The meta-analytic framework for the evaluation of surrogate endpoints in clinical trials

The Meta-analytic Framework for the Evaluation of Surrogate Endpoints in Clinical Trials

Geert Molenberghs[1] **Tomasz Burzykowski**[1] **Ariel Alonso**[1]

Pryseley Assam[1] **Abel Tilahun**[1]

Marc Buyse[1,2]

[1] Hasselt University, Center for Statistics, Diepenbeek, Belgium

[2] International Drug Development Institute, Ottignies Louvain-la-Neuve, Belgium

Abstract

For a number of reasons, surrogate endpoints are considered instead of the so-called true endpoint in clinical studies, especially when such endpoints can be measured earlier, and/or with less burden for patient and experimenter. Surrogate endpoints may occur more frequently than their standard counterparts. For these reasons, it is not surprising that the use of surrogate endpoints in clinical practice is increasing.

Building on the seminal work of Prentice (1989) and Freedman *et al* (1992), Buyse *et al* (2000) framed the evaluation exercise within a meta-analytic setting, in an effort to overcome difficulties that necessarily surround evaluation efforts based on a single trial. In this paper, we review the meta-analytic approach for continuous outcomes, discuss extensions to non-normal and longitudinal settings, as well as proposals to unify the somewhat disparate collection of validation measures currently on the market. Implications for design and for predicting the effect of treatment in a new trial, based on the surrogate, are discussed. Two case studies are analyzed, one in schizophrenia and one in opthalmology.

Some Key Words: Hierarchical model; Likelihood reduction factor; Meta-analysis; Random-effects model; Surrogate endpoint; Surrogate threshold effect.

1 Introduction

The use of surrogate endpoints in the development of new therapies has always been very controversial, 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). For example, in cardiovascular disease, the unsettling discovery that the two major anti arrhythmic drugs encanaide and

flecanaide reduced arrhythmia but caused a more than 3-fold increase in overall mortality stressed the need for caution in using non-validated surrogate markers in the evaluation of the possible clinical benefits of new drugs (CAST 1989). On the other hand, the dramatic surge of the AIDS epidemic, the impressive therapeutic results obtained early on with zidovudine, and the pressure for an accelerated evaluation of new therapies, have all led to the use of CD4 blood count and later of viral load as endpoints that replaced time to clinical events and overall survival (DeGruttola and Tu 1994), in spite of serious concerns about their limitations as surrogate markers for clinically relevant endpoints (Lagakos and Hoth 1992).

Throughout this paper, we use the terms "endpoint" and "marker" interchangeably to refer simply to some random variable that can be measured over the course of the disease process. Variables that are measured early in the course of the disease are often suggested as potential "surrogates" for those that are measured later. The following definitions reflect the commonly accepted use of various terms in the biomedical literature (Biomarkers Definition Working Group 2001). A *clinical endpoint* is a characteristic or variable that reflects how a patient feels, functions, or survives. A *biomarker* is a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention. A *surrogate endpoint* is a biomarker that is intended to substitute for a clinical endpoint. A surrogate endpoint is expected to predict clinical benefit, harm, or lack thereof.

One important reason for the present interest in surrogate endpoints is the advent of a large number of biomarkers that closely reflect the disease process. An increasing number of new drugs have a well-defined mechanism of action at the molecular level, allowing drug developers to measure the effect of these drugs on the relevant biomarkers (Ferentz 2002). There is increasing public pressure for new, promising drugs to be approved for marketing as rapidly as possible, and such approval will have to be based on biomarkers rather than on some long-term clinical endpoint (Lesko and Atkinson 2001). As an illustration of this trend towards early decision-making, recently proposed clinical trial designs use treatment effects on a surrogate endpoint to screen for treatments that show insufficient promise to have a sizeable impact on survival (Royston, Parmar, and Qian 2003). If the approval process is shortened, there will be a corresponding need for earlier detection of safety signals that could point to toxic problems with new drugs. It is a safe bet, therefore, that

the evaluation of tomorrow's drugs will be based primarily on biomarkers, rather than on the longer-term, harder clinical endpoints that have dominated the development of new drugs until now.

It is therefore best to use *validated* surrogates, though one needs to reflect on the precise meaning and extent of validation (Schatzkin and Gail 2002). Like in many clinical decisions, statistical arguments will play a major role, but ought to be considered in conjunction with clinical and biological evidence. At the same time, surrogate endpoints can play different roles in different phases of drug development. While it may be more acceptable to use surrogates in early phases of research, 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. For a biomarker to be used as a "valid" surrogate, a number of conditions must be fulfilled. The ICH Guidelines on Statistical Principles for Clinical Trials state that "In practice, the strength of the evidence for surrogacy depends upon (i) the biological plausibility of the relationship, (ii) the demonstration in epidemiological studies of the prognostic value of the surrogate for the clinical outcome and (iii) evidence from clinical trials that treatment effects on the surrogate correspond to effects on the clinical outcome" (International Conference on Harmonisation 1998).

Two motivating case studies are introduced in Section 2. The meta-analytic evaluation framework is presented in Section 3, in the context of normally distributed outcomes. Extensions to a variety of non-Gaussian settings are discussed in Section 4. Efforts for unifying the scattered suite of validation measures are reviewed in Section 5. Implications for prediction of the effect in a new trial and for designing studies based on surrogates are the topics of Section 6.

2 Motivating Case Studies

2.1 A Meta-analysis of Five Clinical Trials in Schizophrenia

The data come from a meta-analysis of five double-blind randomized clinical trials, comparing the effects of risperidone to conventional anti psychotic agents for the treatment of chronic schizophrenia. The treatment indicator for risperidone versus conventional treatment will be denoted by Z.

Schizophrenia has long been recognized as a heterogeneous disorder with patients suffering from both 'negative' and 'positive' symptoms. Negative symptoms are characterized by deficits in cognitive, affective and social functions, for example poverty of speech, apathy and emotional withdrawal. Positive symptoms entail more florid symptoms such as delusions, hallucinations and disorganized thinking, which are superimposed on mental status (Kay, Fiszbein, and Opler 1987). Several measures can be considered to asses a patient's global condition. Clinician's Global Impression (CGI) is generally accepted as an admittedly subjective clinical measure of change. Here, the change of CGI versus baseline will be considered as the true endpoint T. It is scored on a 7-grade scale used by the treating physician to characterize how well a subject has improved since baseline. Another useful and sufficiently sensitive assessment scales is the Positive and Negative Syndrome Scale (PANSS) (Kay, Opler, and Lindenmayer 1988). The PANSS consists of 30 items that provide an operationalized, drug-sensitive instrument, which is highly useful for both typological and dimensional assessment of schizophrenia. We will use the change versus baseline in PANSS as our surrogate S. The data contain five trials and in all trials, information is available on the investigators that treated the patients. This information is helpful to define group of patients that will become units of analysis.

