot.*

```
proc mixed data=dataset covtest;
class endp patid trial;
model outcome = endp endp*treat / solution noint ;
random endp endp*treat/subject=trial type=un  ;
repeated endp / type=un subject=patid(trial) ;
ods output solutionF=fix CovParms=covar SolutionR=eb;
run;
```

The data structure and variables are identical to those outlined in Section 10.3.1. The `model` statement defines the four fixed-effects in the mean structure, $(\mu_T, \mu_S, \alpha, \beta)$, while the `random` statement defines the structure of the covariance matrix D for the random effects. The `repeated` statement builds up the error covariance matrix Σ in equation (10.2). The estimated covariance matrices are shown in the panel below.



	ERROR SURROGATE	ERROR TRUE
ERROR_SURROG	9.3313	7.4406
ERROR_TRUE	.	9.4772

	INTERCEPT SURROGATE	INTERCEPT TRUE	SLOPE SURROGATE	SLOPE_TRUE
INTER_SURROG	12.4363	9.0861	-0.4732	-0.3845
INTER_TRUE	.	9.6945	-0.4146	-0.2926
SLOPE_SURROG	.	.	9.6047	8.7171
SLOPE_TRUE	.	.	.	10.3347

The lower panel presents the parameter estimates of the covariance matrix D (given in equation 10.10).

$$\widehat{D} = \begin{pmatrix} 12.4363 & 9.0861 & -0.4732 & -0.3845 \\ & 9.6945 & -0.4146 & -0.2926 \\ & & 9.6047 & 8.7171 \\ & & & 10.3347 \end{pmatrix}. \tag{10.14}$$

The trial-level surrogacy measure is estimated using equation (10.12):

$$\hat{R}^2_{trial(f)} = \frac{\begin{pmatrix} -0.3845 \\ 8.7171 \end{pmatrix}^T \begin{pmatrix} 12.4363 & -0.4732 \\ -0.4732 & 9.6047 \end{pmatrix}^{-1} \begin{pmatrix} -0.3845 \\ 8.7171 \end{pmatrix}}{10.3347} = 0.7655. \tag{10.15}$$

The estimated covariance matrix for the residuals, defined in equation 10.2, is given by:

$$\widehat{\Sigma} = \begin{pmatrix} 9.3313 & 7.4406 \\ & 9.4772 \end{pmatrix}, \tag{10.16}$$

The individual-level surrogacy is derived using equation (10.13).

$$\hat{R}^2_{indiv} = \frac{7.4406^2}{9.3313 \times 9.4772} = 0.6260. \tag{10.17}$$

10.3.4 Reduced Mixed-effects Model

A special case of matrix in equation (10.10) is obtained when the random effects models in equation (10.11) do not contain the random intercepts m_{Si} and m_{Ti} at all. Using a simplified two-stage representation, the first-stage model can then be simplified to:

$$\begin{cases} S_{ij} = \mu_S + \alpha_i Z_{ij} + \varepsilon_{Sij}, \\ T_{ij} = \mu_T + \beta_i Z_{ij} + \varepsilon_{Tij}, \end{cases} \tag{10.18}$$

where, ε_{Sij} and ε_{Tij} are zero-mean normally distributed error terms with covariance matrix given in equation (10.2). The second-stage model is reduced to:

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} a_i \\ b_i \end{pmatrix}, \tag{10.19}$$

with, $(a_i, b_i)^T$ following a zero-mean normal distribution with covariance matrix

$$D_r = \begin{pmatrix} d_{aa} & d_{ab} \\ & d_{bb} \end{pmatrix}. \tag{10.20}$$

Combining the above two-stage, we have reduced mixed-effects given by :

$$\begin{cases} S_{ij} = \mu_S + (\alpha + a_i)Z_{ij} + \varepsilon_{Sij}, \\ T_{ij} = \mu_T + (\beta + b_i)Z_{ij} + \varepsilon_{Tij}. \end{cases} \tag{10.21}$$

With reduced random effects models in equation (10.21), the trial-level surrogacy is given by:

$$R^2_{trial(r)} = R^2_{b_i|a_i} = \frac{d^2_{ab}}{d_{aa}d_{bb}}, \tag{10.22}$$

and the individual-level surrogacy is given by:

$$R^2_{indiv} = \frac{d^2_{ST}}{d_{SS}d_{TT}}. \tag{10.23}$$

The SAS Macro `%CONTRANRED`

The macro `%CONTRANRED` is used to perform the analysis.

```
%CONTRANRED(data=simreduced,true=true,surrog=surr,trt=treat,
trial=trial,patid=patientId,looa=0).
```

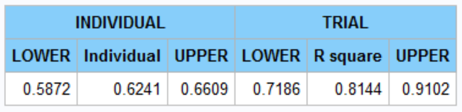
The macro's arguments were presented in Section 10.2.

Data Analysis and Output

Similar to full random effects models, convergence problems arise. Simulated data were used to generate numerical and graphical outputs. The following parameters were used to simulate the data: 1000 observations from 50 trials were generated from a multivariate normal distribution with the mean vector $(\mu_S, \mu_T, \alpha, \beta) = (5, 3, 5, 4)$, and covariance matrices given by:

$$D = \begin{pmatrix} 10 & 9 \\ & 10 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 3 & 2.4 \\ & 3 \end{pmatrix}.$$

Parameter estimates for trial- and individual-level surrogacy measures obtained for the reduced mixed-effects model are equal to $\hat{R}^2_{trial(r)}$ = 0.8144 (0.7186, 0.9102) and $\hat{R}^2_{indiv}$ = 0.6241 (0.5872, 0.6609), respectively (Panel below).



INDIVIDUAL			TRIAL		
LOWER	Individual	UPPER	LOWER	R square	UPPER
0.5872	0.6241	0.6609	0.7186	0.8144	0.9102

Figure 10.6 shows the empirical Bayed estimates for the random effects.

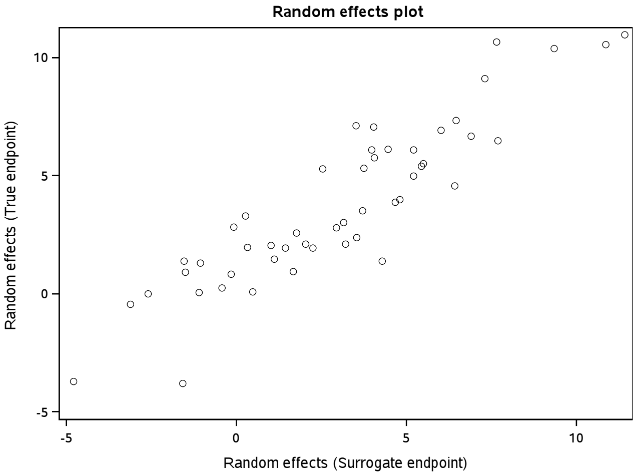


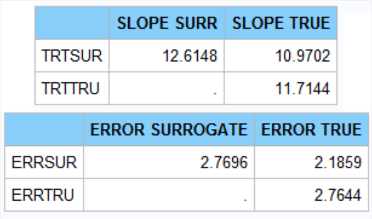
Figure 10.6: *Empirical Bayes estimates for the trial-specific random effects.*

SAS Codes for the Reduced Mixed-effects Model

The SAS code to fit the reduced mixed-effects model is:

```
proc mixed data=dataset covtest;
class endp patid trial;
model outcome = endp endp*treat / solution noint ;
random  endp*treat/subject=trial type=un  ;
repeated endp / type=un subject=patid(trial) ;
ods output solutionF=fix CovParms=covar SolutionR=eb;
run;
```

Note that, compared to the full mixed-effects model, the `random` statement was changed from `random endp endp*treat/subject=trial type=un` to `random endp*treat/subject=trial type=un` while the `model` and `repeated` statements remain the same. This implies that the covariance matrix for the trial-specific treatment effects is a 2 × 2 covariance matrix, while the `repeated` statement defines the error covariance matrix Σ in equation (10.2). The estimated matrices are shown below.



	SLOPE SURR	SLOPE TRUE
TRTSUR	12.6148	10.9702
TRTTRU	.	11.7144

	ERROR SURROGATE	ERROR TRUE
ERRSUR	2.7696	2.1859
ERRTRU	.	2.7644

Trial-level surrogacy is estimated using equation (10.22),

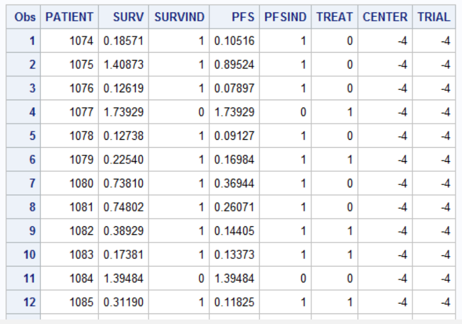
$$\hat{R}^2_{trial(r)} = \frac{10.9702^2}{12.6148 \times 11.7144} = 0.8144, \tag{10.24}$$

and individual-level surrogacy is derived as follows (see equation (10.23)):

$$\hat{R}^2_{indiv} = \frac{2.1859^2}{2.7696 \times 2.7644} = 0.6241. \tag{10.25}$$

10.4 Analysis of a Surrogacy Setting with Two Survival Endpoints

In this section, we discuss about settings in which the two endpoints are time-to-event variables. We focus on three applications. Two of the applications are based on a two-stage approach while the third application is based on a joint modeling of two time-to-event endpoints. For illustration we used the ovarian cancer study with overall survival time and progression-free survival as the true and surrogate endpoint, respectively. See Section 9.1.2 for more information about the data. A partial printout of the data is given below. The data for each patient appear in a single line.



Obs	PATIENT	SURV	SURVIND	PFS	PFSIND	TREAT	CENTER	TRIAL
1	1074	0.18571	1	0.10516	1	0	-4	-4
2	1075	1.40873	1	0.89524	1	0	-4	-4
3	1076	0.12619	1	0.07897	1	0	-4	-4
4	1077	1.73929	0	1.73929	0	1	-4	-4
5	1078	0.12738	1	0.09127	1	0	-4	-4
6	1079	0.22540	1	0.16984	1	1	-4	-4
7	1080	0.73810	1	0.36944	1	0	-4	-4
8	1081	0.74802	1	0.26071	1	0	-4	-4
9	1082	0.38929	1	0.14405	1	1	-4	-4
10	1083	0.17381	1	0.13373	1	1	-4	-4
11	1084	1.39484	0	1.39484	0	0	-4	-4
12	1085	0.31190	1	0.11825	1	1	-4	-4

10.4.1 A Two Stage Approach(I)

In case the true endpoint T_{ij} and the surrogate S_{ij} endpoint are time-to-event endpoints, the approach used in Section 10.3.1 have to be replaced by model for two correlated time-to-event random variables (Burzykowski *et al.*, 2001). The chosen model should provide an association measurement between the two time-to-event variables. We start our discussion with a two-stage approach. It is assumed that the data from a single trial with many centers (or data from many trials, where trial is considered as unit of analysis) are available. With this setting, Buyse *et al.* (2011) proposed a validation approach in which the estimated treatment effects on both endpoints must be correlated. To test this condition, the center-specific (trial-specific) Cox proportional hazard models in equation (10.26) were used:

$$\begin{cases} S_{ij}(t) = S_{i0}(t)\exp(\alpha_i Z_{ij}), \\ T_{ij}(t) = T_{i0}(t)\exp(\beta_i Z_{ij}), \end{cases} \tag{10.26}$$

where, $S_{i0}(t)$ and $T_{i0}(t)$ are trial-specific baseline hazard functions, Z_{ij} is a treatment indicator for the jth individual in the ith trial. The parameters β_i and α_i are trial-specific treatment effects.

One way to account for variation in trials size is to use the number of patients in each trial in a weighted linear regression of the form:

$$\widehat{\beta}_i = \lambda_0 + \lambda_1 \widehat{\alpha}_i + \varepsilon_i. \tag{10.27}$$

A second approach to account for the variability between trials is to use a robust sandwich estimate of Lin and Wei (1989) for the covariance matrix of the parameter estimates for treatment effects in equation (10.26) and follow the approach proposed by Van Houwelingen *et al.* (2002). The two approaches are implemented in the macro `%TWOSTAGECOX` discussed below. As before, the coefficient of determination from equation (10.27) is used as a trial-level surrogacy measure. A leave-one-out analysis can be performed in order to asses stability of the estimated trial-level surrogacy measure.

The SAS Macro %TWOSTAGECOX

The model in equation (10.26) can be fitted using the SAS macro `%TWOSTAGECOX`. It has the following generic form:

```
%TWOSTAGECOX(data=ovarian,true=surv,trueind=survind,
             surrog=pfs,surrogind=pfsind,trt=treat,
             trial=center,patid=patient,common=1,
             robust=1,looa=1);
```

The macro's specific arguments are

- `trueind`: censoring indicator for the true endpoint (1=event, 0=censoring).
- `surrogind`: censoring indicator for the surrogate (1=event, 0=censoring).
- `common`: an option allows the user to choose between trial-specific baseline hazard function (common=0), or common baseline hazard function (common=1) in the first-stage.
- `robust`: an option allows the user to obtain the robust (or adjusted) R^2_{trial} in the output (1 for the robust; 0 for non-robust).

The rest of the arguments were discussed in Section 10.2.

Data Analysis and Output

The macro `%TWOSTAGECOX` produces two exploratory plots: the patients' distribution in the trials by treatment arms (shown in Figure 10.7 left panel) and the Kaplan-Meier curves for the true and surrogate endpoints (shown in Figure 10.7 right panel).

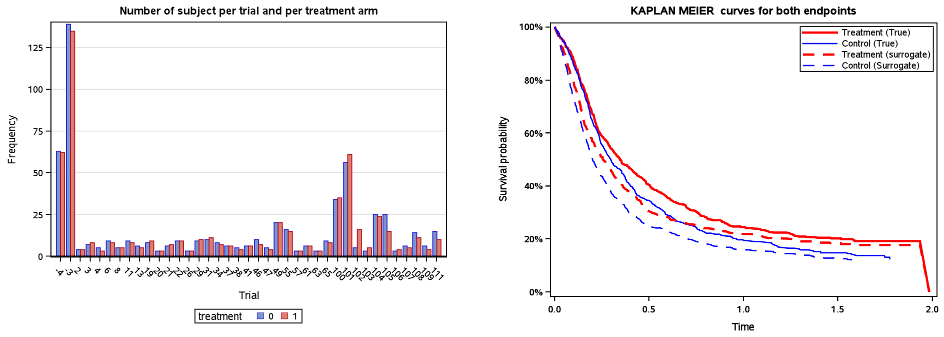
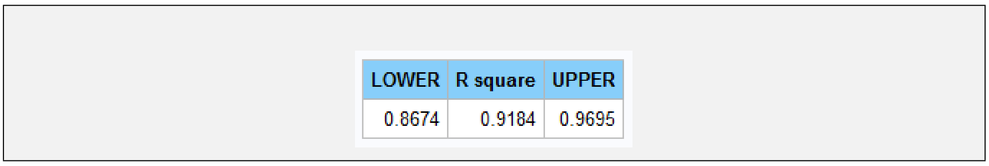


Figure 10.7: *The Ovarian Cancer study. Left Panel: Patients' distribution by treatment arms within centers . Right Panel: Kaplan Meier curves for the survival time (true endpoint) and progression-free survival (surrogate endpoint) for the two treatment groups CP and CAP.*

As shown below, the estimated trial-level surrogacy is equal to $\hat{R}^2_{trial} = 0.9184$ (0.8674, 0.9695). This implies that progression-free survival time is a good surrogate for the overall survival time. This can be clearly seen from the left panel of Figure 10.8. Leave-one-out analysis results are shown in the right panel of Figure 10.8.



LOWER	R square	UPPER
0.8674	0.9184	0.9695

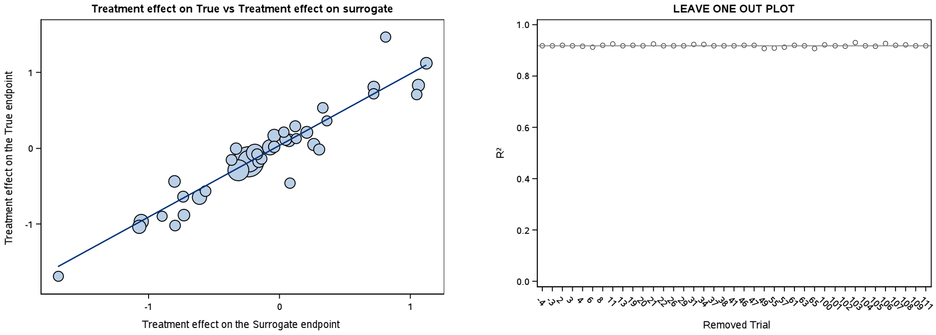


Figure 10.8: *The Ovarian Cancer study. Left Panel: Parameter estimates from the second-stage model. Right Panel: Leave-one-out analysis plot.*

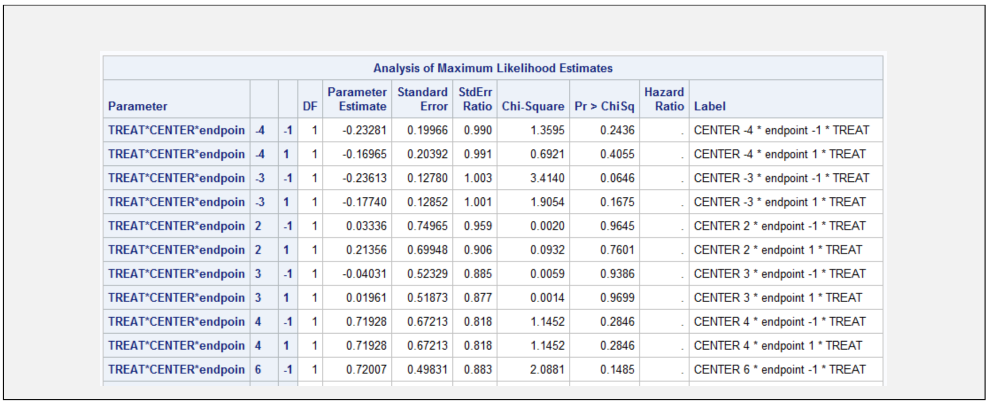
SAS Code for the First-stage Model

The Cox proportional hazard model formulated in equation (10.26) can be fitted using SAS procedure `PHREG`.

```
proc phreg data=firststage covs(aggregate) covout outest=mat;
class  center endpoint/param=glm;
model outcome*status(0)=  endpoint*treat*center ;
strata center endpoint;
id patid ;
run;
```

It is assumed that there are two records per subject in the input dataset, the first one corresponds to the surrogate endpoint and the second one to the true endpoint. The option `covs(aggregate)` applies the robust sandwich estimate of Lin and Wei (1989) for the covariance matrix that can be used in the second-stage to correct for the uncertainty in the estimated parameters. The `outest` option creates an output SAS data set containing estimates of the regression coefficients and the option `covout` adds the estimated covariance matrix of the parameter estimates to the `outest` data set. In the model statement, `status` is the censoring indicator variable (0=censoring, 1=event). In the `strata` statement we specify the variables that determine the stratification.

The parameter estimates obtained from model in equation (10.26) are shown in the panel below.



Analysis of Maximum Likelihood Estimates

Parameter			DF	Parameter Estimate	Standard Error	StdErr Ratio	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
TREAT*CENTER*endpoin	-4	-1	1	-0.23281	0.19966	0.990	1.3595	0.2436	.	CENTER -4 * endpoint -1 * TREAT
TREAT*CENTER*endpoin	-4	1	1	-0.16965	0.20392	0.991	0.6921	0.4055	.	CENTER -4 * endpoint 1 * TREAT
TREAT*CENTER*endpoin	-3	-1	1	-0.23613	0.12780	1.003	3.4140	0.0646	.	CENTER -3 * endpoint -1 * TREAT
TREAT*CENTER*endpoin	-3	1	1	-0.17740	0.12852	1.001	1.9054	0.1675	.	CENTER -3 * endpoint 1 * TREAT
TREAT*CENTER*endpoin	2	-1	1	0.03336	0.74965	0.959	0.0020	0.9645	.	CENTER 2 * endpoint -1 * TREAT
TREAT*CENTER*endpoin	2	1	1	0.21356	0.69948	0.906	0.0932	0.7601	.	CENTER 2 * endpoint 1 * TREAT
TREAT*CENTER*endpoin	3	-1	1	-0.04031	0.52329	0.885	0.0059	0.9386	.	CENTER 3 * endpoint -1 * TREAT
TREAT*CENTER*endpoin	3	1	1	0.01961	0.51873	0.877	0.0014	0.9699	.	CENTER 3 * endpoint 1 * TREAT
TREAT*CENTER*endpoin	4	-1	1	0.71928	0.67213	0.818	1.1452	0.2846	.	CENTER 4 * endpoint -1 * TREAT
TREAT*CENTER*endpoin	4	1	1	0.71928	0.67213	0.818	1.1452	0.2846	.	CENTER 4 * endpoint 1 * TREAT
TREAT*CENTER*endpoin	6	-1	1	0.72007	0.49831	0.883	2.0881	0.1485	.	CENTER 6 * endpoint -1 * TREAT

10.4.2 A Two Stage Approach(II)

In this section, we discuss the evaluation approach, proposed by Buyse *et al.* (2011), to compute the individual-level surrogacy. The underlying idea behind this evaluation approach is that S_{ij} can be considered as a valid surrogate for T_{ij} with respect to a new treatment if the pair of endpoints scores sufficiently highly on the validations measures. Buyse *et al.* (2011) proposed to use the Kaplan Meier (KM) estimates (for both endpoints) at fixed time points and to estimate the correlation between the KM estimates of the two endpoints. Kaplan Meier estimates per trial and per endpoint are estimated using the following formula:

$$S(t_i) = \prod_{t_i \leq t} (1 - \frac{d_i}{n_i}). \tag{10.28}$$

Here, $S(t_i)$ is the estimated survival probability, d_i is the number of patients who had an event at time t_i, and n_i is the number of patients who are at risk at that time. The estimated values are denoted by $\hat{\beta}_i$, $\hat{\alpha}_i$ for the true and surrogate endpoints respectively. Figure 10.9 shows schematically how to choose the time points used to compute the KM estimates on both endpoint (for a given trial).

