Marker Methodology With Focus On Hierarchical Outcomes

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Wetenschappen, richting Wiskunde
te verdedigen door

Abel Tilahun

Promotor: Prof. dr. Geert Molenberghs
Co-promotor: Prof. dr. Ariel Alonso

Contents

List of Abbreviations

CGI : Clinician's Global Impression

HAMD : Hamilton Depression Scale

ITA : Information Theoretic Approach

LRF : Likelihood Reduction Factor

MCA : Middle Cerebral Arteries

NCC : Non linear Correlation Coefficient

PANSS : Positive and Negative Syndrome Scale

RF : Random Forest

RTA : Regression Tree Analysis

SCD : Sickel Cell Disease

SPCA : Supervised Principal Component Analysis

SPLS : Supervised Partial Least Squares Analysis

SVM : Support Vector Machine

TCD : Transccranial Dopler

List of Tables

List of Figures

Publications

1. **Tilahun, A.**, Assam, P., Alonso, A., and Molenberghs, G. (2007) Flexible surrogate marker evaluation from several randomized clinical trials with continuous endpoints, using R and SAS. Computational Statistics and Data Analysis, 51, 4152-4163.

2. **Tilahun, A.**, Assam, P., Alonso, A., and Molenberghs, G. (2008) Information-theory based surrogate marker evaluation from several randomized clinical trials with binary endpoints, Using SAS. Journal of Biopharmaceutical Statistics, 18, 326–341.

3. **Tilahun, A.**, Maringwa, J. ,Geys, H., Alonso, A., Raeymaekers, L., Molenberghs, G., Kieboom, G., Drinkenburg, P., Bijnens, L. (2009). Investigating Association Between Behavior, Corticosterone, Heart Rate, and Blood Pressure in Rats Using Surrogate Marker Evaluation Methodology. Journal of Biopharmaceutical Statistics, 19, 001-017.

4. Assam, P., **Tilahun, A.**, Alonso, A., and Molenberghs, G. (2007) Information-theory based surrogate marker evaluation from several randomized clinical trials with continuous true and binary surrogate endpoints. Clinical Trials, 04, 587-597.

5. Molenberghs, G., Burzykowski, T., Alonso, A., Assam, P., **Tilahun, A.**, and Buyse, M. (2008). The meta-analytic framework for the evaluation of surrogate endpoints in clinical trials. Journal of Statistical Planning and Inference, 138, 432-449.

6. **Tilahun, A.**, Lin, D., Shkedy ,Z., Geys, H., Alonso, A., Molenberghs, G., Kieboom, G., Drinkenburg, P., Bijnens, L. (2008). Genomic biomarkers for depression: feature specific and joint biomarkers. Statistics in Biopharmaceutical Research. Manuscript accepted for publication.

7. Assam, P., **Tilahun, A.**, Alonso, A., and Molenberghs, G. (2008). Using Earlier Measures in a Longitudinal Sequence as Potential Surrogate for a Later One. Computational Statistics and Data Analysis. Manuscript submitted for publication.

8. Molenberghs,G. Burzykowski,T., Alonso,A., Assam, P., **Tilahun,A.**, Buyse,M. (2008). Unified Framework for the Evaluation of Surrogate Endpoints in Mental-Health Clinical Trials. Manuscript submitted for publication

1

Introduction

1.1 Introduction

In recent times, considerable success has been achieved in the development of new drugs for wide ranging diseases affecting human race. The process of drug development has evolved into an extremely complex procedure. The first step in the process of drug development is identifying promising compounds. Once a compound has been isolated for further scrutiny, it enters a rigorous testing and evaluation stage, the so-called pre-clinical phase. This stage is designed to assess the chemical properties of the new drug as well as to determine the steps for synthesis and purification. In this stage, the toxicological and pharmacological effects of the drug are evaluated through in-vitro and in-vivo animal testing. If a compound is thought to be safe and effective as a chemical agent, then it will be approved to move to a clinical trial stage. Once approved for clinical studies, a three-phase process begins where safety and efficacy are continually assessed with increased scrutiny and an increasing patient population. At

all stages of drug development, the efficacy of the drug is assessed through its effect on clinically meaningful variables that are sensitive to detect treatment effects. However, such variables might increase the complexity and/or the duration of a clinical trial, either because they are costly, difficult to measure, require a long follow up time, or require a large sample size due to low incidence of the event. These problems might be avoided through replacing the true endpoints by other ones, measured earlier or in a more convenient fashion, which here will be termed as surrogate endpoints.

The Biomarkers Definitions Working Group gives the following definitions for clinical endpoint, biomarker and surrogate endpoint respectively. A *clinical endpoint* is a characteristic or variable that reflects how a patient feels, functions, or survives. A *biomarker* is a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention. Depending on their intended use, biomarkers are further classified as therapeutic and prognostic. A therapeutic biomarker which is also called a predictive biomarker is a biomarker that informs the treatment effect on the clinical endpoint. A prognostic biomarker on the other hand, is a biomarker that informs the clinical outcome, independent of treatment. It provides information about natural course of the disease in individual with or without treatment under study. A *surrogate endpoint* is a biomarker that is intended to substitute a clinical endpoint. A surrogate endpoint is expected to predict clinical benefit, harm, or lack thereof (Biomarkers Definition Working Group 2001). A surrogate endpoint, as compared to true endpoints like survival, can often be measured earlier, easier, and more frequently and is less subject to competing risks. To this end, surrogate endpoints come into play in a number of contexts in place of the endpoint of interest, referred commonly to as the true or clinical endpoint. One important reason for the present interest

in surrogate endpoints is the advent of a large number of biomarkers that closely reflect the disease process. An increasing number of new drugs have a well-defined mechanism of action at the molecular level, allowing drug developers to measure the effect of these drugs on the relevant biomarkers (Ferentz 2002). There is increasing public pressure for new, promising drugs to be approved for marketing as rapidly as possible, and such approval will have to be based on biomarkers rather than on some long-term clinical endpoint (Lesko and Atkinson 2001). If the approval process is shortened, there will be a corresponding need for earlier detection of safety signals that could point to toxic problems with new drugs. It is a safe bet, therefore, that the evaluation of tomorrow's drugs will be based primarily on biomarkers, rather than on the longer-term, harder clinical endpoints that have dominated the development of new drugs until now. It is therefore best to use *validated* surrogates, though one needs to reflect on the precise meaning and extent of validation (Schatzkin and Gail 2002). Like in many clinical decisions, statistical arguments will play a major role, but ought to be considered in conjunction with clinical and biological evidence. For a biomarker to be used as a "valid" surrogate, a number of conditions must be fulfilled. The ICH Guidelines on Statistical Principles for Clinical Trials state that "In practice, the strength of the evidence for surrogacy depends upon (i) the biological plausibility of the relationship, (ii) the demonstration in epidemiological studies of the prognostic value of the surrogate for the clinical outcome and (iii) evidence from clinical trials that treatment effects on the surrogate correspond to effects on the clinical outcome" (International Conference on Harmonisation 1998).

Ideally, there should be guidelines to declare a marker a useful surrogate for a clinical endpoint. Several methods have been suggested for the formal evaluation of surrogate markers. Some of these methods are based on a single trial while others,

which are gaining momentum in the present day, are based on meta-analytic concepts. The first formal approach to evaluate markers is attributed to Prentice (1989), who has given a definition of surrogate endpoints, followed by a series of operational criteria to check whether the definition is fulfilled. Freedman, Graubard, and Schatzkin (1992) have supplemented the hypothesis-testing-based criteria, which necessarily depend on the power of the test performed, with a quantity to be estimated. They suggested the use of the so-called *proportion of treatment effect explained* (PTE) by the surrogate as an alternative means of validation. The PTE faces serious drawbacks, against the background of which Buyse and Molenberghs (1998) have suggested the use of another quantity, the *relative effect* (RE), defined as the ratio of the treatment effect on the true endpoint to that on the surrogate endpoint. In turn, the RE is open to severe criticism as well. First, the RE's confidence intervals, like the ones for PTE, tend to be wide. While this could in principle be overcome, there is a second, more severe problem in the sense that the RE is useful for prediction of the true treatment effect from the surrogate treatment effect only when the relationship between both is multiplicative. This may be rightfully viewed as restrictive and, in any case, cannot be verified from a single trial. Moving away from single trial based methods leaves us with the meta-analytic alternative. With in the context of meta-analytic approach, two possible views are possible when evaluating a marker. The first deals with the individual patient level and is connected with the biological pathway from the surrogate to the true endpoint and is termed as individual level surrogacy. This, however, does not necessarily mean a marker is useful to capture the treatment effect in the setting of a clinical trial. Therefore, a second view, focusing on the treatment effect is necessary and possible (Fleming and DeMets 1996). Precisely, this level quantifies the association between the treatment effects on the marker and the clinical endpoint which

is termed as trial level surrogacy. Buyse *et al* (2000) and Burzykowsky *et al* (2004), among others, have presented a meta-analytic modeling framework, within which both forms of validation can be undertaken. Recently, pre-clinical microarray experiments have become an increasingly common laboratory tools to investigate the activity of thousands of genes simultaneously and their response to a certain treatment. The main objectives for microarray studies vary from application to application, but most of them revolve around identifying a group of genomic biomarkers for a particular clinical outcome of interest. Note that, we can observe two main differences concerning the evaluation of biomarkers, between the clinical trials and the microarray setting. The first difference is that the surrogacy problem in the microarray setting consists of thousands of potential biomarkers and one response which in most clinical trials is not the case. The second difference is that surrogacy in the microarray setting is needed to be tested and not just to be evaluated as has been the case in the clinical trial setting surrogate marker validation although some surrogate marker validation exercises have testing procedures. Appreciating these differences however, we still can establish analogies between the individual and trial level surrogacy concepts and the prognostic and predictive biomarkers respectively. The selection of prognostic biomarkers can be carried out using existing methods within the the surrogate marker methodology literature with no or little modifications. However, since most microarray experiments are based on a single trial, selection of therapeutic biomarkers might not be easily carried out with the existing methods developed for surrogate marker validation at a trial level which calls for other alternative approaches.

1.2 Structure of the Thesis

The main focus of this thesis is developing marker methodology for hierarchical outcomes. More specifically, emphasis will be given to the statistical validation of surrogate endpoints. The thesis is organized in to two main parts. The first part deals with the statistical validation of surrogate endpoints of various types while the second part focuses on the selection and evaluation of genomic biomarkers. The methods summarized in this thesis will be applied to real life data sets. It is therefore logical to start the thesis by laying out the different case studies used to demonstrate the practical use of the methods. Following the motivational case studies, we outline a concise review of the meta-analytic approach to surrogate marker validation for two cross-sectional normally distributed endpoints. Here, we raise computational and other issues which have not been addressed before. This chapter uses the first publication in the list given earlier as the main reference. The information-theoretic unification, which enables the validation of normally as well as non-normally distributed outcomes will follow the meta-analytic approach. The description of this approach will then be followed by three chapters focused on the assessment of its performance for the cases of mixture of binary and continuous, two binary endpoints and a mixture of time-to-event and cross-sectional endpoints respectively. In the former two of the stated three chapters, simulation studies will be carried out to compare the performance of the information-theoretic approach against a probit formulation. The later mimics the meta-analytic approach designed for two normally distributed endpoints by making use of latent variable formulation. These two chapters use publications 2 and 4 respectively as main references. On the third of these three chapters, three information theory based methods will be compared with one another on a simulation setting where the proportional hazard assumption is violated. These information theory based approaches

will be followed by a method for a mixture of longitudinal and cross-sectional endpoints. The methods, which were originally designed for the case of two longitudinal endpoints, will be modified to accommodate a mixture of cross-sectional and longitudinal outcomes which will be used interchangeably as surrogate and true endpoints. The main reference for this chapter is publication list 3. Once the methods are tuned to deal with a mixed longitudinal and cross-sectional outcome, they will be used to select optimal number of repeated measures in a new surrogate marker setting where both the true and surrogate endpoints come from a single continuous longitudinal sequence. The cumulated earlier measures of the sequence will be used as surrogates for later measurements. The approach takes the cost of waiting time and inclusion of extra time points by devising an objective function. The same methodology then will be used for a binary longitudinal sequence. In both cases simulated data will be used to aid understand the application of the methods under different scenarios. The main reference for this chapter is publication 7. This chapter will wind up the validation of surrogate endpoints part of the thesis.

The second part of the thesis starts with the method of selection and evaluation of genomic biomarkers. Different statistical methods will be utilized to select and evaluate feature-specific as well as joint biomarkers for depression. The methods in this chapter mainly focus on selecting genes which exhibit a linear association with the clinical outcome and use publication 6 as main reference. The assumption of linearity might be restrictive and hence alternative methods for gene selection will be the concern of the chapter that follows. Different parametric and non-parametric models will be used to select genomic biomarkers. The two chapters on the selection and evaluation of biomarkers focus on the so called prognostic biomarkers. The fact that microarray experiments are single trial in nature prohibited the use of trial level

surrogacy measure to quantify the association between the treatment effect on the potential biomarker and the clinical outcome. Thus a Bayesian approach was entertained through which an R-square type measure similar to the trial level surrogacy is given. This chapter winds up the second part of the thesis and will be followed by a general discussion and future research chapter. Some analytical derivations and statistical software will be given in the subsequent appendices.

2

Motivational Case Studies

2.1 Motivating Case Study

In this chapter we will present the case studies used to elaborate the different methodologies summarized in this thesis. Unless stated, for the evaluation of surrogate endpoints part, Z represents a binary treatment indicator, S and T stand for surrogate and true endpoints respectively. For the selection and evaluation of biomarkers part, the same notations represent treatment , biomarker and clinical outcome respectively.

2.1.1 The Age-related Macular Degeneration Study

This is a clinical trial involving patients with age-related macular degeneration (ARMD), a condition in which patients progressively lose vision. Overall, 1186 patients from 114 sites participated in the trial. Patients' visual acuity was assessed using standardized vision charts displaying lines of five letters of decreasing size, which patients had to read from top to bottom. The visual acuity was measured by the a visual

acuity score. The sites in which patients were treated will be considered as units of analysis. Some of the sites participating in the trial enrolled patients only to one of the two treatment arms. These sites were excluded from considerations. A total of 82 sites were thus available for analysis, with a number of individual patients per center ranging from 2 to 19 (424 patients overall).

2.1.2 A Meta-analysis of Five clinical Trials in Schizophernia

The data come from a meta-analysis of five double-blind randomized clinical trials, comparing the effects of risperidone to conventional antipsychotic agents for the treatment of chronic schizophrenia. Schizophrenia has long been recognized as a heterogeneous disorder with patients suffering from both 'negative' and 'positive' symptoms. Negative symptoms are characterized by deficits in cognitive, affective and social functions for example poverty of speech, apathy and emotional withdrawal. Positive symptoms entail more florid symptoms such as delusions, hallucinations and disorganized thinking, which are superimposed on mental status (Kay, Fiszbein, and Opler 1987). Several measures can be considered to asses a patient's global condition. Clinician's Global impression (CGI) is generally accepted as an admittedly subjective clinical measure of change. Here, the change of CGI from baseline will be considered as the true endpoint. It is scored on a 7-grade scale used by the treating physician to characterize how well a subject has improved since baseline. Another useful and sufficiently sensitive assessment scales is the Positive and Negative Syndrome Scale (PANSS) (Kay, Opler, and Lindenmayer 1988). The PANSS consists of 30 items that provide an operationalized, drug-sensitive instrument, which is highly useful for both typological and dimensional assessment of schizophrenia (Kay, Opler, and Lindenmayer 1988). We will use the change from baseline in PANSS as our surrogate

endpoint. The data contains five trials and in all trials, information is available on the investigators that treated the patients. This information is helpful to define group of patients that will become units of analysis.

2.1.3 A Study on Stress Related Disorders

The data come from a preclinical rat experiment on a compound under development for stress-related disorders. The objective of the experiment was to identify the effect of the compound on stress hormones and a series of physiological variables. In the experiment, stress is induced by forcing a rat to swim for 15 minutes in a bath of 20 cm high lukewarm water of 25 degrees Celsius, according to a protocol as described by De Groote and Linthorst (2007). The experiment was designed according to a latin square crossover design with 4 periods and 4 treatment groups (vehicle without stress, vehicle with stress, compound without stress, compound with stress). Forty-five minutes after randomization, the rats were injected with either a vehicle or the compound under consideration. Ten minutes later, half of the rats injected with the vehicle and half of the rats injected with the compound, were subjected to so-called "swim stress", also depending on group membership. For all eight animals, measurements were analyzed in order to quantify their stress level. Telemetry measurements (such as heart rate and blood pressure) were recorded continuously and averaged every 5 minutes. Seventeen blood samples were taken in a fully automated way, leaving the animals completely undisturbed and following a well-defined scheme to sample blood plasma from which corticosterone (CORT) was later extracted and quantified. And finally, rats were also screened for their behavior in a 10 minutes interval by means of a video monitor. For each rat, the percentage of time it has been active is thus determined. The recording of behavior was done twice: a first time at 25 minutes after injection and a second

time at 50 minutes after the end of the swim stress.

2.1.4 A Meta-analysis of Ten Clinical Trials in Acute Migraine

This is a meta-analysis of 10 early phase trials assessing the efficacy of several therapies for the treatment of acute migraine crises. Each trial was placebo-controlled and aimed at evaluating one of three experimental treatments. Two trials also included an active control arm. Overall, 801 patients were available, recruited over 38 different centers, with between 1 and 86 patients enrolled per center. Severity of headache and migraine-related symptoms were measured prior to and at several occasions after the dose administration. Severity was rated on a four-grade intensity scale (0 =no, 1 =mild, 2 =moderate, 3 =severe). Clinically relevant endpoints for efficacy included pain-free (pain score=0) and pain relief (pain score<= 1) two hours post-dose. The main goal is to identify what symptoms are typically associated with migraine episodes, such as, for example, nausea, vomiting, increased sensitivity to light, i.e., photophobia, as well as to sound, i.e., phonophobia.

2.1.5 Stroke Study on Children with Sickle Cell Disease

The data results from a clinical trial involving 2323 children with sickle cell disease (SCD). Children with this disease are susceptible to having a stroke at some time in their lives. One measure that is commonly used for risk estimation for stroke is the so called Transcranial Doppler, or TCD. This measure provides a simple risk estimation for stroke in children with SCD. The unit of measure is velocity in centimeters per second, estimated by Doppler ultrasound, from the higher of the 2 middle cerebral arteries (MCAs), and it represents a physiological marker of the speed of blood flow in the artery. Blood flow velocity can be increased by reduced lumen diameter, as in stenosis or vasospasm, and/or by increased volume flow through the artery. In this

study diastolic and systolic TCD velocities are measured for each patient repeatedly over time although the majority of the patients have only one measurement. In addition to the TCD velocity measures, the time to first stroke, the age and gender of the patients is also recorded. The main objective of the study is to asses if the measured velocities can be used as possible surrogates for the time to stroke.

2.1.6 A Case Study in Depression

A growing number of theories are tested to elucidate the cognitive, molecular and psycho-physiological underpinnings of brain dynamics in clinical depression. A recent approach to the study of depression has been in molecular patient profiling and neuro-degenerative risk factors. However, these theories are usually tested using single variables, e.g., depression and alpha, depression and heart rate. This study aims to combine the cognitive, psycho-physiological and molecular profiling variables, usually studied in isolation, in depression. It is envisaged that the integration of commonly studied indices of depression and molecular patient profiling offer the chance of better understanding the biomarkers of major depression and that these biomarkers may be applied to develop and guide more efficient drug development and testing programs. One way of measuring the severity of depression is through the use of the Hamilton Depression Scale (HDS or HAMD). It is a test measuring the severity of depressive symptoms in individuals, often those who have already been diagnosed with a depressive disorder. The HAMD is used to assess the severity of depressive symptoms present in both children and adults. It is oftentimes used as an outcome measure for depression in evaluations of antidepressant psychotropic medications and is a standard measure of depression used in research of the effectiveness of depression therapies and treatments. It can be administered, for example, prior to the start of

medication and then again during follow-up visits, so that medication dosage can be changed in part based on the patient's test score. The overall objective of this trial is to identify possible biomarkers for depression. 31 patients had been followed up 4–6 weeks after commencement of treatment with antidepressants. Also, 15 control individuals had been followed up 4–6 weeks after their first visit. However, of the 31 depressed patients which had measurements after treatment, after removing the missing values in metabolites and gene expression, complete information was available for 14 patients for the analysis of metabolites and 19 patients for analysis of gene expression. There were in total 17502 genes and 269 metabolites. In addition to the gene and metabolite measures, storage time of the samples, age, gender, and season when the samples were collected and whether or not the subjects fasted were recorded for each patient.

2.1.7 Behavioral Study in Rats

This is a randomized pre-clinical experiment on behavioral study. For each subject information is available about a treatment group, a clinical endpoint, and gene-expression. The aim of the analysis is to identify genes, which can be used as genomic biomarkers, i.e., can be used in order to predict the clinical outcome, and/or are related to treatment. The behavioral study is an experiment for compulsive checking disorder. The disorder is induced by treating the animals with a chemical compound. Twenty-four rats were randomized equally into two groups. The first group received the active compound (T), while the second was given a solvent (P). After receiving treatment, the rats had to complete an open field test. The data indicated how often a rat went back to its home base in the open field. The home base was defined as the area where the animal spent the longest cumulative time. Animals showing the

signs of the disorder (meaning that the compound has successfully induced the symptoms, characteristic of the disorder) were characterized by displaying, for example, an increased frequency of visits to the home base. The clinical outcome of the experiment is the total number of visits the rats made to the home base. After completion of the experiment, a sample was taken from the thalamus part of the brain of the rats and used to obtain microarray measurements for 5644 genes. The data, from the Affymetrix Rat Genome U34A arrays, were summarized using the Affymetrix microarray suite software (MAS) Version 5.0, and normalized using quantile normalization. The aim of the study was to investigate whether one could identify gene changes that were correlated with the compound, i.e., the symptoms of the disorder, and thereby indirectly discover genes that are involved in this disease.

Part I

Validation of Surrogate Endpoints

3

Meta-Analytic Framework of Surrogate Marker Validation

The evaluation of surrogate endpoints can be carried out with either a single trial or within a meta-analytic frame work. Although the single trial based methods are relatively easy in terms of implementation, they are surrounded with difficulty as there evidently is replication at the patient level, but not at the level of the treatment effect which prohibits the computation of the trial levels surrogacy which in most situations is the most important part of the evaluation process. In light of the difficulties surrounding the single trial based methods, the use of the meta-analytic approach becomes imperative. The meta-analytic approach has been originally formulated for two continuous, normally distributed outcomes, and extended in the meantime to a large set of outcome types, ranging from continuous, binary, ordinal, time-to-event, and longitudinally measured outcomes. In this chapter we briefly review the methodology for the case of normally distributed outcomes, followed by the simplified modeling

approaches suggested by Tibaldi *et al* (2003).

3.1 Meta-Analytic Approach for Continuous Outcomes

The meta-analytic approach is based on a hierarchical two-level model. Both a fixed-effects and a random-effects view can be taken. Let T_{ij} and S_{ij} be the random variables denoting the true and surrogate endpoint for the jth subject in the ith trial, and let Z_{ij} be the indicator variable for treatment. First, consider the following fixed-effects models:

$$S_{ij} = \mu_{Si} + \alpha_i Z_{ij} + \varepsilon_{Sij}, \tag{3.1}$$

$$T_{ij} = \mu_{Ti} + \beta_i Z_{ij} + \varepsilon_{Tij}, \tag{3.2}$$

where μ_{Si} and μ_{Ti} are trial-specific intercepts, α_i and β_i are trial-specific effects of treatment Z_{ij} on the endpoints in trial i, ε_{Sij} and ε_{Tij} are correlated error terms, assumed to be zero-mean normally distributed with covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ & \sigma_{TT} \end{pmatrix}. \tag{3.3}$$

A classical hierarchical, random-effects modeling strategy can also be adopted in the following manner:

$$S_{ij} = \mu_S + m_{Si} + \alpha Z_{ij} + a_i Z_{ij} + \varepsilon_{Sij}, \tag{3.4}$$

$$T_{ij} = \mu_T + m_{Ti} + \beta Z_{ij} + b_i Z_{ij} + \varepsilon_{Tij}. \tag{3.5}$$

Here, μ_S and μ_T are fixed intercepts, α and β are fixed treatment effects, m_{Si} and m_{Ti} are random intercepts, and a_i and b_i are random treatment effects in trial i for the surrogate and true endpoints, respectively. The random effects (m_{Si}, m_{Ti} ,a_i , b_i) are assumed to be mean-zero normally distributed with covariance matrix

$$D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ & d_{TT} & d_{Ta} & d_{Tb} \\ & & d_{aa} & d_{ab} \\ & & & d_{bb} \end{pmatrix}. \tag{3.6}$$

The error terms ε_{Sij} and ε_{Tij} follow the same assumptions as in the fixed effects models. In addition, following the fixed effect models (3.1) and (3.2), we can specify

$$\begin{pmatrix} \mu_{Si} \\ \mu_{Ti} \\ \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} m_{Si} \\ m_{Ti} \\ a_i \\ b_i \end{pmatrix}, \tag{3.7}$$

where the second term on the right hand side of (3.7) is assumed to follow a zero-mean normal distribution with covariance matrix (3.6). After fitting the above models, the surrogate marker evaluation is captured by means of two quantities, the trial-level and individual-level R^2, respectively. The former quantifies the association between the treatment effects on the true and surrogate endpoints at the trial level. The latter measures the association at the level of the individual patient and after adjustment for the treatment effect. The former is given by:

$$R^2_{\text{trial}} = R^2_{b_i|m_{Si},a_i} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}}. \tag{3.8}$$

The above quantity is unitless and, at the condition that the corresponding variance-covariance matrix is positive definite, lies within the unit interval. The models (3.1) and (3.2) can be referred to as the full fixed effects models and it is possible to simplify them. The reduced versions of these models are obtained by replacing the fixed trial-specific intercepts, one for each endpoint, common to all trials. The reduced mixed effect models result from removing the random trial-specific intercepts m_{Si} and m_{Ti}

from models (3.4) and (3.5). The R^2 for the reduced models is then calculated as follows:

$$R^2_{\text{trial(r)}} = R^2_{b_i|a_i} = \frac{d^2_{ab}}{d_{aa}d_{bb}}. \tag{3.9}$$

A surrogate could thus be adopted when R^2_{trial} is sufficiently large. Arguably, rather than using a fixed cutoff above which a surrogate would be adopted, there always will be clinical and other judgement involved in the decision process. The R^2_{indiv} is based on (3.3) and takes the following form:

$$R^2_{\text{indiv}} = R^2_{\varepsilon_{Ti}|\varepsilon_{Si}} = \frac{\sigma^2_{ST}}{\sigma_{SS}\sigma_{TT}} \tag{3.10}$$

Note that, here, trial is considered as experimental unit which can be replaced by center, investigator or any other suitable experimental unit, depending on the nature of the study conducted. The issue of the unit of analysis is discussed in Section 3.3.1 and has been thoroughly studied by Cortiñas *et al* (2004).

3.2 Simplified Modeling Strategies

Though the hierarchical modeling discussed earlier is elegant, it often poses a considerable computational challenge (Burzykowski, Molenberghs, and Buyse 2005). To address this problem, Tibaldi *et al* (2003) suggested several simplifications of the above strategy, briefly outlined here. These authors considered three possible dimensions along which simplifications can be undertaken. The first dimension is what they called the *trial dimension*. This dimension provides a choice between treating the trial-specific effects as fixed or random. If the trial-specific effects are chosen to be fixed, a two-stage approach is adopted. The first-stage model will take the form (3.1) and (3.2) and at the second stage, the estimated treatment effect on the true endpoint is regressed on the treatment effect on the surrogate and the intercept associated with

the surrogate endpoint as

$$\widehat{\beta}_i = \widehat{\lambda}_0 + \widehat{\lambda}_1 \widehat{\mu}_{Si} + \widehat{\lambda}_2 \widehat{\alpha}_i + \varepsilon_i. \tag{3.11}$$

The trial-level $R^2_{\text{trial(f)}}$ then is the coefficient of determination obtained by regressing $\widehat{\beta}_i$ on $\widehat{\mu}_{Si}$ and $\widehat{\alpha}_i$, whereas $R^2_{\text{trial(r)}}$ is obtained from the coefficient of determination resulting from regressing $\widehat{\beta}_i$ on $\widehat{\alpha}_i$ only. The individual-level value is calculated as in (3.10) using the estimates from (3.3). Note here that (r) and (f) are indicators that the trial-level association is obtained based on the reduced or the full model respectively. The second option is to consider the trial-specific effects as random. How one then proceeds is related to the so-called *endpoint dimension*. Indeed, though natural to assume the two endpoints correlated, this choice does increase computational complexity. The desirability to accommodate the bivariate nature of the outcome is associated with interest in R^2_{indiv}, which is in some cases of secondary importance. At the same time, there is also a possibility to estimate it by making use of the information-theoretic approach which will be discussed in the next chapter. Depending on the choice made on the endpoint dimension, two directions can be followed. The first one involves a two-stage approach with univariate models (3.4) and (3.5) at first stage. A second stage model consists of a normal regression with the random treatment effect on the true endpoint as response and the random intercept and random treatment effect on the surrogate as covariates. The second direction is based on a full random effects (hierarchical) model as discussed in Section 3.1. If in the trial dimension, the trial-specific effects are considered to be fixed, then models (3.1) and (3.2) are fitted separately. Similarly, if the trial-specific effects are considered random, then models (3.4) and (3.5) are fitted separately, i.e., the corresponding error terms in the two models are assumed to be independent. Except when a bivariate mixed-modeling approach is followed, there is a need to adjust for the heterogeneity

in the amount of information contributed by the various trials. This is the subject of the *measurement error dimension*. One can either ignore this phenomenon or weight the trial-specific contributions according to trial size. This gives rise to a weighted linear regression model (3.11) in the second stage.

3.3 Computational Considerations

In this section, we will address a number of computational issues and considerations, such as the choice of the unit for analysis, the effect of treatment coding, the possible occurrence of ill-conditioned and non-positive definite variance-covariance matrices.

3.3.1 Unit of Analysis

A cornerstone of the meta-analytic method is the choice of the unit of analysis such as, for example, trial, center, or investigator. This choice may depend on practical considerations, such as the information available in the data set at hand, experts' considerations about the most suitable unit for a specific problem, the amount of replication at a potential unit's level, and the number of patients per unit. From a technical point of view, the most desirable situation is where the number of units and the number of patients per unit is sufficiently large. This issue has been discussed by Cortiñas *et al* (2004).

3.3.2 Treatment Coding

When there is a treatment variable included in the model two choices need to be made at analysis time. First, the treatment variable can be considered continuous or discrete (a class variable). Second, when a continuous route is chosen, it is relevant to reflect on the actual coding, $0/1$ and $-1/+1$ being the most commonly encountered ones. For models with treatment occurring as fixed effect only, these choices are essentially

irrelevant, since all choices lead to an equivalent model fit, with parameters from one situation to another connected by simple linear transformations. Note that this is not the case, of course, for more than three treatment arms. However, of more importance for us here is the impact the choices can have on the hierarchical model. Indeed, while the marginal model resulting from (3.4)–(3.5) is invariant under such choices, this is not true for the hierarchical aspects of the model, such as, for example, the R^2 measures derived at the trial level. Indeed, a $-1/+1$ coding ensures the same components of variability operate in both arms, whereas a $0/1$ coding, for a positive definite D matrix, forces the variability in the experimental arm to be greater than or equal to the variability in the standard arm. Both situations may be relevant, and therefore it is of importance to illicit views on this issue from the study's investigators.

3.3.3 Ill-Conditioned Variance-Covariance Matrix

When the full bivariate random effect is used, the R^2_{trial} is computed from the variance-covariance matrix (3.6). It is sometimes possible that this matrix be ill-conditioned and/or non-positive definite. In such cases, the resulting quantities computed based on this matrix might not be trustworthy. One way to asses the ill-conditioning of a matrix is by reporting its condition number, i.e., the ratio of the largest over the smallest eigenvalue. A large condition number is an indication of ill-conditioning. The most pathological situation occurs when at least one eigenvalue is equal to zero. This corresponds to a positive semi-definite matrix, which occurs, for example, when a boundary solution is obtained. Thus, in the validation process, it is necessary to check the D matrix for absence or presence of these issues.

3.4 Simulation Study

To asses the impact of using an incorrect treatment coding, a small simulation involving 12 different combinations of trial size and number of individuals per trial has been performed. The data were generated based on the following model:

$$S_{ij} = 45 + m_{Si} + (3 + a_i)Z_{ij} + \varepsilon_{Sij}, \tag{3.12}$$

$$T_{ij} = 50 + m_{Ti} + (5 + b_i)Z_{ij} + \varepsilon_{Tij}. \tag{3.13}$$

Here a_i and b_i are random treatment effects in trial i for the surrogate and true endpoints, respectively. The random effects (m_{Si}, m_{Ti} ,a_i , b_i) are assumed to be mean-zero normally distributed with covariance matrix

$$D = \begin{pmatrix} 3 & 2.4 & 0 & 0 \\ 2.4 & 3 & 0 & 0 \\ 0 & 0 & 3 & 2.7 \\ 0 & 0 & 2.7 & 3 \end{pmatrix}. \tag{3.14}$$

The error terms ε_{Sij} and ε_{Tij} are assumed to be zero mean random variables with variance-covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} 3 & 2.4 \\ 2.4 & 3 \end{pmatrix}. \tag{3.15}$$

The number of trials was fixed to either 10, 20 or 50 with each trial involving either 10,20,40 or 60 subjects giving rise to 12 different scenarios. For each combination, 100 datasets were generated for both treatment codings. The datasets were then analyzed with the correct treatment coding, i.e., the treatment coding with which the data were generated, as well as with the opposite coding. For each case the median condition number and the percentage of positive definite variance-covariance matrices are counted. The results of these simulations are displayed in Tables 3.1 and 3.2.

The simulation has revealed that, for a small number of analysis units and/or a small number of subjects per analysis unit, the wrong treatment coding could result in a high degree of uncertainty in the resulting variance-covariance matrix. For the 0/1 coding, the effect is noticed even when the correct coding was followed to do the analysis, i.e. there was high degree of uncertainty even when the data were analyzed with the correct 0/1 coding for small sample sizes. The effect, however, seems to vanish with increasing repetition of the unit of analysis and number of subjects per unit of analysis. If we consider a median condition number of 100 as an arbitrary cutoff value, we notice that we require a minimum of 20 trials to achieve a condition number less than 100 for 0/1 coding. This number, however, reduces to only 10 trials to reach a condition number less than 100 for $-1/+1$ coding. With respect to the positive-definitness of the variance-covariance matrix, the percentage of positive-definite matrices increases with increase in the sample size for both treatment coding schemes. However, the $-1/+1$ produced relatively a higher percentage of positive definite matrices even for small samples as compared to the 0/1 coding where the percentage of positive definite matrices is low even for moderately higher sample sizes. Based on the results of this simulation, it seems reasonable to consider the $-1/+1$ treatment coding and chose a reasonable unit of analysis to avoid the numerical problems and achieve positive definiteness in the variance-covariance matrix.

3.5 Application to the Case Studies

Two case studies introduced in the motivating case studies chapter, Sections (2.1.1) and (2.1.2) concerning schizophrenia and Age Related Macular Degeneration are analyzed here. Let us start with the schizophrenia study. Here, trial seems the natural unit of analysis. Unfortunately, the number of trials is not sufficient to apply the

Table 3.1: *Simulation results for* $-1/1$ *treatment coding.*

	simulation strategy		**% positive-definite**		**median condition number**	
simulation #	# trials	# subjects	correct	incorrect	correct	incorrect
1	10	10	42	41	3.44E+16	3.71E+17
2	10	20	66	65	178.00	403.10
3	10	40	91	91	78.36	172.86
4	10	60	98	98	81.23	158.39
5	20	10	90	90	52.43	138.62
6	20	20	97	98	43.33	102.34
7	20	40	100	100	34.87	101.55
8	20	60	100	100	32.97	84.41
9	50	10	100	100	27.55	84.56
10	50	20	100	100	26.54	80.64
11	50	40	100	100	24.28	75.01
12	50	60	100	100	24.92	72.86

full meta-analytic approach. The use of trial as unit of analysis for the simplified methods might also entail problems. The second stage involves a regression model based on only five points, which might give overly optimistic or at least unreliable R^2 values. The other possible unit of analysis for this study is 'investigator'. There were 176 investigators who each treated between 2 and 60 patients. The use of investigator as unit of analysis is also surrounded with problems. Although a large number of investigators is convenient to explain the between investigator variability, because there are few patients per investigators for some investigators, the resulting within-unit variability might not be estimated correctly.

The basic meta-analytic approach and the corresponding simplified strategies have

Table 3.2: *Simulation results for 0/1 treatment coding.*

	simulation strategy		**% positive-definite**		**median condition number**	
simulation #	# trials	# subjects	correct	incorrect	correct	incorrect
1	10	10	10	10	5.44E+16	3.71E+17
2	10	20	25	25	4.09E+16	9.03E+16
3	10	40	57	58	304.05	1184.91
4	10	60	68	68	196.44	436.48
5	20	10	38	38	2.79E+16	6.6E+16
6	20	20	62	62	136.94	560.39
7	20	40	89	89	51.17	186.94
8	20	60	97	97	38.32	166.40
9	50	10	70	71	67.83	225.77
10	50	20	93	93	34.18	158.24
11	50	40	100	100	27.31	134.00
12	50	60	100	100	25.56	127.24

been applied to this data set. The results are displayed in Table 3.3. Investigator and trial were both used as units of analysis. However, as there were only five trials, it became difficult to base the analysis on trial as unit of analysis in the case of the full bivariate random-effects approach. The results have shown a remarkable difference in the two cases. Consistently, in all of the different simplifications, the R^2_{trial} values were found to be higher when trial was used as unit of analysis as expected since the second stage model involved a simple linear regression based on only five data points. Furthermore, it is noted that, when investigator is used as unit of analysis, the R^2_{trial} values are higher when the reduced model is used as compared to the the case where the full model used. The is an indication that the investigator-specific intercept terms

for the surrogate model do convey information and unless there is special reason, full model is to be preferred. The opposite result observed when trials are used as unit of analysis is also explained in the same manner. The bivariate full random effects model does not converge when trial is used as the unit of analysis. This might be due to lack of sufficient information to compute all sources of variability. The reduced bivariate random effects model converged for both cases, but the resulting variance-covariance matrices were not positive-definite and were ill conditioned, as can be seen from the very large value of the condition number. Consequently, the results of the bivariate random effects model should be treated with caution as there might be high uncertainty attached to the results obtained based upon these ill-conditioned matrices. If we concentrate on the results based on investigator as unit of analysis, we observe a low level of surrogacy of PANSS for CGI, with R^2_{trial} ranging roughly between 0.5 and 0.68 for the different simplified models. This result, however, has to be coupled with other findings based on expert opinion to fully guarantee the validation of PANSS as possible surrogate for the CGI. Turning to R^2_{indiv}, it ranges between 0.4904 and 0.5230, depending on the method of analysis, which is relatively low. To conclude, based on the investigators as unit of analysis, PANSS does not seem a good surrogate for the CGI. For the ARMD study, the only available unit of analysis was center. There were 36 centers which treated between 2 and 18 patients. Note that these data has been analyzed by Buyse *et al* (2000) with a treatment coding of 0 and 1 for the placebo and treatment arms, respectively. Here, the $-1/+1$ coding was used and thus slightly different results are obtained. The basic meta-analytic approach and the corresponding simplified modeling strategies have also been applied to this dataset and the results are displayed in Table 3.4 for the $-1/+1$ coding and in Table 3.5 for the $0/1$ coding. For the ARMD study, the R^2_{trial} ranges roughly between 0.64 and

0.8, except for the full bivariate random effects models where we find $\widehat{R}^2_{\text{trial}} = 0.9999$. However, the corresponding variance-covariance matrices were non-positive definite and have very large condition number, a sign of high uncertainty surrounding the latter estimate. Hence, it cannot be trusted. Based on the findings, it is possible to say that assessment of change in visual acuity at 6 months does not seem to be a very strong surrogate for the same assessment at 1 year.

3.6 Discussion

In this chapter we reviewed the meta-analytic strategy for validating a surrogate endpoint. The choice of unit of analysis and corresponding computational issues that need to be given due attention have also been addressed. The choice of unit of analysis in applying the meta-analytic approach is a very important issue to be considered. There might be a large difference in the findings depending on the unit of analysis chosen. The optimal unit of analysis is the one for which there is a sufficient number of repetition and each unit has sufficiently large number of individuals within it. Ideally, the choice of unit of analysis should be based on both statistical and subject-matter considerations. The treatment coding also needs to be given serious consideration, in consultation with experts who may be able to formulate an opinion on the possible variability of the two treatment arms. It is also equally important to give due attention to the variance covariance matrices based upon which the association measures are computed. Because an ill-condition or non-positive definite variance-covariance matrix could yield an inflated association measure which could be misleading. A small simulation study and analysis of two real datasets supported these points.

Table 3.3: *Schizophrenia study. Results of the trial-level (R^2_{trial}) surrogacy analysis.*

	Fixed effects		**Random effects**	
Unit of analysis	Unweighted	Weighted	Unweighted	Weighted
Full Model				
Univariate approach				
Investigator	0.5887	0.5608	0.5488	0.5447
Trial	0.9641	0.9636	0.9849	0.9909
Bivariate approach				
Investigator	0.5887	0.5608	0.9898∗	
Trial	0.9641	0.9636	—	
Reduced Model				
Univariate approach				
Investigator	0.6707	0.5927	0.5392	0.5354
Trial	0.8910	0.8519	0.7778	0.8487
Bivariate approach				
Investigator	0.6707	0.5927	0.9999∗	
Trial	0.7418	0.8367	0.9999∗	

∗: *The variance-covariance matrix is ill-conditioned; in particular, at least one eigenvalue is very close to zero.The condition numbers for the three models with ill-condition matrices, from top to bottom are 3.415E+18, 2.384E+18 and 1.563E+18 respectively.*

Table 3.4: *ARMD data. Results of the trial-level (R^2_{trial}) surrogacy analysis $-1/+1$ coding.*

	Fixed effects		**Random effects**	
Unit of analysis	Unweighted	Weighted	Unweighted	Weighted
			Full Model	
			Univariate approach	
Center	0.6922	0.6963	0.6605	0.7959
			Bivariate approach	
Center	0.6922	0.6963		0.9999∗
			Reduced Model	
			Univariate approach	
Center	0.6409	0.6562	0.6772	0.7929
			Bivariate approach	
Center	0.6409	0.6562		0.9999∗

∗: *The variance-covariance matrix is ill-conditioned; in particular, at least one eigenvalue is very close to zero.The condition numbers for Full and Reduced Bivariate random effects models are 1.109E+17 and 1.965E+18 respectively*

Table 3.5: *ARMD data. Results of the trial-level (R^2_{trial}) surrogacy analysis* 0/1 *coding.*

	Fixed effects		**Random effects**	
Unit of analysis	Unweighted	Weighted	Unweighted	Weighted
Full Model				
Univariate approach				
Center	0.692	0.693	0.664	0.801
Bivariate approach				
Center	0.692	0.693	—	
Reduced Model				
Univariate approach				
Center	0.776	0.758	0.659	0.786
Bivariate approach				
Center	0.776	0.758	—	

4

Information-theoretic Approach

The meta-analytic framework of surrogate marker validation has given another dimension, the trial level, which made it possible to validate a surrogate endpoint in terms of its capacity to convey message about the treatment effect on the true endpoint. However, this approach is computationally intensive and also requires different hierarchical models for different types of outcomes. To circumvent this problem and give a unified approach, Alonso and Molenberghs (2007) have introduced the information-theoretic approach. This approach is simple to apply and can be used for a variety of outcome combinations. In this chapter we will outline this approach through which extension, and therefore unification for a variety of outcomes can be attained Alonso and Molenberghs (2005).