2.2 Age-related Macular Degeneration Study (ARMD)

This is a clinical trial involving patients with age-related macular degeneration, who progressively lose vision. Overall, 190 patients from 42 centers participated in the trial. Patients' visual acuity was assessed using standardized vision charts displaying lines of five letters of decreasing size, which patients had to read from top to bottom. The visual acuity was measured by the number of letters correctly read. The binary indicator for treatment is set to $Z = -1$ for placebo and $Z = 1$ for interferon-α. The surrogate endpoint S is visual acuity 6 months after starting treatment while the true endpoint T is the change in visual acuity at 1 year. In the analysis, the centers in which patients were treated will be considered as units of analysis. Six out of 42 centers participating in the trial enrolled patients only to one of the two treatment arms. These centers were excluded from consideration. A total of 36 centers were thus available for analysis, with a number of individual patients per center ranging from 2 to 18 (183 patients overall).

3 A Meta-analytic Framework for Normally Distributed Outcomes

Several methods have been suggested for the formal evaluation of surrogate markers, some based on a single trial with others, currently gaining momentum, of a meta-analytic nature. The first formal single trial approach to validate markers is due to Prentice (1989), who gave a definition of the concept of a surrogate endpoint, followed by a series of operational criteria. Freedman *et al* (1992) augmented Prentice's hypothesis-testing based approach, with the estimation paradigm, through the so-called *proportion of treatment effect explained*. In turn, Buyse and Molenberghs (1998) added two further measures: the *relative effect* and the *adjusted association*. All of these proposals are hampered by the fact that they are single-trial based, in which there evidently is replication at the patient level, but not at the level of the trial.

3.1 A Meta-Analytic Approach

Although the single trial based methods are relatively easy in terms of implementation, they are surrounded with the difficulties stated at the end of the previous section. Therefore, several authors, such as Daniels and Hughes (1997), Buyse *et al* (2000), and Gail *et al* (2000) have introduced the meta-analytic approach. This section briefly outlines the methodology, followed by simplified modeling approaches as suggested by Tibaldi *et al* (2003).

The meta-analytic approach was formulated originally for two continuous, normally distributed outcomes, and extended in the meantime to a large collection of outcome types, ranging from continuous, binary, ordinal, time-to-event, and longitudinally measured outcomes (Burzykowski, Molenberghs, and Buyse 2005). First, we focus on the continuous case, where the surrogate and true endpoints are jointly normally distributed.

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

$$S_{ij} = \mu_{Si} + \alpha_i Z_{ij} + \varepsilon_{Sij}, \tag{1}$$

$$T_{ij} = \mu_{Ti} + \beta_i Z_{ij} + \varepsilon_{Tij}, \tag{2}$$

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

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ & \sigma_{TT} \end{pmatrix}. \tag{3}$$

In addition, we can decompose

$$\begin{pmatrix} \mu_{Si} \\ \mu_{Ti} \\ \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} m_{Si} \\ m_{Ti} \\ a_i \\ b_i \end{pmatrix}, \tag{4}$$

where the second term on the right hand side of (4) is assumed to follow a zero-mean normal distribution with covariance matrix

$$D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ & d_{TT} & d_{Ta} & d_{Tb} \\ & & d_{aa} & d_{ab} \\ & & & d_{bb} \end{pmatrix}. \tag{5}$$

A classical hierarchical, random-effects modeling strategy results from the combination of the above two steps into a single one:

$$S_{ij} = \mu_S + m_{Si} + \alpha Z_{ij} + a_i Z_{ij} + \varepsilon_{Sij}, \tag{6}$$

$$T_{ij} = \mu_T + m_{Ti} + \beta Z_{ij} + b_i Z_{ij} + \varepsilon_{Tij}. \tag{7}$$

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

After fitting the above models, surrogacy is captured by means of two quantities: trial-level and individual-level coefficients of determination. The former quantifies the association between the treatment effects on the true and surrogate endpoints at the trial level, while the latter measures

the association at the level of the individual patient, after adjustment for the treatment effect. The former is given by:

$$R^2_{\text{trial}} = R^2_{b_i|m_{Si},a_i} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}}. \tag{8}$$

The above quantity is unitless and, at the condition that the corresponding variance-covariance matrix is positive definite, lies within the unit interval.

Apart from estimating the strength of surrogacy, the above model can also be used for prediction purposes. To this end, observe that $(\beta + b_0|m_{S0}, a_0)$ follows a normal distribution with mean and variance:

$$E(\beta + b_0|m_{S0}, a_0) = \beta + \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} \mu_{S0} - \mu_S \\ \alpha_0 - \alpha \end{pmatrix}, \tag{9}$$

$$\text{Var}(\beta + b_0|m_{S0}, a_0) = d_{bb} - \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}. \tag{10}$$

A prediction can be made using (9), with prediction variance (10). Of course, one has to properly acknowledge the uncertainty resulting from the fact that parameters are not known but merely estimated. We return to this issue in Section 6.

Models (1) and (2) are referred to as the full fixed-effects models. It is sometimes necessary, for computational reasons, to contemplate a simplified version. A reduced version of these models is obtained by replacing the fixed trial-specific intercepts by a common one. Thus, the reduced mixed effect models result from removing the random trial-specific intercepts m_{Si} and m_{Ti} from models (6) and (7). The R^2 for the reduced models then is:

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

A surrogate could be adopted when R^2_{trial} is sufficiently large. Arguably, rather than using a fixed cutoff above which a surrogate would be adopted, there always will be clinical and other judgment involved in the decision process. The R^2_{indiv} is based on (3) and takes the following form:

$$R^2_{\text{indiv}} = R^2_{\varepsilon_{Ti}|\varepsilon_{Si}} = \frac{\sigma^2_{ST}}{\sigma_{SS}\sigma_{TT}}. \tag{11}$$

3.2 Simplified Modeling Strategies

Though the above hierarchical modeling is elegant, it often poses a considerable computational challenge (Burzykowski, Molenberghs, and Buyse 2005). To address this problem, Tibaldi *et al* (2003) suggested several simplifications, briefly outlined here. These authors considered three possible dimensions along which simplifications can be undertaken.

The first choice is 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 (1)–(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

$$\widehat{\beta}_i = \widehat{\lambda}_0 + \widehat{\lambda}_1 \widehat{\mu}_{Si} + \widehat{\lambda}_2 \widehat{\alpha}_i + \varepsilon_i. \tag{12}$$

The trial-level $R^2_{\text{trial(f)}}$ then is obtained by regressing $\widehat{\beta}_i$ on $\widehat{\mu}_{Si}$ and $\widehat{\alpha}_i$, whereas $R^2_{\text{trial(r)}}$ is obtained from regressing $\widehat{\beta}_i$ on $\widehat{\alpha}_i$ only. The individual-level value is calculated as in (11), using the estimates from (3).

The second option is to consider the trial-specific effects as random. Depending on whether the endpoints are considered jointly or separately (see next paragraph), two directions can be followed. The first one involves a two-stage approach with at the first stage univariate models (6)–(7). 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 model.