KM estimates on the true and the surrogate endpoints at given time t_T, and t_S , with $t_T > t_S$, are used to fit the following linear regression in order to test for their association.

$$\widehat{\beta}_i = \lambda_0 + \lambda_1 \widehat{\alpha}_i + \varepsilon_i. \tag{10.29}$$

The effective sample size at the time point considered for KM estimates (the number

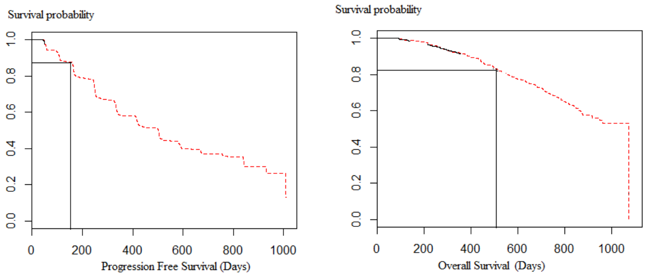


Figure 10.9: *KM courves for the true and the surrogate endpoints for a given time point.*

of deaths prior to the time point plus the number of patients at risk at the time point) can be used for each trial as a weight. The coefficient of determination, R^2, obtained from the linear regression model in equation (10.29) can be used to quantify the surrogacy measure at individual-level.

The SAS Macro `%TWOSTAGEKM`

The model discussed in this section can be fitted using the SAS macro `%TWOSTAGEKM`. The macro has the following general form:

```
%TWOSTAGEKM (data=ovarian,true=surv,trueind=survind,surrog=pfs,
             surrogind=pfsind,trt=treat,trial=center,
             upsurr=1,uptrue=2)
```

The macro's arguments are:

- `upsurr`: the time point at which the KM estimates are computed on the surrogate endpoint.
- `uptrue:` the time point at which the KM estimates are computed on the true endpoint.

The rest of the arguments have been defined in Sections 10.2 and 10.4.1.

Data Analysis and Output

The `%TWOSTAGEKM` macro produces the Kaplan-Meier curves for both endpoints (Figure 10.10).

For the example in this section, we use the KM estimates at one year for progression-free survival and the KM estimates at two years for overall survival. The estimated

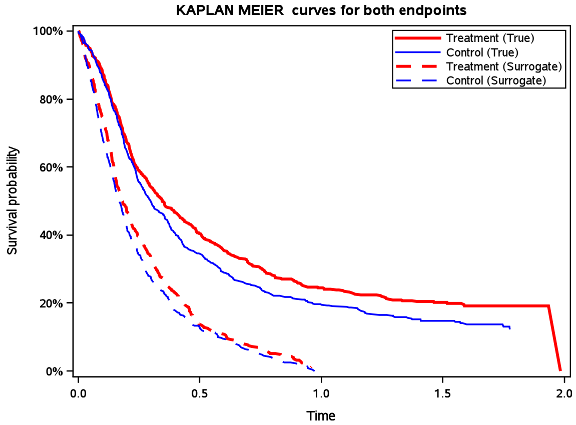
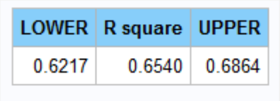


Figure 10.10: *Kaplan-Meier curve for the true endpoint at 2 years and for the surrogate endpoint at 1 year.*

individual-level surrogacy measure for the ovarian cancer study is shown in the panel below and is equal to $\hat{R}^2_{indiv} = 0.6540\ (0.6217,\ 0.6864)$. This indicates that progression-free survival at one year is a surrogate of moderate value for the overall survival at two years. This can be seen in Figure 10.11 which presents the KM estimates of progression-free survival at one year versus the KM estimates for overall survival at two years.



LOWER	R square	UPPER
0.6217	0.6540	0.6864

SAS Codes for Trial Specific KM Estimates (At a Given Time Point)

The survival probability based on the KM curves can be estimated using SAS procedure `lifetest` in the following way:

```
proc lifetest data=ovar;
time surv * survind(0);
strata treat;
where surv<=2;
by trial;
run;
```

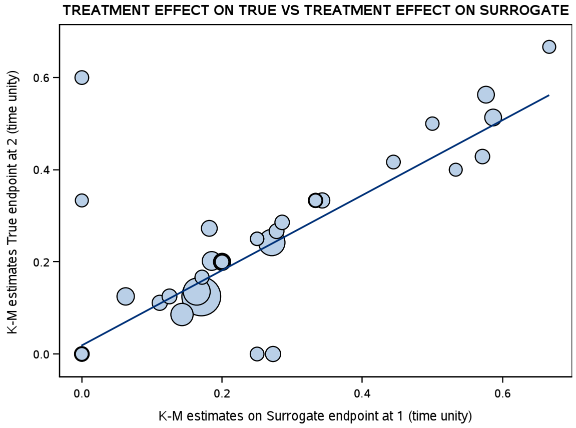
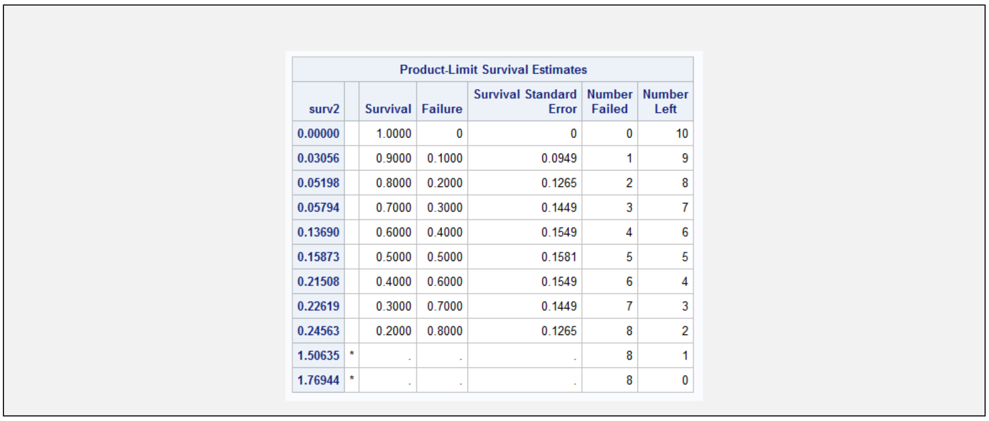


Figure 10.11: *KM estimates at 2 years on the true endpoint vs KM estimates at 1 year on the surrogate endpoint.*

The variables `surv` and `survind` are the survival time and the censoring indicator, respectively. The statement `strata` produces the results for the active treatment and control groups separately. The statement `where` statement allows to select observations for which the survival time is less or equal to 2 years (for the true endpoint). The panel below shows KM estimates from one trial.



Product-Limit Survival Estimates

surv2		Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.00000		1.0000	0	0	0	10
0.03056		0.9000	0.1000	0.0949	1	9
0.05198		0.8000	0.2000	0.1265	2	8
0.05794		0.7000	0.3000	0.1449	3	7
0.13690		0.6000	0.4000	0.1549	4	6
0.15873		0.5000	0.5000	0.1581	5	5
0.21508		0.4000	0.6000	0.1549	6	4
0.22619		0.3000	0.7000	0.1449	7	3
0.24563		0.2000	0.8000	0.1265	8	2
1.50635	*	.	.	.	8	1
1.76944	*	.	.	.	8	0

10.4.3 A Joint Model for Survival Endpoints

In this section, a bivariate copula model proposed by Burzykowski *et al.* (2001) is used to measure the association between the true endpoint and the surrogate endpoint. This model is more satisfactory in that the correlations reflect the whole time axis instead of Kaplan-Meier estimates at specific time points.
More specifically, Burzykowski *et al.* (2001) used a joint survival function for (S_{ij}, T_{ij})

given by:

$$F(s,t) = P(S_{ij} \geq s, T_{ij} \geq t) = C_\theta\{F_{Sij}(s), F_{Tij}(t)\}, \quad s,t \geq 0, \tag{10.30}$$

where, F_{Sij} and F_{Tij} denote marginal survival functions for both endpoints (overall survival and progression-free survival) and C_θ is a copula, i.e., a bivariate distribution function on $[0,1]^2$ which allows correlated probabilities to be modeled. The marginal survival functions are given by:

$$\begin{cases} F_{S_{ij}}(s) = \exp\{-\int_0^s \lambda_{Si}(x)\exp(\alpha_i Z_{ij})\,dx\}, \\ F_{T_{ij}}(t) = \exp\{-\int_0^t \lambda_{Ti}(x)\exp(\beta_i Z_{ij})\,dx\}, \end{cases} \tag{10.31}$$

where, λ_{Si} and λ_{Ti} are trial-specific marginal baseline hazard functions and α_i and β_i are trial-specific treatment effects. At the second-stage, a joint model is formulated to the treatment effects

$$\begin{pmatrix}\alpha_i\\ \beta_i\end{pmatrix} = \begin{pmatrix}\alpha\\ \beta\end{pmatrix} + \begin{pmatrix}a_i\\ b_i\end{pmatrix}, \tag{10.32}$$

where the second term on the right hand side of equation (10.32) is assumed to follow a zero-mean normal distribution with a covariance matrix given by:

$$D = \begin{pmatrix} d_{aa} & d_{ab} \\ & d_{bb}\end{pmatrix}. \tag{10.33}$$

The quality of the surrogate S at the trial-level is assessed based on the coefficient of determination given by:

$$R^2_{trial} = \frac{d^2_{ab}}{d_{aa}d_{bb}}. \tag{10.34}$$

To assess the quality of the surrogate at the individual-level, a measure of association between S_{ij} and T_{ij}, calculated while adjusting the marginal distributions of the two endpoints for both the trial and treatment effects, is needed. Burzykowski *et al.* (2001) proposed to use Kendall's τ as it only depends on the copula function C_θ and is independent of the marginal distribution of S_{ij} and T_{ij}:

$$\tau = 4\int_0^1\int_0^1 C_\theta(F_{S_{ij}}, F_{T_{ij}})C_\theta(dF_S, dF_T) - 1. \tag{10.35}$$

It describes the strength of the association between the two endpoints remaining after adjustment, through the marginal models in equation (10.31), for the trial and the treatment effects.

The SAS Macro %COPULA

The SAS macro %COPULA can be used to conduct the analysis discussed in this section in the following way:

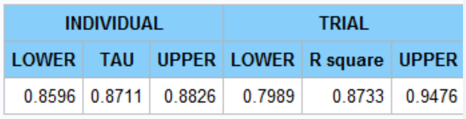
```
%COPULA(data=covar,true=Surv,trueind=survind,surrog=pfs,
        surrogind=pfsind,trt=treat,center=center,trial=center,
        vars=,patid=patientid,copula=clayton,adjustment=Weighted)
```

The macro's arguments are :

- `center`: unique id (continuous) for units, for which specific treatment effects are estimated.
- `trial`: unique id (continuous) for groups of the units, for which common ”baselines” are used.
- `vars`: macro variable containing possible covariates for adjustment of the Weibull and proportional odds models. The names of these covariates have to be passed to the program trough the macro variable `"vars"`.
- `copula`: variable allowing the user to choose one of the three different copulas (`clayton`, `houggard`, `placket`).
- `adjustment`: Adjustment method used to compute the R^2_{trial} (Weighted, Unweighted, adjustedr2, adjustedrcorr, adjustedr2f).

Data Analysis and Output

The exploratory plots produced by the macro %COPULA are shown in Figure 10.7.



INDIVIDUAL			TRIAL		
LOWER	TAU	UPPER	LOWER	R square	UPPER
0.8596	0.8711	0.8826	0.7989	0.8733	0.9476

For the ovarian cancer study, the surrogacy measures $\hat{R}^2_{indiv} = 0.8711$ (0.8596, 0.8826) and $\hat{R}^2_{trial} = 0.8733$ (0.7989, 0.9476) indicate that the progression-free survival is a valid surrogate for the overall survival time at both trial- and individual-level surrogacy. The treatment effects plot in Figure 10.12 can be used to visualize trial-level surrogacy.

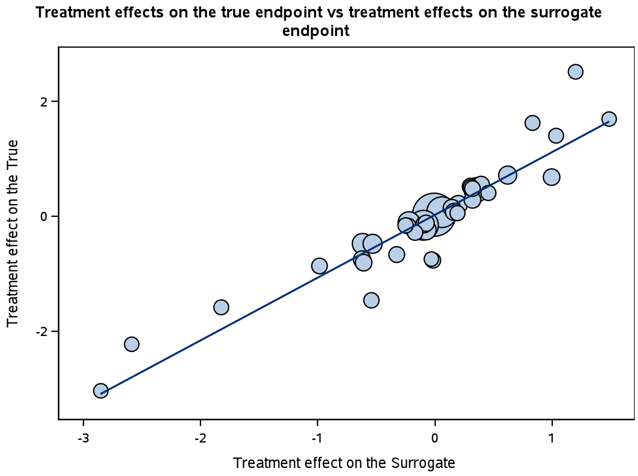


Figure 10.12: *The Ovarian Cancer study. Treatment effects on the true endpoints (survival time) versus treatment effects on the surrogate endpoints (Progression free survival). Circles areas are proportional to trial size.*

10.5 A Continuous (Normally-Distributed) and a Survival Endpoint

Model Formulation

In this section, it is assumed that the true endpoint, T, is a failure-time random variable and the surrogate, S, is a normally-distributed continuous variable. Note that the described approach is applicable also in the reverse case, i.e., with a failure-time surrogate and a continuous true endpoint.

Alonso *et al.* (2016) assumed that the true endpoint T is a failure-time random variable and the surrogate S is a normally-distributed, continuous variable. For each of $j = 1, \ldots, n_i$ patients from trial i ($i = 1, \ldots, N$) the quadruplets $(X_{ij}, \Delta_{ij}, S_{ij}, Z_{ij})$ is obtained, where X_{ij} is a possibly censored version of survival time T_{ij} and Δ_{ij} is the censoring indicator assuming value of 1 for observed failures and 0 otherwise.

The marginal model for S_{ij} is the classical linear regression model:

$$S_{ij} = \alpha_{0,i} + \alpha_i Z_{ij} + \varepsilon_{ij}, \tag{10.36}$$

where, ε_{ij} is normally distributed with mean zero and variance σ_i^2.

For T_{ij}, the proportional hazard model is given by:

$$\lambda_{ij}(t|Z_{ij}) = \lambda_i(t)\exp(\beta_i Z_{ij}), \tag{10.37}$$

here, β_i are trial-specific effects of treatment Z and $\lambda_i(t)$ is a trial-specific baseline hazard function.

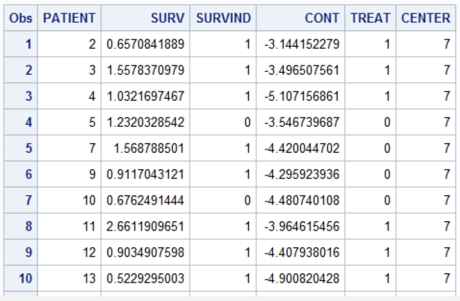
If a parametric (e.g., Weibull-distribution-based) baseline hazard is used in equation (10.37), the joint distribution function defined by the copula and the marginal models in equation (10.36) and (10.37) allows to construct the likelihood function for the observed data and obtaining estimates of the treatment effects α_i and β_i.

The individual-level surrogacy can be evaluated by using Kendall's τ or Spearman's ρ (see Section 10.4.3). The trial-level surrogacy is assessed using the correlation coefficient between the estimated treatment effects α_i and β_i.

If the individual-level association is not of immediate interest, one may base analysis on the marginal models in equation (10.36) and (10.37), without specifying the baseline hazards in the latter. When fitting the models, it is worth to estimate the variance-covariance matrix of the estimated treatment effects $\widehat{\alpha}_i$ and $\widehat{\beta}_i$ while taking into account the association between S and T.

Data Structure

The advanced prostate cancer data described in Section 9.1.5 was used for illustration. The true endpoint is overall survival time and the surrogate endpoint is the logarithm of prostate specific antigen (PSA) measured at about 28 days. The data structure for the survival-normal setting is shown below. The data for each subject appears in a single line.



Obs	PATIENT	SURV	SURVIND	CONT	TREAT	CENTER
1	2	0.6570841889	1	-3.144152279	1	7
2	3	1.5578370979	1	-3.496507561	1	7
3	4	1.0321697467	1	-5.107156861	1	7
4	5	1.2320328542	0	-3.546739687	0	7
5	7	1.568788501	1	-4.420044702	0	7
6	9	0.9117043121	1	-4.295923936	0	7
7	10	0.6762491444	0	-4.480740108	0	7
8	11	2.6611909651	1	-3.964615456	1	7
9	12	0.9034907598	1	-4.407938016	1	7
10	13	0.5229295003	1	-4.900820428	1	7

The SAS Macro `%NORMSURV`

The SAS macro `%NORMSURV` can be used in order to fit the models specified in equation (10.36) and (10.37). For the Prostate cancer data we use:

```
%NORMSURV(data=prostate,true=surv,trueind=survind,surrog=cont,
trt=treat,trial=center,patid=patientId,copula=houggard,
    adjustement=weighted)
```

The specification of the macro's arguments is the same as the specification presented in Section (10.4.3).

Data Analysis and Output

The exploratory plots produced by the macro %NORMSURV are shown in Figure 10.13. The histogram in the upper right panel suggests that the logarithm of PSA at 28 days is normally distributed. The scatter plot for the survival time and the continuous surrogate in the lower panel reveals a weak association (ignoring censoring on the true endpoint).

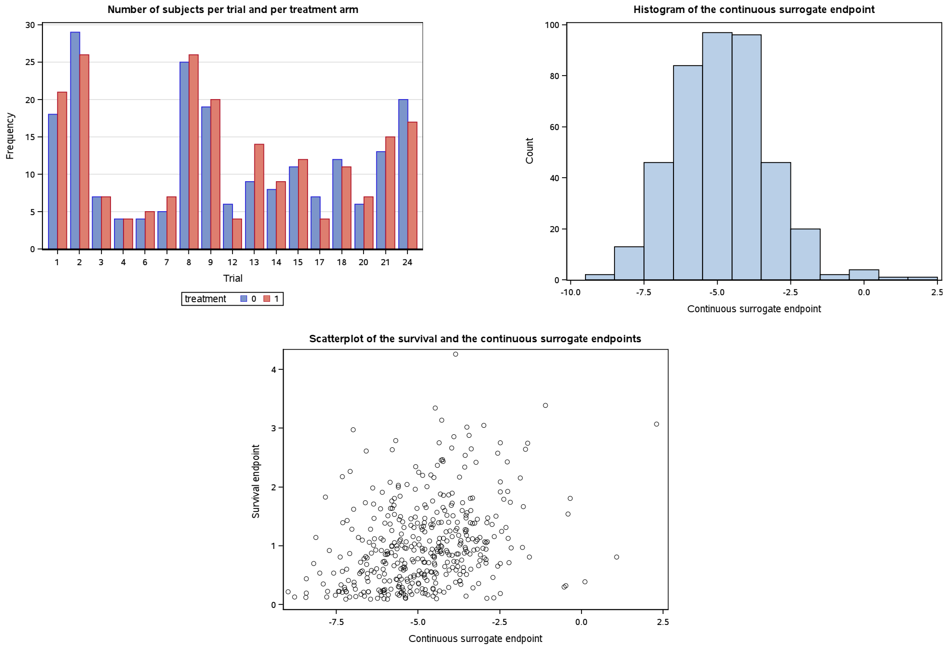
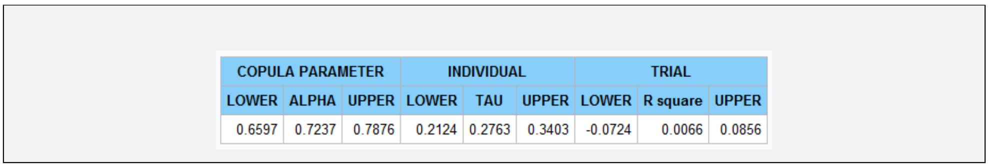


Figure 10.13: *Prostate Cancer study. Descriptive plots for the. Upper Left Panel: Patients distribution by treatment arms across trials. Upper Right Panel: Histogram for the continuous surrogate endpoint. Lower Panel: Scatter plot between the survival time (ignoring censoring) true endpoint and the continuous surrogate endpoint.*

Individual and trial-level surrogacy, Kendall's $\tau = 0.2763$ (0.2124, 0.3403) and $R^2_{trial} =$ 0.0066 (-0.0724, 0.0856), shown in the panel below indicate that the logarithm of PSA after 28 days is a weak surrogate to overall survival time for the prostate cancer data. The Houggard copula parameter is presented in the output as well.



COPULA PARAMETER			INDIVIDUAL			TRIAL		
LOWER	ALPHA	UPPER	LOWER	TAU	UPPER	LOWER	R square	UPPER
0.6597	0.7237	0.7876	0.2124	0.2763	0.3403	-0.0724	0.0066	0.0856

Figure 10.14 shows the parameter estimates for the treatment effects for both surrogate and true endpoints that were used to estimate trial-level surrogacy.

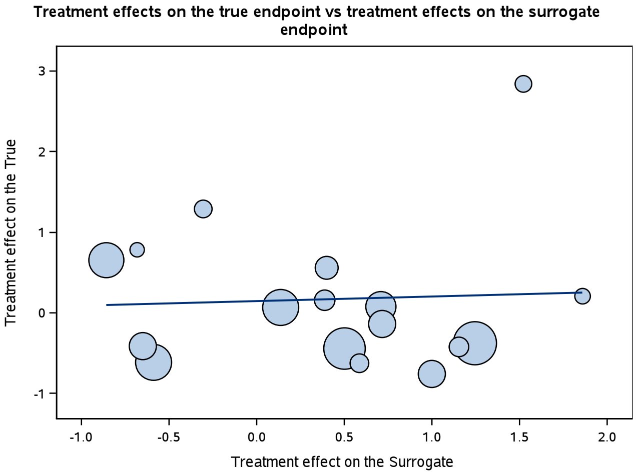


Figure 10.14: *Prostate Cancer study. Evaluation of trial-level surrogacy . Treatment effects upon the true endpoints (log hazard ratio) versus treatment effects upon the continuous surrogate. The circle areas are proportional to the sample size of the trial.*

10.6 Validation Using Joint Modeling of Time-to-event and a Binary Endpoint

The setting we consider in this section consists of a binary surrogate endpoint and a time-to-event true endpoint. A joint model, proposed by Burzykowski *et al.* (2004), is formulated for the true endpoint T_{ij} and a latent normally distributed variable $\tilde{S}_{ij}$. The binary surrogate is defined by:

$$S_{ij} = \begin{cases} 1 & \text{if} \quad \tilde{S}_{ij} > 0, \\ 0 & \text{if} \quad \tilde{S}_{ij} \leq 0. \end{cases} \tag{10.38}$$

For the surrogate endpoint, S_{ij}, a logistic regression model is assumed.

$$logit\{P(S_{ij} = 1|Z_{ij})\} = \gamma_i + \alpha_i Z_{ij}. \tag{10.39}$$

The marginal cumulative distribution function of $\tilde{S}_{ij}$, given $Z_{ij} = z$, is denoted by $F_{\tilde{S}_{ij}}(s; z)$.

