1

Introduction

1.1 Introduction

A disease or medical condition is an abnormal condition of an organism that impairs bodily functions, associated with specific symptoms and signs. It may be caused by external factors, such as invading organisms, or it may be caused by internal dysfunctions, such as autoimmune diseases. In human beings, "disease" is often used more broadly to refer to any condition that causes extreme pain, dysfunction, distress, social problems, and or death to the person afflicted, or similar problems for those in contact with the person (Johnson 2002). Diseases in general, particularly chronic diseases, deprive individuals of their health and productive potential. From countries perspective, chronic diseases reduce life expectancy and ultimately economic productivity, thus depleting the quality and quantity of countries labor force (WHO 2005).

Due to the need for effective, safe, and affordable pharmaceuticals to control, cure

or eradicate diseases, especially those that cause high mortality and morbidity, drug regulatory agencies have been put in place to enforce a standard drug development process. The first step of the process, referred to as *drug discovery*, determines the target disease, develops hypothesis for a mechanism of treatment, as well as determine feasibility of producing and evaluating the selected compounds. Next comes the *screening*, where combination chemistry is used to obtain many possible compounds, which are tested for activity via high throughput screening. This is followed by the *Pre-clinical phase*, which is an animal study used to assess the safety and biological activities of the selected compound, together with lethal and normal dose levels for short and long term use. If promising results are obtained from the pre-clinical stage, a request is made to the appropriate regulatory agency to allow human exposure to the experimental drug. Upon approval from a regulatory agency, the investigational new drug is then tested and evaluated through a series of clinical trials. Table 1.1 provides a summary of the drug development process, with an indication of the sample size and or duration required at each stage.

1.2 The Concept of a Surrogate Endpoint

The drug development process is known to be complex, costly, and time-consuming (DiMasi et al. 1994, Kaitin and Healy 2000, Kaitin and DiMasi 2000). The process is also risky in that most compounds that undergo clinical testing are abandoned without obtaining marketing approval (Table 1.1). The rising costs of drug development and the challenges of new and re-emerging diseases are putting considerable demands on efficiency in the drug candidates selection process. Thus, careful consideration of all factors that have a significant impact on the process is needed, to appropriately allocate research and development resources. A very important factor influencing

the duration and complexity of the drug development process is the choice of the endpoint, which will be used to asses the efficacy of the drug or treatment. Two main criteria to select the endpoint are its sensitivity to detect treatment effects and its clinical relevance to goals of the study (Fleming 1996). The relevance depends on the purpose of the stage of the drug development in question. For example, evidence of the biological activity of a drug in Phase II trials, or a definitive evaluation of clinical benefit to patients in Phase III trials (Burzykwoski, Molenberghs and Buyse 2005).

Often the most sensitive and relevant clinical endpoint, which will be called the "clinical" or "true" endpoint, might be difficult to use in a clinical trial. Thus, use of the true endpoint in such cases might increase the complexity and or the duration of a clinical trial, especially when measurement of such endpoints are:

- costly – for example, expensive equipment for measuring nitrogen, potassium, and water contents in patient's body is required to diagnose "cachexia";
- difficult – for example, involving compound measures such as encountered in quality-of-life;
- requires a long follow-up-time – for example, survival in early-stage cancers;
- requires a large sample size due to a low incidence of the event – for example, short-term mortality in patients with suspected acute myocardial infarction.

To overcome these problems, a seemingly attractive solution is to replace the true endpoint by another one, which is measured earlier, more conveniently, or more frequently. Such "replacement" endpoints are termed "surrogate" endpoints (Ellenberg and Hamilton 1989). In many health care studies, the so-called biomarkers are used to examine organ functions or other aspects of health. An effective strategy is then proper selection and application of biomarkers for efficacy. From a regulatory per-

spective, a biomarker is considered acceptable for efficacy determination only after its establishment as a valid indicator of clinical benefit, i.e., after its validation as a surrogate marker (Burzykowski, Molenberghs, and Buyse 2005).

These considerations naturally lead to the need of proper definitions. An important step came from the Biomarker Definitions Working Group (2001), their definitions nowadays widely accepted and adopted. They defined the most credible indicator or drug response, a *clinical endpoint*, as 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. A *surrogate endpoint* is a biomarker that is intended to substitute for a clinical endpoint, which is expected to predict clinical benefit, harm, or lack thereof.

Biomarkers will continue to play an increasingly important role in improving the effectiveness of drug research and development. This is due to, but not limited to, the rapidly progressing understanding of the molecular basis of disease processes. Once a proposed biomarker has been validated, it can be used to diagnose disease risk, presence of disease in an individual, or to tailor treatments for the disease in an individual (choices of drug treatment or administration regimes). In evaluating potential drug therapies, a biomarker may be used as a surrogate for a clinical endpoint such as survival or irreversible morbidity.

1.3 Controversies, Motivation, and Validation of Surrogate Endpoints

Surrogate endpoints have been used in medical research for a long time, owing to the possible benefits for the duration of a clinical trial (Fleming and DeMets 1996,

Cardiac Arrhythmia Suppresion Trial 1989). However, in spite of the potential advantages, the use of surrogate endpoints in the development of new therapies has always been very controversial. This is partly owing to a number of unfortunate historical instances where treatments showing a highly positive effect on a surrogate endpoints were ultimately shown to be detrimental to the subjects clinical outcome, and conversely, some instances of treatments conferring clinical benefit without measurable impact on presumed surrogates (Fleming and DeMets, 1996).

The best known case is the approval by the Food and Drug Administration (FDA) of two antiarrhythmic drugs: encainide, and flecainide. The drugs reduced arrhythmia but caused a more than three fold increase in overall mortality, there by stressing the need for caution in using non-validated biomarkers in the evaluation of the possible clinical benefits of new drugs (Cardiac Arrhythmia Suppresion Trial 1989). Another example is the use of CD4 blood count as a surrogate endpoint for time to clinical events and overall survival (Lagakos and Hoth 1992), in spite of concern about its limitations as a surrogate marker for clinically relevant endpoints (DeGruttola et al. 1997).

The main reason behind failures was the incorrect perception that surrogacy simply follows from the association between a potential surrogate endpoint and the corresponding clinical endpoint. What is required to replace the clinical endpoint by the surrogate is that the effect of the treatment on the surrogate endpoint reliably predicts the effect on the clinical endpoint. This condition was not checked in the early attempts, consequently leading to negative opinions about the use of surrogates in the evaluation of treatment efficacy (Fleming 1994, Ferentz 2002).

Even with the controversies as mentioned above, the use of surrogate endpoints is still being considered due to some of the following reasons:

- An increasing number of new drugs have well-defined mechanisms of action at molecular level, allowing drug developers to measure the effect of these drugs on the relevant biomarkers (Lesko and Atkinson 2001).

- There is also increasing public pressure for fast approval of promising drugs, which will have to be based on biomarkers rather than on long-term, costly clinical endpoints (Dunn and Mann 1999). This is especially so for diseases that could become a serious threat to public health or the patients (quality of) life.

- The duration and sample size of clinical trials aimed at evaluating the therapeutic efficacy of new drugs are often insufficient to detect rare or late adverse effects (Jones 2001, Heise et al. 1997); using surrogate endpoints in this context might allow one to obtain information about such effects even during the clinical testing phase.

- Shortening the duration of clinical trials not only can decrease the cost of the evaluation process but also limit potential problems with noncompliance and missing data, which are more likely in longer studies (Burzykowksi, Molenberghs, and Buyse 2005, Verbeke and Molenberghs 2000).

- Another important point is the fact that regulatory agencies around the globe, in particular in the United States, in Europe and in Japan, have set mechanisms available for accelerated approval based on surrogate endpoints (Burzykowksi, Molenberghs, and Buyse 2005).

The use of surrogate endpoints is more accepted in early phases of clinical research. Using them to substitute for the clinical endpoint in all clinical research past a certain point is, however, a topic of ongoing debate. Molenberghs et al.(2008), argue that

one should be much more restraint using them as substitutes for the true endpoint in pivotal phase III trials, since the latter might imply replacing the true endpoint by a surrogate for all future studies as well, a far-reaching decision.

While the huge potential of surrogate endpoints to accelerate and improve the quality of clinical trials is unquestioned, the above considerations indicate that it is crucial to use validated surrogates (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, as well as practical and economic considerations. Consequently, the International Conference on Harmonisation (ICH) Guidelines on Statistical Principles for Clinical Trials states that "In practice the strength of the evidence for surrogacy depends upon (1) the biological plausibility of the relationship, (2) the demonstration in epidemiological studies of the prognostic value of the surrogate for the clinical outcome and (3) evidence from clinical trials that treatment effects on the surrogate correspond to effects on the clinical outcome" (International Conference on Harmonisation 1998). This work focuses on the statistical validation of surrogate endpoints, with a lot of emphasis on randomized clinical trials.

Several methods have been suggested for the formal evaluation of surrogate markers. Prentice (1982) proposed a formal definition of surrogate endpoints with a set of criteria for their validation. Freedman, Graubard, and Schatzkin (1992) supplemented Prentice's proposal with the estimation paradigm, by introducing measures, the *proportion explain*, which can be used to evaluate surrogate endpoints. Quantification allows us consider surrogate endpoints which are, in some sense, less than perfect but possibly strong enough to be able to still be of use. Buyse et al. (2000) further decomposed the proportion explained into the *relative effect* and *adjusted as-*

sociation and argued in favor of these quantities instead. These strategies are based on a single trial setting. Thus, although appealing they rest on strong and unverifiable assumptions, leading several authors (Daniels and Hughes 1997, Buyse et al. 2000) to propose methods based on information coming from several units or trials. Using hierarchical linear models, Buyse et al. (2000) defined surrogacy in terms of the quality of trial-level (R^2_{Trial}) and individual-level (R^2_{ind}) association between a potential surrogate and a clinical endpoint, both of a coefficient of determination type. A surrogate will be said to be good when both (R^2_{Trial}) and (R^2_{ind}) are sufficiently high. What is sufficiently high remains to be determined in collaboration with clinical and biopharmaceutical arguments.

Validation of surrogate endpoints within the meta-analytic frame work of Buyse et al. (2000) has been extended to settings involving non-Gaussian outcomes (Burzykowksi, Molenberghs, and Buyse 2005). In the present work, we will review and investigate the advantages and problems, especially computational issues, related with the meta-analytic approach, with a focus on Gaussian outcomes. Additionally, we will extend the said validation framework to a setting where the true endpoint is the ultimate assessment in a sequence of repeated measures, considering earlier measures as a potential surrogate endpoint.

In addition to the severe computational issues which may be encountered with the meta-approach, different settings have led to different measures at the individual-level. To overcome these limitations, Alonso and Molenberghs (2007) used information theory to create a unified framework for surrogate markers evaluation base on a measure of information, R^2_{h}. We investigate the performance of this approach and extend its use to settings consisting of non-Gaussian outcomes, with focus on event-time type outcomes.

1.4 Structure of the Thesis

This work is based on the development of marker methodology with focus on event-time type clinical endpoints. In particular, emphasis will be given to the statistical validation of surrogate endpoints. The general structure of the thesis is divided in two parts.

The first part consists of four chapters dedicate to a meta-analytic validation framework for surrogate markers. Chapter 2 introduces a basic set of notation to be used throughout the thesis, motivating case studies, as well as datasets used to illustrate some methods proposed in subsequent chapters. A concise history of surrogate marker evaluation is covered in Chapter 3, together with a meta-analytic approach proposed by Buyse et al.(2000) for continuous outcomes, and its extension to settings with mixtures of other types of outcomes. We investigate some computational issues related to the meta-analytic approach through simulation studies in Chapter 4. In Chapter 5, we review a meta-analytic approach validation method for two repeatedly measured outcomes (specifically longitudinal outcomes), and modify it to accommodate a mixture of longitudinal and cross-sectional outcomes. The modified version is then applied to a setting where the true endpoint is the ultimate assessment in a sequence of repeated measures, with subsequences of earlier measures as potential surrogate endpoints. This is done while carefully weighing the length and cost reducing potential against loss in precision and the risk of an inappropriate decision.

The second part of the thesis consists of six additional chapters; dedicated to a different approach to surrogacy based on information theory by Alonso and Molenberghs (2007). This approach leads to an intuitive interpretation, is applicable to a wide range of situations, and provides a unified framework as some of the previous proposals in the first part of the thesis follow as special cases of this approach. A

brief review of the information-theoretic approach (ITA) is the focus of Chapter **??**. In Chapter **??**, we investigate the performance of the information-theoretic approach in a setting with mixed continuous clinical and binary surrogate endpoints, through a simulation study. A limited discussion of its performance for two binary endpoints, as well as an application to a case study are also presented. In Chapter **??**, we extend the information-theoretic approach to settings with event-time type clinical endpoints. This extension hinges greatly on a proportional hazards assumption (Cox 1972). In Chapter **??**, we investigate the performance of the ITA in a setting with event-time clinical endpoint and a cross-sectional continuous surrogate, when the proportional hazards assumption is violated.

The motivation or objective for Chapter **??** is two-fold; firstly, it illustrates how the ITA can be employed in a setting with mixed event-time and cross-sectional endpoints. Secondly, it provides us with an idea of how robust the ITA may be to the PH assumption. However, cross-sectional biomarkers are not the most common biomarkers for event-time clinical endpoints. For example, let us consider a study in advanced colorectal cancer. While a researcher waits for the occurrence of the event of a clinical endpoint, for example death, it is customary to:

1. record the event of a biomarker, which is expected to occur before that of the clinical endpoint. For example, progression free survival. This leads to a setting with two event-time endpoints and is focus of Chapter **??**;

2. take repeated measurements of other biomarkers (e.g. tumor size), hence possible surrogate markers, over time. Thus, yielding a longitudinal surrogate for an event-time clinical endpoint. Applying the ITA to such a setting is the covered in Chapter **??**.

Finally, Chapter **??** presents some conclusions, recommendations, and perspectives on further research.

Table 1.1: A representation of the drug development process, with an indication of the sample size and or duration required at each stage. Column labels are: 'Pre-clinical' for the pre-clinical phase, 'Phases I, II, III, and IV' for phase I-IV clinical trials, and 'R.A.' for regulatory agency.

	Pre-clinical		**Phase I**	**Phase II**	**Phase III**		**R.A.**	**Phase IV**
Years	3.5 - 6.5		1 - 1.5	2	3 - 3.5		1.5 - 2.5	
Population	Laboratory and Animal Studies		20 - 80 healthy volunteers	100 - 300 patient volunteers	1,000 - 3,000 patient volunteers			
Purpose	Assess safety and biological activity	File IND with R.A.	Determine safety and dosage	Evaluate effectiveness, look for side effects	Confirm effectiveness, monitor adverse reactions for long term use	File NDA with R.A.	Review process / approval	Additional post-marketing testing
Success Rate	5,000 compounds evaluated			5 enter clinical trials			1 approved	

2

Notation and Motivating Studies

In this chapter, we begin by introducing a basic set of notation to be used throughout the thesis, as well as datasets that will be used to illustrate the various validation methods described in the subsequent chapters.

2.1 Notation

We adopt the following notation: T and S are random variables that respectively denote the clinical and surrogate endpoints, C the censoring times for event-time type endpoints, and Z is a binary indicator variable for the treatment. The fundamental settings considered corresponds to a multi-center trial or a meta-analysis of trials. Hence, the (T, S, Z) notation will be supplemented using indices $i = 1, \ldots, N$ for the ith center or trial, $j = 1, \ldots, n_i$, and to denote the jth subject enrolled in the ith

center or trial. Where appropriate, the indices $k = 1, \ldots, K$ will be used to denote the kth measurement on subject j in the ith center or trial, measured at time point t_{ik}. Some sections and or motivating studies are based on a single trial. The variables and indices maintain their respective description as mentioned earlier above.

2.2 Motivating Studies

This subsection provides description of data from randomized and non-randomized clinical trials in different therapeutic areas will be used throughout many of the chapters.

2.2.1 Ophthalmology: Age-related Macular Degeneration Trial

Age-related macular degeneration is a condition in which patients progressively lose vision. This is a multi-center clinical trial with 42 centers and a total of 190 patients with age-related macular degeneration (Pharmacological Therapy for Macular Degeneration Study Group 1997). Six out of the 42 trials enrolled patients only to one of the two treatment arms. Thus, only 36 centers were available for analysis result to a total of 183 patients, with a number of individual patients per center ranging from 2 to 18.

Patients with macular degeneration progressively lose vision. In the trial, the patients visual acuity was assessed at different time points through their ability to read lines of letters on standardized vision charts. These charts display lines of five letters of decreasing size, which the patient

Patients' visual acuity was assessed at different time points through their ability to read lines of letters on standardized vision charts. These charts display lines of

five letters of decreasing size, which patients had to read from top (largest letters) to bottom (smallest letters). Each line with at least four letters correctly read is called one 'line of vision. The visual acuity was measured by the total number of letters correctly read. The treatment indicator (Z) is set to 0 for placebo and to 1 for interferon-α. The surrogate endpoint S is the change in visual acuity at 6 months after starting treatment, while the true endpoint T is the change in visual acuity at 1 year. Both the continuous and binary versions of these endpoints will be considered in subsequent chapters. Appropriate details will be provided in the corresponding chapters. When treated as continuous variables, the endpoints will be assumed to follow a normal distribution.

2.2.2 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. The 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 a 7-grade scale used by the treating physician to

characterize how well a subject has improved since baseline. The Positive and Negative Syndrome Scale (PANSS) (Kay, Opler, and Lindenmayer 1988) is another useful and sufficiently sensitive assessment. It 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). Since the package insert in most countries recommends that risperidone is most effective at doses ranging from 4 t0 6 mg/day, only patients that received either these doses of risperidone or an active control are included in the dataset. Depending on the trial treatment was administered for a duration of 4 to 8 weeks and measurements were taken on weeks 1, 2, 3, 4, 6, and 8.

For this study, we will evaluate the change from baseline in PANSS as a surrogate endpoint for the change in CGI from baseline. The data contains five trials and in all trials, information is available on the investigators that treated the patients. Like the trials, the investigators will also be considered as the units of analysis. For most of the analysis applied to this dataset, we will restrict attention to the last observed scores during treatment. Additionally, the repeatedly measured PANSS outcome will be used in Chapter 5.

2.2.3 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.

2.2.4 Four Meta-analysis of 28 Clinical Trials in Advanced Colorectal Cancer

The data are from 28 advanced colorectal cancer trials (Advanced Colorectal Cancer Meta- Analysis Project, 1992, 1994; Meta-Analysis Group in Cancer, 1996, 1998). The individual patient data were collected by the Meta-Analysis Group in Cancer between 1990 and 1996 to obtain an overall quantitative assessment of the value of several experimental treatments in advanced colorectal cancer. In the four meta-analyses, the comparison was between an experimental treatment and a control treatment. The control treatments, referred to hereafter as FU bolus, were similar across the four meta-analyses and consisted of fluoropyrimidines (5FU or FUDR) given as a bolus intravenous injection. The experimental treatments, referred to hereunder as experimental FU, differed across the four meta-analyses and consisted of 5FU modulated by leucovorin (Advanced Colorectal Cancer Meta-Analysis Project, 1992), of 5FU modulated by methotrexate (Advanced Colorectal Cancer Meta-Analysis Project, 1994), of 5FU given in continuous infusion (Meta-Analysis Group in Cancer, 1998) and of hepatic arterial infusion of FUDR for patients with metastases confined to the liver (Meta-Analysis Group in Cancer, 1996). As noted by Daniels and Hughes (1997), the use of an experimental treatment that varies among the trials can be defended on the grounds of generalizability of the results of the validation process to future clinical

trials and treatments. The experimental treatments in our example might be considered as representatives of the modifications of the standard fluoropyrimidine-based regimen in advanced colorectal cancer.

Several of the 28 trials were multi-armed. In total, 33 randomized comparisons were considered in the four meta-analyses. Individual patient data were available for 27 of the comparisons (in 24 studies). From now on, we shall refer to each of the comparisons as a separate trial. The total size in the trials ranged from 15 to 382 patients. The true (T) and surrogate (S) endpoints will be survival time and tumor response, respectively. A binary version of S, indicating complete/partial response, as well as a categorical version with four categories (complete response, partial response, stable disease, progression)(World Health Organisation 1979), will be considered. The binary indicator for treatment (Z) will be set to 0 for FU bolus and 1 for experimental FU.

2.2.5 Advanced Colon Cancer: A Meta-analysis of 10 Clinical Trials

Most patients with metastatic colorectal cancer die as a result of their disease. The ultimate goal of chemotherapy is to cure the disease, or failing that, to improve patient symptoms, quality of life, and OS. It seems justified, therefore, to use OS to assess the efficacy of chemotherapies for advanced colorectal cancer. However, patient death can be observed only after prolonged follow-up, and with the increasing number of active compounds available in this disease, any effect of first-line therapies on OS may be confounded or diminished by the effects of subsequent therapies. It is therefore of interest to investigate whether progression free survival (PFS) could replace OS as the primary end point in randomized trials for the treatment of patients with advanced colorectal cancer (Buyse et al. 2007).

These data were used in a meta-analysis of 10 randomized trials in advanced colorectal cancer. Individual patient data were available from 7 trials comparing fluouracil (FU) + leucovorin with FU alone, with a total of 1,744 patients. Additionally, data for 1,345 patients from 3 trials comparing fluouracil (FU) + leucovorin with raltitrexed were available. These trials accrued patients between 1981 and 1990 with a median follow-up of 30.4 months (Buyse et al. 2007). Survival Analyses PFS and OS analyses were based on all randomly assigned patients using the intention-to-treat approach. PFS was defined as the time from random assignment to progressive disease (as assessed in each individual trial) or death from any cause. OS was defined as the time from random assignment to death from any cause.

2.2.6 Early Colon Cancer: A Meta-analysis of 11 Clinical Trials

Meta-Analysis Group In Cancer (MAGIC). Modulation of fluorouracil by leucovorin in patients with advanced colorectal cancer: an updated meta-analysis. Journal of Clinical Oncology 2004; 22: 376675.

2.2.7 Advanced Ovarian Cancer: A Meta-analysis of Four Clinical Trials

You have this data from Tomasz, therefore analyze it and add results to the appropriate section.

2.2.8 Stroke Study on Children with Sickle Cell Disease

The data results from a clinical trial involving 2323 children with sickle cell disease (SCD). Stroke is the second leading cause of death in children with SCD (Stroke Prevention Trial in Sickle Cell Anemia (STOP) trial). The brain and lungs are among the

organs susceptible to serious damage in SCD. Early detection of dysfunction may allow intervention to reduce risk of further damage. Transcranial Doppler ultrasonography (TCD), studies have been used extensively to evaluate children with SCD (Seibert et al. 1998). An abnormally high blood flow velocity by TCD in the middle cerebral or internal carotid arteries is associated with an increased risk of stroke.

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, TCD velocities are measured for each patient repeatedly over time, with the number of repeated measurements corresponding to the number of examinations a patient has undergone. The number of examinations per patient range from 1 to 13. About 36% and 0.04% of the total number of patients have 1 and 13 measurements, respectively. At each examination, blood flow velocities are measured for each of several arterial segments in the brain. Actually, multiple measurements are made in each segment but only the highest for a given segment is recorded, and used to evaluate stroke risk. The following blood velocities, as well as sensible functions of these velocities, were recorded and will be evaluated as surrogates for time-to-stroke:

- Maximum mean TCD velocity in qualifying segments (MaxVel);
- Maximum systolic TCD velocity in qualifying segments (MaxS);
- Maximum diastolic TCD velocity in qualifying segments (MaxD);
- Maximum mean TCD velocity on right in qualifying segments (MaxR);
- Maximum mean TCD velocity on left in qualifying segments (MaxL).

- Sum of maximum systolic and diastolic TCD velocities (MaxSD);
- Sum of maximum right and left TCD velocities (MaxRL).

In addition to the TCD velocity measures, the time to first stroke, which is the clinical endpoint, and the age of the patients are also recorded. The data contains information about two categories of patients. The first, consists of 301 children who were randomized to either receive a control arm or chronic transfusion arm. Only about 10% of them experienced stroke, i.e., 90% of the observations are censored. We refer to this data as the *Randomized* SCD dataset. The second category consists of 2193 children who were screened and only 4% of them experienced stroke. We henceforth refer to this category as the *Screened* SCD dataset. The randomization group of the children was also recorded for the randomized SCD data.

.

Part I

A Meta-analytic Validation Framework for Surrogate Markers

3

Surrogate Markers Validation

3.1 Brief History of Surrogate Endpoint Validation

One of the most important factors influencing the duration and complexity of the process of developing new treatment is the choice of the endpoint, which will be used to assess the efficacy of a treatment. In Chapter 1, we identified some conditions motivating the use of surrogate endpoints (Section 1.2) as well as controversies surrounding their use (Section 1.3). Thus, necessitating the use of validated surrogate endpoints. Several authors have argued that if a biomarker is to serve as a surrogate for a clinical endpoint, there should be a causal relationship between them (Lagakos and Hoth 1992, Fleming and DeMets 1996). Unfortunately, causality is generally extremely difficult to test for, and it ought to be understood that the statistical criteria,

developed developed to validate a surrogate marker, provide indirect evidence only about the causality of the relationship between the marker and the clinical endpoint.

A first source of evidence is provided by the association, at the level of the individual patient, between the marker and the clinical endpoint. One would expect a good surrogate endpoint to have a strong association with the clinical endpoint at the individual level, reflecting some biological pathway from the biomarker to the clinical endpoint. We will henceforth refer to this as *individual-level surrogacy*. However, a good correlate is not automatically a good surrogate (Fleming and DeMets 1996). Another source of evidence is needed to quantify the association, at the level of a trial, between the effects of a treatment on the marker and on the clinical endpoint. We will refer to this as *trial-level surrogacy*. As mentioned earlier, it is crucial to validate (or more realistically evaluate) a surrogate before using it to replace a clinical endpoint. Several formal methods for this purpose have been proposed (Prentice 1989, Freedman, Graubard, and Schatzkin 1992, Daniels and Hughes 1997, Buyse and Molenberghs 1998, Buyse et al. 2000, Gail et al. 2000).

Prentice (1989) proposed to define a surrogate endpoint as "a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint." Symbolically this definition can be written as

$$f(S|Z) = f(S) \iff f(T|Z) = f(T),$$

where $f(X)$ denotes the probability distribution of random variable X and $f(X|Z)$ denotes the probability distribution of X conditional on the value of Z. Prentice proposed four operational criteria to check if a triplet (T, S, Z) fulfills the definition (Prentice 1989, Burzykwoski et al. 2005). This definition involves the triplet (T, S, Z), hence the endpoint S is a surrogate for T only with respect to the effect of some specific

treatment Z, except if S were a *perfect* surrogate for T.