Though natural to assume the two endpoints 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 by making use of the correlation between the residuals from two separate univariate models. 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 fixed, models (1)–(2) are fitted

separately. Similarly, if the trial-specific effects are considered random, models (6)–(7) are fitted separately, i.e., the corresponding error terms in the two models are assumed independent.

When the univariate approach and/or the fixed-effects approach are chosen, there is a need to adjust for the heterogeneity in information content between trial-specific contributions. One way of doing so is weighting the contributions according to trial size. This gives rise to a weighted linear regression model (12) in the second stage.

In summary, the simplified strategies perform rather well, especially when outcomes are of a continuous nature (Cortiñas *et al* 2004), and are a valuable addition to the fully specified hierarchical model, for those situations where the latter is infeasible or less reliable.

3.3 Unit of Analysis

A cornerstone of the meta-analytic method is the choice of 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, experts' considerations about the most suitable unit for a specific problem, the amount of replication at a potential unit's level, and the number of patients per unit. From a technical point of view, the most desirable situation is where the number of units and the number of patients per unit is sufficiently large. This issue has been discussed by Cortiñas *et al* (2004). Of course, in cases where one has to resort to simplified strategies, one has to reflect carefully on the status of the results obtained. Arguably, they may not be as reliable as one might hope for, and one should undertake every effort possible to increase the amount of information available. Clearly, even an analysis based on a simplified strategy, especially in the light of good performance, may support efforts to make more data available for analysis.

3.4 Treatment Coding

Most of the work reported in Burzykowski, Molenberghs, and Buyse (2005) is for a dichotomous treatment indicator. 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 connected by simple linear transformations. Note that this is not the case, of course, for more than three treatment arms. However, of more importance for us here is the impact the choices can have on the hierarchical model. Indeed, while the marginal model resulting from (6)–(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 it is of importance to illicit views from the study's investigators.

3.5 Ill-conditioned and Non-positive Definite Variance-covariance Matrix

When the full bivariate random effect is used, the R^2_{trial} is computed from the variance-covariance matrix (5). It is sometimes possible that this matrix be ill-conditioned and/or non-positive definite. In such cases, the resulting quantities computed based on this matrix might not be trustworthy. One way to asses the ill-conditioning of a matrix is by reporting its condition number, i.e., the ratio of the largest over the smallest eigenvalue. A large condition number is an indication of ill-conditioning. The most pathological situation occurs when at least one eigenvalue is equal to zero. This corresponds to a positive semi-definite matrix, which occurs, for example, when a boundary solution is obtained. While it is hard to definitively identify the reason for a zero eigenvalue, insufficient information, either in terms of the number of trials, the simple size within trials, or both, may often be the cause and deserving of careful assessment. Using the simplified methods is certainly an option in this case; apart from providing a solution to the problem, it may give a handle on the problem at hand.

3.6 Application to the Case Studies

3.6.1 A Meta-analysis of Five Clinical Trials in Schizophernia

Let us start with the schizophrenia study. Here, trial seems the natural unit of analysis. Unfortunately, the number of trials is not sufficient to apply the 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, each treating 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 some investigators treated few patients, the resulting within-unit variability might not be estimated correctly.

The basic meta-analytic approach and the corresponding simplified strategies have been applied, with results displayed in Table 1. 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. The bivariate full random effect model does not converge when trial is used as the unit of analysis. This might be due to lack of sufficient information to compute all sources of variability. The reduced bivariate random effects model converged for both cases, but the resulting variance-covariance matrices were not positive-definite and were ill-conditioned, as can be seen from the very large value of the condition number. Consequently, the results of the bivariate random effects model should be treated with caution. If we concentrate on the results based on investigator as unit of analysis, we observe a low level of surrogacy of PANSS for CGI, with R^2_{trial} ranging roughly between 0.5 and 0.68 for the different simplified models. This result, however, has to be coupled with other findings based on expert opinion to fully guarantee the validation of PANSS as possible surrogate for 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 promising

surrogate for CGI.

3.6.2 Age-related Macular Degeneration Study

For the ARMD study, the only available unit of analysis was center. There were 36 centers which treated between 2 and 18 patients. Note that these data has been analyzed by Buyse *et al* (2000) with a treatment coding of 0 and 1 for the placebo and treatment arms, respectively. Here, the $-1/+1$ coding was used and thus slightly different results obtain. 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 2 when the $-1/+1$ coding is used, and in Table 3, when the 0/1 coding is employed. 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, the results cannot be trusted. Based on the findings, it is possible to say that assessment of visual acuity at 6 months does not seem to be a very strong surrogate for the same assessment at 1 year. While the impact of the coding is clear from the results, there are no substantive changes in the conclusions. Nevertheless, we recommend the $-1/+1$ coding, since it is sensible to assume the overall variance is similar in both arms, whereas the 0/1 coding forces the variance to be larger in the experimental arm. In conclusion, one has to be aware that results can be obtained that look reasonable but are not trustworthy. Hence, the diagnostic tools, such as the condition number, will be a valuable role.

4 Non-Gaussian Endpoints

Statistically speaking, the surrogate endpoint and the clinical endpoint are realizations of random variables. As will be clear from the formalism in Section 3, one is in need of the joint distribution of these variables. The easiest, but not the only, situation is where both are Gaussian random variables, but one also encounters binary (e.g., CD4+ counts over 500/mm3, tumor shrinkage), categorical (e.g., cholesterol levels <200 mg/dl, 200-299 mg/dl, 300+ mg/dl, tumor response as complete response, partial response, stable disease, progressive disease), censored continuous (e.g., time to undetectable viral load, time to cardiovascular death), longitudinal (e.g., CD4+ counts

over time, blood pressure over time), and multivariate longitudinal (e.g., CD4+ and viral load over time jointly, various dimensions of quality of life over time) endpoints. The models used to validate a surrogate for a clinical endpoint will depend on the type of variables observed in the problem at hand. Table 4 shows some examples of potential surrogate endpoints in various diseases. In what follows, we will briefly discuss the settings of binary endpoints, failure-time endpoints, the combination of an ordinal and a survival endpoint, and longitudinal endpoints.

4.1 Binary Endpoints

Renard *et al* (2002) have shown that extension to this situation is easily done using a latent variable formulation. That is, one posits the existence of a pair of continuously distributed latent variable responses $(\widetilde{S}_{ij},\ \widetilde{T}_{ij})$ that produce the actual values of (S_{ij}, T_{ij}). These unobserved variables are assumed to have a joint normal distribution and the realized values follow by double dichotomization. On the latent-variable scale, we obtain a model similar to (1)–(2) and in the matrix (3) the variances are set equal to unity in order to ensure identifiability. This leads to the following model:

$$\begin{cases} \Phi^{-1}(P[S_{ij}=1|Z_{ij}, m_{S_i}, a_i, m_{T_i}, b_i]) &= \mu_S + m_{S_i} + (\alpha + a_i)Z_{ij}, \\ \Phi^{-1}(P[T_{ij}=1|Z_{ij}, m_{S_i}, a_i, m_{T_i}, b_i]) &= \mu_T + m_{T_i} + (\beta + b_i)Z_{ij}, \end{cases}$$

where Φ denotes the standard normal cumulative distribution function. Renard *et al* (2002) used pseudo-likelihood methods to estimate the model parameters. Similar ideas have been used in the case one of the endpoints is continuous, with the other one binary or categorical (Burzykowski, Molenberghs, and Buyse 2005, Ch. 6).