To model the effect of treatment on the marginal distribution of T_{ij}, Burzykowski *et al.* (2004) proposed to use a proportional hazard model of the form:

$$\lambda_{ij}(t|Z_{ij}) = \lambda_i(t)\exp(\beta_i Z_{ij}), \tag{10.40}$$

where, β_i are trial-specific treatment effects, $\lambda_i(t)$ is a trial-specific baseline hazard function. The marginal cumulative distribution function of T_{ij}, with $Z_{ij} = z$, is denoted by $F_{T_{ij}}(t; z)$. The joint cumulative distribution of T_{ij} and $\tilde{S}_{ij}$, given $Z_{ij} = z$, is generated by one parameter copula function C_θ (Burzykowski *et al.*, 2004):

$$F_{T_{ij},\tilde{S}_{ij}}(t, s; z) = C_\theta\{F_{T_{ij}}(t; z), F_{\tilde{S}_{ij}}(s; z), \theta\}. \tag{10.41}$$

Here, C_θ is a distribution function on $[0, 1]^2$ with $\theta \in \mathbb{R}^1$.

The two-stage approach proposed by Burzykowski *et al.* (2004) consists of maximum likelihood estimation for θ and the trial-specific treatment effects α_i and β_i at the first-stage while at the second-stage, it is assumed that:

$$\begin{pmatrix}\alpha_i\\ \beta_i\end{pmatrix} = \begin{pmatrix}\alpha\\ \beta\end{pmatrix} + \begin{pmatrix}a_i\\ b_i\end{pmatrix}. \tag{10.42}$$

The second term on the right hand side of equation (10.42) is assumed to follow a bivariate normal distribution with zero-mean and covariance matrix given by:

$$D = \begin{pmatrix}d_{aa} & d_{ab}\\ & d_{bb}\end{pmatrix}. \tag{10.43}$$

Hence, the trial-level surrogacy is estimated by:

$$R^2_{trial} = \frac{d^2_{ab}}{d_{aa}d_{bb}}. \tag{10.44}$$

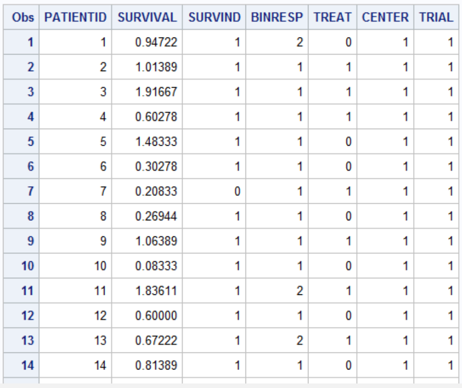
To assess the quality of the surrogate endpoint at the individual-level, a measure of association between S_{ij} and T_{ij} is needed. Burzykowski *et al.* (2004) proposed to use the bivariate Plackett copula. This particular choice was motivated by the fact that, for the Plackett copula, the association parameter θ takes the form of a (constant) global odds ratio.

$$\theta = \frac{P(T_{ij} > t, S_{ij} > k)P(T_{ij} \le t, S_{ij} \le k)}{P(T_{ij} > t, S_{ij} \le k)P(T_{ij} \le t, S_{ij} > k)}. \tag{10.45}$$

For a binary surrogate, it is just the odds ratio for responders versus non-responders (assuming k=2 indicates the response).

Data Structure

We use the Colorectal cancer study for illustration (see Section 9.1.3). The true endpoint is overall survival and the surrogate endpoint is a two-category tumor response: patients with complete or partial response are considered as responders and patients with stable or progressive disease are considered as non-responders.
The data structure for the survival-binary setting is shown below. The data for each subject appears in a single line in which time-to-event (`surv`) and censoring status (`survind`) are given to the true endpoint and the response status (`binresp`) is the surrogate endpoint. The unit for the analysis is the `trial`.



Obs	PATIENTID	SURVIVAL	SURVIND	BINRESP	TREAT	CENTER	TRIAL
1	1	0.94722	1	2	0	1	1
2	2	1.01389	1	1	1	1	1
3	3	1.91667	1	1	1	1	1
4	4	0.60278	1	1	1	1	1
5	5	1.48333	1	1	0	1	1
6	6	0.30278	1	1	0	1	1
7	7	0.20833	0	1	1	1	1
8	8	0.26944	1	1	0	1	1
9	9	1.06389	1	1	1	1	1
10	10	0.08333	1	1	0	1	1
11	11	1.83611	1	2	1	1	1
12	12	0.60000	1	1	0	1	1
13	13	0.67222	1	2	1	1	1
14	14	0.81389	1	1	0	1	1

The SAS Macro %SURVCAT

The joint model discussed above is implemented in the macro `%SURVCAT` that, for the Colorectal cancer data, is called as follows:

```
%SURVCAT(data=colorectal,true=Surv,trueind=survind,
         surrog=responder,trt=treat,center=center,
         trial=center,vars=,patid=patientid)
```

The macro's arguments were discussed in Section 10.2 and 10.4.1. The argument `surrog:` measurement of the ordinal categorical surrogate (levels 1, 2, ..., K).

Data Analysis and Output

The `%SURVCAT` macro produces three default exploratory plots shown in Figure 10.15. The Kaplan-Meier curves (by endpoint, for only binary surrogate) in the upper left panel indicates that there is no difference between the treatment arms across the level of the surrogate endpoint. The box plots for the survival times in the upper right panel reveal the same pattern. The number of patients per trial and treatment arms is shown in the lower panel.

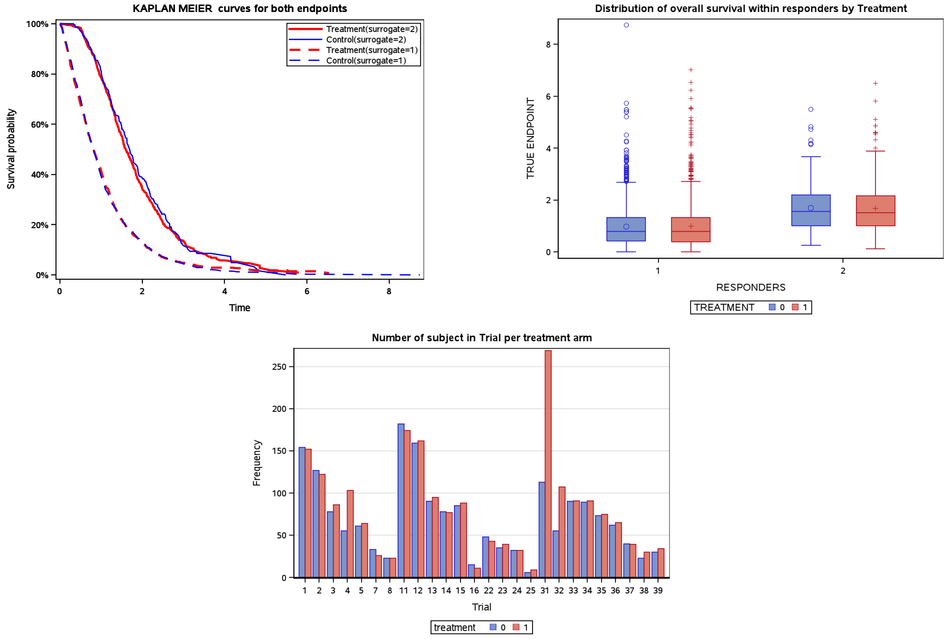
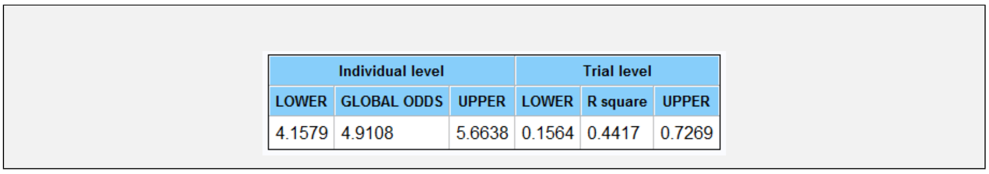


Figure 10.15: *Colorectal Cancer study. Descriptive plots. Upper Left Panel: KM curves stratified by the binary surrogate endpoint. Upper Right Panel: Survival time distribution by treatment arm across the levels of the binary surrogate. Lower Panel: distribution of patients per trial and treatment arm.*

Individual- and trial-level surrogacy, Global odds = 4,9108 (4.15794, 5.6638) and R^2_{trial} = 0.4417 (0.1564, 0.7269), shown in the panel below indicate that two-category tumor response is a surrogate of moderate value to overall survival for the colorectal cancer data.



Individual level			Trial level		
LOWER	GLOBAL ODDS	UPPER	LOWER	R square	UPPER
4.1579	4.9108	5.6638	0.1564	0.4417	0.7269

Figure 10.16 shows the parameter estimates for the treatment effects for both surrogate and true endpoints that were used to estimate trial-level surrogacy.

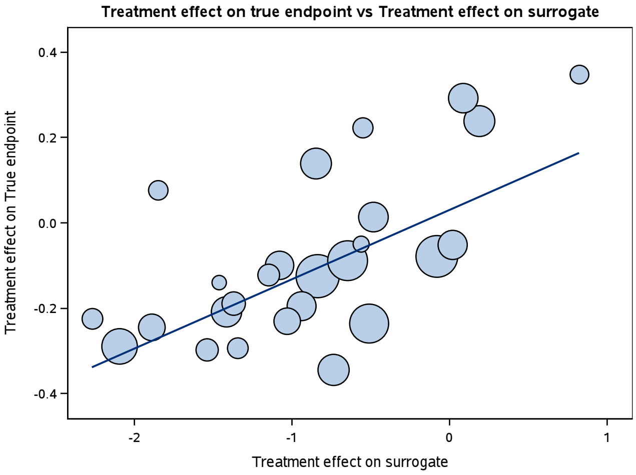


Figure 10.16: *Colorectal Cancer study. Evaluation of trial-level surrogacy. Treatment effects upon the true endpoints (log hazard ratio) versus treatment effects upon the binary surrogate (log odds ratio). Circle areas are proportional to the trial size.*

10.7 Validation Using a Joint Model for Continuous and Binary Endpoints

Similar to the previous section, we assume an underlying latent normally distributed surrogate endpoint, $\tilde{S}_{ij}$, and an observed surrogate given by (Van Sanden *et al.*, 2012):

$$S_{ij} = \begin{cases} 1 & \text{if} \quad \tilde{S}_{ij} > 0, \\ 0 & \text{if} \quad \tilde{S}_{ij} \leq 0. \end{cases} \tag{10.46}$$

A joint model is assumed for the latent surrogate variable $\tilde{S}_{ij}$ and the true endpoint T_{ij},

$$\begin{cases} T_{ij} = \mu_{Ti} + \beta_i Z_{ij} + \varepsilon_{T_{ij}}, \\ \tilde{S}_{ij} = \mu_{Si} + \alpha_i Z_{ij} + \varepsilon_{\tilde{S}_{ij}}. \end{cases} \tag{10.47}$$

Here,

$$\begin{pmatrix} \varepsilon_{T_{ij}} \\ \varepsilon_{\tilde{S}_{ij}} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{TT} & \sigma_{ST} \\ & 1 \end{pmatrix}\right]. \tag{10.48}$$

Model formulation for the observed binary outcome S_{ij} and the continuous outcome

T_{ij} are given by:

$$\begin{cases} T_{ij} = \mu_{Ti} + \beta_i Z_{ij} + \varepsilon_{Tij}, \\ logit\{P(S_{ij} = 1|Z_{ij})\} = \mu_{Si} + \alpha_i Z_{ij}. \end{cases} \tag{10.49}$$

The correlation between the measurements of the two outcomes can be modeled directly using the covariance matrix of the residuals, specified in equation (10.48) and a measure for individual-level surrogacy is given by:

$$R^2_{indiv} = \frac{\sigma^2_{ST}}{1 \times \sigma_{TT}}. \tag{10.50}$$

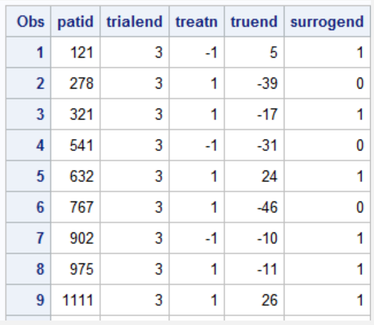
Note that for this surrogacy setting, the measure for individual-level surrogacy is the adjusted association between the true endpoint and the latent surrogate endpoint. Trial-level surrogacy is estimated using a second-stage model (see for example Section 10.3.2).

Data Structure

For the analysis presented in this section we use the Schizophrenia study for illustration (Section 9.1.4). The true endpoint is the PANSS score. For the binary surrogate we use the CGI score which was dichotomized in the following way:

$$S_{ij} = \begin{cases} 1 & \text{if CGI}_{ij} \text{ changed 3 points from baseline,} \\ 0 & \text{otherwise.} \end{cases} \tag{10.51}$$

Patient's data appears in a single line.



Obs	patid	trialend	treatn	truend	surrogend
1	121	3	-1	5	1
2	278	3	1	-39	0
3	321	3	1	-17	1
4	541	3	-1	-31	0
5	632	3	1	24	1
6	767	3	1	-46	0
7	902	3	-1	-10	1
8	975	3	1	-11	1
9	1111	3	1	26	1

The SAS Macro `%NORMALBIN`

The SAS macro `%NORMALBIN` can be used to fit the joint model specified in equation (10.49). For the Schizophrenia study the macro is called as follows:

```
%NORMALBIN(data=schizo,true=panss,surrog=cgi,trt=trtmnt,
           trial=investid,patid=patientid)
```

The macro's arguments have been defined in sections 10.2.

Data Analysis and Output

Descriptive plots produced by the macro include the distribution of the patients by treatment arms (shown in Figure 10.17 left panel) and the distribution of PANSS score per treatment arms across the levels of the surrogate endpoint in the right panel of Figure 10.17.

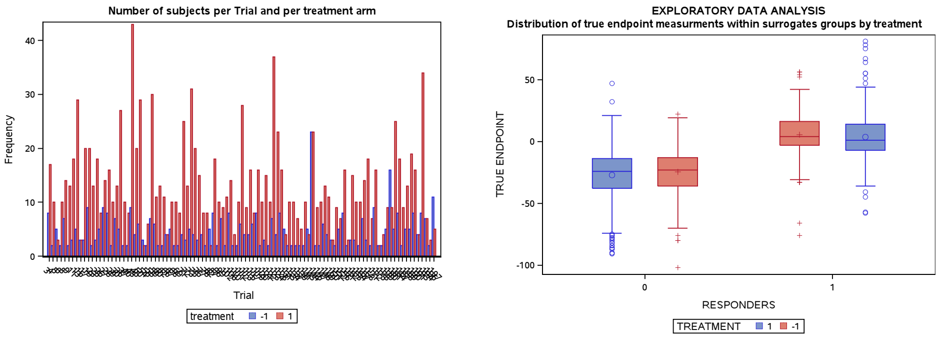
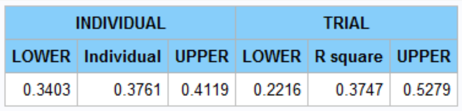


Figure 10.17: *The schizophrenia study. Exploratory plots for a normal-binary surrogacy setting. Left Panel: Patients distribution by treatment arm. Right Panel: PANSS score distribution by treatment arm across the levels of CGI scores.*

Individual- and trial-level surrogacy measures are equal to $\hat{R}^2_{ind}$ = 0.3761 (0.3403, 0.4119) and $\hat{R}^2_{trial}$= 0.3747 (0.2216, 0.5279), respectively, implying that CGI is a poor surrogate to PANSS score.



INDIVIDUAL			TRIAL		
LOWER	Individual	UPPER	LOWER	R square	UPPER
0.3403	0.3761	0.4119	0.2216	0.3747	0.5279

Figure 10.18 shows a scatter plot of the trial-specific parameter estimates for the treatment effects used in the second-stage model for the evaluation of trial-level surrogacy.

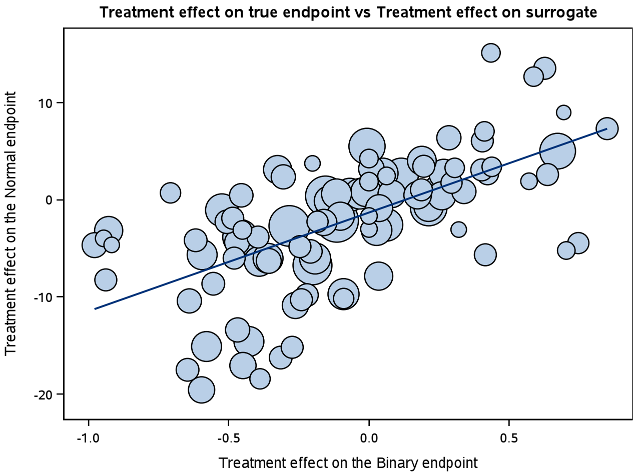


Figure 10.18: *The Second stage model for the schizophrenia study. Parameter estimates for the treatment effects upon the surrogate (log(odds ratio)) and the true endpoints.*

SAS Codes for the First-stage Model

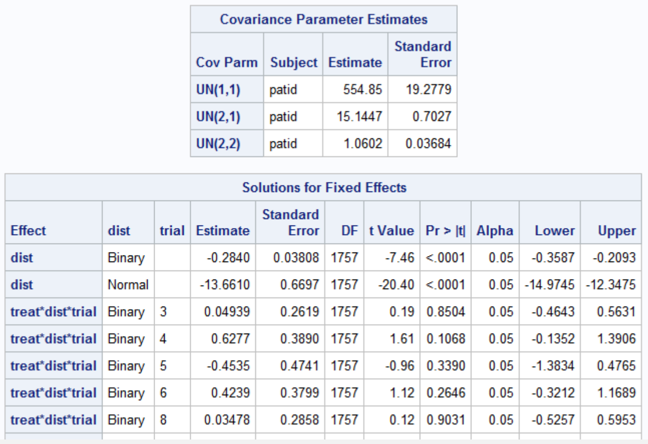
In this section, we discuss the implementation of the first-stage model using SAS procedure `GLIMMIX`. Note that the macro `%NORMALBIN` uses the same implementation, although, it is not visible for the user. The joint model in equation (10.49) is fitted using the following code.

```
proc glimmix data=norbin ;
class patientid endp trial;
model response(event=’1’) = endp endp*treat*trial /
          noint s dist=byobs(endp) link=byobs(lin) cl;
random _residual_  / subject=patientid type=un cl;
run;
```

It is assumed that there are two records per subject in the input data set, the first one corresponding to the surrogate endpoint and the second one to the true endpoint. The `response` variable contains the observed measurements on the continuous true endpoint and the binary surrogate endpoints for each patients. The statement `event=1` specifies the event category for the binary surrogate. The probability of the event category (`event=1`) is modeled.

For the mean structure, the variable `endp` allows to obtain endpoint specific intercepts (common intercepts), while the interaction term `endp*treat*trial` allows to obtain trial-specific treatment effects for both surrogate and true endpoints. The option `noint` requests that no intercept be included in the mean structure (since these are

defined by `endp`).
The argument `dist` specifies the built-in (conditional) probability distribution of the data (normal distribution for the true endpoint, and the binomial distribution for the surrogate endpoint). The statement `dist=byobs(endp)` designates a variable whose values identify the distribution to which an observation belongs while the statement `link=byobs(variable)` designates a variable whose values identify the link function associated with each endpoint (i.e. identity link for the continuous endpoint and logit link for the binary endpoint).
The statement and argument `random _residual_` specify residuals covariance structures. Finally, `subject` argument identifies subject for the analysis while `type` is used to define the covariance matrix (a 2 × 2 matrix in our case). The panel below displays the covariance matrix and parameter estimates for the fixed-effects for some trials.



Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
UN(1,1)	patid	554.85	19.2779
UN(2,1)	patid	15.1447	0.7027
UN(2,2)	patid	1.0602	0.03684

Solutions for Fixed Effects

Effect	dist	trial	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
dist	Binary		-0.2840	0.03808	1757	-7.46	<.0001	0.05	-0.3587	-0.2093
dist	Normal		-13.6610	0.6697	1757	-20.40	<.0001	0.05	-14.9745	-12.3475
treat*dist*trial	Binary	3	0.04939	0.2619	1757	0.19	0.8504	0.05	-0.4643	0.5631
treat*dist*trial	Binary	4	0.6277	0.3890	1757	1.61	0.1068	0.05	-0.1352	1.3906
treat*dist*trial	Binary	5	-0.4535	0.4741	1757	-0.96	0.3390	0.05	-1.3834	0.4765
treat*dist*trial	Binary	6	0.4239	0.3799	1757	1.12	0.2646	0.05	-0.3212	1.1689
treat*dist*trial	Binary	8	0.03478	0.2858	1757	0.12	0.9031	0.05	-0.5257	0.5953

10.8 Validation Using Joint Model for Two Binary Endpoints

To extend the methodology used for continuous endpoints to the case of binary endpoints, Renard *et al.* (2002) adopted a latent variable approach. They assumed that the observed binary variables (S_{ij}, T_{ij}) are obtained from dichotomizing unobserved continuous variables $(\tilde{S}_{ij}, \tilde{T}_{ij})$. The realized value of S_{ij} (T_{ij}) equals 1 if $\tilde{S}_{ij} > 0$ $(\tilde{T}_{ij} > 0)$, and 0 otherwise. It is assumed that the latent variables, representing the continuous underlying values of the surrogate and the true endpoints for the jth subject in the ith trial, follow a random effects model at latent scale given by:

$$\begin{cases} \tilde{S}_{ij} = \mu_S + m_{S_i} + (\alpha + a_i)Z_{ij} + \tilde{\varepsilon}_{S_{ij}}, \\ \tilde{T}_{ij} = \mu_T + m_{T_i} + (\beta + b_i)Z_{ij} + \tilde{\varepsilon}_{T_{ij}}. \end{cases} \tag{10.52}$$

Here, μ_s and μ_T are fixed intercepts, α and β are fixed treatment effects, m_{Si} and m_{Ti} are random (i.e., trial-specific) intercepts, a_i and b_i are random treatment effects and $\tilde{\varepsilon}_{S_{ij}}$ and $\tilde{\varepsilon}_{T_{ij}}$ are error terms. The random effects are zero-mean normally distributed with covariance matrix D (see equation 10.10). The error terms are assumed to follow a bivariate normal distribution with zero-mean and covariance matrix given by:

$$\sum = \begin{pmatrix} 1 & \rho_{ST} \\ & 1 \end{pmatrix}. \tag{10.53}$$

The model formulated in equation (10.52) leads to a joint probit model:

$$\begin{cases} \Phi^{-1}(P[S_{ij} = 1|Z_{ij}, m_{Si}, a_i, m_{Ti}, b_i]) = \mu_s + m_{Si} + (\alpha + a_i)Z_{ij}, \\ \Phi^{-1}(P[T_{ij} = 1|Z_{ij}, m_{Si}, a_i, m_{Ti}, b_i]) = \mu_T + m_{Ti} + (\beta + b_i)Z_{ij}. \end{cases} \tag{10.54}$$

Where, Φ denotes the standard normal cumulative distribution function. Similar to the normal-normal setting, a reduced fixed-effects model in which the random intercepts and slopes are excluded and assuming common intercepts can be formulated as:

$$\begin{cases} \Phi^{-1}(P[S_{ij} = 1|Z_{ij}]) = \mu_s + \alpha_i Z_{ij}, \\ \Phi^{-1}(P[T_{ij} = 1|Z_{ij}]) = \mu_T + \beta_i Z_{ij}. \end{cases} \tag{10.55}$$

Individual-level surrogacy can be estimated using the adjusted association based on the covariance matrix in equation (10.53). This implies that for the binary-binary setting this level of surrogacy should be interpreted at the scale of the linear predictors. Trial-specific treatment effects upon the true and the surrogate endpoints can be used in the second-stage to fit a linear regression model of the form:

$$\hat{\beta}_i = \gamma_0 + \gamma_1 \hat{\alpha}_i + \varepsilon_i. \tag{10.56}$$

Similar to previous sections, the trials sizes are used as weights in order to account for the variability due to difference in trial sizes. The trial-level surrogacy measure is equal to the coefficient of determination from model in equation (10.56).

Data Structure

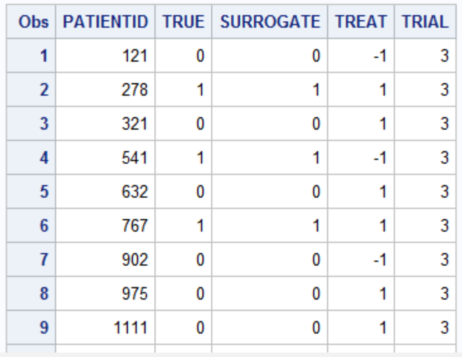
The schizophrenia study was used for illustration in this setting (see Section 9.1.4). The true endpoint is the PANSS score and the surrogate endpoint is the CGI score.

The binary endpoints PANSS/CGI reflect the presence or absence of clinically relevant change in schizophrenic symptomatology. Clinically relevant change is defined as a reduction of 20% or more in the PANSS scores, i.e, 20% reduction in post-treatment scores relative to baseline scores, or a change of 3 points in the original CGI scale (Kane *et al.*, 1988; Leucht *et al.*, 2005). Hence, the true and surrogate binary endpoints are defined, respectively, as:

$$T_{ij} = \begin{cases} 1 & \text{if PANSS}_{ij} \text{ reduced at least 20\% from baseline,} \\ 0 & \text{otherwise,} \end{cases}$$

$$S_{ij} = \begin{cases} 1 & \text{if CGI}_{ij} \text{ changed of 3 points from baseline,} \\ 0 & \text{otherwise.} \end{cases}$$

A partial printout of the data is given below.



Obs	PATIENTID	TRUE	SURROGATE	TREAT	TRIAL
1	121	0	0	-1	3
2	278	1	1	1	3
3	321	0	0	1	3
4	541	1	1	-1	3
5	632	0	0	1	3
6	767	1	1	1	3
7	902	0	0	-1	3
8	975	0	0	1	3
9	1111	0	0	1	3

The SAS Macro `%BINBIN`

The reduced fixed-effects model formulated in equation (10.55) can be fitted using the SAS macro `%BINBIN` that has the following general call:

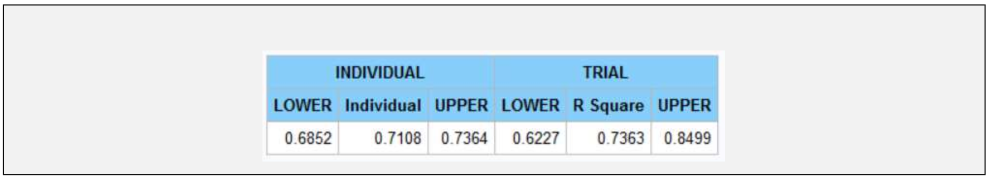
```
%BINBIN (data=schizo,true=cgi,surrog=panss,trt=trtmnt,
         trial=investid,patid=patientid,looa=1)
```

The macro's arguments have been defined in sections 10.2.

Data Analysis and Output

Parameter estimates for individual- and trial-level surrogacy are equal to $\hat{R}^2_{ind}$= 0.7108 (0,6852, 0.7364) and $\hat{R}^2_{trial}$= 0.7363 (0.6227, 0.8499), respectively. This indicates that

CGI is a surrogate of moderate value for PANSS at both individual- and trial-level surrogacy.



INDIVIDUAL			TRIAL		
LOWER	Individual	UPPER	LOWER	R Square	UPPER
0.6852	0.7108	0.7364	0.6227	0.7363	0.8499

Figure 10.19 shows the estimated treatment effects upon both endpoints with the fitted regression line.

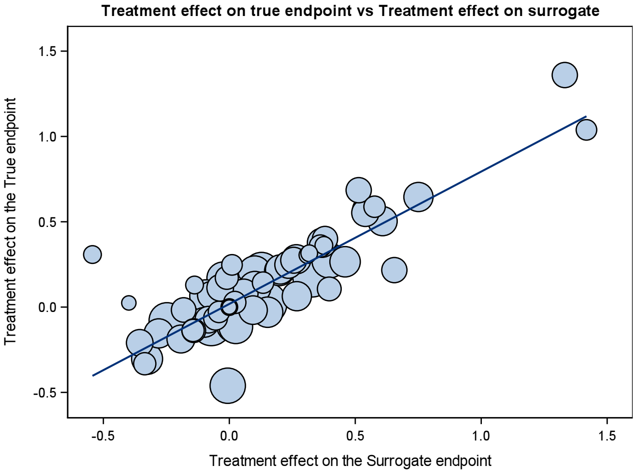


Figure 10.19: *Schizophrenia study. Estimation of trial-level surrogacy.*

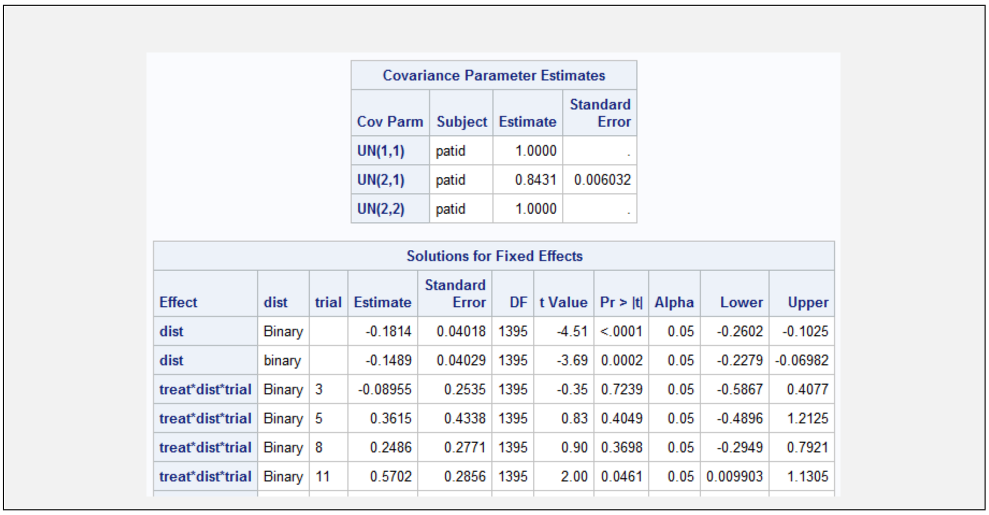
SAS Codes for the First-stage Model

The SAS code to fit reduced fixed-effects model formulated in equation (10.52) can be written as follows:

```
proc glimmix data=binbin ;
class patientid endp trial;
model response(event='1') = endp endp*treat*trial /
          noint s dist=byobs(endp) link=byobs(lin) cl;
random _residual_  / subject=patientid type=un cl;
run;
```

It is assumed that there are two records per subject in the input dataset, the first one corresponding to the surrogate endpoint and the second one to the true endpoint.

The `response` variable contains the observed categories on both endpoints for each patients. The statement `event=1` specifies the event category for both endpoints. The statement `dist=byobs(endp)` defines the distribution for each endpoint and the link function to be used is specified by link=byobs(lin).
The statement `random _residual_` specifies the residuals covariance structures from which individual-level surrogacy is derived. The output from the above codes is shown in the panel below for some trials.



Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
UN(1,1)	patid	1.0000	.
UN(2,1)	patid	0.8431	0.006032
UN(2,2)	patid	1.0000	.

Solutions for Fixed Effects

Effect	dist	trial	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
dist	Binary		-0.1814	0.04018	1395	-4.51	<.0001	0.05	-0.2602	-0.1025
dist	binary		-0.1489	0.04029	1395	-3.69	0.0002	0.05	-0.2279	-0.06982
treat*dist*trial	Binary	3	-0.08955	0.2535	1395	-0.35	0.7239	0.05	-0.5867	0.4077
treat*dist*trial	Binary	5	0.3615	0.4338	1395	0.83	0.4049	0.05	-0.4896	1.2125
treat*dist*trial	Binary	8	0.2486	0.2771	1395	0.90	0.3698	0.05	-0.2949	0.7921
treat*dist*trial	Binary	11	0.5702	0.2856	1395	2.00	0.0461	0.05	0.009903	1.1305

For the schizophrenia study, the estimated covariance matrix is given by:

$$\hat{\Sigma} = \begin{pmatrix} 1.0000 & 0.8431 \\ & 1.0000 \end{pmatrix}. \tag{10.57}$$

Hence, individual-level surrogacy is equal to $\hat{R}^2_{indiv} = 0.8431^2 = 0.7108$.

10.9 Validation Using the Information-theoretic Approach

10.9.1 Individual-level Surrogacy

In this Section, the information-theoretic approach for the evaluation of surrogate endpoints proposed by Alonso and Molenberghs (2007) is briefly discussed. This approach allows to evaluate surrogacy at individual- and trial-level in a general surrogacy setting. We concisely present the setting and illustrate the use of a SAS macro for a normal-normal setting. We consider a multi-trial setting and the following models for the true endpoint:

$$\begin{cases} M_0 : g_T\{E(T_{ij}|Z_{ij})\} = \mu_{T_i} + \beta_i Z_{ij}, \\ M_1 : g_T\{E(T_{ij}|Z_{ij}, S_{ij})\} = \theta_{0_i} + \theta_{1i} Z_{ij} + \theta_{2i} S_{ij}. \end{cases} \tag{10.58}$$

Let, G_i^2 be the likelihood ratio test statistic to compare models M_0 and M_1 in equation (10.58) within the ith trial. The association between both endpoints is quantified using the *likelihood reduction factor* (LRF) given by:

$$LRF = 1 - \frac{1}{N} \sum_i \exp\left(-\frac{G_i^2}{n_i}\right), \tag{10.59}$$

here, N is the total number of the trials, n_i is trial-specific sample size. As pointed out by Alonso and Molenberghs (2007), the LRF ranges between 0 and 1. The case with LRF=0 indicates that the surrogate and the true endpoint are independent in each trial.

10.9.2 Trial-level Surrogacy

Trial-level surrogacy can be estimated using a two-stage approach. At the first-stage, the following models are formulated for the two endpoints.

$$\begin{cases} g_T\{E(S_{ij}|Z_{ij})\} = \mu_{S_i} + \alpha_i Z_{ij}, \\ g_T\{E(T_{ij}|Z_{ij})\} = \mu_{T_i} + \beta_i Z_{ij}. \end{cases} \tag{10.60}$$

Here, μ_{T_i} and μ_{S_i} are trial-specific intercepts, α_i and β_i are trial-specific treatment effects. Note that the models can be fitted with common intercepts (i.e., reduced fixed-effects models). At the second-stage, the parameter estimates obtained from equation (10.60) are used to fit two linear regression models given by:

$$\begin{cases} M_0 : \hat{\beta}_i = \gamma_0 + \varepsilon_{0i}, \\ M_1 : \hat{\beta}_i = \theta_0 + \theta_1 \hat{\mu}_{si} + \theta_2 \hat{\alpha}_i + \varepsilon_{1i}. \end{cases} \tag{10.61}$$

The error terms ε_{0i} and ε_{1i} are normally distributed with zero-mean and constant variances σ_0^2 and σ_1^2, respectively. When the reduced fixed-effects models are used in equation (10.60), $\hat{\mu}_{S_i}$ is dropped from equation (10.61). The trial-level surrogacy is estimated by:

$$R_{ht}^2 = 1 - \exp\left(-\frac{G^2}{N}\right). \tag{10.62}$$

Where, G^2 is the likelihood ratio test statistic comparing the two models in equation (10.61).

10.9.3 Evaluation of Surrogacy for Two Continuous Endpoints

We use the ARMD study to illustrate the analysis for the normal-normal surrogacy setting using the information-theoretic approach. As before, the true endpoint is visual acuity 52 weeks after the start of the treatment (`Diff52`) and the surrogate endpoint is the visual acuity 24 weeks after the start of the treatment (`Diff24`).

The SAS Macro `%NORMNORMINFO`

The models for two normally distributed endpoints can be fitted using the SAS macro `%NORMNORMINFO`. The macro fit the models formulated in equations (10.58) and (10.61) in order to estimate both individual- and trial-level surrogacy.

```
%NORMNORMINFO(data=ARMD,true=Diff54,surrog=Diff24,
treat=treat,trial=center,patid=patientid,weighted=1,
model="full",boot=10)
```

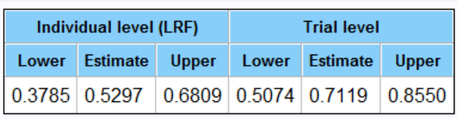
Arguments specific for the `%NORMNORMINFO` are:

- `model:` the model used in equations (10.60) and (10.61) (”reduced” or ”full”).
- `boot:` is the number of bootstrap samples used to construct the confidence intervals for the parameter estimates of the surrogacy measures.

Other arguments have been defined in sections 10.2.

Data Analysis and Output

The `%NORMNORMINFO` macro produces exploratory plots displaying the distribution of the patients per trial (see, for example, left panel in Figure 10.2). Parameter estimates for the LRF and trial-level surrogacy are shown in the panel below.



Individual level (LRF)			Trial level		
Lower	Estimate	Upper	Lower	Estimate	Upper
0.3785	0.5297	0.6809	0.5074	0.7119	0.8550

The estimated individual- and trial-level surrogacy are equal to $\hat{R}^2_h = 0.5297$ (0.3785, 0.6809), and $\hat{R}^2_{ht} = 0.7119$ (0.5074, 0.8550), respectively. Both surrogacy measures

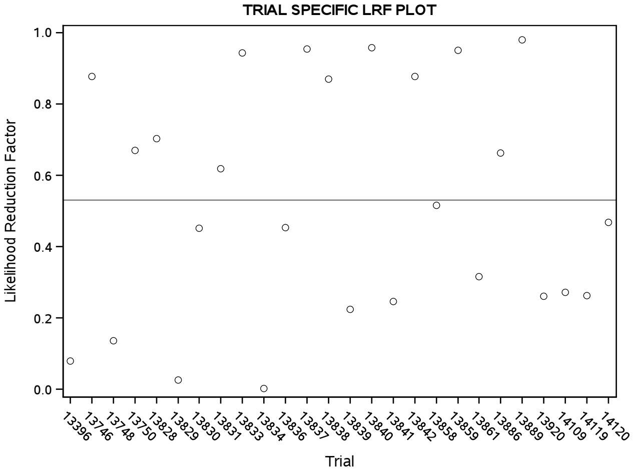


Figure 10.20: Trial-specific likelihood reduction factor (the horizontal line: the overall LRF).

indicate that the visual acuity 24 weeks after the start of the treatment is a surrogate of moderate value for the visual acuity 52 weeks after the start of the treatment.

As a sensitivity analysis, trial specific LRF is presented in Figure 10.20.

Note that the macro `%NORMNORMINFO` uses the same model formulation as the R function `FixedContContIT` (i.e, fixed effects models for two continuous endpoints).

10.9.4 Other Surrogacy Settings

SAS macros that were developed for the evaluation of surrogacy using the information-theoretic approach are presented in Table 10.2. Note that, as discussed in previous sections in the chapter, data structure depends on the surrogacy setting.

Surrogacy Setting	**SAS Macro**	**Data structure**
Normal-Normal	`%NORMNORMINFO`	10.3.1
Normal-Binary	`%NORMBININFO`	10.7
Survival-Survival	`%SURVSURVINFO`	10.4
Survival-Binary	`%SURVBININFO`	10.6
Binary-Binary	`%BINBININFO`	10.8

Table 10.2: *SAS macros available for analyses using the information-theoretic approach.*

Chapter 11

Surrogacy in Cloud Computing

11.1 The `Surrogate` Shiny App

Shiny is an R package (available on CRAN) developed by RStudio which allows to create web-based applications from R-code. The surrogate Shiny App was developed as an online shiny application for the evaluation of surrogate endpoints in randomized clinical trials and has the same capacity, in terms of the methods implemented in the App, as the `surrogate` R package.
The surrogate Shiny App can be used on a local computer or online using the shiny cloud platform. Other cloud platforms, such as Amazon Web Service or google cloud platform can be used as well. In contrast with the `surrogate` R package, the user does not need to install R in order to conduct the analysis. The surrogate Shiny App is a graphical user interface (GUI) and the user is not exposed to the R code behind the analysis.
The surrogate Shiny App can be found on the Shiny Cloud at:

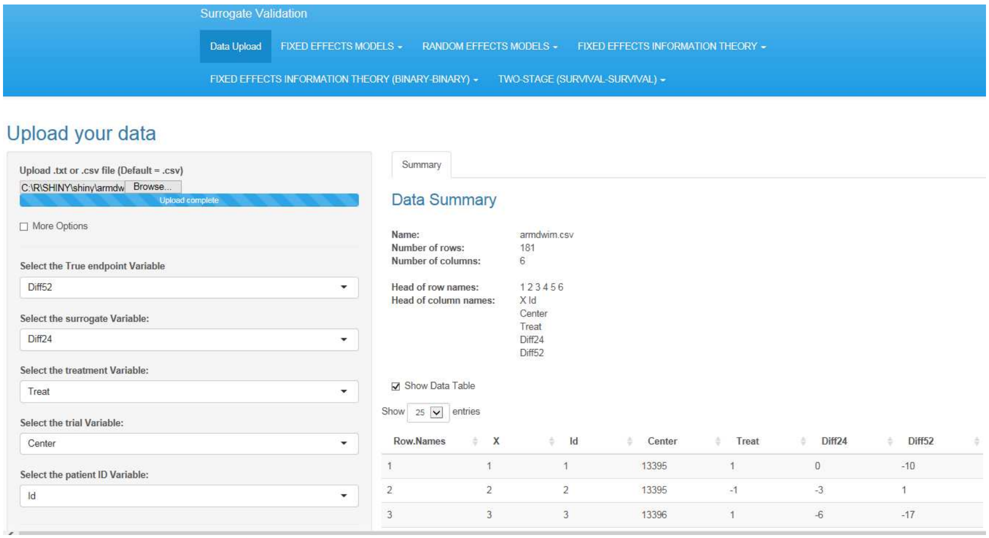
```
https://uhasselt.shinyapps.io/surrogate
```

In addition, the App is available as a stand alone version in a `SurrShiny.zip` file that can be downloaded from:

```
http://ibiostat.be/online-resources
```

In this chapter we briefly illustrate the capacity of surrogate Shiny App for selected methods that were discussed in Chapter 10. For each method, we present the GUI screen that can be used to conduct the analysis and the corresponding R code from the `surrogate` package to perform an identical analysis. The code is presented only for clarity and it is not needed for the surrogate Shiny App. The capacity of the surrogate Shiny App is illustrated using case studies for three surrogacy settings: two continuous endpoints (Section 11.2 and 11.4.2), two survival endpoints (Section 11.3) and two binary endpoints (Section 11.4.3).

The first-step of the analysis requires to upload the data to the App. Figure 11.1 shows the data loading screen for the ARMD data. Similar to Chapter 10, we need to specify the true and surrogate endpoints (`Diff52` and `Diff24`, respectively), the treatment (`Treat`), the unit of analysis for which R^2_{trial} will be calculated (`center`), and the patient's identification number (`Id`).