Freedman, Graubard, and Schatzkin (1992) supplemented Prentice's proposal with the estimation paradigm, by introducing measures which can be used to evaluate surrogate endpoints. They proposed the so-called *proportion explained*, $PE = 1 - \frac{\beta_S}{\beta}$, where β_S and β denote the treatment effect obtained from the models $f(T|S, Z)$ and $f(T|S)$ respectively. Quantification allows us consider surrogate endpoints which are, in some sense, less than perfect but possibly strong enough to be able to still be of use. Freedman, Graubard, and Schatzkin (1992) argue that one of Prenctice criterion raises conceptual problems since it requires the statistical test for the treatment to be nonsignificant after adjustment for the surrogate. Freedman proposed to calculate the proportion of the treatment effect on the true endpoint captured by the surrogate, PE. Freedman suggested that a good surrogate is one for which PE is close to one. However, Buyse and Molenberghs (1998) showed that PE can be decomposed into three different quantities: (1) the *relative effect* RE, which expresses the relationship between the treatment effects on the surrogate and the true endpoint at the trial level; (2) the *adjusted association* ρ_Z., which is a measure of association between the surrogate and the true endpoint at the individual level; and (3) a ratio of variances, which is a nuisance parameter. They then proposed to evaluate the individual- and trial-level surrogacy using adjusted association and relative effect, respectively.

Wide confidence intervals of RE are encountered in practice and a multiplicative assumption, which can not be verified using data from a single trial, is necessary to predict the treatment effect on T for a new trial (Buyse and Molenberghs 1998, Buyse et al. 2000, Molenberghs et al. 2000). It therefore seems more meaningful to view the validation problem from a hierarchical (or multilevel, or, meta-analytic) point of view. Many authors (Freedman, Graubard, and Schatzkin 1992, Daniels

and Hughes 1997, Albert et al. 1997, Buyse et al. 2000) propose methods based on information coming from several units or trials to increase the accuracy of the validation process. A first proposal, using a Bayesian approach, was given by Daniels and Hughs (1997). Buyse et al. (2000) extended these ideas using the theory of linear mixed-effects models. Gail et al. (2000) extended it further using generalized estimation methodology.

Validation within the meta-analytic frame work of Buyse et al. (2000) has been extended to non-normal settings, as well as to settings with mixed data-type endpoints (Burzykwoski et al. 2005). In what follows, we describe the approach as proposed by Buyse et al. (2000) and apply it to some motivating studies.

3.2 A Meta-analytic Validation Framework for Continuous Outcomes

Here, we discuss the foundations of the meta-analytic approach to the validation of surrogate endpoints. The surrogate and true endpoints are assumed to be jointly normally distributed. We assume to have data from N trials at our disposition, in the ith of which n_i subjects are enrolled. 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. The approach is based on a hierarchical two-stage model. Two distinct modeling strategies can be followed, based on a two-stage fixed-effects representation on the one hand and random effects on the other hand. The first stage, of the fixed-effects two-stage model, is based upon a fixed effects model:

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

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

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, and ε_{Sij} and ε_{Tij} are correlated error terms, assumed to be zero-mean normally distributed with covariance matrix

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

At the second stage, we assume

$$\begin{pmatrix} \mu_{Si} \\ \mu_{Ti} \\ \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} m_{Si} \\ m_{Ti} \\ a_i \\ b_i \end{pmatrix}, \tag{3.4}$$

where the second term on the right hand side of (3.4) is assumed to follow a zero-mean normal distribution with covariance matrix

$$D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ & d_{TT} & d_{Ta} & d_{Tb} \\ & & d_{aa} & d_{ab} \\ & & & d_{bb} \end{pmatrix}. \tag{3.5}$$

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

$$S_{ij} = \mu_S + m_{Si} + \alpha Z_{ij} + a_i Z_{ij} + \varepsilon_{Sij}, \tag{3.6}$$

$$T_{ij} = \mu_T + m_{Ti} + \beta Z_{ij} + b_i Z_{ij} + \varepsilon_{Tij}. \tag{3.7}$$

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 (3.5). The error terms ε_{Sij} and ε_{Tij} follow the same assumptions as in the fixed effects models (3.1)–(3.2).

Although the two-stage fixed-effects and the random-effects models rest on different assumptions about the nature of the experiments being analyzed, the two approaches yield discrepant results only in pathological situations. In this setting the two approaches are similar and the-stage procedure can be used to introduce random effects (Laird and Ware 1982, Verbeke and Molenberghs 2000). Pragmatic arguments may also guide the choice between random and fixed effects.

Buyse et al. (2000) argue that it is of interest to investigate how the treatment effect on the true endpoint can be predicted by the treatment effect on the surrogate, especially in a new trial $i = 0$. The authors observe that $(\beta + b_0|m_{S0}, a_0)$ follows a normal distribution with mean and variance

$$E\left(\beta + b_0|m_{S0}, a_0\right) = \beta + \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} \mu_{S0} - \mu_S \\ \alpha_0 - \alpha \end{pmatrix} \quad (3.8)$$

$$Var\left(\beta + b_0|m_{S0}, a_0\right) = d_{bb} - \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}. \quad (3.9)$$

Based on the idea that the conditional variance (3.9) of a "good" surrogate, at the trial-level, is close to zero, Buyse et al. (2000) proposed to assess surrogacy at the trial-level by the coefficient of determination:

$$R^2_{\text{trial}(f)} = R^2_{b_i|m_{Si},a_i} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}}. \quad (3.10)$$

This coefficient measures how precisely the effect of treatment on the true endpoint can be predicted, provided that the treatment effect on the surrogate endpoint has been observed in a new trial ($i = 0$). Coefficient (3.10) has two properties desirable for its interpretation. It 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 (3.6) and (3.7). The R^2 for the reduced models is then calculated as follows:

$$R^2_{\text{trial}(r)} = R^2_{b_i|a_i} = \frac{d^2_{ab}}{d_{aa}d_{bb}}. \tag{3.11}$$

Throughout this text, we will use R^2_{trial} to refer to $R^2_{\text{trial(f)}}$, unless otherwise stated. At the individual-level, to study how an individual's surrogate score is predictive for the true score, Buyse et al. (2000) defined the coefficient of determination based on (3.3) as

$$R^2_{\text{indiv}} = R^2_{\varepsilon_{Ti}|\varepsilon_{Si}} = \frac{\sigma^2_{ST}}{\sigma_{SS}\sigma_{TT}}. \tag{3.12}$$

R^2_{ind} is the squared correlation between both endpoints once we have adjusted for treatment and trial.

Following the developments reviewed above, a surrogate is termed *"trial-level valid"* if $R^2_{\text{trial(f)}}$ (or $R^2_{\text{trial(r)}}$) is sufficiently close to one, and *"individual-level valid"* if R^2_{ind} is sufficiently close to one. A surrogate is termed *"valid"* if it is both trial-level and individual-level valid. Even though the measures provide a quantification of surrogacy, there remains the important question as to how large is large. It is tough to provide hard guidance and, arguably, decisions will have to be taken based on a number of quantitative (statistical) and qualitative (clinical and biological) arguments combined.

3.3 Simplified Modeling Strategies

The hierarchical modeling discussed earlier is elegant, but it often poses considerable computational challenges, particularly when the number of trials or the trial sizes are

small (Burzykowski, Molenberghs, and Buyse 2005). In order to explore approximate strategies with better computational properties, Tibaldi et al. (2003) suggested several simplifications of the above strategy. These authors considered three possible dimensions along which simplifications can be made.

3.3.1 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)–(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

$$\widehat{\beta}_i = \widehat{\lambda}_0 + \widehat{\lambda}_1 \widehat{\mu}_{Si} + \widehat{\lambda}_2 \widehat{\alpha}_i + \varepsilon_i. \tag{3.13}$$

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 (3.12) using the estimates from (3.3). The second option is to consider the trial-specific effects as random. 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 (3.6)–(3.7) at the 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 3.2.

3.3.2 Endpoint dimension

Though natural to assume the two endpoints to be correlated, this can lead to computational difficulties in fitting the models. The need for the bivariate nature of the outcome is associated with R^2_{indiv}, which is in some cases of secondary importance. In addition, there is also a possibility to estimate it through the information-theoretic approach (Alonso and Molenberghs 2007), which is the focus of the second part of this thesis. Thus, further simplification can be achieved by fitting separate models for the true and surrogate endpoints, the so-called univariate approach.

If in the trial dimension, the trial-specific effects are considered to be fixed, then models (3.1)–(3.2) are fitted separately. Similarly, if the trial-specific effects are considered random, then models (3.6)–(3.7) are fitted separately, i.e., the corresponding error terms in the two models are assumed to be independent.

3.3.3 Measurement error dimension

When the univariate approach from the endpoint dimension and/or the fixed effects approach from the trial dimension are chosen, there is a need to adjust for the heterogeneity in information content between trial-specific contributions. One way to do so is weighting the contributions according to trial size. Thus, the researcher can either ignore this phenomenon or take it into account, giving rise to a weighted linear regression model (3.13) in the second stage.

3.4 Unit of Analysis

In addition to the convergence issues which Tibaldi et al. (2003) circumvented through the simplified modeling strategies, the choice of units also poses practical problems as many clinical trials usually consist of few trials and many centers and or investigators.

It is therefore imperative to extend the the framework to a three-level model, as well as assess the impact of omitting one of the levels in such a three-way hierarchy. Cortiñas et al. (2004) investigated several strategies to deal with these issues. The authors observed that the performance of the strategies depends on the sample sizes, as well as on the magnitude of variability present at different levels.

Thus, 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.

Based on results from a simulation study, Cortiñas et al. (2004) provided some justification to the use of, e.g., centers instead of trials as the units of analysis in practical applications of the meta-analytic approach to the validation of surrogate markers. The next chapter provides some of the other computational issues encountered when applying the meta-analytic approach, which had not been investigated prior to work on the thesis.

3.5 Application to a Motivating Study

Here, a motivating study in schizophrenia introduced in Chapter 2, Section (2.2.2), is analyzed using the meta-analytic approach and the simplified modeling strategy, while considering both trials and centers as the unit of analysis. Trial seems the natural unit of analysis. Unfortunately, the number of trials, 5, is not sufficient to apply the 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 been applied to this data set. The results are displayed in Table 4.5. 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 because 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 (see Chapter 4), 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 a 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.

3.6 Discussion

This chapter described an approach to provide a quantitative assessment of the value of a surrogate. we have briefly reviewed the history of surrogate markers validation, beginning with the definition of a surrogate and a testing paradigm, to the estimation paradigm and the meta-analytic view. We then went on to succinctly describe a meta-analytic approach for continuous outcomes as proposed by Buyse et al. (2000). It evaluates the "validity" of a surrogate in terms of coefficients of determination, which are intuitively appealing quantities in the unit interval.

Some practical issues, in particular the complexity of the required models and the choice of analysis, associated with the implementation of this approach were also reviewed. Fitting the models required within the meta-analytic framework by Buyse

et al. (2000) is not always a trivial task. Tibaldi et al. (2003) eluded this through the simplified modeling strategy, by considering simplifications along the trial, endpoint, and measurement error dimensions.

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.

A motivating case study in schizophrenia further highlighted the importance of considering the simplified modeling strategy, as well as the choice of analysis. While taking these points into account, we conclude based on the investigators as unit of analysis, that PANSS does not seem a good surrogate for the CGI.

.

.

Table 3.1: *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.*

4

Some Computational Issues in the Meta-analytic Approach

While we like to underscore the integrity of the meta-analytic framework of Buyse et al. (2000), important questions remain open, apart from those already addressed by Tibaldi et al. (2003) and Cortiñas et al. (2004) presented in the previous chapter. The use of complex hierarchical models implies that different surrogacy measures are proposed for different types of outcomes, especially at the individual level. The second part of this thesis focuses entirely on this issue. In addition, the models considered reflect practice within later phase clinical trials, in the sense that, apart from treatment assignment, no other explanatory information is used. It is conceivable, especially in early phase trials and in preclinical research, that more elaborate models be used, incorporating explanatory (baseline) covariates, biological, pharmacokinetic, or pharmacodynamic information. There are more practical issues such as, but not limited to, variable coding and model formulation.

In Section 4.1, we investigate the effect of treatment coding and ill-conditioned variance-covariance matrices, associated with the estimation of $R^2_{\text{trial(f)}}$, on the meta-analytic approach to surrogate markers validation. Section 4.2 deals with the impact of baseline covariates, on the meta-analytic approach to surrogate markers validation.

4.1 Treatment Coding and Ill-Conditioned Variance-covariance Matrices

In this section, we discuss results from two simulation studies performed to investigate the effect of the treatment coding and ill-conditioned D matrices on the meta-analytic approach to surrogate markers validation. The first study, which we henceforth refer to as *Study I*, investigates the relation between the treatment coding and ill-conditioned D matrices. The second study, *Study II*, investigates the difference in performance of the validation for both treatment coding, for which D is constrained to be positive definite.

4.1.1 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 connected to another 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 (3.6)–(3.7) 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.

4.1.2 Ill-conditioned Variance-covariance Matrix

When the full bivariate random effect is used, the R^2_{trial} is computed from the variance-covariance matrix (3.5). It is sometimes possible that this matrix be ill-conditioned and/or non-positive definite. In such cases, the resulting quantities ($R^2_{\text{trial(f)}}$ or $R^2_{\text{trial(r)}}$) computed based on this matrix might not be trustworthy. One way to asses the ill-conditioning of a matrix is by computing 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.

4.1.3 Simulation Study I

To asses the impact of using an incorrect treatment coding and its relation to ill-conditioned D matrices, a small simulation involving 12 different combinations of the number of trials and number of individuals per trial has been performed. The data

were generated based on the following model:

$$S_{ij} = 45 + m_{Si} + (3 + a_i)Z_{ij} + \varepsilon_{Sij}, \tag{4.1}$$

$$T_{ij} = 50 + m_{Ti} + (5 + b_i)Z_{ij} + \varepsilon_{Tij}. \tag{4.2}$$

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

$$D = \begin{pmatrix} 3 & 2.4 & 0 & 0 \\ 2.4 & 3 & 0 & 0 \\ 0 & 0 & 3 & 2.7 \\ 0 & 0 & 2.7 & 3 \end{pmatrix}. \tag{4.3}$$

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

$$\mathbf{\Sigma} = \begin{pmatrix} 3 & 2.4 \\ 2.4 & 3 \end{pmatrix}. \tag{4.4}$$

The chosen values for D and Σ yield $R^2_{\text{trial}} = 0.81$ and $R^2_{\text{indiv}} = 0.64$, respectively. 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 samples 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 4.3 and 4.4.

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 for the variance-covariance matrix.

4.1.4 Simulation Study II

The performance of the validation within the meta-analytic framework for both treatment coding, for which D is constrained to be positive definite, was assessed using percentage-bias as well as the percentage of samples that converge. A more extensive simulation, relative to *Study I*, involving different combinations of the number of trials, number of individuals per trial, trial-level and individual-level surrogacy was performed. The data were generated based on the models (4.1) and (4.2). However, the random effects (m_{Si}, m_{Ti} ,a_i , b_i) are assumed to be mean-zero normally distributed with covariance matrix

$$D = 3 * \begin{pmatrix} 1 & 0.8 & 0 & 0 \\ 0.8 & 1 & 0 & 0 \\ 0 & 0 & 1 & \sqrt{\rho_t} \\ 0 & 0 & \sqrt{\rho_t} & 1 \end{pmatrix}. \tag{4.5}$$

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

$$\mathbf{\Sigma} = 3 * \begin{pmatrix} 1 & \sqrt{\rho_i} \\ \sqrt{\rho_i} & 1 \end{pmatrix}. \tag{4.6}$$

The following values of the fixed parameters were used:

N = 10, 20, and 50 for each sample;

n_i = 10, 20, 60, and 100 for each trial;

ρ_t = 0.30, 0.50 and 0.90, resulting in trial-level surrogacy (R^2_{trial}) values of 0.30, 0.50 and 0.90. These values are assumed to reflect a 'poor,' 'moderate,' and 'good' surrogate at the trial level, respectively. Investigating the performance of the estimators under these conditions is essential, as a good estimator is expected to appropriately distinguish between such surrogates;

ρ_i = 0.30, 0.50 and 0.90, resulting in individual-level surrogacy (R^2_{indiv}) values of 0.30, 0.50 and 0.90, similarly to the trial-level surrogacy.

Full combination of the simulation parameters gave rise to a total of 135 simulation settings. Within each setting, 500 samples were generated. The samples were generated using both the 0/1 and $-1/1$ treatment coding and analyzed with the corresponding correct and the opposite (incorrect) coding. The percentage-bias was calculated as the difference between a corresponding estimate and the true value expressed as a percentage of the true value. The results of these simulations are displayed in Tables 4.5 – Table 4.8.

In general, the percentage of samples that converged with the D martix constrained to be positive-definite increases with increase in the number of trials and trial size, as earlier observed in *Study I*. When the data was generated based on the 0/1 coding,

the percentage of the samples that converged for the correct and incorrect analyses ranged between (31% and 100%) and (30% and 100%), respectively. The percentage of convergence ranged between (54% and 100%) and (50% and 100%) for the correct and incorrect analyses, respectively, when the samples are generated base on the $-1/1$ coding.

*** Figure around here, fully describe the axes in the caption ***

In more detail, on the one hand, when the samples were generated with the $0/1$ coding, the correct $(0/1)$ analysis had more converged samples than the incorrect $(-1/1)$ analysis in only 34 of the 135 settings. Figure *** shows the distribution of the difference in the number of converged samples between the correct and incorrect analyses, over all settings. Note that the highest difference in the number of converged samples within a setting is 15, either way. On the other hand, for samples generated with the $-1/1$ coding, the correct $(-1/1)$ analysis had more converged samples in all 135 settings (Figure ***). Note that in some settings, for example $(R^2_{\text{trial}} = 0.3,\ R^2_{\text{indiv}} = 0.3,\ N = 10,\ n_i = 10)$, the difference is greater than 200, which is more than 40% of the total number of samples in each setting. Therefore in terms of convergence, the $-1/1$ coding performs much better than the $0/1$ coding.

Additionally, for samples generated and analyzed with the $0/1$ coding, all settings with $(N = 10,\ R^2_{\text{trial}} = 0.9)$ had median condition number > 100 (range: 109– 359). While for samples analyzed with the $-1/1$ coding, all settings with $R^2_{\text{trial}} = 0.9$ had median condition number > 100 (range: 106– 519), irrespective of the values of the other simulation parameters. Samples generated and analyzed with the $-1/1$ all settings with $(N = 10,\ R^2_{\text{trial}} = 0.9)$ had median condition number > 100 (range: 100– 125). Also, for samples analyzed with the $0/1$ coding, all settings with $R^2_{\text{trial}} = 0.9$ had median condition number > 100 (range: 87– 217), irrespective of the values of

the other simulation parameters. Again, the $-1/1$ generally yields smaller condition numbers, leading to a well conditioned D, than the $0/1$ coding.

Focusing on the percentage-bias, very similar results are observed from the corresponding correct and incorrect analysis, for both the $0/1$ and $-1/1$ coding (Tables 4.5 – Table 4.8). The percentage-bias can be as high as 70% for small R^2_{trial} values (0.3), however, it is generally less than 20% for higher values of R^2_{trial}. Irrespective of the treatment coding and parameter settings considered in the simulation, the percentage-bias of estimates for R^2_{indiv} is generally less than 0.1%. This is expected as the treatment coding does not have substantial effect on Σ. Also, the smallest sample in the simulation has 100 subjects, which apparently may be enough to estimate R^2_{indiv}.

The results indicate that the treatment coding has a substantial impact on the convergence rate and the stability of the estimated D matrix. However, for samples that attain convergence with D constrained to be positive-definite, similar performance with respect to percentage-bias are observed. It is therefore advisable to use the $-1/1$ treatment coding when the full bivariate random effects model is used to evaluate a surrogate endpoint. Of course, it should be confirmed that this is in line with expert opinion for a given application. Additionally, it is essential that enough replication be available at each level of the hierarchical data, and the models should be fitted with D constrained to be *strictly* positive-definite.

4.2 Impact of Baseline Covariates

As earlier mentioned in Chapter 1 (Section 1.3), using surrogate endpoints to substitute for the clinical endpoint in all clinical research past a certain point is a topic of ongoing debate. The use of surrogate endpoint is more accepted in early phases of

clinical research. Furthermore, the models used in early phase trials and in preclinical research usually incorporate explanatory or baseline covariates. We dedicate this section to investigating the impact of a baseline binary predictor, say for example Gender, in a very simple setting where the effect of the baseline predictor is assumed to be similar on both endpoints. Also, we assume that there is no interactions between the binary baseline predictor and the treatment effects. Additionally, a similar but very limited simulation, with respect to the number of generated samples, was performed for a continuous baseline predictor.

The data were generated based on the following model:

$$S_{ij} = 45 + 3.2 * \text{Gender}_{ij} + m_{Si} + (3 + a_i)Z_{ij} + \varepsilon_{Sij}, \quad (4.7)$$

$$T_{ij} = 50 + 3.2 * \text{Gender}_{ij} + m_{Ti} + (5 + b_i)Z_{ij} + \varepsilon_{Tij}. \quad (4.8)$$

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) and the error terms (ε_{Sij}, ε_{Tij}) are both assumed to be mean-zero normally distributed with covariance matrices defined as in (4.5) and (4.6), respectively. Excepting the addition of the binary baseline predictor, as shown in (4.7) and (4.8), the simulation setting is similar to that of *Study II* in Section 4.1.4. Also, following the results obtained from *Study II*, only the $-1/1$ treatment coding was employed. The generated samples were then analyzed taking into account the effect of the baseline predictor, *Correct*, and ignoring the effect of the baseline predictor, *Ignored*. The *Correct*-analysis is based on (4.7) and (4.8), while the *Ignored*-analysis is based on (4.1) and (4.2). Results from the simulation are presented in Tables 4.9 and 4.10.

4.3 Application to a Motivating Study

The motivating study in age related macular degeneration introduced in Chapter 2, Section 2.2.1, is analyzed using the meta-analytic approach based on both the $-1/1$ and the $0/1$ coding of the treatment effect Z_{ij}. 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.

Table 4.1: *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*

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 4.1

Table 4.2: *ARMD data. Results of the trial-level (R^2_{trial}) surrogacy analysis* $0/1$ *coding. A — symbol indicates non-convergence.*

	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	—	

for the $-1/+1$ coding and in Table 4.2 for the $0/1$ coding. 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.

4.4 Discussion

This Chapter investigates some computational issues that may be encountered in practice, while employing the meta-analytic approach to surrogate markers for cross-sectional continuous outcomes. In practice, some researchers prefer the $0/1$ coding of 'dummy variables,' while others prefer the $-1/1$. The difference in preference

exacerbated by the fact that these choices lead to an equivalent model fit, for models with the 'dummy' occurring as fixed effects only. When the full-bivariate random-effect models (3.1) and (3.2) are used, it is relevant to reflect on the actual coding of the treatment variable. Indeed, results from simulation studies presented earlier indicate that in general, convergence is attained more with the $-1/1$ coding relative to the $0/1$ coding. Also, for models that converged, the $-1/1$ coding leads to well-conditioned D matrices, hence R^2_{trial}, relatively faster than the $0/1$ coding as the number of trials and trial size increase.

Furthermore, the use of surrogate endpoint is more accepted in early phase trials and preclinical research, which usual incorporate explanatory or baseline covariates. Because the models considered include the treatment assignment as the only explanatory information, it is necessary to have some idea of the possible impact of not including baseline covariates in the surrogate endpoint validation models. To this effect, a simulation study for a simple scenario; similar baseline covariate effect on both endpoints and no interactions, was performed. *** Results of Simulation *****

A motivating case study in Ophthalmology supplemented results from simulation indicating the relevance of the choice of the treatment coding. This may be a delicate issue as it is easy to overlook.

Table 4.3: *Study I: 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

.