4.2 Two Failure-time Endpoints

Assume now that S_{ij} and T_{ij} are failure-time endpoints. Model (1)–(2) is replaced by a model for two correlated failure-time random variables. Burzykowski *et al* (2001) used copulas to this end (Clayton 1978, Hougaard 1986). Precisely, one assumes the joint survivor function of $(S_{ij},\ T_{ij})$ is written as:

$$F(s,\ t) = P(S_{ij} \geq s,\ T_{ij} \geq t) = C_\delta\{F_{Sij}(s),\ F_{Tij}(t)\}, \qquad s,\ t \geq 0, \tag{13}$$

where $(F_{Sij},\ F_{Tij})$ denote marginal survivor functions and C_δ is a copula, i.e., a distribution function on $[0,\ 1]^2$ with $\delta \in R^1$.

When the hazard functions are specified, estimates of the parameters for the joint model can be obtained using maximum likelihood. Shih and Louis (1995) discuss alternative estimation methods. The association parameter is generally hard to interpret. However, it can be shown (Genest and McKay 1986) that there is a link with Kendall's τ:

$$\tau = 4 \int_0^1 \int_0^1 C_\delta(u,\, v) C_\delta(du,\, dv) - 1,$$

providing an easy measure of surrogacy at the individual level. At the second stage R^2_{trial} can be computed based on the pairs of treatment effects estimated at the first stage.

4.3 An Ordinal Surrogate and a Survival Endpoint

Assume that T is a failure-time random variable and S is a categorical variable with K ordered categories. To propose validation measures, similar to those introduced in the previous section, Burzykowski *et al* (2004) also used bivariate copulas, combining ideas of Molenberghs, Geys, and Buyse (2001) and Burzykowski *et al* (2001). One marginal distribution is a proportional odds logistic regression, while the other is a proportional hazards model. The Plackett copula (Dale 1986) was chosen to capture the association between both endpoints. The ensuing global odds ratio is relatively easy to interpret.

4.4 Longitudinal Endpoints

Most of the previous work focuses on univariate responses. Alonso *et al* (2003) showed that going from a univariate setting to a multivariate framework represents new challenges. The R^2 measures proposed by Buyse *et al* (2000), are no longer applicable. Alonso *et al* (2003) based their calculations of surrogacy measures on a two-stage approach rather than a full random effects approach. They assume that information from $i = 1, \ldots, N$ trials is available, in the ith of which, $j = 1, \ldots, n_i$ subjects are enrolled and they denoted the time at which subject j in trial i is measured as t_{ijk}. If T_{ijk} and S_{ijk} denote the associated true and surrogate endpoints, respectively, and Z_{ij} is a binary indicator variable for treatment then along the ideas of Galecki (1994), they proposed the following joint model, at the first stage, for both responses

$$\begin{cases} T_{ijk} = \mu_{Ti} + \beta_i Z_{ij} + g_{Tij}(t_{ijk}) + \varepsilon_{Tijk}, \\ \\ S_{ijk} = \mu_{Si} + \alpha_i Z_{ij} + g_{Sij}(t_{ijk}) + \varepsilon_{Sijk}, \end{cases} \tag{14}$$

where μ_{Ti} and μ_{Si} are trial-specific intercepts, β_i and α_i are trial-specific effects of treatment Z_{ij} on the two endpoints and g_{Tij} and g_{Sij} are trial-subject-specific time functions that can include treatment-by-time interactions. They also assume that the vectors, collecting all information over time for patient j in trial i, $\widetilde{\varepsilon}_{T_{ij}}$ and $\widetilde{\varepsilon}_{S_{ij}}$ are correlated error terms, following a mean-zero multivariate normal distribution with covariance matrix

$$\Sigma_i = \begin{pmatrix} \Sigma_{TTi} & \Sigma_{TSi} \\ \Sigma'_{TSi} & \Sigma_{SSi} \end{pmatrix} = \begin{pmatrix} \sigma_{TTi} & \sigma_{TSi} \\ \sigma_{TSi} & \sigma_{SSi} \end{pmatrix} \otimes R_i. \tag{15}$$

Here, R_i is a correlation matrix for the repeated measurements.

If treatment effect can be assumed constant over time, then (8) can still be useful to evaluate surrogacy at the trial level. However, at the individual level the situation is totally different, the R^2_{ind} no longer being applicable, and new concepts are needed.

Using multivariate ideas, Alonso *et al* (2003) proposed the *variance reduction factor* (VRF) to capture individual-level surrogacy in this more elaborate setting. They quantified the relative reduction in the true endpoint variance after adjustment by the surrogate as

$$VRF_{\text{ind}} = \frac{\sum_i \{\text{tr}(\Sigma_{TTi}) - \text{tr}(\Sigma_{(T|S)i})\}}{\sum_i \text{tr}(\Sigma_{TTi})}, \tag{16}$$

where $\Sigma_{(T|S)_i}$ denotes the conditional variance-covariance matrix of $\widetilde{\varepsilon}_{T_{ij}}$ given $\widetilde{\varepsilon}_{S_{ij}}$: $\Sigma_{(T|S)i} = \Sigma_{TTi} - \Sigma_{TSi}\Sigma^{-1}_{SSi}\Sigma'_{TSi}$. Here, Σ_{TTi} and Σ_{SSi} are the variance-covariance matrices associated with the true and surrogate endpoint respectively and Σ_{TSi} contains the covariances between the surrogate and the true endpoint. Alonso *et al* (2003) showed that the VRF_{ind} ranges between zero and one, and that $VRF_{\text{ind}} = R^2_{\text{ind}}$ when the endpoints are measured only once.

An alternative proposal is

$$\theta_p = \sum_i \frac{1}{Np_i} \text{tr}\left\{\left(\Sigma_{TTi} - \Sigma_{(T|S)i}\right)\Sigma^{-1}_{TTi}\right\}. \tag{17}$$

Structurally, both VRF and θ_p are similar, the difference being the reversal of summing the trace and calculating the ratio. In spite of this strong structural similarity the VRF is not symmetric in S and T and it is only invariant with respect to linear orthogonal transformations, whereas θ_p is both symmetric and invariant with respect to the broader class of linear bijective transformations.

A common problem of all previous proposals is that they are strongly based on the normality assumption and extensions to non-normal settings are difficult. To overcome this limitation, Alonso *et al* (2005), introduced a new parameter, the so-called R^2_Λ, to evaluate surrogacy at the individual level when both responses are measured over time or in general when multivariate or repeated measures are available

$$R^2_\Lambda = \frac{1}{N}\sum_i (1-\Lambda_i), \tag{18}$$

where: $\Lambda_i = \dfrac{|\Sigma_i|}{|\Sigma_{TTi}|\,|\Sigma_{SSi}|}$. This parameter not only allows the detection of more general patterns of association but can also be extended to more general settings that those defined by the normal distribution. They proved that R^2_Λ ranges between zero and one, and that in the cross-sectional case $R^2_\Lambda = R^2_{\text{ind}}$. These authors have shown that $R^2_\Lambda = 1$ whenever there is a deterministic relationship between two linear combinations of both endpoints, allowing the detection of strong associations in cases where the VRF or θ_p would fail in doing so.