Surrogate Validation
Data Upload
FIXED EFFECTS MODELS
RANDOM EFFECTS MODELS
FIXED EFFECTS INFORMATION THEORY
FIXED EFFECTS INFORMATION THEORY (BINARY-BINARY)
TWO-STAGE (SURVIVAL-SURVIVAL)
Browse...
Upload complete
Summary

Upload your data

Upload .txt or .csv file (Default = .csv)
C:\R\SHINY\shiny\armdw

□ More Options

Select the True endpoint Variable: Diff52

Select the surrogate Variable: Diff24

Select the treatment Variable: Treat

Select the trial Variable: Center

Select the patient ID Variable: Id

Data Summary

Name:	armdwim.csv
Number of rows:	181
Number of columns:	6
Head of row names:	1 2 3 4 5 6
Head of column names:	X Id Center Treat Diff24 Diff52

☑ Show Data Table

Show 25 entries

Row.Names	X	Id	Center	Treat	Diff24	Diff52
1	1	1	13395	1	0	-10
2	2	2	13395	-1	-3	1
3	3	3	13396	1	-6	-17

Figure 11.1: *The ARMD data and variables for the analysis are specified in the left panel. A short summary of the data and a partial printout are shown in the right panel.*

11.2 Two Continuous Endpoints: The Reduced Fixed-effects Model

In Chapter 10, we discussed the reduced fixed-effects model for two continuous endpoints. The model can be formulated as:

$$\begin{cases} S_{ij} = \mu_S + \alpha_i Z_{ij} + \varepsilon_{Sij}, \\ T_{ij} = \mu_T + \beta_i Z_{ij} + \varepsilon_{Tij}. \end{cases} \tag{11.1}$$

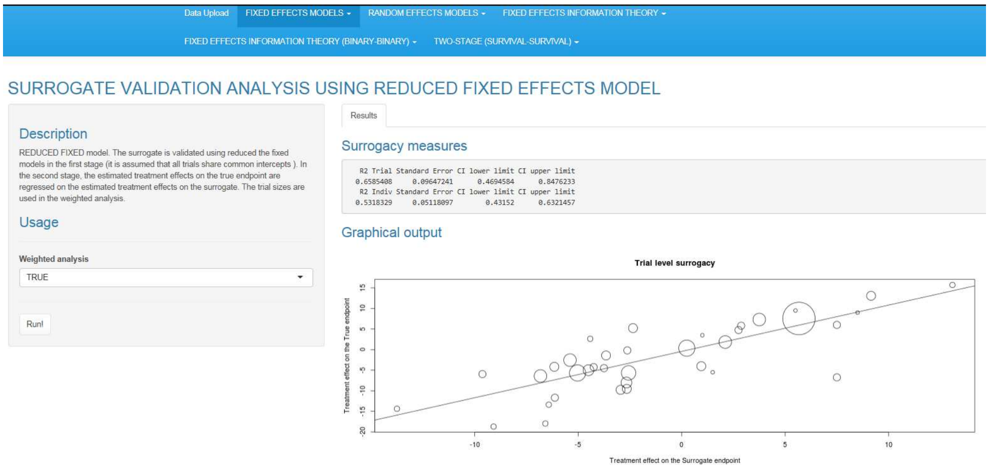
For the reduced fixed-effects model, as shown in Chapter 10, trial-level surrogacy is assessed using the coefficient of determination obtained by fitting a linear regression model of the form:

$$\widehat{\beta}_i = \lambda_0 + \lambda_1 \widehat{\alpha}_i + \varepsilon_i, \tag{11.2}$$

where, $\widehat{\beta}_i$ and $\widehat{\alpha}_i$ are the trial-specific estimated treatment effects upon T_{ij} and S_{ij}, respectively. The error terms, ε_i, are normally distributed with mean zero and a constant variance. Individual-level surrogacy is assessed by the squared correlation between S and T after adjusting for trial-specific treatment effects and is given by:

$$\hat{R}^2_{indiv} = \frac{\sigma^2_{ST}}{\sigma_{SS}\sigma_{TT}}. \tag{11.3}$$

Once the variables specification is complete (see the left panel in Figure 11.1), we can choose the model to be fitted using the command bar in the upper part in Figure 11.1. The surrogate Shiny App produces a default output shown in Figure 11.2. For the ARMD study, $\hat{R}^2_{indiv} = 0.5318$ (0.4315, 0.6321) and $\hat{R}^2_{trial} = 0.6585$ (0.4695, 0.8476). In case that other statistics are of interest we can use the R package `surrogate` to produce them.


Data Upload
FIXED EFFECTS MODELS ▾
RANDOM EFFECTS MODELS ▾
FIXED EFFECTS INFORMATION THEORY ▾
FIXED EFFECTS INFORMATION THEORY (BINARY-BINARY) ▾
TWO-STAGE (SURVIVAL-SURVIVAL) ▾
Results
Run!
Trial level surrogacy
Treatment effect on the True endpoint
-20
-15
-10
-5
0
5
10
15
-10
-5
0
5
10
Treatment effect on the Surrogate endpoint

SURROGATE VALIDATION ANALYSIS USING REDUCED FIXED EFFECTS MODEL

Description

REDUCED FIXED model. The surrogate is validated using reduced the fixed models in the first stage (it is assumed that all trials share common intercepts). In the second stage, the estimated treatment effects on the true endpoint are regressed on the estimated treatment effects on the surrogate. The trial sizes are used in the weighted analysis.

Usage

Weighted analysis: TRUE

Surrogacy measures

```
 R2 Trial Standard Error CI lower limit CI upper limit
0.6585408     0.09647241      0.4694584      0.8476233
 R2 Indiv Standard Error CI lower limit CI upper limit
0.5318329     0.05118097        0.43152      0.6321457
```

Graphical output

Figure 11.2: *The ARMD study. Default output for the reduced fixed-effects model.*

The reduced fixed-effects model specified in equation (11.1) and in the surrogate Shiny App in Figure 11.1 are identical to the model fitted using the function `BifixedContCont` below:

```
Sur<-BifixedContCont(Dataset=ARMD, Surr=Diff24, True=Diff52,
                     Treat=Treat, Trial.ID=Center,
                     Pat.ID=Id,Model="Reduced", Weighted=TRUE)
```

11.3 Two Time-to-event Endpoints: A Two-stage Approach

The analysis for the surrogacy setting of two time-to-event endpoints was discussed in Chapter 10, where a joint modeling and two-stage approach were used to estimate individual- and trial-level surrogacy. The two-stage approach is implemented in the surrogate Shiny App. We use the Ovarian study for illustration (see Section 9.1.2). The specification of the data is shown in the upper panel of Figure 11.3. In the same screen, we select the tab *Two-stage (Survival-Survival)* in order to perform the analysis, to select the censoring indicators and to choose the model used at the second stage (weighted or unweighed). The output is presented in the lower panel of Figure 11.3. The estimated trial-level surrogacy is equal to $\hat{R}^2_{trial} = 0.9184$ (0.8674, 0.9695).

For the Ovarian data, the two-stage model discussed above is identical to the model specified in the function `TwoStageSurvSurv`.

```
Sur<-TwoStageSurvSurv(Dataset=ovarian, Surr=pfs,
SurrCens=PfsInd,True=surv,TrueCens=SurvInd,
Treat=Treat,Trial.ID=Center)
```

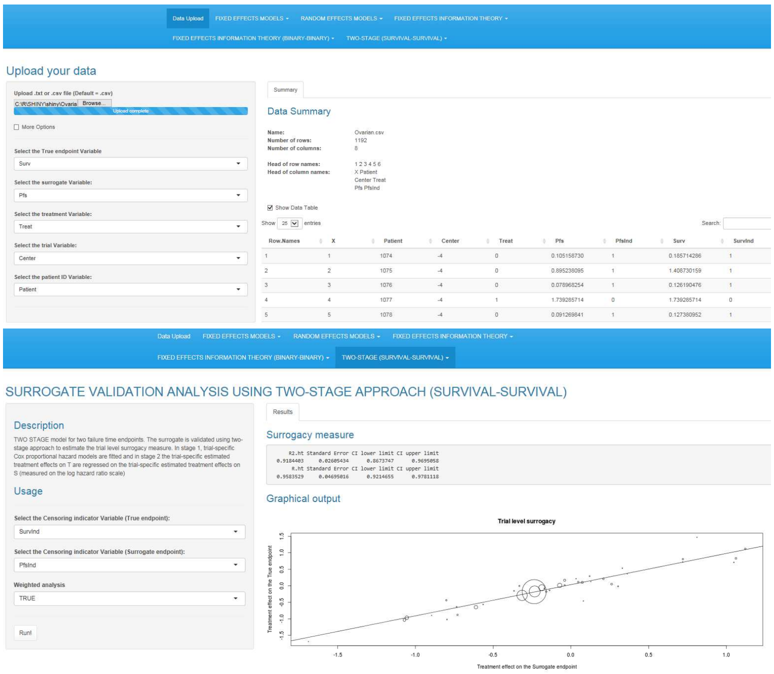
11.4 Information-theoretic Approach

11.4.1 Individual- and Trial-level Surrogacy

The information theoretic approach was discussed in Chapter 10. For a multi-trial setting, we consider two models for the true endpoint T_{ij},

$$\begin{cases} M_0 : g\{E(T_{ij}|Z_{ij})\} = \mu_{T_i} + \beta_i Z_{ij}, \\ M_1 : g\{E(T_{ij}|Z_{ij}, S_{ij})\} = \delta_{0_i} + \delta_{1i} Z_{ij} + \delta_{2i} S_{ij}. \end{cases} \tag{11.4}$$

Here, g is an appropriate link function. For the remainder of this section we briefly discuss the surrogacy measures implemented in the surrogate Shiny App. For an elaborate discussion about the modeling approach and the derivation of the surrogacy


Data Upload
FIXED EFFECTS MODELS
RANDOM EFFECTS MODELS
FIXED EFFECTS INFORMATION THEORY
FIXED EFFECTS INFORMATION THEORY (BINARY-BINARY)
TWO-STAGE (SURVIVAL-SURVIVAL)
Browse...
Upload complete
Summary
Search:
Results
Run!
Trial level surrogacy
Treatment effect on the True endpoint
Treatment effect on the Surrogate endpoint
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

Upload your data

Upload .txt or .csv file (Default = .csv)

C:\R\SHINY\shiny\Ovaria

□ More Options

Select the True endpoint Variable

Surv

Select the surrogate Variable:

Pfs

Select the treatment Variable:

Treat

Select the trial Variable:

Center

Select the patient ID Variable:

Patient

Data Summary

Name:	Ovarian.csv
Number of rows:	1192
Number of columns:	8
Head of row names:	1 2 3 4 5 6
Head of column names:	X Patient Center Treat Pfs PfsInd

☑ Show Data Table

Show 25 entries

Row.Names	X	Patient	Center	Treat	Pfs	PfsInd	Surv	SurvInd
1	1	1074	-4	0	0.105158730	1	0.185714286	1
2	2	1075	-4	0	0.895238095	1	1.408730159	1
3	3	1076	-4	0	0.078968254	1	0.126190476	1
4	4	1077	-4	1	1.739285714	0	1.739285714	0
5	5	1078	-4	0	0.091269841	1	0.127380952	1

SURROGATE VALIDATION ANALYSIS USING TWO-STAGE APPROACH (SURVIVAL-SURVIVAL)

Description

TWO STAGE model for two failure time endpoints. The surrogate is validated using two-stage approach to estimate the trial level surrogacy measure. In stage 1, trial-specific Cox proportional hazard models are fitted and in stage 2 the trial-specific estimated treatment effects on T are regressed on the trial-specific estimated treatment effects on S (measured on the log hazard ratio scale)

Usage

Select the Censoring indicator Variable (True endpoint):

SurvInd

Select the Censoring indicator Variable (Surrogate endpoint):

PfsInd

Weighted analysis

TRUE

Surrogacy measure

```
     R2.ht Standard Error CI lower limit CI upper limit
0.9184403     0.02605434      0.8673747      0.9695058
      R.ht Standard Error CI lower limit CI upper limit
0.9583529     0.04695016      0.9214655      0.9781118
```

Graphical output

Figure 11.3: *The Ovarian study. Evaluation of trial-level surrogacy (using a two-stage model) for a surrogacy setting with two time-to-event endpoints. Upper Panel: Loading the data for the Ovarian study. True endpoint: overall survival. Surrogate endpoint: progression free survival. Lower Panel: Estimation of trial-level surrogacy.*

measures we refer to Chapter 10. An information theoretic measure for individual-level surrogacy for a multi-trial setting is given by:

$$R_h^2 = 1 - \frac{1}{N}\sum_{i=1}^{N} exp\left(\frac{L_{1_i} - L_{0_i}}{n_i}\right),$$

where, L_{k_i} is the -2 log likelihood of model M_k, $k = 0, 1$ defined in equation (11.4) and n_i is the sample size of the ith trial. For a single trial setting (i.e., $N = 1$) the surrogacy measure is reduced to:

$$R_h^2 = 1 - exp\left(\frac{L_1 - L_0}{n}\right).$$

To estimate the trial-level surrogacy measure, the following models are fitted:

$$\begin{cases} g\{E(S_{ij}|Z_{ij})\} = \mu_{S_i} + \alpha_i Z_{ij}, \\ g\{E(T_{ij}|Z_{ij})\} = \mu_{T_i} + \beta_i Z_{ij}. \end{cases} \tag{11.5}$$

At the second stage, the trial-specific parameter estimates from the model defined in equation (11.5) are used to fit the following linear regression models:

$$\begin{cases} M_0 : \hat{\beta}_i = \gamma_0 + \varepsilon_{0i}, \\ M_1 : \hat{\beta}_i = \gamma_0 + \gamma_1 \hat{\mu}_{S_i} + \gamma_2 \hat{\alpha}_i + \varepsilon_{1i}. \end{cases} \tag{11.6}$$

A trial-level surrogacy measure is given by:

$$R^2_{ht} = 1 - exp\left(-\frac{G^2}{N}\right), \tag{11.7}$$

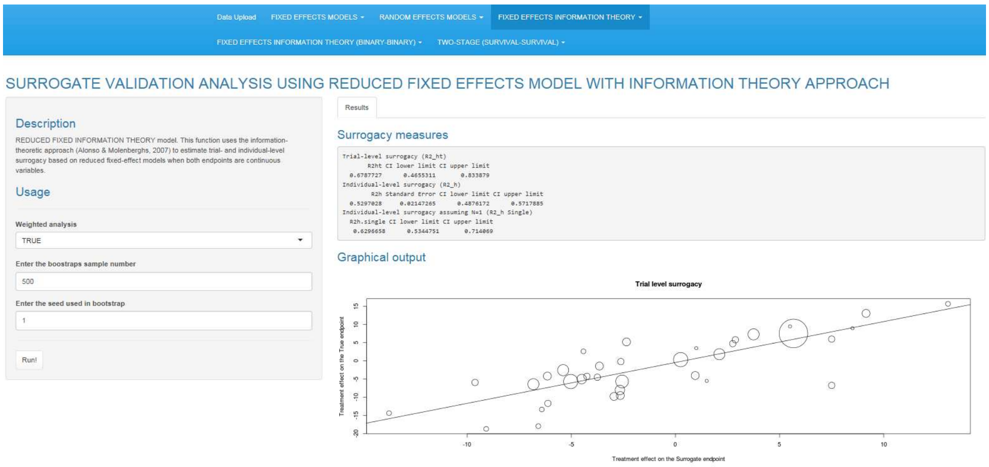
where, G^2 is the likelihood ratio test statistic comparing the models M_0 and M_1 in equation (11.6) and N is the number of trials.

11.4.2 Information-theoretic Approach for Two Continuous Endpoints

The information theoretic approach for two continuous endpoints was applied to the ARMD data. The function `FixedContContIT` was used in order to estimate both individual- and trial-level surrogacy in the following way:

```
Sur<-FixedContContIT(Dataset=ARMD, Surr=Diff24,
                     True=Diff52, Treat=Treat,
                     Trial.ID=Center, Weighted=TRUE,
                     Pat.ID=Id, Model="Reduced",
                     Number.Bootstraps=500,Seed=1)
```

An identical model can be fitted using the surrogate Shiny App. Figure 11.1 shows the specification of the variables for the ARMD data in the data loading screen of the surrogate App. Note that this specification is identical to the one used in the previous section since both examples use the same data. Figure 11.4 presents the output. The number of bootstrap samples (`Number.Bootstraps=500`) and the seed (`Seed=1`) are specified in the left side of Figure 11.4. As shown in Chapter 10, for the ARMD data, trial-level surrogacy and individual-level surrogacy measures are equal to $\hat{R}^2_{ht} = 0.6788$ (0.4655, 0.8338) and $\hat{R}^2_h = 0.5297$ (0.4876, 0.5718), respectively.


Data Upload
FIXED EFFECTS MODELS ▾
RANDOM EFFECTS MODELS ▾
FIXED EFFECTS INFORMATION THEORY ▾
FIXED EFFECTS INFORMATION THEORY (BINARY-BINARY) ▾
TWO-STAGE (SURVIVAL-SURVIVAL) ▾
Results
Run!
Trial level surrogacy
Treatment effect on the True endpoint
Treatment effect on the Surrogate endpoint
-20
-15
-10
-5
0
5
10
15
-10
-5
0
5
10

SURROGATE VALIDATION ANALYSIS USING REDUCED FIXED EFFECTS MODEL WITH INFORMATION THEORY APPROACH

Description

REDUCED FIXED INFORMATION THEORY model. This function uses the information-theoretic approach (Alonso & Molenberghs, 2007) to estimate trial- and individual-level surrogacy based on reduced fixed-effect models when both endpoints are continuous variables.

Usage

Weighted analysis: TRUE

Enter the boostraps sample number: 500

Enter the seed used in bootstrap: 1

Surrogacy measures

```
Trial-level surrogacy (R2_ht)
      R2ht CI lower limit CI upper limit
  0.6787727      0.4655311       0.833879
Individual-level surrogacy (R2_h)
       R2h Standard Error CI lower limit CI upper limit
  0.5297028     0.02147265      0.4876172      0.5717885
Individual-level surrogacy assuming N=1 (R2_h Single)
  R2h.single CI lower limit CI upper limit
   0.6296658      0.5344751       0.714069
```

Graphical output

Figure 11.4: *Analysis if the ARMD study using the information-theoretic approach for two continuous endpoints.*

11.4.3 Information-theoretic Approach for Two Binary Endpoints

For the Schizophrenia study, presented in Chapter 10, the two binary endpoints were defined as:

$$T_{ij} = \begin{cases} 1 & \text{PANSS reduction (at least -20\%),} \\ 0 & \text{otherwise,} \end{cases}$$

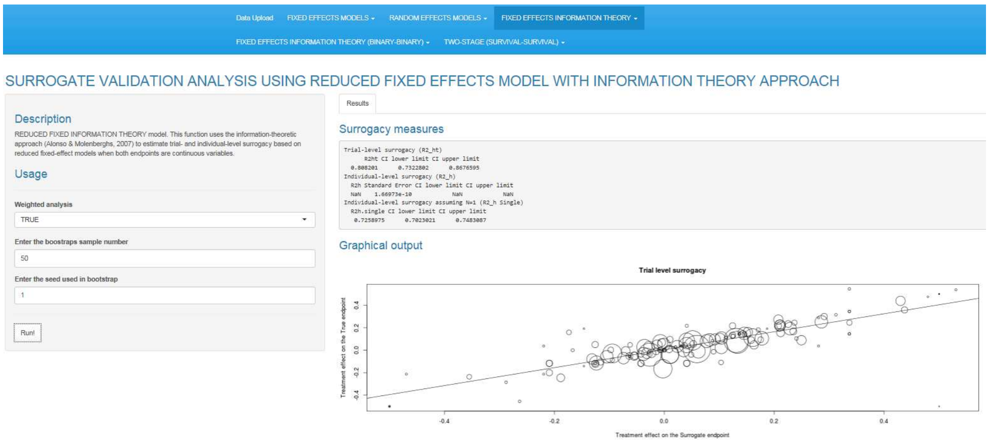
$$S_{ij} = \begin{cases} 1 & \text{CGI change of 3 points,} \\ 0 & \text{otherwise.} \end{cases}$$

In R, using the `surrogate` package, for a multi-trial setting with two binary endpoints, the function `FixedBinBintIT` can be used in order to estimate both individual- and trial-level surrogacy measures. For the Schizophrenia study, the function is called in the following way:

```
Sur<-FixedBinBinIT(Dataset=Schizo, Surr=Panss_Bin,
                   True=CGI_Bin, Treat=Treat,
                   Trial.ID=InvestId, Weighted=TRUE,
                   Pat.ID=Id, Model="Reduced",
                   Number.Bootstraps=500,Seed=1)
```

With the surrogate Shiny App, the following specifications should be used in the data loading screen in Figure 11.1: the true (`CGI_bin`) and the surrogate endpoints

(`PANSS_bin`), the treatment variable (`Treat`), the unit for which R^2_{ht} will be calculated (`InvestId`), and the patient's identification number (`Id`). In the same screen, the tab *Fixed effects information theory (Binary-Binary)* is selected in order to perform the analysis. The number of bootstrap samples (`Number.Bootstraps=500`) and the seed (`Seed=1`) are specified in the left side of Figure 11.5. For the Schizophrenia study, trial- and individual-level surrogacy measures are equal to $\hat{R}^2_{ht} = 0.8213$ (0.7469, 0.87864) and $\hat{R}^2_h = 0.3305$ (0.2992, 0.3623), respectively.


Data Upload
FIXED EFFECTS MODELS
RANDOM EFFECTS MODELS
FIXED EFFECTS INFORMATION THEORY
FIXED EFFECTS INFORMATION THEORY (BINARY-BINARY)
TWO-STAGE (SURVIVAL-SURVIVAL)
Results
Run!
Trial level surrogacy
Treatment effect on the True endpoint
Treatment effect on the Surrogate endpoint
-0.4
-0.2
0.0
0.2
0.4

SURROGATE VALIDATION ANALYSIS USING REDUCED FIXED EFFECTS MODEL WITH INFORMATION THEORY APPROACH

Description

REDUCED FIXED INFORMATION THEORY model. This function uses the information-theoretic approach (Alonso & Molenberghs, 2007) to estimate trial- and individual-level surrogacy based on reduced fixed-effect models when both endpoints are continuous variables.