4.1 The Likelihood Reduction Factor

Estimating individual-level surrogacy, as the previous developments clearly show, has frequently been based on a variance-covariance matrix coming from the distribution of the residuals. However, if we move away from the normal distribution, it is not always clear how to quantify the association between both endpoints after adjusting for treatment and trial effect. To address this problem, Alonso *et al* (2005) and Alonso and Molenberghs (2007) considered the following generalized linear models

$$g_T\{E(T_{ij})\} = \mu_{Ti} + \beta_i Z_{ij}, \tag{4.1}$$

$$g_T\{E(T_{ij}|S_{ij})\} = \theta_{0i} + \theta_{1i} Z_{ij} + \theta_{2i} S_{ij}, \tag{4.2}$$

where g_T is an appropriate link function, μ_{Ti} are the trial-specific intercepts and β_i are trial-specific effects of treatment Z on the true endpoint in trial i. θ_{0i} and θ_{1i} are trial-specific intercepts and effects of treatment on the true endpoint when the surrogate endpoint is known. Note that (4.1) and (4.2) can be readily extended to incorporate more complex settings. Other extensions, such as non-linearity between S_{ij} and $g_T\{E(T_{ij})\}$ are possible. Without loss of generality, we assume a linear relationship between S_{ij} and $g_T\{E(T_{ij})\}$. If the trial-specific effects are considered random, we extend (4.1) and (4.2) to appropriate generalized linear mixed-effects models

$$g_T\{E(T_{ij})\} = \mu_T + m_{Ti} + \beta Z_{ij} + b_i Z_{ij}, \tag{4.3}$$

$$g_T\{E(T_{ij}|S_{ij})\} = \theta_0 + c_{Ti} + \theta_1 Z_{ij} + a_i Z_{ij} + \theta_{2i} S_{ij}, \tag{4.4}$$

where μ_T and β are a fixed intercept and treatment effect on the true endpoint, while m_{Ti} and b_i are a random intercept and treatment effects on the true endpoint. θ_0 and θ_1 are a fixed intercept and treatment effect on the true endpoint when the surrogate is known, and c_{Ti} and a_i are a random intercept and treatment effects on the true

endpoint when the surrogate is known. Observe that, in the case where the true endpoint is continuous and normally distributed, (4.1) and (4.2) reduce to normal regression models and (4.3) and (4.4) reduce to linear mixed models. On the other hand, when the true endpoint is binary, (4.1) and (4.2) reduce to logistic regression models. Let us turn to the so-called *likelihood reduction factor* (LRF). Alonso and Molenberghs (2007) used the LRF to evaluate individual level surrogacy, which is obtained by

$$LRF = 1 - \frac{1}{N}\sum_i \exp\left(-\frac{G_i^2}{n_i}\right), \tag{4.5}$$

where G_i^2 denotes the log-likelihood ratio test statistic to compare (4.1) and (4.2) or (4.3) and (4.4) within trial i. Alonso *et al* (2005) established a number of properties for LRF, in particular its ranging in the unit interval and, importantly, its reduction to the individual level surrogacy measure used in the cross-sectional case.

4.2 An Information-theoretic Unification

This proposal avoids the needs for a joint, hierarchical model, and allows for unification across different types of endpoints. The entropy of a random variable (Shannon 1948), a good measure of randomness or uncertainty, is defined in the following way for the case of a discrete random variable Y, taking values $\{k_1, k_2, \ldots, k_m\}$, and with probability function $P(Y = k_i) = p_i$:

$$H(Y) = \sum_i p_i \log\left(\frac{1}{p_i}\right). \tag{4.6}$$

The differential entropy $h_d(X)$ of a continuous variable X with density $f_X(x)$ and support S_{f_X} equals

$$h_d(Y) = -E[\log f_X(X)] = -\int_{S_{f_X}} f_X(x)\log f_X(x)dx. \tag{4.7}$$

The joint and conditional (differential) entropies are defined in an analogous fashion. Defining the information of a single event as $I(A) = \log p_A$, the entropy is $H(A) = -I(A)$. No information is gained from a totally certain event, $p_A \approx 1$, so $I(A) \approx 0$, while an improbable event is informative. $H(Y)$ is the average uncertainty associated with P. Entropy is always non-negative, satisfies $H(Y|X) \leq H(Y)$ for any pair of random variables, with equality holding under independence, and is invariant under a bijective transformation Cover and Tomas (1991). Differential entropy enjoys some but not all properties of entropy: it can be infinitely large, negative, or positive, and is coordinate dependent. For a bijective transformation $Y = y(X)$, it follows $h_d(Y) = h_d(X) - \mathrm{E}_Y\left(\log\left|\frac{dx}{dy}(y)\right|\right)$. We can now quantify the amount of uncertainty in Y, expected to be removed if the value of X were known, by $I(X,Y) = h_d(Y) - h_d(Y|X)$, the so-called *mutual information*. It is always non-negative, zero if and only if X and Y are independent, symmetric, invariant under bijective transformations of X and Y, and $I(X,X) = h_d(X)$. The mutual information measures the information of X, shared by Y. Let us now introduce the entropy-power (Shannon 1948) for comparison of continuous random variables. Let X be a continuous n-dimensional random vector. The entropy-power of X is

$$\mathrm{EP}(X) = \frac{1}{(2\pi e)^n} e^{2h(X)}. \tag{4.8}$$

The differential entropy of a continuous normal random variable is $h(X) = \frac{1}{2}\log\left(2\pi\sigma^2\right)$, a simple function of the variance and, on the natural logarithmic scale: $\mathrm{EP}(X) = \sigma^2$. In general, $\mathrm{EP}(X) \leq \mathrm{Var}(X)$ with equality if and only if X is normally distributed. We can now define an information-theoretic measure of association Schemper and Stare (1996):

$$R_h^2 = \frac{\mathrm{EP}(Y) - \mathrm{EP}(Y|X)}{\mathrm{EP}(Y)}, \tag{4.9}$$

which ranges in the unit interval, equals zero if and only if (X, Y) are independent, is symmetric, is invariant under bijective transformation of X and Y, and, when $R_h^2 \to 1$ for continuous models, there is usually some degeneracy appearing in the distribution of (X,Y). There is a direct link between R_h^2 and the mutual information: $R_h^2 = 1 - e^{-2I(X,Y)}$. For Y discrete: $R_h^2 \leq 1 - e^{-2H(Y)}$, implying that R_h^2 then has an upper bound smaller than 1; we then redefine

$$\widetilde{R}_h^2 = \frac{R_h^2}{1 - e^{-2H(Y)}},$$

reaching 1 when both endpoints are deterministically related.

We can now redefine surrogacy, while preserving previous proposals as special cases. While we will focus on individual-level surrogacy, all results apply to the trial level too. Let $Y = T$ and $X = S$ be the true and surrogate endpoints, respectively. We consider S a good surrogate for T at the individual (trial) level, if a "large" amount of uncertainty about T (the treatment effect on T) is reduced when S (the treatment effect on S) is known. Equivalently, we term S a good surrogate for T at the individual level, if our lack of knowledge about the true endpoint is substantially reduced when the surrogate endpoint is known. A meta-analytic framework, with N clinical trials, produces N_q different R_{hi}^2, and a meta-analytic R_h^2 given by:

$$R_h^2 = \sum_{i=1}^{N_q} \alpha_i R_{hi}^2 = 1 - \sum_{i=1}^{N_q} \alpha_i e^{-2I_i(S_i,T_i)},$$

where $\alpha_i > 0$ for all i and $\sum_{i=1}^{N_q} \alpha_i = 1$ can be entertained. Different choices for α_i lead to different proposals, producing an uncountable family of parameters. This opens the additional issue of finding an *optimal* choice. In particular, for the cross-sectional normal-normal case, Alonso and Molenberghs (2006) have shown that $R_h^2 = R_{\text{indiv}}^2$. Finally, when the true and surrogate endpoints have distributions in the exponential family, then LRF $\xrightarrow{P} R_h^2$ when the number of subjects per trial goes to

infinity. Alonso and Molenberghs (2007) developed asymptotic confidence intervals for R_h^2, based on the idea of Kent (1983), to build confidence intervals for $2I(T,S)$. Let $\hat{a} = 2n\widehat{I}(T,S)$, where n is the number of patients. Define $\kappa_{1:\alpha}$(a) and $\delta_{1:\alpha}$(a) by $P\left(\chi_2^1(\kappa_{1:\alpha}(a)) \geq a\right) = \alpha$ and $P\left(\chi_2^1(\delta_{1:\delta}(a)) \leq a\right) = \alpha$. Here, χ_1^2 is a chi-squared random variable with 1 degree of freedom. If $P\left(\chi_2^1(0) \geq a\right) = \alpha$ then we set $\kappa_{1:\alpha}(a) = 0$. A conservative two-sided $1-\alpha$ asymptotic confidence interval for R_h^2 is

$$\sum_i \alpha_i \left[n_i^{-1}\kappa_{1:\alpha}^i(\hat{a}), n_i^{-1}\delta_{1:\alpha}^i(\hat{a})\right], \tag{4.10}$$

where $1-\alpha_i$ is the Bonferroni confidence level for the trial intervals Alonso and Molenberghs (2007). This asymptotic interval has considerable computational advantage with respect to the bootstrap approach used by Alonso *et al* (2005). Although ITA involves substantial mathematics, its implementation in practice is fairly straightforward and less computer-intensive than the meta-analytic approach.

5

Mixture of Continuous and Binary Outcomes

In one of the preceding chapters, we have considered the meta-analytic approach which has been formulated originally for two continuous, normally distributed outcomes. We have raised some concerns related to the units of analysis and computational issues that need to be addressed. In addition to the meta-analytic approach, we have also outlined the information-theoretic approach which allows unification across different endpoint combinations. The meta-analytic approach has been extended for other type of outcome combinations. One such extension involves the mixture of continuous and binary outcomes as given by Molenberghs Geys, and Buyse (2001). In this chapter we provide a review of the method for a continuous-binary endpoint combination and outline the application of the information-theoretic approach to the mixed continuous-binary case and asses its performance using a simulation study.

5.1 Methods for Mixed Continuous-Binary Endpoints

Statistical problems where various outcomes of a combined nature are observed are common, especially with normally distributed outcomes on the one hand and binary or categorical outcomes on the other hand. Emphasis may be on the determination of the entire joint distribution of both outcomes or on specific aspects, such as the association in general or correlation in particular between both outcomes. Here we focus on the combination of continuous and binary outcomes. We start with a bivariate non-hierarchical setting, which can always be expressed as the product of a marginal distribution of one of the responses and the conditional distribution of the remaining response given the former one. The main problem with this approach is that no easy expressions for the association between both endpoints are available. Thus, we opt for a symmetric treatment of both endpoints. Let us focus on the case where the true endpoint is continuous and the surrogate is binary, the reverse case being entirely similar. Generalized linear mixed models for endpoints of different data types are challenging Molenberghs and Verbeke (2005). Hence, we concentrate on two-stage fixed-effects models. In the first stage, let $\widetilde{S}_{ij}$ be a latent variable of which S_{ij} is the dichotomized version. A bivariate normal model for $\widetilde{S}_{ij}$ and T_{ij} is given by Molenberghs, Geys, and Buyse (2001):

$$\widetilde{S}_{ij} = \mu_{Si} + \alpha_i Z_{ij} + \varepsilon_{Sij}, \tag{5.1}$$

$$T_{ij} = \mu_{Ti} + \beta_i Z_{ij} + \varepsilon_{Tij}, \tag{5.2}$$

where μ_{Si} and μ_{Ti} are trial-specific intercepts, α_i and β_i are trial-specific effects of treatment Z_{ij} on the endpoints in trial i, and ε_{Si} and ε_{Ti} are correlated error terms,

assumed to be zero-mean normally distributed with covariance matrix

$$\Sigma = \begin{pmatrix} \frac{1}{(1-\rho^2)} & \frac{\rho\sigma}{\sqrt{(1-\rho^2)}} \\ & \sigma \end{pmatrix}. \tag{5.3}$$

The variance of $\widetilde{S}_{ij}$ is chosen for computational reasons. Using a probit formulation like Molenberghs Geys, and Buyse (2001) and owing to the replication at the trial level, we can impose a distribution on the trial-specific parameters. At the second stage, we assume model (3.7) given in section 3.1 of Chapter 3.

Measures to assess the quality of the surrogate both at the trial and individual level are obtained as in (3.8) and (3.10). Interpretation of these measures and decision making follows the logic laid out in Section 3.1 of Chapter 3.

5.2 A Simulation Study

The direct consequence of the fact that the models used in the information-theoretic approach are univariate models is that, the models can be easily fitted using any standard regression software. However, the performance of this approach has not been studied in the mixed continuous and binary endpoint settings. In the next section, insight into the performance of this approach, together with its asymptotic interval, is offered through a simulation study. Let us first lay out the design of the simulation study, whereafter the results are described.

5.2.1 Design of the Simulation Study

Due to the computational difficulties encountered in practice with the bivariate random effects models required for the meta-analytic approach by Buyse *et al* (2000), ITA becomes an interesting option to consider in practice. As stated earlier, the performance of the later has not been investigated in the mixed continuous binary setting, and is the focus of this section. Here, we outline the procedures followed in

generating the data used for simulation. The data were generated based on models (3.12)–(3.13) and the corresponding variance-covariance matrices. After generating continuous outcomes based on the above models, a binary surrogate is obtained by dichotomizing the resulting continuous surrogate using the fixed intercept as cut-off point. The dichotomized surrogate takes value 1 if the corresponding continuous surrogate is greater than μ_S and zero otherwise. This formulation assumes trial-level and individual-level R^2 values of 0.9 and 0.64, respectively, at the continuous scale. It is important to note that this value of the individual-level R^2 is the squared correlation between the latent unobservable continuous surrogate endpoint and the observable true endpoints. However, the situation is totally different at the trial-level. Based on (5.1) and (5.2), Alonso *et al* (2005) showed that the relationship between the treatment effects on the latent-continuous and observed-binary surrogate endpoints is linear. Hence, the value of the trial-level R^2 (0.9) is valid both for the latent and observed surrogate. The number of trials was fixed to either 5, 10, 20 or 30. There were 2 sets of trial sizes used, the first set consists of 10, 20, 40 or 60, which we term *small trial size*. The second set consists of 100, 150, 200 or 300, termed *large trial size*. A full combination of the number of trials and trial sizes was obtained. In each case, 100 runs were performed, assuming either models (4.1) and (4.2) or models (4.3) and (4.4). Apart from the primary objectives to investigate the performance of ITA as well as comparing the bootstrap percentile intervals with the asymptotic interval by Alonso and Molenberghs (2007), there are two secondary objectives. The first is to investigate the impact of alternative link functions, at the individual-level, on the performance of ITA. Thus, both probit and logit link functions were implemented in all settings. Second, both linear and non-linear (splines) functions were considered, at the trial-level, to explore the assumption of linearity between treatment effects.

5.2.2 Simulation Results

Tables 5.2 and 5.3 present a selection of the simulation results. We focus on both large and small numbers of trials and numbers of subjects per trial. Both individual-level and trial-level R^2 measures are included. ITA yields estimates of surrogacy at the individual-level, bounded above by 0.3. Hence, the approach yields estimates substantially lower than the value assumed when generating the datasets, 0.64. This phenomenon is observed in all settings considered in the simulation study. However, it should be noted that the value of 0.64 is the individual-level surrogacy at the latent scale, whereas ITA estimates asses the individual-level surrogacy at the observed scale. Also, it is expected that dichotomizing a continuous variable leads to information loss, which would imply that results obtained from the continuous and discrete version should not generally be expected to be in agreement with each other. Unlike the individual level, Alonso *et al* (2002) showed that the trial-level surrogacy at the latent scale translates equally to the observed scale. For small trial sizes ITA tends to underestimate the trial-level surrogacy. Nevertheless, the models perform considerably well for large trial sizes. The mixed-effect models, (4.3) and (4.4), outperform the fixed-effect models, (4.1) and (4.2), in all simulation settings considered. However, the mixed-models had some convergence issues, which were not encountered with the fixed-effect models. Even so, the percentage of non-convergence is smaller than 10% within each simulation setting. Generally, increasing the number of trials has little effect on the surrogacy measures, although increasing the trial size appears to yield better estimates for the surrogacy measures. Also, it is not advisable to use very small number of trials, as it may overestimate or not provide enough data points to reliably assess the trial-level surrogacy. The 95% asymptotic intervals are tighter than the 95% percentile bootstrap intervals for all simulation settings considered. The discrepancy

Table 5.1: *Age-related macular degeneration trial. Estimates (standard error) of the individual-level (R^2_{indiv}) and trial-level (R^2_{trial}) surrogacy analysis based on the conventional and information-theoretic approach.*

		probit link		**logit Link**	
level	type	fixed	mixed	fixed	mixed
information-theoretic approach					
trial	line	0.33 (0.14)	0.49 (0.13)	0.32 (0.13)	0.48 (0.13)
	spline	0.33 (0.13)	0.49 (0.12)	0.32 (0.13)	0.48 (0.13)
individual		0.23 (0.13)	0.27 (0.13)	0.23 (0.13)	0.27 (0.13)
conventional meta-analytic approach (2-stage fixed effects)					
trial				0.42 (0.13)	
individual				0.44 (0.09)	

between these intervals reduces with increases in the number of trials and trial sizes. Further, the choice of an appropriate link function appears to have little influence on the results. We Observed that more than 97% of the samples have differences below 0.1. Also, almost identical results were obtained in each sample when the spline and linear functions were considered at the trial level as more than 93% of the samples have differences inferior to 0.04.

5.3 Application to the Case Study

The case study on Age related Macular Degeneration introduced in Chapter 2 will now be analyzed. The two-stage meta-analytic approach and the corresponding ITA models have been applied to this dataset and results displayed in Table 5.1. Extension of the meta-analytic approach to the mixed continuous and binary endpoints, using two-stage fixed-effects model yields R^2_{indiv}=0.42 (s.e. 0.13) and R^2_{trial}=0.44 (s.e. 0.09).

Thus, the loss of at least two lines of vision at 6 months is a relatively poor surrogate for visual acuity at 1 year, a conclusion in synchrony with the one reached by Buyse *et al* (2000) at the continuous level. At the individual level, ITA yield estimates of R^2_{indiv} ranging from 0.2319 to 0.2735. It should be noted that we do not have information about the degree of under-estimation of R^2_{indiv} by ITA at the observed scaled. As mentioned earlier, research on this issue is still ongoing. Nevertheless, the very low values obtained indicate that the loss of at least two lines of vision at 6 months may not be a good surrogate for visual acuity at 1 year, at the individual level. ITA yields estimates of R^2_{trial} ranging from 0.3211 to 0.4864. This indicates that the loss of at least two lines of vision at 6 months does not seem to be a very good surrogate for visual acuity at 1 year, at the trial level. It should be noted that the size of the largest unit of analysis (center) was only 18, though. Thus, there may be a considerable degree of under-estimation on the estimates of R^2_{trial}. There appears to be no difference between the probit and logit link functions on these data. Also, the line and spline models yield similar results, indicating that the linearity assumption at the trial level may be a plausible one. Furthermore, the mixed models generally have higher estimates for surrogacy measures than the fixed models, hence, exhibiting a lower degree of underestimation.

5.4 Discussion

In this chapter, we reviewed the extension of the meta-analytic strategy of Buyse *et al* (2000), to a mixed binary and continuous endpoints, and the information-theoretic approach for validating surrogate endpoints. Combination of the latter with combined-type outcomes is novel. The meta-analytic approach and its extension are mathematically appealing, but encounter practical and/or computational issues. The

information-theoretic approach on the other hand involves substantial mathematics yet it is more practically feasible than the meta-analytic approach as it depends on simple univariate models. We primarily investigated the performance of the ITA for combined continuous and binary endpoints, through a simulation study. Generally, this approach underestimates the measures of surrogacy. The underestimation reduces with increase in both the number of trials and trial sizes. However, the simulation study showed that the degree of underestimation is higher with very small trial sizes, even for large number of trials. The model proposed by Alonso *et al* (2005) for a general setting, which is based on fixed-effects models, was outperformed by its extension to generalized linear mixed models, which has as its basis two univariate mixed models. Quite similar results were obtained by extending the linear relationship between the true and surrogate endpoints to non-linear, spline-based models, at the trial level. Thus, it may be reasonable to assume a linear relationship between the treatment effects on the true and surrogate endpoints. Asymptotic confidence intervals for surrogacy measures (R^2_{indiv} and R^2_{trial}) performed better than bootstrap confidence intervals, in the sense of being generally more narrow. On the other hand, the asymptotic confidence intervals are computationally advantageous and are tighter than the bootstrap confidence intervals. Arguably, a fully formal comparison would be of interest; we view this a topic for further research. The choice of link function appears to have little influence on the estimates of the surrogacy measures. Particularly, the logit and probit link functions gave similar estimates in all settings considered in the simulation study. This is also supported by the fact that these link functions gave almost identical estimates when applied to the motivational case study. These finding are not surprising in view of their well-known relationship. The meta-analytic strategy for evaluating surrogacy faces computational problems, which are largely al-

leviated by the information-theoretic approach. On the other hand, the latter may be biased downards in smaller trials. Therefore, it is advisable to reserve the use of ITA for larger trial sizes. Also, the extended generalized linear mixed models are recommended. Clearly, the use of validation methods, such as the ones proposed in this chapter, whether based on R^2, other association measures, or ITA, is but one component of the broader surrogate endpoint evaluation picture.

Table 5.2: *Simulation study results for individual level surrogacy.*

Individual-level surrogacy.				
# trials	**# subjects**	R^2_{indiv}	**bootstrap c.i.**	**asymptotic c.i.**
Univariate fixed-effects model.				
5	10	0.15	(0.00;0.38)	(0.05;0.33)
5	60	0.16	(0.05;0.28)	(0.10;0.23)
30	10	0.16	(0.09;0.24)	(0.10;0.23)
30	60	0.16	(0.12;0.21)	(0.14;0.19)
5	100	0.15	(0.06;0.25)	(0.11;0.21)
5	300	0.15	(0.08;0.27)	(0.13;0.19)
30	100	0.16	(0.12;0.19)	(0.14;0.18)
30	300	0.16	(0.12;0.20)	(0.15;0.17)
Univariate mixed-effects model.				
5	10	0.16	(0.02;0.39)	(0.06;0.37)
5	60	0.16	(0.05;0.28)	(0.10;0.23)
30	10	0.18	(0.11;0.26)	(0.12;0.25)
30	60	0.17	(0.12;0.21)	(0.14;0.20)
5	100	0.15	(0.06;0.25)	(0.11;0.21)
5	300	0.16	(0.08;0.27)	(0.13;0.19)
30	100	0.16	(0.12;0.20)	(0.14;0.18)
30	300	0.16	(0.12;0.20)	(0.15;0.17)

Table 5.3: *Simulation study results for the trial level surrogacy.*

Trial-level surrogacy.				
# trials	**# subjects**	R^2_{trial}	**bootstrap c.i.**	**asymptotic c.i.**
Univariate fixed-effects model.				
5	10	0.48	(0.00;0.95)	(0.12;0.81)
5	60	0.59	(0.03;0.95)	(0.13;0.87)
30	10	0.41	(0.13;0.62)	(0.17;0.64)
30	60	0.51	(0.19;0.70)	(0.26;0.71)
5	100	0.81	(0.01;0.99)	(0.29;0.93)
5	300	0.82	(0.15;0.98)	(0.30;0.96)
30	100	0.71	(0.46;0.84)	(0.47;0.85)
30	300	0.75	(0.54;0.87)	(0.52;0.88)
Univariate mixed-effects model.				
5	10	0.42	(0.00;0.95)	(0.13;0.78)
5	60	0.59	(0.00;0.94)	(0.14;0.84)
30	10	0.43	(0.04;0.62)	(0.18;0.63)
30	60	0.53	(0.28;0.75)	(0.28;0.73)
5	100	0.82	(0.08;0.98)	(0.30;0.94)
5	300	0.88	(0.34;0.99)	(0.41;0.98)
30	100	0.78	(0.27;0.90)	(0.54;0.87)
30	300	0.82	(0.51;0.92)	(0.61;0.90)

6

A Binary Surrogate for a Binary True Endpoint

The previous chapters dealt with the case of two normally distributed outcomes and the case of a mixture of a normal and binary outcomes. We have noticed that, the challenge of quantifying the association measures was more pronounced for the case of the mixed continuous-binary endpoints. However, the introduction of the information-theoretic approach has given a great deal of flexibility through which it has been possible to quantify the individual level surrogacy in a simpler manner than dealing with a probit formulation. In this chapter we move one step further and consider the case of two binary outcomes. Unfortunately there is no simple tractable bivariate model, similar to the bivariate normal distribution, which will enable us to directly apply the method used for the case of two normally distributed outcomes. However, similar to the previous chapter, we can attempt to quantify the individual level association through the use of a probit formulation on the one hand and the

information-theoretic approach as a simplified alternative. We first consider the meta-analytic approach for two binary outcomes which is based on a probit formulation and proceed to the information-theoretic approach.

6.1 The Meta-Analytic Approach for 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, resting on the assumption that the observed binary variables result from dichotomizing an unobserved continuous variable based on the threshold chosen. Assume a pair of latent variables $(\widetilde{S}_{ij}, \widetilde{T}_{ij})$, representing the continuous, underlying values of the surogate and true endpoints for subject j in trial i, follwoing a random-effects model at the latent scale:

$$\widetilde{S}_{ij} = \mu_S + m_{Si} + \alpha Z_{ij} + a_i Z_{ij} + \varepsilon_{Sij}, \tag{6.1}$$

$$\widetilde{T}_{ij} = \mu_T + m_{Ti} + \beta Z_{ij} + b_i Z_{ij} + \varepsilon_{Tij}, \tag{6.2}$$

where 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 ε_{Sij} and ε_{Tij} are error terms. The random effects are zero-mean normally distributed with covariance matrix D given in (3.6), which is the same matrix we considered for the case of two normally distributed outcomes and probit formulation of the mixture of a binary and continuous outcomes. The error terms are also zero-mean normally distributed with covariance matrix:

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho_{ST} \\ \rho_{ST} & 1 \end{pmatrix}.$$

The implied model for the observed binary outcomes is then given by:

$$\Phi^{-1}[P(S_{ij}=1|m_{Si},m_{Ti},a_i,b_i)] = \mu_S + m_{Si} + \alpha Z_{ij} + a_i Z_{ij}, \quad (6.3)$$

$$\Phi^{-1}[P(T_{ij}=1|m_{Si},m_{Ti},a_i,b_i)] = \mu_T + m_{Ti} + \beta Z_{ij} + b_i Z_{ij}, \quad (6.4)$$

where Φ denotes the standard normal cumulative distribution function. Formulation (6.1)–(6.2) allows the use of the coefficient of determination (3.8) as the trial-level R^2, whereas the individual-level R^2_{indiv} is equal to the square of ρ_{ST}.

6.1.1 Parameter Estimation

Models (6.1)–(6.2) belong to the class of so-called generalized linear mixed models (Molenberghs and Verbeke 2005), with probit link even though the logit link is more generally used for binary outcomes. Molenberghs and Verbeke (2005) discuss a variety of commonly used estimation methods, including maximum likelihood with numerical integration over the random effects, penalized quasi-likelihood, marginal pseudo-likelihood, and Laplace approximation. These methods suffer to various extents from computational complexity and severe bias (Rodríguez and Goldman 1995, Molenberghs and Molenberghs 2005). For the specific case of the probit link, as in (6.3)–(6.4), Renard *et al* (2002) have suggested the use of so-called maximum pairwise likelihood (MPL), a form of pseudo-likelihood (Molenberghs and Verbeke 2005). Let us briefly describe this method. Assembling all parameters into the vector $\boldsymbol{\Theta}$, the contribution of the ith trial ($i = 1, \ldots, N$) to the likelihood, conditional on $\boldsymbol{b_i} = (m_{Si}, m_{Ti}, a_i, b_i)^T$, is

$$L_i(\boldsymbol{\Theta}|\boldsymbol{b_i}) = \prod_{j=1}^{n_i} P(S_{ij}, T_{ij}|\boldsymbol{b_i}). \quad (6.5)$$

Maximum likelihood estimation follows from integrating (6.5) over $\boldsymbol{b_i}$, summing over all subjects, taking the logarithm, and maximizing

$$\ell(\boldsymbol{\Theta}) = \sum_{i=1}^{N} \ln \int L_i(\Theta|\boldsymbol{b_i})\phi(\boldsymbol{b_i};\boldsymbol{D})d\boldsymbol{b_i} \tag{6.6}$$

over $\boldsymbol{\Theta}$. Here, $\phi(\boldsymbol{b_i};\boldsymbol{D})$ denotes the mean-zero multivariate normal density with covariance matrix $\boldsymbol{D}$. The intractable nature of (6.6) dictates the use of one or other form of approximation, as mentioned earlier. Renard *et al* (2002) suggested the use of maximum pairwise likelihood (MPL), a pseudo-likelihood approach based on replacing the likelihood by a product of conditional and/or marginal densities. In our particular case, the proper likelihood contribution of trial i is replaced by all possible pairwise margins. Detailed overviews of the methodology can be found in Molenberghs and Verbeke (2005) and Burzykowski, Molenberghs, and Buyse (2005).

6.2 Drawbacks and Simplified Modeling Strategies

While it is technically possible to fit the bivariate probit model, the use of which necessitated by the pairwise likelihood approach, there still are a number of drawbacks associated with the approach outlined. First, the resulting surrogate marker evaluation measures apply to the postulated latent variables rather than to the observed binary variables. Second, the computational burden still is considerable. Third, the approach might result in an ill-conditioned variance-covariance matrix, thence calling the reliability of the association measures derived into question. In light of these difficulties, it is beneficial to switch towards ITA. The ITA approach can be adapted to accommodate the case of two binary outcomes by choosing an appropriate link function such as logit or probit in the models (4.1) and (4.2). The individual level surrogacy can then be quantified by using (4.5). As stated earlier, one issue arising is that, for discrete random variables the measure of association based on ITA has

an upper bound smaller than one, as shown by Alonso and Molenberghs (2007), who therefore suggested the use of an adjusted version:

$$R^2_{\text{hadj}} = \frac{R^2_h}{1 - \exp[-2H(Y)]}, \tag{6.7}$$

where $H(y)$ is the log-likelihood of the true endpoint divided by the total number of subjects. ITA ideas can be applied to compute the trial-level R^2_{trial} too, using the fully hierarchical model for continuous outcomes (Buyse *et al* 2000). The resulting $R^2_{h,\text{trial}}$ will take the same form (4.5), with now G^2_i the likelihood ratio statistics for comparing models relating treatment effect on the true endpoint, with and without adjusting for the treatment effect on the surrogate endpoint. Since this second-stage model is for continuous endpoints, the issue of an upper bound smaller than one does not crop up.

6.3 Simulation Study

Though the ITA approach has been applied to a case study involving binary outcomes before, no objective evaluation has been performed to investigate the performance of this approach through simulation studies. Here we will assess the performance of the information-theoretic approach in comparison with the bivariate probit model, first laying out the design of our simulation study and then summarizing the results.

6.3.1 Design of Simulation Study

The data were generated based on model (6.3)–(6.4). The parameters were set equal to $\mu_S = 0.5$, $\mu_T = 0.45$, $\alpha = 0.05$, and $\beta = 0.03$. Values assumed for the covariance

matrices are:

$$\Sigma = \begin{pmatrix} 3 & 2.4 \\ & 3 \end{pmatrix}, \qquad D = \begin{pmatrix} 3 & 2.4 & 0 & 0 \\ & 3 & 0 & 0 \\ & & 3 & 2.84605 \\ & & & 3 \end{pmatrix}.$$

After generating continuous outcomes based on the above models, the corresponding binary variables are obtained by dichotomizing the resulting continuous outcomes using the fixed intercepts as cut-off points, setting values exceeding the intercept to 1 and 0 otherwise. These model choices imply R^2_{trial}=0.90 and $R^2_{\text{indiv}} = 0.64$, at the continuous scale.

6.3.2 Simulation Results

Several combination of the number of trials and trial sizes was considered. In each case, 100 runs were performed. We further distinguish between the bivariate and univariate models on the one hand, and mixed- versus fixed-effects models on the other hand. The mixed models take the form of the probit model in the bivariate situation and the Generalized Linear Mixed Model (GLMM) in the univariate case. The simulation results are displayed in Tables 6.2–6.17. Let us first consider association at the trial level. The simulation reveal that the full bivariate random effects model and its univariate counterpart are consistent in that both models produce surrogacy measures approaching the true values with the number of trials and the number of subjects increasing. However, the corresponding fixed-effects models lead to underestimation, even for larger sample sizes. Even though the full bivariate model leads to measures at the latent scale, since it measures the association between the treatment effects on the two endpoints, we expect it to be preserved at the explicit scale. This claim is corroborated by the results from the univariate mixed model, which operates at the observed binary scale. It is also noteworthy that there is not much difference

between the ITA and the conventional approach of regressing the treatment effect on the true endpoint on the treatment effect on the surrogate endpoint. Turning to the individual-level association, the full bivariate random-effects and bivariate fixed-effect models result in individual-level measures close to the true value, i.e., the theoretical value at the latent scale. However, they are hard to translate from the latent scale to the explicit one. ITA is a convenient way out of this problem. An important observation is that the values reported with ITA are substantially smaller than their latent counterparts, in line with expectation: switching from the latent scale to the explicitly observed scale reduces association. This is a manifestation of the fact that important information is lost when switching from a continuous to a binary scale. Of course, in a real study, the binary variables are the only ones observed and it is therefore fair to assert that the ITA is a fair representation of reality, whereas the other methods are overly optimistic.

6.4 Application to the Case Study

The acute migrane data introduced in Chapter 2, Sections 2.1.4 was analyzed using the methods introduced in the previous sections of this chapter. Of the symptoms studied: nausea, vomiting, photophobia, phonophobia, the photobia symptom had the highest trial-level surrogacy. Results for both the trial- and individual-level surrogacy are presented in Table 6.1. Both point estimates as well as 95% confidence intervals are presented. Observe that the univariate and bivariate fixed-effects models result in smaller R^2_{trial} than the random-effects counterpart. However, the latter is unreliable since it is found to be based on an ill-conditioned covariance matrix, in the sense of a grossly inflated leading eigenvalue. Basing our conclusions on the univariate mixed effect model, in the simulations found to work well, it is fair to assert that, at

the trial level, the presence of photophobia is a good surrogate for migraine severity, i.e., the corresponding R^2_{trial} may be considered sufficiently high. The reasonably good agreement between the treatment effects at both levels, and in addition the absence of obvious outliers, is clear from Figure 6.1, a so-called bubble plot, displaying a scatter of the pairs of treatment effects for each unit. The size of the circles, or bubbles, is proportional to the number of patients per unit. The R^2_{indiv} for the bivariate fixed and mixed models are higher than their univariate counterparts. This is expected for the same reason as explained in Section 6.3.2, i.e., one is at the latent scale, whereas the ITA works at the interpretationally more relevant explicitly observed scale.

6.5 Discussion

In this chapter, we have considered the bivariate probit model approach, and the information-theoretic approach for validating surrogate endpoints. The use of the latter framework with binary data has given a substantial simplicity of application and ease of interpretation of the resulting association measure. The meta-analytic framework through the probit formulation, where individual-level surrogacy is expressed at the latent level, leads to overestimation of the said quantity. Since the ITA operates at the explicitly observed scale, it provides a fairer and more useful quantity. Additionally, the computational complexity of the full random-effects meta-analytic framework has led to the use of simplifying frameworks, trading the random effects for fixed effects on the one hand and/or bivariate, joint modeling of both endpoints by two univariate, separate models. These simplifications work well when the number of trials and the number of subjects per trials is large, indicating one should in practice carefully consider the unit of analysis. Applying the proposed methodology to acute

Table 6.1: *Acute Migraine Study. Estimates (confidence intervals) for trial-level and individual-level surrogacy for the photophobia symptom.*

Trial-level surrogacy			
Fixed effects		Random effects	
Unweighted	Weighted	Unweighted	Weighted
Univariate approach			
0.7579	0.7579	0.8112	0.8886
(0.5712;0.8817)	(0.5712;0.8817)	(0.6367;0.9066)	(0.8134;0.9567)
Bivariate approach			
0.7336	0.7336	0.9587*	
(0.5426;0.8688)	(0.5426;0.8688)	(0.6966;1.000)	

Individual-level surrogacy	
Fixed effects	Random effects
Univariate approach (ITA based)	
0.5016	0.5885
(0.4354;0.5681)	(0.5221;0.6540)
Bivariate approach (probit, latent scale)	
0.8959	0.8664
(0.8822;0.9095)	(0.6042;1.000)

$*$: *This value is unreliable due to ill-conditioning of the variance-covariance matrix from which it was calculated.*

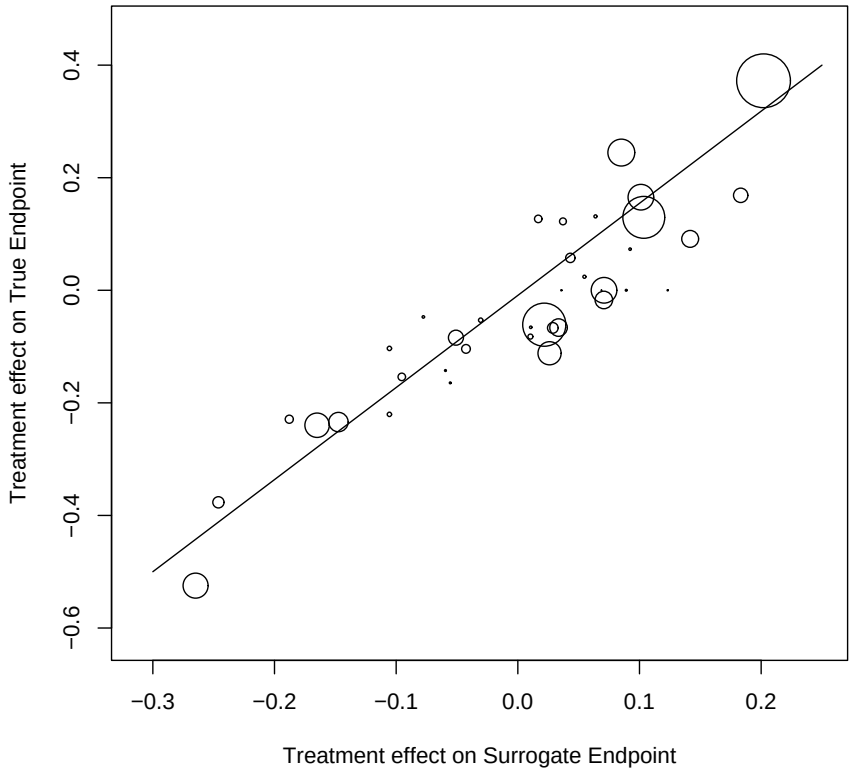

Figure 6.1: *Acute Migraine Study. Bubble plot of trial-specific treatment effect on the surrogate versus true endpoints. The size of the bubbles corresponds to the size of the trial*

Table 6.2: *Simulation study. Univariate mixed-effects model for large trial sizes, individual-level surrogacy.*

	strategy			**confidence intervals**	
sim. #	# trials	# subjects	R^2_{indiv}	**percentile**	**asymptotic**
1	5	100	0.2358	(0.1392;0.3489)	(0.1687;0.3104)
2	5	150	0.2255	(0.1152;0.3402)	(0.1710;0.2852)
3	5	200	0.2252	(0.1250;0.3378)	(0.1776;0.2767)
4	5	300	0.2211	(0.1354;0.3102)	(0.1822;0.2627)
5	10	100	0.2256	(0.1523;0.3158)	(0.1782;0.2769)
6	10	150	0.2215	(0.1536;0.3100)	(0.1827;0.2629)
7	10	200	0.2190	(0.1613;0.3068)	(0.1854;0.2547)
8	10	300	0.2172	(0.1634;0.2821)	(0.1896;0.2461)
9	20	100	0.2298	(0.1866;0.2778)	(0.1955;0.2661)
10	20	150	0.2274	(0.1751;0.2853)	(0.1994;0.2568)
11	20	200	0.2249	(0.1878;0.2635)	(0.1947;0.3089)
12	20	300	0.2213	(0.1854;0.2748)	(0.1936;0.2856)
13	30	100	0.2340	(0.1971;0.2816)	(0.2006;0.2502)
14	30	150	0.2289	(0.1891;0.2723)	(0.2058;0.2528)
15	30	200	0.2287	(0.1824;0.2603)	(0.2086;0.2493)
16	30	300	0.2220	(0.1839;0.2561)	(0.2054;0.2225)

migraine trial data has shown that photophobia is a reasonably good surrogate at the trial level, whereas its surrogacy at the individual level may be called into question. This finding is of interest and may spark of further investigation from a clinical and biopharmaceutical perspective.