5 Towards a Unified Approach

The longitudinal method of the previous section, while elegant, hinges upon normality. First using the likelihood reduction factor (Section 5.1) and then an information-theoretic approach (Section 5.2), extension, and therefore unification, will be achieved.

5.1 The Likelihood Reduction Factor

Estimating individual-level surrogacy, as the previous developments clearly show, has frequently been based on a variance-covariance matrix coming from the distribution of the residuals. However, if we move away from the normal distribution, it is not always clear how to quantify the association between both endpoints after adjusting for treatment and trial effect. To address this problem, Alonso *et al* (2004b) considered the following generalized linear models in the ith trial

$$g_T(T_{ij}) = \mu_{T_i} + \beta_i Z_{ij}, \tag{19}$$

$$g_T(T_{ij}) = \theta_{0i} + \theta_{1i} Z_{ij} + \theta_{2i} S_{ij}. \tag{20}$$

The longitudinal case would be covered by considering particular functions of time in (19) and (20). Consider G_i^2 as the log-likelihood ratio test statistics to compare (19) with (20) in trial i, and quantify the association between both endpoints at the individual level using a scaled likelihood reduction factor (LRF)

$$\text{LRF} = 1 - \frac{1}{N}\sum_i \exp\left(-\frac{G_i^2}{n_i}\right). \tag{21}$$

Alonso *et al* (2004b) established a number of properties for LRF, in particular its ranging in the unit interval, and its reduction to R_Λ^2 in the longitudinal and to R_{ind}^2 in the cross-sectional case.

5.2 An Information-theoretic Unification

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

$$H(Y) = \sum_i p_i \log\left(\frac{1}{p_i}\right). \tag{22}$$

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

$$h_d(Y) = -E[\log f_X(X)] = -\int_{S_{f_X}} f_X(x)\log f_X(x)dx. \tag{23}$$

The joint and conditional (differential) entropies are defined in an analogous fashion. Defining the information of a single event as $I(A) = \log p_A$, the entropy is $H(A) = -I(A)$. No information is gained from a totally certain event, $p_A \approx 1$, so $I(A) \approx 0$), while an improbable event is informative.

$H(Y)$ is the average uncertainty associated with P. Entropy is always non-negative, satisfies $H(Y|X) \leq H(Y)$ for any pair of random variables, with equality holding under independence, and is invariant under a bijective transformation (Cover and Tomas 1991). Differential entropy enjoys some but not all properties of entropy: it can be infinitely large, negative, or positive, and is coordinate dependent. For a bijective transformation $Y = y(X)$, it follows $h_d(Y) = h_d(X) - \text{E}_Y\left(\log\left|\frac{dx}{dy}(y)\right|\right)$.

We can now quantify the amount of uncertainty in Y, expected to be removed if the value of X were known, by $I(X,Y) = h_d(Y) - h_d(Y|X)$, the so-called *mutual information*. It is always non-negative,

zero if and only if X and Y are independent, symmetric, invariant under bijective transformations of X and Y, and $I(X,X) = h_d(X)$. The mutual information measures the information of X, shared by Y.

We will now introduce the entropy-power (Shannon 1948) for comparison of continuous random variables. Let X be a continuous n-dimensional random vector. The entropy-power of X is

$$\mathrm{EP}(X) = \frac{1}{(2\pi e)^n} e^{2h(X)}. \tag{24}$$

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

We can now define an information-theoretic measure of association (Schemper and Stare 1996):

$$R_h^2 = \frac{\mathrm{EP}(Y) - \mathrm{EP}(Y|X)}{\mathrm{EP}(Y)}, \tag{25}$$

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

$$R_{h\max}^2 = \frac{R_h^2}{1 - e^{-2H(Y)}},$$

reaching 1 when both endpoints are deterministically related.

We can now redefine surrogacy, while preserving previous proposals as special cases. While we will focus on individual-level surrogacy, all results apply to the trial level too. Let $Y = T$ and $X = S$ be the true and surrogate endpoints, respectively. We consider S a good surrogate for T at the individual (trial) level, if a "large" amount of uncertainty about T (the treatment effect on T) is reduced when S (the treatment effect on S) is known. Equivalently, we term S a good surrogate for T at the individual level, if our lack of knowledge about the true endpoint is substantially reduced when the surrogate endpoint is known.

A meta-analytic framework, with N clinical trials, produces N_q different R^2_{hi}, and hence we propose a meta-analytic R^2_h:

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

where $\alpha_i > 0$ for all i and $\sum_{i=1}^{N_q} \alpha_i = 1$. Different choices for α_i lead to different proposals, producing an uncountable family of parameters. This opens the additional issue of finding an *optimal* choice. In particular, for the cross-sectional normal-normal case, Alonso and Molenberghs (2006) have shown that $R^2_h = R^2_{\text{ind}}$. The same holds for R^2_Λ, defined in (14) for the longitudinal case. Finally, when the true and surrogate endpoints have distributions in the exponential family, then LRF $\xrightarrow{P} R^2_h$ when the number of subjects per trial goes to infinity.

5.3 Fano's Inequality and the Theoretical Plausibility of Finding a Good Surrogate

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

$$\mathrm{E}\left[(T - g(S))^2\right] \geq \mathrm{EP}(T)(1 - R^2_h) \tag{26}$$

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, (26) 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.

5.4 Application to Case Studies

5.4.1 A Meta-analysis of Five clinical Trials in Schizophernia

We will treat CGI as the true endpoint and PANSS as surrogate, although the reverse would be sensible, too. In practice, these endpoints are frequently dichotomized in a clinically meaningful way. Our binary true endpoint $T = \mathrm{CGId} = 1$ for patients classified from "Very much improved"

to "Improved", and 0 otherwise. The binary surrogate $S = \text{PANSSd} = 1$ for patients with at least 20 points reduction versus baseline, and 0 otherwise. We will start from probit and Plackett-Dale models and compare results with the ones from the information-theoretic approach.