Table 4.4: *Study I: 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

Table 4.5: *Study II: Simulation results for 0/1 treatment coding, for R^2_{indiv}=0.3 and R^2_{indiv}=0.5*

			$R^2_{\mathbf{indiv}} = 0.3$				$R^2_{\mathbf{indiv}} = 0.5$			
			Correct		**Incorrect**		**Correct**		**Incorrect**	
N	n_i	R^2_{trial}	$\widehat{R}^2_{\text{trial}}$	$\widehat{R}^2_{\text{indiv}}$	$\widehat{R}^2_{\text{trial}}$	$\widehat{R}^2_{\text{indiv}}$	$\widehat{R}^2_{\text{trial}}$	$\widehat{R}^2_{\text{indiv}}$	$\widehat{R}^2_{\text{trial}}$	$\widehat{R}^2_{\text{indiv}}$
10	10	0.3	0.683	0.030	0.670	0.033	0.733	0.010	0.703	0.006
20	10	0.3	0.333	0.017	0.343	0.017	0.337	0.002	0.337	0.002
50	10	0.3	0.140	0.000	0.140	0.000	0.117	0.000	0.117	0.000
10	20	0.3	0.497	0.013	0.490	0.010	0.520	0.002	0.513	0.002
20	20	0.3	0.293	-0.003	0.293	-0.003	0.270	-0.002	0.270	-0.002
50	20	0.3	0.097	0.003	0.097	0.003	0.087	0.002	0.087	0.002
10	40	0.3	0.420	0.003	0.417	0.003	0.420	0.000	0.427	0.000
20	40	0.3	0.250	0.003	0.250	0.003	0.240	0.000	0.240	0.000
50	40	0.3	0.080	0.003	0.080	0.003	0.073	0.002	0.073	0.002
10	60	0.3	0.417	0.007	0.420	0.007	0.407	0.002	0.407	0.002
20	60	0.3	0.237	0.003	0.237	0.003	0.230	0.000	0.230	0.000
50	60	0.3	0.073	0.003	0.073	0.003	0.070	0.002	0.070	0.002
10	10	0.5	0.196	0.037	0.188	0.047	0.242	0.016	0.226	0.014
20	10	0.5	0.116	0.027	0.116	0.023	0.128	0.004	0.122	0.004
50	10	0.5	0.062	0.000	0.062	0.000	0.046	0.000	0.046	0.000
10	20	0.5	0.150	0.013	0.162	0.017	0.168	0.002	0.170	0.002
20	20	0.5	0.120	-0.003	0.120	-0.003	0.106	-0.002	0.106	-0.002
50	20	0.5	0.038	0.003	0.038	0.003	0.030	0.002	0.030	0.002
10	40	0.5	0.134	0.003	0.128	0.003	0.134	0.000	0.128	0.000
20	40	0.5	0.104	0.003	0.104	0.003	0.096	0.000	0.096	0.000
50	40	0.5	0.028	0.003	0.028	0.003	0.024	0.002	0.024	0.002
10	60	0.5	0.146	0.007	0.146	0.007	0.134	0.004	0.134	0.004
20	60	0.5	0.094	0.003	0.094	0.003	0.088	0.000	0.088	0.000
50	60	0.5	0.026	0.003	0.026	0.003	0.022	0.002	0.022	0.002
10	10	0.9	-0.056	0.050	-0.070	0.067	-0.042	0.022	-0.049	0.026
20	10	0.9	-0.069	0.037	-0.070	0.027	-0.033	0.016	-0.036	0.012
50	10	0.9	-0.016	0.013	-0.020	0.013	-0.006	0.004	-0.006	0.004
10	20	0.9	-0.057	0.027	-0.050	0.027	-0.023	0.000	-0.032	-0.002
20	20	0.9	-0.020	0.010	-0.020	0.010	-0.006	0.002	-0.008	0.002
50	20	0.9	0.003	0.003	0.002	0.003	0.004	0.000	0.004	0.002
10	40	0.9	-0.011	0.017	-0.009	0.017	-0.006	0.004	-0.004	0.002
20	40	0.9	0.002	0.007	0.001	0.007	0.004	0.002	0.003	0.002
50	40	0.9	0.006	0.003	0.006	0.003	0.004	0.002	0.004	0.002
10	60	0.9	-0.006	0.013	-0.007	0.010	-0.001	0.006	-0.002	0.004
20	60	0.9	0.008	0.003	0.008	0.003	0.006	0.002	0.006	0.002
50	60	0.9	0.003	0.003	0.003	0.003	0.002	0.002	0.002	0.002

Table 4.6: *Study II: Simulation results for 0/1 treatment coding, for R^2_{indiv}=0.9*

			Correct		**Incorrect**	
N	n_i	R^2_{trial}	$\widehat{R}^2_{\text{trial}}$	$\widehat{R}^2_{\text{indiv}}$	$\widehat{R}^2_{\text{trial}}$	$\widehat{R}^2_{\text{indiv}}$
10	10	0.3	0.690	-0.002	0.680	-0.002
20	10	0.3	0.287	-0.001	0.290	-0.001
50	10	0.3	0.083	0.000	0.083	0.000
10	20	0.3	0.520	-0.001	0.507	-0.001
20	20	0.3	0.243	0.000	0.243	0.000
50	20	0.3	0.073	0.000	0.073	0.000
10	40	0.3	0.423	0.000	0.423	0.000
20	40	0.3	0.223	0.000	0.223	0.000
50	40	0.3	0.063	0.000	0.063	0.000
10	60	0.3	0.393	0.000	0.390	0.000
20	60	0.3	0.217	0.000	0.217	0.000
50	60	0.3	0.067	0.000	0.067	0.000
10	10	0.5	0.250	-0.002	0.230	-0.002
20	10	0.5	0.086	-0.001	0.088	0.000
50	10	0.5	0.020	0.000	0.018	0.000
10	20	0.5	0.156	-0.001	0.154	0.000
20	20	0.5	0.086	0.000	0.086	0.000
50	20	0.5	0.020	0.000	0.020	0.000
10	40	0.5	0.130	0.000	0.130	0.000
20	40	0.5	0.084	0.000	0.084	0.000
50	40	0.5	0.018	0.000	0.018	0.000
10	60	0.5	0.118	0.000	0.118	0.000
20	60	0.5	0.080	0.000	0.080	0.000
50	60	0.5	0.020	0.000	0.020	0.000
10	10	0.9	0.004	-0.001	0.000	-0.001
20	10	0.9	0.001	-0.001	0.000	-0.001
50	10	0.9	-0.002	0.000	-0.002	0.000
10	20	0.9	0.001	-0.001	-0.001	-0.001
20	20	0.9	-0.001	0.000	-0.001	0.000
50	20	0.9	-0.001	0.000	-0.001	0.000
10	40	0.9	0.001	0.000	0.001	0.000
20	40	0.9	0.000	0.000	0.000	0.000
50	40	0.9	0.001	0.000	0.001	0.000
10	60	0.9	-0.003	0.000	-0.003	0.000
20	60	0.9	0.000	0.000	0.000	0.000
50	60	0.9	0.000	0.000	0.000	0.000

Table 4.7: *Study II: Simulation results for −1/1 treatment coding, for R^2_{indiv}=0.3 and R^2_{indiv}=0.5*

			$R^2_{\mathbf{indiv}} = 0.3$				$R^2_{\mathbf{indiv}} = 0.5$			
			Correct		**Incorrect**		**Correct**		**Incorrect**	
N	n_i	R^2_{trial}	$\widehat{R}^2_{\text{trial}}$	$\widehat{R}^2_{\text{indiv}}$	$\widehat{R}^2_{\text{trial}}$	$\widehat{R}^2_{\text{indiv}}$	$\widehat{R}^2_{\text{trial}}$	$\widehat{R}^2_{\text{indiv}}$	$\widehat{R}^2_{\text{trial}}$	$\widehat{R}^2_{\text{indiv}}$
10	10	0.3	0.520	0.007	0.513	0.010	0.490	-0.004	0.493	-0.004
20	10	0.3	0.240	0.007	0.240	0.007	0.230	0.000	0.230	0.000
50	10	0.3	0.083	0.000	0.083	0.000	0.080	0.000	0.080	0.000
10	20	0.3	0.403	0.003	0.403	0.003	0.403	0.000	0.403	0.000
20	20	0.3	0.233	0.000	0.233	0.000	0.227	-0.002	0.227	-0.002
50	20	0.3	0.073	0.003	0.073	0.003	0.070	0.002	0.070	0.002
10	40	0.3	0.377	0.000	0.377	0.000	0.377	-0.002	0.377	-0.002
20	40	0.3	0.227	0.003	0.227	0.003	0.223	0.000	0.223	0.000
50	40	0.3	0.067	0.003	0.067	0.003	0.067	0.002	0.067	0.002
10	60	0.3	0.367	0.007	0.367	0.007	0.363	0.002	0.363	0.002
20	60	0.3	0.217	0.003	0.217	0.003	0.213	0.000	0.213	0.000
50	60	0.3	0.067	0.003	0.067	0.003	0.067	0.002	0.067	0.002
10	10	0.5	0.162	0.010	0.166	0.010	0.162	-0.004	0.156	-0.004
20	10	0.5	0.098	0.007	0.098	0.007	0.090	0.000	0.090	0.000
50	10	0.5	0.030	0.000	0.030	0.000	0.028	0.000	0.028	0.000
10	20	0.5	0.132	0.007	0.130	0.007	0.126	0.000	0.126	0.000
20	20	0.5	0.094	0.000	0.094	0.000	0.090	-0.002	0.090	-0.002
50	20	0.5	0.024	0.003	0.024	0.003	0.022	0.002	0.022	0.002
10	40	0.5	0.118	0.000	0.118	0.000	0.118	-0.002	0.118	-0.002
20	40	0.5	0.088	0.003	0.088	0.003	0.086	0.000	0.086	0.000
50	40	0.5	0.022	0.003	0.022	0.003	0.020	0.002	0.020	0.002
10	60	0.5	0.116	0.007	0.116	0.007	0.112	0.002	0.112	0.002
20	60	0.5	0.084	0.003	0.084	0.003	0.082	0.000	0.082	0.000
50	60	0.5	0.020	0.003	0.020	0.003	0.020	0.002	0.020	0.002
10	10	0.9	-0.016	0.033	-0.020	0.047	-0.010	0.006	-0.011	0.010
20	10	0.9	-0.002	0.013	-0.002	0.010	0.001	0.000	0.001	0.002
50	10	0.9	0.003	0.000	0.003	0.000	0.001	0.000	0.001	0.000
10	20	0.9	-0.004	0.007	-0.006	0.007	-0.003	0.000	-0.004	0.000
20	20	0.9	0.004	0.000	0.004	0.000	0.003	-0.002	0.003	-0.002
50	20	0.9	0.001	0.003	0.001	0.003	0.001	0.002	0.001	0.002
10	40	0.9	0.001	0.003	0.001	0.003	0.001	0.000	0.001	0.000
20	40	0.9	0.002	0.003	0.002	0.003	0.001	0.000	0.001	0.000
50	40	0.9	0.002	0.003	0.002	0.003	0.001	0.002	0.001	0.002
10	60	0.9	0.000	0.007	0.000	0.007	-0.001	0.004	-0.001	0.004
20	60	0.9	0.002	0.003	0.002	0.003	0.001	0.000	0.001	0.000
50	60	0.9	0.001	0.003	0.001	0.003	0.001	0.002	0.001	0.002

Table 4.8: *Study II: Simulation results for −1/1 treatment coding, for R^2_{indiv}=0.9*

			Correct		**Incorrect**	
N	n_i	R^2_{trial}	$\widehat{R}^2_{\text{trial}}$	$\widehat{R}^2_{\text{indiv}}$	$\widehat{R}^2_{\text{trial}}$	$\widehat{R}^2_{\text{indiv}}$
10	10	0.3	0.473	-0.002	0.473	-0.002
20	10	0.3	0.213	-0.001	0.213	-0.001
50	10	0.3	0.070	0.000	0.070	0.000
10	20	0.3	0.390	-0.001	0.390	-0.001
20	20	0.3	0.220	0.000	0.220	0.000
50	20	0.3	0.067	0.000	0.067	0.000
10	40	0.3	0.370	0.000	0.370	0.000
20	40	0.3	0.220	0.000	0.220	0.000
50	40	0.3	0.063	0.000	0.063	0.000
10	60	0.3	0.357	0.000	0.357	0.000
20	60	0.3	0.210	0.000	0.210	0.000
50	60	0.3	0.063	0.000	0.063	0.000
10	10	0.5	0.140	-0.002	0.142	-0.002
20	10	0.5	0.076	-0.001	0.076	-0.001
50	10	0.5	0.020	0.000	0.020	0.000
10	20	0.5	0.116	-0.001	0.116	-0.001
20	20	0.5	0.084	0.000	0.084	0.000
50	20	0.5	0.018	0.000	0.018	0.000
10	40	0.5	0.112	0.000	0.112	0.000
20	40	0.5	0.084	0.000	0.084	0.000
50	40	0.5	0.018	0.000	0.018	0.000
10	60	0.5	0.108	0.000	0.108	0.000
20	60	0.5	0.080	0.000	0.080	0.000
50	60	0.5	0.018	0.000	0.018	0.000
10	10	0.9	-0.002	-0.002	-0.001	-0.002
20	10	0.9	-0.002	-0.001	-0.002	-0.001
50	10	0.9	-0.001	0.000	-0.001	0.000
10	20	0.9	-0.006	-0.001	-0.006	-0.001
20	20	0.9	-0.001	0.000	-0.001	0.000
50	20	0.9	-0.001	0.000	-0.001	0.000
10	40	0.9	-0.002	0.000	-0.002	0.000
20	40	0.9	0.000	0.000	0.000	0.000
50	40	0.9	0.000	0.000	0.000	0.000
10	60	0.9	-0.004	0.000	-0.004	0.000
20	60	0.9	-0.001	0.000	-0.001	0.000
50	60	0.9	0.000	0.000	0.000	0.000

Table 4.9: *Simulation results for the impact of a binary baseline predictor with similar effect on both endpoints, for R^2_{indiv}=0.3 and R^2_{indiv}=0.5*

			$R^2_{\mathbf{indiv}} = 0.3$				$R^2_{\mathbf{indiv}} = 0.5$			
			Correct		**Ignored**		**Correct**		**Ignored**	
N	n_i	R^2_{trial}	$\widehat{R}^2_{\text{trial}}$	$\widehat{R}^2_{\text{indiv}}$	$\widehat{R}^2_{\text{trial}}$	$\widehat{R}^2_{\text{indiv}}$	$\widehat{R}^2_{\text{trial}}$	$\widehat{R}^2_{\text{indiv}}$	$\widehat{R}^2_{\text{trial}}$	$\widehat{R}^2_{\text{indiv}}$
10	10	0.3	0.507	0.010	0.520	0.007	0.493	-0.004	0.490	-0.004
20	10	0.3	0.243	0.007	0.240	0.007	0.230	0.000	0.230	0.000
50	10	0.3	0.083	0.000	0.083	0.000	0.077	0.000	0.080	0.000
10	20	0.3	0.407	0.003	0.403	0.003	0.400	0.000	0.403	0.000
20	20	0.3	0.233	0.000	0.233	0.000	0.227	-0.002	0.227	-0.002
50	20	0.3	0.073	0.003	0.073	0.003	0.070	0.002	0.070	0.002
10	40	0.3	0.377	0.000	0.377	0.000	0.377	-0.002	0.377	-0.002
20	40	0.3	0.227	0.003	0.227	0.003	0.223	0.000	0.223	0.000
50	40	0.3	0.067	0.003	0.067	0.003	0.067	0.002	0.067	0.002
10	60	0.3	0.367	0.007	0.367	0.007	0.363	0.002	0.363	0.002
20	60	0.3	0.217	0.003	0.217	0.003	0.213	0.000	0.213	0.000
50	60	0.3	0.067	0.003	0.067	0.003	0.067	0.002	0.067	0.002
10	10	0.5	0.162	0.013	0.162	0.010	0.158	-0.004	0.162	-0.004
20	10	0.5	0.098	0.007	0.098	0.007	0.090	0.000	0.090	0.000
50	10	0.5	0.030	0.000	0.030	0.000	0.026	0.000	0.028	0.000
10	20	0.5	0.130	0.007	0.132	0.007	0.126	0.000	0.126	0.000
20	20	0.5	0.094	0.000	0.094	0.000	0.090	-0.002	0.090	-0.002
50	20	0.5	0.024	0.003	0.024	0.003	0.022	0.002	0.022	0.002
10	40	0.5	0.116	0.000	0.118	0.000	0.116	-0.002	0.118	-0.002
20	40	0.5	0.088	0.003	0.088	0.003	0.086	0.000	0.086	0.000
50	40	0.5	0.022	0.003	0.022	0.003	0.020	0.002	0.020	0.002
10	60	0.5	0.116	0.007	0.116	0.007	0.112	0.002	0.112	0.002
20	60	0.5	0.084	0.003	0.084	0.003	0.082	0.000	0.082	0.000
50	60	0.5	0.020	0.003	0.020	0.003	0.020	0.002	0.020	0.002
10	10	0.9	-0.016	0.033	-0.016	0.033	-0.010	0.010	-0.010	0.006
20	10	0.9	-0.002	0.010	-0.002	0.013	0.001	0.000	0.001	0.000
50	10	0.9	0.003	0.000	0.003	0.000	0.001	0.000	0.001	0.000
10	20	0.9	-0.006	0.010	-0.004	0.007	-0.003	0.000	-0.003	0.000
20	20	0.9	0.004	0.000	0.004	0.000	0.002	-0.002	0.003	-0.002
50	20	0.9	0.001	0.003	0.001	0.003	0.001	0.002	0.001	0.002
10	40	0.9	0.001	0.003	0.001	0.003	0.001	0.000	0.001	0.000
20	40	0.9	0.002	0.003	0.002	0.003	0.001	0.000	0.001	0.000
50	40	0.9	0.002	0.003	0.002	0.003	0.001	0.002	0.001	0.002
10	60	0.9	0.000	0.007	0.000	0.007	-0.001	0.004	-0.001	0.004
20	60	0.9	0.002	0.003	0.002	0.003	0.001	0.000	0.001	0.000
50	60	0.9	0.001	0.003	0.001	0.003	0.001	0.002	0.001	0.002

Table 4.10: *Simulation results for the impact of a binary baseline predictor with similar effect on both endpoints, for R^2_{indiv}=0.9*

				Correct		**Ignored**	
N	n_i	R^2_{trial}		$\widehat{R}^2_{\text{trial}}$	$\widehat{R}^2_{\text{indiv}}$	$\widehat{R}^2_{\text{trial}}$	$\widehat{R}^2_{\text{indiv}}$
10	10	0.3	0.9	0.473	-0.002	0.473	-0.002
20	10	0.3	0.9	0.213	-0.001	0.213	-0.001
50	10	0.3	0.9	0.070	0.000	0.070	0.000
10	20	0.3	0.9	0.390	-0.001	0.390	-0.001
20	20	0.3	0.9	0.220	0.000	0.220	0.000
50	20	0.3	0.9	0.067	0.000	0.067	0.000
10	40	0.3	0.9	0.370	0.000	0.370	0.000
20	40	0.3	0.9	0.220	0.000	0.220	0.000
50	40	0.3	0.9	0.063	0.000	0.063	0.000
10	60	0.3	0.9	0.357	0.000	0.357	0.000
20	60	0.3	0.9	0.210	0.000	0.210	0.000
50	60	0.3	0.9	0.063	0.000	0.063	0.000
10	10	0.5	0.9	0.142	-0.002	0.140	-0.002
20	10	0.5	0.9	0.076	-0.001	0.076	-0.001
50	10	0.5	0.9	0.020	0.000	0.020	0.000
10	20	0.5	0.9	0.114	-0.001	0.116	-0.001
20	20	0.5	0.9	0.082	0.000	0.084	0.000
50	20	0.5	0.9	0.020	0.000	0.018	0.000
10	40	0.5	0.9	0.112	0.000	0.112	0.000
20	40	0.5	0.9	0.084	0.000	0.084	0.000
50	40	0.5	0.9	0.018	0.000	0.018	0.000
10	60	0.5	0.9	0.108	0.000	0.108	0.000
20	60	0.5	0.9	0.080	0.000	0.080	0.000
50	60	0.5	0.9	0.018	0.000	0.018	0.000
10	10	0.9	0.9	-0.002	-0.002	-0.002	-0.002
20	10	0.9	0.9	-0.002	-0.001	-0.002	-0.001
50	10	0.9	0.9	-0.001	0.000	-0.001	0.000
10	20	0.9	0.9	-0.004	-0.001	-0.006	-0.001
20	20	0.9	0.9	-0.001	0.000	-0.001	0.000
50	20	0.9	0.9	-0.001	0.000	-0.001	0.000
10	40	0.9	0.9	-0.002	0.000	-0.002	0.000
20	40	0.9	0.9	0.000	0.000	0.000	0.000
50	40	0.9	0.9	0.000	0.000	0.000	0.000
10	60	0.9	0.9	-0.004	0.000	-0.004	0.000
20	60	0.9	0.9	-0.001	0.000	-0.001	0.000
50	60	0.9	0.9	0.000	0.000	0.000	0.000

5

Earlier Measures in a Longitudinal Sequence as Potential Surrogate for a Later One

5.1 Introduction

Thus far, the previous chapters have focused on the methodologically appealing case of cross-sectional normally distributed endpoints. In many practical applications, situations abound where repeated measurements are encountered on either or both endpoints. Repeated measures of a quantitative (bio)marker are nowadays 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 measurement or status of a patient, then the (bio)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 (bio)marker may lead to reduction of the study's length and/or cost.

Going to a fully multivariate framework presents new challenges, as the R^2 measures introduced in Chapter 3 are no longer applicable. In the cross-sectional cases, one assumes that only one potential surrogate is available and that treatment effect on both responses is constant. These assumptions can fail when a patient is measured repeatedly over time. Alonso et al. (2003, 2004) extended the work by Buyse et al. (2000) to a setting where both the surrogate and true endpoints are measured repeatedly over time, using models for bivariate longitudinal data. Their proposal also enables us to evaluate surrogacy when more than one surrogate variable is available for the analysis. It serves as a basis for the work presented in this chapter. Additionally, surrogate-marker evaluation endeavors that have been performed thus far involved two different endpoints (Buyse *et al.* 2000, Burzykowski, Molenberghs, and Buyse 2005), 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 time-to-event) or of a mixed nature (e.g., an ordinal surrogate, such as tumor response, for a time-to-event endpoint, such as overall survival).

In contrast, the scenario under investigation 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 patient, 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 and reduces to the scenario discussed in Chapter 3.

The challenge is to determine the number of repeated measures that are required to sufficiently 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 of 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.

Section 5.2, presents a canonical correlation approach for two repeatedly measured endpoints and its modification to the scenario of interest. Section 5.3 provides, from a theoretical point of view, the performance of an objective function for two important special cases, compound symmetric structure and first-order auto-regressive process. Section **??** provides details on the design and results of our simulation study, and provide a perspective on the conclusions that can be drawn from it. In Section 5.5, we briefly introduce a constrained maximization problem. Section 5.7 contains the results of the case studies' analysis. It is not unusual for some measurement sequences in longitudinal study to terminate early for reasons outside the control of the investigator. Also, intermediate scheduled measurements might be missed. Thus, in Section 5.6, a limited simulation study was performed to get an idea of the impact of missingness on the scenarios considered in this chapter. It should be noted that focus is

on individual level surrogacy.

5.2 Longitudinal Endpoints and Surrogacy

In Section 5.2.1, we present a concise description of the meta-analytic approach to surrogate marker evaluation for repeated measurements, using canonical correlations as proposed by Alonso et al. (2004). This will be followed in Section 5.2.2 by a modification to the scenario where early measurements on a longitudinal endpoints are treated as a surrogate for the final measurement. In Section 5.2.3, the modified version will be applied in the determination of an optimal number of repeated measurements required to accurately predict the true endpoint. We zoom in on the development of an objective function in Section 5.2.4.

We shall assume that information from $i = 1, \ldots, N$ trials is available, in the i^{th} of which, $j = 1, \ldots, n_i$ subjects are included. We shall further denote the time points at which each subject in trial i is measured as t_{ik}. Let T_{ijk} and S_{ijk} denote the associated true and surrogate endpoints at time k, respectively, and Z_{ij} is a binary indicator variable for treatment.

5.2.1 Two Repeatedly Measured Endpoints

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 these methods can be adapted to the situation where one of the two outcomes is cross-sectional. The authors based the calculation of surrogacy measures on a two-stage approach (Section 3.3.1), rather than a full random-effects approach (Section 3.2), which would take into account both the repeated measures and the multi-trial nature of the data, in order to reduce numerical complexity. Additionally, the two-stage approach has shown good performance in both statistical and compu-

tational terms (Burzykowski et al. 2005). Precisely, following the ideas of Galecki (1994), Alonso et al. (2003) considered the following joint model at the first stage for the true and surrogate endpoints:

$$\begin{aligned} T_{ijk} &= \mu_{T_i} + \beta_i Z_{ij} + f(t_{ijk}) + \varepsilon_{Tijk}, \\ S_{ijk} &= \mu_{S_i} + \alpha_i Z_{ij} + f(t_{ijk}) + \varepsilon_{Sijk}, \end{aligned} \tag{5.1}$$

where $(\mu_{T_i}, \mu_{S_i}, \beta_i, \alpha_i)$ are intercepts and treatment effects on the true and surrogate endpoints, respectively, $f(t_{ijk})$ is a 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. Furthermore, note that even though in practice T_{ij} and S_{ij} are frequently measured at the same time points, model (5.1) would let us approach situations in which this condition does not hold. The error terms $(\varepsilon_{Tijk}, \varepsilon_{Sijk})$ are assumed to follow a zero-mean normal distribution with patterned variance-covariance matrix

$$\Sigma_i = \begin{pmatrix} \Sigma_{TTi} & \Sigma_{TSi} \\ \Sigma_{STi} & \Sigma_{SSi} \end{pmatrix}, \tag{5.2}$$

where Σ_{TT_i} and Σ_{SS_i} denote the variance-covariance matrices associated with the true and the surrogate endpoints, respectively, and $\Sigma_{TS_i} = \Sigma^T_{ST_i}$ contains the covariances between the measurements for the true and the surrogate endpoints. In some practical settings, Σ_i can be modeled as the Kronecker product of a general correlation matrix that captures the association within the sequences and an unstructured 2×2 matrix that captures the association between the sequences (Galecki 1994).

Due to the longitudinal nature of the endpoints, Alonso et al. (2004) extended the ideas of Buyse et al. (2000) for capturing individual-level surrogacy, based on coefficients of determination, to a multivariate version using the concept of canonical

correlation. Based on model (5.2), these authors obtained the canonical correlations, ρ_{ik} from $\Sigma_{TT_i}^{-1}\,\Sigma_{TS_i}\Sigma_{SS_i}^{-1}\Sigma_{TS_i}^{T}$ and proposed a family of measures to evaluate surrogacy at the individual level. This so-called Ω family is defined as

$$\Omega = \left\{ \vartheta : \vartheta = \sum_i \sum_k \alpha_{ik}\rho_{ik}^2, \quad \text{where:} \quad \alpha_{ik} > 0 \quad \forall (i,k), \quad \sum_i \sum_k \alpha_{ik} = 1 \right\}. \quad (5.3)$$

An important member of the Ω family is the *Variance Reduction Factor* (VRF) originally introduced by Alonso et al. (2003) and defined as

$$VRF_{\text{ind}} = \frac{\sum_i \{\text{tr}(\Sigma_{TT_i}) - \text{tr}(\Sigma_{T|S_i})\}}{\sum_i \text{tr}(\Sigma_{TT_i})}, \quad (5.4)$$

where $\Sigma_{T|S_i}$ denotes the conditional variance-covariance matrix of $\varepsilon_{T_{ij}}$ given $\varepsilon_{S_{ij}}$, i.e. $\Sigma_{T|S} = \Sigma_{TT} - \Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}$. Henceforth, the index i maybe dropped from the notation, for simplicity, when particular attention is given to the individual-level surrogacy. 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 (5.4), 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 (5.1) as follows:

$$R^2_\Lambda = 1 - \frac{|\Sigma|}{|\Sigma_{TT}| \cdot |\Sigma_{SS}|}. \tag{5.5}$$

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 one 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. In all cases, VRF or R^2_Λ estimates close to one are indicative of 'good' surrogacy at the individual-level, with the reverse holding for values close to zero.

At the second stage, the trial-level surrogacy can be estimated by regressing the estimated treatment effect on the true endpoint on the treatment effect on the sur-

rogate, as described in Section 3.3.1. Furthermore, it has been shown (Burzykwoski et al. 2005) that the VRF can be adequately applied at the trial level when (1) the treatment effect cannot be assumed constant over time and (2) the prediction of the treatment effect on the true endpoint can be substantially improved by using information about the treatment effect on an entire set of possibly relevant variables at the same time.