Table 6.3: *Simulation study. Univariate mixed-effects model for large trial sizes, trial-level surrogacy.*

	strategy			**confidence intervals**	
sim. #	# trials	# subjects	R^2_{trial}	**percentile**	**asymptotic**
1	5	100	0.9014	(0.5550;0.9959)	(0.5719;0.9901)
2	5	150	0.9028	(0.5693;0.9985)	(0.5803;0.9903)
3	5	200	0.8995	(0.5416;0.9978)	(0.5704;0.9998)
4	5	300	0.9092	(0.5418;0.9993)	(0.6014;0.9907)
5	10	100	0.8716	(0.3962;0.9683)	(0.6005;0.9729)
6	10	150	0.8870	(0.4939;0.9718)	(0.6283;0.9781)
7	10	200	0.8878	(0.4767;0.9760)	(0.6312;0.9780)
8	10	300	0.8864	(0.5575;0.9753)	(0.6322;0.9770)
9	20	100	0.8686	(0.7271;0.9432)	(0.6809;0.9583)
10	20	150	0.8722	(0.7406;0.9442)	(0.6869;0.9597)
11	20	200	0.8762	(0.7222;0.9493)	(0.6942;0.9612)
12	20	300	0.8834	(0.7574;0.9548)	(0.7079;0.9640
13	30	100	0.8596	(0.7548;0.9397)	(0.7066;0.9436)
14	30	150	0.8631	(0.7659;0.9428)	(0.7119;0.9454)
15	30	200	0.8713	(0.7761;0.9441)	(0.7259;0.9493)
16	30	300	0.8767	(0.7977;0.9415)	(0.7344;0.9520)

Table 6.4: *Simulation study. Univariate fixed-effects model for large trial sizes, individual-level surrogacy.*

	strategy			**confidence intervals**	
sim. #	# trials	# subjects	R^2_{indiv}	**percentile**	**asymptotic**
1	5	100	0.2232	(0.1135;0.3360)	(0.1577;0.2966)
2	5	150	0.2194	(0.1297;0.3354)	(0.1654;0.2787)
3	5	200	0.2183	(0.1082;0.3345)	(0.1714;0.2692)
4	5	300	0.2183	(0.1185;0.3070)	(0.1796;0.2597)
5	10	100	0.2134	(0.1388;0.2955)	(0.1671;0.2638)
6	10	150	0.2161	(0.1367;0.3109)	(0.1777;0.2572)
7	10	200	0.2132	(0.1557;0.3042)	(0.1798;0.2485)
8	10	300	0.2142	(0.1509;0.2963)	(0.1867;0.2429)
9	20	100	0.2149	(0.1693;0.2627)	(0.1814;0.2504)
10	20	150	0.2161	(0.1609;0.2699)	(0.1885;0.2449)
11	20	200	0.2152	(0.1818;0.2578)	(0.1913;0.2401)
12	20	300	0.2134	(0.1781;0.2692)	(0.1939;0.2337)
13	30	100	0.2172	(0.1830;0.2635)	(0.1897;0.2461)
14	30	150	0.2158	(0.1736;0.2603)	(0.1933;0.2393)
15	30	200	0.2174	(0.1769;0.2491)	(0.1978;0.2378)
16	30	300	0.2148	(0.1768;0.2524)	(0.1962;0.2200)

Table 6.5: *Simulation study. Univariate fixed-effects model for large trial sizes, trial-level surrogacy.*

	strategy			**confidence intervals**	
sim. #	# trials	# subjects	R^2_{trial}	**percentile**	**asymptotic**
1	5	100	0.8462	(0.3826;0.9966)	(0.4770;0.9799)
2	5	150	0.8575	(0.3961;0.9987)	(0.5016;0.9809)
3	5	200	0.8520	(0.3065;0.9981)	(0.5089;0.9756)
4	5	300	0.8864	(0.5697;0.9986)	(0.5570;0.9873)
5	10	100	0.7514	(0.3991;0.9823)	(0.4084;0.9347)
6	10	150	0.7791	(0.3896;0.9818)	(0.4570;0.9436)
7	10	200	0.8057	(0.4401;0.9729)	(0.4911;0.9531)
8	10	300	0.8199	(0.3564;0.9853)	(0.5174;0.9567)
9	20	100	0.7049	(0.2900;0.9236)	(0.4464;0.8761)
10	20	150	0.7123	(0.3697;0.9361)	(0.4554;0.8809)
11	20	200	0.7321	(0.4283;0.9588)	(0.4793;0.8924)
12	20	300	0.7503	(0.4652;0.9316)	(0.5058;0.9007
13	30	100	0.6780	(0.4351;0.8713)	(0.4528;0.8403)
14	30	150	0.6997	(0.5016;0.8793)	(0.4795;0.8541)
15	30	200	0.7304	(0.4631;0.9175)	(0.5211;0.8719)
16	30	300	0.7639	(0.5283;0.9317)	(0.5673;0.8913)

Table 6.6: *Simulation study. Bivariate fixed-effects model for large trial sizes, trial-level surrogacy.*

	strategy			**confidence intervals**	
sim. #	# trials	# subjects	R^2_{trial}	**percentile**	**asymptotic**
1	5	100	0.8500	(0.3796;0.9965)	(0.4828;0.9802)
2	5	150	0.8609	(0.4050;0.9988)	(0.5049;0.9817)
3	5	200	0.8592	(0.3285;0.9974)	(0.5140;0.9778)
4	5	300	0.8893	(0.5515;0.9996)	(0.5633;0.9877)
5	10	100	0.7735	(0.4328;0.9823)	(0.4385;0.9428)
6	10	150	0.7930	(0.3886;0.9867)	(0.4724;0.9491)
7	10	200	0.7930	(0.4443;0.9809)	(0.5177;0.9594)
8	10	300	0.8240	(0.4396;0.9805)	(0.5413;0.9632)
9	20	100	0.8375	(0.3261;0.9158)	(0.4829;0.8934)
10	20	150	0.7352	(0.4798;0.9307)	(0.5082;0.9043)
11	20	200	0.7545	(0.4747;0.9579)	(0.5348;0.9149)
12	20	300	0.7746	(0.5489;0.9360)	(0.5669;0.9244)
13	30	100	0.7126	(0.4928;0.8711)	(0.4948;0.8624)
14	30	150	0.7444	(0.5919;0.8856)	(0.5363;0.8815)
15	30	200	0.7764	(0.5390;0.9184)	(0.5819;0.8992)
16	30	300	0.8059	(0.6373;0.9204)	(0.6244;0.9155)

Table 6.7: *Simulation study. Bivariate fixed-effects model for large trial sizes, individual-level surrogacy.*

	strategy			**confidence intervals**
sim. #	# trials	# subjects	R^2_{indiv}	**percentile**
1	5	100	0.6843	(0.5058;0.8273)
2	5	150	0.6685	(0.5275;0.8218)
3	5	200	0.6642	(0.5141;0.7981)
4	5	300	0.6976	(0.5570;0.7440)
5	10	100	0.6521	(0.5497;0.7631)
6	10	150	0.6636	(0.5576;0.7748)
7	10	200	0.6527	(0.5872;0.7118)
8	10	300	0.6474	(0.5326;0.8485)
9	20	100	0.6640	(0.6009;0.7280)
10	20	150	0.6574	(0.5893;0.7339)
11	20	200	0.6517	(0.6041;0.7013)
12	20	300	0.6449	(0.6129;0.6910)
13	30	100	0.6682	(0.6066;0.7222)
14	30	150	0.6562	(0.6046;0.7109)
15	30	200	0.6495	(0.6136;0.7002)
16	30	300	0.6481	(0.6163;0.6845)

Table 6.8: *Simulation study. Bivariate mixed-effects model for large trial sizes, trial-level surrogacy.*

	strategy			**confidence intervals**
sim. #	# trials	# subjects	R^2_{indiv}	**percentile**
1	5	100	0.9433	(0.5005;1.0000)
2	5	150	0.9431	(0.4636;1.0000)
3	5	200	0.9477	(0.4611;1.0000)
4	5	300	0.9325	(0.5021;1.0000)
5	10	100	0.9291	(0.5706;0.9989)
6	10	150	0.9337	(0.6616;0.9996)
7	10	200	0.9306	(0.5499;0.9999)
8	10	300	0.9243	(0.4903;0.9998)
9	20	100	0.9236	(0.7458;0.9997)
10	20	150	0.9230	(0.7820;0.9996)
11	20	200	0.9196	(0.7602;0.9948)
12	20	300	0.9235	(0.7940;0.9977)
13	30	100	0.9152	(0.7932;0.9947)
14	30	150	0.9064	(0.7610;0.9963)
15	30	200	0.9079	(0.7914;0.9896)
16	30	300	0.9082	(0.7729;0.9984)

Table 6.9: *Simulation study. Bivariate mixed-effects model for large trial sizes, individual-level surrogacy.*

	strategy			**confidence intervals**
sim. #	# trials	# subjects	R^2_{indiv}	**percentile**
1	5	100	0.6863	(0.4814;0.9044)
2	5	150	0.6763	(0.5007;0.8370)
3	5	200	0.6681	(0.5305;0.7944)
4	5	300	0.6650	(0.5204;0.7967)
5	10	100	0.6379	(0.4949;0.7852)
6	10	150	0.6468	(0.5146;0.7929)
7	10	200	0.6359	(0.5409;0.7331)
8	10	300	0.6390	(0.5274;0.7345)
9	20	100	0.6376	(0.5555;0.7510)
10	20	150	0.6397	(0.5607;0.7421)
11	20	200	0.6398	(0.5692;0.7119)
12	20	300	0.6338	(0.5723;0.6929)
13	30	100	0.6392	(0.5695;0.7095)
14	30	150	0.6367	(0.5725;0.7006)
15	30	200	0.6398	(0.5779;0.7043)
16	30	300	0.6339	(0.5833;0.6804)

Table 6.10: *Simulation study. Univariate mixed-effects model for small trial sizes, individual-level surrogacy.*

	strategy			**confidence intervals**	
sim. #	# trials	# subjects	R^2_{indiv}	**percentile**	**asymptotic**
1	5	10	0.2661	(0.0108;0.6816)	(0.0917;0.5070)
2	5	20	0.2709	(0.0642;0.5322)	(0.1312;0.4442)
3	5	40	0.2407	(0.0998;0.3983)	(0.1390;0.3608)
4	5	60	0.2365	(0.1292;0.3504)	(0.1516;0.3340)
5	10	10	0.2990	(0.0902;0.4920)	(0.1503;0.4765)
6	10	20	0.2574	(0.1373;0.4244)	(0.1525;0.3793)
7	10	40	0.2448	(0.1434;0.3589)	(0.1696;0.3289)
8	10	60	0.2398	(0.1515;0.6434)	(0.1779;0.3078)
9	20	10	0.3090	(0.1824;0.4241)	(0.1955;0.4360)
10	20	20	0.2853	(0.1456;0.4078)	(0.2054;0.3727)
11	20	40	0.2497	(0.1764;0.3313)	(0.1947;0.3089)
12	20	60	0.2381	(0.1783;0.3023)	(0.1936;0.2856)
13	30	10	0.3362	(0.1964;0.4583)	(0.2394;0.4406)
14	30	20	0.2753	(0.1839;0.3524)	(0.2100;0.3456)
15	30	40	0.2513	(0.1985;0.3174)	(0.2059;0.2995)
16	30	60	0.2388	(0.1823;0.2788)	(0.2022;0.2775)

Table 6.11: *Simulation study. Univariate mixed-effects model for small trial sizes, trial-level surrogacy.*

	strategy			**confidence intervals**	
sim. #	# trials	# subjects	R^2_{trial}	**percentile**	**asymptotic**
1	5	10	0.7037	(0.0332;0.9949)	(0.3199;0.9293)
2	5	20	0.8350	(0.2446;0.9944)	(0.4456;0.9758)
3	5	40	0.8625	(0.2305;0.9951)	(0.5228;0.9744)
4	5	60	0.8880	(0.5288;0.9978)	(0.5531;0.9849)
5	10	10	0.7450	(0.3825;0.9375)	(0.3836;0.9347)
6	10	20	0.7981	(0.4189;0.9637)	(0.4691;0.9517)
7	10	40	0.8464	(0.4927;0.9605)	(0.5462;0.9675)
8	10	60	0.8545	(0.4582;0.9704)	(0.5751;0.9674)
9	20	10	0.7130	(0.4327;0.8718)	(0.4455;0.8845)
10	20	20	0.7895	(0.5521;0.9110)	(0.5502;0.9237)
11	20	40	0.8234	(0.5882;0.9447)	(0.6067;0.9383)
12	20	60	0.8427	(0.6495;0.9501)	(0.6374;0.9470)
13	30	10	0.7225	(0.4826;0.8965)	(0.5098;0.8674)
14	30	20	0.7803	(0.5781;0.9044)	(0.5862;0.9018)
15	30	40	0.8228	(0.6908;0.9179)	(0.6482;0.9250)
16	30	60	0.8414	(0.7109;0.9188)	(0.6761;0.9348)

Table 6.12: *Simulation study. Univariate fixed-effects model for small trial sizes, individual-level surrogacy.*

	strategy			**confidence intervals**	
sim. #	# trials	# subjects	R^2_{indiv}	**percentile**	**asymptotic**
1	5	10	0.2274	(0.0036;0.5809)	(0.0684;0.4663)
2	5	20	0.2376	(0.0489;0.4606)	(0.1072;0.4059)
3	5	40	0.2206	(0.0893;0.3844)	(0.1229;0.3382)
4	5	60	0.2203	(0.1263;0.3297)	(0.1379;0.3160)
5	10	10	0.2330	(0.0555;0.3973)	(0.1027;0.4009)
6	10	20	0.2092	(0.1091;0.3693)	(0.1144;0.3244)
7	10	40	0.2166	(0.1143;0.3334)	(0.1454;0.2978)
8	10	60	0.2213	(0.1388;0.3281)	(0.1614;0.2877)
9	20	10	0.2015	(0.0696;0.3304)	(0.1081;0.3157)
10	20	20	0.2236	(0.1084;0.3348)	(0.1513;0.3055)
11	20	40	0.2163	(0.1458;0.3014)	(0.1647;0.2731)
12	20	60	0.2142	(0.1569;0.2859)	(0.1755;0.2602
13	30	10	0.2231	(0.0886;0.3534)	(0.1413;0.3177)
14	30	20	0.2147	(0.1305;0.2945)	(0.1557;0.2804)
15	30	40	0.2174	(0.1668;0.2843)	(0.1746;0.2636)
16	30	60	0.2149	(0.1662;0.2593)	(0.1797;0.2523)

Table 6.13: *Simulation study. Univariate fixed-effects model for small trial sizes, trial-level surrogacy.*

	strategy			**confidence intervals**	
sim. #	# trials	# subjects	R^2_{trial}	**percentile**	**asymptotic**
1	5	10	0.74967	(0.2237;0.9994)	(0.3429;0.9572)
2	5	20	0.78668	(0.1593;0.9970)	(0.4210;0.9585)
3	5	40	0.81303	(0.1523;0.9995)	(0.4644;0.9675)
4	5	60	0.83546	(0.3234;0.9980)	(0.4714;0.9779)
5	10	10	0.67267	(0.2214;0.9198)	(0.3019;0.9035)
6	10	20	0.66771	(0.1851;0.9623)	(0.3120;0.8984)
7	10	40	0.71278	(0.1988;0.9589)	(0.3563;0.9190)
8	10	60	0.73415	(0.2825;0.9809)	(0.3988;0.9237)
9	20	10	0.63657	(0.2892;0.8413)	(0.3533;0.8409)
10	20	20	0.64009	(0.3417;0.8394)	(0.3567;0.8434)
11	20	40	0.66795	(0.3826;0.8884)	(0.3944;0.8579)
12	20	60	0.66832	(0.3344;0.9317)	(0.3996;0.8560)
13	30	10	0.63242	(0.3971;0.8232)	(0.3980;0.8109)
14	30	20	0.62815	(0.3844;0.8106)	(0.3913;0.8089)
15	30	40	0.64394	(0.4076;0.8313)	(0.4129;0.8179)
16	30	60	0.66659	(0.4111;0.8379)	(0.4408;0.8323)

Table 6.14: *Simulation study. Bivariate fixed-effects model for small trial sizes, trial-level surrogacy.*

	strategy			**confidence intervals**	
sim. #	# trials	# subjects	R^2_{trial}	**percentile**	**asymptotic**
1	5	10	0.7975	(0.2130;0.9995)	(0.3842;0.9699)
2	5	20	0.8059	(0.1797;0.9975)	(0.4409;0.9635)
3	5	40	0.8148	(0.2834;0.9997)	(0.4572;0.9699)
4	5	60	0.8427	(0.3645;0.9988)	(0.4844;0.9795)
5	10	10	0.6749	(0.1515;0.9104)	(0.3131;0.9013)
6	10	20	0.6512	(0.2183;0.9637)	(0.2993;0.8905)
7	10	40	0.7099	(0.2898;0.9607)	(0.3543;0.9172)
8	10	60	0.7392	(0.2288;0.9773)	(0.3999;0.9268)
9	20	10	0.5941	(0.1759;0.8188)	(0.3082;0.8139)
10	20	20	0.6197	(0.2706;0.8273)	(0.3348;0.8307)
11	20	40	0.6709	(0.3847;0.8803)	(0.3967;0.8601)
12	20	60	0.6833	(0.3681;0.9425)	(0.4170;0.8646)
13	30	10	0.5675	(0.3313;0.7811)	(0.3280;0.7654)
14	30	20	0.6147	(0.4083;0.8148)	(0.3750;0.8004)
15	30	40	0.6505	(0.4540;0.8338)	(0.4199;0.8226)
16	30	60	0.6834	(0.4700;0.8532)	(0.4602;0.8435)

Table 6.15: *Simulation study. Bivariate fixed-effects model for small trial sizes, individual-level surrogacy.*

	strategy			**confidence intervals**
sim. #	# trials	# subjects	R^2_{indiv}	**percentile**
1	5	10	0.8298	(0.0492;0.9999)
2	5	20	0.7763	(0.2469;0.9999)
3	5	40	0.7099	(0.4134;0.9999)
4	5	60	0.6976	(0.4737;0.9999)
5	10	10	0.8461	(0.4281;0.9999)
6	10	20	0.7459	(0.4346;0.9999)
7	10	40	0.7110	(0.4852;0.9181)
8	10	60	0.6970	(0.5326;0.8485)
9	20	10	0.8179	(0.4765;0.9999)
10	20	20	0.7662	(0.4927;0.9187)
11	20	40	0.6967	(0.5409;0.8161)
12	20	60	0.6729	(0.5835;0.7787)
13	30	10	0.8487	(0.5242;0.9999)
14	30	20	0.7529	(0.4083;0.8148)
15	30	40	0.6995	(0.6036;0.8002)
16	30	60	0.6755	(0.5984;0.7589)

Table 6.16: *Simulation study. Bivariate mixed-effects model for small trial sizes, trial-level surrogacy.*

	strategy			**confidence intervals**
sim. #	# trials	# subjects	R^2_{trial}	**percentile**
1	5	10	0.9749	(0.6438;1.0000)
2	5	20	0.9556	(0.4628;1.0000)
3	5	40	0.9477	(0.5341;1.0000)
4	5	60	0.9349	(0.5049;1.0000)
5	10	10	0.9114	(0.5373;1.0000)
6	10	20	0.9252	(0.5364;1.0000)
7	10	40	0.9345	(0.6964;0.9999)
8	10	60	0.9263	(0.4939;0.9999)
9	20	10	0.9078	(0.5867;0.9999)
10	20	20	0.9321	(0.7076;0.9996)
11	20	40	0.9209	(0.6956;0.9998)
12	20	60	0.9240	(0.7508;0.9997)
13	30	10	0.9231	(0.6530;0.9999)
14	30	20	0.9189	(0.7196;0.9993)
15	30	40	0.9125	(0.7512;0.9937)
16	30	60	0.9142	(0.7970;0.9996)

Table 6.17: *Simulation study. Bivariate mixed-effects model for small trial sizes, individual-level surrogacy.*

	strategy			**confidence intervals**
sim. #	# trials	# subjects	R^2_{indiv}	**percentile**
1	5	10	0.8088	(0.1167;1.0000)
2	5	20	0.7758	(0.2494;1.0000)
3	5	40	0.7033	(0.4431;0.9936)
4	5	60	0.6993	(0.4511;0.9491)
5	10	10	0.7632	(0.2686;1.0000)
6	10	20	0.6562	(0.3415;0.9044)
7	10	40	0.6611	(0.4329;0.8421)
8	10	60	0.9243	(0.4708;0.8369)
9	20	10	0.6605	(0.3033;0.9413)
10	20	20	0.6567	(0.4167;0.8494)
11	20	40	0.6608	(0.4904;0.8182
12	20	60	0.6430	(0.5176;0.7935)
13	30	10	0.6340	(0.3267;0.8577)
14	30	20	0.6536	(0.4611;0.8232)
15	30	40	0.6366	(0.5332;0.7540)
16	30	60	0.6349	(0.5445;0.7378)

7

Cross-Sectional Surrogate for Time-to-Event True Endpoint

In the last couple of years, a considerable amount of research has been devoted to the statistical validation of surrogate endpoints of various types. The information-theoretic approach by Alonso *et al.* (2007) has shed light on the possibility of using a unified platform for the validation of surrogate endpoints of both normally as well as non-normally distributed outcomes. The performance of this method, for time-to-event outcomes, however, has been less optimal, specially in the presence of substantial censoring. In this chapter we will compare the performance of the information-theoretic approach of Alonso *et al.* (2007) with the measure of explained variation by Kent and O'Quigely (1988) and that of Xu and O'Quigely (1999) through a simulation study and later the methods will be applied to a case study.

7.1 The Information-theoretic Approach for Time-to-Event Endpoint

The information-theoretic approach suggested in the previous chapters can be applied to a case of time-to-event true endpoint with slight modification. The univariate models used to relate the expected value of the true endpoint to the treatment only and to the surrogate endpoint and treatment can be altered to two appropriate models for survival type outcome. Two such models could be the Cox-proportional hazards model or an accelerated failure time model. If we chooses the Cox-proportional hazards model, we can consider models (7.1) and (7.2). Then, depending on wether we consider the number of subjects or number of events, we will end up with two different measures of associations. Lets denote the two measures by R_k^2 and R_n^2 with k and n representing the number of events and number of subjects respectively.

$$\lambda(t|z, s : \beta) = \lambda_0(t) \exp(\beta_1 z + \beta_2 s) \tag{7.1}$$

$$\lambda(t|z : \beta) = \lambda_0(t) \exp(\beta_1 z) \tag{7.2}$$

.

$$R_k^2 = 1 - \exp\left(\frac{-G^2}{k}\right) \tag{7.3}$$

$$R_n^2 = 1 - \exp\left(\frac{-G^2}{n}\right), \tag{7.4}$$

where G^2 is the likelihood ratio statistics to compare the two models and S and T represent the surrogate and true endpoints respectively.

7.2 The Kent and O'Quigley Measure of Explained Variation

The survival analysis context adds a complication due to the presence of censoring. In the absence of censored observations, a standard estimate of information gain will be provided by the inverse of the number of subjects times the usual likelihood ratio statistic which is equivalent to the Likelihood reduction factor approach of Alonso *et al.* (2007). However, censoring is found to have a substantial effect on this measure. Another alternative approach is a method due to Kent and O'Quigley (1988). These authors have introduced a measure of explained variation for censored survival outcome using the concept of the information gain approach of Kent (1983). They developed these ideas, obtaining simple, multiple and partial coefficients for the situation of proportional hazards regression. Their approach was based upon the idea of transforming a general proportional hazards model to a specific one of Weibull form. In this section we will outline this measure as discussed in Kent and O'Quigely (1988). Without loss of generality lets first assume that there are no censored observations. Consider two random variables X and Y and let G(x) denote the marginal distribution of X and let the conditional distribution of Y given X be modelled by:

$$Y = -\sigma\mu - \sigma\beta^t X + \sigma\varepsilon, \tag{7.5}$$

where the error ε follows some specified distribution with probability density function $f(y)$ and it is independent of X. If we assume a normal distribution with a zero mean and a unit variance for the error term, the above model represents the usual linear regression. If however, we chose a Gumbel density for f, it gives a weibull regression model for $T = e^Y$ where T is the survival time. Now let $\Theta = (\beta, \mu, \sigma^2)$ denote the parameters of the model with $\sigma > 0$ and $\beta = (\beta_{\mathbf{1}}, \beta_{\mathbf{2}})$ a 2-dimensional vector. Let

$\Theta_1 = (\beta, \mu, \sigma^2)$ denote the true values of the parameters. Consider two hypotheses $\mathbf{H_0}$: $\beta_\mathbf{1} = \mathbf{0}$ and $\mathbf{H_1}$: no restrictions on β. The objective here is to measure the dependence between Y and X_1 after allowing the regression on X_2. Now denote Θ_0 be the value of Θ maximizing the expected log-likelihood:

$$\Phi(\Theta, \Theta_1) = \int\int \log\{f(y|x;\theta)\} f(y|x;\theta_1) dy G(dx) \tag{7.6}$$

over Θ satisfying the null hypothesis. A measure of the distance between the null hypothesis and the alternative hypothesis is given by the Kullback & Liebler information gain

$$\Gamma = \Gamma(\mathbf{H_1}, \mathbf{H_0}, \mathbf{G}) = \mathbf{2}\{\mathbf{\Phi}(\theta_\mathbf{1};\theta_\mathbf{1}) - \mathbf{\Phi}(\theta_\mathbf{1};\theta_\mathbf{0})\}. \tag{7.7}$$

Following this construction, Kent(1983) proposed (7.8) as a measure of dependence between Y and X_1 after allowing the regression on X_2.

$$\rho_W^2 = 1 - e^{-\Gamma}. \tag{7.8}$$

Note that, Kent and O'Quigely denoted their dependence measure by ρ_W^2 in order to emphasize the relationship to the Weibull distribution. They stated that in principle other possible accelerated failure time and proportional hazard models could be considered. The reasons they site for the choice of the weibull distribution is that, the weibull distribution results in a tractable expected log-likelihood and can be viewed as a proportional hazard model. Now within the context of surrogate marker validation involving a time-to-event true endpoint and a cross-sectional surrogate endpoint, we can calculate the measure suggested by setting T=Y, S=X_1 and Z=X_2. A detailed account of the method can be found in Kent and O'Quigely (1988).

7.3 Xu and O'Quigley Measure of Explained Variation

The Kent and O'Quigley measure of explained variation discussed earlier suffers from two main drawbacks: computational complexity and its inability to accommodate time-dependent covariates. Citing these shortcomings, Xu and O'Quigley (1999) developed a similar measure based on information gain and using the conditional distribution of the covariates given the failure times. This measure accommodates time-dependent covariates and is computable using standard softwares for fitting a Cox model. Extensions to multiple covariates are immediate.

Despite the fact that it is computationally simple and can accommodate time-varying covariates, Xu and O'Quigley argue that, one important difficulty with the approach of Alonso *et al.* (2007) is how to adequately deal with censoring. They state that for low levels of censoring this may not be an issue of much concern but for high levels it would be useful to have a coefficient that explains the proportion of variation captured by the surrogate and which is not impacted by the censoring mechanism. This is in fact quite easily achieved and amounts to working with the same quantities described in Alonso *et al.* (2007) and weighting them differently Xu and O'Quigely (2005).

Let S, T and Z denote the surrogate, time-to-event true endpoint and the binary treatment indicator respectively as defined before. Now lets outline the steps involved in quantifying the Xu and O'Quigley measure. First consider models (7.1) and (7.2). Now from the partial likelihood estimates under (7.1), we can compute $\pi_j(t;\hat{\beta})$ and $\pi_j(t;0)$ as follows

$$\pi_j(t;\hat{\beta}) = \frac{y_j(t)\exp(\hat{\beta}_1 z + \hat{\beta}_2 s)}{\sum_{j=1}^{n} y_j(t)\exp(\hat{\beta}_1 z + \hat{\beta}_2 s)}, \tag{7.9}$$

$$\pi_j(t;0) = \frac{y_j(t)}{\sum_{j=1}^{n} y_j(t)} \tag{7.10}$$

where $y_j(t)$ is the risk indicator. Then we compute $P(t_i)$, the jump of the Kaplan-Meier curve at time t_i. Now consider the quantity:

$$\hat{\Gamma}_2(\hat{\beta}) = 2\sum_{i=1}^{k} P(t_i)\sum_{j=1}^{n} \pi_j(t;\hat{\beta})\log(\frac{\pi_j(t;\hat{\beta})}{\pi_j(t;0)}) \tag{7.11}$$

from which we can compute $\rho^2_{Z,S}$=1 − exp($-\hat{\Gamma}_2(\hat{\beta})$). In a similar manner, using the partial likelihood estimates under (7.2), we can proceed to compute $\pi_j(t;\hat{\beta})$ and $\pi_j(t;0)$ as follows:

$$\pi_j(t;\hat{\beta}) = \frac{y_j(t)\exp(\hat{\beta}_1 z)}{\sum_{j=1}^{n} y_j(t)\exp(\hat{\beta}_1 z)}, \tag{7.12}$$

$$\pi_j(t;0) = \frac{y_j(t)}{\sum_{j=1}^{n} y_j(t)}. \tag{7.13}$$

Following the above procedures, compute $\hat{\Gamma}_2(\hat{\beta})$ from which ρ^2_z= $1-\exp(-\hat{\Gamma}_2(\hat{\beta}))$ can be computed. Finally using the relationship that $1-\rho^2_{z,s}$=$(1-\rho^2_z)\times(1-\rho^2_{s|z})$, we will be able to quantify the desired measure of association which is $\rho^2_{s|z}$ from the quantities calculated above. We will denote the measure of Xu and O'Quigely by ρ^2_{xu}. For a detailed description of the method please refer to Xu and O'Quigely(1999) and Xu and O'Quigely (2005).

7.4 A Simulation Study

In the next section, insight into the performance of the information-theoretic approach of Alonso *et al.* (2007) together with the measure of explained variation by Kent

and O'Quigely (1988) and Xu and O'Quigley (1999), is offered through a simulation study. We first lay out the design of the simulation study, whereafter the results are described.

7.4.1 Design of the Simulation Study

The focus of this section is to design a simulation study to compare the performance of the methods discussed in the previous sections. To simplify matters and for ease of comparison with the linear correlation coefficient and moreover to asses the robustness of the models against the violation of the proportional hazard assumption, we will assume a log-normal distribution for the the time-to-event outcome. This allows us to generate data easily from a bivariate normal distribution assuming a normal distribution for the cross-sectional surrogate endpoint. After generating outcomes in this format, the survival outcome is obtained by taking the exponential of the resulting continuous random variable. Censoring was introduced using a uniform distribution. The percentage of censored observations was set be either 0, 10, or 35 which we considered as the small to moderate level of censoring and another set which contains 50, 75 or 90 percent censored observations which represents a high to extreme number of censored observations. The individual-level R^2 values were set to be 0.36, 0.64, or 0.81. The number of subjects was fixed to be either 20, 50, 100, 200 or 1000. In each case, 100 runs were performed, where the three methods discussed are computed.

7.4.2 Simulation Results

The simulation results are displayed in Tables 7.1 and 7.2. For small to moderate level of censoring, all the methods seem to perform adequately with the estimated measures approaching the true value from which the data were generated with increase in sample size. For percentages of censoring ranging between 35% to 50%, R_k^2 and

ρ^2_{xu} tend to slightly overestimate the association measure whereas the R^2_n provides underestimated association measure even for substantially large sample size. The overestimation of the R^2_k and ρ^2_{xu} becomes larger as we move to 75% censoring and gets worse when the percentage of censoring is as high as 90%, while the R^2_n pointing in the opposite direction. The ρ^2_w results in underestimated association measures as the percentage of censoring gets larger, but the underestimation subsides as the sample size increases except for high level of association between the surrogate and the true endpoint. Note also that, as theoretically expected, with no censoring, the R^2_k and ρ^2_w didn't give similar results.

7.5 Application to the Case Study

The three measures of association discussed earlier were applied to the case study in stroke of children with sickle cell disease introduced in Chapter 2, Section 2.1.5. The response of interest was time to first stroke for which several competing surrogate endpoints were considered. From the results summarized in Table 7.3, we can learn that none of the potential surrogates have a reasonable degree of association with the response. R^2_k and ρ^2_{xu} produced relatively higher measures of association, while ρ^2_w resulted in very small measures in agreement with the results of the simulation study. Taking the fact that there is a 90% censoring into consideration and inline with the results of the simulation study, we tend to trust the ρ^2_w measure to provide a reasonable estimate as compared to the other two measures of association. Our conclusion based on this association measure reflects absence of overlapping information between the potential surrogates and the true endpoint.

7.6 Discussion

In this chapter we have compared the performance of three information theory based methods for the evaluation of a cross-sectional surrogate for a time-to-event true endpoint when the assumption of proportional hazard is violated. Substantial literature exists concerning the performance of these methods when the proportional hazard assumption is full filled. Within the context of validating a time-to-event surrogate endpoint for a time-to-event true endpoint, a small simulation study was performed to compare the performance of the Xu and O'Quigely measure against the measure suggested by Alonso *et al.* (2007) for varying percentages of censoring. The simulation results have revealed that for most instances the Xu and O'Quigely measure outperforms the measure of Alonso *et al.* (2007) given that the proportional hazard assumption is satisfied. Schemper and Stare (1996) compared several measures of explained variation including the Kent and O'Quigely measure of association for a Cox proportional hazards model. They have found that among the other methods suggested, the measure of Kent and O'Quigely was unaffected by censoring even for substantial percentage of censoring. It is therefore possible to assume that, except the information-theoretic approach of Alonso *et al.* (2007), the other two methods have given very promising results even for large percentage of censoring under the assumption of proportional hazard. Note however that, not many studies have been conducted to see how these methods fair when the proportional hazards assumption is questionable. The first method due to Alonso *et al.* (2007), which was found to perform well for many non-normally distributed outcomes, seems less adept for the case of survival outcomes. This is specially true for the case of excess censoring and small sample sizes, which was also the case even when the proportional hazard assumption was full filled. The Kent and O'Quigely measure is found to perform very well even

for moderately large percentage of censoring for reasonable number of observations. A noticeable drawback of the method by Kent and O'Quigely, apart from its computational difficulty, is its inadequacy to accommodate time varying covariates. The third method which is due to Xu and O'Quigely is more flexible in terms of allowing the inclusion of time-varying covariates. This method however, is highly dependent on the proportional hazard assumption. It was however found that, even when the proportional hazard assumption does not hold, the method has given acceptable level of bias for large sample sizes. Of the three methods, the method by Alonso *et al.* (2007) has the least computational difficulty. The version of this method which uses the number of events rather than the total number of subjects is to be preferred. This version was found to perform well in small and medium level of censoring and large sample sizes. The methods were evaluated for the case of single trial setting, however all of them can be used for the meta-analytic setting with little modification. In conclusion, for a case of cross-sectional surrogate with no time varying covariates, the measure of Kent and O'Quigely is a promising choice even when the proportional hazard assumption is questionable. If we resort our attention to the case study, we can learn that none of the potential surrogates have enough information to be able to predict the true endpoint.

Table 7.1: *Simulation results for 0%, 15% and 35% censored observations. (n: Sample size ; R^2_k: R^2 based on ITA with number of events is denominator; R^2_n: R^2 based on ITA with number of subjects as denominator; ρ^2_w: R^2 based on the Kent and O'Quigely measure of dependence; ρ^2_{xu} : R^2 based on the Xu and O'Quigely measure of dependence;*

	Censoring =0%				Censoring =15%				Censoring =35 %			
					$R^2 = 0.36$							
n	R^2_k	R^2_n	ρ^2_w	ρ^2_{xu}	R^2_k	R^2_n	ρ^2_w	ρ^2_{xu}	R^2_k	R^2_n	ρ^2_w	ρ^2_{xu}
20	0.3459	0.3459	0.4048	0.3467	0.3782	0.3329	0.4156	0.3633	0.4245	0.3018	0.4250	0.4027
50	0.3329	0.3329	0.3690	0.3337	0.3616	0.3201	0.3814	0.3486	0.4097	0.2870	0.3972	0.3911
100	0.3292	0.3292	0.3572	0.3298	0.3540	0.3105	0.3649	0.3448	0.3956	0.2752	0.3777	0.3799
200	0.3334	0.3334	0.3522	0.3339	0.3643	0.3216	0.3674	0.3562	0.4110	0.2890	0.3832	0.3975
1000	0.3333	0.3333	0.3465	0.3338	0.3572	0.3152	0.3552	0.3525	0.4032	0.2812	0.3697	0.3949
					$R^2 = 0.64$							
20	0.5904	0.5904	0.6686	0.5913	0.6213	0.5634	0.6726	0.6030	0.6696	0.5120	0.6714	0.6431
50	0.5965	0.5965	0.6418	0.5974	0.6266	0.5713	0.6495	0.6076	0.6795	0.5156	0.6602	0.6524
100	0.5992	0.5992	0.6314	0.5999	0.6301	0.5723	0.6366	0.6157	0.6825	0.5195	0.6493	0.6569
200	0.6088	0.6088	0.6310	0.6094	0.6430	0.5859	0.6420	0.6301	0.6907	0.5311	0.6519	0.6649
1000	0.6122	0.6122	0.6267	0.6129	0.6418	0.5849	0.6335	0.6321	0.6908	0.5278	0.6425	0.6744
					$R^2 = 0.81$							
20	0.7436	0.7436	0.8203	0.7445	0.7640	0.7090	0.8151	0.7454	0.8028	0.6508	0.7984	0.7755
50	0.7692	0.7692	0.8105	0.7700	0.7942	0.7423	0.8157	0.7761	0.8329	0.6792	0.8209	0.8043
100	0.7771	0.7771	0.8042	0.7777	0.8022	0.7495	0.8071	0.7876	0.8376	0.6876	0.8102	0.8121
200	0.7873	0.7873	0.8046	0.7878	0.81337	0.7620	0.8101	0.8009	0.8469	0.7012	0.8133	0.8240
1000	0.7911	0.7911	0.8006	0.7917	0.81420	0.7635	0.8036	0.8036	0.8493	0.7019	0.8069	0.8328

Table 7.2: *Simulation results for 50%, 75% and 90% censored observations. (n: Sample size ; R_k^2: R^2 based on ITA with number of events as denominator; R_n^2: R^2 based on ITA with number of subjects as denominator; ρ_w^2: R^2 based on the Kent and O'Quigely measure of dependence; ρ_{xu}^2 : R^2 based on the Xu and O'Quigely measure of dependence;*

	Censoring =50%				Censoring =75 %				Censoring =90 %			
					$R^2 = 0.36$							
n	R_k^2	R_n^2	ρ_w^2	ρ_{xu}^2	R_k^2	R_n^2	ρ_w^2	ρ_{xu}^2	R_k^2	R_n^2	ρ_w^2	ρ_{xu}^2
20	0.4496	0.2741	0.4336	0.4259	0.5294	0.1980	0.3516	0.4915	0.5444	0.1081	0.2231	0.005
50	0.4331	0.2580	0.4031	0.4104	0.5142	0.1779	0.3765	0.4889	0.5941	0.1024	0.1868	0.5900
100	0.4302	0.2514	0.3887	0.4127	0.5190	0.1746	0.3777	0.4907	0.6622	0.1067	0.2522	0.6317
200	0.4357	0.2586	0.3847	0.4198	0.5319	0.1749	0.3818	0.5078	0.6396	0.1006	0.3098	0.6118
1000	0.4333	0.2528	0.3728	0.4231	0.5196	0.1655	0.3618	0.5074	0.6415	0.0933	0.3419	0.6251
					$R^2 = 0.64$							
20	0.6793	0.4568	0.6574	0.6457	0.7232	0.3103	0.4790	0.6560	0.7560	0.1736	0.3213	0.7883
50	0.7013	0.4681	0.6605	0.6668	0.7649	0.3238	0.5958	0.7259	0.8062	0.1810	0.3297	0.7984
100	0.7107	0.4717	0.6534	0.6850	0.7900	0.3288	0.6311	0.7803	0.8717	0.1939	0.4267	0.8351
200	0.7201	0.4858	0.6536	0.6952	0.8008	0.3339	0.6420	0.7741	0.8749	0.1911	0.5206	0.8435
1000	0.7212	0.4807	0.6436	0.7024	0.7997	0.3270	0.6260	0.7803	0.8819	0.1845	0.5932	0.8635
					$R^2 = 0.81$							
20	0.8136	0.5897	0.7851	0.7670	0.8357	0.4019	0.6334	0.7508	0.8299	0.2194	0.4157	0.8455
50	0.8485	0.6244	0.8206	0.8205	0.8864	0.4371	0.7501	0.8540	0.8960	0.2414	0.4052	0.8277
100	0.8570	0.6317	0.8119	0.8275	0.9056	0.4522	0.7774	0.8802	0.9492	0.2644	0.5719	0.9290
200	0.8660	0.6500	0.8132	0.8438	0.9167	0.4642	0.8017	0.8933	0.9547	0.2694	0.6412	0.9331
1000	0.8700	0.6487	0.8062	0.8520	0.9181	0.4598	0.7901	0.9009	0.9597	0.2632	0.7609	0.9464

Table 7.3: *Results of the case study. (R_k^2: R^2 based on ITA with number of events as denominator; R_n^2: R^2 based on ITA with number of subjects as denominator; ρ_w^2: R^2 based on the Kent and O'Quigely measure of dependence; ρ_{xu}^2 : R^2 based on the Xu and O'Quigely measure of dependence*

Surrogate	R_k^2	R_n^2	ρ_w^2	ρ_{xu}^2
Maximum diastolic TCD velocity	0.5594	0.0787	0.0231	0.6777
Maximum mean TCD velocity on right	0.4718	0.0618	0.0142	0.6824
Maximum systolic TCD velocity	0.6245	0.0933	0.0215	0.6944
Maximum of maximum mean TCD velocity on left and right	0.6448	0.0983	0.0245	0.6727
Difference between maximum systolic and dystolic TCD velocity	0.3728	0.0456	0.0340	0.2999
Maximum mean TCD velocity	0.6245	0.0933	0.0215	0.7491

8

Mixture of Longitudinal and Cross-Sectional Endpoints

Thus far, we have considered the case of two cross-sectionally measured outcomes and tried to quantify the individual and trial level surrogacy measures. In practice however, we will encounter cases where one or both of the endpoints of interest is multivariate in nature and specifically longitudinally measured. This induces a new challenge in terms of quantifying the desired measures of association. The methods that have been suggested for the univariate cases may not be directly applicable to this situation. In addition to adapting the methods to the case of longitudinal outcomes, we are also challenged with fitting the appropriate model for the time course. The concern of this chapter is therefore to revisit the methods suggested for the case of two continuous longitudinal outcomes and adapt them for cases where either of the two outcomes is cross-sectionally measured. We start by giving a concise description of the various methods used in validating a surrogate endpoint, for the

case of two longitudinal outcomes and then adapt the methods to the case of a mixture of longitudinal and cross-sectional outcomes, with focus on individual level surrogacy.