In line with Section 4.1, we formulate two continuous latent variables $(\widetilde{\text{CGI}}_{ij}, \widetilde{\text{PANSS}}_{ij})$ assumed to follow a bivariate normal distribution. The following probit model can be fitted

$$\begin{pmatrix} \widetilde{\mu}_{ij}^T \\ \widetilde{\mu}_{ij}^S \\ \ln(\sigma^2) \\ \ln\left(\dfrac{1+\widetilde{\rho}}{1+\widetilde{\rho}}\right) \end{pmatrix} = \begin{pmatrix} \widetilde{\mu}_{T_i} + \widetilde{\beta}_i Z_{ij} \\ \widetilde{\mu}_{S_i} + \widetilde{\alpha}_i Z_{ij} \\ c_{\sigma^2} \\ c_{\widetilde{\rho}} \end{pmatrix}, \tag{27}$$

where $\widetilde{\mu}_{ij}^T = E(\widetilde{\text{CGI}}_{ij})$, $\widetilde{\mu}_{ij}^S = E(\widetilde{\text{PANSS}}_{ij})$, $\text{Var}(\widetilde{\text{CGI}}_{ij}) = 1$, $\sigma^2 = \text{Var}(\widetilde{\text{PANSS}}_{ij})$ and $\widetilde{\rho} = \text{corr}(\widetilde{\text{CGI}}_{ij}, \widetilde{\text{PANSS}}_{ij})$ denotes the correlation between the true and surrogate endpoint latent variables. We can then use the estimated values of $(\widetilde{\mu}_{S_i}, \widetilde{\alpha}_i, \widetilde{\beta}_i)$ to evaluate trial level surrogacy through the R^2_{trial}. At the individual level, $\widetilde{\rho}^2$ is used to capture surrogacy.

Alternatively, the Dale (1986) formulation can be used, based on

$$\begin{pmatrix} \text{logit}(\pi_{ij}^T) \\ \text{logit}(\pi_{ij}^S) \\ \ln(\psi) \end{pmatrix} = \begin{pmatrix} \mu_{T_i} + \beta_i Z_{ij} \\ \mu_{S_i} + \alpha_i Z_{ij} \\ c_\psi \end{pmatrix} \tag{28}$$

where $\pi_{ij}^T = \text{E}(\text{CGId}_{ij})$, $\pi_{ij}^S = \text{E}(\text{PANSSd}_{ij})$ and ψ is the global odds ratio associated to both endpoint. As before, the estimated values of $(\mu_{S_i}, \alpha_i, \beta_i)$ can be used to evaluate surrogacy at the trial level and the individual level surrogacy is quantified using the global odds ratio.

In the information-theoretic approach the following three models are fitted independently

$$\Phi(\pi_{ij}^T) = \mu_{T_i} + \beta_i Z_{ij}, \tag{29}$$

$$\Phi(\pi_{ij}^{T|S}) = \mu_{T_i}^S + \beta_i^S Z_{ij} + \gamma_{ij} S_{ij}, \tag{30}$$

$$\Phi(\pi_{ij}^S) = \mu_{S_i} + \alpha_i Z_{ij}, \tag{31}$$

where $\pi_{ij}^T = \text{E}(\text{CGId}_{ij})$, $\pi_{ij}^{T|S} = \text{E}(\text{CGId}_{ij}|\text{PANSSd}_{ij})$, $\pi_{ij}^S = \text{E}(\text{PANSSd}_{ij})$ and Φ denotes the cumulative standard normal distribution. At the trial level, the estimated values of $(\mu_{S_i}, \alpha_i, \beta_i)$ obtained from (29) and (31) can be used to calculate the R^2_{trial}, whereas at the individual level we can quantify surrogacy using R_h^2. As it was stated before, the LRF is a consistent estimator of R_h^2,

however, in principle other estimators could be used as well. We will then quantify surrogacy at the individual level by $\hat{R}_h^2 = 1 - \exp\left(-G^2/n\right)$, where G^2 is the loglikelihood ratio test to compare (29) with (30) and n denotes total number of patients. Furthermore, when applied to the binary-binary setting, Fanos's inequality takes the form

$$P(T \neq S) \geq \frac{1}{\log|\Psi|}\left[H(T) - 1 + \frac{1}{2}\ln(1 - R_h^2)\right],$$

where $\Psi = \{0, 1\}$ and $|\Psi|$ denotes the cardinal of Ψ. Here, again, Fano's inequality gives a lower bound for the probability of incorrect prediction.

Table 5 shows the results at the trial and individual level obtained with the different approaches described above. At the trial level, all the methods produced very similar values for the validation measure. In all cases, $R^2_{\text{trial}} \simeq 0.50$. It is also remarkable that the probit approach, in spite of being based on treatment effects defined at a latent level, produced a R^2_{trial} value similar to the ones obtained with the information–theoretic and Plackett-Dale approaches. However, as Alonso *et al* (2003) showed, there is a linear relationship between the mean parameters defined at the latent level and the mean parameters of the model based on the observable endpoints and that could explain the agreement between the probit and the other two procedures. Therefore, at the trial level, we could conclude that knowing the treatment effect on the surrogate will reduce our uncertainty about the treatment effect on the true endpoint by 50%.

At the individual level, the probit approach gives the strongest association between the surrogate and the true endpoint. Nevertheless, this value describes the association at an unobservable latent level, rendering its interpretation more awkward than with information theory, since it is not clear how this latent association could be relevant from a clinical point of view or how it could be translated into an association for the observable endpoints. The Plackett-Dale procedure quantifies surrogacy using a global odds ratio, making the comparison between this method and the others more difficult. Note that even though odds ratios are widely used in biomedical fields the lack of an upper bound makes difficult their interpretation in this setting.

On the other hand, the value of the $R^2_{h\max}$ illustrates that the surrogate can merely explain 39% of our uncertainty about the true endpoint, a relatively low value. Additionally, the lower bound for Fano's inequality clearly shows that using the value of PANSS to predict the outcome on CGI

would be misleading in at least 8% of the cases. Even though this value is relatively low, it is only a lower bound and the real probability of mistake could be much larger.

At the trial level, the information-theoretic approach produces results similar to the ones from the conventional methods, but does so by means of models that are generally much easier to fit. At the individual level, the information-theoretic approach avoids the problem common with the probit model in that the correlation of the latter is formulated at the latent scale and therefore less relevant for practice. In addition, the information-theoretic measure ranges between 0 and 1, circumventing interpretational problems arising from using the unbounded Plackett-Dale based odds ratio.

5.4.2 The Age-related Macular Degeneration Study

Consider two dichotomized outcomes: visual acuity at 6 months (S) and visual acuity at 1 year (T), defined as increase versus decrease of number of letters correctly read on the vision chart. Again, (29)–(31) are fitted independently, where now $S_{ij} = \text{vis6}_{ij}$ and $T_{ij} = \text{vis12}_{ij}$ are the dichotomized visual acuity, for the j^{th} patient in the i^{th} trial, at 6 months and one year, respectively. We also use the notation $\pi_{ij}^T = \text{E}(\text{vis12}_{ij})$, $\pi_{ij}^{T|S} = \text{E}(\text{vis12}_{ij}|\text{vis6}_{ij})$, and $\pi_{ij}^S = \text{E}(\text{vis6}_{ij})$. Assuming that the association between both variables is constant we can estimate the individual level-surrogacy computing the $R_h^2 = 1 - e^{-2I(X,Y)}$. By way of sensitivity analysis, the assumption of a constant covariance structure was relaxed. The results obtained were virtually identical and therefore omitted. The LRF is computed as in the previous section.