5.2.2 A Longitudinal Surrogate for a Cross-sectional True Endpoint

We now consider a setting where the surrogate endpoint is repeatedly measured over time with K repeated measures and that the true endpoint is cross-sectional. Model (5.1) then takes the form:

$$\begin{aligned} T_j &= \mu_T^* + \alpha^* Z_j + \varepsilon_{Tj}, \\ S_{jk} &= \mu_S^* + \beta^* Z_j + f(t_{jk}) + \varepsilon_{Sjk}. \end{aligned} \tag{5.6}$$

There are some important differences between (5.6) and the joint model for two longitudinal outcomes given in (5.1). First, there is a difference in the number of parameters when modeling the surrogate and true endpoints. Secondly, computational issue is induced as the variance-covariance matrix of the error term $(\varepsilon_{S_{jk}}, \varepsilon_{T_j})^T$, 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.,

$$\Sigma = \begin{pmatrix} \sigma_{TT} & \Sigma_{TS} \\ \Sigma_{ST} & \Sigma_{SS} \end{pmatrix}, \tag{5.7}$$

where 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. From (5.7), the VRF_{indiv} for longitudinal surrogate and a cross-sectional true endpoint can be computed as

$$VRF_{\text{indiv}} = \frac{\text{tr}(\sigma_{TT}) - \text{tr}(\sigma_{T|S})}{\text{tr}(\sigma_{TT})}. \tag{5.8}$$

Here, $\sigma_{T|S}$ denotes the conditional variance of T given S: $\sigma_{T|S} = \sigma_{TT} - \Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}$. Using this expression, (5.8) can be re-written as

$$VRF_{\text{indiv}} = \frac{\text{tr}(\sigma_{TT}) - \text{tr}(\sigma_{TT} - \Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST})}{\text{tr}(\sigma_{TT})}. \tag{5.9}$$

Note that all matrices involved in the computation of VRF_{ind} are of dimension (1×1) and hence the trace reduces to the corresponding scalar, offering the opportunity to simplify (5.9):

$$VRF_{\text{indiv}} = \frac{\sigma_{TT} - \sigma_{TT} + \Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}}{\sigma_{TT}} = \frac{\Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}}{\sigma_{TT}}. \tag{5.10}$$

Note that $VRF_{\text{indiv}} = 0$ if and only if $\Sigma_{ST} = 0$, i.e., if and only if S and T are independent. Intuitively, (5.10) 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 (5.6) and the corresponding variance-covariance matrix (5.7). The R^2_{Λ} for a longitudinal surrogate and a cross-sectional endpoint is given by

$$R^2_{\Lambda,\, ST} = 1 - \frac{|\Sigma\,|}{|\sigma_{TT}\,| \cdot |\Sigma_{SS}\,|}, \tag{5.11}$$

where σ_{TT}, Σ_{SS}, and Σ are as defined in (5.7). Note that

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

and, substituting this in (5.11), we obtain

$$\begin{aligned} R^2_{\Lambda_{ST}} &= 1 - \frac{|\sigma_{TT} - \Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}|}{|\sigma_{TT}|} \\ &= 1 - \frac{\sigma_{TT} - \Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}}{\sigma_{TT}} \\ &= \frac{\Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}}{\sigma_{TT}}, \end{aligned} \tag{5.12}$$

since all matrices involved are of dimension one. It is then obvious from (5.12) and (5.10) that R^2_Λ and VRF_{ind} are equal for a continuous longitudinal surrogate and a cross-sectional continuous true endpoint. Henceforth, we shall focus on VRF_{indiv} only in this chapter.

5.2.3 The Optimal Number of Repeated Measurements

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

$$Y_{ijk} = (\beta_0 + b_{1i}) + (\beta_1 + b_{2i})Z_{ij} + \beta_2 t_{ik} + \beta_3 Z_{ij} t_{ik} + \varepsilon_{ijk}, \tag{5.13}$$

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.

In many applications, assuming a constant treatment effect over time will be unrealistic. We assume a linear treatment effect over time, constant across trial, but

extension of which is straightforward (Alonso et al. 2004). Let us formally define our surrogate and true endpoints, based on (5.13). 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 (5.13). This leads to model (5.6) and its corresponding variance covariance matrix (5.7), from which we can compute the measure of surrogacy of the initial m measures for the final outcome using either (5.10) or (5.12).

The challenge is to determine the number of repeated measures that are required to sufficiently 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.

What makes this setting peculiar is the fact that the true endpoint is the ultimate assessment in a sequence of repeated measures. It is then appealing to consider earlier measures, either in isolation or several combined, as a potential surrogate endpoint. The length and cost reducing potential has to be weighed carefully against loss in precision and the risks of an inappropriate decision regarding a new compounds fate. In the next section we develop a quantitative criteria to do so.

5.2.4 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):

$$FC = NC_1 + NKC_2. \tag{5.14}$$

Here, FC represents the fixed total financial cost, N is the total number of patients in the study, K is the number of planned repeated measurements per subjects, C_1 is the cost of recruiting a patient to the study, and C_2 is the cost per measurement and per subject. Let $R = C_1/C_2$ be the ratio of both costs; usually the cost of recruiting a patient to the study is higher than the cost per measurement, i.e., $R > 1$. We can then re-write (5.14) 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, $1 - VRF_{\text{indiv}}(m)$, 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 (5.15), 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:

$$CPR_0(m) = w_1 \cdot [1 - VRF_{\text{ind}}(m)] + (1 - w_1) \cdot \frac{R + m}{R + K}, \tag{5.15}$$

The quantity $CPR_0(m)$ balances the lack of surrogacy, $1 - VRF_{\text{ind}}(m)$, on the one hand, and the proportion of total financial cost required to take m measurements, $(R+m)/(R+K)$, on the other hand. Retaining more measurements reduces the first term, because the VRF will go up, but at the same time leads to an increase in the

cost term. The relative importance attributed to the terms is captured by the weight w_1, with a user-assigned value between 0 and 1. The number m_o is determined as that minimizing $CPR(m)$.

It goes without saying that the time of measurement, which is not explicitly accounted for in (5.15), plays an important role in longitudinal data. It is therefore sensible to consider some extensions incorporating the time of measurement. 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 (5.15). One can therefore elect to introduce a third term accounting for time lag:

$$CPR_I(m) = w_1 \cdot [1 - VRF_{\text{ind}}(m)] + w_2 \cdot \frac{R+m}{R+K} + w_3 \cdot \frac{t_m - t_0}{t_k - t_0}, \tag{5.16}$$

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, (5.16) takes the form

$$CPR_I(m) = w_1 \cdot [1 - VRF_{\text{ind}}(m)] + w_2 \cdot \frac{R+m}{R+K} + w_3 \cdot \frac{M}{K}. \tag{5.17}$$

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

$$CPR_{II}(m) = w_1 \cdot [1 - VRF_{\text{ind}}(m)] + w_2 \cdot \frac{R+m}{R+K} + w_3 \cdot \frac{M-1}{K}. \tag{5.18}$$

Although extended, 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 (5.15) will be employed, unless otherwise stated.

5.3 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: compound symmetry and first-order auto-regressive process. Derivations of the results used in this section are provided in more details in Appendix A.

5.3.1 Compound Symmetry Structure

Although stringent, the compound symmetry structure is amongst the most commonly used covariance structures. Assume that the covariance structure of (5.13) is compound symmetry, 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. In this setting, it is trivial to show that (see Section A.1.1),

$$VRF_{\text{indiv}}(m) = \frac{m\rho^2}{1+(m-1)\rho}. \tag{5.19}$$

Let us study the predictive characteristics of this case. It follows that $VRF_{\text{indiv}}(m)$ 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:

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

Elementary calculations show that $g(\rho)$ is a decreasing function of ρ, 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. Conversely, if $\rho = 0$ then all the observations are independent and no sensible prediction is possible.

Coherent with the nature of compound symmetry, the position in the sequence of the m observations that constitute the surrogate is totally irrelevant. From (5.15) and (5.19), it is trivial to show that in this setting the CPR function takes the form

$$CPR(m) = w_1 \cdot \frac{(1-\rho)(1+m\rho)}{1+(m-1)\rho} + (1-w_1) \cdot \frac{R+m}{R+K}, \tag{5.20}$$

of which the extremes are easy to determine: (5.20) reaches its minimum at m_+ and m_- when $\rho > 0$ and $\rho < 0$ respectively, where

$$m_\pm = -\left(\frac{1-\rho}{\rho}\right) \pm \sqrt{\frac{w_1(R+K)(1-\rho)}{1-w_1}}. \tag{5.21}$$

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, it is recommended that $m_\pm$ 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 expected because when the cost of recruiting patients is much higher than taking more measurements on subjects, the obvious way to increase precision is through taking more measurement per subject. Additionally, an increase in the correlation ρ between measurements leads to a decrease in m_+ when the weight assigned to precision is small to moderate. As the weight increases, the value of m_+ increases for ρ in $[0; 0.5]$ and decreases in

$[0.5; 1]$. Also, an increase in K generally leads to a slight increase in m_+.

5.3.2 First-order Auto-regressive Process

In statistics and signal processing, an auto-regressive (AR) model is often used to model and predict various types of natural phenomena. The first-order auto-regressive (AR(1)) process is frequently encountered in longitudinal data, with ρ^t the correlation between two measurements, t time units apart. Let Σ_{SS} be 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 $VRF_{\text{ind}}(m) = \rho^{2(K-m)}$ (see Section A.1.2).

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. Although this may reduce the cost of taking measurements, it implies increase in waiting-time or duration of the study. 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:

$$CPR(m) = w_1 \cdot \left(1 - \rho^{2(K-m)}\right) + (1 - w_1) \cdot \frac{R+m}{R+K}. \tag{5.22}$$

Interestingly, (5.22) 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.

5.4 Simulation Study

The previous results are enlightening. However, many other covariance structures commonly encountered in practice such as the Toeplitz and unstructured, amongst others, are not 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. A simulation study was performed to investigate further these issues, with focus on the two association structures of Section 5.3.

5.4.1 Data Generation

Equally spaced longitudinal data were generated based on (5.13), 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 first-order autoregressive, AR(1), or compound symmetry, CS. The variance in Σ_L was assumed constant and the correlation between successive measurements was set to either 0.3, 0.6, 0.71, 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 (5.13).

The data generation scheme discussed earlier 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 as described in above (Section 5.4.1), analyzed and the optimal number of measurements determined as described in Section 5.2.

In principle, simulations based on 100 runs are in jeopardy of large Monte Carlo errors. However, because we predominantly determine the optimal number of measurements, a discrete quantity, there is little gain to be expected from increasing the number of runs.

5.4.2 Simulation Study Results

The results of the simulation for the case of $R = 4$ and $K = 10$ are summarized in Tables 5.1–5.4. 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 5.3, 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 (Table 5.2), then 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.

Note that, in Table 5.2, missing entries are not due to convergence issues. Actually those spaces are left for conveniently putting the results for CS and UN in one table. For example, for CS structure with a weight of 0.7 and correlation of 0.71, when the analysis was done with a CS, time point 3 was selected as optimal with 100 percent of the samples. Whereas, when the analysis is conducted with UN, time points 2 and 4 were also selected as optimal with percentages of 14 and 6 respectively. These two time points were not picked before and hence the space corresponding to time points 2 and 4 is left blank in the columns corresponding to CS.

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 Tables 5.3–5.4. For the Toeplitz structure, up to 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%.

5.5 Constrained Maximization

It is not uncommon to encounter circumstances in which clinical trials are faced with budget and or time 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: $(R+m)/(R+K) \leq \delta_1$ 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. Details on the derivations for the results used in this section are provided in Section A.3 of Appendix A.

Table 5.1: *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

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 $\rho > 0$, 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).

Table 5.2: *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 5.3: *Simulation study. Results for the optimal number of measurements with: unstructured covariance and Toeplitz correlation structure with slowly declining correlation (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

To enhance insight, we carried out a limited set of simulations for both AR(1) and CS. Results are summarized in Table 5.5. The simulation revealed that as R increases, the optimal M diminishes. This is in line with intuition because the total cost and the number of patients in the study are fixed and hence to maintain a low cost, the only

Table 5.4: *Simulation study. Results for the optimal number of measurements with: 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
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

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,

Table 5.5: *Simulation study for constraint maximization. Results for the optimal number of measurements for ρ= 0.3 with CS and AR(1). (δ: percentage of cost available; R: cost ratio; m_o: optimal number of measurements; $VRF_{ind}(m)$: individual-level surrogacy for the optimal number of measurements m_o.)*

CS				AR(1)			
δ	R	m_0	$VRF_{\text{ind}}(m)$	δ	R	m_0	$VRF_{\text{ind}}(m)$
0.2	1	1	0.10979	0.2	1	1	2.29E-14
0.3	1	2	0.13882	0.3	1	2	4.44E-09
0.4	1	3	0.16273	0.4	1	3	4.11E-08
0.5	4	3	0.17758	0.5	4	3	4.11E-08
0.6	4	4	0.19843	0.6	4	4	2.85E-07
0.8	4	7	0.21728	0.8	4	7	0.000584203
0.6	10	2	0.19843	0.6	10	2	4.44E-09
0.7	10	4	0.20691	0.7	10	4	2.85E-07
0.8	10	6	0.21728	0.8	10	6	4.68688E-05
0.9	10	8	0.22203	0.9	10	8	0.007548459

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.

5.6 Impact of Missingness

In practice, it is not unusual for some measurement sequences in a longitudinal study to terminate early for reasons outside the control of the investigator. Subjects so affected are referred to as dropouts. Additionally, intermediate scheduled measurements might be missed, the so-called intermittent missing values (Verbeke and Molenberghs 2000, Molenberghs and Verbeke 2005, Molenberghs and Kenward 2008). Dropout is a common occurrence in clinical trials. Indeed, the vast majority of incomplete sequences in clinical trials are of a dropout type, with a relative minor

fraction of incompletely sequences exhibiting intermittent missingness. Thus we put emphasis on the problem of dropout.

It might be necessary to accommodate missingness into the modeling process, as they may have impact on statistical inference. Many methods have been formulated such as the *selection models* (Little and Rubin 1987, 2002) as opposed to *pattern-mixture models* (Little 1993, 1994) and *shared-parameter models* (Wu and Carroll 1988, Wu and Bailey 1988, 1989). Here, we focus on the selection models and the terminology of Rubin (1976) and Little and Rubin (1987). Key concepts are (1) *missing completely at random* (MCAR), if the missingness is independent of both unobserved and observed data, (2) *missing at random* (MAR) if, conditional on the observed data, the missingness is independent of the unobserved measurements, and (3) *missingness not at random* (MNAR), when neither MCAR nor MAR applies.

In this section we performed a limited simulation study to assess the impact of missingness on the prediction of the final outcome of a longitudinal response using earlier measures from the same sequence. Again, the CS and AR(1) covariance structures were used with ρ set to either 0.3, 0.5 or 0.9. First, data were generated with ten repeated measures, which we refer to as the full-data. Next, artificial missingness was imposed using a logistic model to mimic the three missing data mechanisms; MCAR, MAR and MNAR. The VRF_{ind} was calculated base on complete cases (CC), available cases (AC) , multiple imputation (MI) and the full-data. Bias and mean square error (MSE) were used to assess the magnitude of the effect.

It should be noted that the objective here is to investigate the impact of missingness, under the three missingness mechanisms, using some methods commonly use to handle missingness. The issue of handling missingness in a optimal manner together with sensitivity analysis is not of particular concern. All methods used, CC, AC and

MI, exhibited bias for all missingness mechanisms considered. The bias was highest under the MNAR mechanism. This is not unexpected as the methods considered are only valid under the MCAR or MAR mechanisms. MI had the lowest bias, closely followed by AC, while CC generally showed relatively larger bias. The discrepancy between these methods is high when the data were generated with moderate to low values of ρ. This makes sense intuitively, as for very high values of ρ, say > 0.9, the first few observed measurements may be enough to predict the final outcome, especially for the CS structure. It is therefore crucial to use methods that appropriately account for missingness when applying this method to incomplete data.

5.7 Application to the Case Study

Two motivating studies introduced in Section 2.2.1 and Section 2.2.2 are analyzed here and the results displayed in Tables 5.6 and 5.7, respectively. For the data coming from the opthalmology experiment (Section 2.2.1), measurements of visual acuity were taken at baseline and every sixth week there after up to 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.

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. In conclusion, even though necessarily a bit subjective, it seems that 3 measurements leads to reasonably good quality, while reducing the study time to a third.

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 this case, with similar logic as in the previous case study, it appears that two measurements provides reasonable results, while leading to a 50% study-time reduction.

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 patients 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 (5.17) and (5.18). Results can be found in Table 5.7 for schizophrenia and Table 5.8 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.

5.8 Discussion

Our simulation study involved varying numbers of trials and subjects within trials. Unlike conventional surrogate marker validation, which involves two separate outcomes where one is used as a potential surrogate for the other, here we have studied

Table 5.6: *Case study in opthalmology. Results for the optimal number of measurements based on cost function (5.16). (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				

a scenario where there is a single outcome only, 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 individual-level surrogacy was assessed using the canonical correlation approach, introduced by Alonso et al. (2004) and discussed in Section 5.2. 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,

Table 5.7: *Case study in schizophrenia. Results for the optimal number of measurements based on cost function (5.16) and modified cost function (5.17). (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 (5.16)				Cost function (5.17)					
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

an objective function has been utilized. The objective function has two parts, which takes 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

Table 5.8: *Case study in opthalmology. Results for the optimal number of measurements based on modified cost function (5.17) and (5.18); (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 (5.17)												
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
0.1	0.3	0.6	1	0.18	1	0.18	1	0.18	1	0.18	1	0.18
0.3	0.3	0.4	1	0.18	1	0.18	1	0.18	1	0.18	1	0.18
0.4	0.3	0.3	2	0.34	2	0.34	2	0.34	2	0.34	2	0.34
0.5	0.3	0.2	3	0.45	3	0.45	3	0.45	3	0.45	8	0.85
0.6	0.3	0.1	8	0.85	8	0.85	8	0.85	9	0.91	9	0.91
Modified cost function (5.18)												
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
0.1	0.3	0.6	1	0.18	1	0.18	1	0.18	1	0.18	1	0.18
0.3	0.3	0.4	1	0.18	1	0.18	1	0.18	1	0.18	1	0.18
0.4	0.3	0.3	2	0.18	2	0.18	2	0.34	2	0.34	2	0.34
0.5	0.3	0.2	3	0.45	3	0.45	3	0.45	3	0.45	3	0.45
0.6	0.3	0.1	8	0.85	8	0.85	8	0.85	9	0.91	9	0.91

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. Similarly, when it is deemed better to detect a condition early rather than late. A possible extension of our work would be to incorporate the cost of a failure to detect the condition early, when treatments are more effective or when a change to an alternative therapy may be more beneficial than when such a switch is effectuated at a later stage.

The results of the simulation study for two data-generation schemes, based on CS and AR(1), respectively, 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 CPR 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.

In practice, missing data are frequently encountered with longitudinal studies. Results from a limited simulation indicate that when using this method with incomplete data, is it important to employ models that appropriately handle the missingness. Additionally, misspecification of the covariance structure can have substantial effect on VRF_{ind}. Simulation results (not shown) indicated that when data were generated using CS and analyzed as AR(1) and vice versa, substantial bias were obtained relative to using the 'correct' covariance structure. Also, analyzing the data using an unstructured covariance structure minimizes bias, when the data were generated with either CS or AR(1). Therefore it is advisable to determine a plausible covariance structure through model building, starting with an unstructured covariance matrix, when applying these methods.

Our simulation study, while relatively broad, is intrinsically limited, as is the case for every simulation study. A number of extensions could be considered. First, while the first-order autoregressive structure applies to equally spaced measures only, this is

not the case for the compound symmetry and unstructured covariances. In principle, further structures for unbalanced data, such as general special functions, as available in the SAS System, could be considered. Second, our derivations crucially rely on the continuous nature of the outcome, and hence on the linearity of the expressions involved, enabling the derivation of explicit expressions.

Should the outcome be non-Gaussian, then relevant model choices are generalized estimating equations (GEE, Liang and Zeger 1986) or generalized linear mixed models (GLMM, Breslow and Clayton 1993), for example. A review of this and additional methodology is provided in Molenberghs and Verbeke (2005). Such models, however, raise a number of complexities. The presence of a mean-variance link and the non-linear nature of the link function defeats the derivation of explicit analytical expressions like in the continuous case. Of course, one might make progress through the use of approximate expressions, or by way of Monte-Carlo-based evaluations. These are just two examples of how extensions could be considered.

Indeed, GEE was employed while performing a simulation for predicting the final outcome of a binary longitudinal sequence using cumulative earlier measurements, subject to cost and time constraints. Simulation results (not shown here)indicate that 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. These results are similar to with the results obtained for the case of continuous longitudinal outcome.

Part II

Surrogate Marker Evaluation from an Information Theory Perspective

6

Information-Theoretic Approach to Surrogate Markers Validation

To recapitulate, methodologies discussed in the previous chapters where based on the meta-analytic approach of Buyse et al. (2000) and its extensions. This methodology stemmed from the work of Prentice (1989) and Freedman, Graubard, and Schatzkin (1992), who propose a formal definition of surrogacy and outline a validation strategy. Even though their ideas are appealing, a drawback, shared with all single-trial approaches, is that they rest on strong and unverifiable assumptions. Therefore several authors (Daniels and Hughes 1997, Buyse et al. 2000) proposed methods based on information coming from several units or trials. Using hierarchical linear models, Buyse et al. (2000) defined individual-level (R^2_{ind}) and trial-level (R^2_{trial}) surrogacy measures, both of a coefficient of determination type.

A drawback of the meta-analytic paradigm is that several individual-level measures have been proposed for different settings. For normally distributed endpoints, Buyse et al. (2000), using a bivariate normal regression model, defined the R^2_{ind}as the correlation between the surrogate and true endpoint after adjusting for treatment and trial effects. In the binary-binary setting, Renard et al. (2002) used the correlation between two latent variables $R^2_{\text{ind}} = \text{corr}(\widetilde{S}, \widetilde{T})$ to define individual-level surrogacy and alternatively define $R^2_{\text{ind}} = \psi$, the global odds ratio between both endpoints estimated from a bivariate Plackett-Dale model. Considering a longitudinal surrogate for a time-to-event clinical endpoint, Renard et al. (2003) used the so-called model-based curve $R^2_{\text{ind}}(t)$, using ideas stemming from the model proposed by Henderson et al. (2000), which can be defined relative to any time over the course of measurement of the surrogate. Also, Burzykowski et al. (2001) used Kendal's tau to characterize the individual-level surrogacy for failure time endpoints. Using multivariate ideas and canonical correlations, the so-called R^2_Λ has been proposed to evaluate surrogacy when both responses are measured longitudinally (Burzykowski, Molenberghs, and Buyse, 2005; Alonso, Geys, and Molenberghs, 2006). This R^2_Λ can be incorporated into a more general framework allowing for interpretation in terms of canonical correlations of the error vectors, based on which these authors define a family of individual-level parameters.

The scenarios mentioned highlight a limitation of the meta-analytic methodology so far: different settings require different definitions and in some of these settings, the association is measured at a latent level, hampering interpretation. Furthermore, more than one measure for individual-level surrogacy can be defined for some settings, an immediate example being the binary-binary setting. Additionally, in all cases, a joint and often nonstandard model for both endpoints is needed, frequently represent-

ing a serious computational burden, even for normally distributed endpoints (Tilahun et al. 2007, Chapters 4 and 3).

To overcome these limitations, Alonso et al. (2004b) extend the R^2_Λ to nonnormal settings using a scaled likelihood reduction factor (LRF). Later in 2007, Alonso and Molenberghs (2007) used information theory to create a unified framework for surrogate markers evaluation base on a measure of information, R^2_h. These authors showed that the LRF is a consistent estimator of R^2_h, thereby, leading to a definition of surrogacy with an intuitive interpretation and applicable in a wide range of situations. Another strong point for the LRF is its easy applicability in practice. This is a direct consequence of the fact that it is based on routine measures produced by standard softwares.

This chapter serves as basis for all the chapters in this part of the thesis. Therefore, a succinct introduction on information theory is provided in Section 6.1, while some important concepts of information theory are briefly described in Section 6.2. Section 6.3 focuses on surrogate marker validation from an information-theoretic perspective, as proposed by Alonso and Molenberghs (2007). Finally, some implications of the information theoretic approach are discussed in Section 6.4.

6.1 Information Theory

Information theory is a branch of electrical engineering, applied mathematics, and the mathematical theory of probability and statistics, involving the quantification of information. Its abstract formulations are applicable to any probabilistic or statistical system of observations. Since its inception it has broadened to find applications in many other areas, including natural language processing, modern communication theory, networks other than communication networks and statistical inference.

In spirit and concept, information theory has its mathematical roots connected with the idea of entropy used in thermodynamics and statistical mechanics. It was studied by James Clerk Maxwell (1867), Ludwig Boltzman (1877) and Leo Szilard (1929) who identified entropy with information. Nyquist (1924) recognized the logarithmic nature of the measure of information and made an early attempt to formalize the theory. A major contribution in this area came in 1948 when Shannon published a remarkable paper on the properties of information sources and the communication channels: "Mathematical Theory of Communication."

It is worth considering the fact that the most commonly used estimating technique, maximum likelihood, is usefully viewed as an empirical version of information. This follows since the usual log-likelihood divided by the sample size, which can be taken as a fixed constant, provides a consistent estimate of the information (O'Quigley 2008). R. A. Fisher's well-known measure of the amount of information supplied by data about an unknown parameter is the first use of information in statistics. Further, Kullback and Leibler (1951) studied another statistical information measure, the so-called information gain, involving two probability distributions associated with the same experiment. The later plays an important role in this work as will be seen in subsequent sections.

6.2 Some Concepts and Information-theoretic Measure of Association

The entropy, a good measure of randomness or uncertainty, of a random variable is defined in terms of its probability distribution. Shannon (1948) defined entropy for a discrete random variable, Y, taking values $\{k_1, k_2, \ldots, k_m\}$ and with probability

function $P(Y = k_i) = p_i$ as

$$H(Y) = \sum_i p_i \log\left(\frac{1}{p_i}\right). \tag{6.1}$$

$H(Y)$ is the average uncertainty associated with P. 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. In general, we could use the alternative definition $H(Y) = E[-\log P(Y)]$. We can also define the joint and conditional entropy as $H(X, Y) = E[-\log P(X, Y)]$ and $H(Y|X) = E[-\log P(Y|X)]$, respectively. Entropy is always nonnegative, 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). Shannon's entropy can be extended to situations where the random variable is continuous, the so-called differential entropy. The differential entropy $h_d(X)$ of a continuous variable X with density f_X and support S_{f_X} equals

$$h_d(Y) = -E[\log f_X(X)] = -\int_{S_{f_X}} f_X(x) \log f_X(x) dx. \tag{6.2}$$

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 $Y = g(X)$, it follows that $h_d(Y) = h_d(X) - \mathrm{E}_Y\left(\log\left|\frac{dx}{dy}(g)\right|\right)$. Nevertheless, beyond all these limitations, it is possible to define some very important and meaningful measures based on the concept of differential entropy that are positive and invariant by bijective transformations.