Usage

Weighted analysis: TRUE

Enter the boostraps sample number: 50

Enter the seed used in bootstrap: 1

Surrogacy measures

```
Trial-level surrogacy (R2_ht)
     R2ht CI lower limit CI upper limit
  0.808201      0.7322802      0.8676595
Individual-level surrogacy (R2_h)
  R2h Standard Error CI lower limit CI upper limit
  NaN    1.66973e-10           NaN            NaN
Individual-level surrogacy assuming N=1 (R2_h Single)
  R2h.single CI lower limit CI upper limit
   0.7258975      0.7023021      0.7483087
```

Graphical output

Figure 11.5: *Analysis of the Schizophrenia study using the information-theoretic approach for two binary endpoints.*

Chapter 12

Discussion and Future research

12.1 Part I: Development of Metabolic Biomarker for Cancer

In the first part of the dissertation, we discussed the use of metabolic data for the detection of lung and breast cancer. As stated before, screening for cancer (lung or breast) at an early stage before a patient develops clinical symptoms and when the treatment is most effective should benefit the patient by increasing his/her quality of life and life expectancy (Bourzac, 2014; Shlomi *et al.*, 2014; Wood *et al.*, 2012). An appropriate screening test should be cost-effective. According to Field *et al.* (2013a,b) the benefit-risk balance is maximized when high-risk target population is selected for screening. In Chapter 3 and 4, we applied several classification methods to construct metabolic signatures for early detection of cancer. PLS-DA became a standard method for classification based on metabolic data mainly due to the fact that friendly-user softwares are available for this particular method. Other methods for classification are implemented in R software but the application requires a knowledge in R programing. Therefore, development of a publicly available software, which include more classification methods but does not require a deep knowledge of R programing will benefit the scientific community who are interested in the development of metabolic signatures. Our intention is to develop an R shiny app which will offer a friendly-user software tool.

In chapter 5, we studied the added predictive value of metabolic data, in addition to the clinical variables, to predict the lung cancer status of a patient. Current risk models are of the form:

$$log[P(Y = 1|Z)] = \gamma_0 + \sum_{j=1}^{p} \gamma_j z_{ij},$$

where, **Z** is a matrix containing clinical risk factors. In chapter 5, the above risk model was extended by adding the metabolic data:

$$log[P(Y = 1|Z, X)] = \gamma_0 + \sum_{j=1}^{p} \gamma_j z_{ij} + \sum_{k=1}^{m'} \beta_k x_{ik},$$

Since the matrix **X** might contain more variables than observations, penalized models were used in order to select a subset of the metabolic variables. Random forest method was used as well in order to select important variables. Applications on real data showed that the omics data improve performance measures.

12.2 Part II: High Dimensional Biomarkers in Drug Discovery

In chapters 6, 7, and 8, we focused on integrated analysis of multi-source data in the drug discovery experiments. A gene-by-gene analysis was used to identify causal structures in high-dimensional data. For each gene, different causal structures were assumed and the causal structure corresponding to the highest posterior probability was retained. Selection and evaluation of genes which can be used as potential biomarkers in the drug discovery process can help the development team to better understand the mechanism of action of a new set of compounds and substantially shorten the development time. This approach can be implemented in the production pipeline to different number of chemical structure of interest, genes and biological assays (efficacy or toxicity related).

In chapter 7, we discussed different methods which can be used if the interest is to identify set(s) of genes that could be used to predict the outcome of interest. Supervised principal component analysis (SPCA), penalized regression models using the lasso and the elastic net penalties were used to construct gene signature. We have shown that the feature selection procedure can be done in order to maximize a surrogacy measure. For example, for the SPCA, the squared informational coefficient of correlation (SICC, Alonso and Molenberghs, 2007) was used. $R_h^2(U(X))$ is calculated by comparing models:

$$M_0 : E(Y_i|Z_i) = \gamma_0 + \gamma_1 Z_i,$$
$$M_1 : E(Y_i|Z_i, U(X)) = \alpha_0 + \alpha_1 Z_i + \alpha_2 U(X).$$

Here, $U(X)$ is latent and can be estimated using SPCA.

$$R_h^2(\hat{U}(X)) = 1 - \exp\left(\frac{-G^2}{n}\right),$$

where, G^2 is the likelihood ratio statistic comparing models M_0 and M_1, n is the sample size. We have shown that a similar approach can be implemented when lasso or elastic net are used as well. Note that in case lasso or elastic net penalties are used, only the genes are penalized.

The analysis presented in Chapter 7 was focused on the construction of the predictive model and not on inference. A re-sampling based inference procedure can be developed in order to test the following null hypothesis:

$$H_0 : R_h^2 = 0.$$

A rejection of the null hypothesis implies that the omics signature has an added predictive value.

12.3 Part III: Software Development

In the third part of the dissertation, we focused on the validation of surrogate endpoints in randomized clinical trials. We presented new software tools, both SAS and R, that can be used for validation of surrogate endpoints in different settings.

The SAS macro system was written in a generic way so it can be developed in the future to a SAS procedure which will focus on surrogacy. Furthermore, the macros system was written for a wide range of users starting with those who are familiar with surrogacy but may have a limited knowledge about statistical modeling and SAS programing.

The surrogate Shiny App does not require any knowledge in programing. It can be used on a local computer or online (smartphone, tablet, laptop or pc) using the shiny cloud platform. Other cloud platforms, such as Amazon Web Service or google cloud platform are possible as well. In contrast to the `surrogate` R package, the user does not need to install R in order to conduct the analysis. The surrogate shiny app was written in order to increase the accessibility of surrogate software to users who are interested to preform an analysis for a validation a surrogate endpoints but do not have the expertise in both statistical modeling and software. The shiny app offers an easy tool to use for the data analysis and standard outputs and allows the user to conduct a high quality analysis in the same surrogacy settings which are included in the R package surrogate.

Bibliography

Adams, C. P. and Brantner, V. V. (2006) Estimating the cost of new drug development: is it really $ 802 million? *Health Affairs*, **25**, 420–428.

Akaike, H. (1974) A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, **19**, 716–723.

Alonso, A. and Molenberghs, G. (2007) Surrogate marker evaluation from an information theory perspective. *Biometrics*, **63**, 180–186.

Alonso, A. A., Bigirumurame, T., Burzykowski, T., Buyse, M., Leacky, M., Perualila, N. J., Molenberghs, G., Wim, V. d. E. and Shkedy, Z. (2016) *Applied Surrogate Endpoint Evaluation Methods with SAS and R*. Chapman & Hall/CRC Biostatistics Series.

Alter, O., Brown, P. O. and Botstein, D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, **97**, 10101–10106.

Amaratunga, D., Cabrera, J. and Shkedy, Z. (2014) *Exploration and Analysis of DNA Microarray and Other High-Dimensional Data*, vol. Second Edition. Wiley Series in Probability and Statistics.

Ambroise, C. and McLachlan, G. J. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the national academy of sciences*, **99**, 6562–6566.

Asiago, V. M., Alvarado, L. Z., Shanaiah, N., Gowda, G. N., Owusu-Sarfo, K., Ballas, R. A. and Raftery, D. (2010) Early detection of recurrent breast cancer using metabolite profiling. *Cancer research*, **70**, 8309–8318.

Association, A. P. (2000) Diagnostic and statistical manual of mental disorders. 4th text revision ed. *Washington, DC: American Psychiatric Association*, 553–557.

Bach, P. B., Kattan, M. W., Thornquist, M. D., Kris, M. G., Tate, R. C., Barnett, M. J., Hsieh, L. J. and Begg, C. B. (2003) Variations in lung cancer risk among smokers. *Journal of the National Cancer Institute*, **95**, 470–478.

Bach, P. B., Mirkin, J. N., Oliver, T. K., Azzoli, C. G., Berry, D. A., Brawley, O. W., Byers, T., Colditz, G. A., Gould, M. K., Jett, J. R. *et al.* (2012) Benefits and harms of CT screening for lung cancer: a systematic review. *Jama*, **307**, 2418–2429.

Baek, S., Tsai, C.-A. and Chen, J. J. (2009) Development of biomarker classifiers from high-dimensional data. *Briefings in bioinformatics*, **10**, 537–546.

Bain, J. R., Stevens, R. D., Wenner, B. R., Ilkayeva, O., Muoio, D. M. and Newgard, C. B. (2009) Metabolomics applied to diabetes research moving from information to knowledge. *Diabetes*, **58**, 2429–2443.

Bair, E., Hastie, T., Paul, D. and Tibshirani, R. (2006) Prediction by supervised principal components. *Journal of the American Statistical Association*, **101**, 119–137.

Barderas, M. G., Laborde, C. M., Posada, M., de la Cuesta, F., Zubiri, I., Vivanco, F. and Alvarez-Llamas, G. (2011) Metabolomic profiling for identification of novel potential biomarkers in cardiovascular diseases. *BioMed Research International*, **2011**, 9.

Barton, R. H., Nicholson, J. K., Elliott, P. and Holmes, E. (2008) High-throughput 1h nmr-based metabolic analysis of human serum and urine for large-scale epidemiological studies: validation study. *International journal of epidemiology*, **37**, i31–i40.

Bayne, C. K. (1999) Chemometric techniques for quantitative analysis. *Technometrics*, **41**, 173–174.

Beckonert, O., Keun, H. C., Ebbels, T. M., Bundy, J., Holmes, E., Lindon, J. C. and Nicholson, J. K. (2007) Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature protocols*, **2**, 2692–2703.

Benjamini, Y. and 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)*, 289–300.

van den Bergh, K. A., Essink-Bot, M.-L., Bunge, E. M., Scholten, E. T., Prokop, M., van Iersel, C. A., van Klaveren, R. J. and de Koning, H. J. (2008) Impact of computed tomography screening for lung cancer on participants in a randomized controlled trial (NELSON trial). *Cancer*, **113**, 396–404.

Bervoets, L., Louis, E., Reekmans, G., Mesotten, L., Thomeer, M., Adriaensens, P. and Linsen, L. (2015) Influence of preanalytical sampling conditions on the ^{1}H-NMR metabolic profile of human blood plasma and introduction of the standard preanalytical code used in biobanking. *Metabolomics*, **11**, 1197–1207.

Bigirumurame, T., Pushpike, T., Louis, E., Liene, B., Karolien, V., de Jonge, E., Thomeer, M., Mesotten, L., Stinissen, P., Vanderzande, D., Shkedy, Z., Adetayo, K. and Adriaensens, P. (2016) Analysis and statistical validation of ^{1}H-NMR metabolite profiles for early detection of breast cancer. *(Submitted)*.

Biomarkers Definitions Working Group (2001) Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacol & Therapeutics*, **69**, 89–95.

Bollen, K. A. (1989) *Structural Equations with Latent Variables*. Wiley.

Boulesteix, A.-L. (2004) Pls dimension reduction for classification with microarray data. *Statistical applications in genetics and molecular biology*, **3**, 1–30.

Boulesteix, A.-L. and Hothorn, T. (2010) Testing the additional predictive value of high-dimensional molecular data. *BMC bioinformatics*, **11**, 78.

Boulesteix, A.-L., Janitza, S., Kruppa, J. and König, I. R. (2012) Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **2**, 493–507.

Boulesteix, A.-L. and Sauerbrei, W. (2011) Added predictive value of high-throughput molecular data to clinical data and its validation. *Briefings in bioinformatics*, **12**, 215–229.

Boulesteix, A.-L. and Strimmer, K. (2007) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in bioinformatics*, **8**, 32–44.

Bourzac, K. (2014) Diagnosis: early warning system. *Nature*, **513**, S4–S6.

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

Bovey, F. A., Mirau, P. A. and Gutowsky, H. (1988) *Nuclear magnetic resonance spectroscopy.* Elsevier.

Boyle, P., Levin, B. *et al.* (2008) *World cancer report 2008.* IARC Press, International Agency for Research on Cancer.

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

Brett, G. (1968) The value of lung cancer detection by six-monthly chest radiographs. *Thorax*, **23**, 414–420.

Bryan, K., Brennan, L. and Cunningham, P. (2008) Metafind: A feature analysis tool for metabolomics data. *BMC bioinformatics*, **9**, 470.

Bühlmann, P. and Hothorn, T. (2007) Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 477–505.

Burnham, K. P. and Anderson, D. R. (2003) *Model selection and multimodel inference: a practical information-theoretic approach*, vol. Second Edition. Springer Science & Business Media.

Burnham, K. P. and Anderson, D. R. (2004) Multimodel inference understanding AIC and BIC in model selection. *Sociological methods & research*, **33**, 261–304.

Burzykowski, T., Molenberghs, G. and Buyse, M. (2004) The validation of surrogate end points by using data from randomized clinical trials: a case-study in advanced colorectal cancer. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **167**, 103–124.

Burzykowski, T., Molenberghs, G. and Buyse, M. (2005) *The evaluation of surrogate endpoints.* New York: Springer-Verlag.

Burzykowski, T., Molenberghs, G., Buyse, M., Geys, H. and Renard, D. (2001) Validation of surrogate end points in multiple randomized clinical trials with failure time end points. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **50**, 405–422.

Buyse, M., Michiels, S., Squifflet, P., Lucchesi, K. J., Hellstrand, K., Brune, M. L., Castaigne, S. and Rowe, J. M. (2011) Leukemia-free survival as a surrogate end point for overall survival in the evaluation of maintenance therapy for patients with acute myeloid leukemia in complete remission. *Haematologica*, **96**, 1106–1112.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D. and Geys, H. (2000) The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, **1**, 49–67.

Buyse, M., Vangeneugden, T., Bijnens, L., Renard, D., Burzykowski, T., Geys, H. and Molenberghs, G. (2003) Validation of biomarkers as surrogates for clinical endpoints. *Drugs and the Pharmaceutical Sciences*, **132**, 149–168.

Cantor, J. R. and Sabatini, D. M. (2012) Cancer cell metabolism: one hallmark, many faces. *Cancer discovery*, **2**, 881–898.

Cassidy, A., Duffy, S. W., Myles, J. P., Liloglou, T. and Field, J. K. (2007) Lung cancer risk prediction: a tool for early detection. *International journal of cancer*, **120**, 1–6.

Cassidy, A., Myles, J. P., van Tongeren, M., Page, R., Liloglou, T., Duffy, S. and Field, J. (2008) The LLP risk model: an individual risk prediction model for lung cancer. *British journal of cancer*, **98**, 270–276.

Chen, W., Zu, Y., Huang, Q., Chen, F., Wang, G., Lan, W., Bai, C., Lu, S., Yue, Y. and Deng, F. (2011) Study on metabonomic characteristics of human lung cancer using high resolution magic-angle spinning ^{1}H-NMR spectroscopy and multivariate data analysis. *Magnetic resonance in medicine*, **66**, 1531–1540.

Chen, X., Wang, L., Smith, J. D. and Zhang, B. (2008) Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes. *Bioinformatics*, **24**, 2474–2481.

Chen, Y., Ma, Z., Li, A., Li, H., Wang, B., Zhong, J., Min, L. and Dai, L. (2015) Metabolomic profiling of human serum in lung cancer patients using liquid chromatography/hybrid quadrupole time-of-flight mass spectrometry and gas chromatography/mass spectrometry. *Journal of cancer research and clinical oncology*, **141**, 705–718.

Cheng, L. L., Burns, M. A., Taylor, J. L., He, W., Halpern, E. F., McDougal, W. S. and Wu, C.-L. (2005) Metabolic characterization of human prostate cancer with tissue magnetic resonance spectroscopy. *Cancer research*, **65**, 3030–3034.

Claeskens, G. and Hjort, N. L. (2008) *Model selection and model averaging*. Cambridge University Press Cambridge.

Clarke, P. A., te Poele, R. and Workman, P. (2004) Gene expression microarray technologies in the development of new therapeutic agents. *European journal of cancer*, **40**, 2560–2591.

Cowlrick, I., Hedner, T., Wolf, R., Olausson, M. and Klofsten, M. (2011) Decision-making in the pharmaceutical industry: analysis of entrepreneurial risk and attitude using uncertain information. *R&D Management*, **41**, 321–336.

Cuffe, S., Moua, T., Summerfield, R., Roberts, H., Jett, J. and Shepherd, F. A. (2011) Characteristics and outcomes of small cell lung cancer patients diagnosed during two lung cancer computed tomographic screening programs in heavy smokers. *Journal of Thoracic Oncology*, **6**, 818–822.

Danner, D., Hagemann, D. and Fiedler, K. (2015) Mediation analysis with structural equation models: Combining theory, design, and statistics. *European Journal of Social Psychology*, **45**, 460–481.

De Bin, R., Sauerbrei, W. and Boulesteix, A.-L. (2014) Investigating the prediction ability of survival models based on both clinical and omics data: two case studies. *Statistics in medicine*, **33**, 5310–5329.

De Gruttola, V., Fleming, T., Lin, D. and Coombs, R. (1997) Perspective: validating surrogate markers are we being naive? *Journal of Infectious Diseases*, **175**, 237–246.

DiMasi, J. A., Hansen, R. W. and Grabowski, H. G. (2003) The price of innovation: new estimates of drug development costs. *Journal of health economics*, **22**, 151–185.

Duarte, I. F., Rocha, C. M., Barros, A. S., Gil, A. M., Goodfellow, B. J., Carreira, I. M., Bernardo, J., Gomes, A., Sousa, V. and Carvalho, L. (2010) Can nuclear magnetic resonance (NMR) spectroscopy reveal different metabolic signatures for lung tumours? *Virchows Archiv*, **457**, 715–725.

Duarte, I. F., Rocha, C. M. and Gil, A. M. (2013) Metabolic profiling of biofluids: potential in lung cancer screening and diagnosis. *Expert review of molecular diagnostics*, **13**, 737–748.

Dudoit, S., Fridlyand, J. and Speed, T. P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association*, **97**, 77–87.

Duffy, M. J. (2006) Serum tumor markers in breast cancer: are they of clinical value? *Clinical chemistry*, **52**, 345–351.

Ebeling, F., Stieber, P., Untch, M., Nagel, D., Konecny, G., Schmitt, U., Fateh-Moghadam, A. and Seidel, D. (2002) Serum cea and ca 15-3 as prognostic factors in primary breast cancer. *British journal of cancer*, **86**, 1217–1222.

Efron, B. and Tibshirani, R. (1997) Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, **92**, 548–560.

Eliassen, A. H., Spiegelman, D., Xu, X., Keefer, L. K., Veenstra, T. D., Barbieri, R. L., Willett, W. C., Hankinson, S. E. and Ziegler, R. G. (2012) Urinary estrogens and estrogen metabolites and subsequent risk of breast cancer among premenopausal women. *Cancer research*, **72**, 696–706.

Ellenberg, S. S. and Hamilton, J. M. (1989) Surrogate endpoints in clinical trials: cancer. *Statistics in medicine*, **8**, 405–413.

Emwas, A.-H. M., Salek, R. M., Griffin, J. L. and Merzaban, J. (2013) NMR-based metabolomics in human disease diagnosis: applications, limitations, and recommendations. *Metabolomics*, **9**, 1048–1072.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D. M., Forman, D. and Bray, F. (2015) Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. *International Journal of Cancer*, **136**, E359–E386.

Fiedler, K., Schott, M. and Meiser, T. (2011) What mediation analysis can (not) do. *Journal of Experimental Social Psychology*, **47**, 1231–1236.

Field, J. K., Chen, Y., Marcus, M. W., Mcronald, F. E., Raji, O. Y. and Duffy, S. W. (2013a) The contribution of risk prediction models to early detection of lung cancer. *Journal of surgical oncology*, **108**, 304–311.

Field, J. K. and Duffy, S. (2008) Lung cancer screening: the way forward. *British journal of cancer*, **99**, 557–562.

Field, J. K., Oudkerk, M., Pedersen, J. H. and Duffy, S. W. (2013b) Prospects for population screening and diagnosis of lung cancer. *The Lancet*, **382**, 732–741.

Fischer, H. and Heyse, S. (2005) From targets to leads: the importance of advanced data analysis for decision support in drug discovery. *Current opinion in drug discovery & development*, **8**, 334–346.

Friebolin, H. and Becconsall, J. K. (1993) *Basic one-and two-dimensional NMR spectroscopy*. VCH Weinheim.

Friedman, J., Hastie, T. and Tibshirani, R. (2001) *The elements of statistical learning*, vol. 2. Springer series in statistics Springer, Berlin.

Gartlehner, G., Thaler, K., Chapman, A., Kaminski-Hartenthaler, A., Berzaczy, D., Van Noord, M. G. and Helbich, T. H. (2013) Mammography in combination with breast ultrasonography versus mammography for breast cancer screening in women at average risk. *The Cochrane database of systematic reviews*, **4**, CD009632.

Giskeødegård, G. F., Grinde, M. T., Sitter, B., Axelson, D. E., Lundgren, S., Fjøsne, H. E., Dahl, S., Gribbestad, I. S. and Bathen, T. F. (2010) Multivariate modeling and prediction of breast cancer prognostic factors using mr metabolomics. *Journal of proteome research*, **9**, 972–979.

Goeman, J. J., Oosting, J., Cleton-Jansen, A.-M., Anninga, J. K. and Van Houwelingen, H. C. (2005) Testing association of a pathway with survival using gene expression data. *Bioinformatics*, **21**, 1950–1957.

Goeman, J. J., Van De Geer, S. A., De Kort, F. and Van Houwelingen, H. C. (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**, 93–99.

Göhlmann, H. and Talloen, W. (2009) *Gene expression studies using Affymetrix microarrays*. CRC Press.

Goldstraw, P., Crowley, J., Chansky, K., Giroux, D. J., Groome, P. A., Rami-Porta, R., Postmus, P. E., Rusch, V., Sobin, L. and for the Study of Lung Cancer International Staging Committee, I. A. (2007) The iaslc lung cancer staging project: proposals for the revision of the tnm stage groupings in the forthcoming (seventh) edition of the tnm classification of malignant tumours. *Journal of thoracic oncology*, **2**, 706–714.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R. and Caligiuri, M. A. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Greenland, S., Pearl, J. and Robins, J. M. (1999) Causal diagrams for epidemiologic research. *Epidemiology*, **10**, 37–48.

Grepin, C. and Pernelle, C. (2000) High-throughput screening. *Drug discovery today*, **5**, 212–214.

Gromski, P. S., Muhamadali, H., Ellis, D. I., Xu, Y., Correa, E., Turner, M. L. and Goodacre, R. (2015) A tutorial review: Metabolomics and partial least squares-discriminant analysis–a marriage of convenience or a shotgun wedding. *Analytica chimica acta*, **879**, 10–23.