8.1 Measures for Two Longitudinal Outcomes

We begin with the review of the variance reduction factor and the R^2_Λ, suggested by Alonso *et al.* (2003) for the case of two repeatedly measured outcomes, where after we show how theses methods can be adapted to the situation where one of the two outcomes is cross-sectional. Let us assume that there are n subjects enrolled for a particular study and further suppose that t_{jk} is the time at which the kth measurement of the jth subject is taken. Let T_{jk} and S_{jk} be the true and the surrogate endpoints, respectively, and let Z_j be a binary treatment indicator. Now, consider the following joint model for the true and surrogate endpoints:

$$\begin{aligned} T_{jk} &= \mu_T + \alpha Z_j + f(t_{jk}) + \varepsilon_{Tjk}, \\ S_{jk} &= \mu_S + \beta Z_j + f(t_{jk}) + \varepsilon_{Sjk}, \end{aligned} \tag{8.1}$$

where (μ_T, μ_S, α,β) are intercepts and treatment effects on the true and surrogate endpoints, respectively, $f(t_{jk})$ is a flexible function in time which can be modeled as fractional polynomial, penalized spline, or any flexible function in time. In principle, it is possible for the two endpoints to depend on time through different functions, in which case we will have $f_T(t_{jk})$ and $f_S(t_{jk})$ for the true and surrogate endpoint respectively. However, without loss of generality, let us assume that both depend on time through the same function. The error terms (ε_{Tjk}, ε_{Sjk}) are assumed to follow a zero-mean normal distribution with patterned variance-covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_{TT} & \Sigma_{TS} \\ \Sigma_{ST} & \Sigma_{SS} \end{pmatrix}, \tag{8.2}$$

with obvious notation. In this setting, Alonso *et al.* (2003) proposed to quantify the individual-level surrogacy using the so-called *variance reduction factor*, which is

defined as

$$VRF = \frac{\text{tr}(\Sigma_{TT}) - \text{tr}(\Sigma_{T|S})}{\text{tr}(\Sigma_{TT})}, \tag{8.3}$$

where $\Sigma_{T|S}$ denotes the conditional variance-covariance matrix of T_{jk} given S_{jk}, i.e., $\Sigma_{T|S} = \Sigma_{TT} - \Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}$. Furthermore, these authors have shown that the VRF satisfies a set of properties that makes it practically applicable: (i) VRF ranges between zero and one; (ii) $VRF = 0$ if and only if the true and the surrogate endpoints are independent; (iii) $VRF = 1$ if and only if there exists a deterministic relationship between the true and surrogate endpoint; and (iv) $VRF = R^2$ in the cross-sectional setting. Note that, at the individual level, interest lies in the prediction of the true endpoint given the surrogate endpoint. In this regard, property (ii) shows that if the VRF equals zero, then no sensible prediction is possible, whereas a perfect prediction is attained if VRF equals one, as indicated by property (iii). Property (iv) establishes the link between this approach and the one suggested by Buyse *et al.* (2000) for univariate outcomes. As can be seen from (8.3), the VRF summarizes the variability of the two endpoints using the trace of the corresponding variance-covariance matrices. In multivariate analysis, there is no unique way of defining a generalized variance, the trace is one of the classical ways of doing so, while another common definition uses the determinant. Interestingly, using the trace or the determinant to summarize the variability of the endpoints has important ramifications for analysis and leads to two totally separate measures with different interpretations. To this end, Alonso *et al.* (2003) have suggested another measure, the so-called R^2_Λ, which uses this alternative definition of the generalized variance. Like the VRF, this measure can be derived based on Model (8.1), as follows:

$$R^2_\Lambda = 1 - \frac{|\Sigma|}{|\Sigma_{TT}| \cdot |\Sigma_{SS}|}. \tag{8.4}$$

The authors have shown that this measure also enjoys desirable properties: (i) R^2_Λ is symmetric and invariant with respect to linear bijective transformations; (ii) R^2_Λ ranges between zero and one; (iii) $R^2_\Lambda = 0$ if and only if the error terms are independent; (iv) $R^2_\Lambda = 1$ if and only if there exist a and b so that $a^T\varepsilon_{S_{jk}} = b^T\varepsilon_{T_{jk}}$ with probability one; and (v) $R^2_\Lambda = R^2$ in the cross-sectional setting. All of these properties, except the fourth property are shared with the VRF. The fourth property, however, differs in important ways from the VRF. Indeed, whereas the VRF takes the value 1 when there is a deterministic relationship between both endpoints, R^2_Λ is 1 whenever there is a deterministic relationship between two linear combinations of both endpoints, allowing us to uncover strong association in cases where the VRF might fail to do so. This is not a disadvantage of one or the other proposal, but rather underscores them focusing on different aspects. The expression for R^2_Λ clearly shows that, unlike the VRF, this measure treats both endpoints symmetrically.

8.2 A Longitudinal Surrogate for a Cross-Sectional True Endpoint

Let us assume that the surrogate endpoint is repeatedly measured over time with K repeated measures and that the true endpoint is cross-sectional. Model (8.1) takes the form:

$$\begin{aligned} T_j &= \mu^*_T + \alpha^* Z_j + \varepsilon_{Tj}, \\ S_{jk} &= \mu^*_S + \beta^* Z_j + f(t_{jk}) + \varepsilon_{Sjk}. \end{aligned} \tag{8.5}$$

Notice that there are some important differences between (8.5) and the joint model for two longitudinal outcomes given in (8.1). One dissimilarity is that there is a difference in the number of parameters when modeling the surrogate and true endpoints. A second one, a computational issue is induced such that, the variance-covariance matrix of the error term $(\varepsilon_{S_{ij}}, \varepsilon_{T_{ij}})^T$, Σ, cannot be modeled using a Kronecker product of

two matrices like suggested in Galecki (1994), as there are no repeated measurements within the true endpoint. Thus, Σ has to be modeled as one matrix using either a compound symmetry, first-order autoregressive, spatial or another type of covariance structure. Nevertheless, Σ can still be subdivided into four sub-matrices, i.e.,

$$\Sigma = \begin{pmatrix} \sigma_{TT} & \Sigma_{TS} \\ \Sigma_{ST} & \Sigma_{SS} \end{pmatrix}. \tag{8.6}$$

Here, σ_{TT} denotes the variance of the true endpoint, Σ_{TS} is a $(1 \times K)$ vector containing the covariances between the true endpoint and the surrogate endpoint at different time points, and Σ_{SS} is a $(K \times K)$ variance-covariance matrix associated with the longitudinal surrogate endpoint. Then, the VRF_{indiv} for longitudinal surrogate and a cross-sectional true endpoint denoted by VRF_{ST}^{LC}, with a subscript 'L' ('C') reminiscent of 'longitudinal' ('cross-sectional'), can be computed as

$$VRF_{ST}^{LC} = \frac{\text{tr}(\sigma_{TT}) - \text{tr}(\sigma_{T|S})}{\text{tr}(\sigma_{TT})}, \tag{8.7}$$

where $\sigma_{T|S}$ denotes the conditional variance of T given S: $\sigma_{T|S} = \sigma_{TT} - \Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}$. Using this expression, (8.7) can be re-written as

$$VRF_{ST}^{LC} = \frac{\text{tr}(\sigma_{TT}) - \text{tr}(\sigma_{TT} - \Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST})}{\text{tr}(\sigma_{TT})}. \tag{8.8}$$

Note that all matrices involved in the computation of VRF_{ST}^{LC} are of dimension (1×1) and hence the trace reduces to the corresponding scalar, offering the opportunity to simplify (8.8):

$$VRF_{ST}^{LC} = \frac{\sigma_{TT} - \sigma_{TT} + \Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}}{\sigma_{TT}} = \frac{\Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}}{\sigma_{TT}}. \tag{8.9}$$

Notice that $VRF_{ST}^{LC} = 0$ if and only if $\Sigma_{ST} = 0$, i.e., if and only if when S and T are independent.

Intuitively, (8.9) quantifies how much of the total variability of the true endpoint is explained by the surrogate endpoint, after adjusting for treatment effects and repeated measures of the surrogate endpoint. Resorting our attention to R^2_Λ, let us again consider Model (8.5) and the corresponding variance-covariance matrix (8.6). The R^2_Λ for a longitudinal surrogate and a cross-sectional endpoint is given by

$$R^{2,LC}_{\Lambda,ST} = 1 - \frac{|\Sigma\,|}{|\sigma_{TT}\,| \cdot |\Sigma_{SS}\,|}, \tag{8.10}$$

where σ_{TT}, Σ_{SS}, and Σ are as defined in (8.6). Note that

$$|\Sigma| = |\Sigma_{SS}| \cdot |\Sigma_{T|S}| = |\Sigma_{SS}| \cdot |\sigma_{TT} - \Sigma_{TS}\Sigma^{-1}_{SS}\Sigma_{ST}|,$$

and, substituting this in (8.10), we obtain

$$\begin{aligned} R^{2,LC}_{\Lambda_{ST}} &= 1 - \frac{|\sigma_{TT} - \Sigma_{TS}\Sigma^{-1}_{SS}\Sigma_{ST}\,|}{|\sigma_{TT}\,|} \\ &= 1 - \frac{\sigma_{TT} - \Sigma_{TS}\Sigma^{-1}_{SS}\Sigma_{ST}}{\sigma_{TT}} \\ &= \frac{\Sigma_{TS}\Sigma^{-1}_{SS}\Sigma_{ST}}{\sigma_{TT}}, \end{aligned} \tag{8.11}$$

since all matrices involved are of dimension one.

8.3 A Cross-Sectional Surrogate for a Longitudinal True Endpoint

Next, let us consider a role reversal, such that the true endpoint is repeatedly measured over time with K repeated measures, whilst having the surrogate endpoint in cross-sectional form. Model (8.1) then becomes:

$$\begin{aligned} T_{jk} &= \mu^*_T + \beta^* Z_j + f(t_{jk}) + \varepsilon_{Tjk}, \\ S_j &= \mu^*_S + \alpha^* Z_j + \varepsilon_{Sj}. \end{aligned} \tag{8.12}$$

The error terms $(\varepsilon_{T_{jk}}, \varepsilon_{S_j})$ are zero-mean normally distributed with variance-covariance matrix:

$$\Sigma = \begin{pmatrix} \Sigma_{TT} & \Sigma_{TS} \\ \Sigma_{ST} & \sigma_{SS} \end{pmatrix}. \tag{8.13}$$

Now, the VRF_{indiv} for this case is

$$\begin{aligned} VRF_{ST}^{CL} &= \frac{\text{tr}(\Sigma_{TT}) - \text{tr}(\Sigma_{T|S})}{\text{tr}(\Sigma_{TT})} \\ &= \frac{\text{tr}(\Sigma_{TT}) - \text{tr}(\Sigma_{TT} - \Sigma_{TS}\sigma_{SS}^{-1}\Sigma_{ST})}{\text{tr}(\Sigma_{TT})} \\ &= \frac{\text{tr}(\Sigma_{TT}) - \text{tr}(\Sigma_{TT}) + \text{tr}(\Sigma_{TS}\sigma_{SS}^{-1}\Sigma_{ST})}{\text{tr}(\Sigma_{TT})} \\ &= \frac{\text{tr}(\Sigma_{TS}\Sigma_{ST})}{\sigma_{SS}.\text{tr}(\Sigma_{TT})}. \end{aligned} \tag{8.14}$$

From (8.9) and (8.14), it is clear that there is asymmetry in the VRF calculations. Results differ depending on which of the two endpoints is the cross-sectional one. This is in line with our expectations. In the case of a longitudinal true endpoint, the VRF measures the ability of the cross-sectional endpoint to predict the longitudinal outcome at each time point, whereas when the longitudinal sequence is treated as surrogate endpoint, the VRF measures the adequacy of the longitudinal sequence to predict the cross-sectional outcome. It is therefore imperative to determine in advance which of the two outcomes is treated as true when applying this procedure to quantify association. Either way, a VRF value close to one indicates that the surrogate is a 'good' predictor of the true endpoint at the individual level, while values close to zero indicate 'poor' prediction. In any case however, the values of the VRF have to be complemented with expert opinion before passing judgment on the adequacy of the surrogate to predict the true endpoint. In the same manner lets us consider model

(8.12). The R^2_Λ for a longitudinal true and a cross-sectional surrogate endpoint is

$$\begin{aligned} R^{2,CL}_{\Lambda,ST} &= 1 - \frac{|\Sigma|}{|\Sigma_{TT}| \cdot |\sigma_{SS}|} \\ &= 1 - \frac{|\Sigma_{TT}| \cdot |\sigma_{SS} - \Sigma_{ST}\Sigma^{-1}_{TT}\Sigma_{TS}|}{|\Sigma_{TT}| \cdot |\sigma_{SS}|} \\ &= 1 - \frac{|\sigma_{SS} - \Sigma_{ST}\Sigma^{-1}_{TT}\Sigma_{TS}|}{|\sigma_{SS}|} \\ &= 1 - \frac{\sigma_{SS} - \Sigma_{ST}\Sigma^{-1}_{TT}\Sigma_{TS}}{\sigma_{SS}} \\ &= \frac{\Sigma_{ST}\Sigma^{-1}_{TT}\Sigma_{TS}}{\sigma_{SS}}. \end{aligned} \tag{8.15}$$

Comparing (8.11) with (8.15) establishes that $R^{2,LC}_{\Lambda,ST} = R^{2,CL}_{\Lambda,ST}$. In the first case, we used σ_{TT} and Σ_{SS} as component variances, of scalar and matrix type, respectively. These roles are reversed in the current, second case. Nevertheless, we obtain the same final expression for R^2_Λ as is, of course, entirely in line with the original, symmetric definition (8.4) of the quantity. Furthermore, note that R^2_Λ and VRF are equal when the surrogate is longitudinal and the true endpoint cross-sectional. This implies that, only the VRF with the surrogate cross-sectional and the true endpoint longitudinal will be different from all of the others, that than coincide. This again highlights the feature that, for a longitudinal true endpoint, the VRF studies prediction of the entire sequence, while the R^2_Λ assesses how well an optimal linear combination of the true endpoint profile can be predicted. Both may be useful, but definitely are different. Moreover, one would expect the VRF to be well below the R^2_Λ in many applications, since prediction of an entire longitudinal sequence from a cross-sectional quantity is a tall order, whereas it might well be feasible to predict a particular linear combination. The choice between the two measures lies in the objective to be attained. If the objective is to measure the strength of the surrogate to predict the

entire sequence of the true endpoint, then VRF will be an ideal choice. However, when this seems an attainable goal or when we are rather interested in predicting some linear combination of the true endpoint, then we can resort to R^2_Λ. Note that standard error of the estimates can be calculated using either a delta method or bootstrap (Efron, Bradley and Tibshirani, 1993).

8.4 Flexible Linear Mixed Modeling

The setting considered here include both cross-sectional and longitudinal outcomes and hence appropriate modeling of the longitudinal outcome is called for. Let us first give a brief introduction to the analysis of longitudinal data.

8.4.1 Longitudinal Data Analysis

Since we are in the framework of continuous longitudinal data, modeling can be done by way of a linear mixed model. The general linear mixed-effects model can be represented as (Verbeke and Molenberghs, 2000):

$$\begin{cases} \boldsymbol{Y}_j = X_j\boldsymbol{\beta}_j + Z_j\boldsymbol{b}_j + \boldsymbol{\varepsilon}_j \\ \boldsymbol{b}_j \sim N(\boldsymbol{0}, \boldsymbol{G}), \quad \boldsymbol{\varepsilon}_j \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_j), \quad \boldsymbol{b}_1, \ldots, \boldsymbol{b}_N, \quad \boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_N \text{ are independent,} \end{cases} \tag{8.16}$$

where $\boldsymbol{Y}_j$ $(j = 1, \ldots, n)$ is the m_i-dimensional response vector of measurements for subject j, X_j and Z_j are $m_j \times p$- and $m_j \times q$-dimensional matrices of known covariates (e.g., time), respectively, $\boldsymbol{\beta}_j$ is a p-dimensional vector of fixed effects, $\boldsymbol{b}_j$ is q-dimensional subject-specific vector of random effects and $\boldsymbol{\varepsilon}_j$ is an m_j-dimensional vector of residuals. The matrix $\boldsymbol{G}$ is a general $q \times q$ covariance matrix and $\boldsymbol{\Sigma}_j$ is an $m_j \times m_j$ covariance matrix. Often, $\boldsymbol{\Sigma}_j$ is assumed to equal $\sigma^2_\varepsilon \boldsymbol{I}_{m_j}$, resulting in the so-called conditional independence model. Note that when the response is cross-sectional, the general model reverts to the usual regression model wherein subject-specific effects are

dropped. The evolution over time can be captured by specifying parametric functions, such as, for example, linear, quadratic or even higher-order polynomials in the vector X_j. These effects may well be included in the random-effects vector Z_j as well. However, it is not difficult to imagine cases where obtaining a suitable parametric form adequately describing the mean is a challenge. Although our primary goal is to quantify the association between various outcomes via surrogate marker validation methods, proper modeling of the mean evolution in time is necessary. One can get rid of the need to specify a parametric model through use of flexible modeling techniques, an issue taken up further in the following section.

8.4.2 Flexible Modeling Techniques

Postulating a parametric function to model the mean evolution may be difficult and/or restrictive. An appealing alternative is to model the time evolution using some flexible smooth function. In this section, we briefly discuss linear mixed models to model longitudinal data (Verbeke and Molenberghs, 2000) with the time trend determined by some flexible smooth function in the form of either penalized smoothing splines (Eilers and Marx, 1996; Verbyla *et al.*, 1999; Ruppert, Wand, and Carroll, 2003) or fractional polynomials (Royston and Altman, 1994).

Penalized Smoothing Splines

Use of penalized splines results in a semi-parametric smooth function, the term 'semi-parametric' here referring to the feature that the model combines parametric and non-parametric aspects. We provide a brief description of the model as is usually encountered with longitudinal data.

Let Y_{jk} denote the response taken from subject j at time t_{jk} $(k = 1, \ldots, K)$. The model of interest can be expressed as: $Y_{jk} = f(t_{jk}) + \varepsilon_{jk}$, for a smooth function

$f(\cdot)$. Restricting focus to the truncated lines basis, which is simple in formulation and performs adequately in many circumstances (Ngo and Wand, 2004), the penalized-spline representation can be written as:

$$Y_{jk} = \beta_0 + \beta_1 t_{jk} + \sum_{q=1}^{Q} b_q (t_{jk} - \kappa_q)_+ + \varepsilon_{jk}, \tag{8.17}$$

where $\kappa_1, \ldots, \kappa_Q$ are a set of distinct knots in the range of t_{jk}, $t_+ = \max(0, t)$, and $b_q \sim N(0, \sigma_b^2)$. The knot points are selected as equally spaced quantiles of time (Ruppert *et al.*, 2003). For ease of development, we adopt the following matrix notation. Let

$$\boldsymbol{Y}_j = \left[\; y_{jk} \; \right]_{1 \le j \le n, 1 \le k \le K}, \quad \boldsymbol{X}_j = \left[\; 1 \quad t_{jk} \; \right]_{1 \le j \le n, 1 \le k \le K}, \qquad \boldsymbol{\beta} = \left[\; \beta_0 \quad \beta_1 \; \right]'.$$

Further, define:

$$\boldsymbol{Z}_j = \left[\; (t_{jk} - \kappa_k)_+ \; \right]_{1 \le j \le n, \; 1 \le k \le K, \; 1 \le \kappa \le Q}, \boldsymbol{b} = \left[\; b_1, \ldots, b_Q \; \right]', \boldsymbol{\varepsilon}_j = \left[\; \varepsilon_{11}, \ldots, \varepsilon_{nK} \; \right]'$$

using this notation, a stacked version of (8.17) becomes $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}$. The correspondence between the penalized spline smoother and the optimal predictor in a mixed model framework is a key feature in fitting the models. This connection offers the opportunity of using ordinary software packages for mixed models, such as, for example, SPlus, SAS, or R. Fitting penalized splines by the linear mixed model approach has some appealing advantages, such as automatic determination of the smoothing parameter, a unified framework for inference, and the flexibility with which the models can be extended (Faes *et al.*, 2006).

Fractional Polynomials

As an alternative to capturing the time trend as mentioned in Section 8.4.2, the so-called fractional polynomial approach may be used. Fractional polynomials provide an extension to classical polynomials allowing for non-integer powers to the time

covariate, thereby adding greater flexibility in capturing rather complex non-linear relationships. A brief description of fractional polynomials is given below. Let $\boldsymbol{t} = (t_{j1}, \ldots, t_{jK})$ denote the set of time points pertaining to subject j. Royston and Altman (1994) define a fractional polynomial of degree m by

$$\phi_m(\boldsymbol{t}; \boldsymbol{\beta}, \boldsymbol{p}) = \sum_{r=0}^{m} \beta_r H_r(\boldsymbol{t}), \tag{8.18}$$

where m is a positive integer and $\boldsymbol{p} = (p_1, \ldots, p_m)$ is a real-valued set of powers such that $p_1 \leq \cdots \leq p_m$ and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_m)$ are real-valued coefficients. For $r = 0$, $H_0(\boldsymbol{t}) = 1, p_0 = 0$, and for $r = 1, \ldots, m$:

$$H_r(\boldsymbol{t}) = \begin{cases} \boldsymbol{t}^{p_r} & \text{if } p_r \neq p_{r-1}, \\ H_{r-1}(\boldsymbol{t}) \ln(\boldsymbol{t}) & \text{if } p_r = p_{r-1}. \end{cases}$$

As mentioned in Royston and Altman (1994), polynomials of a degree higher than 2 or 3 are rarely encountered in practice. The best power transformation is frequently found among the members of the list $\{-2, -1, -0.5, 0, 0.5, 1, \ldots, \max(3, m)\}$. While it is possible to incorporate other powers, there is a danger coming with including large negative powers, in the sense that individual extreme observations will influence the fit too much (Royston, Parmar, and Qian 2003). Note that the fractional polynomial model has been defined in its generic form and in analogy with penalized-splines models; extension to include covariates other than time is possible.

8.5 Application to the Case Study

The case study on stress related disorders introduced in Chapter 2 section 2.1.5 was analyzed here. The objective was to assess the association between the different responses before and after stress was induced. Thus, the results for pre-stress and post-stress correspond to the associations measured between the different responses

before and after the stress with the treatment variable (Z), having two possible values (1: active treatment, 0:vehicle) for pre-stress and having four different possible values after stress. Figure 8.1 shows the group-specific mean profiles of CORT measurements, averaged over the four treatment periods. The plot depicts the average CORT values per treatment group at each time point, essentially showing how, on average, CORT values evolve over time in each treatment group. The need for flexible modeling tools is apparent, since finding a suitable or rather an acceptable classical parametric model is not an easy task. Hence, as mentioned before, we discuss results emanating from an application of surrogate marker validation methodology in conjunction with flexible modeling techniques (spline and fractional-polynomial based), meant to appropriately capture trends over time. For purposes of comparison, an unstructured mean model or a full factorial structure for time is also considered. However, this approach often yields excessively large numbers of parameters, thereby rendering it less desirable. The variance-covariance matrices, based upon which the VRF and R^2_Λ are computed, are estimated using maximum likelihood. The variance-covariance matrices can assume general structures unless the data suggests otherwise. In such cases, simple covariance structures, such as auto-regressive or compound symmetry, might be considered. For the purpose of our application, a number of models with different variance-covariance structures has been fitted. The best model, here being an unstructured variance-covariance structure, can be chosen using a conventional likelihood ratio test and/or Akaike's Information Criterion. The results of the analysis for the association of telemetry and behavior as well as that of CORT and behavior are summarized in Table 8.1 with bootstrap standard errors and in Table 8.2 with asymptotic standard errors, respectively. We should like to point out that it is not a trivial task to derive a closed-form expression for the standard errors of VRF and

R^2_Λ for the particular case we have considered. However, fortunately, Alonso *et al* (2006) have shown that the VRF and R^2_Λ are special cases of the so-called *Likelihood Reduction Factor*, which is based on the information-theory approach. These authors have derived an asymptotic solutions for LRF. Hence, by virtue of the relationship of these measures with the LRF, we have been able to provide asymptotic standard errors based on the information-theory approach. There are no general guidelines as to how large a VRF and R^2_Λ should be in order to be considered sufficiently large. However, since the VRF and R^2_Λ are R-square type measures, it might be possible to make some general remarks concerning the degree of association based on their magnitude. Since such a degree of association arguably would vary from application to application, the final decision has to be made in consultation with the experts, regardless their value. Having this in mind, from the results for the pre- and post-stress, we might infer that there is a rather weak relationship between behavior and CORT. However, strong and moderate relationships were observed between heart rate and behavior, and between blood pressure and behavior, respectively. Recall that behavior is measured cross-sectionally while CORT, heart rate, and blood pressure are longitudinal outcomes. In this regard, when the cross-sectional outcome was used as a possible surrogate for the longitudinal outcomes, the VRF produced very low values, as anticipated in the previous section. Indeed, it is very difficult to predict the subtleties and richness of a longitudinal sequence from a single, cross-sectional measure. We consider this a desirable feature of the VRF. The R^2_Λ on the other hand, states that, although still small for some of the endpoints, there is better hope to predict a particular linear combination of the longitudinal outcomes from the cross-sectional outcome. As such, VRF and R^2_Λ both provide useful but totally *different* pieces of information. When there is role reversal, that is, when the longitudinal outcomes

were treated as a possible surrogates for the cross-sectional outcome, the VRF values coincided with the R^2_Λ. This underscores that the VRF does not treat both endpoints symmetrically. The R^2_Λ, however, stayed the same even when there was role reversal, as expected from its construction. The higher VRF and R^2_Λ values obtained when the longitudinally measured heart rate and blood pressure were used as surrogate endpoints for the cross-sectionally measured behavior, establish the possibility of predicting behavior using some linear combination of the longitudinal sequence. Zooming in on the association between telemetry and CORT, both longitudinal in nature, we learn that there is a very weak association, with a maximum $\widehat{R}^2_\Lambda = 0.2314$ and maximum $\widehat{VRF} = 0.0513$, between the three modeling approaches. This is an indication that there is a very limited overlap in information between both outcomes, inhibiting comfortable prediction of one from the other. In conclusion, the analysis has revealed that the longitudinally measured CORT level offers limited opportunity for prediction of activity, which is measured by the degree of alertness expressed in terms of the percentage of minutes the rats have been awake. We learn that heart rate and blood pressure are weakly related to CORT but have a strong predictive ability for behavior. The results advice against the use of activity to predict the longitudinal CORT level, heart rate, and blood pressure at each time point.

8.6 Discussion

In this chapter, we have adapted surrogate marker evaluation methods, originally designed to handle two repeated measures sequences, to the case of one cross-sectional and one longitudinal outcome, where either of these can be used as the surrogate. The methods have been applied to quantifying association between longitudinally measured CORT level, heart rate, and blood pressure, with cross-sectional behavior

measured by the level of activity, expressed as the percentage of time experimental rats have been active after exposure to treatment followed by stress. The methods appear to work adequately for this particular mix of longitudinal and cross-sectional endpoints. The various theoretical properties of the methods have manifested themselves in the results of the data analysis. In particular, it has been nicely confirmed that the VRF focuses on the prediction of a longitudinal sequence *as a whole* by a cross-sectional outcome, while R^2_Λ is concerned with the prediction of an *optimal linear combination* of the longitudinal outcome. In the case of two longitudinal outcomes, the optimal linear combinations from the two outcomes are the first canonical variates. In the context of a longitudinal true and cross-sectional surrogate endpoint, the optimal linear combination could be the first principal component or any other summary measure of the longitudinal measurements, thereby maximally retaining information. Thus, optimality in this context refers to finding a linear combination that best summarizes the repeated measures. The longitudinal outcomes were modeled using flexible modeling tools such as fractional polynomials, penalized splines, and a general unstructured mean where the time trend is not modeled but rather an analysis-of-variance type approach is followed. This offers the possibility of fitting different models and then selecting the best one according to some model selection tool such as, for example, Akaike's Information Criterion. It is, indeed, important to conduct proper modeling before moving into quantifying surrogacy, because the results may critically depend on the model's goodness-of-fit. In all cases, VRF or R^2_Λ estimates close to one are indicative of 'good' surrogacy, with the reverse holding for values close to zero. Evidently, it is difficult to provide general advice as to how large is large enough. Arguably, the statistical evaluation of a surrogate can be an important component in the decision making process, but at least equally important

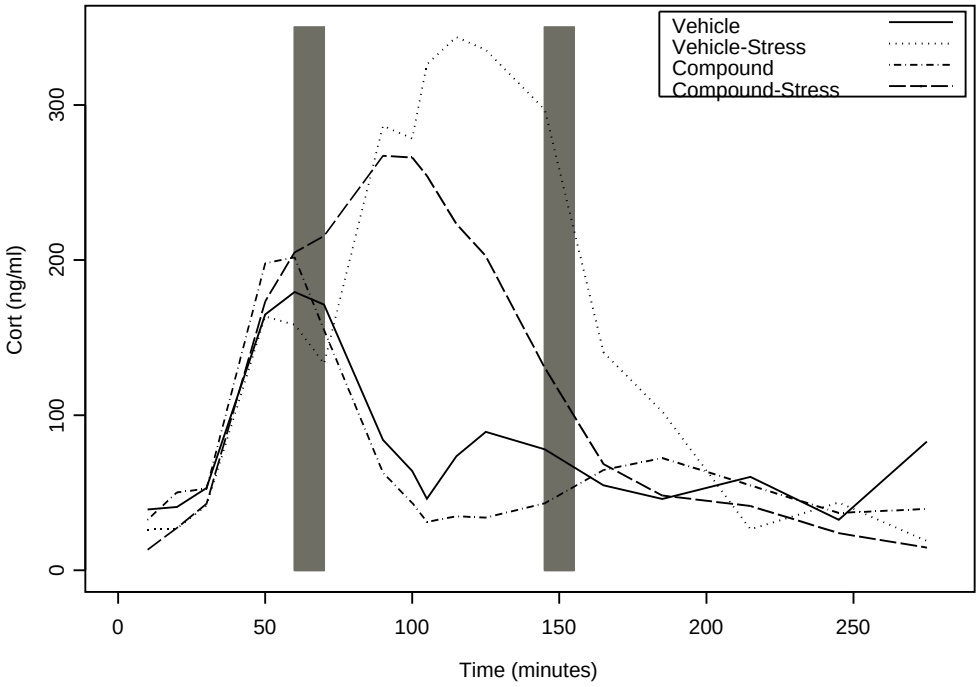

Figure 8.1: *Group-specific mean profiles of CORT values, averaged over different treatment periods. The shaded regions indicate the time windows in which activity was measured before and after the stress induction.*

is expert opinion coming in from pharmacological, biological, clinical, ethical, and health economy considerations.

Table 8.1: R^2_{indiv} *values(bootstrap standard errors) under pre-stress and post-stress conditions, for a variety of true and surrogate endpoints, using unstructured, fractional polynomial, and penalized splines models, and based on both* VRF *and* R^2_Λ.

endpoint		unstructured		fract. pol.		pen. splines	
true	surrogate	VRF	R^2_Λ	VRF	R^2_Λ	VRF	R^2_Λ
				Pre-stress			
behavior	CORT	0.433(0.1803)	0.433(0.1803)	0.372(0.1547)	0.372(0.1547)	0.402(0.1818)	0.402(0.1818)
CORT	behavior	0.060(0.0314)	0.433(0.1813)	0.039(0.020)	0.372(0.1547)	0.026(0.290)	0.402(0.1818)
behavior	heart rate	0.807(0.0928)	0.807(0.0928)	0.816(0.1116)	0.816(0.1116)	0.798(0.1793)	0.798(0.1793)
heart rate	behavior	0.119(0.0568)	0.807(0.0928)	0.069(0.0624)	0.816(0.1116)	0.071(0.0689)	0.798(0.1793)
behavior	blood pressure	0.571(0.1916)	0.571(0.1916)	0.586(0.1781)	0.586(0.1781)	0.408(0.2146)	0.408(0.2146)
blood pressure	behavior	0.081(0.0246)	0.571(0.1916)	0.073(0.0468)	0.586(0.1781)	0.011(0.0369)	0.408(0.2146)
				Post-stress			
behavior	CORT	0.386(0.1889)	0.386(0.1889)	0.499(0.2095)	0.499(0.2095)	0.359(0.1190)	0.359(0.1190)
CORT	behavior	0.038(0.0248)	0.386(0.1889)	0.045(0.0984)	0.499(0.2095)	0.032(0.0273)	0.359(0.1190)
behavior	heart rate	0.913(0.0498)	0.913(0.0498)	0.984(0.0263)	0.984(0.0263)	—	—
heart rate	behavior	0.227(0.0868)	0.913(0.0498)	0.126(0.0755)	0.984(0.0263)	—	—
behavior	blood pressure	0.343(0.1041)	0.343(0.1041)	0.513(0.2050)	0.513(0.2050)	—	—
blood pressure	behavior	0.079(0.055)	0.343(0.1041)	0.160(0.1288)	0.513(0.2050)	—	—

Table 8.2: R^2_{indiv} *values (asymptotic standard errors) under pre-stress and post-stress conditions, for a variety of true and surrogate endpoints, using unstructured, fractional polynomial, and penalized-splines models, and based on both* VRF *and* R^2_Λ.

endpoint		unstructured		fract. pol.		pen. splines	
true	surrogate	VRF	R^2_Λ	VRF	R^2_Λ	VRF	R^2_Λ
				Pre-stress			
behavior	CORT	0.433(0.1178)	0.433(0.1178)	0.372(0.1174)	0.372(0.1174)	0.402(0.1179)	0.402(0.1179)
CORT	behavior	0.060(0.0632)	0.433(0.1178)	0.039(0.0533)	0.372(0.1174)	0.026(0.0463)	0.402(0.1179)
behavior	heart rate	0.807(0.0724)	0.807(0.0724)	0.816(0.0702)	0.816(0.0702)	0.798(0.0745)	0.798(0.0745)
heart rate	behavior	0.119(0.0850)	0.807(0.0724)	0.069(0.0669)	0.816(0.0702)	0.071(0.0677)	0.798(0.0745)
behavior	blood pressure	0.571(0.1105)	0.571(0.1105)	0.586(0.1091)	0.586(0.1091)	0.408(0.1179)	0.408(0.1179)
blood pressure	behavior	0.081(0.0717)	0.571(0.1105)	0.073(0.0685)	0.586(0.1091)	0.011(0.0823)	0.408(0.1179)
				Post-stress			
behavior	CORT	0.386(0.1177)	0.386(0.1177)	0.499(0.1156)	0.499(0.1156)	0.359(0.1171)	0.359(0.1171)
CORT	behavior	0.038(0.0528)	0.386(0.1177)	0.045(0.0563)	0.499(0.1156)	0.032(0.00497)	0.359(0.1171)
behavior	heart rate	0.913(0.0415)	0.913(0.0415)	0.984(0.0108)	0.984(0.0108)	—	—
heart rate	behavior	0.227(0.1063)	0.913(0.0415)	0.126(0.0868)	0.984(0.0108)	—	—
behavior	blood pressure	0.343(0.1164)	0.343(0.1164)	0.513(0.1149)	0.513(0.1149)	—	—
blood pressure	behavior	0.079(0.0709)	0.343(0.1164)	0.160(0.0947)	0.513(0.1149)	—	—

9

Optimal Number of Repeated Measurements

Surrogate-marker validation exercises that have been considered thus far involved two different endpoints, where one endpoint is a candidate surrogate and the other is a true endpoint. Such endpoints may be of the same nature (e.g., both continuous, binary, or longitudinal) or of a mixed nature (e.g., a binary surrogate, for a continuous true endpoint, a continuous surrogate for time-to-event true endpoint). In contrast, the scenario we deal with here has only one endpoint, measured repeatedly over time. We are then interested in the predictive potential of the earlier clinical measurements for the later ones, and in particular for the last one. This can be placed within the surrogate-marker evaluation context, by considering the accumulated first few repeated measurements as potential surrogates and the outcome, for example at the final measurement occasion, as the true endpoint. Thus, for each subject, the surrogate is a vector of repeated measurements and the true endpoint is a scalar. The

situation where the surrogate is a single early measurement is, of course, merely a special case. The challenge is to determine the number of repeated measures that are required to adequately predict the true endpoint. It is evident that collecting more repeated measurements enhances prediction. However, more repeated measurements imply longer study periods and increase cost. Thus, there must be a balance between cost and precision. The objective we want to address in this chapter is threefold. First, existing surrogate-marker evaluation procedures will be tuned to accommodate the present scenario. Second, selection of an optimal number of repeated measurements will be effectuated using an objective function, designed as a weighted function of financial cost and predictive precision. The objective function allows tuning to the specific needs of a particular case study. Third, a simulation study is conducted to investigate the performance of the proposed procedure under different covariance structures for the repeated measures.

9.1 Measure of Surrogacy

In this chapter we are dealing with a special case that deviates from the main stream surrogate marker validation where two separate outcomes are entertained. Note however that, a closer look into the problem reveals that the situation is similar to the case of a longitudinal surrogate for a cross-sectional outcome discussed in Chapter 8. The only difference is that, here the problem is formulated based on a single repeatedly measured endpoint which will be subdivided into a surrogate and true endpoint. As a consequence, we have a repeatedly measured (longitudinal) surrogate and a singly measured (cross sectional) true endpoint. Thus the measures derived in the previous chapter for the case of a longitudinal surrogate endpoint for a cross-sectional true endpoint can directly be applied to this situation.

9.1.1 Optimal Number of Repeated Measurements

In this scenario, we are interested in predicting the outcome of a subject at a specified point in time given an accumulated number of repeated measurements of the outcome at an earlier point in time. Along this idea, let us denote by Y_{ijk} the k^{th} measurement, $k = 1 \ldots K$, of subject j, $j = 1 \ldots n_i$, in trial i, $i = 1 \ldots N_t$. We shall further assume that the following model holds

$$Y_{ijk} = (\beta_0 + b_{1i}) + (\beta_1 + b_{2i})Z_{ij} + \beta_2 t_{ik} + \beta_3 Z_{ij} t_{ik} + \varepsilon_{ijk}, \tag{9.1}$$

where Z_{ij} and t_{ik} are binary treatment indicator and the time at which measurements are taken, (b_{1i}, b_{2i}) are trial specific effects assumed to follow a normal distribution with mean zero and variance covariance matrix D_L, and the error vector ε_{ijk} is assumed to follow a normal distribution with mean zero and variance covariance matrix Σ_L. Our model assumes a linear treatment effect over time, which is equal for all trials but can be extended to more complex model, if need be, as proposed by Alonso *et al.* (2004d). Let us formally define our surrogate and true endpoints, based on (9.1). Suppose we intend to investigate if the first cumulated m measurements, where $1 \leq m \leq K - 1$, are a good predictor for the outcome measured at time K. Our surrogate endpoint is then the m dimensional vector of measurements $\widetilde{S}_{ij}^T = (Y_{ij1}, \ldots, Y_{ijm})$, and our true endpoint is the measurement Y_{ijK}, i.e., $S_{ijk} = Y_{ijk}$ $(k = 1, \ldots, m-1)$ and $T_{ij} = Y_{ijK}$, where the indices i, j, and k are defined as in (9.1). This leads to model (8.5) and its variance covariance matrix from which we can compute the measure of surrogacy of the initial m measures for the final outcome using (8.9).

9.1.2 Cost Function and Optimal Number of Measurements

To determine the optimal number of measurements (m_o), we will consider the following cost function, introduced by Winkens *et al.* (2005):

$$FC = NC_1 + NKC_2. \tag{9.2}$$

Here, FC represents the fixed total financial cost, N is the total number of subjects in the study, K is the number of planned repeated measurements per subjects, C_1 is the cost of recruiting a subject to the study, and C_2 is the cost per measurement and per subject. Let $R = C_1/C_2$ be the ratio of the two costs; usually the cost of recruiting a subject to the study is higher than the cost per measurement, i.e., $R > 1$. We can then re-write (9.2) as $FC = NC_2(R + K)$. Suppose now that, instead of taking K measurements, we take m, $1 \leq m \leq K-1$, measurements and use this information to predict the outcome at the K^{th} time point, the financial cost for the m measurements is then given by $FC(m) = NC_1 + NmC_2$. Thus, the proportion of the total financial cost required to take m measurement is $PFC(m) = (R + m)/(R + K)$. It is easy to show that the variance of the prediction, based on m observations, of the outcome at the last time point takes the form $[1 - VRF_{\text{ind}}(m)]\sigma_{TT}$. Note further that σ_{TT} is constant, irrespective of the number of repeated measurements used as a surrogate; thus a standardized version of the prediction variance, $1 - VRF_{\text{ind}}(m)$, will be used. Finally, a weighted linear combination of the prediction variance and the financial cost can be used to define an objective function as shown in (9.3), with weights w_1 and $(1 - w_1)$, respectively. An advantage of standardizing the prediction variance and financial cost for a given number of repeated measurements m is the relative ease of specifying w_1, compared to using the non-standardized versions:

$$CPR_0(m) = w_1 \cdot [1 - VRF_{\text{ind}}(m)] + (1 - w_1) \cdot \frac{R + m}{R + K}. \tag{9.3}$$

The number m_o is determined as that minimizing $CPR(m)$. Let us consider some extensions. The objective function assumes that the cost of each measurement is the same, which may be unrealistic for some situations; for example, when patients have to stay in a hospital or health institute, where the waiting time may incur additional costs, a feature not accommodated by (9.3). One can therefore elect to introduce a third term accounting for time lag:

$$CPR_I(m) = w_1 \cdot [1 - VRF_{\text{ind}}(m)] + w_2 \cdot \frac{R+m}{R+K} + w_3 \cdot \frac{t_m - t_0}{t_k - t_0}. \tag{9.4}$$

If the repeated measures are equidistant with time lag $\triangle$, then $t_m = t_0 + \triangle M$ and $t_k = t_0 + \triangle K$. Hence, (9.4) takes the form

$$CPR_I(m) = w_1 \cdot [1 - VRF_{\text{ind}}(m)] + w_2 \cdot \frac{R+m}{R+K} + w_3 \cdot \frac{M}{K}. \tag{9.5}$$

If in addition we assume that the waiting cost for the first measurement is zero, then:

$$CPR_{II}(m) = w_1 \cdot [1 - VRF_{\text{ind}}(m)] + w_2 \cdot \frac{R+m}{R+K} + w_3 \cdot \frac{M-1}{K}. \tag{9.6}$$

These objective functions assume that the cost is constant across treatment arms, whether of a placebo, standard-therapy, or experimental nature. When deemed unrealistic, appropriate modifications can be implemented. Arguably, the choice of a cost function will have to balance simplicity with it being a realistic representation of reality. In what follows, objective function (9.3) will be employed, unless otherwise stated.