A concept analogous to conditional entropy can also be defined for continuous outcomes. Let (X, Y) be two continuous random variables with joint density f_{XY}. Further, we will denote by f_Y, f_X, and $f_{Y|X}$ the marginal densities of Y, X, and $Y|X$, respectively. The expected conditional entropy of Y given X can be defined as

$h_d(Y|X) = E_x[h_d(Y|X = x)]$. It represents the uncertainty in Y that is expected to remain if the value of X is known (Alonso and Molenberghs 2007). Based on these concepts, the amount of uncertainty in Y that is expected to be removed if the value of X is known can be quantified by $I(X,Y) = h_d(Y) - h_d(Y|X)$, the so-called *mutual information*. $I(X,Y)$ is always non-negative, zero if and only if X and Y are independent, symmetric, invariant under bijective transformations of X and Y, and $I(X,X) = h_d(X)$. Intuitively, if (X, Y) are independent, then their mutual information should be zero. On the other hand, if X and Y are identical, then knowing X reveals nothing new about Y or vice versa, therefore, the mutual information should be the same as the information conveyed by X (or Y) alone, which is given by the entropy of the random variable. Also, it can be shown that $I(X,Y)$ equals the Kullback-Leibler discrepancy measure (Alonso and Molenberghs 2007), which quantifies the *distance* between the joint distribution of X and Y and the product of their marginal distribution.

The concept of entropy-power, introduced by Shannon (1948) for comparison of continuous random variables, can be defined for a continuous n-dimensional random vector, X, as

$$\mathrm{EP}(X) = \frac{1}{(2\pi e)^n} e^{2h(X)}. \tag{6.3}$$

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. For other distributions where variance itself may not be the best measure of uncertainty or lack of knowledge (dispersion), the concept of entropy-power, might be viewed as having more generality (O'Quigley 2008). We can now define an information-

theoretic measure of association (Schemper and Stare 1996):

$$R^2_{\text{h}} = \frac{\text{EP}(Y) - \text{EP}(Y|X)}{\text{EP}(Y)}, \tag{6.4}$$

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 $R^2_{\text{h}} \to 1$ for continuous outcomes, there is usually some degeneracy appearing in the distribution of (X,Y). There is a direct link between R^2_h and the mutual information:

$$R^2_{\text{h}} = 1 - e^{-2I(X,Y)}. \tag{6.5}$$

For Y discrete: $R^2_{\text{h}} \leq 1 - e^{-2H(Y)}$, implying that R^2_{h} will have an upper bound that can generally be smaller than 1, given by $1 - e^{-2H(Y)}$. Consequently, the calibrated quantity

$$\widetilde{R}^2_{\text{h}} = \frac{R^2_{\text{h}}}{1 - e^{-2H(Y)}} \tag{6.6}$$

is more meaningful, reaching 1 when both endpoints are deterministically related. Note that $H(Y)$ is the log-likelihood of the true endpoint divided by the total number of subjects.

6.3 Surrogate Marker Validation from an Information-Theoretic Perspective

We can now redefine surrogacy in a simple and intuitive manner, while preserving previous proposals as special cases. While we will focus on individual-level surrogacy, all results apply to the trial level too. Let us denote by $Y = T$ and $X = S$ the true and surrogate endpoints, respectively. Base on the information-theoretic measure of association, Alonso and Molenberghs (2007) stated the following definition for surrogacy:

> *S is called 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.*

This definition, in spite of being based on formal concepts rooted in information theory, is simple and intuitive, because the idea behind surrogacy is to reduce our lack of knowledge about a true endpoint through the use of a surrogate alternative. Thus, R^2_{h} is a valuable tool to evaluate surrogacy in practice. $R^2_{\mathrm{h}} \approx 1$ implies that our potential surrogate is promising, and could be interpreted as follows: once the surrogate is known, almost all of our uncertainty about the true endpoint will be removed. On the other hand, $R^2_{\mathrm{h}} \approx 0$ evidences a poor surrogate, unable to reduce our uncertainty about the true endpoint.

A meta-analytic framework, with N clinical trials, produces N_q different R^2_{hi}, and a meta-analytic R^2_h can be obtained as:

$$R^2_h = \sum_{i=1}^{N_q} \alpha_i R^2_{hi} = 1 - \sum_{i=1}^{N_q} \alpha_i e^{-2I_i(S_i,T_i)}, \tag{6.7}$$

where $\alpha_i > 0$ for all i and $\sum_{i=1}^{N_q} \alpha_i = 1$, i.e., a convex combination. Different choices for α_i lead to different proposals, producing an uncountable family of parameters.

$$\Omega_h = \left\{ \theta_h : \theta_h = 1 - \sum_{i\,=\,1}^{N_q} \alpha_i e^{-2I_i\,(S_i,\,T_i)}, \quad \text{where:} \quad \alpha_i > 0 \text{ and } \sum_i \alpha_i = 1 \right\}. \tag{6.8}$$

This opens the additional issue of finding, in a well-defined way, the best member of Ω_h that should be more appropriate in a certain practical situation, if it actually exists. Alonso and Molenberghs (2007) showed that (6.7) reduces to previous proposals as

special cases for well chosen choices of α_i. When applied in the cross-sectional normal setting and considering the bivariate model introduced by Buyse et al. (2000) the R^2_{h} equals the R^2_{ind}. In the longitudinal normal case and based on the model considered in Burzykowski et al. (2005) and Chapter 5 (Section 5.2.1), the $R^2_{\mathrm{h}} = R^2_{\Lambda}$ when $\alpha_i = N^{-1}$ and Ω_h reduces to the Ω_Λ family defined by these authors.

It is often the case in practice that different trials in meta-analysis have different sizes. Since univariate models are used to evaluate surrogacy in the information-theoretic approach, there is a need to adjust for the heterogeneity in information content between trial-specific contributions (Tibaldi et al. 2003, Pryseley et al. 2007, Section 3.3.3). A common way to account for a variable amount of information per trial is by weighting the contributions according to trial size.

Furthermore, when the information-theoretic ideas are applied at the trial level using the fully hierarchical approach considered by Buyse et al. (2000), the R^2_{h} reduces to the R^2_{trial} proposed by these authors. This connection allows us to reinterpret R^2_{trial} from an information point.

Based on the ideas of Kent (1983) to build confidence intervals for $2I(T,S)$, Alonso and Molenberghs (2007) developed asymptotic confidence intervals for R^2_h. Let $\hat{a} = 2n\widehat{I}(T,S)$, where n is the number of patients. Define $\kappa_{1:\alpha}(\mathrm{a})$ and $\delta_{1:\alpha}(\mathrm{a})$ by $P\left\{\chi^2_1\left[\kappa_{1:\alpha}(a)\right] \geq a\right\} = \alpha$ and $P\left\{\chi^2_1\left[\delta_{1:\delta}(a)\right] \leq a\right\} = \alpha$. Here, χ^2_1 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^2_h is

$$\sum_i \alpha_i \left[n_i^{-1}\kappa^i_{1:\alpha}(\hat{a}), n_i^{-1}\delta^i_{1:\alpha}(\hat{a})\right], \tag{6.9}$$

where $1-\alpha_i$ is the Bonferroni confidence level for the trial intervals. This asymptotic interval has considerable computational advantage relative to the bootstrap approach used by Alonso et al. (2005). Through simulation studies, it was observed that 95%

asymptotic intervals were tighter than 95% percentile bootstrap intervals, and the discrepancy between the asymptotic and bootstrap intervals reduces with increase in the number of trials and trial sizes (Pryseley et al. 2007).

6.4 Some Implications of the Informational Approach

The link between information theory and surrogate marker validation may have many interesting implications. We briefly mention some of them and discuss the theoretical plausibility of finding a good surrogate.

6.4.1 Prentice Criteria

Prentice (1989) put forward one of the most interesting ideas to approach surrogate marker validation. However, base on an information-theoretic point of view, Prentice's main criterion $f(T|S,T) = f(T|S)$ implies that once we have adjusted for the surrogate, the true endpoint carries no further information about the treatment (Alonso and Molenberghs 2007). Essentially, it implies that there are no pathways from the treatment to the true endpoint which bypass the surrogate. We believe this is quite a strong condition and that in practice the mechanisms that lead from the treatment to the true endpoint will be frequently more complicated than what this model states. Additionally, even when this condition is satisfied, the data processing inequality (DPI) (Cover and Tomas 1991), seems to illustrate that Prentice's main criterion entails that a treatment effect on T implies a treatment effect on S rather than the other way around (Alonso and Molenberghs 2007).

6.4.2 The Proportion Explained

In spite of contributing the important switch from a hypothesis testing to an estimation paradigm, the PE is not considered to be the final answer (Freedman, 2001). Among other limitations, the PE is not restricted to the unit interval and can lead to counterintuitive results in some scenarios (Burzykowski et al. 2005). On the other hand, if we apply the information-theoretic ideas at the trial level using the fully hierarchical approach considered by Buyse et al. (2000), the R^2_{h} defined before will take the form

$$R^2_{\mathrm{ht}} = 1 - \frac{\mathrm{EP}(\beta_i|\alpha_i)}{\mathrm{EP}(\beta_i)},$$

where $(\beta_i,\ \alpha_i)$ denote the treatment effects on the surrogate and true endpoints, respectively, at the i^{th} trial and $\mathrm{EP}(\beta_i|\alpha_i)$ and $\mathrm{EP}(\beta_i)$ are the power entropies associated with the conditional distribution $\beta_i|\alpha_i$ and the marginal distribution of β_i, respectively. Note that the R^2_{ht} and the PE are structurally very similar (see Section 3.1). Although R^2_{ht} and PE are very close in spirit, R^2_{ht} overcomes some of the limitations of the PE as it is based on a meta-analytic framework, thereby avoiding in this way all the drawbacks common to all the single-trial approaches. Therefore, it is clear at this stage that the information-theoretic approach allows us to redefine the PE in a more meaningful and principled way.

6.4.3 Theoretical Plausibility of Finding a Good Surrogate

As mentioned and illustrated earlier in the first part of the thesis, Buyse et al. (2000) approach the surrogate marker validation problem from a prediction perspective. Their coefficients of determination aims at quantifying, under certain assumptions, how well this prediction is for normally distributed endpoints. However, it is neither straightforward nor trivial to generalize this idea beyond normal endpoints.

Alonso and Molenberghs (2007) provided a more general result for R^2_{h} base on Fano's inequality.

Fano's inequality shows the relationship between entropy and prediction:

$$\mathrm{E}\left[(T - g(S))^2\right] \geq \mathrm{EP}(T)(1 - R^2_h), \tag{6.10}$$

where $\mathrm{EP}(T) = \dfrac{1}{2\pi e} e^{2h(T)}$. Note that nothing has been assumed about the distribution of our responses and no specific form has been considered for the prediction function g. Also, (6.10) shows that the predictive quality strongly depends on the characteristics of the endpoint, specifically on its power-entropy. Fano's inequality states that the prediction error increases with $\mathrm{EP}(T)$ and therefore, if our endpoint has a large power-entropy then a surrogate should produce a large R^2_h to have some predictive value. This means that, for some endpoints, the search for a good surrogate can be a dead end street: the larger the entropy of T the more difficult it is to predict. Studying the the power-entropy before trying to find a surrogate is therefore advisable.

6.5 The Likelihood Reduction Factor (LRF)

The previous sections of this chapter introduced information theory and described theoretical concepts leading to the validation of surrogate markers from an information theory perspective. The next question becomes, 'how do we estimate R^2_{h} from data in practice?' Alonso and Molenberghs (2007) showed that the LRF of Alonso et al. (2004a) is a consistent estimator of the R^2_{h}. In principle, other estimators could be used as will be shown in subsequent chapters. Here we focus on the LRF.

As previous developments and the first part of the thesis clearly showed, estimating individual-level surrogacy 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

$$g_T\{E(T_{ij})\} = \mu_{Ti} + \beta_i Z_{ij}, \tag{6.11}$$

$$g_T\{E(T_{ij}|S_{ij})\} = \theta_{0i} + \theta_{1i} Z_{ij} + \theta_{2i} S_{ij}, \tag{6.12}$$

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 (6.11) and (6.12) can be readily extended to incorporate more complex settings. For example, longitudinal data are easily incorporated by including functions of time in (6.11) and (6.12). Other extensions, such as non-linearity between S_{ij} and $g_T\{E(T_{ij})\}$ are possible, more generally we have $g_T\{E(T_{ij}|S_{ij})\} = \theta_{0i} + \theta_{1i} Z_{ij} + f(S_{ij})$. In most part of the thesis, we assume a linear relationship between S_{ij} and $g_T\{E(T_{ij})\}$, but consider extensions of (6.11) and (6.12) in the light of simplified modeling strategy, as presented by Tibaldi et al. (2003).

The trial dimension provides a choice between treating the trial-specific effects as fixed or random. The former is often chosen out of necessity, when the latter is too challenging. If the trial-specific effects are chosen fixed, then (6.11) and (6.12) are used to validate the surrogate endpoint. On the other hand, if the trial-specific effects are considered random, we extend (6.11) and (6.12) to appropriate generalized linear

mixed-effects models

$$g_T\{E(T_{ij})\} = \mu_T + m_{Ti} + \beta Z_{ij} + b_i Z_{ij}, \quad (6.13)$$

$$g_T\{E(T_{ij}|S_{ij})\} = \theta_0 + c_{Ti} + \theta_1 Z_{ij} + a_i Z_{ij} + \theta_{2i} S_{ij}, \quad (6.14)$$

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.

Let us turn to the so-called *likelihood reduction factor* (LRF). Observe that, in the case where the true endpoint is continuous and normally distributed, (6.11) and (6.12) reduce to normal regression models. On the other hand, when the true endpoint is binary or counts, (6.11) and (6.12) reduce to logistic or Poisson regression models. Similarly, setting the natural link function g_T equal to the identity, logit, and logarithmic functions in (6.13) and (6.14) leads to the linear mixed model for continuous data, the logistic-normal model for binary data, and the Poisson-normal model for counts, respectively. Alonso and Molenberghs (2007) used the LRF to evaluate individual level surrogacy, which is obtained by

$$LRF = 1 - \frac{1}{N}\sum_i \exp\left(-\frac{G_i^2}{n_i}\right), \quad (6.15)$$

where G_i^2 denotes the log-likelihood ratio test statistic to compare (6.11) and (6.12) or (6.13) and (6.14) within trial i. We can think of (6.15) as a sample estimate of a general measure of association between both endpoints based on the information gain about the true endpoint by using the surrogate. Alonso et al. (2005) established a number of properties for LRF, in particular its ranging in the unit interval and, importantly, its reduction to R^2_{ind} and R^2_Λ in the cross-sectional normal-normal and

the longitudinal-longitudinal cases, respectively. Furthermore, (6.6) has been recommended for settings with discrete clinical endpoints. Although there are formulas for calculating $H(Y)$ for most parametric distributions, an estimate for $H(Y)$ in a given sample is the log-likelihood from (6.11) without the treatment effect scaled by the total number of subjects. Finally, asymptotic confidence intervals for LRF can be obtained from (6.9) by replacing $\widehat{a}$ with G_i^2.

6.6 Discussion

Stemming from interesting ideas of Prentice (1989) and Freedman et al. (1992), the meta-analytic approach of Buyse et al. (2000), partly devoted to frame the evaluation in a multi-trial framework, led to definitions in terms of the quality of trial- and individual-level association between a potential surrogate and a true endpoint. A drawback is that different settings have led to different measures at the individual level, as the R^2 measures coming from the framework of Buyse et al. (2000), do not readily generalize to settings with nonnormal outcomes. Alonso and Molenberghs (2007) used information theory to create a unified framework, leading to a definition of surrogacy with an intuitive interpretation, offering interpretational advantages, and applicable in a wide range of situations. Indeed, the information-theoretic measure, R_{h}^2, serves as a theoretical basis for the R^2 based measures within the meta-analytic framework, as it is applicable to a wide variety of settings (normal, binary, categorical, and longitudinal outcomes) and reduces to the quantities previously introduced in the literature.

Furthermore, while the meta-analytic approach is elegant, it faces computational problems, which are largely alleviated by the information-theoretic approach, given that it is only based on univariate models that can be fitted with standard software

packages. Also, when there are more than two arms in the clinical trials under consideration, one has the choice between calculating the validation measures using all arms simultaneously. Indeed, the information-theoretic developments carry through when Z represents a nominal covariate rather than a sole binary variable. Alternatively, the measures can be calculated for every pair of arms deemed of interest.

The information-theoretic approach can easily incorporate baseline covariates for early-stage clinical trials, as well as assess the joint effect of multiple surrogate endpoints on a true endpoint. Note that, by parsimoniously using information, the information-theoretic approaches may lead to tighter confidence intervals than in the hierarchical-model framework. This is an advantage, in addition to increased generality and flexibility (Alonso and Molenberghs 2008).

Again, 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. Clinical and biopharmaceutical arguments will have to be juxtaposed with evidence from the surrogate marker evaluation analysis. Therefore, it is hard to specify a universal cutoff for R^2-based measures, above which a potential surrogate be deemed "valid." Further, once a surrogate has been adopted, or even before, it is important to assess how it will perform in a *new* trial. Burzykowski and Buyse (2006) proposed the so-called *surrogate threshold effect* (STE). Their method is intended for deriving a sample size large enough for a treatment effect on the surrogate endpoints to translate into a meaningful and significant effect on the true endpoint. Methods discussed in this chapter are situated within the meta-analytic framework. This is also true for the recent work by Baker (2006), who proposed the use of average prediction error based on easy-to-implement regression models.

Given that R^2_{h} reduces to the R^2_{trial} at the trial level, we shall mainly focus at the individual level when applying this method in subsequent chapters. The following two chapters investigate the performance of the information-theoretic approach in a setting where either the clinical or both endpoints are binary.

7

Information-Theoretic Validation with Mixed Continuous and Binary Endpoints

Validation of surrogate markers within the meta-analytic framework of Buyse et al. (2000) has been extended to settings with nonnormal endpoints. In this chapter, we review an extension to mixed continuous and binary outcomes (Section 7.1) as well as apply the information-theoretic approach to this setting. As discussed in the previous chapter, the information-theoretic measure, R^2_{h} reduces to the quantities previously introduced in the literature. Additionally, Alonso et al. (2005) established that its estimator, LRF, reduces to R^2_{ind} and R^2_{Λ} in the cross-sectional normal-normal and the longitudinal-longitudinal cases, respectively.

However, the performance of the information-theoretic approach has not been investigated with a binary endpoint. It should be noted that unlike the case for normally distributed endpoints, the individual-level surrogacy within the meta-analytic framework is quantified at a latent scale (Molenberghs, Geys, and Buyse 2001, Section 7.1). A consequence of the fact that the observed binary endpoint is considered to be a dichotomized version of a latent continuous variable. On the other hand, ITA quantifies surrogacy at the observed scale, thereby motivating the investigation of its performance in such a setting. Here, we investigate the performance of ITA in the continuous-binary setting through a simulation study, as well as the operational characteristics of various ways to derive confidence intervals for measures of surrogacy. Also, the assumption of linearity, inherent in the models, is put to the test. The simulation study is reported in Section 7.2.

7.1 Mixed Continuous and Binary Endpoints

It is not uncommon to encounter statistical problem where various outcomes of a combined nature are observed, 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. Burzykowski, Molenberghs, and Buyse (2005) review extensions of the meta-analytic approach, ranging over continuous, binary, ordinal, time-to-event, and longitudinally measured 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 response. One can choose either the continuous or the discrete outcome for the marginal model (Molenberghs and Lesaffre 1994). 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, using bivariate generalized linear mixed models. We focus on the case where the true endpoint is continuous and the surrogate is binary. The reverse case is entirely similar, because the meta-analytic approach treats both endpoints symmetrically.

The generalized linear mixed model (GLMM) (Breslow and Clayton 1993) is the most frequently used random-effects model in the context of discrete repeated measurements. It is a straightforward extension of the generalized linear model for univariate data to the context of clustered measurements, and there is a wide range of software available for fitting these models. However, generalized linear mixed models for endpoints of different data types are challenging (Molenberghs and Verbeke 2005). Thus, we concentrate on a two-stage fixed-effects model with a binary surrogate and a continuous clinical endpoint. 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):

$$\widetilde{S}_{ij} = \mu_{Si} + \alpha_i Z_{ij} + \varepsilon_{Sij}, \tag{7.1}$$

$$T_{ij} = \mu_{Ti} + \beta_i Z_{ij} + \varepsilon_{Tij}, \tag{7.2}$$

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

$$\Sigma = \begin{pmatrix} \frac{1}{(1-\rho^2)} & \frac{\rho\sigma}{\sqrt{(1-\rho^2)}} \\ & \sigma \end{pmatrix}. \tag{7.3}$$

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

$$\begin{pmatrix} \mu_{Si} \\ \mu_{Ti} \\ \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} m_{Si} \\ m_{Ti} \\ a_i \\ b_i \end{pmatrix}, \tag{7.4}$$

where the second term on the right hand of (7.4) is assumed to follow a zero-mean normal distribution with dispersion matrix (3.5). Measures to assess the quality of the surrogate both at the trial and individual level are obtained as in (3.10) and (3.12). Interpretation of these measures and decision making follows the logic laid out in Section 3.2.

7.2 A Simulation Study

The information-theoretic approach is easily applicable because its measure of surrogacy, LRF, can be readily calculated from routine outputs provided by univariate models implemented in standard regression software. The primary objectives of this simulation study is to investigate the performance of this approach, as well as its corresponding asymptotic confidence interval, in the mixed continuous binary setting. Prior to that, we lay out the design of the simulation study followed by a summary of the results obtained.

7.2.1 Design of the Simulation Study

Owing 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. However, 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 model (3.6)–(3.7). Choices made are $\mu_S = 0.5$, $\mu_T = 0.45$, $\alpha = 0.05$, and $\beta = 0.03$. Values assumed for the covariance matrices are:

$$\Sigma = \begin{pmatrix} 3 & 2.4 \\ & 3 \end{pmatrix}, \qquad D = \begin{pmatrix} 3 & 2.4 & 0 & 0 \\ & 3 & 0 & 0 \\ & & 3 & 2.85 \\ & & & 3 \end{pmatrix}.$$

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. The above model 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. One should also keep in mind that correlation between observed binary outcomes can have an upper bound less than one (Aerts et al. 2002). However, the situation is totally different at the trial-level. Based on (7.1) and (7.2), 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 (6.11) and (6.12) or (6.13) and (6.14). In each of these models, an appropriate and commonly used link function, g_T, is the identity link as the clinical endpoint is assumed to be normally distributed. Obviously, the surrogate, S_{ij}, is the corresponding binary outcome and the information-theoretic measures of surrogacy are obtained as described in Chapter 6.

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. Results are shown for the probit link at the individual-level and linear function at the trial-level. Histograms are used to depict results of the secondary objectives.

7.2.2 Simulation Results

The simulation results are presented in Tables 7.2 – 7.9. Both individual-level and trial-level R^2 measures are included, as based on both the mixed-effects models (6.13) and (6.14) as well as the fixed-effects models (6.11) and (6.12). The tables have columns indicating the number of trials, the trial size, median R^2, percentile bootstrap and asymptotic 95% confidence intervals.

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 assess 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, ≤ 60, ITA tends to underestimate the trial-level surrogacy. Nevertheless, the models perform considerably well for large trial sizes, ≥ 100. The mixed-effect models, (6.13) and (6.14), outperform the fixed-effect models, (6.11) and (6.12), 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 a 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 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. Figure 7.1 shows a plot of magnitude of pooled differences of the trial-level estimates between the logit and probit links, for the fixed-effect models with large sample size. Observe 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 con-

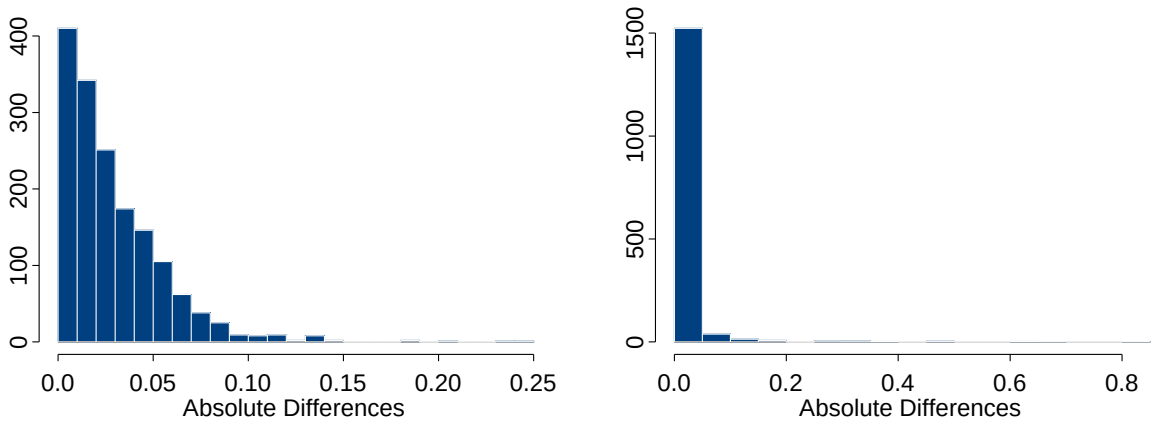

Figure 7.1: *Frequency histogram showing absolute differences obtained for each sample. (a) Difference between probit and logit link functions. (b) Linearity assumption at trial level: linear versus spline-based.*

sidered at the trial level. This is supported by Figure 7.1, as more than 93% of the samples have differences inferior to 0.04.

7.3 Analysis of the ARMD Data

The motivating study introduced in Section 2.2.1, coming from a study on age-related macular degeneration, will now be analyzed. Patients with macular degeneration progressively lose vision. In the trial, the patients visual acuity was assessed through their ability to read lines of letters on standardized vision charts. These charts display lines of five letters of decreasing size. Each line with at least four letters correctly read is called one 'line of vision. The patients visual acuity is the total number of letters correctly read. Here we investigate if the loss of two lines of vision at 6 months (binary) could be used as a surrogate for the visual acuity at 1 year (treated as a continuous endpoint).