Guo, Y., Hastie, T. and Tibshirani, R. (2007) Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, **8**, 86–100.

Guy, W. (1976) Clinical global impression scale. *The ECDEU Assessment Manual for Psychopharmacology*, **338**, 218–222.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002) Gene selection for cancer classification using support vector machines. *Machine learning*, **46**, 389–422.

Hasan, N., Kumar, R. and Kavuru, M. S. (2014) Lung cancer screening beyond low-dose computed tomography: the role of novel biomarkers. *Lung*, **192**, 639–648.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015) *Statistical learning with sparsity: the lasso and generalizations*. CRC Press.

Henschke, C. I., McCauley, D. I., Yankelevitz, D. F., Naidich, D. P., McGuinness, G., Miettinen, O. S., Libby, D. M., Pasmantier, M. W., Koizumi, J., Altorki, N. K. *et al.* (1999) Early lung cancer action project: overall design and findings from baseline screening. *The Lancet*, **354**, 99–105.

Hoggart, C., Brennan, P., Tjonneland, A., Vogel, U., Overvad, K., Østergaard, J. N., Kaaks, R., Canzian, F., Boeing, H. and Steffen, A. (2012) A risk model for lung cancer incidence. *Cancer Prevention Research*, **5**, 834–846.

Holmes, E., Wilson, I. D. and Nicholson, J. K. (2008) Metabolic phenotyping in health and disease. *Cell*, **134**, 714–717.

Horeweg, N., Scholten, E. T., de Jong, P. A., van der Aalst, C. M., Weenink, C., Lammers, J.-W. J., Nackaerts, K., Vliegenthart, R., ten Haaf, K., Yousaf-Khan, U. A. *et al.* (2014) Detection of lung cancer through low-dose CT screening (NELSON): a prespecified analysis of screening test performance and interval cancers. *The Lancet Oncology*, **15**, 1342–1350.

Hori, S., Nishiumi, S., Kobayashi, K., Shinohara, M., Hatakeyama, Y., Kotani, Y., Hatano, N., Maniwa, Y., Nishio, W. and Bamba, T. (2011) A metabolomic approach to lung cancer. *Lung cancer*, **74**, 284–292.

Höskuldsson, A. (1988) Pls regression methods. *Journal of chemometrics*, **2**, 211–228.

Howlader, N., Noone, A., Krapcho, M., Garshell, J., Neyman, N., Aletkruse, S., Kosary, C., Yu, M., Ruhl, J., Tatalovich, Z., Cho, H., Mariotto, A., Lewis, D., Chen, H., Feuer, E. and Cronin, K. (2012) Seer cancer statistics review, 1975-2010,[online] national cancer institute. bethesda, md, http://seer.cancer.gov/csr/1975_2010/, based on november 2012 seer data submission, posted to the seer web site, april 2013.

Hughes, J., Rees, S., Kalindjian, S. and Philpott, K. (2011) Principles of early drug discovery. *British journal of pharmacology*, **162**, 1239–1249.

Humphrey, L. L., Teutsch, S. and Johnson, M. (2004) Lung cancer screening with sputum cytologic examination, chest radiography, and computed tomography: an update for the us preventive services task force. *Annals of Internal Medicine*, **140**, 740–753.

Hürlimann, M. and Griffin, D. (2000) Spin dynamics of carr–purcell–meiboom–gill-like sequences in grossly inhomogeneous B0 and B1 fields and application to NMR well logging. *Journal of Magnetic Resonance*, **143**, 120–135.

Iurlaro, R., León-Annicchiarico, C. L. and Muñoz-Pinedo, C. (2014) Regulation of cancer metabolism by oncogenes and tumor suppressors. *Methods Enzymol*, **542**, 59–80.

Jansson, J., Willing, B., Lucio, M., Fekete, A., Dicksved, J., Halfvarson, J., Tysk, C. and Schmitt-Kopplin, P. (2009) Metabolomics reveals metabolic biomarkers of crohn's disease. *PloS one*, **4**, e6386.

Jöreskog, K. G. (1993) Testing structural equation models. *Sage focus editions*, **154**, 294–294.

Kane, J., Honigfeld, G., Singer, J. and Meltzer, H. (1988) Clozapine for the treatment-resistant schizophrenic: a double-blind comparison with chlorpromazine. *Archives of general psychiatry*, **45**, 789–796.

Kochhar, S., Jacobs, D. M., Ramadan, Z., Berruex, F., Fuerholz, A. and Fay, L. B. (2006) Probing gender-specific metabolism differences in humans by nuclear magnetic resonance-based metabonomics. *Analytical biochemistry*, **352**, 274–281.

Kohavi, R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, vol. 2 of *IJCAI'95*, 1137–1143.

Kraljevic, S., Stambrook, P. J. and Pavelic, K. (2004) Accelerating drug discovery. *EMBO reports*, **5**, 837–842.

Kramer, R. (1998) *Chemometric techniques for quantitative analysis*. CRC Press.

Kroemer, G. and Pouyssegur, J. (2008) Tumor cell metabolism: cancer's achilles' heel. *Cancer cell*, **13**, 472–482.

Kuiper, R., Hoijtink, H. and Silvapulle, M. (2011) An akaike-type information criterion for model selection under inequality constraints. *Biometrika*, **98**, 495–501.

Kuiper, R. M., Gerhard, D. and Hothorn, L. A. (2014) Identification of the minimum effective dose for normally distributed endpoints using a model selection approach. *Statistics in Biopharmaceutical Research*, **6**, 55–66.

Larke, F. J., Kruger, R. L., Cagnon, C. H., Flynn, M. J., McNitt-Gray, M. M., Wu, X., Judy, P. F. and Cody, D. D. (2011) Estimated radiation dose associated with low-dose chest CT of average-size participants in the national lung screening trial. *American Journal of Roentgenology*, **197**, 1165–1169.

Lennon, G. G. (2000) High-throughput gene expression analysis for drug discovery. *Drug Discovery Today*, **5**, 59–66.

Lenz, E., Bright, J., Wilson, I., Hughes, A., Morrisson, J., Lindberg, H. and Lockton, A. (2004) Metabonomics, dietary influences and cultural differences: a ^{1}H-NMR-based study of urine samples obtained from healthy british and swedish subjects. *Journal of pharmaceutical and biomedical analysis*, **36**, 841–849.

Leucht, S., Kane, J. M., Kissling, W., Hamann, J., Etschel, E. and Engel, R. (2005) Clinical implications of brief psychiatric rating scale scores. *The British Journal of Psychiatry*, **187**, 366–371.

Li, R., Tsaih, S.-W., Shockley, K., Stylianou, I. M., Wergedal, J., Paigen, B. and Churchill, G. A. (2006) Structural model analysis of multiple quantitative traits. *PLoS Genet*, **2**, e114.

Lin, D., Fleming, T., De Gruttola, V. *et al.* (1997) Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in medicine*, **16**, 1515–1527.

Lin, D., Shkedy, Z., Yekutieli, D., Amaratunga, D. and Bijnens, L. (2012) *Modeling dose-response microarray data in early drug development experiments using R: order-restricted analysis of microarray data.* Springer Science & Business Media.

Lin, D. Y. and Wei, L.-J. (1989) The robust inference for the cox proportional hazards model. *Journal of the American Statistical Association*, **84**, 1074–1078.

Lindon, J. C. and Nicholson, J. K. (2008) Spectroscopic and statistical techniques for information recovery in metabonomics and metabolomics. *Annu. Rev. Anal. Chem.*, **1**, 45–69.

Loehlin, J. C. (1998) *Latent variable models: An introduction to factor, path, and structural analysis .* Lawrence Erlbaum Associates Publishers.

Lonn, E. (2001) The use of surrogate endpoints in clinical trials: focus on clinical trials in cardiovascular diseases. *Pharmacoepidemiology and drug safety*, **10**, 497–508.

Louis, E., Adriaensens, P., Guedens, W., Bigirumurame, T., Baeten, K., Vanhove, K., Vandeurzen, K., Darquennes, K., Vansteenkiste, J., Dooms, C., Ziv, S., Liesbet, M. and Michiel, T. (2016) Detection of lung cancer through metabolic changes measured in blood plasma. *Journal of Thoracic Oncology*, **11**, 516–523.

Louis, E., Adriaensens, P., Guedens, W., Vanhove, K., Vandeurzen, K., Darquennes, K., Vansteenkiste, J., Dooms, C., de Jonge, E. and Thomeer, M. (2015a) Metabolic phenotyping of human blood plasma: a powerful tool to discriminate between cancer types? *Annals of Oncology*, mdv499.

Louis, E., Bervoets, L., Reekmans, G., De Jonge, E., Mesotten, L., Thomeer, M. and Adriaensens, P. (2015b) Phenotyping human blood plasma by ^{1}H-NMR: a robust protocol based on metabolite spiking and its evaluation in breast cancer. *Metabolomics*, **11**, 225–236.

MacLachlan, G. (1992) *Discriminant analysis and statistical pattern recognition.* New York: Wiley.

Mamas, M., Dunn, W. B., Neyses, L. and Goodacre, R. (2011) The role of metabolites and metabolomics in clinically applicable biomarkers of disease. *Archives of toxicology*, **85**, 5–17.

Mayr, L. M. and Bojanic, D. (2009) Novel trends in high-throughput screening. *Current opinion in pharmacology*, **9**, 580–588.

McPherson, K., Steel, C. and Dixon, J. (2000) Breast cancer epidemiology, risk factors, and genetics. *British Medical Journal*, **321**, 624–628.

Meiboom, S. and Gill, D. (1958) Modified spin-echo method for measuring nuclear relaxation times. *Review of scientific instruments*, **29**, 688–691.

Meinshausen, N. (2007) Relaxed lasso. *Computational Statistics & Data Analysis*, **52**, 374–393.

Melamed, M. R., Flehinger, B. J., Zaman, M. B., Heelan, R. T., Perchick, W. A. and Martini, N. (1984) Screening for early lung cancer. results of the memorial sloan-kettering study in new york. *CHEST Journal*, **86**, 44–53.

Morgan, S. L. and Winship, C. (2007) *Counterfactuals and causal inference*. Cambridge University Press.

Mortimer, A. M. (2007) Symptom rating scales and outcome in schizophrenia. *The British Journal of Psychiatry*, **191**, s7–s14.

Mulshine, J. and Sullivan, D. (2005) Clinical practice. lung cancer screening. *The New England journal of medicine*, **352**, 2714–2720.

Munoz-Pinedo, C., El Mjiyad, N. and Ricci, J. (2012) Cancer metabolism: current perspectives and future directions. *Cell death & disease*, **3**, e248.

National Lung Screening Trial Research Team (2011a) The national lung screening trial: Overview and study design1. *Radiology*, **258**, 243–253.

National Lung Screening Trial Research Team (2011b) Reduced lung-cancer mortality with low-dose computed tomographic screening. *The New England journal of medicine*, **365**, 395–409.

National Lung Screening Trial Research Team (2013) Results of initial low-dose computed tomographic screening for lung cancer. *The New England journal of medicine*, **368**, 1980–1991.

Nguyen, D. V. and Rocke, D. M. (2002) Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18**, 39–50.

Nicholson, J., Lindon, J. and Holmes, E. (1999) ”metabonomics”:understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*, **29**, 1181–1189.

Oakman, C., Tenori, L., Biganzoli, L., Santarpia, L., Cappadona, S., Luchinat, C. and Di Leo, A. (2011) Uncovering the metabolomic fingerprint of breast cancer. *The international journal of biochemistry & cell biology*, **43**, 1010–1020.

O'Connell, T. M. (2012) Recent advances in metabolomics in oncology. *Bioanalysis*, **4**, 431–451.

Omura, G., Buyse, M., Marsoni, S., Bertelsen, K., Conte, P., Jakobsen, A. and Vermorkin, J. (1991) Cyclophosphamide plus cisplatin plus adriamycin versus cyclophosphamide, doxorubicin, and cisplatin chemotherapy of ovarian carcinoma: a meta-analysis. *Journal of Clinical Oncology*, **9**, 1668–1674.

Overall, J. E. and Gorham, D. R. (1962) The brief psychiatric rating scale. *Psychological reports*, **10**, 799–812.

Pan, Z. and Raftery, D. (2007) Comparing and combining NMR spectroscopy and mass spectrometry in metabolomics. *Analytical and bioanalytical chemistry*, **387**, 525–527.

Park, M. Y. and Hastie, T. (2007) L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**, 659–677.

Parkin, D. M., Pisani, P. and Ferlay, J. (1999) Global cancer statistics. *CA: a cancer journal for clinicians*, **49**, 33–64.

Parkin, D. M., Pisani, P. and Ferlay, J. (2005) Global cancer statistics. *CA: a cancer journal for clinicians*, **55**, 74–108.

Pearl, J. (2009) Causal inference in statistics: An overview. *Statistics Surveys*, **3**, 96–146.

Pérez-Enciso, M. and Tenenhaus, M. (2003) Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. *Human genetics*, **112**, 581–592.

Perualila, T. N., Kasim, A., Talloen, W., Göhlmann, H., Verbist, B. and Shkedy, Z. (2016) A joint modeling approach for uncovering associations between gene expression, bioactivity and chemical structure in early drug discovery to guide lead selection and genomic biomarker development. *Statistical Applications in Genetics and Molecular Biology*, **00**, 00–000.

Peter, B. and Bernard, L. (2008) *World cancer report 2008.* World Health Organization.

Pharmacological Therapy for Macular Degeneration Study Group (1997) Interferon α-IIA is ineffective for patients with choroidal neovascularization secondary to age-related macular degeneration. results of a prospective randomized placebo-controlled clinical trial. *Archives of Ophthalmology*, **115**, 865–872.

Pike, M. C., Spicer, D. V., Dahmoush, L. and Press, M. F. (1993) Estrogens progestogens normal breast cell proliferation and breast cancer risk. *Epidemiologic reviews*, **15**, 17–35.

Raji, O. Y., Agbaje, O. F., Duffy, S. W., Cassidy, A. and Field, J. K. (2010) Incorporation of a genetic factor into an epidemiologic model for prediction of individual risk of lung cancer: the liverpool lung project. *Cancer Prevention Research*, **3**, 664–669.

Renard, D., Geys, H., Molenberghs, G., Burzykowski, T. and Buyse, M. (2002) Validation of surrogate endpoints in multiple randomized clinical trials with discrete outcomes. *Biometrical Journal*, **44**, 921–935.

Ripley, B. (1996) *Pattern recognition and neural networks.* Cambridge university press, Cambridge, UK.

Rocha, C. M., Carrola, J., Barros, A. S., Gil, A. M., Goodfellow, B. J., Carreira, I. M., Bernardo, J., Gomes, A., Sousa, V., Carvalho, L. *et al.* (2011) Metabolic signatures of lung cancer in biofluids: NMR-based metabonomics of blood plasma. *Journal of proteome research*, **10**, 4314–4324.

Schwarz, G. (1978) Estimating the dimension of a model. *The annals of statistics*, **6**, 461–464.

Sciacovelli, M., Gaude, E., Hilvo, M. and Frezza, C. (2014) The metabolic alterations of cancer cells. *Methods Enzymol*, **542**, 1–23.

Shlomi, D., Ben-Avi, R., Balmor, G. R., Onn, A. and Peled, N. (2014) Screening for lung cancer: time for large-scale screening by chest computed tomography. *European Respiratory Journal*, **44**, 217–238.

Sickles, E. (1991) Screening for breast cancer with mammography. *Clinical Imaging*, **15**, 253–260.

Singh, M. M. and Kay, S. R. (1975) A comparative study of haloperidol and chlorpromazine in terms of clinical effects and therapeutic reversal with benztropine in schizophrenia. theoretical implications for potency differences among neuroleptics. *Psychopharmacologia*, **43**, 103–113.

Sitter, B., Bathen, T. F., Singstad, T. E., Fjøsne, H. E., Lundgren, S., Halgunset, J. and Gribbestad, I. S. (2010) Quantification of metabolites in breast cancer patients with different clinical prognosis using hr mas mr spectroscopy. *NMR in Biomedicine*, **23**, 424–431.

Slawski, M., Daumer, M. and Boulesteix, A.-L. (2008) CMA–a comprehensive bioconductor package for supervised classification with high dimensional data. *BMC bioinformatics*, **9**, 439.

Smolinska, A., Blanchet, L., Buydens, L. M. and Wijmenga, S. S. (2012) NMR and pattern recognition methods in metabolomics: from data acquisition to biomarker discovery: a review. *Analytica chimica acta*, **750**, 82–97.

Smyth, G. K. (2004) Linear models and empirical bayes methods for assessing diferential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**, 1–25.

Smyth, G. K. (2005) Limma: linear models for microarray data. In: *Bioinformatics and computational biology solutions using R and Bioconductor*, 397–420. Springer.

Sone, S., Li, F., Yang, Z., Takashima, S., Maruyama, Y., Hasegawa, M., Wang, J., Kawakami, S. and Honda, T. (2000) Characteristics of small lung cancers invisible on conventional chest radiography and detected by population based screening using spiral CT. *The British journal of radiology*, **73**, 137–145.

Sone, S., Takashima, S., Li, F., Yang, Z., Honda, T., Maruyama, Y., Hasegawa, M., Yamanda, T., Kubo, K. and Hanamura, K. (1998) Mass screening for lung cancer with mobile spiral computed tomography scanner. *The Lancet*, **351**, 1242–1245.

Spitz, M. R., Etzel, C. J., Dong, Q., Amos, C. I., Wei, Q., Wu, X. and Hong, W. K. (2008) An expanded risk prediction model for lung cancer. *Cancer Prevention Research*, **1**, 250–254.

Spitz, M. R., Hong, W. K., Amos, C. I., Wu, X., Schabath, M. B., Dong, Q., Shete, S. and Etzel, C. J. (2007) A risk model for prediction of lung cancer. *Journal of the National Cancer Institute*, **99**, 715–726.

Starkuviene, V. and Pepperkok, R. (2007) The potential of high-content high-throughput microscopy in drug discovery. *British journal of pharmacology*, **152**, 62–71.

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

Szymańska, E., Saccenti, E., Smilde, A. K. and Westerhuis, J. A. (2012) Double-check: validation of diagnostic statistics for pls-da models in metabolomics studies. *Metabolomics*, **8**, 3–16.

Tammemagi, C. M., Pinsky, P. F., Caporaso, N. E., Kvale, P. A., Hocking, W. G., Church, T. R., Riley, T. L., Commins, J., Oken, M. M. and Berg, C. D. (2011) Lung cancer risk prediction: prostate, lung, colorectal and ovarian cancer screening trial models and validation. *Journal of the National Cancer Institute*, **103**, 1058–1068.

Tammemagi, M. C., Katki, H. A., Hocking, W. G., Church, T. R., Caporaso, N., Kvale, P. A., Chaturvedi, A. K., Silvestri, G. A., Riley, T. L. and Commins, J. (2013) Selection criteria for lung-cancer screening. *New England Journal of Medicine*, **368**, 728–736.

Tammemagi, M. C. and Lam, S. (2014) Screening for lung cancer using low dose computed tomography. *BMJ*, **348**, g2253.

Tibaldi, F., Abrahantes, J. C., Molenberghs, G., Renard, D., Burzykowski, T., Buyse, M., Parmar, M., Stijnen, T. and Wolfinger, R. (2003) Simplified hierarchical linear models for the evaluation of surrogate endpoints. *Journal of Statistical Computation and Simulation*, **73**, 643–658.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tilahun, A., Lin, D., Shkedy, Z., Geys, H., Alonso, A., Peeters, P., Talloen, W., Drinkenburg, W., Göhlmann, H. and Gorden, E. (2010) Genomic biomarkers for depression: Feature-specific and joint biomarkers. *Statistics in Biopharmaceutical Research*, **2**, 419–434.

Tsay, J.-C. J., DeCotiis, C., Greenberg, A. K. and Rom, W. N. (2014) Current readings: blood-based biomarkers for lung cancer. In: *Seminars in thoracic and cardiovascular surgery*, 328–334. Elsevier.

Upadhyay, M., Samal, J., Kandpal, M., Singh, O. V. and Vivekanandan, P. (2013) The warburg effect: insights from the past decade. *Pharmacology & therapeutics*, **137**, 318–330.

Van Houwelingen, H., Arends, L. and Stijnen, T. (2002) Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medecine*, **4**, 589–624.

Van Sanden, S., Shkedy, Z., Burzykowski, T., Göhlmann, H. W., Talloen, W. and Bijnens, L. (2012) Genomic biomarkers for a binary clinical outcome in early drug development microarray experiments. *Journal of biopharmaceutical statistics*, **22**, 72–92.

Verbist, B., Klambauer, G., Vervoort, L., Talloen, W., Shkedy, Z., Thas, O., Bender, A., Göhlmann, H. W., Hochreiter, S. and Consortium, Q. (2015) Using transcriptomics to guide lead optimization in drug discovery projects: Lessons learned from the qstar project. *Drug discovery today*, **20**, 505–513.

Weir, C. J. and Walley, R. J. (2006) Statistical evaluation of biomarkers as surrogate endpoints: a literature review. *Statistics in medicine*, **25**, 183–203.

Wen, T., Gao, L., Wen, Z., Wu, C., Tan, C. S., Toh, W. Z. and Ong, C. N. (2013) Exploratory investigation of plasma metabolomics in human lung adenocarcinoma. *Molecular BioSystems*, **9**, 2370–2378.

Wood, D. E., Eapen, G. A., Ettinger, D. S., Hou, L., Jackman, D., Kazerooni, E., Klippenstein, D., Lackner, R. P., Leard, L., Leung, A. N. *et al.* (2012) Lung cancer screening. *Journal of the National Comprehensive Cancer Network*, **10**, 240–265.

Zhaoa, Y. R., Xiea, X., de Koningb, H. J., Malic, W. P., Vliegentharta, R. and Oudkerka, M. (2011) NELSON lung cancer screening study. *Cancer Imaging*, **11**, S79–S84.

Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 301–320.