Table 6 shows the results at both levels. All of the estimated values are too low to make visual acuity at 6 months a reliable surrogate. At the trial-level, $\widehat{R}^2_{\text{trial}} = 0.38$, which clearly shows that an accurate prediction of treatment effect at one year based on the treatment effect observed at 6 months is not possible. It is clear that when the outcome at 6 months is sufficiently large, then the prediction of the month 12 outcome, together with its prediction limits, may contain useful information. While this would hold for every R^2 larger than zero, the closer it is to zero, the larger and hence the more unrealistic will the surrogate endpoint value have to be. Switching to $R^2_{h\max}$, we do obtain some evidence of a weak association at the individual-level.

6 Prediction and Design Aspects

An important application of surrogacy evaluation is the prediction of treatment effect on the true endpoint *without measuring the latter*, supplemented with appropriate quantification of uncertainty. We will review the work done in this respect by Burzykowski and Buyse (2006).

Two components contribute to such a prediction: (a) information obtained in the validation process based on trials $i = 1, \ldots, N$, used to fit model (1)–(2), and (b) the estimate of the effect of Z on S in a new trial $i = 0$ providing data on the surrogate endpoint but not on the true endpoint. We can then fit the following linear model to the surrogate outcomes S_{0j}:

$$S_{0j} = \mu_{S0} + \alpha_0 Z_{0j} + \varepsilon_{S0j}. \tag{32}$$

Based on this, we observe that the treatment effect on the true endpoint, $(\beta + b_0 | m_S, a_0)$, follows a normal distribution with mean linear in μ_{S0}, μ_S, α_0, and α, and variance

$$\text{Var}(\beta + b_0 | m_{S0}, a_0) \quad = \quad (1 - R^2_{\text{trial}})\text{Var}(b_0), \tag{33}$$

where m_{S0} and a_0 are the surrogate-specific random intercept and treatment effect in the new trial, respectively, and $\text{Var}(b_0)$ denotes the unconditional variance of the trial-specific random effect. Group the fixed-effects parameters and variance components into ϑ, with $\widehat{\vartheta}$ the corresponding estimates. The prediction variance can then be written as:

$$\text{Var}(\beta + b_0 | \mu_{S0}, \alpha_0, \vartheta) \approx f\{\text{Var}(\widehat{\mu}_{S0}, \widehat{\alpha}_0)\} + f\{\text{Var}(\widehat{\vartheta})\} + (1 - R^2_{\text{trial}})\text{Var}(b_0), \tag{34}$$

where $f\{\text{Var}(\widehat{\mu}_{S0}, \widehat{\alpha}_0)\}$ and $f\{\text{Var}(\widehat{\vartheta})\}$ are functions of the asymptotic variance-covariance matrices of $(\widehat{\mu}_{S0}, \widehat{\alpha}_0)^T$ and $\widehat{\vartheta}$, respectively. The third term on the right of (34), describes the prediction's variability if μ_{S0}, α_0, and ϑ were known. The first two terms describe the contribution to the variability due to the need for estimation. It is useful to consider three scenarios.

Scenario 1. Estimation error in both the meta-analysis and the new trial. In the realistic case where the parameters in (1)–(2) and (32) are estimated, the prediction variance is (34), showing that the practical variability reduction in estimating $\beta + b_0$, coming from m_{S0} and a_0, will always be smaller than indicated by R^2_{trial}, which measures the "potential" validity of a surrogate endpoint

at trial level, assuming precise knowledge of the parameters in (1)–(2) and (32), e.g., obtained from a infinite number of trials of infinite size. See also Scenario 3.

Scenario 2. Estimation error only in the meta-analysis. This theoretical construct, requiring an infinite-sized new trial, provides information of practical interest since then the parameters of (32) are known and (34) reduces to

$$\text{Var}(\beta + b_0|\mu_{S0}, \alpha_0, \vartheta) \approx f\{\text{Var}(\widehat{\vartheta})\} + (1 - R^2_{\text{trial}})\text{Var}(b_0), \tag{35}$$

to be interpreted as the minimum variance achievable in the prediction of $\beta + b_0$. In practice, the meta-analysis will be finite and the first term on the right hand side of (35) will be present. This lead Gail *et al* (2000) to conclude that the use of surrogates, validated through the meta-analytic approach, will always be less efficient than the direct use of the true endpoint, but then still there can be gain in sample size, trial length, and/or number of life years.

Scenario 3. No estimation error. If the parameters of (1)–(2) and (32) were known, the prediction variance would reduce to (33), which is clearly linked with (26). This situation is of theoretical relevance for the insight it provides.

6.1 Surrogate Threshold Effect

We will outline the proposal of Burzykowski and Buyse (2006) for normally distributed endpoints. Assume that the prediction of $\beta + b_0$ can be made independently of μ_{S0}, then the conditional mean of $\beta + b_0$ is a simple linear function of α_0, while the conditional variance can be written as

$$\text{Var}(\beta + b_0|\alpha_0, \vartheta) \quad = \quad \text{Var}(b_0)\left(1 - R^2_{\text{trial(r)}}\right). \tag{36}$$

The $R^2_{\text{trial(r)}}$ in (36) is the squared correlation coefficient between b_i and a_i. In Scenario 2, the prediction variance is (36). In practice, an estimate $\widehat{\vartheta}$ is used and then prediction variance (35) ought to be applied:

$$\text{Var}(\beta + b_0|\alpha_0, \vartheta) \approx f\{\text{Var}(\widehat{\vartheta})\} + (1 - R^2_{\text{trial(r)}})\text{Var}(b_0), \tag{37}$$

which can be approximated by a quadratic function in α_0 (Verbeke and Molenberghs 2000).

Let us consider a (1-γ)100% prediction interval for $\beta + b_0$:

$$l(\alpha_0), u(\alpha_0) \equiv E(\beta + b_0|\alpha_0, \vartheta) \pm z_{1-\frac{\gamma}{2}} \sqrt{\text{Var}(\beta + b_0|\alpha_0, \vartheta)}, \tag{38}$$

where $z_{1-\gamma/2}$ is the $(1-\gamma/2)$ quantile of the standard normal distribution. One might then compute a value of α_0 such that

$$l(\alpha_0) = 0, \tag{39}$$

and term this value the *surrogate threshold effect* (STE). In some settings the upper limit is needed. The larger the prediction variance, the larger the absolute value of STE. From a clinical point of view, a large STE points to the need for observing a large treatment effect on the surrogate endpoint, which may cast doubts on a surrogate's usefulness, even when its $R^2_{\text{trial(r)}} \simeq 1$.

Both (38) and $l(\alpha_0)$ can be constructed using the variances in either (36) or (37), producing two versions of STE. The version obtained from (36) will be denoted by $\text{STE}_{\infty,\infty}$, the ∞ signs indicating that the measure assumes knowledge of ϑ and α_0. It captures a surrogate's "potential" validity. The notation $\text{STE}_{N,\infty}$ is used when variance (37) is employed, with N indicating the need for the estimation of ϑ. $\text{STE}_{N,\infty}$ captures the surrogate's "practical" validity. A surrogate may be potentially but not practically valid.