Table 7.1: *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)	

Again, it is natural to consider center as the unit of analysis, as the data comes from a multi-center trial. There were 36 centers, each treating between 2 and 18 patients. The two-stage meta-analytic approach and the corresponding ITA models, described in Section 7.1 and Chapter 6 respectively, have been applied to this dataset and results displayed in Table 7.1. Figure 7.2 graphically presents the surrogate by true endpoint treatment effect pairs for all centers. The circles are proportional in surface to the trial size, so as to graphically indicate the centers' relative importance.

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-

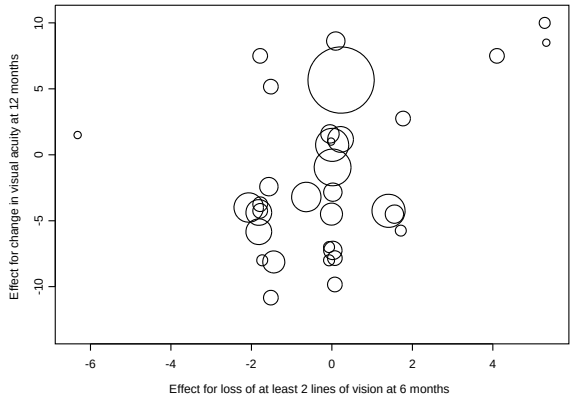

Figure 7.2: *Age-related macular degeneration trial. Treatment effect on the surrogate endpoint by treatment effect on the true endpoint, for each center. Circle surfaces are proportional to the centers' sample sizes.*

estimation of R^2_{indiv} by ITA at the observed scaled. As mentioned earlier, it may be bounded above by a number less than one. 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.

7.4 Discussion

In this chapter, we reviewed the meta-analytic strategy of Buyse et al. (2000), its extension to 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 involves substantial mathematics yet it is more practically feasible than the meta-analytic approach as it depends on simple univariate models.

Here, we investigated the performance of the ITA for combined continuous and binary endpoints, particularly continuous surrogate and binary true 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 we referred to as fixed-effects models, was outperformed by its extension to generalized linear 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}) developed by Alonso and Molenberghs (2007) performed better than bootstrap confidence intervals used by Alonso et al. (2005) in the sense of being generally more narrow. On the other hand, the asymptotic confidence intervals are computationally advan-

tageous 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 alleviated 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.

Table 7.2: *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	0.1537	(0.0600;0.2512)	(0.1099;0.2095)
2	5	150	0.1580	(0.0816;0.2706)	(0.1245;0.2073)
3	5	200	0.1649	(0.0826;0.2719)	(0.1309;0.2030)
4	5	300	0.1553	(0.0821;0.2680)	(0.1341;0.1926)
5	10	100	0.1652	(0.0974;0.2299)	(0.1269;0.1993)
6	10	150	0.1610	(0.0875;0.2427)	(0.1335;0.1927)
7	10	200	0.1651	(0.0801;0.2362)	(0.1375;0.1887)
8	10	300	0.1640	(0.0961;0.2353)	(0.1408;0.1827)
9	20	100	0.1624	(0.1119;0.2107)	(0.1356;0.1870)
10	20	150	0.1607	(0.1094;0.2106)	(0.1403;0.1823)
11	20	200	0.1596	(0.1162;0.2055)	(0.1433;0.1797)
12	20	300	0.1590	(0.1103;0.2069)	(0.1459;0.1756)
13	30	100	0.1633	(0.1203;0.1967)	(0.1417;0.1839)
14	30	150	0.1592	(0.1159;0.1992)	(0.1434;0.1777)
15	30	200	0.1578	(0.1134;0.1953)	(0.1440;0.1736)
16	30	300	0.1601	(0.1215;0.1961)	(0.1473;0.1715)

.

Table 7.3: *Simulation study. Univariate mixed-effects model for large trial sizes, trial-level surrogacy.*

	strategy			**confidence intervals**	
sim. #	# trials	# subjects	R^2_{indiv}	**percentile**	**asymptotic**
1	5	100	0.8159	(0.0752;0.9804)	(0.3016;0.9360)
2	5	150	0.8323	(0.0864;0.9823)	(0.3305;0.9481)
3	5	200	0.8619	(0.1321;0.9870)	(0.3545;0.9594)
4	5	300	0.8807	(0.3362;0.9884)	(0.4123;0.9761)
5	10	100	0.7696	(0.0043;0.9211)	(0.3614;0.8849)
6	10	150	0.7975	(0.0144;0.9417)	(0.3905;0.8899)
7	10	200	0.7973	(0.1295;0.9422)	(0.3876;0.9031)
8	10	300	0.7870	(0.0374;0.9502)	(0.4019;0.8984)
9	20	100	0.7628	(0.1500;0.8841)	(0.4583;0.8701)
10	20	150	0.8060	(0.4690;0.9018)	(0.5229;0.9073)
11	20	200	0.8075	(0.2821;0.9092)	(0.5171;0.8975)
12	20	300	0.8289	(0.2745;0.9170)	(0.5626;0.9177)
13	30	100	0.7770	(0.2732;0.9041)	(0.5440;0.8706)
14	30	150	0.7835	(0.4593;0.8963)	(0.5672;0.8892)
15	30	200	0.7976	(0.5068;0.9130)	(0.5823;0.8941)
16	30	300	0.8158	(0.5118;0.9234)	(0.6072;0.9024)

Table 7.4: *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	0.1511	(0.0575;0.2483)	(0.1068;0.2054)
2	5	150	0.1553	(0.0803;0.2688)	(0.1222;0.2045)
3	5	200	0.1659	(0.0796;0.2696)	(0.1293;0.2011)
4	5	300	0.1542	(0.0811;0.2663)	(0.1330;0.1913)
5	10	100	0.1609	(0.0939;0.2294)	(0.1242;0.1961)
6	10	150	0.1586	(0.0861;0.2410)	(0.1318;0.1907)
7	10	200	0.1626	(0.0790;0.2366)	(0.1362;0.1873)
8	10	300	0.1634	(0.0952;0.2343)	(0.1400;0.1817)
9	20	100	0.1588	(0.1086;0.2089)	(0.1326;0.1836)
10	20	150	0.1576	(0.1071;0.2095)	(0.1383;0.1801)
11	20	200	0.1581	(0.1144;0.2045)	(0.1418;0.1781)
12	20	300	0.1578	(0.1092;0.2061)	(0.1448;0.1745)
13	30	100	0.1602	(0.1159;0.1943)	(0.1387;0.1806)
14	30	150	0.1569	(0.1135;0.1973)	(0.1414;0.1755)
15	30	200	0.1563	(0.1120;0.1934)	(0.1425;0.1720)
16	30	300	0.1590	(0.1202;0.1952)	(0.1463;0.1705)

Table 7.5: *Simulation study. Univariate fixed-effects model for large trial sizes, trial-level surrogacy.*

	strategy			**confidence intervals**	
sim. #	# trials	# subjects	R^2_{indiv}	**percentile**	**asymptotic**
1	5	100	0.8134	(0.0133;0.9852)	(0.2938;0.9274)
2	5	150	0.8004	(0.1640;0.9796)	(0.2884;0.9560)
3	5	200	0.8342	(0.2109;0.9906)	(0.3237;0.9576)
4	5	300	0.8167	(0.1457;0.9827)	(0.3027;0.9564)
5	10	100	0.7350	(0.2608;0.9163)	(0.3272;0.9118)
6	10	150	0.7797	(0.1995;0.9308)	(0.3619;0.9177)
7	10	200	0.7668	(0.1808;0.9478)	(0.3693;0.9166)
8	10	300	0.7755	(0.1742;0.9066)	(0.3645;0.9129)
9	20	100	0.7075	(0.4191;0.8659)	(0.4117;0.8692)
10	20	150	0.7283	(0.4732;0.8865)	(0.4548;0.8903)
11	20	200	0.7399	(0.4631;0.8786)	(0.4549;0.8893)
12	20	300	0.7543	(0.4872;0.8850)	(0.4713;0.8960)
13	30	100	0.7145	(0.4588;0.8440)	(0.4730;0.8516)
14	30	150	0.7200	(0.4596;0.8530)	(0.4818;0.8545)
15	30	200	0.7223	(0.5393;0.8673)	(0.4956;0.8643)
16	30	300	0.7492	(0.5445;0.8747)	(0.5225;0.8754)

Table 7.6: *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	0.1501	(0.0003;0.3844)	(0.0497;0.3338)
2	5	20	0.1443	(0.0255;0.3379)	(0.0670;0.2777)
3	5	40	0.1589	(0.0552;0.3060)	(0.0974;0.2560)
4	5	60	0.1607	(0.0471;0.2810)	(0.0983;0.2251)
5	10	10	0.1503	(0.0520;0.3253)	(0.0679;0.2846)
6	10	20	0.1676	(0.0855;0.2697)	(0.0945;0.2559)
7	10	40	0.1608	(0.0737;0.2506)	(0.1068;0.2196)
8	10	60	0.1576	(0.0903;0.2436)	(0.1152;0.2077)
9	20	10	0.1558	(0.0627;0.2696)	(0.0878;0.2451)
10	20	20	0.1601	(0.0907;0.2138)	(0.1067;0.2203)
11	20	40	0.1623	(0.1055;0.2223)	(0.1216;0.2024)
12	20	60	0.1619	(0.1099;0.2113)	(0.1286;0.1949)
13	30	10	0.1572	(0.0916;0.2443)	(0.0986;0.2287)
14	30	20	0.1634	(0.0994;0.2297)	(0.1207;0.2148)
15	30	40	0.1588	(0.1019;0.2114)	(0.1283;0.1945)
16	30	60	0.1641	(0.1197;0.2096)	(0.1373;0.1919)

Table 7.7: *Simulation study. Univariate fixed-effects model for small trial sizes, trial-level surrogacy.*

	strategy			**confidence intervals**	
sim. #	# trials	# subjects	R^2_{indiv}	**percentile**	**asymptotic**
1	5	10	0.4767	(0.0010;0.9480)	(0.1236;0.8055)
2	5	20	0.5559	(0.0293;0.9429)	(0.1364;0.8745)
3	5	40	0.6232	(0.0800;0.9593)	(0.1429;0.9012)
4	5	60	0.5870	(0.0267;0.9536)	(0.1333;0.8739)
5	10	10	0.3973	(0.0333;0.7484)	(0.0870;0.7478)
6	10	20	0.4644	(0.0567;0.8768)	(0.1371;0.7872)
7	10	40	0.4659	(0.0143;0.8405)	(0.1297;0.7898)
8	10	60	0.5583	(0.0857;0.8743)	(0.1767;0.8338)
9	20	10	0.4139	(0.0649;0.7116)	(0.1483;0.6732)
10	20	20	0.4745	(0.0820;0.7412)	(0.1853;0.7194)
11	20	40	0.4979	(0.1206;0.8008)	(0.2218;0.7467)
12	20	60	0.5251	(0.1257;0.7265)	(0.2155;0.7424)
13	30	10	0.4141	(0.1317;0.6158)	(0.1712;0.6360)
14	30	20	0.4522	(0.2472;0.6932)	(0.2231;0.6853)
15	30	40	0.4855	(0.1755;0.6816)	(0.2367;0.6964)
16	30	60	0.5092	(0.1934;0.6993)	(0.2570;0.7143)

Table 7.8: *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	0.1595	(0.0235;0.3929)	(0.0596;0.3667)
2	5	20	0.1594	(0.0404;0.3332)	(0.0736;0.2920)
3	5	40	0.1669	(0.0567;0.2979)	(0.1008;0.2612)
4	5	60	0.1635	(0.0475;0.2784)	(0.1011;0.2291)
5	10	10	0.1712	(0.0553;0.3263)	(0.0789;0.3054)
6	10	20	0.1777	(0.0924;0.2682)	(0.1007;0.2652)
7	10	40	0.1633	(0.0805;0.2515)	(0.1106;0.2246)
8	10	60	0.1598	(0.0968;0.2401)	(0.1178;0.2110)
9	20	10	0.1773	(0.0823;0.2881)	(0.1053;0.2708)
10	20	20	0.1718	(0.1057;0.2253)	(0.1163;0.2329)
11	20	40	0.1683	(0.1099;0.2262)	(0.1262;0.2080)
12	20	60	0.1657	(0.1127;0.2138)	(0.1318;0.1986)
13	30	10	0.1799	(0.1135;0.2604)	(0.1164;0.2528)
14	30	20	0.1750	(0.1101;0.2380)	(0.1294;0.2256)
15	30	40	0.1637	(0.1070;0.2135)	(0.1330;0.1999)
16	30	60	0.1687	(0.1220;0.2137)	(0.1403;0.1952)

Table 7.9: *Simulation study. Univariate mixed-effects model for small trial sizes, trial-level surrogacy.*

	strategy			**confidence intervals**	
sim. #	# trials	# subjects	R^2_{indiv}	**percentile**	**asymptotic**
1	5	10	0.4173	(0.0010;0.9497)	(0.1272;0.7803)
2	5	20	0.5699	(0.0021;0.9589)	(0.1306;0.8506)
3	5	40	0.5990	(0.0136;0.9389)	(0.1306;0.8702)
4	5	60	0.5914	(0.0024;0.9395)	(0.1359;0.8416)
5	10	10	0.4024	(0.0038;0.7753)	(0.0886;0.7027)
6	10	20	0.4395	(0.0330;0.8527)	(0.1204;0.7794)
7	10	40	0.4754	(0.0047;0.8042)	(0.1250;0.7709)
8	10	60	0.5226	(0.0280;0.8562)	(0.1576;0.8047)
9	20	10	0.4212	(0.0017;0.6803)	(0.1470;0.6495)
10	20	20	0.4769	(0.0406;0.7370)	(0.1886;0.7057)
11	20	40	0.5218	(0.0373;0.7701)	(0.2167;0.7307)
12	20	60	0.5352	(0.0034;0.7554)	(0.2218;0.7201)
13	30	10	0.4268	(0.0415;0.6198)	(0.1753;0.6284)
14	30	20	0.4505	(0.0105;0.7169)	(0.2288;0.6677)
15	30	40	0.5175	(0.1498;0.7066)	(0.2636;0.7106)
16	30	60	0.5322	(0.2769;0.7543)	(0.2756;0.7267)

8

Information-Theoretic Valdiation with Binary Endpoints

The previous chapter highlights some of the issues encountered with extending the validation of surrogate markers within the meta-analytic approach by Buyse et al. (2000), as well as the information-theoretic approach. The main issue being that the individual-surrogacy is quantified at a latent scale for the meta-analytic approach. Even though ITA quantifies the individual-level surrogacy at the observed scale, it may be bounded above by a value less than one, which is usually unknown. Nevertheless, ITA is still a viable alternative, especially when a more complex functional form is necessary to describe the relation between both treatment effects (endpoints) at the trial (individual) level. Additionally, results from the previous chapters indicate that the random-effects model outperforms the fixed-effects model. Therefore, generalized

linear mixed models will have to be considered, even with the information-theoretic approach, when the clinical endpoint is nonnormal.

Considering a binary surrogate for a binary clinical endpoint yields a setting with some peculiar issues: (1) the individual-level surrogacy is quantified at the latent scale; (2) correlation between binary outcomes can be highly constrained, owing to but not limited to, the mean-variance relationship; (3) Rodríguez and Goldman (1995) demonstrated that both Penalized Quasi-likelihood (PQL) and Marginal Quasi-likelihood (MQL), two commonly used and software implemented methods for GLMM based on a linear Taylor expansion, may be seriously biased when applied to binary response data. Thus, we base our estimation for the information-theoretic approach on the maximum likelihood with numerical integration over the random effects (quadrature estimation method).

In this chapter, we review the meta-analytic approach for binary endpoints (Section 8.1) base on the maximum pairwise likelihood (MPL), proposed by Renard et al. (2002) to reduce estimation bias. Further, we investigate how the information-theoretic approach with the quadrature estimation method performs in this setting, and how it fairs against the meta-analytic approach base on MPL. As mentioned earlier, we consider random-effects models, in contrast to marginal and conditional models, to cope with the hierarchical structure of the meta-analytic data, combined with a probit formulation for the pair of surrogate and clinical endpoints.

8.1 Meta-analytic Validation for Binary Endpoints

Extending the meta-analytic approach for continuous endpoints, Renard et al. (2002) adopted a latent variable perspective. They posit the existence of a pair of latent variables $(\widetilde{S}_{ij}, \widetilde{T}_{ij})$ that are continuously distributed and related to the actual pair of

responses (S_{ij}, T_{ij}) through a certain threshold, which can be taken to be 0 without loss of generality. This approach motivates a wide class of models for binary data, of which the standard logistic and probit regression models are special cases (Cox and Snell 1989). With the additional assumption that $(\widetilde{S}_{ij}, \widetilde{T}_{ij})$ is zero-mean normally distributed with covariance matrix Σ, we consider the following random-effects model for the latent variables:

$$\widetilde{S}_{ij} = \mu_S + m_{Si} + \alpha Z_{ij} + a_i Z_{ij} + \widetilde{\varepsilon}_{Sij}, \tag{8.1}$$

$$\widetilde{T}_{ij} = \mu_T + m_{Ti} + \beta Z_{ij} + b_i Z_{ij} + \widetilde{\varepsilon}_{Tij}, \tag{8.2}$$

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 (3.5). Due to identifiability issues, the covariance matrix of the zero-mean normally distributed error terms can be written, without loss of generality, as:

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

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

$$\Phi^{-1}[P(S_{ij} = 1 | m_{Si}, m_{Ti}, a_i, b_i)] = \mu_S + m_{Si} + \alpha Z_{ij} + a_i Z_{ij}, \tag{8.3}$$

$$\Phi^{-1}[P(T_{ij} = 1 | m_{Si}, m_{Ti}, a_i, b_i)] = \mu_T + m_{Ti} + \beta Z_{ij} + b_i Z_{ij}, \tag{8.4}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. Formulation (8.1)–(8.2) allows the use of the coefficient of determination defined in Chapter 3 Section 3.2 without any further modification. However, their interpretation is bound by the postulated latent variables. Thus, the R^2_{trial} is calculated as shown in (3.10), whereas the individual-level R^2_{indiv} is equal to ρ^2_{ST}.

Models (8.3)–(8.4) take the form of a multilevel probit model and belongs to the class of generalized linear mixed models (Breslow and Clayton 1993). 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 Verbeke 2005). Because interest lies in variance-covariance parameters, another approach to parameter estimation is preferable, while keeping the computational burden as low as possible. A procedure fulfilling such requirements is maximum pairwise likelihood (MPL), a form of pseudo-likelihood (Molenberghs and Verbeke 2005). For the specific case of the probit link, as in (8.3)–(8.4), Renard et al. (2002) suggested the use of MPL. For detailed description of MPL see Besag (1975), Lindsay (1988), Molenberghs and Verbeke (2005), and Burzykowski, Molenberghs, and Buyse (2005).

Even with the use of MPL, while fitting the bivariate probit models 8.3–8.4, one may still encounter ill-conditioned variance-covariance matrices as well as convergence issues. Additionally, even without these problems, interpretation of the surrogacy measures may be hampered as they apply to the postulated latent variables rather than to the observed binary variables of interest. Considering these difficulties, the ITA approach is a reasonable alternative. As mentioned earlier, ITA easily handles nonnormal endpoints by choosing an appropriate link function g_T in models (6.11) and (6.12), or (6.13) and (6.14). Given that the clinical endpoint is discrete, the individual-level surrogacy can be quantified based on the expression (6.6), as R^2_{h} may be bounded above by a number less than one. The trial-level surrogacy can be quantified using (6.5) as the second-stage model is for continuous endpoints. Of

course, estimates are obtained using the LRF as described in Section 6.5. In the following section, we investigate the performance of these methods via a simulation study.

8.2 Simulation Study

We expect MPL to provide more stable estimates and be computationally less intensive than PQL and MQL, within the meta-analytic approach. Also, to evade computational and interpretation issues, we recommended the information-theoretic approach based on univariate models. In this section, we investigate the performance of both methods in a setting with a binary surrogate for a binary clinical endpoint, base on a probit formulation.

8.2.1 Design of the Simulation Study

Data were generated based on model (8.1)–(8.2) with parameters set equal to $\mu_S = 0.5$, $\mu_T = 0.45$, $\alpha = 0.05$, and $\beta = 0.03$. The following values assumed for the covariance matrices are:

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

leading to trial- and individual-level surrogacy of 0.90 and 0.64 respectively. 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. It is assumed that a success $S_{ij} = 1$ and $T_{ij} = 1$, respectively, is recorded if $\widetilde{S}_{ij} > \mu_S$ and $\widetilde{T}_{ij} > \mu_T$, respectively, and a failure, $S_{ij} = 0$ and $T_{ij} = 0$, respectively otherwise. 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. Further, we 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), (6.13) and (6.14) with a probit link, in the univariate case. Additionally, the univariate fixed-effects models take the form of the generalized linear models (6.11) and (6.12) with a probit link.

8.2.2 Simulation Results

The simulation results are displayed in Tables 8.2–8.17. Focusing at the trial level, the simulation reveals that the full bivariate random effects model and ITA (univariate) mixed-effects models are consistent in that both models produce surrogacy measures approaching the true values as the number of trials and trial size increases. However, the corresponding fixed-effects models lead to underestimation, even for larger sample sizes. Although the full bivariate model leads to measures at the latent scale, we expect it to be preserved at the observed scale (Alonso et al. 2002). Indeed, 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.

At the individual-level, 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. This comes as no surprise as these models quantify the individual-level surrogacy at the latent scale. However, it is not trivial to translate these estimates from the latent scale to the explicit one. On the otherhand, this issue

is circumvented by the ITA. An important observation is that the values reported with ITA are substantially smaller than their latent counterparts, in line with expectation: (1) switching from the latent scale to the explicitly observed scale reduces association (2) association between correlated binary outcomes is usually bounded by a value less than one. This may be, but not limited to, a manifestation of the fact that important information is lost when switching from a continuous to a binary scale. Customarily, the binary variables are the only ones observed in some real life studies. In this light, the ITA is a fair representation of reality, whereas the other methods may be overly optimistic.

8.3 Application to the Case Study

The meta-analysis of 10 early phase trials assessing the efficacy of several therapies for the treatment of acute migraine crises, introduces in Chapter 2, Section 2.2.3, was analyzed using the methods described in the previous sections of this chapter together with considerations discussed in Section 3.3. Of the symptoms studied: nausea, vomiting, photophobia, phonophobia, the photophobia symptom had the highest trial-level surrogacy. Both point estimates and 95% confidence intervals for both the trial- and individual-level surrogacy are presented in Table 8.1. The univariate and bivariate fixed-effects models result in a smaller R^2_{trial} than their random-effects counterpart. However, R^2_{trial} estimates from the bivariate randome-effects model are unreliable as they are based on an ill-conditioned covariance matrix.

The univariate mixed-effect model performed acceptably well in the simulation, thus, we base our conclusion on results from this model. Therefore, considering that R^2_{trial} for photophobia at the trial level is sufficiently high, the presence of photophobia is a 'good' surrogate for migraine severity. The reasonably good agreement

between the treatment effects on both endpoints, and in addition the absence of obvious outliers, is clear from Figure 8.1. The figure displays a scatter plot of the pairs of treatment effects for each unit. The size of the circles 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 and is supported by the simulation results in Section 8.2.2. This is a consequence of the fact that former is at the latent scale, whereas the ITA works at the interpretationally more relevant explicitly observed scale.

8.4 Discussion

Unlike the previous chapters thus far, the clinical endpoint considered in this chapter is a nonnormal outcome, binary. Here, we focused on settings with a binary surrogate for a binary clinical endpoint. This poses some challenges which had not been addressed in the previous chapters, especially with respect to estimation methods and interpretation. Based on a probit formulation, we considered the meta-analytic approach based on a bivariate probit random-effects model, using maximum pairwise likelihood for parameter estimation to minimize bias. Although bias was minimized, interpretation was an issue as validation within this framework yield estimates for measures of surrogacy at a latent scale rather than the observed scale.

The information-theoretic approach was also employed, eluding both the computational and interpretation issues. However, unlike its previous applications in this thesis, an appropriate link function had to specified for both the surrogate and clinical endpoints. Apparently, the probit link function was considered, although other

Table 8.1: *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.*

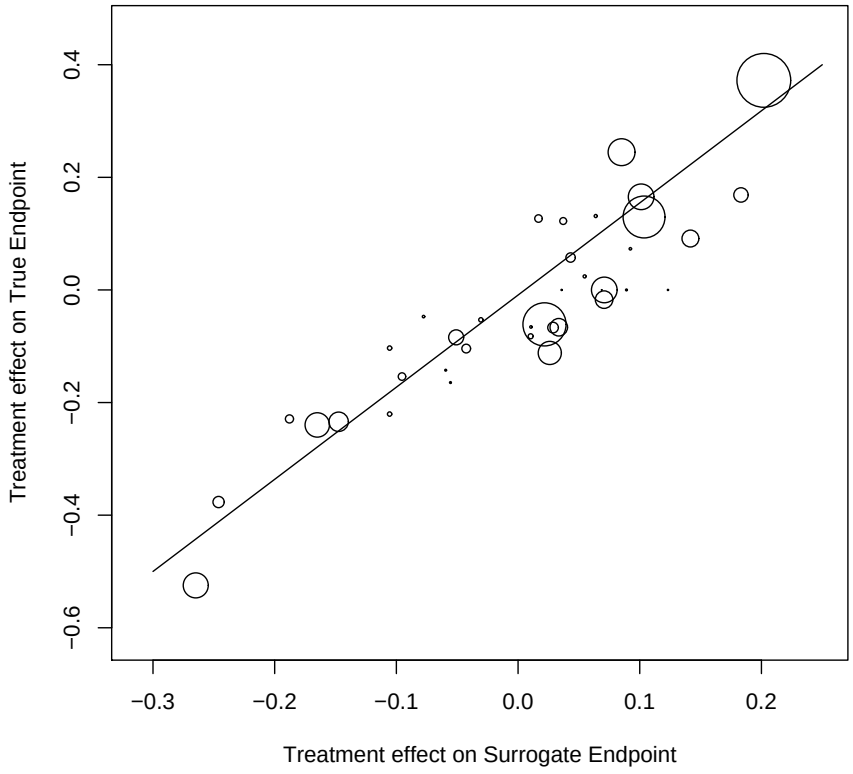

Figure 8.1: *Acute Migraine Study. Bubble plot of trial-specific treatment effect on the surrogate versus true endpoints. The size of the bubbles corresponds to the size of the trial*

appropriate links such as the logit or complementary log-log can be considered as well.