Appendix A

Supplementary for Chapter 3

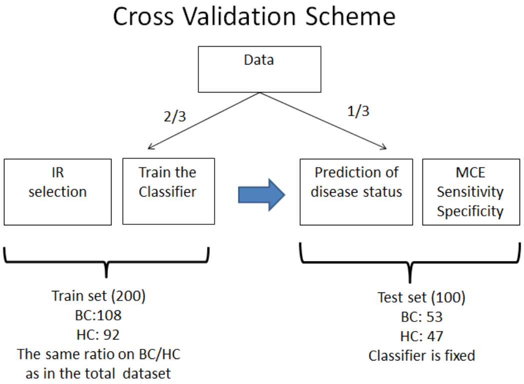


Figure A.1: *Scheme of the 3-fold cross validation procedure.*

	3	12	20	30	40	43
LDA	0.28	0.24	0.24	0.24	0.24	0.24
DLDA	0.28	0.26	0.26	0.26	0.26	0.26
fda	0.22	0.21	0.21	0.22	0.22	0.22
PLSLDA	0.28	0.24	0.28	0.28	0.28	0.30
SVM	0.26	0.24	0.22	0.22	0.22	0.23
RF	0.33	0.28	0.26	0.26	0.24	0.26
QDA	0.17	0.17	0.17	0.22	0.28	0.30

Table A.1: *Median misclassification error of BC as HC obtained for the seven different classification methods and different 'top-k-based' classifiers (build by the top 3, 8, 20, 30, 40, and 43 IRs selected by the Limma t-test).*

	3	12	20	30	40	43
LDA	0.72	0.76	0.76	0.76	0.76	0.76
DLDA	0.72	0.74	0.74	0.74	0.74	0.74
FDA	0.78	0.79	0.79	0.78	0.78	0.78
PLSLDA	0.72	0.76	0.72	0.72	0.72	0.70
SVM	0.74	0.76	0.78	0.78	0.78	0.77
RF	0.67	0.72	0.74	0.74	0.76	0.74
QDA	0.83	0.83	0.83	0.78	0.72	0.70

Table A.2: *Median sensitivity obtained for the seven different classification methods and different 'top-k-based' classifiers (build by the top 3, 8, 20, 30, 40, and 43 IRs selected by the Limma t-test).*

	3	12	20	30	40	43
LDA	0.69	0.70	0.72	0.74	0.74	0.74
DLDA	0.66	0.63	0.64	0.64	0.63	0.63
FDA	0.63	0.67	0.69	0.70	0.72	0.74
PLSLDA	0.70	0.69	0.69	0.70	0.72	0.72
SVM	0.74	0.74	0.74	0.75	0.76	0.76
RF	0.74	0.78	0.79	0.80	0.81	0.81
QDA	0.55	0.58	0.62	0.68	0.74	0.74

Table A.3: *Median specificity obtained for the seven different classification methods and different 'top-k-based' classifiers (build by the top 3, 8, 20, 30, 40, and 43 IRs selected by the Limma t-test).*

	3	12	20	30	40	43
	MCE					
LDA	0.39	0.37	0.33	0.30	0.30	0.30
DLDA	0.40	0.41	0.40	0.41	0.40	0.40
FDA	0.48	0.43	0.37	0.33	0.30	0.31
PLSLDA	0.38	0.37	0.35	0.34	0.34	0.34
SVM	0.34	0.31	0.30	0.26	0.25	0.26
RF	0.41	0.34	0.30	0.26	0.26	0.25
QDA	0.43	0.48	0.37	0.30	0.22	0.22
	SPECIFICITY					
LDA	0.61	0.63	0.67	0.70	0.70	0.70
DLDA	0.60	0.59	0.60	0.59	0.60	0.60
FDA	0.52	0.57	0.63	0.67	0.70	0.69
PLSLDA	0.62	0.63	0.65	0.66	0.66	0.66
SVM	0.66	0.69	0.70	0.74	0.75	0.74
RF	0.59	0.66	0.70	0.74	0.74	0.75
QDA	0.57	0.52	0.63	0.70	0.78	0.78
	SENSITIVITY					
LDA	0.45	0.59	0.65	0.67	0.68	0.68
DLDA	0.46	0.57	0.61	0.64	0.66	0.66
FDA	0.54	0.64	0.67	0.70	0.72	0.70
PLSLDA	0.43	0.52	0.57	0.57	0.57	0.57
SVM	0.38	0.53	0.59	0.63	0.65	0.64
RF	0.48	0.57	0.57	0.59	0.59	0.60
QDA	0.50	0.77	0.76	0.74	0.67	0.65

Table A.4: *Median misclassification error of BC as HC (top), sensitivity (bottom) and specificity (middle) obtained by the different classification methods for several 'top-k-based' classifiers (k = 3, 8, 20, 30, 40, and 43) for which the IRs were selected by the Limma t-test out of the remaining 56 IRs.*

Appendix B

Supplementary for Chapter 5

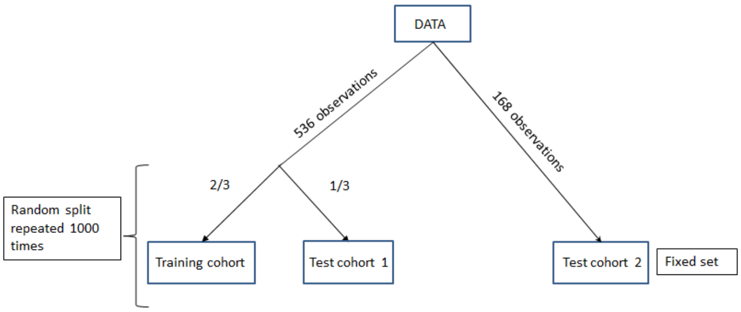


Figure B.1: *Illustration of the cross validation procedure using lasso and elastic net penalties.*

Appendix C

Supplementary Results for Chapter 8

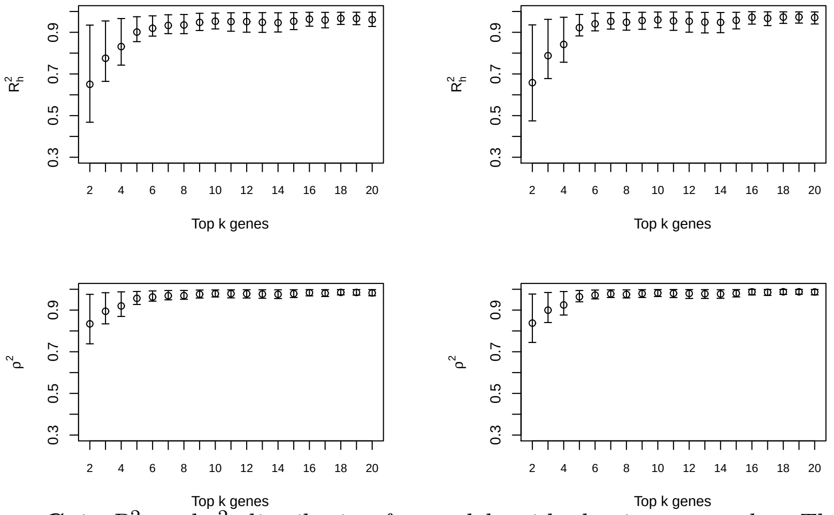


Figure C.1: *R_h^2 and ρ^2 distribution for models with elastic net penalty. The mixing parameter $\alpha = 0, 0.1$.*

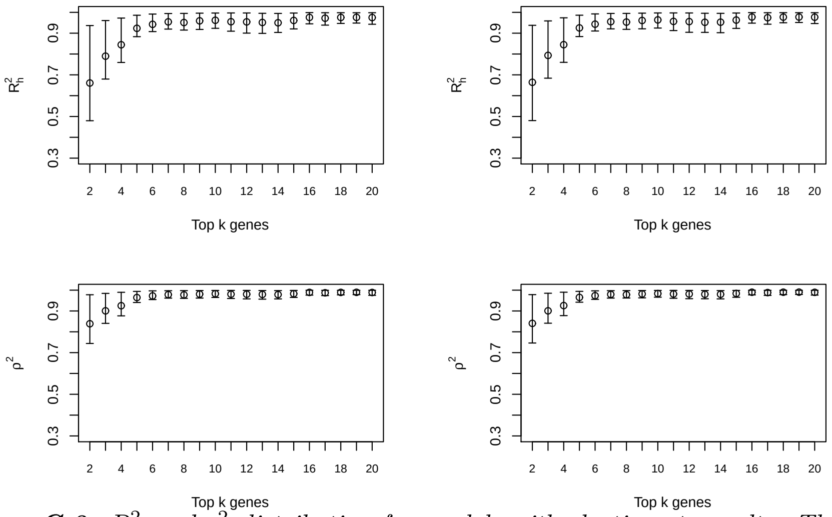


Figure C.2: *R^2_h and ρ^2 distribution for models with elastic net penalty. The mixing parameter $\alpha = 0.2, 0.3$.*

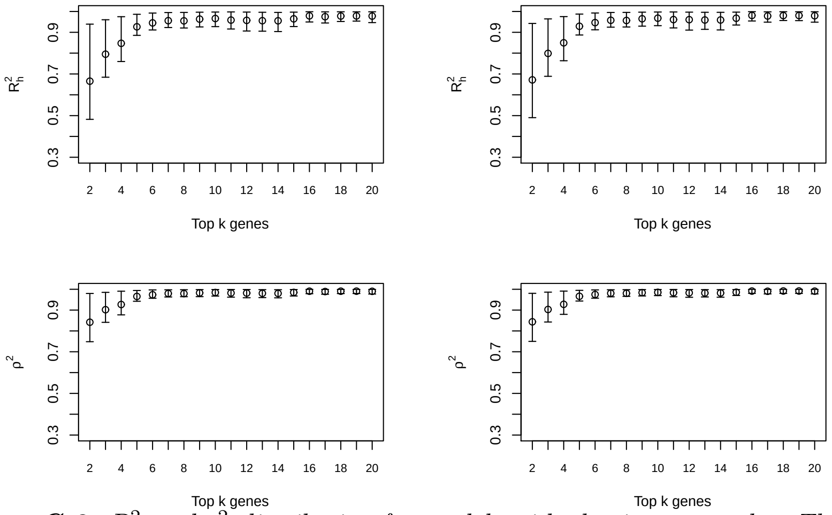


Figure C.3: *R^2_h and ρ^2 distribution for models with elastic net penalty. The mixing parameter $\alpha = 0.4, 0.6$.*

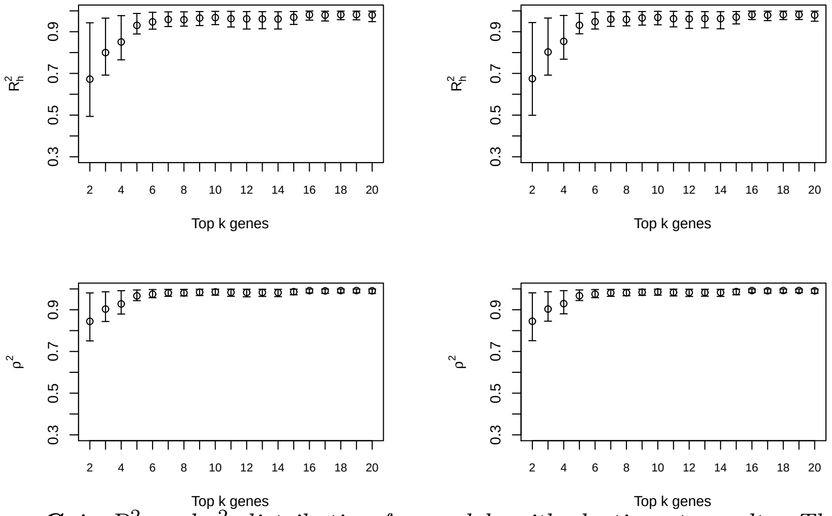


Figure C.4: *R_h^2 and ρ^2 distribution for models with elastic net penalty. The mixing parameter $\alpha = 0.7, 0.8$.*

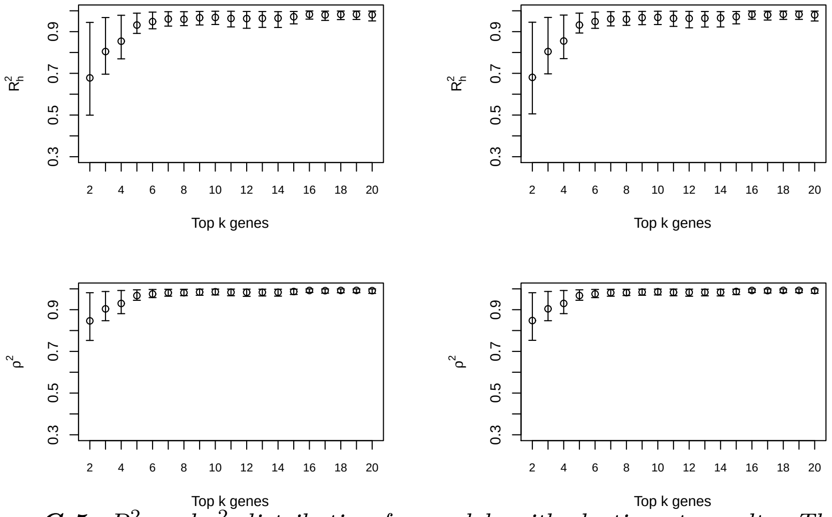


Figure C.5: *R_h^2 and ρ^2 distribution for models with elastic net penalty. The mixing parameter $\alpha = 0.9, 1$.*

Samenvatting

Een biomerker is een eigenschap die objectief gemeten en geëvalueerd wordt als een indicator van normale of pathogene biologische processen of als farmacologische respons op therapeutische of andere interventies Biomarkers Definitions Working Group (2001). In deze thesis werd eerst ingegaan op metabolische biomerkers gericht op de verbetering van bestaande diagnostische procedures, screening instrumenten en modellen met als doel de identificatie van patiënten met een verhoogd risico op de ontwikkeling van borst- en longkanker.

Wereldwijd is borstkanker de meest gediagnosticeerde kanker en voornaamste oorzaak van kankergerelateerde sterfte bij vrouwen. Borstkanker wordt 100 maal meer gediagnosticeerd bij vrouwen dan bij mannen en de meerderheid van vergevorderde borstkankers wordt vastgesteld bij vrouwen ouder dan 50. Momenteel zijn er verschillende complementaire technieken voor de diagnose en opvolging van borstkanker voorhanden. Deze technieken betreffen mammografie, lichamelijk onderzoek, ultrasonografie, MRI en biomerker tests op basis van bloed. Van deze technieken wordt mammografie nog steeds beschouwd als gouden standaard. In hoofdstuk 3 werd nucleic magnetic resonance metabolomics (NMRM) als complementaire methode statistisch onderzocht. NMRM heeft de mogelijkheid om de diagnose van borstkanker te vervroegen tot vóór klinische of radiologische detectie van de symptomen. Verschillende classificatie methoden werden vergeleken om de test op te bouwen. Een classificatie test opbouwen, gebaseerd op slechts een beperkt aantal metabolieten, bleek zeer moeilijk. De belangrijkste oorzaak van deze beperking zal gezocht moeten worden in de verwevenheid van de metabolieten in de biochemische routes.

In hoofdstuk 4 werd dieper ingegaan op longkanker. Longkanker is wereldwijd één van de meest voorkomende kwaadaardige weefsels en wordt meestal pas vastgesteld in een geavanceerd stadium. Dit wordt verklaard door het uitblijven van symptomen tijdens de vroege stadia van de ziekte, waarin behandeling het meest effectief zou zijn. Hierdoor zijn curatieve behandelingen voor longkanker eerder schaars.

Het screenen naar longkanker in een vroeg stadium, vóór patiënten klinische symptomen ontwikkelen, zou patiënten dus ten goede komen door de verbetering van zijn/haar levenskwaliteit en -verwachting. De kosten-baten verhouding wordt gemaximaliseerd wanneer de hoog-risico doelgroep geselecteerd kan worden voor screening.

Screeningstechnieken voor longkanker, zoals radiografie van de borstkas (CXR), sputum cytologie en low-dose computed tomography (LDCT), worden gekenmerkt door hoge vals-positieve ratio's. Dit leidt tot emotionele stress, nodeloze financiële kosten en zelfs gezondheidsrisico's voor gezonde personen. Blootstelling aan onnodige straling, biopsie en chirurgische procedures zijn namelijk gerelateerd aan verhoogde ziekte en sterfte kansen.

Omwille van de hoge vals-positieve ratio, is er zeer veel interesse in de verbetering van de accuraatheid van de huidige risico modellen. Dit kan door de opname van aan longkanker gerelateerde biomerkers in de test, die de selectie van hoog-risico individuen geschikt voor LDCT screening mogelijk maken. Een metabolische signatuur voor longkanker werd ontwikkeld op basis van verschillende classificatie methoden. Deze signatuur liet ons toe om 82% van de longkanker patiënten en 89% van de controle groep correct te classificeren. Bovendien onderzochten we in hoofdstuk 5 het voordeel van metabolische data bovenop epidemiologische en klinische variabelen in een risico model voor longkanker. De resultaten tonen aan dat de toevoeging van metabolische data potentieel tot de verbetering van de identificatie van hoog-risico individuen leidt. Deze individuen kunnen vervolgens geselecteerd worden voor longkanker screening.

In hoofdstukken 7 en 8 beschouwden we het gebruik van biomerkers in de vroege stadia van geneesmiddelenontdekking en ontwikkeling. Deze fasen zijn typisch zeer tijdrovend en duur. Hierdoor wordt een project, gericht op de ontwikkeling van een compound, niet zelden pas stopgezet nadat er al substantiële middelen aan gespendeerd zijn. Daarom is het cruciaal om onbruikbare compounds zo vroeg mogelijk in het ontwikkelingsproces te identificeren om tijd en financiële middelen, in overbodige volgende fases, te kunnen besparen. Hoog-dimensionale biologische data, dewelke snel en relatief goedkoop kunnen vergaard worden, zouden in dit geval nuttig kunnen blijken om het inzicht in de moleculaire basis van ziekten te versnellen. Deze data laten daarenboven ook toe de efficiëntie en toxiciteit van kandidaat geneesmiddelen te onderzoeken.

Het ontdekkings- en ontwikkelingsproces van geneesmiddelen genereert verschillende bronnen van hoog-dimensionale data zoals onder meer high-troughput screening (HTS), chemical structures gene expression en image-based high-content screening (HCS). In onze studie hebben we door middel van structural equation modeling (SEM) gebruik gemaakt van de relatie tussen drie courante data bronnen in geneesmiddelen ontdekking studies (gen expressie data, bio-activiteitsdata en chemische structuur data). Recent, heeft Verbist *et al.* (2015) aangetoond dat het gebruik van transciptomic biomerkers voor de activiteit van een bepaalde compound een vroeg inzicht geeft in het actiemechanisme van een specifieke (of specifieke verzameling van) compound(s). SEM laat toe om verschillende oorzakelijke modellen te onderzoeken, die de relatie beschrijven tussen de chemische structuur en biologische activiteit met gen expressie als de voorgestelde mediator. In het bijzonder, SEM ontbindt het totale effect van de chemische structuur op de biologische activiteitsvariabele in directe en indirecte effecten. Een indirect effect is het effect van de chemische structuur op de bio-activiteitsvariabele via gen expressie. De zo verkregen resultaten lieten ons toe om genen te groeperen in verschillende oorzakelijke structuren gebaseerd op de hoogste a-posteriori kansen.

In hoofdstukken 10 en 11 richten we ons op de evaluatie van surrogaat eindpunten in klinische studies. Klinische studies zijn studies die opgezet worden om de therapeutische effectiviteit van nieuwe geneesmiddelen te evalueren. Tijdens zulke studies wordt gebruik gemaakt van eindpunten die de concrete voordelen voor patiënten zoveel mogelijk weerspiegelen. Onder zulke eindpunten wordt ziekte uitkomst, tijd tot een bepaalde gebeurtenis, dood, enz., gerekend. In vele gevallen vereisen deze studies een zeer groot aantal patiënten en omvatten deze een lange periode van studie. Surrogaat eindpunten zijn daarom zeer interessant voor zowel onderzoekers als patiënten, daar ze de beoordelingstijd aanzienlijk kunnen verminderen terwijl ze toelaten om de effectiviteit van een nieuw medicijn vast te stellen Burzykowski *et al.* (2005).

In sommige gevallen dient de onderzoeker een beroep te doen op een surrogaat eindpunt omdat het klinische eindpunt niet beschikbaar is, moeilijk te meten is of een kostelijke, invasieve of oncomfortabele procedure vereist. Ellenberg and Hamilton (1989) definiëren een surrogaat eindpunt als: ëen eindpunt dat kan gebruikt worden als alternatief voor andere eindpunten in de evaluatie van experimentele behandelingen of andere interventies”. Dit eindpunt is nuttig wanneer het vroeger, gemakkelijker of frequenter kan gemeten worden dan de eindpunten van interesse (het zo genoemd klinisch eindpunt). Maar, het dient uiteraard ook klinisch relevant en biologisch plausibel te zijn. Alvorens een surrogaat eindpunt een klinisch eindpunt kan vervangen in de evaluatie van een experimentele behandeling, moet het formeel ’gevalideerd’ worden. De belangrijkste reden voor validatie van een surrogaat eindpunt is om toe te laten een voorspelling van het behandelingseffect te maken. Deze voorspelling gebeurt in termen van het ware eindpunt gebaseerd op het effect van de behandeling op het surrogaat eindpunt. Uiteraard dient dit met voldoende precisie te gebeuren om een veilig onderscheid te kunnen maken tussen effecten die klinisch van belang zijn en effecten die dat niet zijn. Desondanks is er geen standaard software voorhanden om zulke validatie analyses uit te voeren. In hoofdstukken 10 en 11 introduceren we twee software producten, SAS en R gebaseerd, voor de analyse van surrogaat eindpunten in gerandomiseerde klinische studies. Deze software is ontwikkeld vanuit het standpunt dat de eindgebruikers niet noodzakelijk statistici zullen zijn. Beide producten voorzien gebruiksvriendelijke en eenvoudig te interpreteren standaard output die enkel de hoofdresultaten van de analyse bevat.

Het R product (surrogate Shiny App) kan gebruikt worden op een lokale computer of online, enkel gebruik makend van het shiny cloud platform. Andere cloud platformen, zoals Amazon Web Service of google cloud platform, kunnen ook gebruikt worden. Het is een graphical user interfase (GUI) waardoor de gebruiker niet wordt blootgesteld aan de achterliggende R code.