9.2 Some Important Special Cases

In this section, we aim to aid understanding of the nature of the cost functions through theoretical considerations for two special, important cases. The detailed derivations of the expressions for the associations measures are given in the appendix A.

9.2.1 Compound Symmetry Structure

Assume that the covariance structure of (9.1) is compound symmetry (CS), i.e., $\Sigma_L = \sigma(1-\rho)I_K + \sigma\rho J_K$, where σ denotes the variance of the response at each time point, ρ is the correlation between two observations, I_K is a K-dimensional identity matrix and J_K is a K-dimensional square matrix of ones. It is easy to show that, in this setting,

$$VRF_{\text{ind}}(m) = \frac{m\rho^2}{1+(m-1)\rho}.$$

Let us study the predictive characteristics of this case. It follows that $VRF_{\text{ind}}(m)$ is an increasing function of m as far as $\rho \neq 0, 1$ and, therefore, the more observations we include in $\widetilde{S}_{ij}$, the more precise our prediction of T_{ij} will be. Turning to ρ, the question is how the correlation influences the amount of information that $\widetilde{S}_{ij}$ brings about T_{ij}. To usefully study this, let us calculate the additional information that one extra observation will bring, quantified using the ratio:

$$g(\rho) = \frac{VRF_{\text{ind}}(m+1)}{VRF_{\text{ind}}(m)} = \left(\frac{m+1}{m}\right)\left(\frac{1+(m-1)\rho}{1+m\rho}\right).$$

Some elementary calculations show that $g(\rho)$ is a decreasing function of ρ and therefore, the higher the correlation the less we gain by taking additional observations, rather an intuitive result. Indeed, if the correlation is very high, then all the measurements are nearly deterministically related, and having observed one or a few of them will allow us to predict with high precision all the others. For instance, in the extreme case when $\rho = 1$ the $VRF_{\text{ind}}(m+1) = VRF_{\text{ind}}(m)$ for all m and the first observation will be sufficient to predict the true endpoint without error. Coherent with the nature of compound symmetry, the position in the sequence of the m observations that constitute the surrogate is totally irrelevant. It is easy to show that in

this setting the CPR function takes the form

$$CPR(m) = w_1 \cdot \frac{(1-\rho)(1+m\rho)}{1+(m-1)\rho} + (1-w_1) \cdot \frac{R+m}{R+K}, \tag{9.7}$$

of which the extremes are easy to determine: (9.7) reaches its minimum at m_+ and m_- when $\rho > 0$ and $\rho < 0$ respectively, where

$$m_\pm = -\left(\frac{1-\rho}{\rho}\right) \pm \sqrt{\frac{w_1(R+K)(1-\rho)}{1-w_1}}. \tag{9.8}$$

Obviously, in many practical situations, $m_\pm$ will not be integers, in which case they will have to be rounded. There is also a possibility for $m_\pm$ to assume a negative value for some combinations of K, ρ, R, and w_1. When this happens, $m_\pm$ should be set to one. Zooming in on m_+ reveals that, when less weight is assigned to the precision part of the cost function, an increase in R has little influence on m_+ but its influence increases as more weight is assigned to precision. This is to be expected because when the cost of recruiting subjects is much higher than taking more measurements on subjects, the obvious way to increase precision is through taking more measurement per subject. An increase in the correlation ρ between measurement leads to a decrease in m_+ when the weight assigned to precision is small to moderate. When the weight increases, the value of m_+ increase for ρ in $[0; 0.5]$ and decreases in $[0.5; 1]$. Also, a increase in K generally leads to a slight increase in m_+.

9.2.2 First-order Auto-regressive Process

Another association structure frequently encountered in longitudinal data is the first-order auto-regressive one, with ρ^t the correlation between two measurements, t time units apart. In this case, Σ_{SS} is also an $(m \times m)$ AR(1) matrix, $\Sigma_{ST} = \Sigma_{TS}^T = \rho^{K-m}\delta_1^T$ with $\delta_1^T = (\rho^{m-1}, \ldots, 1)$ and $\sigma_{TT} = \sigma$. It then follows that $VRF_{\text{ind}}(m) = \rho^{2(K-m)}\sigma\delta_1^T\Sigma_{SS}^{-1}\delta_1$. Further, using the expression for the inverse of an AR(1) matrix (Graybill 1983), one can prove that $\sigma\delta_1^T\Sigma_{SS}^{-1}\delta_1 = 1$ and therefore $VRF_{\text{ind}}(m) =$

$\rho^{2(K-m)}$. Like in the compound-symmetry case, here the $VRF_{\text{ind}}(m)$ is an increasing function of m. However, unlike before, it is also an increasing function of ρ, implying that the higher ρ, the more advantageous it is to include more observations into the surrogate. This is again a very intuitive result. This is intuitively plausible because, under AR(1), the correlation decreases rapidly with time lag; hence it is recommendable to consider surrogate outcomes that are collected sufficiently closely to the true endpoint. More generally, the position of the surrogate measures within the sequence of repeated measures is now relevant. For instance, if we now consider as the surrogate marker a sub-sequence of m observations starting at time point $s+1$, then $VRF_{\text{ind(s+1)}}(m) = \rho^{2(K-s-m)}$. Obviously, $VRF_{\text{ind(s+1)}}(m) \geq VRF_{\text{ind}}(m)$, for $s \geq 1$, and therefore considering m observations closer to the true endpoint will result in a surrogate with more predictive power. In this scenario, the CPR function takes the form:

$$CPR(m) = w_1 \cdot \left(1 - \rho^{2(K-m)}\right) + (1 - w_1) \cdot \frac{R+m}{R+K}. \tag{9.9}$$

Interestingly, (9.9) does not reach its minimum value in the interval $(1, K-1)$ and therefore $CPR(m)$ will always lead to choosing the first observation only if the cost is the impelling criterion or choosing the entire $K-1$ sequence if prediction is the more important factor. This result also holds if the longitudinal surrogate sequence is started at a time point different from the first one. Thus, the $CPR(m)$ seems to indicate that in this scenario the surrogate should contain one observation only and therefore, the most rational choice would be to consider a value sufficiently close to the true endpoint so that a reasonable level of precision can be achieved in the prediction. Obviously, the closer this observation is to the true endpoint the better the prediction will be but the longer we will have to wait. A compromise between these two considerations should be found in this setting using external elements such

as, for example, expert opinion.

9.3 Simulation Study

Even though the previous results are enlightening, not all cases are analytically tractable. Moreover, even in those cases where analytic results are obtainable it is still of great interest to study the performance of the proposed method when parameters have to be estimated. To this end, a simulation study was performed to investigate further these issues, with focus on the two association structures namely compound symmetry and autoregressive of order one.

9.3.1 Design of Simulation Study

Equally spaced longitudinal data were generated based on (9.1) and using a two-stage approach. In the first stage, random trial-specific intercepts and treatment effects, b_{1i} and b_{2i} respectively, were generated from a zero-mean normal distribution with covariance matrix

$$D_L = \begin{pmatrix} 1.5 & 2.098 \\ 2.098 & 3.26 \end{pmatrix}.$$

Additionally, error terms ε_{ijk} were generated from a zero-mean normal distribution with covariance matrix Σ_L, either AR(1) or CS. The variance in Σ_L was assumed constant and the correlation between successive measurements was set to either 0.3, 0.6, or 0.9. The fixed-effects vector was set to $\boldsymbol{\beta}^T = (2.5, 4.3, 0.78, 3.5)$. Using these, the outcomes were obtained from (9.1). The data generation scheme assumes that the treatment-by-time interaction is constant across trials. To increase flexibility, a more general framework, where the treatment effect is allowed to randomly vary over time and across trials was adopted. The first stage now involved generation of random trial-specific time effects and random slopes, in addition to random trial-specific intercepts and treatment effects, b_{1i} and b_{2i}, from a zero-mean normal distribution

with covariance matrix

$$D_L = \begin{pmatrix} 1.0 & 0.8 & 0.00 & 0.00 \\ 0.8 & 1.0 & 0.00 & 0.00 \\ 0.0 & 0.0 & 1.00 & 0.95 \\ 0.0 & 0.0 & 0.95 & 1.00 \end{pmatrix}.$$

The error terms were, again, generated from a zero-mean normal distribution with AR(1) or CS covariance matrix Σ_L, The outcome vector Y_{ijk} then takes the form:

$$Y_{ijk} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})Z_{ij} + (\beta_2 + b_{2i})t_{ij} + (\beta_3 + b_{3i})Z_{ij}t_{ij} + \varepsilon_{ijk}.$$

The number of trials was set to either 10, 20, 30, or 40. Two sets of trial sizes were considered. The first set of smaller trial sizes consists of 20, 40, and 60 subjects per trial. The second set of larger trial sizes consists of 100, 200, and 300 subjects per trial. The simulation consists of a full combination of the specified correlation values, covariance matrix structures, number of trials, and trial sizes. For each combination, 100 datasets (samples) where generated, analyzed and the optimal number of measurements determined.

9.3.2 Simulation Study Results

The results of the simulation for the case of $R = 4$ and $K = 10$ are summarized in Tables 9.1–9.3. In the tables, $VRF_{\text{ind}}(m_o)$ is the usual individual-level surrogacy for the optimal number of measurements, while $VRF_{\text{ind}}(K-1)$ corresponds to the entire $K-1$ sequence being used as a surrogate. Furthermore, f represents the percentage of datasets that resulted in a given m_o as the optimal number of measurements. The weight, w_1, was set to either 0.3, 0.5, or 0.7. Let us focus on the first data-generation scheme, where the treatment-by-time interaction is assumed constant across trials. We learn that the $VRF_{\text{ind}}(m)$ increases with increasing number of repeated measurements. When the data are generated under AR(1) but analyzed using an unstructured

covariance matrix, the optimal number of time points was chosen to be either 1 or 9, depending on the weights assigned. When the correlation was set to 0.9, assigning more weight to precision or equal weights to both precision and financial cost requires all 9 repeated measurements to minimize the objective function. For the other possible values of the correlation, i.e., 0.30, 0.60, or 0.71, if more weight is assigned to financial cost or equal weights are assigned to financial cost and precision then the optimum simply is the first measurement only. However, the entire sequence is needed when progressively more weight is assigned to the precision. This result is in agreement with Section 9.2, where we have shown that, under AR(1), $CPR(m)$ does not reach its minimum value in the interval $(1, K-1)$ and therefore it will always lead to taking either only one observation or the entire $K-1$ subsequence. Hence, this result carries over to the simulation setting, in spite of the added variability coming from parameter estimation. When the data are generated using CS and analyzed with either unstructured or CS, 1, 2, 3, or 4 repeated measurements may be required to predict the outcome at the last time point, with differing percentages of the sample depending on the weight assigned. When less weight is assigned to precision, the first observation is selected and the optimal number of measurements equals one, for both CS and unstructured. In the second data-generation scheme, where treatment effects are allowed to vary, the same results followed, for both AR(1) and CS. We also gave some consideration to the Toeplitz, or banded, structure, where the correlation between pairs of measurements varies with the time lag between them, in an unstructured way, but is independent of the actual times at which the measurements are taken. Furthermore, an AR(1)-type structure was assumed where the decline in autocorrelation is expressed in terms of the square root of the time lag, denoted by AR(1)-Sq. The results are summarized in Table 9.3. For the Toeplitz structure up to

Table 9.1: *Simulation study. Results for the optimal number of measurements with AR(1). (ρ: correlation between successive time measurements; w_1: weight assigned to the precision part of the objective function; m_o: optimal number of measurements; $VRF_{ind}(m)$: individual-level surrogacy for the optimal number of measurements; $VRF_{ind}(K-1)$: expected value of individual-level surrogacy; f: percentage of datasets resulting in m_o is 100% in all cases.)*

		$VRF_{\text{ind}}(m)$				$VRF_{\text{ind}}(m)$	
w_1	m_o	as AR(1)	as CS	w_1	m_o	as AR(1)	as CS
$\rho = 0.30$ & $VRF_{\text{ind}}(K-1) = 0.09$				$\rho = 0.71$ & $VRF_{\text{ind}}(K-1) = 0.50$			
0.7	1	0.00003	0.0006	0.7	9	0.50	0.50
0.5	1	0.00003	0.0006	0.5	1	0.0032	0.0032
0.3	1	0.00003	0.0006	0.3	1	0.0032	0.0032
$\rho = 0.60$ & $VRF_{\text{ind}}(K-1) = 0.36$				$\rho = 0.90$ & $VRF_{\text{ind}}(K-1) = 0.81$			
0.7	9	0.36	0.42	0.7	9	0.81	0.81
0.5	1	0.07	0.07	0.5	9	0.81	0.81
0.3	1	0.07	0.07	0.3	1	0.15	0.15

five time points and for the unstructured matrix up to six time points were selected as optimum, depending on the weight assigned to the precision part of the cost function. For the AR(1)-Sq structure, the optimal time point swings between taking the first measurement or the entire sequence. However, it picks the first time point as optimal more often, except when the weight assigned to precision is as high as 70% and correlation values are 0.60 and 0.90. For a correlation of 0.30, it invariably picks the first time point only, even when the weight is as high as 70%.

9.4 Constrained Maximization

There are circumstances in which clinical trials are faced with budget constraints and yet are expected to produce acceptable results. This predicament motivates the use of constraint maximization to arrive at an optimal number of subjects and/or repeated measures per subject, thereby not exceeding the budget available. Translated to our setting, we aim at maximizing the individual level surrogacy measure, subject to cost and time constraints. We first maximize $VRF_{\text{ind}}(m)$ subject to $(R+m)/(R+K) \leq \delta_1$ and then later subject to two constraints given by: $(R+m)/(R+K) \leq \delta_1$ and $(t_m - t_0)/(t_k - t_0) \leq \delta_2$, where both δ_1 and δ_2 assume values between zero and one. Without loss of generality, if we assume that the measurements are equally spaced with fixed time interval $\triangle$, then $t_m = t_0 + \triangle M$ and $t_k = t_0 + \triangle K$ and hence the second constraint reduces to $M/K \leq \delta_2$. Using a Lagrange multiplier for the first optimization problem, one can show that, for CS with positive ρ, the optimal number of repeated measures required for a percentage budget of δ_1 is given as:

$$M = \begin{cases} \delta_1(R+K) - R & \text{if } (R+1) - \delta_1(R+k) \leq \frac{1}{\rho}, \\ 2\left(\frac{1-\rho}{\rho}\right) - \delta_1(R+K) + R & \text{if } (R+1) - \delta_1(R+k) \geq \frac{1}{\rho}. \end{cases}$$

In a similar manner, for AR(1) with $\rho \geq 0$, the optimal number of repeated measures for a given percentage of the budget is $M = \delta_1(R+K) - R$. If we now maximize the association measure subject to both budget and time constraint, we find $M = \min[\delta_1(R+K) - R, \delta_2 K]$ for the optimal number of repeated measures for both CS and AR(1). To enhance insight, we carried out a limited set of simulations for both AR(1) and CS. The simulation has revealed that as R increases, the optimal M diminishes. This is in line with intuition because the total cost and the number of subjects in the study are fixed and hence to maintain a low cost, the only option is to reduce the number of repeated measures. It also follows that, for some values of

R, it is not possible to obtain a value of M for which the percentage of cost incurred is lower than the specified δ value. In such cases, only the first time point or the entire sequence could be taken, depending on the magnitude of M. In this context, it is also worth noting that, although there is no difference in the optimal number of repeated measures for CS and AR(1), the same number of repeated measures in the two covariance structures will nevertheless not yield identical $VRF_{\text{ind}}(m)$ values.

9.5 Application to the Case Study

Two case studies introduced in Chapter 2, Sections 2.1.1 and 2.1.2 are analyzed here and the results displayed in Tables 9.4 and 9.5, respectively. For the data coming from the opthalmology experiment, measurements of visual acuity were taken at baseline and every sixth week there after up to the 54^{th} week giving 10 repeated measures. For the schizophrenia study, the PANSS values were measured at five different time points, taken at the baseline and every two weeks thereafter. In both cases, the objective is to predict the ultimate measurement using earlier ones from the sequence, thereby accounting for cost. In both cases, an unstructured variance-covariance matrix fits the data best. Now focusing attention on the data coming from the opthalmology experiment, we find that, with increasing weight attributed to precision: the first one; the first and the second; the first, the second, and the third; the first eight; or all nine time points were required to optimally predict the final measurement. Note that one time unit corresponds to 6 weeks. Thus, for example, taking the first three time points amounts to using measurements from 18 weeks to predict a response at the 54^{th} week. For the schizophrenia experiment, first, to stabilize the variance, a linear transformation of the outcome and a non-linear transformation of time, taking the form $Y_{ij} = -3.5675 + 0.0484 \cdot \text{PANSS}_{ij}$ and $t_{j,\text{new}} = e^{-t_j/4}$, respectively, were

applied. It follows that, with increasing weight assigned to precision: the first one; the first and the second; or all four time points were required to optimally predict the final measurement. In line with intuition, in both cases, the number of time points required also changes with increasing R. Setting $R = 0$ corresponds to assuming that subjects are recruited at no cost or when interest is solely with the cost per additional measurement occasion. To accommodate the waiting time in the decision making process, we also studied the optimal number of time points based on the modified cost functions (9.5) and (9.6). Results can be found in Table 9.5 for schizophrenia and Table 9.6 for opthalmology. The modified functions lead to the same results when $R = 0$, but, as R increases, the modified cost functions are more prudent and tend to select less time points.

9.6 Discussion

In this chapter, unlike conventional surrogate marker validation, which involves two separate outcomes where one is used as a potential surrogate for the other, we have studied a scenario where there is only a single outcome, measured repeatedly over time. The objective was to assess the performance of accumulated measures of an equally spaced longitudinal sequence as a possible surrogate for a final outcome and to determine the optimal number of repeated measures required to adequately attain 'good' surrogacy. The determination of the optimal number of measurements requires striking a balance between precision and cost of incorporating a long sequence of repeated measures. To this end, an objective function has been utilized. The objective function has two parts, which take care of the cost and precision components. The importance of both components is gauged through the use of weights. Whenever it is felt that the importance of precision outweighs cost, more weight will be assigned

to the precision part and vice versa. The objective function can be modified to accommodate other possible sources of cost. One such cost is the cost of waiting time. This can be incorporated through a third component which accounts for the time lag between the start of the study and the optimal time point. This calls for assigning three possible weights, corresponding to financial cost, time cost, and precision cost, respectively. The results of the simulation study for two data-generation schemes, based on CS and AR(1), have revealed that, depending on the correlation structure of the data and the weights assigned, the first few repeated measures or the entire $K-1$ sequence might be needed to adequately predict the outcome at the last time point. Assuming that the outcome has an AR(1) structure, we showed theoretically and via simulations that either only the first measurement or the entire $K-1$ sequence is required to predict the true endpoint, depending on the weights chosen and the level of the AR(1) correlation. This is a very interesting characteristic of the first-order auto-regressive structure. Our results illustrate that here no balance between precision and cost is possible, because the objective function always leads to the two extreme situations. If precision is the driving requirement, then the entire $K-1$ subsequence is the best option, whereas if cost if the impelling factor then the surrogate should never contain more than a single observation. In such a situation, the best strategy will be to use only one measurement, located somewhere in the interval $(1, K-1)$. Obviously if the observation is taken at the end of the sequence, more predictive power will be achieved but a longer waiting time will also be needed. Arguably, a decision should then be taken based on other field related factors and the opinion of the experts in the area will be important. Moreover, at most six measurements, about 60% of the entire sequence, are required to adequately predict the final measurement if the outcome has a CS or a Toeplitz structure, or a general structure with slowly

decaying correlation between repeated measures. Based on these findings, it seems promising to use the proposed approach to balance between cost and precision in the process of evaluating the performance of a few repeated measures taken early as possible surrogates to adequately predict the outcome and/or treatment effect of the final measure.

Table 9.2: *Simulation study. Results for the optimal number of measurements with CS. (ρ: correlation between successive time measurements; w_1: weight assigned to the precision part of the objective function; m_o: optimal number of measurements; $VRF_{ind}(m)$: individual-level surrogacy for the optimal number of measurements; $VRF_{ind}(K-1)$: expected value of individual-level surrogacy; f: percentage of datasets resulting in m_o.)*

		as CS		as UN	
w_1	m_o	$VRF_{\text{ind}}(m)$	f	$VRF_{\text{ind}}(m)$	f
	$\rho = 0.30$ & $VRF_{\text{ind}}(K-1) = 0.24$				
0.7	1	0.11	18	0.10	18
0.7	2	0.14	6	0.12	34
0.7	3	0.16	60	0.17	22
0.7	4	0.19	16	0.19	26
0.5	1	0.09	100	0.09	100
0.3	1	0.09	100	0.09	100
	$\rho = 0.60$ & $VRF_{\text{ind}}(K-1) = 0.56$				
0.7	3	0.49	60	0.48	62
0.7	4	0.52	40	0.51	38
0.5	1	0.37	30	0.37	18
0.5	2	0.44	70	0.43	82
0.3	1	0.36	100	0.36	100
	$\rho = 0.71$ & $VRF_{\text{ind}}(K-1) = 0.68$				
0.7	2			0.58	14
0.7	3	0.62	100	0.62	80
0.7	4			0.64	6
0.5	3			0.62	70
0.5	4			0.64	6
0.5	1	0.51	30		
0.5	2	0.58	70	0.57	24
0.3	1	0.50	100	0.50	100
	$\rho = 0.90$ & $VRF_{\text{ind}}(K-1) = 0.89$				
0.7	2	0.85	100	0.85	100
0.5	1	0.81	100	0.81	100
0.3	1	0.81	100	0.81	100

Table 9.3: *Simulation study. Results for the optimal number of measurements with: unstructured covariance; Toeplitz correlation structure with slowly declining correlation; and AR(1) with square root of time lag analyzed as conventional AR(1). (w_1: weight assigned to the precision part of the objective function; m_o: optimal number of measurements; $VRF_{ind}(m)$: individual-level surrogacy for the optimal number of measurements; $VRF_{ind}(K-1)$: expected value of individual-level surrogacy; f: percentage of datasets resulting in m_o.)*

w_1	m_o	$VRF_{\text{ind}}(K-1)$	f
Unstructured			
$VRF_{\text{ind}}(K-1) = 0.995$			
0.1	1	0.53	100
0.3	1	0.53	100
0.5	4	0.86	92
0.5	5	0.91	8
0.7	6	0.96	100
0.6	4	0.86	29
0.6	5	0.91	57
0.6	6	0.96	14
Toeplitz			
$VRF_{\text{ind}}(K-1) = 0.75$			
0.1	1	0.15	100
0.3	2	0.16	80
0.3	3	0.22	20
0.5	4	0.38	100
0.6	4	0.38	98
0.6	5	0.42	2
0.7	5	0.42	100

w_1	m_o	$VRF_{\text{ind}}(K-1)$	f
AR(1)-Sq			
$\rho = 0.30$ & $VRF_{\text{ind}}(K-1) = 0.22$			
0.1	1	0.0016	100
0.3	1	0.0016	100
0.5	1	0.0016	100
0.6	1	0.0016	100
0.7	1	0.0016	100
AR(1)-Sq			
$\rho = 0.60$ & $VRF_{\text{ind}}(K-1) = 0.50$			
0.1	1	0.052	100
0.3	1	0.052	100
0.5	1	0.052	100
0.6	9	0.052	100
0.7	9	0.052	100
AR(1)-Sq			
$\rho = 0.90$ & $VRF_{\text{ind}}(K-1) = 0.86$			
0.1	1	0.21	100
0.3	1	0.21	100
0.5	1	0.21	100
0.6	1	0.21	100
0.7	9	0.86	100

Table 9.4: *Case study in opthalmology. Results for the optimal number of measurements based on cost function (9.3). (w_1: weight assigned to the precision part of the objective function; m_o: optimal number of measurements;$R = C_1/C_2$ be the cost ratio ; $VRF_{ind}(m)$: individual-level surrogacy for the optimal number of measurements; $VRF_{ind}(K-1)$: expected value of individual-level surrogacy.)*

$VRF_{\text{ind}}(K-1) = 0.91$							
w_1	R	m_o	VRF_{ind}	w_1	R	m_o	VRF_{ind}
0.1	0	1	0.18	0.1	4	1	0.18
0.3	0	1	0.18	0.3	4	1	0.18
0.4	0	2	0.34	0.4	4	3	0.45
0.5	0	3	0.45	0.5	4	8	0.85
0.7	0	9	0.91	0.7	4	9	0.91
0.1	1	1	0.18	0.1	6	1	0.18
0.3	1	1	0.18	0.3	6	2	0.34
0.4	1	2	0.34	0.4	6	3	0.45
0.5	1	3	0.45	0.5	6	8	0.85
0.7	1	9	0.91	0.7	6	9	0.91
0.1	2	1	0.18				
0.3	2	1	0.18				
0.4	2	2	0.34				
0.5	2	3	0.45				
0.7	2	9	0.91				

Table 9.5: *Case study in schizophrenia. Results for the optimal number of measurements based on cost function (9.3) and modified cost function (9.5). (w_1: weight assigned to the precision part of the objective function; m_o: optimal number of measurements;$R = C_1/C_2$ be the cost ratio ; $VRF_{ind}(m)$: individual-level surrogacy for the optimal number of measurements; $VRF_{ind}(K-1)$: expected value of individual-level surrogacy.)*

$VRF_{\text{ind}}(K-1) = 0.85$									
Cost function (9.3)				Cost function (9.5)					
w_1	R	m_o	$VRF_{\text{ind}}(m)$	w_1	w_2	w_3	R	m_o	$VRF_{\text{ind}}(m)$
0.1	0	1	0.20	0.1	0.1	0.8	0	1	0.20
0.3	0	1	0.20	0.3	0.1	0.6	0	1	0.20
0.5	0	2	0.59	0.5	0.1	0.4	0	2	0.59
0.7	0	4	0.85	0.7	0.1	0.2	0	4	0.85
0.1	1	1	0.20	0.1	0.1	0.8	1	1	0.20
0.3	1	2	0.59	0.3	0.1	0.6	1	1	0.20
0.5	1	2	0.59	0.5	0.1	0.4	1	2	0.59
0.7	1	4	0.85	0.7	0.1	0.2	1	4	0.85
0.1	2	1	0.20	0.1	0.1	0.8	2	1	0.20
0.3	2	2	0.59	0.3	0.1	0.6	2	1	0.20
0.5	2	2	0.59	0.5	0.1	0.4	2	2	0.59
0.7	2	4	0.85	0.7	0.1	0.2	2	4	0.85
0.1	4	1	0.20	0.1	0.1	0.8	4	1	0.20
0.3	4	2	0.59	0.3	0.1	0.6	4	1	0.20
0.5	4	4	0.85	0.5	0.1	0.4	4	2	0.59
0.7	4	4	0.85	0.7	0.1	0.2	4	4	0.85
0.1	6	1	0.20	0.1	0.1	0.8	6	1	0.20
0.3	6	2	0.59	0.3	0.1	0.6	6	1	0.20
0.5	6	4	0.85	0.5	0.1	0.4	6	2	0.59
0.7	6	4	0.85	0.7	0.1	0.2	6	4	0.85

Table 9.6: *Case study in opthalmology. Results for the optimal number of measurements based on modified cost function (9.5) and (9.6); (w_1-w_3): weights assigned to the precision, financial cost and waiting time parts of the objective function; m_o: optimal number of measurements; $R = C_1/C_2$ be the cost ratio; $VRF_{ind}(m)$: individual-level surrogacy for the optimal number of measurements; $f = 100$: percentage of datasets resulting in m_o, in all cases.)*

Weights			Cost Ratios									
			$R=0$		$R=1$		$R=2$		$R=4$		$R=6$	
w_1	w_2	w_3	m_o	VRF_{ind}	m_o	VRF_{ind}	m_o	VRF_{ind}	m_o	VRF_{ind}	m_o	VRF_{ind}
Modified cost function (9.5)												
0.1	0.1	0.8	1	0.18	1	0.18	1	0.18	1	0.18	1	0.18
0.3	0.1	0.6	1	0.18	1	0.18	1	0.18	1	0.18	1	0.18
0.4	0.1	0.5	2	0.34	2	0.34	2	0.34	2	0.34	2	0.34
0.5	0.1	0.4	3	0.45	3	0.45	3	0.45	3	0.45	3	0.45
0.7	0.1	0.2	9	0.91	9	0.91	9	0.91	9	0.91	9	0.91
0.1	0.2	0.7	1	0.18	1	0.18	1	0.18	1	0.18	1	0.18
0.3	0.2	0.5	1	0.18	1	0.18	1	0.18	1	0.18	1	0.18
0.4	0.2	0.4	2	0.34	2	0.34	2	0.34	2	0.34	2	0.34
0.5	0.2	0.3	3	0.45	3	0.45	3	0.45	3	0.45	3	0.45
0.6	0.2	0.2	8	0.85	8	0.85	8	0.85	9	0.91	9	0.91
Modified cost function (9.6)												
0.1	0.1	0.8	1	0.18	1	0.18	1	0.18	1	0.18	1	0.18
0.3	0.1	0.6	1	0.18	1	0.18	1	0.18	1	0.18	1	0.18
0.4	0.1	0.5	1	0.18	1	0.18	2	0.34	2	0.34	2	0.34
0.5	0.1	0.4	2	0.34	3	0.45	3	0.45	3	0.45	3	0.45
0.7	0.1	0.2	9	0.91	9	0.91	9	0.91	9	0.91	9	0.91
0.1	0.2	0.7	1	0.18	1	0.18	1	0.18	1	0.18	1	0.18
0.3	0.2	0.5	1	0.18	1	0.18	1	0.18	1	0.18	1	0.18
0.4	0.2	0.4	2	0.34	2	0.34	2	0.34	2	0.34	2	0.34
0.5	0.2	0.3	3	0.45	3	0.45	3	0.45	3	0.45	3	0.45
0.6	0.2	0.2	8	0.85	8	0.85	8	0.85	8	0.85	8	0.85

10

Predicting the Final Outcome of a Binary Longitudinal Response

Similar to the case of a continuous longitudinal outcome outlined in the previous chapter, one might be interested in predicting the final outcome of a binary longitudinal sequence using cumulative earlier measurement subject to cost and time constraints. To this end, we will devise the same set of methodology used in the previous chapter to address the stated objective. For the case of predicting the final measurement of the continuous longitudina1 outcome, we have used the VRF as a measure of association between earlier measures and final outcome. For this case however, we will use the information-theoretic approach measure denoted by R_h^2. The objective function, which accounts for cost of time and financial cost coupled with the precision of

prediction, can then take the following format:

$$CPR_I(m) = w_1 \cdot [1 - R_h^2(m)] + w_2 \cdot \frac{R+m}{R+K} + w_3 \cdot \frac{t_m - t_0}{t_k - t_0}, . \qquad (10.1)$$

10.1 Simulation Study

The methodology has been thoroughly discussed in the previous chapter and hence we proceed with a small simulation study to asses how the method fairs within the context of binary longitudinal outcome. We begin with a concise description of the data generation schemes followed.

10.1.1 Generating Binary longitudinal outcome

Data were generated under three different scenarios: the Bahadur model and under the assumption of first order autoregressive and compound symmetric correlation structures for the error terms. Let $Y = (Y_1, ..., Y_T)$ represent a vector of binary responses for any subject. Further, we let the marginal mean of Yt , where $t = 1, ..., T$, be $E(Y_t) = P(Y_t = 1) = \mu_t$. In addition, for any subject, we let the correlation between the two binary responses Y_t and $Y_t^{'}$, $t \neq t^{'}$, be $Corr(Y_t, Y_t^{'}) = \rho_{tt'}$. One method for generating longitudinally correlated binary data is based on the work of Bahadur (1961), who proposed a representation for multivariate binary distributions which is expressed as a joint mass function of $Y_1, ..., Y_T$. Specifically, if all coefficients of order three and higher are ignored,

$$f(Y_1, ..., Y_T) \quad = \quad \left\{ \prod_{t=1}^{T} \mu_t^{y_t} (1 - \mu_t)^{1-y_t} \right\} \left\{ 1 + \sum_{1 \leq t \leq t'} \rho_{tt'} \tilde{Y}_t \tilde{Y}_{t'} \right\}, \qquad (10.2)$$

where $\tilde{Y}_t = (Y_t - \mu_t)/\sqrt{\mu_t(1 - \mu_t)}$ results from simply standardizing Y_t. This joint distribution is used to determine an expression for the conditional probabil-

ity $P(Y_t = 1|Y_{t-1}, ..., Y_1)$, which is subsequently employed to generate a value for Y_t. Unfortunately, Bahadur's (1961) representation is computationally difficult to manipulate when T is large, and becomes even more so when the higher-order coefficients are not ignored. Furthermore, truncation of the representation after the second-order term to reduce computational complexity will come at the expense of limited dependence ranges for the binary responses. However, the Bahadur (1961) representation does offer some flexibility in that in order to generate binary responses it is not necessary to impose standard-type correlation structures on the Y_t, such as autoregressive of order one Patrick J. Farrell and Katrina Rogers-Stewart (2008). In most real life applications however, standard-type correlation structures on the Y_t, such as autoregressive of order one and compound symmetry are common place. Here we briefly introduce the data generation scheme for these two special cases as outlined in Patrick J. Farrell and Katrina Rogers-Stewart (2008). Kanter (1975) used a model which was designed to only generate correlated binary responses according to a stationary autoregressive process of order p, which we will refer to as AR(p). Initially, a value for Y_1 is generated from a Bernoulli distribution with parameter μ. Then, in order to generate $Y_t, t = 2, ..., T$ according to an autoregressive process of order $min(t.1, p)$, Kanter (1975) proposes the model

$$Y_t = \sum_{t'=1}^{min(t-1,p)} U_t^{(t')}(Y_{t-t'} \oplus W_t) \left\{ 1 - \sum_{t'=1}^{min(t-1,p)} U_t^{(t')} \right\} W_t, \qquad (10.3)$$

where $\oplus$ represents addition modulo 2. In addition, $U_t^{(t')}$ and W_t are generated as Bernoulli random variables with parameters $\xi^{(t')}$ and η, which are determined by the values of μ and the correlation parameters. In a spirit similar to Kanter (1975), Lunn and Davies (1998) introduce an efficient model for the generation of stationary binary

response data with mean μ that are correlated according to either a stationary AR(1), or an exchangeable process with parameter ρ. For $AR(1)$ data, Lunn and Davies (1998) suggest to initially generate Y_1 from a Bernoulli distribution with parameter μ, and then Y_t for $t = 2, ..., T$ according to $Y_t = A_t Y_{t-1} + (1 - A_t)B_t$, where A_t and B_t are generated as Bernoulli random variables with parameters ρ and μ, respectively. Since ρ must be treated as a Bernoulli parameter here, it is only possible to generate AR(1) binary sequences with a positive correlation. Generation of exchangeable binary data proceeds in a similar fashion. Initially, Y_0 is generated from a Bernoulli distribution with parameter ρ, and then the sequence is generated using $Y_t = A_t Y_0 + (1 - A_t)B_t$, where A_t and B_t are generated as Bernoulli random variables with parameters $\sqrt{(}\rho)$ and μ, respectively. For this study we have generated data with compound symmetry and Autoregressive of order one following the suggestion of Lunn and Davies (1998) and using Bahadur's formulation.

10.1.2 Simulation Study Results

The results for the simulation study have shown that, for the CS and AR(1) processes, the R^2 values start from relatively small values and gradually increases as the number of repeated measures increases. For the Bahadur model, the R^2 values start at a relatively higher value and increase gradually to even higher values. The optimal number of repeated measures required to adequately predict the final outcome considering cost and waiting time has also been considered. For the Bahadur model up to 8 time points were selected optimal with varying percentages of samples pointing to different number of time points depending on the magnitude of the weight assigned. For the AR(1) and CS structures more or less similar results to the continuous case are observed, i.e., for AR (1) swinging between 1 and 9 time points but with some

percentage of samples pointing to other possible number of measurements as optimal. For CS up to 5 time points were selected as optimal. The effect of R, the cost ratio of recruiting a patient and taking repeated measures was minimal or non-existent.

10.2 Application to the Case Study

The case study on Age related Macular Degeneration introduced in Chapter 2, Sections 2.1.1 was used to demonstrate the application of the method for the prediction of the final outcome of binary longitudinal sequence. The visual acuity measures were first dichotomized by taking the change from baseline values. If the visual acuity shows an increase relative to the baseline it will be set to one and zero otherwise. The dichotomized version of the dataset was used to demonstrate the performance of the method. The results of the analysis for a cost ratio $R = 4$ are shown in Tables 10.1-10.4. The results have shown that either the first measure or the entire sequence are required to adequately predict the final outcome. This is in agreement with the simulation results; as the correlation structure that best fits the data was found to be an AR(1). The effect of the cost ratio is minimal or non-existent.

10.3 Discussion

In this chapter, we have devised the same methodology that was used in the previous chapter to predict the final outcome of a binary longitudinal sequence using earlier measures subject to cost and time constraints. We learn from the results that for the Compound symmetry correlation structure, a varying number of samples have suggested the use of few earlier repeated measures to predict the final outcome. For the Auto regressive of order-one correlation structure on the other hand, either the first time or the entire sequence is required to predict the final outcome depending

Table 10.1: *Results for the optimal number of measurements for the case study in Ophthalmology based on* $CPR_0(m)$.w_1*: weight assigned to the precision part of the objective function;* m_o*: the optimal number of measurements;* R_h^2*: individual-level surrogacy for the optimal number of measurements;* f*: percentage of datasets resulting in* m_o*.)*

w_1	R	R_h^2	m_o	f
0.1	4	0.12221	1	100
0.3	4	0.12221	1	100
0.5	4	0.12221	1	100
0.6	4	0.58070	8	100
0.7	4	0.58070	8	100

on the weight assigned to the precision part of the cost function. The results of the Bahadur model show that up to a maximum of 8 time points might be required to make adequate predictions. The results for the Compound symmetry and Auto regressive of order-one structures are in synchrony with the results obtained for the case of continuous longitudinal outcome. This method proves to be very beneficial if the correlation structure can be assumed to be Compound symmetry or the data can be assumed to follow a Bahadur model since for these cases first few measurements are required to adequately predict the outcome at the end of the study. This is a desirable property, as with only few repeated measures, it will be possible to predict the outcome which could have taken longer and incurred more cost.

Table 10.2: *Results for the optimal number of measurements for the case study in Ophthalmology based on $CPR_I(m)$.w_1-w_3: weights assigned to the different parts of the objective function; m_o: the optimal number of measurements; R_h^2: individual-level surrogacy for the optimal number of measurements; f: percentage of datasets resulting in m_o.)*

w_1	w_2	w_3	R	R_h^2	m_o	f
0.1	0.1	0.8	4	0.12221	1	100
0.3	0.1	0.6	4	0.12221	1	100
0.5	0.1	0.4	4	0.12221	1	100
0.7	0.1	0.2	4	0.58070	8	100
0.1	0.2	0.7	4	0.12221	1	100
0.3	0.2	0.5	4	0.12221	1	100
0.5	0.2	0.3	4	0.12221	1	100
0.6	0.2	0.2	4	0.58070	8	100
0.1	0.3	0.6	4	0.12221	1	100
0.3	0.3	0.4	4	0.12221	1	100
0.5	0.3	0.2	4	0.12221	1	100
0.6	0.3	0.1	4	0.58070	8	100

Table 10.3: *Results for the optimal number of measurements for the case study in Ophthalmology based on $CPR_{II}(m)$.w_1-w_3: weights assigned to the different parts of the objective function; m_o: the optimal number of measurements; R_h^2: individual-level surrogacy for the optimal number of measurements; f: percentage of datasets resulting in m_o.)*

w_1	w_2	w_3	R	R_h^2	m_o	f
0.1	0.1	0.8	4	0.12221	1	100
0.3	0.1	0.6	4	0.12221	1	100
0.5	0.1	0.4	4	0.12221	1	100
0.7	0.1	0.2	4	0.58070	8	100
0.1	0.2	0.7	4	0.12221	1	100
0.3	0.2	0.5	4	0.12221	1	100
0.5	0.2	0.3	4	0.12221	1	100
0.6	0.2	0.2	4	0.58070	8	100
0.1	0.3	0.6	4	0.12221	1	100
0.3	0.3	0.4	4	0.12221	1	100
0.5	0.3	0.2	4	0.12221	1	100
0.6	0.3	0.1	4	0.58070	8	100

Table 10.4: *Results for the optimal number of measurements for the case study in Ophthalmology based on $CPR_{II}(m)$.w_1-w_3: weights assigned to the different parts of the objective function; m_o: the optimal number of measurements; R_h^2: individual-level surrogacy for the optimal number of measurements; f: percentage of datasets resulting in m_o.)*

w_1	w_2	w_3	R	R_h^2	m_o	f
0.1	0.1	0.8	4	0.12221	1	100
0.3	0.1	0.6	4	0.12221	1	100
0.5	0.1	0.4	4	0.12221	1	100
0.7	0.1	0.2	4	0.58070	8	100
0.1	0.2	0.7	4	0.12221	1	100
0.3	0.2	0.5	4	0.12221	1	100
0.5	0.2	0.3	4	0.12221	1	100
0.6	0.2	0.2	4	0.58070	8	100
0.1	0.3	0.6	4	0.12221	1	100
0.3	0.3	0.4	4	0.12221	1	100
0.5	0.3	0.2	4	0.12221	1	100
0.6	0.3	0.1	4	0.58070	8	100

Part II

Selection and Evaluation of Biomarkers

11

Genomic Biomarkers: Feature-specific and Joint Biomarkers

A biomarker can be defined as a physical sign or laboratory measurement that serves as an indicator for biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention Lesko and Atkinson (2001). When the measurement in question is the expression of a gene, we refer to the gene as a genomic biomarker. We can differentiate between feature-specific and joint biomarkers. Feature-specific biomarker refers to a single biomarker used on its own to explain characteristics of the clinical outcome. Whilst a joint biomarker involves a combination of biomarkers combined according to some guideline. The focus of this chapter will therefore be to apply different statistical methods to select and evaluate genomic biomarkers for continuous clinical outcome. We will first introduce approaches that

are designed for selecting feature-specific biomarkers and then move on to two approaches designed for constructing joint biomarkers. The methods described will then be applied to a case study in depression where we select genes and metabolites as both feature-specific and joint biomarkers for depression measured by Hamilton Depression Scale (HAMD).