At the trial level, the meta-analytic approach and the ITA base on univariate mixed-effects models performed acceptably well, while their fixed-effects counterparts exhibited some downward bias. Performances of all models increases with increase in both number of trials and trial sizes. At the individual level, 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. Because the ITA operates at the explicitly observed scale, it provides a fairer and more useful quantity.

Applying the proposed methodology to acute migraine trial data indicate that photophobia is a 'good' surrogate at the trial level, although 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. Finally, given the fact that the ITA performs well when the number of trials and number of subjects per trials are large, the unit of analysis should be carefully considered in practice when applying this method.

Table 8.2: *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)

Table 8.3: *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 8.4: *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 8.5: *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 8.6: *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 8.7: *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 8.8: *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 8.9: *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 8.10: *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 8.11: *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 8.12: *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 8.13: *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 8.14: *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 8.15: *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 8.16: *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 8.17: *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)

9

Information-Theoretic Surrogate Marker Validation for Censored Data

There has been a lot of work in the area of surrogate marker validation since Prentice (1989) put forward a formal definition, some 20 years ago. Buyse et al. (2000) proposed an evaluation in a multi-trial framework leading to definitions in terms of the quality of trial- and individual-level surrogacy, the so-called meta-analytic approach. Extensions within the validation framework of Buyse et al. (2000) highlighted some limitations of the meta-analytic approach (Chapter 6). To circumvent these limitations, Alonso and Molenberghs (2007) used information theory to create a unified framework with an intuitive interpretation, which is applicable to a wide range of situations (normal, binary, categorical, and longitudinal outcomes), and is easy to implement in practice.

In practice, the use of surrogate endpoints is a viable option for the need to develop new drugs as quickly as possible, and regulatory agencies have made various provisions and policies for this. However, most of these provisions are limited to diseases where no effective therapies exist. In particular, the U.S. Food and Drug Administration fast track programs are designed to facilitate the development and expedite the review of new drugs that meet two criteria: (1) are intended to treat serious or life-threatening conditions and (2) demonstrate the potential to address unmet medical needs for the condition (Burzykwoski, Molenberghs and Buyse 2005). The most common clinical endpoint for life-threatening diseases is survival time, which is a failure-time (event-time type or censored) outcome.

Although the concept of entropy, and hence R^2_{h}, can be applied even to event-time variables, the consistent estimator *LRF* proposed by Alonso and Molenberghs (2007) is not suited for settings with event-time clinical endpoints. Indeed, O'Quigley and Flandre (2006) remarked that the *LRF* requires some modification if it is to be useful in general situations, specifically to censored data. These authors proposed an adjustment of the LRF (which we refer to as LRF-a) to serve as a partial remedy for censored data. Additionally, we consider another information theory measure, *explained randomness* and its estimator, which we denote by R^2_{XOQ} (Xu and O'Quigley 1999, O'Quigley 2008). Another interesting measure by Kent and O'Quigley (1998), which we denote henceforth as ρ_W, will be discussed for completeness. It should be noted that these measures will be used to evaluate only the individual-level surrogacy, but prior to that we provide a brief overview of censored data and common models for censored data. We shall use the terms censored data and survival data interchangeably.

9.1 Events Time Data

Survival time T is a positive random variable, typically right skewed and with a non-negligible probability of sampling large values, far above the mean. Additionally, the fact that an ordering $T_1 > T_2$, corresponds to a solid physical interpretation are put forward as reasons by some authors to consider techniques other than the classic techniques of linear regression (O'Quigley 2008). This reasoning is incorrect from a purely statistical viewpoint as using transformations and paying careful attention to the structure of the error, linear models are perfectly adequate for dealing with these situations. The most important particularity of survival data is the presence of censoring. A breakthrough in the modelling of censored data occurred with the use of the "at-risk" function.

It is the censoring that forces us to consider other techniques. Typically, censoring is viewed as a nuisance feature of the data, essentially something that hinders us from estimating what it is we would like to estimate. Thus, we have to make some assumptions about the nature of the censoring mechanism in order for our endeavors to succeed. The assumptions may often be motivated by convenience, in which case it is necessary to give consideration as to how well grounded the assumptions are, as well as to how robust are the procedures to departures from any such assumption. An example of the later is considered in the next chapter. In other cases the assumption may appear natural given the context of interest.

Let us set up some notation for time-to-event data that will be required, for a given trial i. The random variables of interest are represented by the vector (T, C, Z), where T is the survival time, C the censoring times and Z is a p vector of explanatory variables. The vector Z may depend upon time, which is precisely the case for some surrogate endpoints settings as proposed by Alonso and Molenberghs (2008).

Time dependency is sometimes indicated through $Z = Z(t)$. For each individual j we observe $X_j = \min(T_j, C_j)$ and $\delta_j = I(T_j \leq C_j)$, where $I(\cdot)$ is the indicator function. Clearly $P_r(X > x) = P_r(T > x, C > x)$ and we describe censoring as being independent, sometimes referred to as non-informative, whenever

$$P_r(X > x) = P_r(T > x, C > x) = P_r(T > x) P_r(C > x).$$

Such censoring often occurs in animal experimentations where the censoring time, when the study is stopped, for all individuals being censored is equal, *Type I censoring*. Another scenario occurs when the proportion of censoring is determined in advance, *Type II censoring*.

Type III censoring occurs in clinical trials, where a model for the patients' random entry is assumed to be uniform over a fixed study period, and subjects can be censored because (1) the end of the study period is reached, (2) they are lost to follow-up, (3) the subjects fail due to something unrelated to the event of interest. Unlike for *Type I* or *Type II* censoring, the C_j could all be distinct for $j = 1, \ldots, n_i$. For such data, we are unable to estimate the joint distribution of the pair (T, C), only the marginal distributions can be estimated under the independent censoring assumption (Tsiatis 1975). This assumption is strong but may be plausible in some situations. Otherwise, censoring is informative and a model for censoring has to be explicitly introduce when estimating the main quantities of interest. We assume independent censoring mechanism throughout this thesis.

Basic quantities used to describe failure time data is the survival function, given by $S(t) = 1 - F(t) = P(T > t)$, which is the probability of surviving beyond time t, and the hazard function (also referred to as the intensity function)

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}.$$

The hazard function can be interpreted as the instantaneous failure rate, conditional on having survived up to time t. The survival function can be rewritten in terms of the cumulative hazard function

$$S(t) = \exp\left\{-\int_0^t \lambda(u)du\right\} = \exp -\Lambda(t),$$

and can be estimated in a nonparametric way using the product-limit estimator, proposed by Kaplan and Meier (1958). We denote the left continuous version of the Kaplan-Meier estimator of survival by $\hat{S}(t)$ and the Kaplan-Meier estimate of the distribution function by $\hat{F}(t) = 1 - \hat{S}(t)$.

9.1.1 Events Time Data and Information Gain

Let $Y_j(t)$ denote the 'at risk function,' indicating whether subject j is at risk ($Y_j(t) = 1$) or not ($Y_j(t) = 0$) at time t. Our regression model for T, given the observed values of the covariates z of Z, is $f(t|z)$, the conditional density, although it is more frequently expressed equivalently in terms of the hazard. In particular, the proportional hazards (PH) model (Cox 1972) specifies that the intensity function can be written as

$$\lambda(t|Z(t)) = Y(t)\lambda_0(t)\ \exp\{\beta Z(t)\}, \tag{9.1}$$

where λ_0(t) is a fixed but unknown "baseline" hazard function, and β is a $p \times 1$ vector of unknown coefficients to be estimated and is interpretable as the log relative risk for binary covariates. The baseline hazard function $\lambda_0(t)$ can be specified to be of a power form or a constant, in which cases the Weibull and exponential models are recovered (Kalbfleisch and Prentice 1980, Cox and Oakes 1984). More generally, $\lambda_0(t)$ in (9.1) is unspecified, resulting in the use of inferential procedures other than ordinary maximum likelihood (Cox 1975, Andersen and Gill 1982). All parameters can be estimated using standard techniques, maximum likelihood in particular, based on

some conditions on the censoring variable C. This work is based on the independent (non-informative) censoring assumption (Tsiatis 1975, Duchateau and Janssen 2008, O'Quigley 2008). In practice, standard techniques are rarely used, most likely as a consequence of the attractive proposal, based on *partial likelihood*, of Cox (1972).

Now, suppose that the covariate $Z = (\beta_1', \beta_2')$ is a $p + q$ dimensional vector, where Z_1 is a vector of baseline covariates such as treatment effect and Z_2 is a vector of surrogate endpoints, possibly time-dependent. For simplicity, we drop the dependence of the surrogate endpoints on time in the notation. Also β is partitioned in the corresponding fashion as (β_1', β_2'). We seek a measure of partial dependence between Z_2 and T, allowing for the regression of Z_1. Let β_0 denote the true value of β, and denote $H_0 : \beta_2 = 0$ and H_1: no constraint on β. Also, let $\widetilde{\beta}_1$ be the value maximizing β under H_0.

Under model (9.1), when $\beta_2 = 0$ there is no association between T and the surrogate endpoints Z_2. A population measure of strength of association, or the distance between the two models indexed by $\beta = \widetilde{\beta}_1$ and $\beta = \beta_0$, can be provided by twice the Kullback-Leibler information gain $\Gamma_1(H_1, H_0; \beta_0) = 2\left[I_1(\beta_0, \beta_0) - I_1(\widetilde{\beta}_1, \beta_0)\right]$, where

$$I_1(\beta, \beta_0) = \int_{\mathcal{Z}} \int_{\mathcal{T}} \log\{f(t|z;\beta)\} f(t|z;\beta_0) dt dG(z). \tag{9.2}$$

In the above expression, the domains of definition of T and Z are denoted by $\mathcal{T}$ and $\mathcal{Z}$, respectively, and $G(z)$ is the marginal distribution function of Z. Assuming no censoring, a standard estimate of information gain will be provided by n^{-1} times the conventional likelihood ratio statistics. Kent (1986) studied the fitted information, having similar statistical properties, in which $I_1(\beta_0, \beta_0)$ and $I_1(\widetilde{\beta}_1, \beta_0)$ are estimated by

$$\hat{I}_1(\beta, \hat{\beta}) = n^{-1} \sum_{j=1}^{n} \int_{\mathcal{T}} \log\{f(t|Z_j;\beta)\} f(t|Z_j;\hat{\beta}) dt, \tag{9.3}$$

$\beta = \widetilde{\beta}_1$ and $\beta = \hat{\beta}$, a consistent estimate of β_0, respectively. Additionally, the marginal distribution of Z has been replaced by its empirical estimate. Furthermore, the variable t enters into the expression as a dummy variable so that the actual values, some of which we might have not been able to observe in the presence of censoring, do not affect the calculation beyond their effect on the estimation of β (Kent and O'Quigley 1988, O'Quigley, Xu and Stare 2005). Thus, the concept of information gain can be readily applied to right censored data.

We focused on proportional hazards model above because some of the estimators discussed in the subsequent sections are based on the PH assumption. Instead of modelling the hazard function or the survival function, we can also model the survival time directly. This can be done through the loglinear model,

$$\log T_j = \mu + \beta Z_j + \sigma E_j, \tag{9.4}$$

with T_j the event time for subject j, μ the intercept, σ the scale parameter, and finally E_j the random error term for subject j. The random error term is assumed to have a fully specified distribution. For instance, assuming a normal, logistic, and Gumbel distribution for E_j, we obtain respectively the lognormal, loglogistic, and Weibull distributions for the event times. A scaled version of the overall mean μ and the regression coefficients β can be considered,

$$\log T_j = \sigma\mu + \sigma\beta Z_j + \sigma E_j, \tag{9.5}$$

thereby facilitating comparison with the proportional hazards model (9.1) (Duchateau and Janssen 2008).

As mentioned in Section 6.5, estimators for R^2_{h}, other than the LRF, could be used in principle. The following sections of this chapter provide concise description of the estimators considered, beginning with LRF for censored data.

9.2 Likelihood Reduction Factor for Censored Data

Observe that LRF as defined in Section 6.5 uses G_i^2/n_i as an estimator for $2I(X,Y)$ in (6.5), which is a standard estimate for information gain assuming no censoring. Alonso and Molenberghs (2007) showed that, 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, implying that LRF is a consistent estimator of R_h^2. Furthermore, a meta-analytic estimate can be obtained as described in (6.7).

We now consider the application of the concept of LRF to event times endpoint such as survival time with (9.1) as basis. Consequently, for event time endpoints, (6.11) and (6.12) translate respectively to

$$\lambda(t_{ij}|z_{ij};\beta_{1i}) = Y_{ij}(t)\lambda_{0T_i}(t)\ \exp(\beta_{1i}z_{ij}), \tag{9.6}$$

$$\lambda(t_{ij}|z_{ij},s_{ij};\theta_{1i},\theta_{2i}) = Y_{ij}(t)\lambda_{0T_i}(t)\ \exp\left[\theta_{1i}z_{ij}+\theta_{2i}s_{ij}(t)\right], \tag{9.7}$$

and LRF_{i} is estimated similarly to (6.15) with G_{i} being the partial log-likelihood ratio test statistic to compare (9.6) and (9.7). This result applies equally well to multiple surrogates.

9.3 Adjusted Likelihood Reduction Factor (LRF-a)

As a reaction to the work of Alonso et al. (2004), O'Quigley and Flandre (2006) remarked that the LRF as proposed by Alonso et al. requires some modification if it is to be useful in general situations, especially in the presence of censoring or event times clinical endpoints. Indeed, the LRF as it stands depends upon the censoring mechanism, for event time clinical endpoints, even when independent of the failure mech-

anism. The population equivalent of the LRF, R^2_h, fits in with the well-established theory of measuring information gain and explained randomness (O'Quigley 2008). This can be carried out for a time-dependent covariate opening up the way to analyzing the impact of surrogate endpoints.

The main difficulty with the LRF is how to adequately deal with censoring for event time clinical endpoints. This may not be an issue of much concern for low levels of censoring. However, for high levels of censoring, it would be useful to have a coefficient which is not impacted by the censoring mechanism. This is quite easily achieved and amounts to working with the LRF but weighting it differently. The theoretic background of the population equivalent for event time endpoint has been described by Kent (1983), Kent and O'Quigley (1988), Xu and O'Quigley (1999), O'Quigley, Xu, and Stare (2005) and O'Quigley (2008), who propose non-trivial weights of the partial log-likelihood to minimize the effect of censoring. As a simple and easily obtainable approximation, O'Quigley and Flandre (2006) proposed redefining the LRF as

$$\text{LRF-a}_i = 1 - \exp\left(-\frac{G_i^2}{k_i}\right), \tag{9.8}$$

with G_i being the partial log-likelihood ratio test statistic to compare (9.6) and (9.7), and k_i the total number of events experienced in trial i. In the absence of censoring $\text{LRF}_i = \text{LRF-a}_i$, since $n_i = k_i$. In the presence of independent censoring, using n_i instead of k_i leads to a coefficient which will depend upon an independent censoring mechanism regardless of population effects. In particular, it approaches zero as the percentage of censored observation approaches one. More generally, in the presence of independent censoring, LRF-a_i is a better approximation because it can be viewed as the empirical expectation of R^2_h (O'Quigley 2008). This also makes sense intuitively for event time data, given that the information about the parameters of interest increases with increasing numbers of events (k_i) rather than with increase in sample

size (n_i).

9.4 Measure of Explained Variation (ρ^2_{w})

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 developed by Kent (1983). They developed these ideas, obtaining simple, multiple and partial coefficients for proportional hazards regression, which are robust to the percentage of censored observations (O'Quigley, Xu, and Stare 2005). Their approach was based upon the idea of transforming a general proportional hazards model to a specific one of Weibull form.

Without loss of generality let us first assume that there are no censored observations. Consider two random variables Z and Y and let G(z) denote the marginal distribution of Z and let the conditional distribution of Y given Z be modelled by (9.5). If we assume that the probability density function $f(y)$ for the random error term is a Gumbel density, we get a Weibull regression model as mentioned earlier. Now let $\mathbf{\Theta} = (\beta, \mu, \sigma^2)$ 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 $\mathbf{H_0}$: $\beta_1 = 0$ and $\mathbf{H_1}$: $\beta_1 \neq 0$. The objective here is to measure the dependence between Y and Z_2 after allowing the regression on Z_1. Now let Θ_0 be the value of Θ which maximizes the expected log-likelihood:

$$\Phi(\Theta, \Theta_1) = \int \int \log\{f(y|z;\theta)\} f(y|z;\theta_1) dy G(dz) \tag{9.9}$$

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: $\Gamma = \Gamma(\mathbf{H_1}, \mathbf{H_0}, G) = 2\{\Phi(\theta_1;\theta_1) - \Phi(\theta_1;\theta_0)\}$. Kent (1983) proposed the

following measure of dependence between Y and Z_2 after allowing the regression on Z_1

$$\rho^2_W = 1 - e^{-\Gamma}. \tag{9.10}$$

Kent and O'Quigely denoted the measure by ρ^2_W in order to emphasize the relationship to the Weibull distribution. These authors appealed to the concept of fitted information gain (9.3) to provide estimates for (9.9), and hence $\rho_{\mathrm{W}}2$. In principle, other possible accelerated failure time and proportional hazard models could be considered. They choose the Weibull distribution so because the resulting integrals can be worked out explicitly. Thus, expression (9.3) can be evaluated leading to an estimate of ρ_W.

A limitation of using this measure is that the calculations are numerically involved, although a SAS macro has been made available (Heinzl and Stare 2000). Its primary limitation is that it can not be applied to PH models with time-dependent variables. Most of the surrogate marker settings encountered with a time-to-event clinical endpoint lead to time-dependent covariates. Consequently, we shall pay little attention to this measure in subsequent chapters.

9.5 Explained Randomness (R^2_{XOQ})

As mentioned before, O'Quigley and Flandre (2006) remarked that the LRF required some modification, which can be achieved by weighting the quantities contributing to the LRF differently. Although they proposed the LRF-a as a good, simple, and easily attainable upgrade, it is still affected by the percentage of censored observations, although less so than the LRF. This downside encourages the need for other measures of dependence which are more robust to the percentage of censored observations, yet based on the concept of information theory. Here, we recall the work of Xu and O'Quigley (1999, 2000), which adapts the idea of using information gain to obtain

a measure of dependence for proportional hazards regression. We focus on their work because it can be applied to time-dependent covariates as well as to partial dependence.

The fitted information enables us to apply the concept of information gain to right censored data. This observation motivated Kent and O'Quigley (1988) to develop a measure of dependence based on (9.3), which is independent of the percentage of censored observations. Nonetheless, the procedure is not straightforward and inference for the resulting estimate is even less so. In addition, the fact that it is not applicable to time-dependent covariates motivated Xu and O'Quigley (1999) to develop a measure arising naturally in the context of proportional hazards regression. Their measure can be obtained in a straightforward manner, inference is also straightforward, and it easily extends immediately to time-dependent covariates.

Unlike Kent and O'Quigley (1988), who worked with the family of conditional distributions of the survival time T given the covariate vector Z, Xu and O'Quigley (1999) base their work on the conditional distribution of Z given T. This idea was motivated by; (1) for a normal model, the resulting measure of explained randomness is unaltered by the way in which the conditioning is done; (2) for other models, the results will often be close (Kent 1983); (3) being able to predict which subject is to fail at any given failure time is equivalent to being able to predict failure rankings of all failed subjects; and (4) studying the conditional distribution of Z given T, does in fact correspond to the way in which inference is carried out in proportional hazards regression. Additionally, this ensures that the measure is rank invariant to monotonic increasing transformations of time (O'Quigley 2008). Thus, rather than (9.2), consider

$$I_2(\beta, \beta_0) = \int_{\mathcal{T}} \int_{\mathcal{Z}} \log\{g(z|t;\beta)\} g(z|t;\beta_0) dz dF(t), \tag{9.11}$$

where $F(t)$ is the marginal distribution function of T, and $g(z|t)$ is the conditional density or conditional probability function of Z given T. It is assumed, for the most part, that there is enough information in the tail of F. As before, the information gain is given by $\Gamma_2(H_2, H_0; \beta_0) = 2\left[I_2(\beta_0, \beta_0) - I_2(\widetilde{\beta}_1, \beta_0)\right]$ and

$$R^2_{XOQ} = 1 - \exp\left[-\Gamma_2(H_2, H_0; \beta_0)\right].$$

Proceeding in a semiparametric way, the Kaplan-Meier (1958) estimator can be used as a consistent estimator of F in the presence of censoring. Also, consider

$$\pi_j(t; \beta) = \frac{Y_j(t)\exp(\beta Z_j)}{\sum_{l=1}^n Y_l(t)\exp(\beta Z_l)},$$

known to be the conditional probability of choosing individual j, given all the individuals at risk at time t and that one individual is to be selected to fail. The product of the $\pi' s$ over the observed failure times is the partial likelihood (Cox 1972, 1975) of the proportional hazards model, which gives our estimate $\hat{\beta}$. Also, under model (9.1), the conditional distribution function of Z given T is consistently estimated by

$$\hat{P}(Z \leq z|T = t) = \sum_{\{j: Z_j \leq z\}} \pi_j(t; \hat{\beta}).$$

Let $t_1 < \ldots < t_K$ be the distinct failure times. We estimate the conditional distribution of Z given T by $\{\pi_j(t; \hat{\beta})\}_j$, and the marginal distribution of T by the Kaplan-Meier estimate, with $W(t_k)$ being the jump of the Kaplan-Meier curve at time t_k. Then

$$\Gamma_2(H_2, H_0; \beta_0) = 2\int_{\mathcal{T}}\int_{\mathcal{Z}} \log\left[\frac{g(z|t; \beta_0)}{g(z|t; \widetilde{\beta}_1)}\right] g(z|t; \beta_0) dz dF(t) \tag{9.12}$$

can be consistently estimated by

$$\hat{\Gamma}_2(H_2, H_0; \hat{\beta}_0) = 2\sum_{k=1}^K W(t_k) \sum_{j=1}^n \pi_j(t_k; \hat{\beta}_0) \log\left[\frac{\pi_j(t_k; \hat{\beta}_0)}{\pi_j(t_k; \widetilde{\beta}_1)}\right], \tag{9.13}$$

where the second sum is effectively over those subjects that are in the risk set at time t_k. Whenever there is a finite limit τ of our observation time, intuitively no more information is to be gained beyond τ. Therefore, we can look at the information gain conditioning on $T \leq \tau$, which is consistently estimated by the proposed estimator divided by $\sum_1^K P(t_k)$. It inevitably depends on the upper limit τ reflecting its faithfulness to the data, especially since the proportional hazards assumption can often hold only for limited situations. This can be viewed as appropriate in that the conditioning is on the time span actually studied rather than a time span of potential interest.

Additionally, working with the conditional distribution of the covariate Z given the survival time T implies that R^2_{XOQ} will be bounded by a number strictly less than 1 for discrete covariates taking on very few levels, as mentioned before in the final paragraph of Section 6.2. Even so, for the most extreme case of a single binary covariate, the bound is close enough to 1 for the phenomenom to be practically ignored, except for very high values of R^2_{XOQ} (O'Quigley 2008).

Conditioning Z on T leads to great simplification, as the computation for estimating the information gain only involves those quantities routinely calculated in a proportional hazards analysis. Another consequence is the fact that R^2_{XOQ} increases with $|\beta|$ and we can directly infer confidence intervals for R^2_{XOQ} from those for β (O'Quigley 2008). Suppose that a 95% confidence interval for β is $0 < \beta_L < \beta < \beta_U$, then an interval for R^2_{XOQ} is $\left[R^2_{XOQ}(\beta_L), R^2_{XOQ}(\beta_U)\right]$. In practice, we do not know $R^2_{XOQ}(\cdot)$, but, following the consistency results, we can "plug" β_L and β_U into $R^2_{XOQ}(\cdot) = 1 - \exp\left[-\hat{\Gamma}_2(\cdot)\right]$ to obtain an approximation of $R^2_{XOQ}(\beta_L)$ and $R^2_{XOQ}(\beta_U)$. We can similarly deal with the case where $\beta_U < \beta < \beta_L < 0$. See O'Quigely (2008) for full details when β_L and β_U have different signs. The coverage properties will not

be exactly the same but will be close. This aspect will be studied further through a simulation study in the Chapter 11.

We can view R^2_{XOQ} either as an estimator of the explained randomness in Z given T or as an approximation to the explained randomness in the ranks of T given Z. This second interpretation is the one that corresponds most closely to our application, surrogate endpoint validation. Both interpretations are equal for bivariate normal models and are anticipated to be generally close. O'Quigley (2008) discuss some conditions required for equality of both interpretations for proportional hazards regression and noted that it is unlikely to be very far removed from such conditions in practice.

9.6 Discussion

This chapter further extends the application of the information-theoretic approach to event-time clinical endpoints, with focus on proportional hazard regression models. Generally, all measures considered have intuitive interpretation and range between 0 and 1. We described how the LRF can be calculated from these models, as well as an adjusted version, LRF-a, which tries to minimize the impact of censoring. Of the measures considered in this chapter, they are the least computationally complex. Additionally, two other estimators ρ^2_{W} and R^2_{XOQ}, which are robust to the percentage of censored observations, were considered. It is obvious that LRF, LRF-a and ρ^2_{W} are not entirely based on the proportional hazards assumption, unlike R^2_{XOQ}. Thus, ρ^2_{W} appears to be the most suited estimator for R^2_{h} in the presence of censoring and absence of time-dependent surrogates. However, it is computationally involved and inference on its estimates is even more so. Additionally, it is not applicable to time-dependent covariates, which is usually encountered with event-time clinical endpoints.

On the other hand, R^2_{XOQ} is applicable to time-dependent covariates, easy to calculate and inference on its estimates are straight forward. However, its high dependency on the PH assumption is a call for concern. Nevertheless, it is required that a good measure of dependence should perform satisfactorily also when applied to misspecified models. The next chapter investigates how these measures fair when the PH assumption is violated in the presence of censoring. Also, LRF and LRF-a reduce to the multiple R^2 in the case of a normally distributed clinical endpoint, while R^2_{XOQ} is expected to produce a close estimate in such settings. This aspect will also be investigated in the next chapter.