Interestingly, the STE can be expressed in terms of treatment effect on the surrogate necessary to be observed to predict a significant treatment effect on the true endpoint. In a practical application, one would seek a value of STE (preferably, $\text{STE}_{N,\infty}$) well within the range of treatment effects on surrogates observed in previous clinical trials, as close as possible to the (weighted) mean effect.

Apart from the normal-endpoints case reviewed here, Burzykowski and Buyse (2006) derived the STE when, perhaps for numerical convenience, the two-stage approach of Section 3.2 is used. Furthermore, STE can be computed for any type of surrogate. To this end, one merely needs to use an appropriate joint model for surrogate and true endpoints, capable of providing the required treatment effect. Burzykowski and Buyse (2006) presented time-to-event applications.

7 Concluding Remarks

Over the years, a variety of surrogate marker evaluation strategies have been proposed, cast within a meta-analytic framework. With an increasing range of endpoint types considered, such as continuous, binary, time-to-event, and longitudinal endpoints, also the scatter of types of measures proposed has increased. Some of these measures are difficult to calculate from fully specified hierarchical models, which has sparked of the formulation of simplified strategies. We reviewed the ensuing divergence of proposals, which then has triggered efforts of convergence, eventually leading to the information-theoretic approach, which is both general and simple to implement. These developments have been illustrated using data from clinical trials in schizophrenia and opthalmology.

While quantifying surrogacy is important, so is prediction of the treatment effect in a new trial based on the surrogate. Work done in this area has been reviewed, with emphasis on the so-called surrogate threshold effect and the sources of variability involved in the prediction process. A connection with the information-theoretic approach is pointed out.

Even though more work is called for, we believe the information-theoretic appraoch and the surrogate threshold effect are promising paths towards effective assessment and use of surrogate endpoints in practice. Software implementations are available from www.uhasselt.be/censtat.

Acknowledgment

We gratefully acknowledge support from Belgian IUAP/PAI network "Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data".

References

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

Alonso, A., Geys, H., Molenberghs, G., and Vangeneugden, T. (2003). Validation of surrogate

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

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

Alonso, A. and Molenberghs, G. (2006). Surrogate marker evaluation from an information theoretic perspective. *Biometrics*, **00**, 000–000.

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

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

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

Burzykowski, T., Molenberghs, G., Buyse, M., Renard, D., and Geys, H. (2001). Validation of surrogate endpoints in multiple randomized clinical trials with failure-time endpoints. *Applied Statistics*, **50**, 405–422.

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.

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.

Clayton, D.G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, **65**, 141–151.

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.

Dale, J.R. (1986). Global cross ratio models for bivariate, discrete, ordered responses. *Biometrics*, **42**, 909–917.

Daniels, M.J. and Hughes, M.D. (1997). Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine*, **16**, 1515–1527.

DeGruttola, V. and Tu, X.M. (1994). Modelling progression of CD-4 lymphocyte count and its relationship to survival time. *Biometrics*, **50**, 1003–1014.

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

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

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

Gail, M.H., Pfeiffer, R., van Houwelingen, H.C., Carroll, R.J. (2000). On meta-analytic assessment of surrogate outcomes. *Biostatistics* **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.

Genest, C. and McKay, J. (1986). The joy of copulas: bivariate distributions with uniform marginals. *American Statistician*, **40**, 280–283.

Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, **73**, 387–396.

International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (1998). ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials. *Federal Register*, **63**, No. 179, 49583.

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.

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. *Annual Review of Pharmacological Toxicology*, **41**, 347–366.

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

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

Renard, D., Geys, H., Molenberghs, G., Burzykowski, T., and Buyse, M. (2002). Validation of surrogate endpoints in multiple randomized clinical trials with discrete outcomes. *Biometrical*

Journal, **44**, 1–15.

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

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. *Statistics in Medicine*, **15**, 1999–2012.

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

Shih, J.H. and Louis, T.A. (1995). Inferences on association parameter in copula models for bivariate survival data. *Biometrics*, **51**, 1384–1399.

Tibaldi, F.S, Cortiñas Abrahantes, J., Molenberghs, G., Renard, D., Burzykowski, T., Buyse, M., Parmar, M., Stijnen, T., and Wolfinger, R. (2003). Simplified hierarchical linear models for the evaluation of surrogate endpoints. *Journal of Statistical Computation and Simulation*, **73**, 643–658.

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

Table 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.*

Table 2: *Age-related macular degeneration trial. 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	0.692	0.696	0.661	0.796
		Bivariate approach		
Center	0.692	0.696	0.999∗	
		Reduced Model		
		Univariate approach		
Center	0.641	0.656	0.677	0.793
		Bivariate approach		
Center	0.641	0.656	0.999∗	

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

Table 3: *Age-related macular degeneration trial. Results of the trial-level (R^2_{trial}) surrogacy analysis 0/1 coding.*

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

Table 4: *Examples of possible surrogate endpoints in various diseases (Abbreviations: AIDS = acquired immune deficiency syndrome; ARMD = age-related macular degeneration; HIV = human immunodeficiency virus).*

Disease	Surrogate Endpoint	Type	Final Endpoint	Type
Resectable solid tumor	Time to recurrence	Censored	Survival	Censored
Advanced cancer	Tumor response	Binary	Time to progression	Censored
Osteoporosis	Bone mineral density	Longitudinal	Fracture	Binary
Cardiovascular disease	Ejection fraction	Continuous	Myocardial infraction	Binary
Hypertension	Blood pressure	Longitudinal	Coronary heart disease	Binary
Arrhythmia	Arrhythmic episodes	Longitudinal	Survival	Censored
ARMD	6-month visual acuity	Continuous	24-month visual acuity	Continuous
Glaucoma	Intraoccular pressure	Continuous	Vision loss	Censored
Depression	Biomarkers	Multivariate	Depression scale	Continuous
HIV infection	CD4 counts + viral load	Multivariate	Progression to AIDS	Censored

Table 5: *Schizophrenia study. Trial-level and individual-level validation measures (95% confidence intervals). Binary-binary case.*

Parameter	Estimate	95% C.I.
Trial-level R^2_{trial} measures		
1.1 Information-theoretic	0.49	(0.21,0.81)
1.2 Probit	0.51	(0.18,0.78)
1.3 Plackett-Dale	0.51	(0.21,0.81)
Individual-level measures		
R^2_h	0.27	(0.24,0.33)
$R^2_h\text{max}$	0.39	(0.35,0.48)
Probit	0.67	(0.55,0.76)
Plackett-Dale ψ	25.12	(14.66;43.02)
Fano's lower-bound	0.08	

Table 6: *Age-related macular degeneration trial. Trial-level and individual-level validation measures (95% confidence intervals). Binary-binary case.*

Parameter	Estimate	95% C.I.
R^2_{trial}	0.38	(0.15;0.61)
R^2_h	0.26	(0.22;0.37)
$R^2_h\text{max}$	0.50	(0.33;0.60)