11.1 Feature-specific Biomarkers in Microarray Experiments

The main goal in this setting is to identify specific biomarkers for a particular response of interest. As stated earlier, their is an analogy between prognostic biomarker and individual level surrogate defined within the surrogate marker validation framework. The implication of this analogy is that, the methods designed for the evaluation of surrogacy at the individual patient level can be used with no or little modification to select and evaluate prognostic biomarkers. To this end, for univariate outcomes, we can use a joint model of the response and the gene expression in line with the model introduced in chapter (3), for the case of two normally distributed outcomes and quantify the association using the individual-level R^2. Another alternative approach is the information-theoretic approach, which can be used both for normal and non-normal outcomes. The R^2_Λ, introduced for two longitudinal outcomes, can be used for repeatedly measured response and gene expressions. Note however that, all these models are fitted for each gene separately, a procedure often termed "gene-by-gene" analysis. It is also important to highlight that a microarray experiment is equivalent to the single trial setting and hence, the gene specific models used to compute the association measures should be tuned to reflect the single trial setting.

11.2 Joint Biomarkers in Microarray Experiments

In this section, we focus on the question: "How to combine information about expression levels from all genes in the array into one variable?" which we term a "joint biomarker." In the microarray setting, the number of predictors is large compared to the number of observations, and the design matrix of the gene expressions S is likely to be singular, making a regression approach to summarize information into one linear predictor no longer feasible. Several approaches have been developed to cope with this problem. One approach is to perform a principal component analysis (PCA) of the S matrix and then use the principal components of S as regressors on the response of interest Bair *et al* (2006). The orthogonality of the principal components eliminates the multicollinearity problem. A possible strategy is to keep only the first few components. Alternatively, so-called partial least squares regression can be employed Herve Abdi(2003). In the following sections, we briefly outline supervised principal components analysis and partial least squares regression respectively.

11.2.1 Supervised Principal Component Analysis

The Supervised principal Component Analysis (SPCA) relies on the underlying assumption that there is a latent variable $U(\S)$, associated with the response variable T. Because in almost all cases the number of genes is much larger than the number of observations, the first step in the SPCA method is data reduction in which the dimension of the expression matrix $\S$ is reduced. In line with Bair *et al* (2006), a fully supervised method is expected to give the most weight to those genes having the strongest relationship with the response. The SPCA approach ensures that $U(\S)$ will be constructed in such a way that the association between the joint biomarker and the response will be maximized. The SPCA methods consists of three main steps:

- Step 1: Fit one of the gene-specific models and estimate the association measure.
- Step 2: Form a reduced expression matrix consisting of only those genes whose gene specific association measure exceeds a threshold level.
- Step 3: Let $\mathbf{S_R}$ be the reduced matrix. Compute for each matrix the first principal component, $U(\mathbf{S_R})$.

The three steps ensure that selection of the subset of genes from which the first principal component is calculated will maximize the corresponding association measure. This is a crucial point: because the joint biomarker is latent, it is constructed in such a way that it will maximize the association measure. As a consequence, for a given dataset, $U(\mathbf{S_R})$ is the "best" joint biomarker. Once the $U(\mathbf{S_R})$ is computed, its association with the response can be quantified using the measures used for the selection of feature-specific biomarkers.

11.2.2 Supervised Partial Least Squares

Partial least squares (PLS) regression is a technique that generalizes and combines features from principal component analysis and multiple regression. In PCA, the strategy implemented was to keep only a few of the first components. But because the components are to explain S rather than T, there is no guarantee that the principal components, which explain S, are relevant for T. In contrast, PLS regression finds components from S that are also relevant for T. Specifically, PLS regression searches for a set of components, called latent vectors, that perform a simultaneous decomposition of S and T, with the constraint that these components explain as much as possible of the covariance between S and T. This step generalizes PCA. It is followed by a regression step where the decomposition of S is used to predict T. For our particular case, the PLS is used to create a joint biomarker that optimally

explains the response. The procedure is supervised, because the genes to be used for the construction of the factors that explain the response are selected based on the strength of their association with the response. The general steps in SPLS can be summarized as follows:

- Step 1: Fit one of the gene-specific models and estimate the association measure.
- Step 2: Form a reduced expression matrix consisting of only those genes whose gene specific association measure exceeds a threshold level.
- Step 3: Let $\mathbf{S_R}$ be the reduced matrix. Fit a partial least squares regression and take the first factor, $U(\mathbf{S_R})$.
- Step 4: Select the genes with the largest influence on the resulting latent factor in Step 3 and use them to construct the joint biomarker using PLS.

11.3 Application to the case Study

The different methods discussed earlier are applied to the case study in depression introduced in Chapter 2, Sections 2.1.6. We first introduce the results for feature-specific biomarkers and then move to the joint biomarkers based on principal components and partial least squares respectively.

11.3.1 Feature-specific Biomarkers

For all the patients in the study, 17502 genes, 269 metabolites, and a HAMD score were measured before and after treatment with the objective of identifying genes and metabolites as potential biomarkers for depression. Association measures based on a joint model and information-theoretic approaches were computed for all genes and metabolites after correcting for some other variables such as storage time, gender and

age of the patient. Note however that, since all patients are treated, there is no need to adjust for treatment effect in the usual sense. Rather, the treatment effect is accounted for by taking the difference from baseline. The results are summarized in Tables 11.1 and 11.2 and Figures 11.2–11.4. After multiplicity adjustment using the *false discovery rate* (FDR) approach (Benjamini and Hochberg, 1995), two genes (736, 2419) and three metabolites (68, 12, and 67) were found to be significant. As theoretically expected, the results from the joint model and the information-theoretic approach are the same. Figures 11.2 and 11.3 depict the scatter plot of the residuals from the top four genes/metabolites and HAMD score. It is possible to observe that there appears to be a linear association between the HAMD score and the genes/metabolites after adjusting for treatment and other covariates. Figure 11.4 shows density plots of the R^2 values for the whole 17502 genes and 269 metabolites. As expected, the majority of genes/metabolites have very low correlation with the HAMD score after adjusting for the confounders. Instead of taking the changes from baseline, if we directly consider the pre and post treatment HAMD scores and the pre and post gene/metabolite values, we can measure the association between some linear combination of the pre and post treatment HAMD scores and pre and post gene/metabolite values using the R^2_Λ. The results are summarized in Tables 11.3 and 11.4. In these tables, the top 20 genes/metabolites with higher R^2_Λ values are displayed. The coefficients corresponding to the pre/post HAMD and gene/metabolite can be used to construct the linear combination whose association has been quantified by the R^2_Λ. These linear combinations can be viewed as weighted sums of the pre/post measurements, although some of them might not have a clear biological interpretation. However, they still can serve as possible transformations of the pre/post HAMD score and the gene/metabolite measurements that can maximize the association between the two group of measures.

The plot in Figure 11.5 depicts the fact that there is a linear combination of the pre/post HAMD score that is strongly correlated with a linear combination of the pre/post gene expression other than the change from baseline for gene 12161. From the tables, one also notes that different sets of genes/metabolites are selected as top 20 when on the one hand R^2_Λ and on the other hand the information-theoretic or joint model approach were used. This is, however, expected since R^2_Λ quantifies the association between the vector of pre/post HAMD score and pre/post gene/metabolite values. However, in the cases of the of the later two, we related the changes after treatment. Note also that the leave-one-out cross validation results are comparable with the original measures, giving comfort to the validity of the measures. However, it is important to mention that, after adjustment for multiplicity, using (FDR) approach (Benjamini and Hochberg, 1995) none of the genes/metabolites were found to be significant for R^2_Λ.

11.3.2 Joint Biomarkers Using Principal Component Analysis

Up until now, we have been able to identify a set of genes/metabolites in the array as possible biomarkers for the HAMD score. However, instead of taking a particular gene/metabolite as a biomarker, information gain might be achieved if a joint biomarker could be constructed. To this end, we have used the supervised principal component analysis discussed in Section 11.2. Once the top k genes/metabolites are identified based on their correlation with the HAMD score, i.e the subset of k genes/metabolites with the highest threshold level, a principal component involving these top genes/metabolites is constructed as a possible joint biomarker. Three approaches have been followed. The first approach achieves the stated objective by taking the top k gene/metabolites and constructing a first principal component as

joint biomarker profile. If we observe Figure 11.6, and Table 11.5, we can easily notice that when the fifth best gene is included in the construction of the principal component, the correlation between the gene profile and the HAMD score becomes smaller than by merely taking the top four genes and hence a second approach is considered. This is similar to the first approach except that a gene/metabolite will be included as part of the gene/metabolite profile only if it results in an increase of the correlation between the profile and the HAMD score. Using the second gene profiling approach, 6 genes were considered in the construction of the the gene profile, giving an R^2 value of 0.8923, which is higher than taking for example the top 20 genes at once. In a similar manner, for the metabolites, following the first approach, the association increases from the the first to the second metabolites and then declines when the third best metabolite is included as the component of the joint biomarker profile. Thus here also, the second approach was entertained and 9 metabolites were selected, producing an $R^2 = 0.9433$, which is higher than taking the top 20 metabolites to construct a metabolite profile. The third approach involves three steps: (1) construct a principal component based on top k genes/metabolites; (2) re-rank the genes/ metabolites based on their loadings in the principal component; (3) construct a joint biomarker based on the top genes/metabolites with larger weights. The results for the third approach are summarized in Table 11.6.

Apart from the fact that the use of the supervised principal component analysis is tempting, in that it maximizes the measure of association with the response, as opposed to taking a single gene/metabolite, there is a need to perform a significance test on the resulting measure of association as it might not be statistically significant. We have performed a permutation-based test to asses the significance of the correlation between the response and the joint biomarker. As can be observed from

Tables 11.5 and 11.6, there is a significant association between the joint metabolite biomarkers and the response, irrespective of the number of top metabolites considered and the approach followed. On the other hand, none of the joint genomic biomarkers constructed based on the top k genes have a significant correlation with the response. This prompts caution in using the joint biomarker constructed by using the principal components analysis approach without proper significance testing of the inflated association measure.

11.3.3 Joint Biomarkers Using Partial Least Squares

Similar to the supervised principal components analysis discussed earlier, here also we begin with the selection of gene/metabolite-specific biomarkers. The top k genes and metabolites selected based on the information-theoretic approach will be used as inputs for the partial least squares regressions. Two approaches were followed. First, similar to the SPCA, the joint biomarker was created by incorporating the top k genes/metabolites into the construction of the latent factor. As can be observed from Figure 11.6, similar to the SPCA, the association measure increases and decreases with the inclusion of more genes/metabolites. A second approach is carried out as follows. To begin with, all the top k genes/metabolites will be used in the partial least squares regression to create the latent factor that is associated with the HAMD score. Then the genes/metabolites are re-ranked according to the absolute value of their corresponding weights. The joint biomarker is then constructed by selecting the top genes/metabolites, in terms of the absolute value of the weight, one by one starting with the top gene/metabolite until inclusion of gene/metabolites does not improve the amount of variation explained by the joint biomarker. The results are summarized in Tables 11.7 and 11.8. Under this approach, for both the gene and

metabolite expressions, it was observed that the explained variation increases up to a certain limit and then decreases, and seems to stabilizes after a while. This implies that no additional information will be gained about the response by incorporating more genes/metabolites. The test of significance for the joint biomarker based on the PLS approach revealed that only the joint biomarker constructed based on the top 2, 3, or 4 genes is significant, while the joint biomarker involving any number of the top 20 metabolites was significant.

11.4 Discussion

The primary objective of the analysis in this chapter was to select and evaluate gene/metabolite-specific biomarkers for depression and to construct a joint biomarker using information from several genes/metabolites simultaneously. Three modeling approaches have been applied to select and evaluate genes/metabolites that are strongly related to depression as measured by HAMD score. The first two approaches involved measuring the linear association between the pre/post treatment HAMD difference with the pre/post treatment gene/metabolite difference through the use of a joint model and the information-theoretic approach. The two approaches yielded similar results, in agreement with theoretical expectation. But, given the number of potential biomarkers available, which amounts to the number of models that need to be fitted, it seems reasonable to consider the information-theoretic approach that has less computation time as opposed to the approach based on a bivariate model. Furthermore, in the information-theoretic approach, it is possible to distinguish between genes/metabolite with positive and negative association with the response, directly from the model. The other added advantage of the information-theoretic approach is that it can readily be applied to non-normal settings, such as binary and time-

to-event. The third approach, which took a multivariate look in to the data, aims at quantifying a general association between some linear combinations of pre/post HAMD score and the pre/post gene/metabolite, based on the concept of canonical correlation. The coefficients corresponding to the pre/post HAMD score and the gene/metabolite expressions can be used to calculate the linear combinations that could result in the maximum correlation between the gene/metabolite expressions and the HAMD score. Note however that, there exists a possibility of selecting different sets of genes/metabolites as top k gene/metabolites by the first two methods and the multivariate approach. For example, if we consider the genes, we see that only four genes were commonly selected by the three methods. For this particular setting, the difference between the first two approaches and the approach based on the multivariate model which uses R^2_Λ, is that, in the case of the first two approaches, we relate the changes from baseline post-pre gene/metabolites to post-pre HAMD vales and pick those genes/metabolites which give a higher degree of association. Whereas, in later case we look for genes/metabolites where some linear combination of pre and post gene/metabolites values is related to a linear combination of the pre and post HAMD values, and provide the linear combination that maximizes the association. Thus, in similar settings where pre/post measures are taken, the choice between the methods is based on the research question of interest. If interest focuses on finding genes that show an increase or decrease in expression level in relation to an increase/decrease in HAMD score value, then the information-theoretic approach is advisable. However, if the interest is to find any general association between the pre/post HAMD and gene/metabolite levels, then the multivariate model will be appropriate.

In addition to selecting gene/metabolite specific biomarkers, we have attempted to construct a joint biomarker through the use of supervised principal components

analysis and supervised partial least squares regression. The supervised principal component involves two stages. in the first stage, genes/ metabolites exhibiting strong correlation with the HAMD score are selected. The second stage involves creating a first principal component of the top genes/metabolites. Within the framework of supervised principal components approach, we followed two additional alternative approaches. In the first alternative, instead of taking the top k genes as they are, we have created a joint biomarker consisting of a subset of the top 20 genes/metabolites. In this alternative approach, a gene/metabolite will be included as component of the joint biomarker only if its inclusion results in increase in the magnitude of the association of the joint biomarker with the HAMD score. This has resulted in an increase in the measure of association between the joint biomarker and the HAMD score with only 6 genes and 9 metabolites. The second alternative starts with the top 20 genes/metabolites from which a first principal component is constructed. Then, subsets of genes/metabolites are selected to construct the joint biomarker based on the absolute value of their loadings. Although the joint biomarker has given an improved measure of association, a permutation-based test has revealed that the observed measure of association between the joint gene biomarker involving any number of the top 20 genes is not statistically significant. This prompts carrying out an appropriate test of significance of the association measures, be it on a gene/metabolite specific biomarker or on the joint biomarker that involves the top k genes. In a similar manner, we used the supervised partial least squares approach to construct a joint biomarker. First, the top k genes/metabolites were used to construct a factor that has the potential of predicting the HAMD score. But later it was observed that the amount of explained variation in the HAMD score will be higher if only the genes/metabolites with positive weights were used. Another alternative approach was

also entertained. Here, first all the top k genes/metabolites are used in the partial least squares regression and the absolute values of their weights are used to re-rank the genes/metabolites. Then, the genes/metabolites with higher weights are included into the joint biomarker, starting with the top gene until the explained variation starts to decline. A permutation based test was performed to asses the significance of the the association measure based on the last approach. The result has revealed that joint biomarkers involving the top 4 genes and any number of top metabolites were statistically significant.

The comparison between the SPCA and SPLS approaches reveals that, when the reduced matrix is formulated based solely on the univariate association of the individual genes/metabolites, the SPCA approach provides better prediction of the response, although the significance of the inflated association measure is questionable. If instead of taking the order of the genes/metabolites based on their individual association as it is, and rather re-rank them according to their influence (loading) of the joint biomarker, then the SPLS approach results in a large association measure consistently for any number of top genes/metabolites considered in the construction of the joint biomarker. It is also worth noting that different sets of genes/metabolites were deemed important in the construction of the joint biomarker with the two approaches. This however is expected since in the case of the PLS approach, the genes/metabolites are selected such that the correlation between the joint biomarker and the response is maximized. For the SPCA analysis on the other hand, the genes/metabolites are selected based on their contribution to the principal component. But because, individually, the genes/metabolites are correlated with the response, they are expected to have a better association jointly which is manifested in the magnitude of the association measure of the joint biomarker. Thus, in similar situations, if one opts for the SPCA

approach, it is advisable to use some reasonable number of top genes/metabolites as selected by their individual association with the response in the construction of the joint biomarker. However, proper testing of significance for the resulting measure should be carried out. The SPLS approach seems to have better performance when the reduced matrix is reformulated based on the weight of each gene on the latent construct rather than taking the top genes/metabolites as selected by their individual correlation with the response. To avoid the dimensionality problem, as well as to circumvent inclusion of noisy genes/metabolites which might affect the prediction of the response, the procedure can be initiated by first selecting a reasonable number of genes/metabolites based on their individual association but perform further selection of subset of genes/metabolites based on their relative importance in the resulting joint biomarker.

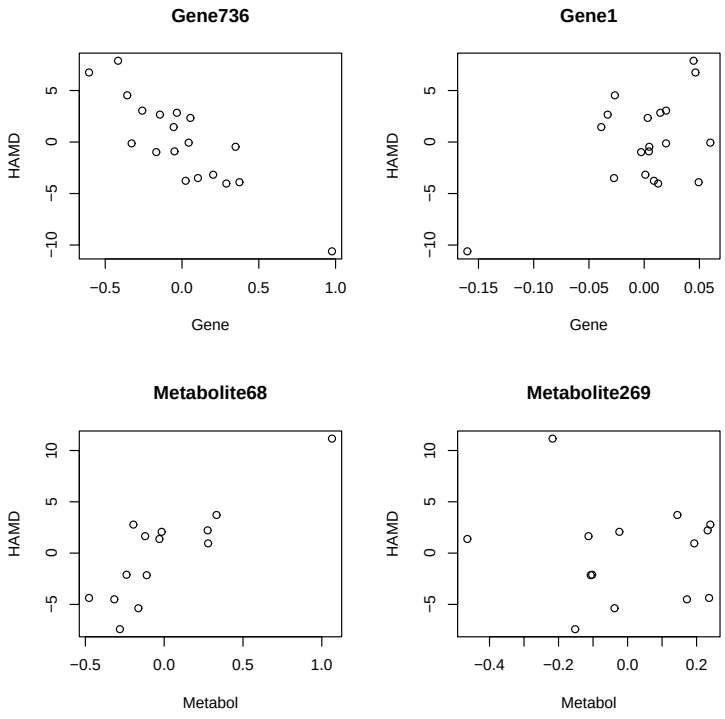

Figure 11.1: *Genes and metabolites with relatively strong association with change from baseline HAMD score (left) and weak association (right) after correcting for covariates.*

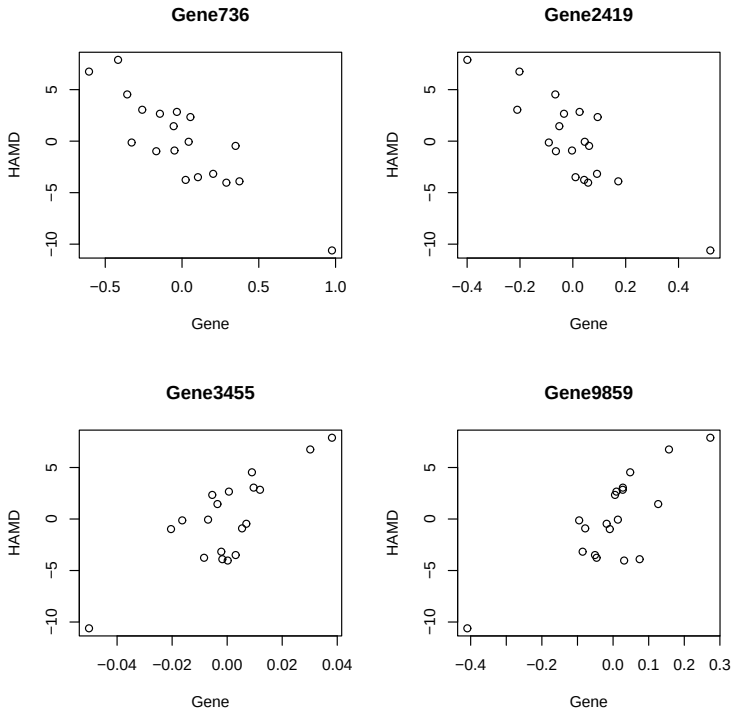

Figure 11.2: *Top four genes based on the informaion-theory approach.*

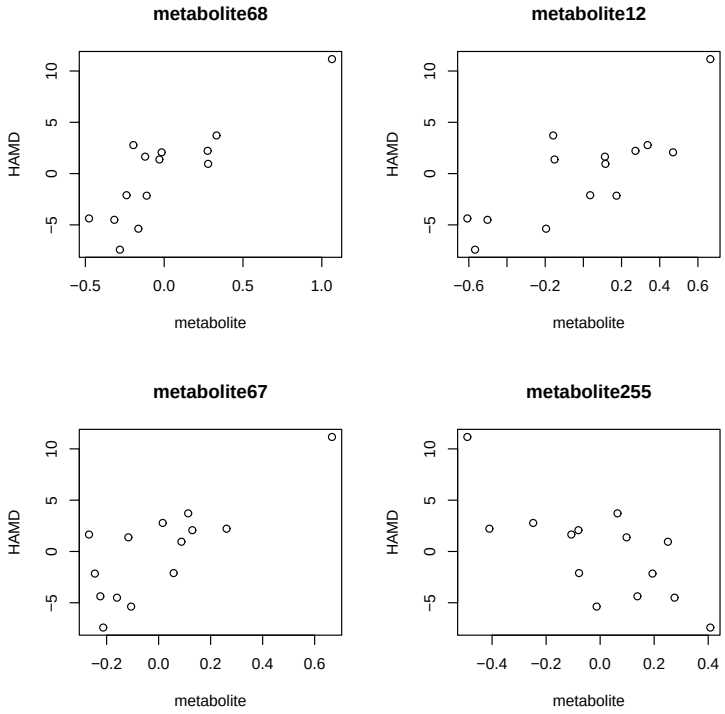

Figure 11.3: *Top four metabolites based on ITA.*

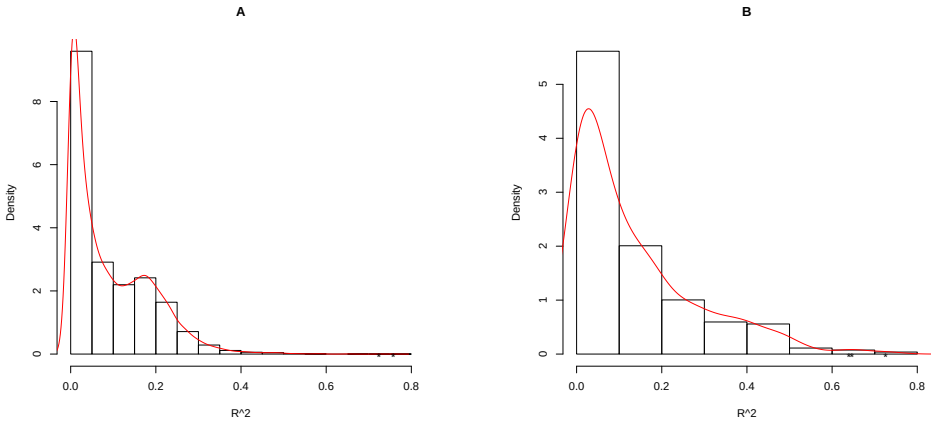

Figure 11.4: *Panel A: distribution of R^2 values based on the ITA approach for Gene expression. Panel B: distribution of R^2 values based on the ITA approach for metabolite expression.*

Table 11.1: *Results for top 20 genes.* R^2*: Association measure based on the information-theory approach, and adjusted association;* R^2_{cr}*:* R^2 *with leave-one-out cross validation;* raw_p*: Raw p-values;* adj_p*: adjusted p-values*

Gene Id	R^2	R^2_{hcr}	raw_p	adj_p
736	0.7579	0.7541	< 0.0001	0.0365
2419	0.7295	0.7243	< 0.0001	0.0426
3455	0.6536	0.6477	< 0.0001	0.1553
9859	0.6507	0.6460	< 0.0001	0.1553
8427	0.5910	0.5906	0.0001	0.3142
1954	0.5881	0.5829	0.0001	0.3142
13988	0.5799	0.5782	0.0002	0.3142
6342	0.5786	0.5728	0.0002	0.3142
6119	0.5771	0.5723	0.0002	0.3142
16073	0.5632	0.5575	0.0002	0.3142
16501	0.5447	0.5380	0.0003	0.3142
16415	0.5394	0.5345	0.0003	0.3142
5543	0.5381	0.5328	0.0003	0.3142
14657	0.5376	0.5355	0.0003	0.3142
9635	0.5276	0.5247	0.0004	0.3142
6195	0.5236	0.5187	0.0004	0.3142
4900	0.5194	0.5193	0.0005	0.3142
12791	0.5178	0.5126	0.0005	0.3142
15294	0.5146	0.5123	0.0005	0.3142
4375	0.5090	0.5018	0.0006	0.3142

Table 11.2: *Results for top 20 metabolites. R^2: Association measure based on the information-theory approach, and adjusted association; R^2_{cr}: R^2 with leave one out cross validation; raw_p: Raw p-values; adj_p: adjusted p-values*

Metabolite Id	R^2	R^2_{cr}	raw_p	adj_p
68	0.7256	0.7164	0.0001	0.0317
12	0.6466	0.6446	0.0005	0.0489
67	0.6400	0.6306	0.0005	0.0489
255	0.5516	0.5493	0.0020	0.1302
21	0.5312	0.5333	0.0026	0.1302
153	0.5011	0.5011	0.0038	0.1302
253	0.4870	0.4872	0.0045	0.1302
194	0.4823	0.4774	0.0048	0.1302
87	0.4782	0.4731	0.0050	0.1302
11	0.4773	0.4724	0.0050	0.1302
130	0.4728	0.4771	0.0053	0.1302
46	0.4606	0.4545	0.0061	0.1341
84	0.4560	0.4481	0.0065	0.1341
144	0.4258	0.4258	0.0091	0.1535
25	0.4195	0.4199	0.0098	0.1535
172	0.4191	0.4180	0.0098	0.1535
139	0.4184	0.4174	0.0099	0.1535
259	0.4102	0.4111	0.0109	0.1535
258	0.4037	0.4080	0.0117	0.1535
262	0.4015	0.4007	0.0119	0.1535

Table 11.3: *Results for top 20 genes based on R^2_Λ. R^2_Λ $R^2_{\Lambda cr}$: The measure of association with and without cross validation; $Hcof_0, Hcof_1$: The coefficients for pre and post treatment HAMD score; $Gcof_0, Gcof_1$: The coefficients for pre and post treatment gene expressions; raw_p: Raw p-values; adj_p: adjusted p-values.*

Gene Id	R^2_Λ	$R^2_{\Lambda cr}$	Hcof_0	Hcof_1	Gcof_0	Gcof_1	raw_p	adj_p
12161	0.9177	0.9075	18.081	-4.4469	-0.1393	0.4462	0.00001	0.2478
9806	0.8871	0.8877	-2.0437	63.2346	-0.06813	0.3694	0.00007	0.2809
4877	0.8862	0.8833	24.782	133.77	0.0015	0.2711	0.00008	0.2809
5324	0.8846	0.8778	2.8306	4.2187	0.0886	-0.3943	0.00008	0.2809
13456	0.8832	0.8682	-3.5597	-50.955	-0.2276	0.4817	0.00009	0.2809
11687	0.8831	0.8706	-28.479	31.169	0.2176	-0.4844	0.00009	0.2809
4078	0.8810	0.8744	5.308	-0.5172	0.2326	-0.2423	0.00011	0.2974
4796	0.8780	0.8733	8.9318	-0.8017	-0.2501	0.4551	0.00014	0.3098
8845	0.8645	0.8551	246.04	13.147	0.2569	-0.4242	0.00026	0.5177
5329	0.8564	0.8560	2.8713	1.3125	0.2318	-0.2392	0.00038	0.5783
3150	0.8551	0.8468	-5.2653	6.0060	-0.0797	0.3837	0.00040	0.5783
736	0.8543	0.8526	-4.1146	2.7613	-0.1902	0.1098	0.00041	0.5783
16964	0.8519	0.8507	12.860	-23.366	-0.1570	0.4605	0.00045	0.5783
16073	0.8482	0.8503	0.1580	1.5688	-0.1217	-0.0525	0.00048	0.5783
9810	0.8477	0.8439	49.596	-17.088	0.0014	0.2713	0.00049	0.5783
8182	0.8452	0.8391	-407.53	494.72	-0.0661	0.3669	0.00050	0.5923
2619	0.8392	0.8404	12.947	226.52	-0.0014	0.2758	0.00060	0.5923
16415	0.8361	0.8376	0.8542	-2.4436	0.0604	0.1723	0.00071	0.5923
9859	0.8336	0.8379	7.0408	-1.8856	-0.2414	0.2774	0.00078	0.5923
8369	0.8318	0.8206	-94.226	43.135	-0.0524	0.3491	0.00080	0.5923

Table 11.4: *Results for top 20 metabolites based on R^2_Λ. R^2_Λ $R^2_{\Lambda cr}$: The measure of association with and without cross validation; $Hcof_0, Hcof_1$, $Mcof_0, Mcof_1$: The coefficients for pre and post treatment HAMD score and for pre and post treatment metabolite expressions respectively; raw_p: Raw p-values; adj_p: adjusted p-values.*

Metabolite Id	R^2_Λ	$R^2_{\Lambda cr}$	Hcof_0	Hcof_1	Mcof_0	Mcof_1	raw_p	adj_p
68	0.9364	0.9308	-2.5773	2.5960	0.1988	-0.3786	0.0032	0.4116
168	0.8484	0.8526	-2.6417	3.2986	-0.1023	0.4584	0.0204	0.7672
158	0.8468	0.8467	-1.0414	3.2146	0.2119	-0.2378	0.0207	0.7672
115	0.8371	0.8464	1.9940	-0.96005	0.1821	-0.0300	0.0244	0.7672
150	0.8080	0.8073	10.444	-5.8443	-0.2912	0.5654	0.0367	0.7672
263	0.8028	0.8114	-0.5355	4.2755	0.2104	-0.3109	0.0040	0.7672
84	0.7727	0.7765	1.0859	-4.4460	-0.1963	0.3867	0.0604	0.7672
239	0.7647	0.7610	-0.4305	1.8738	0.1676	-0.4396	0.0658	0.7672
46	0.7622	0.7873	2.2478	-2.0604	-0.1810	0.4215	0.0675	0.7672
21	0.7605	0.7714	4.8186	-5.8448	0.1583	0.0570	0.0695	0.7672
152	0.7527	0.7520	1.1737	1.1501	-0.1780	0.4264	0.0769	0.7672
67	0.7516	0.7437	-2.7632	5.1035	0.2104	-0.3100	0.0783	0.7672
222	0.7463	0.7486	8.2472	-4.4350	-0.1404	0.4577	0.0849	0.7672
24	0.7461	0.7568	0.4841	1.8812	0.0859	0.2404	0.0852	0.7672
200	0.7181	0.7245	4.2544	0.2299	0.1676	-0.3497	0.1128	0.8784
11	0.7154	0.7294	4.2544	0.2299	0.1676	-0.3497	0.1155	0.8784
73	0.6654	0.7281	0.1145	2.8322	-0.1912	0.4009	0.1185	0.8784
171	0.6971	0.7066	2.0562	-0.3953	0.1188	0.1673	0.1399	0.8784
12	0.6914	0.6989	2.0562	-0.395	0.11887	0.1673	0.1480	0.8784
69	0.6842	0.6938	2.3975	0.1941	-0.12853	0.4604	0.1579	0.8784

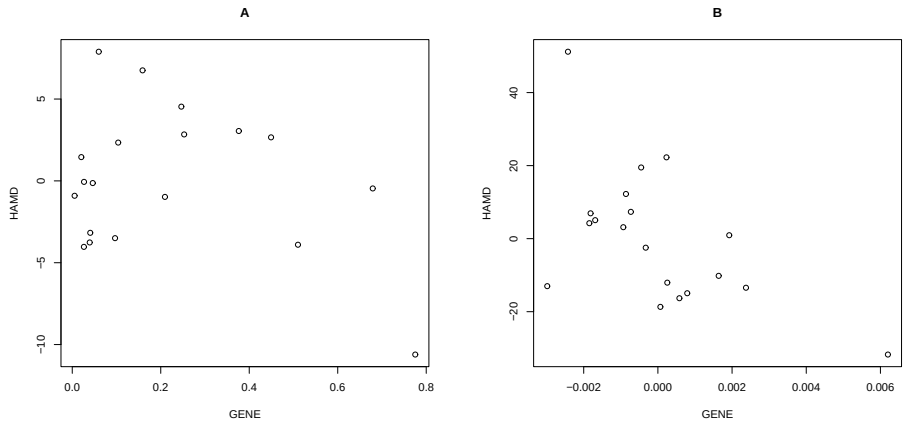

Figure 11.5: *Panel A: Plot of the change from baseline HAMD score versus change from baseline gene expression. Panel B: Plot of optimal linear combination of pre/post HAMD score versus pre/post gene expression for gene 12161.*

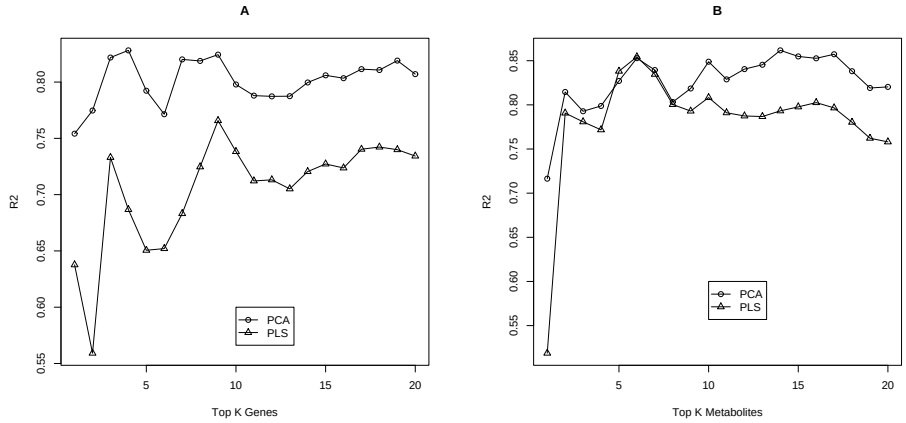

Figure 11.6: *Panel A: Plot of the R^2 measure with SPCA and SPLS, based on leave-one-out cross validation for top genes selected based on R_h^2. Panel B: Plot of the R^2 measure with SPCA and SPLS, based on leave-one-out cross validation for top metabolites selected based on R_h^2.*

Table 11.5: *Results of supervised principal components based on top 20 genes and metabolites selected based on R_h^2. R^2, R_{cr}^2: The measure of association without and with leave one out cross validation; and p-values.*

	Genes			Metabolites		
Top	R^2	R_{cr}^2	p-value	R^2	R_{cr}^2	p-value
2	0.7791	0.7747	0.2280	0.8229	0.8146	0.0000
3	0.8253	0.8218	0.3270	0.8029	0.7927	0.0000
4	0.8301	0.8281	0.2980	0.8072	0.7987	0.0000
5	0.7917	0.7923	0.3430	0.8342	0.8270	0.0003
6	0.7734	0.7714	0.4260	0.8603	0.8529	0.0002
8	0.8210	0.8187	0.1070	0.8118	0.8032	0.0001
10	0.7977	0.7978	0.7330	0.8630	0.8488	0.0001
15	0.8084	0.8059	0.8510	0.8819	0.8548	0.0001
20	0.8090	0.8070	0.7940	0.8452	0.8203	0.0001

Table 11.6: *Results of supervised principal components based on top k genes and metabolites selected based on weights on PCA. R^2, R_{cr}^2: The measure of association without and with leave-one-out cross validation; and p-values.*

	Genes			Metabolites		
Top	R^2	R_{cr}^2	p-value	R^2	R_{cr}^2	p-value
2	0.6597	0.6335	0.3010	0.7168	0.5736	0.0030
3	0.6733	0.6852	0.6060	0.6947	0.6145	0.0030
4	0.7162	0.7321	0.6220	0.7845	0.6382	0.0020
5	0.7006	0.7601	0.6630	0.7205	0.6525	0.0000
6	0.7229	0.7733	0.6680	0.7094	0.6667	0.0000
8	0.7544	0.7864	0.6930	0.6673	0.6830	0.0000
10	0.7586	0.8005	0.7100	0.7250	0.6930	0.0000
15	0.8183	0.8209	0.7950	0.7808	0.7168	0.0000
20	0.8419	0.8372	0.8090	0.8659	0.7540	0.0000

Table 11.7: *Results of supervised partial least squares based on top k genes and metabolites selected based on R_h^2. R^2, R_{cr}^2: The measure of association without and with leave-one-out cross validation; and p-values.*

	Genes			Metabolites		
Top	R^2	R_{cr}^2	p-value	R^2	R_{cr}^2	p-value
2	0.5561	0.5591	0.4480	0.7915	0.7907	0.0000
3	0.7292	0.7330	0.3820	0.7805	0.7808	0.0000
4	0.6797	0.6867	0.5170	0.7715	0.7717	0.0000
5	0.6405	0.6505	0.8380	0.8377	0.8381	0.0000
6	0.6442	0.6520	0.9350	0.8534	0.8544	0.0000
8	0.71946	0.7246	0.8420	0.7986	0.8004	0.0000
10	0.7329	0.7382	0.7330	0.8056	0.8083	0.0000
15	0.7245	0.7272	0.9340	0.7933	0.7977	0.0000
20	0.7314	0.7342	0.9400	0.7521	0.7581	0.0001

Table 11.8: *Results of supervised partial least squares based on top k genes and metabolites selected based on weights on PLS. R^2, R_{cr}^2: The measure of association without and with leave-one-out cross validation; and p-values.*

	Genes			Metabolites		
Top	R^2	R_{cr}^2	p-value	R^2	R_{cr}^2	p-value
2	0.8029	0.7996	0.0370	0.7281	0.7341	0.0000
3	0.8337	0.8297	0.0810	0.7878	0.7783	0.0000
4	0.8779	0.8464	0.0630	0.8607	0.8168	0.0000
5	0.8380	0.8552	0.2040	0.8377	0.8273	0.0000
6	0.7939	0.8603	0.4810	0.8534	0.8479	0.0000
8	0.7696	0.8494	0.7300	0.8168	0.8635	0.0010
10	0.7633	0.8393	0.8360	0.7868	0.8608	0.0020
15	0.7584	0.8406	0.8680	0.7588	0.8554	0.0030
20	0.7314	0.8504	0.9370	0.7521	0.8657	0.0040

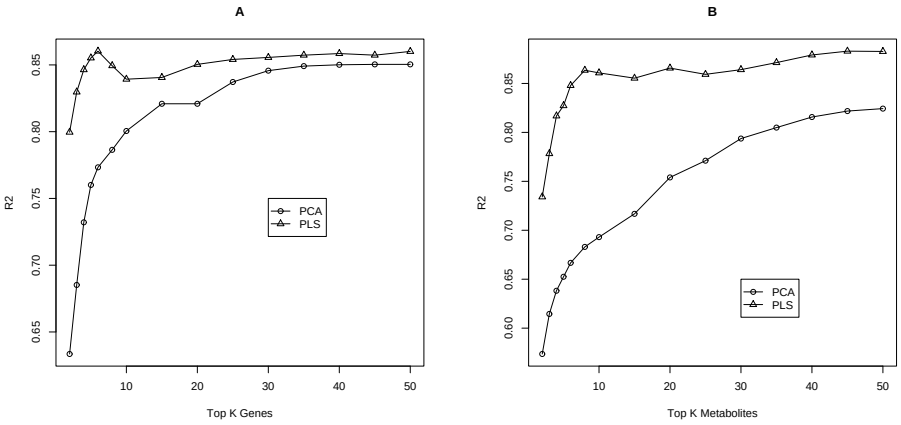

Figure 11.7: *Panel A: Plot of the R^2 measure with SPCA and SPLS, based on leave-one-out cross validation for top genes selected based on weights. Panel B: Plot of the R^2 measure with SPCA and SPLS, based on leave-one-out cross validation for top metabolites selected based on weights.*

12

Alternative Methods For The Selection of Prognostic Biomarkers

The selection and evaluation of genes as prognostic biomarkers requires quantifying the degree of association between the response of interest and the gene expression after correcting for treatment and other possible confounding factors. The associations between the gene expression and the response could be of linear or nonlinear type. If we can possibly assume that there is a linear relationship between the gene expression and the response after accounting for a set of confounding variables, we can use two of the widely used measures of association suggested in the surrogate marker literature namely the adjusted association and the likelihood reduction factor. The same methods have been applied in the previous chapter for selecting genomic biomarkers. The two methods involve either fitting a bivariate model and quantify the association

through individual level R^2 or use an equivalent conditional model which has its basis in the information-theoretic approach and use the likelihood reduction factor. The two methods perform rather well for genes which exhibit a linear association with the response. Practice has however thought us that there are other possible types of relationships with some responses. In this chapter, we will outline some methods that could be used to quantify association with the response without a need to specify functional relationships and revisit the information-theoretic approach with some flexible function to capture possible non-linear relationships with the response.

12.1 Information-theoretic Approach with Penalized Smoothing Splines

Recall that the information-theoretic approach is based on the fit of two univariate models. With in the context of microarray setting, the first model relates the expected value of the clinical outcome to the treatment and/or other confounding variables only, and the second relates the expected value of the clinical outcome to the gene expression as well. So far, we have considered the case where the gene expression enters the model as a covariate in a linear fashion. However, we can incorporate a flexible function to deal with a possible non-linear relationship between the gene expression and the response. One such function is penalized smoothing spline which is discussed in chapter 8 for longitudinal data. Here we provide a brief description of the model for this particular situation. Let T_j denote the response taken from subject j and S_{ij} be the gene expression for the ith gene of subject j . Then the model which relates the response and the gene expression takes the form: $E(T_j|S_{ij}) = \boldsymbol{Z}_j\boldsymbol{\beta}+f(S_{ij})$, for a smooth function $f(\cdot)$. The penalized-spline representation of the model can then

be written as:

$$E(T_j|S_{ij}) = \boldsymbol{Z}_j\boldsymbol{\beta} + \sum_{q=1}^{Q} b_q(S_{ij} - \kappa_q)_+, \tag{12.1}$$

where $\kappa_1, \ldots, \kappa_Q$ are a set of distinct knots in the range of S_{ij}, $S_+ = \max(0, S)$, and $b_q \sim N(0, \sigma_b^2)$. The knot points are selected as equally spaced quantiles of S_{ij} (Ruppert *et al.*, 2003). Similar to the original information-theoretic approach, the measure of association can be quantified by comparing this model with a model that relates the response to treatment and other confounders only. Note however that, if there is no treatment effect or other confounding variables to account for, the two models would be a model relating the response to the gene expression and the second one an intercept model for the response.