10

ITA for Censored Data: Some Computational Issues

To evade computational and interpretational issues encountered by extending validation within the meta-analytic framework of Buyse et al. (2000), Alonso and Molenberghs (2007) developed a unified framework using information theory with an intuitive interpretation, and is applicable in a wide range of situations (normal, binary, categorical, and longitudinal outcomes). For these situations, the authors proposed the LRF, a consistent estimator, to be used in practice. In the previous chapter, we extended the application of the information-theoretic approach to situations with event-time clinical endpoints. Further, we described how the LRF can be calculated for such situations, together with an adjusted version LRF-a proposed by O'Quigley and Flandre (2006). These estimators are expected to be affected by the percentage of censored observations, which is typical of event-time data. Here, we investigate the performance of these estimators both in the presence and absence of censoring.

Furthermore, two estimators robust to the percentage of censoring were considered ρ^2_{W} and XOQ_2 The later depends entirely on the PH assumption, which when satisfied is only an approximate of R^2_{h}. Many simulations have been performed to investigate the performance of these measures under the proportional hazards assumption (Schemper and Stare 1996, O'Quigley, Xu and Stare 2005), demonstrating their robustness to the percentage of censored observations under this assumption.

Within the context of validating a cross-sectional surrogate for a time-to-event clinical endpoint, the objectives of this chapter is two fold (1) to investigate how the said measures perform when the PH assumption is violated, in the presence of censoring, and (2) to apply these measures in the evaluation of individual-level surrogacy, using a motivating study, in a situation with a time-to-event clinical endpoint and a cross-sectional continuous surrogate.

10.1 A Simulation Study

This section grants us insight into the performance of the information-theoretic measures LRF, LRF-a, ρ^2_{W} and R^2_{XOQ} when the PH assumption is violated, with and without censoring.

10.1.1 Design of the Simulation Study

Here we describe the design of the simulation study used to compare the performance of the measures, together with the motivation for the various choices made on the design. As mentioned earlier, we intend to violate the proportional hazards assumption. Also, we intend to investigate how well R^2_{XOQ} approximates the 'classical' R^2 from a traditional normal-linear regression model. This requires the residuals to be normally distributed, as well as the absence of censoring. In light of these require-

ments, we considered model (9.5) and assumed a normal distribution for the random error term resulting to a log-normal model for the clinical endpoint. Therefore, data were generated from a bivariate normal distribution, assuming a normally distributed cross-sectional surrogate endpoint. The event-time clinical outcome was obtained by taking the exponential of the resulting continuous variable.

A uniform distribution was used to induce censoring on the exponentiated clinical outcome. The percentage of censored observations is set to either 0, 10, or 35%, which we consider as the small to moderate level of censoring. Censoring percentages of 50, 75 or 90% were also considered, representing high to extreme number of censored observations. The individual-level R^2 values are set to 0.36, 0.64, or 0.81, corresponding to 'poor', 'moderate' and 'good' surrogacy at the individual level. The number of subjects is fixed to either 20, 50, 100, 200 or 1000. For full combinations of these parameter values, 100 runs are performed in each case, and the measures of interest calculated.

10.1.2 Simulation Results

The simulation results are displayed in Tables 10.1 and 10.2. It can be observed that in the absence of censoring (Censoring = 0%), LRF and LRF-a provide equal estimates as expected. Also, in the absence of censoring, R^2_{XOQ} yields similar estimates to LRF, and hence LRF-a. Generally, all measures seem to perform adequately with the estimates approaching the true value, from which the data were generated, as the sample size increases with small to moderate level of censoring.

The LRF-a and R^2_{XOQ} slightly overestimate the individual-level surrogacy when the percentage of censoring ranges between 35% and 50%. For similar censoring percentages, LRF underestimates the individual-level surrogacy, even for large sample

sizes. The bias exhibited by LRF, LRF-a and R^2_{XOQ} increases 'unacceptably' as the percentage of censoring increases to 75% and becomes worse at the percentage of censoring reaches 90%. ρ^2_{W} underestimates the individual-level surrogacy as the percentage of censoring increases, but the underestimation subsides as the sample size increases except for high level of association between the surrogate and the true endpoint.

These observations indicate that the LRF is highly affected by the percentage of censored observation. Also, as indicated by O'Quigley and Flandre (2006) LRF-a performs well for small to moderate percentage of censoring. However, it exhibits upward bias for high levels of censoring. Surprisingly, R^2_{XOQ} performs acceptably well when the PH assumption is violated and the percentage of censoring is low to moderate. As expected, ρ^2_{W} is robust to the percentage of censoring, especially when the sample size is large. Given that event-time clinical endpoints are usually encountered with time-dependent surrogates, we recommend the use of LRF-a or R^2_{XOQ} when the PH assumption is violated and the percentage of censoring is low to moderate. In case of cross-sectional surrogates, we recommend the use of ρ^2_{W}. Investigating the performance within a surrogate marker context when the PH assumption is plausible is the topic of the next chapter.

In the following section, we illustrate how the information-theoretic approach can be employed to a situation with a cross-sectional surrogate for an event-time clinical endpoint, using the measures discussed in this section.

10.2 Application to a Motivating Study

Estimates of R^2_{XOQ}, LRF-a, LRF, and their respective asymptotic confidence intervals, are used to evaluate the different TCD velocities as potential surrogates for time to

first stroke, in the stroke study on children with sickle cell disease. This motivating study was described in Chapter 2, Section 2.2.8. Cross-sectional surrogates were obtained by taking the median of the respective TCD velocities over time. The median age was included in the models with and without the surrogate, corresponding to models (9.6) and (9.7). Also, the PH assumption was tested for each model and the results indicate that the PH assumption is plausible (Table 10.4). Results for individual-level surrogacy measures are shown in Table 10.3.

Given that the PH assumption is plausible, R^2_{XOQ} is expected to be robust to the percentage of censoring, even when it is as high as 90% (Xu and O'Quigley 1999, O'Quigley, Xu, and Stare 2005). In this light, although LRF-a is expected to be affected by such high degree of censoring, its estimates are close to those obtained from R^2_{XOQ}. Obviously, LRF is substantially affected by the high degree of censoring, as expected. Conclusions will be made base on R^2_{XOQ}, as well as LRF-a. It can be observed that estimates from the screened SCD data are consistently higher than their counterparts from the randomized SCD data. This discrepancy may be due to, but not limited to, effects of possible prognostic factors on the screened SCD data. Also, the estimates based on the screened SCD data have narrower intervals, probably due to its larger sample size relative to the randomized SCD data.

Generally, moderate associations were found at the individual level, indicating that the TCD velocities only explain about 50% of our uncertainty about the time to first stroke. However, the asymptotic confidence intervals do not exclude the plausibility of stronger associations. Therefore, we recommend these biomarkers for further studies.

10.3 Discussion

In the previous chapter we defined the LRF for settings with event-time endpoints and considered a modified version LRF-a, both of which are expected to be affected by high percentages of censoring. Thus, two other measures which are expected to be robust to the percentage of censoring were also consider: ρ^2_{W} and R^2_{XOQ}. Substantial literature is available concerning the performance of the later measures when the proportional hazard assumption is met. They indicate that these measures are robust to censoring, even as high as 90%, under the PH assumption. In this chapter, we (1) investigated the performance of all measures when the PH assumption is violated, with increasing percentage of censoring, and (2) evaluated individual-level surrogacy for the stroke study using the measures of interest. Both tasks were done in the context of validating a cross-sectional surrogate for a time-to-event clinical endpoint.

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. The authors observed that, among the other methods suggested, the measure of Kent and O'Quigely was unaffected by censoring even for a substantial percentage of censoring. Similar results were observed for R^2_{XOQ} (Xu and O'Quigley 1999, O'Quigley, Xu, and Stare 2005). Thus, we assume that these measures yield promising results in quantifying the individual-level surrogacy, even with substantial censoring, under the PH assumption. Generally, results from LRF-a, and especially LRF, should not be trusted when the degree of censoring is high even with when the PH assumption seems plausible.

When the PH assumption is violated in the presence of censoring, LRF is not suitable to handle the censoring especially with small sizes. LRF-a performs much better with low to moderate censoring and high sample sizes. ρ^2_{W} performs satisfac-

torily even for moderately large percentage of censoring and reasonable sample sizes. Surprisingly, R^2_{XOQ} performed acceptably with low to moderate censoring and large sample sizes. This indicates that with low censoring and large sizes R^2_{XOQ} may be a good measure of dependence even without the PH assumption.

These measures were used together with their respective asymptotic confidence intervals to evaluate various cross-sectional TCD velocities as potential surrogates of time to first stroke. Conclusions were mainly based on R^2_{XOQ} as the PH assumption was deemed plausible, and its asymptotic confidence interval can be easily estimated, unlike that of R^2_{XOQ}. We recommend the TCD velocities for further studies because confidence intervals do not exclude stronger associations, although they explain only about 50% of our uncertainty about the time to first stroke.

Table 10.1: *Simulation results for 0%, 15% and 35% censored observations. (n: Sample size ; R^2_k: and R^2_n: R^2 based on ITA with # of events and # of subjects as denominator; ρ^2_w: R^2 of Kent and O'Quigely; ρ^2_{xu} : R^2 of Xu and O'Quigely;*

	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 10.2: *Simulation results for 50%, 75% and 90% censored observations. (n: Sample size ; R^2_k: and R^2_n: R^2 based on ITA with # of events and # of subjects as denominator; ρ^2_w: R^2 of Kent and O'Quigely; ρ^2_{xu} : R^2 of Xu and O'Quigely;*

	Censoring =50%				Censoring =75 %				Censoring =90 %			
					$R^2 = 0.36$							
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	0.4496	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 10.3: Evaluating individual-level surrogacy for the stroke study on children with sickle cell disease.

		Randomized		Screened	
Biomarker	**Measure**	**Estimate**	**95% Asymp. CI**	**Estimate**	**95% Asymp. CI**
MaxL	R^2_{XOQ}	0.574	(0.124; 0.880)	0.668	(0.388; 0.906)
	LRF-a	0.592	(0.199; 0.867)	0.672	(0.477; 0.819)
	LRF	0.082	(0.021; 0.176)	0.023	(0.013; 0.034)
MaxR	R^2_{XOQ}	0.415	(0.041; 0.822)	0.733	(0.490; 0.904)
	LRF-a	0.426	(0.080; 0.764)	0.753	(0.575; 0.874)
	LRF	0.055	(0.008; 0.136)	0.027	(0.016; 0.039)
MaxS	R^2_{XOQ}	0.588	(0.153; 0.886)	0.744	(0.507; 0.912)
	LRF-a	0.604	(0.226; 0.866)	0.751	(0.578; 0.871)
	LRF	0.089	(0.026; 0.184)	0.027	(0.017; 0.040)
MaxD	R^2_{XOQ}	0.531	(0.109; 0.888)	0.694	(0.412; 0.917)
	LRF-a	0.541	(0.166; 0.833)	0.700	(0.512; 0.838)
	LRF	0.075	(0.018; 0.165)	0.024	(0.014; 0.036)
MaxRL	R^2_{XOQ}	0.562	(0.105; 0.903)	0.802	(0.587; 0.930)
	LRF-a	0.563	(0.172; 0.853)	0.812	(0.658; 0.909)
	LRF	0.077	(0.018; 0.169)	0.033	(0.021; 0.047)
MaxSD	R^2_{XOQ}	0.590	(0.154; 0.898)	0.748	(0.494; 0.928)
	LRF-a	0.603	(0.225; 0.866)	0.754	(0.582; 0.873)
	LRF	0.089	(0.025; 0.183)	0.028	(0.017; 0.040)
MaxVel	R^2_{XOQ}	0.560	(0.129; 0.887)	0.717	(0.450; 0.919)
	LRF-a	0.574	(0.196; 0.851)	0.723	(0.541; 0.853)
	LRF	0.082	(0.022; 0.175)	0.025	(0.015; 0.037)

Table 10.4: P-values for tests of the PH assumption in both the model without the respective surrogates 'Red.Mod' and the model with the respective surrogates 'Ful.Mod'.

	Randomized		Screened	
Biomarker	**Red.Mod**	**Ful.Mod**	**Red.Mod**	**Ful.Mod**
MaxL	0.684	0.797	0.155	0.180
MaxR	0.805	0.914	0.114	0.267
MaxS	0.820	0.925	0.152	0.197
MaxD	0.820	0.987	0.152	0.230
MaxRL	0.669	0.907	0.117	0.154
MaxSD	0.820	0.953	0.152	0.231
MaxVel	0.820	0.937	0.152	0.230

11

Information Theoretic Approach to Surrogate Markers Evaluation With Failure Time Endpoints

12

A Longitudinal Surrogate for an Event-Time True Endpoint

13

Concluding Remarks and Further Research

References

Aerts, M., Geys, H., Molenberghs, G., and Ryan, L. (2002). *Topics in Modelling of Clustered Data*. London: Chapman and Hall.

Advanced Colorectal Cancer Meta-Analysis Project (1992). Modulation of 5-fluorouracil by leucovorin in patients with advanced colorectal cancer: evidence in terms of response rate. *J. Clin. Oncol.*, **10**, 896–903.

Advanced Colorectal Cancer Meta-Analysis Project (1994). Meta-analysis of randomized trials testing the biochemical modulation of 5-fluorouracil by methotrexate in metatstatic colorectal cancer. *J. Clin. Oncol.*, **12**, 960–969.

Albert JM, Ioannidis JPA, Reichelderfer P, Conway B, Coombs RW, Crane L, Demasi R, Dixon DO, Flandre P, Hughes MD, Kalish LA, Lartnz K, Lin D, Marschner IC, Munõz A, Murray J, Neaton J, Pettinelli C, Rida W, Taylor JMG, and Welles SL (1998). Statistical issues for HIV surrogate endpoints: point and counterpoint. *Stat. Med.*, **17**, 2435–2462.

Alonso, A., Geys, H., and Molenberghs, G. (2006) A unifying approach for surrogate marker validation based on Prentices criteria. *Statistics in Medicine,* **25;** 205–221.

Alonso, A., Geys, H., Molenberghs, G., and Kenward, M.G. (2003). Validation of

surrogate markers in multiple randomized clinical trials with repeated measures. *Biometrical Journal*, **45**, 931–945.

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

Alonso, A., Molenberghs, G., Burzykowski, T., Renard, D., Geys, H., Shkedy, Z., Tibaldi, F., Abrahantes, J., and Buyse, M. (2004a). Prentices approach and the metaanalytic 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. *Stat. Med.,* **25;** 205–211.

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

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

Biomarkers Definitions Working Group (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin. Pharmacol. Ther.* **69**, 89-95.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Burzykowski T and Buyse M. (2006) Surrogate threshold effect: An alternative measure for meta-analytic surrogate endpoint validation. *Pharmaceutical Statistics*

5, 173–186.

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

Burzykowski T, Molenberghs G, and Buyse M. The validation of surrogate endpoints using data from randomized clinical trials: a case-study in advanced colorectal cancer. *J. Roy. Stat. Soc. A* 2004; 167: 103–124.

Burzykowski, T., Molenberghs, G., Buyse, M., Renard, D., Geys, H., (2001) Validation of surrogate endpoints in multiple randomized clinical trials with failure-time endpoints. *Appl. Statist.,* **50;** 405422.

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.

Buyse M., Burzykowski T., Carroll K., Michiels S. et al. (2007). Progression-Free Survival Is a Surrogate for Survival in Advanced Colorectal Cancer. *Journal of Clinical Oncology*, **25**, 5218–5224.

Cardiac Arrhythmia Suppression Trial (CAST) Investigators (1989). Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infraction. *New England Journal of Medicine*, **321**, 406-412.

Cortiñas Abrahantes, 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 T., and Tomas J. (1991) *Elements of Information Theory.* New York: Wiley.

Daniels MJ and Hughes MD. (1997) Meta-analysis for the evaluation of potential surrogate markers. *Stat. in Med.,* **16**, 1515–1527.

DeGruttola, 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.

DiMasi, J.A., Seibring, M.A., Lasagna, L. (1994). New drug development in the United States from 1963 to 1992. *Clinical Pharmacology & Therapeutics*, **55**, 609–622.

Dunn, N. and Mann, R.D. (1999). Prescription-event and other forms of epidemiological monitoring of side-effects in the UK. *Clinical and Experimental Allergy*, **29**, 217–239.

Ellenberg, S.S. and Hamilton, J.M. (1989) Surrogate endpoints in clinical trials: cancer. *Stat. Med.* **8**, 405–413.

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

Fisher R. (1925) Theory of statistical estimation. *Proc. Cam. Phil. Soc.* **22**, 700–725.

Fleming, T.R. (1994). Surrogate markers in AIDS and cancer trials. *Statistics in Medicine*, **13**, 1423–1435.

Fleming, T.R. (1996) Surrogate endpoints in clinical trials. *Drug Information Journal*, **30**, 545–551.

Fleming, T.R. and DeMets, D.L. (1996) Surrogate endpoints in clinical trials: are we being misled? *Ann. Internal Med.* **125**, 605–613.

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

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

Galecki, A. (1994). General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics: Theory and Methods*, **23**, 3105–3119.

Graybill, F.A. (1983). *Matrices with Applications in Statistics.* Wadsworth, Belmont, CA.

Heise, C., Sampson-Johannes, A., Williams, A., McCormick, F., Von Hoff, D.D., and Kirn, D.H. (1997). ONYX-015, an E1B gene-attenuated adenovirus, causes tumor-specific cytolysis and antitumoral efficacy that can be augmented by standard chemo-therapeutic agents. *Nature Medicine*, **3**, 639–645.

Henderson, R., Diggle, P., and Dobson, A. (2000) Joint modelling of longitudinal measurements and event time data. *Biostatistics,* **1;** 465–480.

International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (1998). ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials. Fed. Regist. 63 (179), 49583.

Johnson, R. (2002). The concept of sickness behavior: a brief chronological account of four key discoveries". *Veterinary Immunology and Immunopathology*, **87**, 443–450.

Jones, T.C. (2001). Call for a new approach to the process of clinical trials and drug registration. *British Medical Journal*, **322**, 920–923.

Kaitin, K.I., DiMasi, J.A. (2000). Measuring the pace of new drug development in the user fee era. *Drug Information Journal*, **34**, 673–680.

Kaitin, K.I., Healy, E.M. (2000). The new drug approvals of 1996, 1997, and 1998. Drug development trends in the user fee era. *Drug Information Journal*, **34**, 1–14.

Kalbfleisch, J. and Prentice, R. L. (1980) *The Statistical Analysis of Failure Time Data*. New York: Wiley.

Kaplan, E. L. and Meier, P. (1958) Non-parametric estimation from incomplete observations *J. Amer. Statist. Assoc.,* **53;** 457–481.

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. (1986) The underlying structure of nonnested hypothesis tests. *Biometrika,* **73:** 333–344.

Kent J. and O'Quigley J. (1988) Measure of dependence for censored survival data. *Biometrika,* **75;** 525–534.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics,* **22;** 7986.

Laird, N.M. and Ware, J.H. (1982) Random effects model for longitudinal data, *Biometrics*, **38**, 963–974.

Lagakos, S.W. and Hoth, D.F. (1992). Surrogate markers in AIDS: Where are we? Where are we going? *Annals of Internal Medicine*, **116**, 599-601.

Lesko, L.J. and Atkinson, A.J. (2001). Use of biomarkers and surrogate endpoints in drug development and regulatory decision making: criteria validation strategies. *Annu. Rev. Pharmacol. Toxicol.* **41**, 347–366.

Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Little, R.J.A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88**, 125–134.

Little, R.J.A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, **81**, 471–483

Little, R.J.A. and Rubin, D.B. (1987) *Statistical Analysis with Missing Data.* New York: John Wiley & Sons.

Little, R.J.A. and Rubin, D.B. (2002) *Statistical Analysis with Missing Data* (2nd ed.). New York: John Wiley & Sons.

Meta-Analysis Group in Cancer (1996). Reappraisal of hepatic arterial infusion in the treatment of nonresectable liver metastases from colorectal cancer. *J. Natn Cancer Inst.*, **88**, 252–258.

Meta-Analysis Group in Cancer (1998). Efficacy of intravenous continuous infusion of 5-fluorouracil compared with bolus administration in patients with advanced colorectal cancer. *J. Clin. Oncol.*, **16**, 301–308.

Molenberghs, G. and Kenward, G.K. (2007). *Missing Data in Clinical Studies.* Chicester: Wiley.

Molenberghs G. and Lesaffre E. (1994) Marginal modelling of correlated ordinal data using a multivariate Plackett distribution. *Journal of the American Statistical Association*, **89**, 633–644.

Molenberghs, G. and Verbeke, G. (2005). *Models for discrete longitudinal data.* New York: Springer.

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.

Molenberghs, G., Buyse, M., Geys, H., Renard, D., Burzykwoski, T. and Alonso, A. (2002). Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Controlled Clinical Trials*, **23**, 607–625.

Molenberghs G, Geys, H, and Buyse M. (2001). Evaluation of surrogate end-points in randomized experiments with mixed discrete and continuous outcomes. *Stat. Med.* **20**, 3023–3038.

Nyquist, H. (1924). Certain factors affecting telegraph speed. *Bell System Technical Journal,* **3;** 324346.

Oakes, D. (1989) Bivariate Survival Models Induced by Frailties *J. Amer. Statist. Assoc.,* **84;** 487–493.

O'Qguigley J. (2008). *Proportional Hazards Regression.* New York: Springer.

O'Qguigley J. and Flandre P. (2006). Quantification of the Prentice criteria for surrogate endpoints. *Biometrics,* **64;** 297–300.

O'Quigley, J., Xu, R. and Stare, J. (2005) Explained randomness in proportional hazards models. *Stat. Med.,* **24;** 479–489.

Pharmacological Therapy for Macular Degeneration Study Group (1997). Interferon α-IIA is ineffective for patients with choroidal neovascularization secondary to age-related macular degeneration. Results of a prospective randomized placebo-controlled clinical trial. *Archives of Ophthalomology*, **115**, 865–872.

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

Pryseley, A., Abel, T., 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,* **4;** 587–597.

Renard, D., Geys, H., Molenberghs, G., Burzykowski, T., and Buyse, M. (2002). Validation of surrogate endpoints in randomized trials with discrete outcomes. *Biometrical Journal,* **30;** 1-15.

Renard, D., Geys, H., Molenberghs, G., Burzykowski, T., Buyse, M., Vangeneugden, T., and Bijnens, L. (2003). Validation of a longitudinally measured surrogate marker for a time-to-event endpoint. *Appl. Statist.,* **30;** 235-247.

Rubin, D.B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592

Sargent, D.J., Wieand, H.J., Haller, D.G., Gray, R., et al. (2005) Disease-free survival versus overall survival as a primary endpoint for adjuvant colon cancer studies: individual patient data from 20,898 patients on 18 randomized trails *Journal of Clinical Oncology*, **23**, 8664–8670.

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

Schemper M, and Stare J. (1996)Explained variation in survival analysis. *Stat Med.* **15(19)**, 1999–2012.

Seibert, J., Glasier, C., Kirby, R. (1998). Transcranial Doppler (TCD), MRA and MRI as a screening examination for cerebrovascular disease in patients with sickle cell anemiaan eight year study. *Pediatr. Radiol.* **28**, 138–142.

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

Szilrd, Le. (1929) On the Decrease in Entropy in a Thermodynamic System by the Intervention of Intelligent Beings. *Zeitschrift fur Physik,* **53;** 840–856.

Tibaldi, F.S., Cortiñas A.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. *J. Stat. Comp. Simul.* **73**, 643–658.

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. *Comp. Stat. Data Anal.,* **51;** 4152–4163.

Tilahun A, Assam P, Alonso A, and Molenberghs G. Flexible surrogate marker evaluation from several randomized clinical trials with continuous endpoints, using R and SAS. *Comp. Stat. Data Anal.* 2007a; 51: 4152–4163.

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.

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

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.

World Health Organization (WHO) (2005). *Preventing chronic diseases: a vital investment*. WHO global report. Geneva: World Health Organization.

Wu, M.C. and Bailey, K.R. (1988) Analysing changes in the presence of informative right censoring caused by death and withdrawal. *Statistics in Medecine*, **7**, 939–955.

Wu, M.C. and Bailey, K.R. (1989) Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics*, **45**, 939–955.

Wu, M.C. and Carroll, R.J. (1988) Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, **44**, 175–188.

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.

A.1 Derivation of The Association Measures

A.1.1 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),

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

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:

$$\Sigma = \sigma \begin{pmatrix} R_{SS} & R_{ST} \\ R_{TS} & R_{TT} \end{pmatrix}, \tag{A.2}$$

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 (A.1) and (A.2) we have:

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

$$\Rightarrow R_\Lambda^2(m) = \frac{m\rho^2}{1+(m-1)\rho}. \tag{A.3}$$

The $R_\Lambda^2(m)$ is a function of m, the number of repeated measurements, if we calculate the derivative of $R_\Lambda^2(m)$ 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 now define

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

which quantifies how much extra information about the true endpoint we get by considering another observation. Note that

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

This last equation implies that $g(\rho)$ is a decreasing function of ρ, i.e, on the one hand, 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 $i + m < k$.

A.1.2 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:

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

In this scenario it has been shown that:

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

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

where $\delta_1^t = (\rho^{m-1}, \rho^{m-2}, ., ., ., \rho, 1)$. Note that R_{SS} is again an $AR(1)$ matrix of dimension m and from Graybill (1983), we have

$$R^{-1}_{SS} = \frac{1}{(1-\rho^2)} \left(\begin{array}{cccccc} 1 & -\rho & 0 & . & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & . & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & . & 0 & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & 0 & . & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & . & \rho & 1 \end{array} \right)$$

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, $C_1^t = (1, -\rho, 0, ., ., ., 0)$, and $C_m^t = (0, ., ., ., ., -\rho, 1)$. 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 $i = 2, ., ., ., m-1$. Additionally, we 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:

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

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

and therefore

$$R_\Lambda^2 = \rho^{2(k-m)}, \tag{A.4}$$

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 precise 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 scenario 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: $R^2_{\Lambda s} = \rho^{2(k-(s+m-1)}$ and obviously $R^2_{\Lambda s} \geq R^2_{\Lambda}$ 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. However, that may imply inquiring more cost and or waiting time.

A.2 Optimal Number of Measurements

A.2.1 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 $1 \leq m \leq k$. It is obvious from (A.3) 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 $\rho > 0$, $\frac{d^2}{dm^2}CPR0(m_1) > 0$ 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 will likely not be integers, thus, we should take the integer that is closest to them.

A.2.2 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:

$$CPR0(m) = w_1 \cdot (1 - R^2_\Lambda(m)) + (1 - w_1) \cdot \frac{R+m}{R+K}, \tag{A.5}$$

where k is the total number of measurements and $1 \leq m \leq k$. From (A.4) we know that $R^2_\Lambda(m) = \rho^{2(k-m)}$ 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 $m= k - 1$.

A.3 Constraint Maximization Problem Derivations