12.2 Nonlinear Correlation Coefficient

Here we give a concise description of the nonlinear correlation coefficient (NCC) measure suggested by Wang *et al* (2005). The authors have tried to demonstrate that the mutual information carried by the rank sequences, which are obtained from the original sequences, is a good measure of nonlinear correlation. They later have developed the measure as a concept called nonlinear correlation coefficient. Given two discreet random variables S and T, for describing the general correlation between two variables except for the correlation coefficient which is used to describe the linear correlation of the two variables, the mutual information concept is used widely which is defined as:

$$I(S;T) = H(S) + H(T) - H(S,T), \tag{12.2}$$

$H(S)$ is the information entropy of the variable S, which is defined as:

$$H(S) = -\sum_{j=1}^{L} p_j \; lnp_j, \tag{12.3}$$

and the joint entropy of the two variables S and T, $H(S,T)$, is defined as

$$H(S,T) = -\sum_{j=1}^{L}\sum_{j=1}^{M} p_j \; lnp_j. \tag{12.4}$$

Wang *et al* (2005) state that mutual information can be thought of as a generalized correlation analogous to the linear correlation coefficient, but sensitive to any relationship, not just linear dependence. But it can be seen from the definition of the mutual information that it does not range in a definite closed interval as the squared correlation coefficient does, which ranges in $[0,1]$ with 0 indicating the minimum linear correlation and 1 indicating the maximum. They have given a revised version of the mutual information, which will be sensitive to the general correlation of two variables as the mutual information does, while ranges within a closed interval $[0,1]$. Considering two discrete variables $S = \{s_j\}1 \leq j \leq n$ and $T = \{t_j\}1 \leq j \leq n$, they are first resorted in ascending order and placed into b ranks with first n/b samples in the first rank, the second n/b samples in the second rank, and so on. Second, the sample pairs, $\{s_j, t_j\}1 \leq j \leq n$, are placed into a $b \times b$ rank grids by comparing the sample pairs to the rank sequences of S and T. The revised joint entropy of the two variables S and T is defined as

$$H^r(S,T) = -\sum_{j=1}^{n}\sum_{k=1}^{L} \frac{nij}{n} \log_b \frac{njk}{n}, \tag{12.5}$$

where njk is the number of samples distributed in the jk^{th} rank grid. And the nonlinear correlation coefficient is defined as

$$NCC(S;T) = H^r(S) + H^r(T) - H^r(S,T), \tag{12.6}$$

where $H^r(S)$ is the revised entropy of the variable S, which is defined as

$$H^r(S) = -\sum_{i=1}^{L} \frac{ni}{N} \log_b \frac{ni}{N}. \tag{12.7}$$

Notice that the number of samples distributed into each rank of S and T is invariant, and the total number of sample pairs is N, so the nonlinear correlation coefficient, (12.6), can be rewritten as

$$NCC(S,T) = 2 + \sum_{i=1}^{b^2} \frac{nij}{N} \log_b \frac{nij}{N}. \tag{12.8}$$

The nonlinear correlation coefficient not only is sensitive to the nonlinear correlation of two variables, but can also describe this relationship with a number that ranges within the closed interval $[0, 1]$. In the maximum correlation condition, sample sequences of the two variables are exactly the same, i.e. $s_j = t_j (j = 1, 2, \ldots, N)$ Wang *et al*(2005).

12.3 Regression Tree Analysis

The regression tree methodology (RTA) is a very well-known and widely used technique for different applications. In contrast to classical regression techniques, for which the relationship between the response and predictors is pre-specified, such as linear or quadratic, and the test is performed to confirm or reject the relationship, regression tree analysis assumes no such relationship Breiman *et al* (1984). It is primarily a method for constructing a set of decision rules on the predictor variables. The rules are constructed by recursively partitioning the data into successively smaller groups with binary splits based on a single predictor variable. Splits for all of the predictors are examined by an exhaustive search procedure and the best split is chosen. For regression trees, the selected split is the one that maximizes the homogeneity of

the two resulting groups with respect to the response variable, the split that maximizes the between-group sum of squares, as in analysis of variance, although other options may be available. The output is a tree diagram with branches determined by the splitting rules and a series of terminal nodes that contain the mean response. The procedure initially grows maximal trees and then uses techniques such as cross-validation to prune the overfitted tree to an optimal size Therneau and Atkinson (1997). We choose as the right- sized" tree the smallest-sized, i.e., least complex, tree of which the cross-validation costs do not differ appreciably from the minimum cross-validation costs Breiman *et al* (1984). In particular, some authors proposed a 1-SE rule" for making this selection, i.e., choose as the right-sized" tree the smallest-sized tree whose cross-validation costs do not exceed the minimum cross-validation costs plus one the standard error of the cross-validation costs for the minimum cross-validation costs tree. RTA has clear advantages over classical statistical methods in that it is effective in uncovering structure in data with hierarchical or non-additive variables. Because no prior assumptions are made about the nature of the relationships among the response and predictor variables, RTA allows for the possibility of interactions and non-linearity among variables.

It should be emphasized here that our main objective is to measure the association between the gene expression and the response after accounting for treatment and other confounding variables. Now collecting the residuals S_{ij} and T_j for the gene expression of the ith gene of jth subject and the response of jth subject respectively from their joint model, the association measure that will be employed with regression tree analysis takes the ideas proposed by Alonso and Molenberghs (2007) into account and, for this particular model (for the final tree), can be written as

$$RD_{treei} = \frac{D(T) - D(T \mid S_i)}{D(T)}, \tag{12.9}$$

where

$$D(T) = \sum_{j}^{n} (T_j - \overline{T}), \tag{12.10}$$

is the deviance. Furthermore, $D(T \mid S_i)$ denotes the deviance of the final pruned tree when the information of the gene expression are accounted for. Assuming that we have v terminal nodes $(M1; M2; \quad Mv)$, $D(T \mid S_i)$ can be calculated as:

$$D(T \mid S_i) = \sum_{h=1}^{v} \left(\sum_{T_j \varepsilon M_m} (T_j - \overline{T_{M_h}})^2 \right), \tag{12.11}$$

where $\overline{T_{M_h}}$ is the mean in terminal node M_h.

12.4 Bagging Regression Trees

Bagging is a technique that can be used with many regression methods so as to reduce the variance associated with prediction, thereby improving the prediction process. The main idea behind bagging is as follows: many bootstrap samples are drawn from the available data, some prediction method is applied to each bootstrap sample, and then the results are combined, by averaging for regression, to obtain the overall prediction, with the variance being reduced due to the averaging. It can be used to improve both the stability and predictive power of regression trees, but its use is not restricted to improving tree-based predictions. Rather, it is a general technique that can be applied in a wide variety of settings to improve predictions. Bagging avoids overfitting by randomizing the input of deterministic learning algorithms in the hope that directions where overfitting occurs for individual predictions cancel out. Whatever overfitting there might be is averaged out when the combining takes place. The main characteristic of bagging that prevents it from being affected by overtraining

is that, in bagging, the training data set is modified randomly and independently at each step. The bootstrap resampling used in bagging is known as a robust technique. Therefore, in expectation, the distribution of a bootstrap sample, which consists of 63,2% of training data, becomes similar to the real data distribution. In this way, bagging is prevented from overtraining. The association measure will then be the median of the list of relative reduction in deviance RD_{tree} of each tree constructed for each bootstrap sample.

12.5 Random Forests

A random forest (RF) is an ensemble of many identically distributed trees generated from bootstrap samples of the original data Breiman (2001). Each tree is constructed via a regression tree algorithm. The simplest random forest with random features is formed by selecting randomly, at each node, a small group of input variables to split on. The size of the group is fixed throughout the process of growing the forest. Each tree is grown by using the RTA methodology without pruning. Some features of random forest worth highlighting are: (1) it is an excellent classifier, comparable in accuracy to support vector machines; (2) it generates an internal unbiased estimate of the generalization error as the forest building progresses; (3) it computes proximities between pairs of cases that can be used in clustering, locating outliers, or by scaling, giving useful views of the data; (4) it is well known that random forests avoid overfitting and it has been demonstrated to have excellent performance in comparison to other machine learning algorithms Cortiñas *et al* (2009). The measure of association will be computed similarly to the case in which bagging methods were used. For the Random Forest cross-validation methods may not be needed since each tree is grown from a bootstrapped sample, on average, about one-third of the observations in the

data set will not be used to grow the tree.

12.6 Support Vector Machine

The term support vector machines (SVM) refers to a family of learning algorithms which is considered as one of the most efficient methods throughout a variety of applications. SVM is a supervised learning technique for classification and regression. SVM can also be applied to regression problems by the introduction of an alternative loss function, (Smola, 1996). The loss function must be modified to include a distance measure. SVM regressions use the ε-insensitive loss function. If the deviation between the predicted and actual values is less than ε, then the regression function is considered good, which can be mathematically expressed as: $-\varepsilon \leq \omega.S_{ij} - b - T_j$. From a geometric point of view, it can be seen as a band of size 2ε around the hypothesis function and any point outside this band is considered as a training error. Suppose the data can be explained by a linear model; the goal is to find a fitting hyperplane $\langle \omega, g_{ij} \rangle + b = 0$. Formally, we need to minimize $\|\omega\|^2/2$, subject to the following constraints:

$$T_j - \langle \omega, S_{ij} \rangle - b \leq \varepsilon, \quad \langle \omega, S_{ij} \rangle - T_j \geq= \varepsilon$$

To account for training errors and the possibility of handling non-linearity, we can map the input data S_{ij} into a, possibly higher-dimensional, so-called feature space $\Phi(S_{ij})$ and introduce some weights to our optimization problem, which now becomes:

$$\mathbf{min} \frac{\|\omega\|^2}{2} + \text{C}. \sum_{i}^{N} (\xi_i + \widehat{\xi}_i),$$

subject to the following constraints:

$$T_j - \langle \omega, \Phi(S_{ij}) \rangle - b \leq \varepsilon + \xi_j,$$
$$\langle \omega, \Phi(S_{ij}) \rangle - t_j \geq= \varepsilon + \xi_j,$$
$$\xi_j, \widehat{\xi_j} \geq 0.$$

We then need to solve a constrained optimization problem. It turns out that, in most cases, it can be solved more easily in its dual formulation. Several kernels can be used such as:

- Polynomial: $(\gamma(\langle S_{ij}, S_{kj} \rangle + \delta)^d$
- Radial basic function (RBF): $exp(\gamma \| S_{ij}, S_{kj} \|^2)$
- Sigmoid: $tanh(\gamma(\langle S_{ij}, S_{kj} \rangle + \delta)$.

One possible way to select a kernel is first to tune all three kernels, using cross-validation, and, finally, the kernel, together with the set of parameters that produce the smallest mean squared error would be retained. In this way, we can control for the risk of overfitting, given that the set of parameters used to obtain the final model are selected using a cross-validation procedure. We can then go on and evaluate the model performance for each of the observations left out in the cross-validated samples and thus the ability of the model to generalize beyond the fitting data. For this application, we have considered the RBF kernel, which can handle the non-linear mapping and have few parameter to be controlled (C between 0.25 and 6, with step of 0.25 and γ between 0.5 and 50 with step of 0.5) Hsu *et al* (2001). The parameters C and γ obtained from the tuning process were then used to estimate the measure of association. For comparison purposes, we have also considered the polynomial kernel. Similar to the case of regression trees, the association measure can be computed using the ratio between the portion of the variability not explained by the model and the total variability of the residuals from the response:

$$RD_{SVMi} = \frac{D(T) - DSVMR(T \mid S_i)}{D(T)}, \tag{12.12}$$

$D(T)$ can be calculated as in (12.10), and $DSVMR(T \mid S_i)$ is the sum of the squares of the differences between the actual value (T_j) and their estimated value obtained when the SVM regression model is employed.

12.7 Application to the Case Study

The methods discussed in the previous sections were applied to the case study in depression. The results are summarized in Tables 12.1- 12.3. The results highlighted that, when there is a noticeable linear relationship, the information-theoretic approach without flexible functional form in the gene expression performs reasonably well. However, when there is some form of nonlinear relationship, this method is less optimal. To augment flexibility to this simple but elegant approach, we have used penalized spline. It is worth mentioning here that, the information-theoretic approach complemented with the splines method still selected the same set of genes that were selected by the linear models. This however is not the failure of the method to deal with non-linear associations but rather can be attributed to the way the knot points were selected. Since there are large number of models to be fitted, the default knot selection was used. With appropriate knot points selected, this method can perform reasonably well for addressing both linear as well as non-linear associations as has been the case for other applications Tilahun *et al* (2007). The rest of the alternative methods have selected different genes with different patterns. The support vector machine approach with radial basis has selected genes with linear as well as nonlinear relationships. The random forest method has given substantially larger values of the association measure compared to the other approaches. The non linear correlation

coefficient of Wang *et al*(2005) has given similar values for the association measure for a large number of genes. This might be attributed to the fact that the method works on the ranks of the genes rather than the actual values. The use of cross-validation is advised as it appeared that the results with and without cross-validation were found to be different.

12.8 Discussion

In this chapter, we have outlined alternative methods for the selection of prognostic genomic biomarkers in line with Cortiñas *et al* (2009) who used the same set of methods to quantify the trial level surrogacy in the context of meta-analytic framework of surrogate marker validation. The main motivation behind the use of these alternative methods in the selection of prognostic biomarkers is the need to deal with other possible types of associations rather than simple linear relationships.

The methods that assume linear relationship between the gene expression and the outcome might come short of selecting genes that exhibit other forms of associations. This could lead to loss of important information which amounts to loss of some important prognostic biomarkers. The comparison of the methods has revealed that the different methods might select different set of genes as potential biomarkers. However some of the methods seem to perform poorly which is reflected by the type of genes they have selected. For example the non-linear correlation coefficient of Wang *et al*(2005) has given similar measures of associations for a large number of genes which might be questionable given that some of the genes do not seem to have any meaningful association with the response. Some other methods, such as the support vector machine with radial basis, on the other hand, picked genes that have portrayed both linear as well as non-linear associations with the response.

Table 12.1: *Results for top 20 genes ITA and NCC*

ITA			Non-linear Correlation		
Gene	*LRF*	LRF_cr	*Gene*	*NCC*	NCC_cr
736	0.7579	0.7541	14771	0.72856	0.58561
2419	0.7295	0.7243	5167	0.71319	0.51359
3455	0.6536	0.6477	4165	0.67247	0.49785
9859	0.6507	0.6460	2891	0.67247	0.51666
8427	0.5910	0.5906	13703	0.65710	0.54923
1954	0.5881	0.5829	13515	0.65710	0.52826
13988	0.5799	0.5782	16929	0.65710	0.52076
6342	0.5786	0.5728	6298	0.65710	0.53576
6119	0.5771	0.5723	10544	0.65710	0.53686
16073	0.5632	0.5575	2391	0.65710	0.55028
16501	0.5447	0.5380	12978	0.65710	0.49359
16415	0.5394	0.5345	11559	0.65710	0.53524
5543	0.5381	0.5328	676	0.65710	0.55969
14657	0.5376	0.5355	5585	0.65710	0.54035
9635	0.5276	0.5247	7654	0.65710	0.56184
6195	0.5236	0.5187	14338	0.65710	0.58147
4900	0.5194	0.5193	6876	0.65710	0.54464
12791	0.5178	0.5126	2514	0.65710	0.53524
15294	0.5146	0.5123	4146	0.65710	0.55807
4375	0.5090	0.5018	13133	0.65710	0.50811

Table 12.2: *Results for top 20 genes Regresion tree Random Forest and Bagging.*

Regresstion Trees			Bagging		Random Forest	
Gene	*RT*	RT_cr	*Gene*	*Bagg*	*Gene*	*RF*
304	0.68016	0.60334	8039	0.61277	7407	0.92480
14338	0.67599	0.57584	6458	0.61232	213	0.91894
4319	0.66621	0.64267	16575	0.60797	11585	0.90820
4739	0.66581	0.46607	13988	0.60713	2696	0.90143
11629	0.66497	0.52607	6228	0.57677	5252	0.89644
6458	0.65962	0.62697	9761	0.57269	15041	0.86346
4618	0.65789	0.57569	5144	0.56375	9447	0.85629
12844	0.65634	0.58129	5363	0.56304	14253	0.84653
16028	0.65504	0.54897	10351	0.56187	6186	0.83971
9761	0.65326	0.61939	11010	0.55900	16886	0.83568
11618	0.65131	0.54502	7489	0.55703	16358	0.83488
10363	0.65109	0.55807	2970	0.55640	11583	0.83402
7829	0.64672	0.50540	736	0.54896	338	0.82453
16575	0.64507	0.62101	6406	0.54806	16135	0.82190
6098	0.64440	0.53311	12484	0.53888	697	0.81716
10401	0.64409	0.50206	44	0.53500	13572	0.81654
5986	0.64351	0.59575	13965	0.53431	9565	0.81394
6692	0.64102	0.60109	307	0.53133	14048	0.80730
1764	0.64060	0.52899	3772	0.52993	16912	0.80180
1025	0.63889	0.58602	3645	0.52628	9995	0.79238

Table 12.3: *Results for top 20 genes SVM with polynomial and Radial Basis*

Polynomial Basis			Radial Basis		
Gene	SVM	SVM_cr	*Gene*	SVM	SVM_cr
6195	0.61628	0.58789	5144	0.73692	0.67544
14157	0.60334	0.58143	736	0.66090	0.65484
4055	0.60155	0.57661	2940	0.65652	0.60360
2903	0.58291	0.56110	10540	0.63615	0.59622
3455	0.57470	0.56472	12483	0.62505	0.59569
6008	0.56447	0.53505	862	0.60827	0.58801
2519	0.55457	0.53672	3455	0.60326	0.57892
16501	0.54994	0.54574	12748	0.60192	0.58665
11096	0.54718	0.52774	11105	0.59653	0.56807
15204	0.54220	0.52994	5215	0.59461	0.57649
6345	0.53957	0.53106	13893	0.59446	0.56181
10455	0.53724	0.51402	13081	0.59082	0.54215
2419	0.53643	0.54212	10351	0.58994	0.57088
6342	0.53045	0.53491	3598	0.58874	0.54205
12023	0.52857	0.50317	7429	0.58796	0.50982
1548	0.52701	0.51101	2419	0.58761	0.58754
1249	0.52617	0.50895	4382	0.58663	0.56366
3290	0.52306	0.50376	12847	0.58637	0.56107
13767	0.52138	0.49973	4126	0.58630	0.56403
6414	0.51725	0.50240	4900	0.58385	0.55803

In a real life application, involving a large number of genes, it might not be feasible to apply these methods all at once. Note however that, the information-theoretic approach with appropriate choice of the knot points and the support vector machine approach with radial basis can handle both linear as well as non-linear associations adequately. Hence these two methods might be suitable candidates for general purposes. The information-theoretic approach with splines can easily be fitted with any software which has the facility to handle linear mixed models. It also takes substantially less computation time which makes it a prime candidate to deal with situations that call for both linear and non-linear associations. The methods suggested in this chapter should be complemented with tests for the significance of the resulting association measures through for example bootstrap methods. However, the bootstrap methods might be time consuming and hence there is a need to devise other methods that might be less computationally intensive which we believe could be an interesting topic for further research. In this regard also, the information-theoretic approach gets the upper hand as there is an asymptotic theory that allows for the construction of asymptotic confidence interval for the estimated measure of association.

13

The Selection and Evaluation of Gene Specific biomarkers: Hierarchical Bayesian Approach

It has to be recalled that depending on the way they are related to the clinical outcome, biomarkers can be classified as prognostic or therapeutic. The selection of prognostic biomarkers can be carried out by using an association measure which quantifies the relationship between the response of interest and the biomarker after adjusting for treatment and other confounding variables. The selection of therapeutic biomarkers on the other hand, requires establishing a relationship between the treatment effects on the clinical outcome and the potential biomarker. In the surrogate marker validation context, the former are referred to as individual level surrogates

while the latter ones refer to what is known as trial level surrogate. The absence of replicates at a trial level in a microarray experiment has prohibited the direct use of some of the methods designed for surrogate marker validation which has led to the use of a Bayesian approach. This approach assumes a bivariate normal distribution for the treatment effects on the potential biomarker and the response of interest from which an R-square type measure can be derived. In this chapter, this approach will be discussed and then will be applied to a case study in behavior. The results are compared with the approach of Lin et.al (2007) which uses the relative reduction in deviance based on regression tree approach. Let us begin with a brief introduction of the method of reduction in relative deviance and later move on to the hierarchical Bayesian modeling.

13.1 Reduction in Relative Deviance

To evaluate the quality of therapeutic biomarkers, Lin et.al (2007) followed the approach of Alonso and Molenberghs (2005) and proposed a measure for therapeutic biomarker, the reduction in relative deviance. The total variability of the response, the deviance, without any information about the gene-expression level can be measured by

$$D(T) = \sum_{j=1}^{n} (T_j - \hat{\mu})^2 , \tag{13.1}$$

where $\hat{\mu} = 1/n \sum_{j=1}^{n} T_j$ and $j = 1, \ldots, n$ indicates the arrays. For a therapeutic biomarker, because gene-expression is differentially expressed, one can use the gene-expression level in order to predict the response level. While a linear regression model is not an appropriate model for this type of a biomarker, a regression tree model (Venables and Ripley 1994), in which the gene-expression is the only predictor, can capture the structure of the data shown in Figure 13.1.

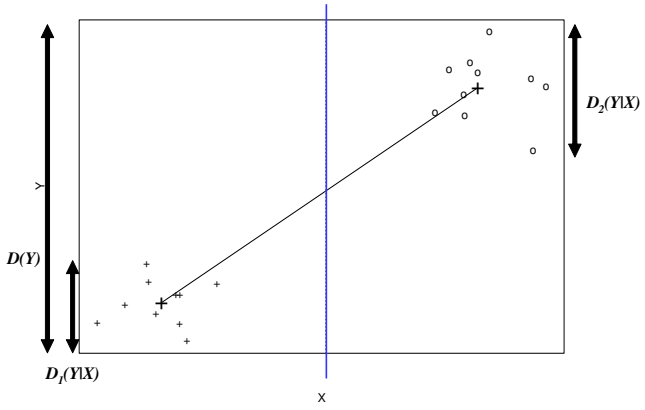

Figure 13.1: *A regression tree model for a hypothetical example with two terminal nodes. The vertical line in the plot indicates the split point in the regression tree.* $D(Y)$ *represents the total variability in the response* Y*, while* $D_1(Y|X)$ *and* $D_2(Y|X)$ *represent the variability within each of the terminal nodes.*

Moreover, because the gene is differentially expressed, we can restrict the tree to two terminal nodes (two final homogenous groups of the response), in which the cutoff point (or the split point) is determined only by the gene-expression level. An example of the cutoff point is shown as the vertical line in Figure 13.1. Let k denote the number of terminal nodes in the tree and let $D(T|S, k = 2)$ denote the sum of deviances for the terminal nodes,

$$\begin{aligned} D(T|S, k = 2) &= D_1(T|S) + D_2(T|S) \\ &= \textstyle\sum_{T_j \in k_1} (T_j - \hat{\mu}_1)^2 + \sum_{T_j \in k_2} (T_j - \hat{\mu}_2)^2 , \end{aligned} \tag{13.2}$$

where $D_1(T|S)$ and $D_2(T|S)$ denote the deviance in each of the terminal nodes, k_1 and k_2 denote the sets of subject indices corresponding to the two terminal nodes,

and $\hat{\mu}_1$ and $\hat{\mu}_2$ are the mean response in the two terminal nodes. The reduction in the deviance, $D(T) - D(T|S, k = 2)$, measures the gain in prediction of the response level using gene-expression, as compared to the case where the gene-expression is not used. In other words, the reduction in deviance measures whether information about the gene-expression is relevant for predicting the response level. The relative deviance reduction, R^2_D, is given by

$$R^2_D = \frac{D(T) - D(T|S)}{D(T)} = \frac{D(T) - D_1(T|S) - D_2(T|S)}{D(T)},$$

hence,

$$R^2_D = \frac{\sum_{j=1}^{n} (T_j - \hat{\mu})^2 - \left[\sum_{T_j \in k_1} (T_j - \hat{\mu}_1)^2 + \sum_{T_j \in k_2} (T_j - \hat{\mu}_2)^2\right]}{\sum_{j=1}^{n} (T_j - \hat{\mu})^2}. \tag{13.3}$$

Following Alonso and Molenberghs' (2005) information theoretic approach, it is easy to see that R^2_D belong to the family of information theoretic association measures. This is a crucial point, as it implies that, although prognostic and therapeutic biomarkers are evaluated using different validity measures both measures can be interpreted in the same way.

13.2 Hierarchical Joint Model for the Gene Expression and the Response

In the relative reduction approach discussed in the previous section, geneomic biomarkers were evaluated according to their quality in predicting the response. In this section we focus on the association between the treatment effects upon the response and the gene expression. In particular we wish to identify genomic biomarkers for which information about the treatment effect upon the biomarker will reduce the uncertainty about the treatment effect upon the outcome of primary interest. In other words we

would like to identify genomic biomarkers for which treatment effect upon biomarker can be used in order to predict the treatment effect upon the response.

Now let us discuss the hierarchical Bayesian joint model for the gene expression and the response, from which both prognostic and therapeutic genes can be tested and evaluated. For this let us consider a single gene, and fit the following bivariate model for the gene expression and the response of interest. Here S_{ij} denotes the gene expression of the i^{th} gene of the j^{th} subject whereas T_j represents the clinical outcome of interest.

$$\begin{pmatrix} S_{ij} \\ T_j \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_S + \boldsymbol{Z}_j\alpha \\ \mu_T + \boldsymbol{Z}_j\beta \end{pmatrix}, \Sigma_i = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ \sigma_{ST} & \sigma_{TT} \end{pmatrix} \right). \tag{13.4}$$

From this model, the selection of prognostic biomarkers can be carried out by using the so called adjusted association which takes the form

$$\rho^2 = \frac{\sigma_{ST}^2}{\sigma_{ST}\sigma_{TT}}. \tag{13.5}$$

Hierarchical Bayesian model allows us to evaluate therapeutic biomarkers by specifying a joint prior distribution , $[\beta, \alpha_i]$, for the treatment effects. Note that both $[T_j, S_{ij}|Z_j]$ and $[\beta, \alpha_i]$ are gene specific which implies that a "gene by gene" analysis is performed. The prior and hyperprior distributions for the parameters in (13.4) will be discussed in the next section. Note that a microarray experiment is equivalent to the single trial setting in the clinical trials framework in the sense that for a gene specific model we have one treatment effect each upon the response and the gene expression. This is in contrast with the multiple trial setting in which one can estimate a trial specific treatment effect for each trial. Using hierarchical Bayesian models, Daniels and Hughes (1997) and Shkedy *et al.* (2005) show that the trial level surrogacy can be evaluated from prior distribution of the treatment effects. In what

follows, we specify the prior distribution for the joint model and show that, similar to the meta analytic approach, a second level of association, which we term "gene level association" can be evaluated from the joint distribution of the treatment effects.

13.2.1 Specification of the Prior Distributions

In order to complete the specification of the hierarchical model we assume independent normal prior to the intercepts, i.e,

$$\begin{aligned} \mu_{S_j} &\sim \mathrm{N}(0, \theta^2_{\mu_{S_j}}), \\ \mu_T &\sim \mathrm{N}(0, \theta^2_{\mu_T}), \end{aligned} \tag{13.6}$$

For the precision parameters in (13.6) flat hyperprior models are specified using Gamma distributions, e.g., $\theta^{-2}_{\mu_S} \sim \mathrm{Gamma}(0.001, 0.001)$, etc. Similar to the model proposed by Daniels and Hughes (1997) and Shkedy *et al.* (2005), we need to specify a prior distribution to model the association between the treatment effects of the two endpoints. We specify a bivariate normal prior distribution

$$\begin{pmatrix} \alpha_j \\ \beta \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, D_j \right), \tag{13.7}$$

with variance-covariance matrix given by

$$D_j = \begin{pmatrix} d_{\alpha_j\alpha_j} & d_{\alpha_j\beta} \\ d_{\alpha_j\beta} & d_{\beta\beta} \end{pmatrix}. \tag{13.8}$$

Within the meta analytic approach, the measure for trial level surrogacy, R^2_{trial} is derive from the covariance matrix D (Buyse *et al.* 2000). We follow this approach and use coefficient of determination, R^2_{gene} , in order to evaluate the association between α_j and β.

$$R^2_{gene} = \frac{d^2_{\alpha_j\beta}}{d_{\alpha_j\alpha_j} d_{\beta\beta}}. \tag{13.9}$$

Indeed, $R^2_{gene} = 1$ indicates a deterministic relationship between the treatment effects while $R^2_{gene} = 0$ indicates that the treatment effects are uncorrelated. Wishart

distributions are assumed as the hyperprior distribution for the variance-covariance matrices in (13.4) and (13.8):

$$D^{-1} \sim \text{Wishart}(R_D), \qquad \Sigma^{-1} \sim \text{Wishart}(R_\Sigma). \tag{13.10}$$

In summary, the gene-level and individual-level associations (used to select prognostic biomarkers) are assessed using the posterior means for the coefficients of determination (13.5) and (13.9), respectively. Note that both (13.5) and (13.9) are gene specific coefficients.

13.3 Model Selection

In order to validate the genes that are selected as therapeutic biomarkers, we can proceed by fitting two models. The first model corresponds to the one discussed earlier which assumes that the treatment effects on the gene expression and the outcome are jointly normally distributed. The second model assumes that the treatment effects are independent which can be formulated by specifying a variance covariance matrix of the form:

$$D_j = \begin{pmatrix} d_{\alpha_j \alpha_j} & 0 \\ 0 & d_{\beta\beta} \end{pmatrix}. \tag{13.11}$$

For each gene, the two models will be fitted and their corresponding DIC values will be compared. Genes whose DIC is smaller for the first model which assumes existence of correlation between the two treatment effects compared to the second model will be considered as potential therapeutic biomarkers.

13.4 Confirmatory Analysis

The model selection approach discussed earlier can be a handy tool to indirectly ascertain whether or not the observed association between the treatment effects α_j and β is statistically significant. In this section, we try to address the same objective using a different approach. From previous experiences and a preliminary analysis of the case study under consideration, we can learn that the only association between the clinical outcome and most of the the gene expressions is treatment induced. Thus a gene expression will be considered a reasonable biomarker if it enables us to answer the question namely "does the gene expression provide information that can be used in order to classify the response into the two treatment groups?". This is equivalent to saying that substantial amount of the information about the treatment effect on the clinical outcome is captured by the gene expression. Now let us consider the following model construction.

$$\begin{pmatrix} X_{ij} \\ Y_i \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_{X_j} + \alpha_i Z_i \\ \mu_Y + \gamma I_i \end{pmatrix}, \Sigma_j \right). \tag{13.12}$$

This model is similar to the previous bivariate model except here I_i, which is an indicator variable determined by the gene expression, such that

$$I_i = \begin{cases} 1, & X_{ik} \leq \theta, \\ 0, & X_{ik} > \theta, \end{cases} \tag{13.13}$$

is used in place of the treatment. The parameter θ is a split point which split the response into two groups. Observations to the left of θ form one group while observations to the right of θ form the second group. We specify a non informative prior for θ

$$\theta \sim \mathrm{U}\left[\min(X_{ik}), \max(X_{ik})\right]. \tag{13.14}$$

Note that models (13.4) and (13.12) both imply that the distribution of the response is a mixture of two distributions. The shift in the mixture in (13.4) is β determined by the treatment group while the shift in (13.12) is γ and it is determined by the gene expression. Hence, the model in (13.12) focuses on the question whether the gene expression cab be used in order to determine the shift of the response. Note that if $Z_i = I_i \;\; i = 1, \ldots, n$ then $\beta = \gamma$ in this case the classification based on the gene expression form the two treatment groups. The procedure will be carried out similar to a leave-one-out approach in a sense that, each time the response of a single subject will be dropped and the model will be fitted with the remaining subjects and then the fitted model will be used to predict the outcome of the subject dropped. For a good biomarker, the predicted values are expected to be closer to the group means of the clinical outcome which manifests the assumptions that the treatment on the clinical outcome is captured by the gene expression.

13.5 Application to the Case Study

The Bayesian approach discussed earlier has been applied to the case study introduced in the Section 2.1.7. The top 20 genes selected based on the association between the treatment effects on the gene expressions with the treatment effects on the response and the top 20 genes selected based on the relative reduction in deviance are displayed in Table 13.1. The results have exhibited that none of the genes qualify to be prognostic biomarkers for the response as the magnitude of the association measures were found to be rather too small. However, the treatment effects on three genes have shown a relatively moderate level of association with the treatment effect on the response. This highlights that there is some hope of using these set of genes as possible therapeutic biomarkers. This is in agreement with what can be seen in

the box plots displayed in Figure 13.2 and 13.3. From these box plots we can see that there is a clear treatment effect for the top three genes and on the response. From Figure 13.5 we can see that for the gene which was selected as a top gene the posterior distribution of the R^2 values is skewed to the left with the majority of the values closer to one. Whilst for the least ranking gene, the distribution is skewed to the right with more small values. The contour plots also exhibit the presence of correlation between the treatment effects on gene 1962 and the response while no apparent pattern is observed for the gene expression of gene 1090. The model selection approach followed has also revealed that the DIC values for the model which assumes correlation between the treatment effects are slightly smaller than the independence model. This gives a guarantee that the assumption might be viable. The confirmatory analysis was performed for the top and least genes namely genes 1962 and 1090. As can be seen from Figure 13.6, for gene 1962, for which the treatment effect have showed a moderate level of association with the treatment effect on the clinical outcome, the predicted values from the model with treatment effect and a model with the gene expressions used in place of the treatment effect, are closer to the mean of each treatment group confirming that the treatment effect on the clinical outcome is captured by the gene expression. For gene 1090, the predicted values based on the treatment effect used as a covariate are closer to the mean of each treatment group while the once predicted with the gene expression used in place of the treatment are clustered to the over all mean inline with expectation. In conclusion, even though there is a relatively moderate association between the treatment effects on the four genes and the outcome, there is very little information left in the gene expressions about the response of interests after adjusting for treatment effect. The comparison of the Bayesian approach and the relative reduction in deviance reveals that, some

of the top genes by both approaches are identical although the rankings differer from one to the other.

13.6 Discussion

In this chapter, we have introduced a Bayesian approach to select prognostic and therapeutic genomic biomarkers. The method can be applied both for a single microarray experiment as well as for the meta-analytic approach. We have tried to establish analogies between the measures of associations defined in the meta-analytic frame work of the surrogate marker validation and the selection and evaluation of biomarkers. The individual level surrogacy, which quantifies the association at the individual patient level can be directly used in the microarray setting to select prognostic biomrkers. The selection of therapeutic biomarkers, using the method designed for trial level surrogacy, however, requires the existence of replications at the trial level which is not a common practice in microarray settings. This problem has motivated the use of the Bayesian approach where the treatment effects from the response of interest and the gene expressions are assumed to follow a bivariate normal distribution. This formulation has enabled us to derive an R-square type measure similar to the adjusted association, which is used to select prognostic biomarkers. From the results we have been able to identify few genes which might be considered as possible therapeutic biomarkers. However, none of the genes qualify to be considered as prognostic biomarkers as reflected by the small magnitude of the association measure relating the response to the gene expression after adjusting for treatment. Some of the top genes selected by the relative reduction in deviance and the Bayesian approach were identical. This however does not necessarily mean that the two methods are equivalent. The reduction in relative deviance quantifies the association between the response

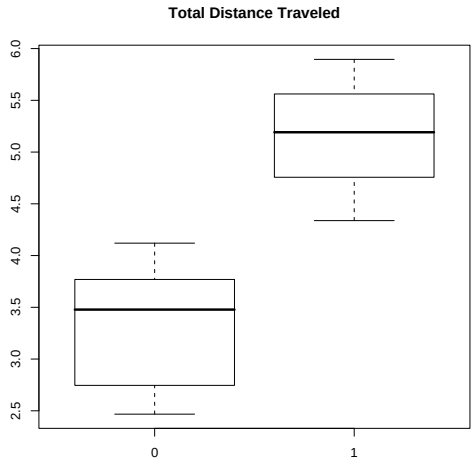

Figure 13.2: *Boxplot of the total distance traveled by the rats.*

and the gene expressions without adjusting for treatment effect under the assumption that the observed association is induced by treatment only. The justification that the genes selected by this approach are therapeutic comes from the very assumption that the observed association is treatment induced and hence the genes contain information about the treatment effect on the response indirectly. The Bayesian approach on the other hand, quantifies the association between the gene expressions and the response after adjusting for the treatment effect. And by further assuming that the treatment effects on the gene expression and the outcome follow a bivariate distribution, it enables us to select genes whose treatment effect gives indication of the treatment effect on the clinical outcome. Note however that, the Bayesian method hinges strongly on the validation of the assumptions made about the the treatment effects and hence care should be taken in making general conclusion about the results.

Table 13.1: *Top 20 genes selected based on R^2_{gene} and RD_{tree}.*

Top Genes Based on Rgene			Top Genes Based on RD		
Gene	R^2_{gene}	ρ^2	*Gene*	*RE*	RD_{tree}
1962	0.6362	0.0564	345	-0.5548	0.7565
60	0.6183	0.0198	1962	-0.4241	0.7565
345	0.6053	0.0545	4447	-2.0463	0.7565
486	0.5460	0.0286	5356	-7.4262	0.7307
59	0.5442	0.0476	486	-0.7625	0.7280
1569	0.2825	0.0058	662	-2.4272	0.6442
5614	0.2728	0.0003	2247	-5.1065	0.6123
4447	0.2658	0.0002	5614	1.8131	0.5769
158	0.2540	0.1033	5216	2.6550	0.5578
2028	0.2521	0.0650	1022	-5.3237	0.5576
214	0.2400	0.0338	214	-2.2181	0.5561
662	0.2172	0.0523	59	-0.7582	0.5548
3899	0.2164	0.0475	60	-0.4819	0.5548
2591	0.2140	0.0015	158	-2.0907	0.4309
4254	0.2114	0.1137	1316	3.0449	0.4181
1263	0.2050	0.1073	522	-5.6379	0.4122
637	0.1964	0.0084	2489	5.8340	0.4099
4320	0.1947	0.0137	3170	-11.6454	0.4088
2697	0.1937	0.0077	4297	3.9599	0.4017
2753	0.1887	0.0005	5352	-6.7483	0.3574

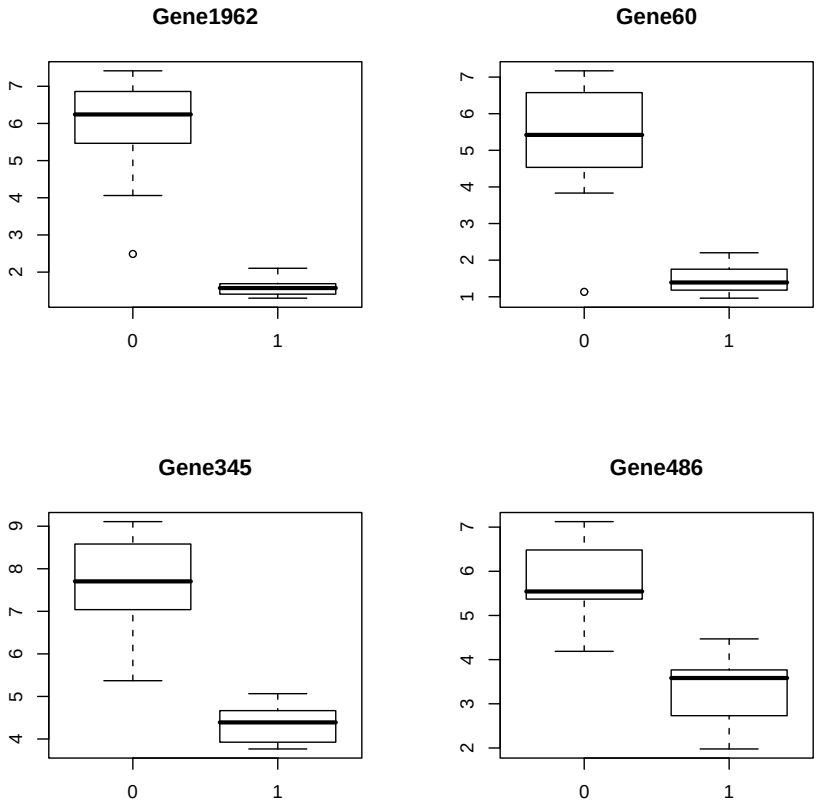

Figure 13.3: *Boxplot of the top four genes selected based on correlation of treatment effects.*

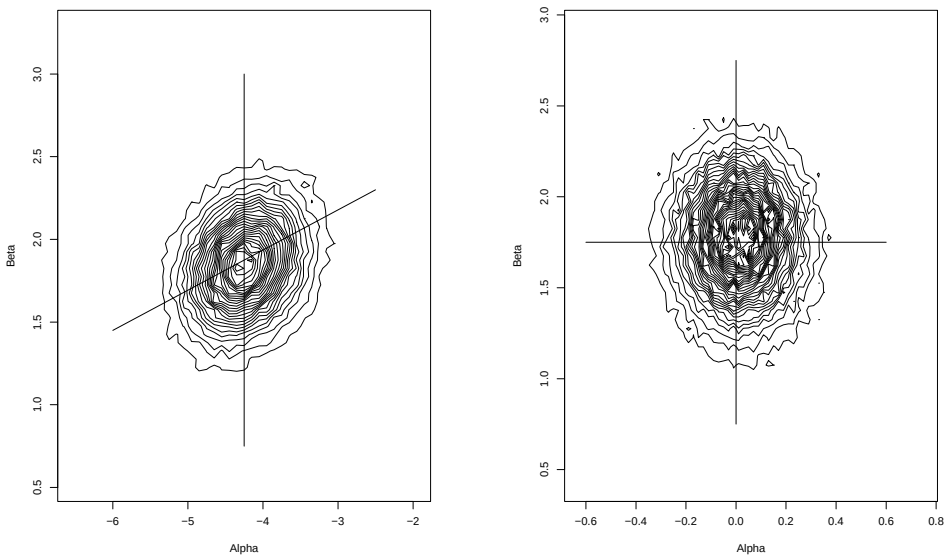

Figure 13.4: *Panel A: contour plot of treatment effect on the response against treatment effect on the gene expression for the top gene. Panel B: contour plot of treatment effect on the response against treatment effect on the gene expression for the lowest gene. gene.*

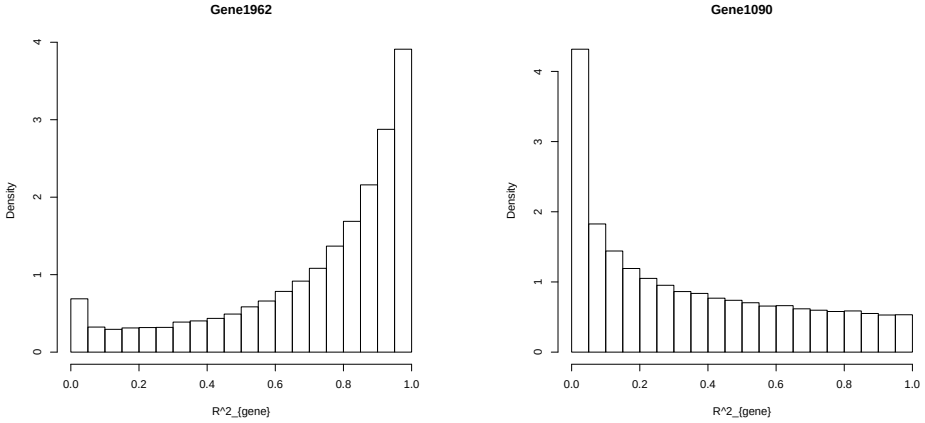

Figure 13.5: *Panel A: Histogram of the R^2_{gene} for top gene. Panel B: Histogram of the R^2_{gene} for the lowest gene.*

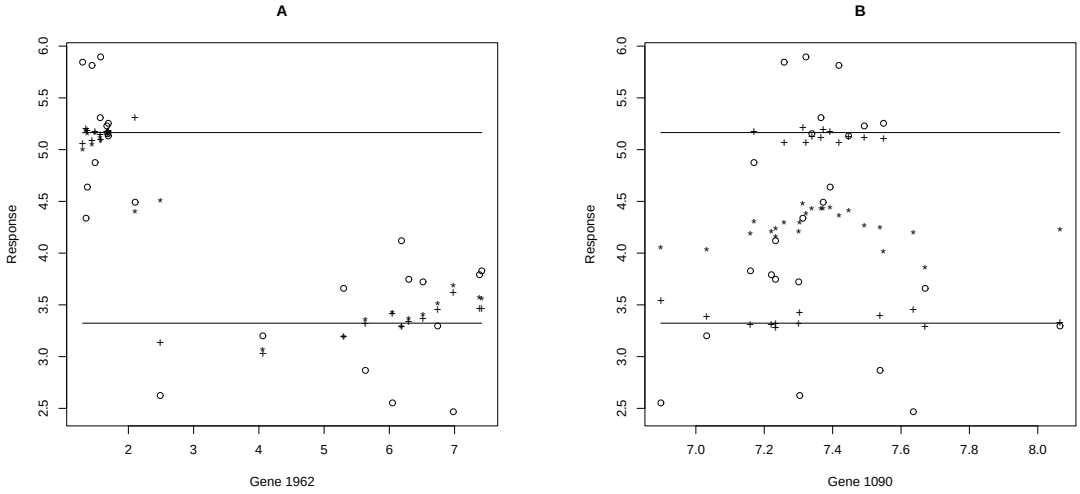

Figure 13.6: *Panel A: Scatter plot of observed versus predicted values for top gene. Panel B: Scatter plot of observed versus predicted values for lowest gene.*

14

Conclusions and Further Research

Statistical methods that can be used for the selection and evaluation of biomarkers which later will be validated to graduate into potent surrogate markers are in dire demand. The selected biomarkers need to go through rigorous testing procedures both statistically as well as biologically to finally end up being reliable replacements for the clinically relevant endpoint. The competitive nature of the pharmaceutical industry and the impending societal demand for urgent drugs for some of the chronic diseases threatening human life add to the need for these methods.

In this thesis we have revised some existing methods, introduced some new once and assessed their merit through simulation studies. The thesis is organized in to two parts. The first part dealt with surrogate marker validation and the second one is devoted to the selection and evaluation of biomarkers mainly genomic biomarkers from microarray experiments. This however is not the natural ordering of the events.

In practice, we first identify a biomarker which later will be promoted to a surrogate endpoint. However, the aim here is to introduce a set of methods that have been designed to validate a surrogate endpoint for the selection and evaluation of biomarkers.

The case of two normally distributed outcomes has been thoroughly investigated by several authors but there were still some subtle issues that were marginalized. Some of these issues are related to computational considerations that need due attention which other wise could cause the conclusions to follow questionable. The treatment coding might appear a trivial task but our simulation study has revealed that it needs to be taken seriously specially when a random effects approach is followed. The choice between a $0/1$ and $-1/+1$ coding schemes has implications in the positive definiteness and ill-conditioning of the variance covariance matrices based up on which the association measures are computed. In most instances, statistical software packages such as SAS, report that the variance-covariance matrix is not positive definite when there is a negative eigenvalue. But ill-conditioned covariance matrices for which there is a huge discrepancy between the smallest and the largest eigenvalues might pass undetected. The association measures based on such matrices will be exaggerated sending a false alarm which could lead to grave consequences. Thus considerable attention should be given to the ill-condition matrices. We have introduced the condition number, which is the ratio of the largest to the smallest eigenvalue, as a possible gauge to determine the influence of the ill-conditioning of the covariance matrix on the magnitude of the association measure.

A move away from the normal-normal setting induces new set of challenges. The fine properties of the bivariate normal distribution will no longer hold true for other bivariate distributions and hence new methods need to be devised to circumvent

these challenges. This has led to the use of the information-theoretic approach which has resulted in a unification across different outcome types. The method was found to work well for a combination of two binary outcomes and mixture of binary and continuous outcomes with a slight downward bias for small samples. However, the measure has shown to be less adept for the case of time-to-event true endpoints. The presence of censored observations introduced a new form of challenge. Two direction were followed, the first is to consider the number of events as denominator in the computation of the likelihood reduction factor and the second one considered the use of the number of subjects rather than events. The later one has resulted in a substantial downward bias while the former pointed in the opposite direction. Of the two however, the one that uses the number of events works reasonably well for small percentage of censoring.

In light of the shortcomings of the information-theoretic approach for time-to-event true endpoints, two other methods were entertained. The two methods are derived based on the information theory measure proposed by Kent (1983). Because the methods were thoroughly investigated for their robustness against censoring under the Cox-proportional hazard assumption, we tried to asses their merit when this assumption is questionable. The simulation studies have reveled that the method due to Kent and O'Quiegely performs reasonably well for the case of time-to-event true endpoint and cross-sectional surrogate endpoint. This method however is criticized for its inability to accommodate time varying covariates and its computational complexity.

Repeated measures of a quantitative marker are commonly obtained in clinical trials. When such measurements have the ability to predict, and/or explain a large proportion of the variability of, future clinical measurements or status of a patient,

then the marker may be used as a surrogate for the final measurements or status of a patient at the end of the study. If this is the case, such a marker may lead to a reduction of the length and hence the cost of the study. For example, instead of taking repeated measurement for a period of say 60 months, it may be possible to use the repeated measurements for the first 24 or 30 months to *accurately* predict the measurement at 60 months; thereby reducing the length of the study by about 50%. This phenomenon was studied in this thesis for the case of binary and continuous longitudinal sequences. Two special cases for the correlation structure namely compound symmetry and Auto-regressive of order one were specially treated for which analytical solutions have been derived. For the compound symmetry and other correlation structures where the correlation between repeated measures decays slowly, few earlier measures were sufficient to adequately predict the final measurement of a longitudinal sequence. For the Auto-regressive of order one correlation structure, depending on the magnitude of the correlation the first measurement or the entire sequence were needed. Logically a large number of earlier repeated measures might provide good prediction of the ultimate measurement but this entails a larger cost. And on the other hand, taking very few measures might hinder the precision of prediction. A balance should be strike between cost and precision. This was handled by introducing a cost function.

The selection and evaluation of biomarkers part of the thesis mainly focused on using the same set of methods that have been devised to validate surrogate endpoints. The concepts related to individual and trial level surrogacy got their analogies in the form of prognostic and therapeutic biomarkers. This analogy laid the foundation for the use of the measures in the surrogate marker to select and evaluate genomic biomarkers. The individual level surrogacy which works at the individual patient

level was directly used with little or no modification to select prognostic biomarkers. The trial level surrogacy measure is based on a meta-analytic framework which has hindered its direct use to select therapeutic biomarkers as most microarray experiments are single trial experiments. This has led to the use of a Bayesian approach which mimics the meta-analytical frame work. The analogy emanates from the use of a distributional assumption around the treatment effects on the clinical outcome and the gene expression. This has led to the derivation of the so-called gene-level association.

As it is not possible to exhaust all possible scenarios and all details within the scenarios considered, some questions are left for further research. The information-theoretic approach as opposed to a probit formulation was appreciated as it quantifies the association at the observed scale of the outcome. But the amount of bias on the individual level association introduced from moving from the latent scale to the observable scale is not clear. This might be further investigated through a simulation study or analytical derivation.

Two competing methods namely partial least squares and principal components were used to construct joint biomarkers with different ways of selecting the set of genes to be used in the construction. The methods were applied to a case study but were not formally tested in a simulation setting and this can lead to one further research. Moreover, several alternatives to the methods that deal with linear associations and a Bayesian approach were entertained for the selection of prognostic and therapeutic biomarkers respectively. These methods however, have not been formally investigated for their merit using a simulation study. We therefore consider this also an interesting area for further research.

References

Abdi, H. (2003). Partial Least Squares Regression; Multivariate analysis. In M. Lewis-Beck, A. Bryman, T. Futing (Eds) *Encyclopedia for research methods for the social sciences. Thousand Oaks Sage.*

Albert, J.M., Ioannidis, J., Reichelderfer, P., Conway, B., Coombs. R., Crane, L., Demasi, R., Dixon. D., Flandre, P., Hughes, M., Kalish, L., Lartnz. K., Lin, D., Marschner. I., Munõz, A., Murray, J., Neaton, J., Pettinelli, C., Rida, W., Taylor, J., and Welles, S. (1998). Statistical issues for HIV surrogate endpoints: point and counterpoint. *Statistics in Medicine*, **17**, 2435–2462.

Alonso, A. and Molenberghs, G. (2007). Surrogate marker evaluation from an information theory perspective. *Biometrics*, **63**, 180–186.

Alonso, A., Geys, H., Molenberghs, G., and Kenward, M. (2004). Validation of surrogate markers in multiple randomized clinical trials with repeated measurements: canonical correlation approach. *Biometrics*, **60**, 845–853.

Alonso, A., Geys, H., Molenberghs, G., and Vangeneugden, T. (2002). Investigating the criterion validity of psychiatric symptom scales using surrogate marker validation methodology. *Journal of Biopharmaceutical statistics*, **12**, 161–178.

Alonso, A., Geys, H., Molenberghs, G., and Vangeneugden, T. (2002). Investigat-

ing the criterion validity of psychiatric symptom scales using surrogate marker validation methodology. *Journal of Biopharmaceutical Statistics* **12**, 161–179.

Alonso, A., Geys, H., Kenward, M.G., Molenberghs, G., and Vangeneugden, T. (2003). Validation of surrogate markers in multiple randomized clinical trials with repeated measurements. *Biometrical Journal*, **45**, 1–15.

Alonso, A., Molenberghs, G., Burzykowski, T., Renard, D., Geys, H., Shkedy, Z., Tibaldi, F., Cortiñas, J., and Buyse, M. (2004). Prentice's approach and the meta-analytic paradigm: a reflection on the role of statistics in the evaluation of surrogate endpoints. *Biometrics*, **60**, 724–728.

Alonso, A., Molenberghs, G., Geys, H., and Buyse. M. (2005). A unifying approach for surrogate marker validation based on Prentice's criteria. *Statistics in Medicine.*, **25**, 205–211.

Amaratunga, D. and Cabrera, J. (2004). Exploration and Analysis of DNA Microarray and Protein Array Data. *New York: John Wiley & Sons.*

Anderson, T.W. (1958). An Introduction to Multivariate Statistical Analysis. *New York: Wiley.*

Bahadur, R. R.(1961). A representation of the joint distribution of responses to n dichotomous items. *In Solomon, H. Studies in Item Analysis and Prediction*, 158-168.

Bair, E., Hastie, T., Paul, D. and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, **101**, 119–137.

Baker, G. (2006). A simple meta-analytic approach for binary surrogate and true endpoints. *Biostatistics*, **7**, 57–70.

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

Breiman, L. (1996a). Bagging predictors. *Machine Learning*, **26**, 123–140.

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

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

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*, **57**, 289–300.

Biomarkers Definitions Working Group. (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical Pharmacology and Therapy*, **69**, 89–95.

Buyse, M. and Molenberghs, G. (1998). The validation of surrogate endpoints in randomized experiments. *Biometrics*, **54**, 1014–1029.

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.

Burzykowski, T. , Molenberghs, G. , and Buyse, M. (2004). The validation of surrogate endpoints using data from randomized clinical trials: a case-study in advanced colorectal cancer. *Journal of Royal Statistical Society Series A*, **167**, 103–124.

Burzykowski, T., Molenberghs, G., and Buyse, M. (2005). The Evaluation of Surrogate Endpoints. *New York: Springer*.

Cortiñas, J., Molenberghs, G., Burzykowski, T., Shkedy, Z., and Renard, D. (2004). Choice of units of analysis and modeling strategies in multilevel hierarchical models. *Computational Statistics and Data Analysis*, **47**, 537–563.

Cover, M. and Thomas, A. (1991). Elements of Information Theory. *New York: John Wiley & Sons*.

Corfu-A Study Group (1995). Phase III randomized study of two fluorouracil combinations with either interferon alfa-2a or leucovorin for advanced colorectal cancer. *Journal of Clinical Oncology*, **13**, 921–928.

Daniels, M.J. and Hughes, M.D. (1997). Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine*, **16**, 1965–1982.

De Gruttola, V., Fleming, T.R., Lin, D.Y. and Coombs, R. (1997). Validating surrogate markers – are we being naive? *Journal of Infectious Diseases*, **175**, 237–246.

De Gruttola, V., Wulfsohn, M. , Fishk, MA. , and Tsiatis, AA. (1993). Modelling the relationship between survival and CD4 lymphocytes in patients with AIDS and AIDS-related complex. *Journal of Acquired Imino Deficiency Syndrome.*, **6**, 359–365.

De Groote, L. and Linthorst, A.C. (2007). Exposure to novelty and forced swimming evoke stressor- dependent changes in extracellular GABA in the rat hippocampus. *Neuroscience*, **148**, 794–805.

Ding, C.G. (1996). On the computation of the distribution of the square of the sample multiple correlation coefficient. *Computational Statistics and Data Analysis*, **22**, 345–350.

Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A., and Vapnik V. (1997). Support vector regression machines. In: Mozer M.C., Jordan M.I., and Petsche T. (Eds.), Advances in Neural Information Processing Systems 9, MIT Press, Cambridge, MA, pp. 155–161.

Efron, B. and Tibshirani, R. (1993). An Introduction to the Bootstrap. *New York: Chapman & Hall.*

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, **92**, 548–560.

Ellenberg, SS. and Hamilton, JM. (1989). Surrogate endpoints in clinical trials: cancer. *Statistics in Medicine*, **8**, 405–413.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* 1996; **11**, 89–121.

Faes, C., Geys, H., Molenberghs, G., Aerts, M., Cadarso-Suarez, C., Acuña, C., and Cano, M. (2006). A flexible method to measure synchrony in neuronal firing. *Journal of American Statistical Association*, **101**, 000–000.

Fisher, R. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, **22**, 700–725.

Fleming, TR. and DeMets, DL. (1996). Surrogate endpoints in clinical trials: are we being misled? *Annals of Internal Medicine*, **125**, 605–613.

Ferentz, A.E. (2002). Integrating pharmacogenomics into drug development. *Pharmacogenomics*, **3**, 453–467.

Freedman, LS., Graubard, BI., and Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine*, **11**, 167–178.

Fletcher R. (1989). Practical Methods of Optimization. *New York: John Wiley.*

Gail, M., Pfeiffer, R., van Houwelingen, HC., Carroll RJ. (2000). On meta-analytic assessment of surrogate outcomes. *Biostatistics*, **1**, 231–246.

Galecki, A.T. (1994). General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics*, **23**, 3105–3120.

Graybill, F.A. (1983). Matrices with Applications in Statistics (2nd ed.) *Belmont, California: Wadsworth.*

Greco, F.A., Figlin, R., York, M., Einhorn, L., Schilsky, R., Marshall, E.M. (1996). Phase III randomized study to compare interferon alfa-2a in combination with fluorouracil versus fluorouracil alone in patients with advanced colorectal cancer. *Journal of Clinical Oncology*, **14**, 2674–2681.

International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials. (1996). *Federal Register*, **63**, **179**, 49583.

Ihaka, R. and Gentleman R. (1996). R: A Language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314.

Kanter, M. (1975). Autoregression for Discrete process Mod 2. *Journal of Applied Probability*, 1975; 371–375

Kay, S.R., Fiszbein, A., and Opler, L.A. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin*, **13**, 261–276.

Kay, S.R., Opler, L.A., and Lindenmayer, J.P. (1988). Reliability and validity of the Positive and Negative Syndrome Scale for schizophrenics. *Psychiatric Research*, **23**, 99-110.

Kent, J. (1983). Information gain and a general measure of correlation. *Biometrika*, **70**, 163–173.

Kent, J., O'Quigley, J. (1988). Measures of dependence for censored survival data. *Biometrika* , **75**, 525-534.

Lesko, L.J. and Atkinson, A.J. (2001). Use of biomarkers and surrogate endpoints in drug development and regulatory decision making: criteria, validation, strategies. *Annual Review of Pharmacological Toxicology*, **41**, 347–366.

Lunn, A.D., and Davies, S.T. (1998). A note on generating correlated binary variables. *Biometrika*, **85**, 487–490.

Liaw, A. and Wiener, M. (2002). Classification and regression by random forest. *The Newsletter of the R Project 2/3*, 18–22.

Maruish, M.R. (1999). The Use of Psychological Testing for Treatment Planning and Outcomes Assessment. *Mahwah, NJ: Lawrence Erlbaum Associates.*

McIntosh, M.W. (1996). The population risk as an explanatory variable in research synthesis of clinical trials, *Statistics in Medicine*, **15**, 1713–1728.

Meyer, D. (2001). Support vector machines, the interface to libsvm in package e1071. *The Newsletter of the R Project 1/3*, 23–26.

Meyer, D., Leisch, F., Hornik, K. (2003). The support vector machine under test. *Neurocomputing*, **55**, 169–186.

Meyer, D., Leisch, F., Hornik, K. (2002). Benchmarking support vector machines. *Technical Report*, **78**, SFB "Adaptive Information Systems and Modeling in Economics and Management Science".

Molenberghs, G., Geys, H., and Buyse, M. (2001). Evaluation of surrogate endpoints in randomized experiments with mixed discrete and continuous outcomes. *Statistics in Medicine*, **20**, 3023–3038.

Molenberghs, G. and Verbeke, G. (2005). Models for Discrete Longitudinal Data. *New York: Springer*.

Moore, D.E., Lees, B.G., and Davey, S.M. (1991). A new method for predicting vegetation distributions using decision tree analysis in a geographic information system. *Journal of Environmental Management*, **15**, 59–71.

Ngo, L. and Wand, M.P. (2004). Smoothing with mixed model software. *Journal of Statistical Software*, **9**, 1–56.

Patrick, J.F., and Katrina, R.S. (2008). Methods for Generating Longitudinally Correlated Binary Data. *International Statistical Review*, **76**, 28–38.

O'Quigley, J., Xu, R., and Stare, J. (2005). Explained randomness in proportional hazards models. *Statistics in medicine*, **24** 479–89.

Oosterlinck, W., Mattelaer, J., Casselman, J., Van Velthoven, R., Derde, M.P., Kaufman, L. (1997) PSA evolution: a prognostic factor during treatment of advanced prostatic carcinoma with total androgen blockade. Data from a Belgian multicentric study of 546 patients. *Acta Urol Belg*, **65**, 63–71.

Ovarian Cancer Meta-Analysis Project. (1991). Cyclophosphamide plus cisplatin versus cyclophosphamide, doxorubicin, and cisplatin chemotherapy of ovarian carcinoma: a meta-analysis. *Journal of Clinical Oncology*, **9**, 1668–1674.

Prentice, R.L. (1989). Surrogate endpoints in clinical trials: definitions and operational criteria. *Statistics in Medicine*, **8**, 431–440.

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.

Rodríguez, G. and Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A*, **158**, 73–89.

Royston, P. and Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Applied Statistics*, **43**, 429–467.

Royston, P., Parmar, M.K.B., and Qian, W. (2003). Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer. *Statistics in Medicine*, **22**, 2239–2256.

Ruppert, D., Wand, M.P., and Carroll, R.J. (2003). Semiparametric Regression. *Cambridge: Cambridge University Press.*

Senn, S. (1993). Cross-over Trials in Medical Research. *Chichester: John Wiley.*

Schatzkin, A. and Gail, M. (2002). The promise and peril of surrogate end points in cancer research. *Nature Reviews Cancer*, **2**, 19–27.

Schatzkin, A, Gail, M., and Freedman, L. (2005), The promise and peril of surrogate endpoint in cancer research. In: *The Evaluation of surrogate endpoints*. T. Burzykowski, T., Molenberghs, G., and Buyse, M. (Eds.) New York: Springer, pp. 349–366.

Shannon C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, **27** 379–423 and 623–656.

Schember, M., Stare, J. (1996). Explained variation in survival analysis. *Statistics in Medicine*, **15**, 1999–2012.

Tibaldi, S., Cortiñas, J., 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 computations and Simulations*, **73**, 643–658.

Tilahun, A., Assam, P., Alonso, A., and Molenberghs, G. (2007a). Flexible surrogate marker evaluation from several randomized clinical trials with continuous endpoints, using R and SAS. *Computational Statistics and Data Analysis*, **51**, 4152–4163.

Tilahun, A., Assam, P., Alonso, A., and Molenberghs, G. (2007b). Information-theory based surrogate marker evaluation from several randomized clinical trials with binary endpoints, Using SAS. *Journal of Biopharmaceutical Statistics.*, **00**, 000–000.

Therneau, T.M. and Atkinson, E.J. (1997). An introduction to recursive partitioning using the rpart routines. *Technical Report*, **61**, Department of Health Science Research, Mayo Clinic, Rochester, New York.

Xu, R., O'Quigley, J. (1999). A measure of dependence for proportional hazards models. *Journal of Nonparametric Statistics*, **12**, 83-107.

Van Houwelingen, H.C., Arends, L.R., and Stijnen, T. (2002). Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine*, **21**, 589–624.

Vapnik, V. (1995). The Nature of Statistical Learning Theory. *New York: Springer*.

Verbeke, G. and Molenberghs, G. (2000). Linear Mixed Models for Longitudinal Data. *New York: Springer*.

Verbyla, A.P., Cullis, B.R., Kenward, M.G., and Welham, S.J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines. *Applied Statistics*, **48**, 269–311.

Verbyla, D.L., (1987). Classification trees: a new discrimination tool. *Canadian Journal of Forestry Research*, **17**, 1150–1152.

Vapnik, V. and Lerner, A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, **24**, 774–780.

Vapnik, V. and Chervonenkis, A. (1964). A note on one class of perceptrons. *Automation and Remote Control*, **25**.

Winkens, B., Schouten, H.J.A, van Breukelen, G.J.P., and Berger, M.P.F. (2005). Optimal time-points in clinical trials with linearly divergent treatment effects. *Statistics in Medicine*, **24**, 3743–3756.

A

Mathematical Derivations

Here we will outline the analytical derivations used in the chapter concerned with the optimal number of repeated measures. It has to be recalled that, in the chapter on the mixed longitudinal and cross-sectional setting, we have shown that, the R^2_Λ and VRF_{ind} are equal for a longitudinal surrogate and a cross-sectional true endpoint, and hence we use R^2_Λ in place of VRF_{ind} for ease of notation.

Derivation of The Association Measures

Compound Symmetry case

Let us assume that we have k longitudinal observations with a mean vector μ and variance covariance matrix Σ_c:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_k \end{pmatrix}, \qquad E(Y) = \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ . \\ \mu_k \end{pmatrix} \qquad V(Y) = \Sigma_c,$$

we will further assume that Σ_c is a kXK compound symmetric matrix , i.e

$$\Sigma_c = \sigma \begin{pmatrix} 1 & \rho & . & . & . & \rho \\ \rho & 1 & . & . & . & \rho \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \rho & \rho & . & . & . & 1 \end{pmatrix} = \sigma(1-\rho)I_k + \sigma\rho J_k,$$

where $J_k = 1_k 1_k^{'}$. It is well known that (Graybill 1983),

$$\|\Sigma_c\| = \sigma^k (1-\rho)^{k-1}(1+(k-1)\rho.$$

We now want to evaluate the performance of the first m observations as a surrogate for the last one. Therefore in this setting we will consider:

$$S = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_m \end{pmatrix} \qquad T = Y_k.$$

$$X = \begin{pmatrix} S \\ T \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_m \end{pmatrix}, \qquad E(X) = \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ . \\ \mu_k \end{pmatrix}$$

and $V(X) = \Sigma$ where Σ is a (m+1)$\times$ (m+1) compound symmetry matrix. Essentially, Σ can be decomposed as:

$$\Sigma = \sigma \begin{pmatrix} R_{SS} & R_{ST} \\ R_{TS} & R_{TT} \end{pmatrix},$$

where:

1. R_{SS} is a compound symmetric correlation matrix.
2. R_{TS}= (ρ, ρ, ..., ρ) is a 1$\times$ m vector and $R_{ST} = R_{TS}^t$
3. R_{TT}=1

The amount of information on T that S brings can be quantified as:

$$R_\Lambda^2 = 1 - \frac{|\Sigma|}{|\Sigma_{TT}| \cdot |\Sigma_{SS}|}.$$

Using (1) and (2) and the expression for the determinant of a compound symmetry matrix given earlier we have:

$$R^2_\Lambda(m) = 1 - \frac{\sigma^{m+1}(1-\rho)(1+m\rho)}{\rho^{m+1}(1-\rho)^{m-1}(1+(m-1)\rho)} = 1 - \frac{(1-\rho)(1+m\rho)}{1+(m-1)\rho}.$$

$$\Rightarrow R^2_\Lambda(m) = \frac{m\rho^2}{1+(m-1)\rho}.$$

The $R^2_\Lambda(m)$ is a function of m, the number of repeated measurements, if we calculate the derivative of $R^2_\Lambda(m)$ with respect to m we get:

$$\frac{d}{dm}R^2_\Lambda(m) = \frac{\rho^2(1-\rho)}{[1+(m-1)\rho]^2} \geq 0.$$

This implies that if $\rho \neq 1$ then $R^2_\Lambda(m)$ is an increasing function of m i.e the more repeated measures we include in S, the more precise our prediction of T will be. However, another important question is concerned with the impact of ρ on this information gain, i.e, how the value of ρ influences the amount of information that S brings about T. To study this issue further, let us consider the additional information that one extra observation will bring. This means, let us consider a new surrogate formed by adding another observation to S. For this new surrogate :

$$R^2_\Lambda(m+1) = \frac{(m+1)\rho^2}{1+m\rho}$$

let us define

$$g(\rho) = \frac{R^2_\Lambda(m+1)}{R^2_\Lambda(m)} = \left(\frac{m+1}{m}\right)\left(\frac{1+(m-1)\rho}{1+m\rho}\right)$$

$g(\rho)$ quantifies how much extra information about the true endpoint we get by considering another observation.

$$g^{'}(\rho) = \left(\frac{m+1}{m}\right)\left(\frac{-1}{[1+m\rho]^2}\right) < 0.$$

This last equation implies that $g(\rho)$ is a decreasing function of ρ, i.e, the higher the correlation between two consecutive observations the less we gain by taking more observations. On the other hand, the lower the ρ the more meaningful it is to consider more observations. Note that $g(\rho)$ will reach its maximum when $\rho = 1$ and in that case:

$$g(1) = \frac{R^2_\Lambda(m+1)}{R^2_\Lambda(m)} = 1 \Leftrightarrow R^2_\Lambda(m+1) = R^2_\Lambda(m)$$

and therefore, adding a new observation will not bring any additional information. Indeed, if $\rho = 1$ then there is deterministic relationship between Y_i and Y_k for all i. Actually, knowing the value of Y_1 would be enough to predict $T = Y_k$ without error.

Conversely, if $\rho = 0$ then $R^2_\Lambda(m) = 0$ for all $m = 1, \ldots k-1$. Obviously in that situation all the observations are independent and no sensible prediction is possible. Finally, it is important to point out that in all the previous analysis the position of the chosen surrogate vector S is totaly irrelevant, i.e, all these results will be equally valid if we consider the following vector: $S^t = (Y_{i+1}, Y_{i+2}, .\,, .\, Y_{i+m})$ with $i + m < k$.

Auto-Regressive of Order One AR(1)

Let us consider the same general settings as in the compound symmetry case with $V(Y) = \Sigma_{AR}$ where Σ_{AR} is now the variance covariance matrix of an $AR(1)$ process, i.e

$$\Sigma_{AR} = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & . & . & \rho^{k-1} \\ \rho & 1 & . & . & . & \rho^{k-2} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \rho^{k-1} & \rho^{k-2} & . & . & . & 1 \end{pmatrix}.$$

Like before we want to evaluate the performance of the first m observations as a surrogate for the last one. For this situation $V(X) = \Sigma$ where

$$\Sigma = \sigma \begin{pmatrix} R_{SS} & \delta \\ \delta^t & 1 \end{pmatrix} = \begin{pmatrix} \sigma R_{SS} & \sigma\delta \\ \sigma\delta^t & \sigma \end{pmatrix} = \begin{pmatrix} \Sigma_{SS} & \Sigma_{ST} \\ \Sigma_{TS} & \Sigma_{TT} \end{pmatrix},$$

here:

1. R_{SS} is an $AR(1)$ $m \times m$ correlation matrix.

2. $\delta^t = (\rho^{k-1}, \rho^{k-2}, \ldots, \rho^{k-m}) = \rho^{k-m}(\rho^{m-1}, \rho^{k-2}, .. \rho, 1)$

so Σ can be written as:

$$\Sigma = \left(\begin{array}{ccccccc|c} 1 & \rho & \rho^2 & . & . & . & \rho^{m-1} & \rho^{k-1} \\ \rho & 1 & \rho & . & . & . & \rho^{m-2} & \rho^{k-2} \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ \rho^{m-1} & \rho^{m-2} & . & . & . & . & 1 & . \\ \hline \rho^{k-1} & \rho^{k-2} & . & . & . & . & \rho^{k-m} & 1 \end{array}\right).$$

In this scenario it has been shown that:

$$R^2_\Lambda = \frac{\Sigma_{TS}\Sigma^{-1}_{SS}\Sigma_{ST}}{\sigma_{TT}} = \frac{\sigma\delta^t(\sigma R_{SS})^{-1}\sigma\delta}{\sigma}$$

$$\Rightarrow R^2_\Lambda = \rho^{2(k-m)} delta^t_1 R^{-1}_{SS}\delta_1$$

where $\delta^t_1 = (\rho^{m-1}, \rho^{m-2}, ., ., ., \rho, 1)$. Note that R_{SS} is again an $AR(1)$ matrix of dimension m and from Gray bill (1983), we have

$$R_{SS}^{-1} = \frac{1}{(1-\rho^2)} \begin{pmatrix} 1 & -\rho & 0 & . & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & . & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & . & 0 & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & 0 & . & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & . & \rho & 1 \end{pmatrix}$$

In general if c_i denotes the i^{th} column of R_{SS} then:

$C_i^t = (0,\, 0,\, .,\, .,\, .,\, 0,\, ,-\rho,\, 1+\rho^2,\, -\rho,\, 0,\, .,\, .,\, .,\, 0)$ $i = 2,\, .,\, .,\, .,\, m-1$ where the first $-\rho$ appears in the $i-1$ component, and $C_1^t = (1, -\rho,\, 0,\, .,\, .,\, .,\, 0) C_m^t = (0, .,\, .,\, .,\, .,\, -\rho,\, 1)$. Using this notation we have that:

$$\delta_1^t R_{SS} = \frac{1}{(1-\rho^2)} (\delta_1^t C_1,\, \delta_1^t C_2,\, .,\, .,\, .,\, \delta_1^t C_m),$$

but : $\delta_1^t C_i = \rho^{m-i}(-\rho) + \rho^{m-i-1}(1+\rho^2) + \rho^{m-i-2}(-\rho)) = -\rho^{m-i+1} + \rho^{m-i-1} + \rho^{m-i+1} - \rho^{m-i-1} = 0$ so $\delta_1^t C_i = 0$ for $i = 2,\, .,\, .,\, .,\, m-1$, we also have :

$$\delta_1^t C_1 = \rho^{m-1} - \rho^{m-1} = 0$$

$$\delta_m^t C_m = -\rho^2 - \rho^{m-1} = 0$$

$$\Rightarrow \delta_1^t R_{SS} = \frac{1}{1-\rho^2} (0,\, 0,\, .,\, .,\, .,\, 1-\rho^2)$$

Finally we have:

$$\delta_1^t R_{SS}^{-1} = \frac{1}{1-\rho^2}(0,\, 0,\, .,\, .,\, .,\, 1-\rho^2) \begin{pmatrix} \rho^{m-1} \\ \rho^{m-2} \\ . \\ . \\ . \\ \rho \\ 1 \end{pmatrix}$$

$$= \frac{1}{1-\rho^2}(1-\rho^2) = 1$$

$$\Rightarrow \delta_1^t R_{SS}^{-1} \delta_1 = 1,$$

and therefore $R_\Lambda^2 = \rho^{2(k-m)}$ where $k = 1, ., ., ., m-1$. Here again $R_\Lambda^2(m)$ is an increasing function of m, i.e., the more observations we take, the more preciese our prediction on the true endpoint will be. Additionally, R_Λ^2 is also an increasing function of ρ and, therefore, the higher the correlation the more meaningful is to take more observations. Unlike in the compound symmetry case, in this scenariao the "position" of the of the surrogate sequence becomes relevant. Indeed, let us assume that we shift the entire sequence in the following way:

$$s_{new} = \begin{pmatrix} Y_s \\ Ys+1 \\ . \\ . \\ . \\ Y_{s+m} \end{pmatrix},$$

with $s + m < k$. In this scenario it is easy to see that: $R_{\Lambda s}^2 = \rho^{2(k-(s+m-1)}$ and obviously $R_{\Lambda s}^2 \geq R_\Lambda^2$ for $s \geq 1$. This implies that considering m observations closer to the true endpoint will result in a surrogate with more predictive power than the

one obtained by using m observations further away from the true endpoint.

Computing the Optimal Number of Measurements

Compound Symmetry case

We have proposed to calculate the optimal number of measurements to predict the true endpoint by minimizing the objective function:

$$CPR0(m) = w_1 \cdot (1 - R^2_\Lambda(m)) + (1 - w_1) \cdot \frac{R + m}{R + K},$$

where k is the total number of measurements and $1 \leq m \leq k$.

We know that, for the compound symmetry case:

$$1 - R^2_\Lambda(m) = \frac{(1 - \rho)(1 + m\rho)}{1 + (m - 1)\rho}.$$

and therefore:

$$CPR0(m) = w_1 \cdot \frac{(1 - \rho)(1 + m\rho)}{1 + (m - 1)\rho} + (1 - w_1) \cdot \frac{R + m}{R + K}.$$

To find the maximum of $CPR0(m)$ we need to solve the score equation:

$$\frac{d}{dm} CPR0(m) = 0.$$

But

$$\frac{d}{dm} CPR0(m) = w_1 \cdot (1 - \rho) \cdot \frac{d}{dm} \left(\frac{1 + m\rho}{1 + (m - 1)\rho} \right) + \frac{1 - w_1}{R + K}$$

$$\frac{d}{dm} \left(\frac{1 + m\rho}{1 + (m - 1)\rho} \right) = \frac{-\rho^2}{[1 + (m - 1)\rho]^2}$$

$$\Rightarrow \frac{d}{dm} CPR0(m) = \frac{-w_1 \cdot (1 - \rho)\rho^2}{[1 + (m - 1)\rho]^2} + \frac{1 - w_1}{R + K}$$

and this implies:

$$\Rightarrow \frac{d}{dm}CPR0(m) \Leftrightarrow \frac{1-w_1}{R+K} = \frac{w_1 \cdot (1-\rho)\rho^2}{[1+(m-1)\rho]^2}$$

Solving this equation with respect to m we get:

$$m_{12} = \left(\frac{-(1-\rho)}{\rho}\right) \pm \sqrt{\frac{(R+k)w_1(1-\rho)}{1-w_1}}$$

so essentially we have two solutions:

$$m_1 = \left(\frac{-(1-\rho)}{\rho}\right) + \sqrt{\frac{(R+k)w_1(1-\rho)}{1-w_1}}$$

$$m_2 = \left(\frac{-(1-\rho)}{\rho}\right) - \sqrt{\frac{(R+k)w_1(1-\rho)}{1-w_1}}$$

The value of m that minimizes $CPR0(m)$ is the one for which its second derivative is positive:

$$\frac{d^2}{dm^2}CPR0(m) = \frac{2 \cdot w_1(1-\rho)\rho^3}{[1+(m-1)\rho]^3}$$

and therefore :

$$\frac{d^2}{dm^2}CPR0(m_1) = \frac{2 \cdot w_1(1-\rho)\rho^3}{\left[\frac{(R+K)w_1\rho^2(1-\rho)}{dm^2}\right]^{3/2}}$$

$$\frac{d^2}{dm^2}CPR0(m_2) = \frac{-2 \cdot w_1(1-\rho)\rho^3}{\left[\frac{(R+K)w_1\rho^2(1-\rho)}{dm^2}\right]^{3/2}}$$

We have then the following case:

1. If $\rho > 0$, $\frac{d^2}{dm^2}CPR0(m_2) > 0$ and m_1 is the optimal
2. If $\rho < 0$, $\frac{d^2}{dm^2}CPR0(m_2) > 0$ and m_2 is the optimal

In a practical situation m_1 and/or m_2 might not necessarily be integers and hence we should take the next closest integer.

Auto-Regressive of Order One AR(1)

Similar to the compound symmetry case, we want to calculate the optimal number of measurements to predict the true endpoint by minimizing the objective function:

$$CPR0(m) = w_1 \cdot (1 - R^2_\Lambda(m)) + (1 - w_1) \cdot \frac{R + m}{R + K}, \tag{A.1}$$

where k is the total number of measurements and $1 \leq m \leq k$.

We know that, for the $AR(1)$ case:

$$R^2_\Lambda(m) = \rho^{2(k-m)},$$

and therefore:

$$CPR0(m) = w_1[1 - \rho^{2(k-m)}] + (1 - w_1)\frac{R + M}{R + K}.$$

Now to find the value of m that maximizes $CPR0(m)$ we need to solve the score equation:

$$\frac{d}{dm}CPR0(m) = 2w_1\rho^{2(k-m)} \log \rho + \frac{1 - w_1}{R + K}.$$

But

$$\frac{d}{dm}CPR0(m) = 0$$

$$\Leftrightarrow -2w_1\rho^{2(k-m)} \log \rho = \frac{1 - w_1}{R + K}$$

$$\Leftrightarrow \rho^{2(k-m)} = \frac{-(1 - w_1)}{2w_1(R + K) \log \rho},$$

and this implies:

$$\Leftrightarrow 2(k-m)\log\rho = \log\left[\frac{-(1-w_1)}{2w_1(R+K)\log\rho}\right]$$

$$\Leftrightarrow (k-m) = \frac{\log\left[\frac{-(1-w_1)}{2w_1(R+K)\log\rho}\right]}{2\log\rho}$$

$$\Leftrightarrow m = k - \frac{\log\left[\frac{-(1-w_1)}{2w_1(R+K)\log\rho}\right]}{2\log\rho}.$$

To ascertain whether m maximizes or minimizes $CPR0(m)$ we need to evaluate the second derivative of the function, we have:

$$\frac{d^2}{dm^2}CPR0(m) = -4w_1(\log\rho)^2\rho^{2(k-m)} < 0,$$

for all m. This result implies that the previous value of m maximizes $CPR0(m)$. This from a practical point of view means that the minimum value of $CPR0(m)$ can only be attained at the two extreme cases i.e m= 1 or m= $k-1$.

B

Software

A list of generic SAS macros that have been used to carry out the analysis in this thesis are given below. For the details concerning how to invoke the respective macros, the data layout and the inputs required please refer to the macros which can be obtained from the center for statistics website or could be requested from the authors.

B.1 Two Continuous Outcomes

The analysis discussed in Chapter 3 can be conducted using the **SURCONCON** macro. The macro allows a choice between the full random effects approach or the simplified modeling strategies. It also allows the choice of four different types of bootstrap based confidence intervals for the trail and individual level surrogacy measures.

B.2 Two Binary Outcomes

The SAS macro **SURBINBIN** can be used to perform the analysis described in chapter 6. Both the meta-analytic approach and information theoretic approach can be carried out using this macro.

B.3 Mixture of Binary and Continuous Outcomes

The SAS macro **SURBINCON** can be used to perform the analysis described in chapter 5. Information theoretic approach with both fixed and random trial specific effects will be used.

B.4 Two Longitudinal Outcomes

The SAS macro **LONG_LONG** computes the VRF and R^2_Λ for the case of longitudinal endpoints with linear time effect or when time is considered as a class variable. For fractional polynomial and smoothing splines additional manipulation is required. A dataset will be produced namely "Bothlong" containing the VRF and R^2_Λ values.

B.5 Longitudinal and Cross-sectional Outcomes

The SAS macro **LONG_CROSS** computes the VRF and R^2_Λ for the case of mixture of longitudinal and cross-sectional endpoints. The macro computes the stated quantities by alternatively using the longitudinal and the cross-sectional outcomes as surrogate endpoints for the other.