Chapter 13
Statistical Validation of Surrogate Markers in Clinical Trials

Ariel Alonso, Geert Molenberghs, and Gerard van Breukelen

Abstract The increasing cost of drug development has raised the demand on the use of biomarkers as surrogate endpoints for the evaluation of new drugs in clinical trials. However, failed past attempts to use surrogate endpoints made it clear that, before deciding on the use of a candidate surrogate endpoint, it is of the utmost importance to investigate its validity. Such validation process has proven challenging for conceptual and practical reasons. In the present chapter, some of the statistical methods introduced for the evaluation of surrogate markers will be discussed. Emphasis will be made on the so-called meta-analytic approach and its information-theoretic version, where information from several units is combined to carry out the validation exercise. The methods will be illustrated using a case study in ophthalmology.

13.1 Motivations and Antecedents

Recent discoveries in medicine and biology are opening an entire range of possibilities for the development of new treatments. However, these unquestionable achievements are also facing us with the challenge of having to evaluate a large number of promising therapies, using increasingly complex and costly clinical trials [2].

One of the most important factors influencing the duration and complexity of modern clinical trials is the choice of the endpoint used to assess drug efficacy. Actually, the most sensitive and relevant clinical endpoint, the so-called "true" endpoint, might often be difficult to use. This can happen, for instance, if measurement of the true endpoint is costly (e.g., to diagnose "cachexia", a condition associated with malnutrition and involving loss of muscle and fat tissue, expensive

A. Alonso (✉) • G. van Breukelen
Department of Methodology and Statistics, Maastricht University, Maastricht, The Netherlands
e-mail: ariel.alonso@maastrichtuniversity.nl; gerard.vbreukelen@maastrichtuniversity.nl

G. Molenberghs
I-BioStat, Universiteit Hasselt, Diepenbeek, Belgium

KU Leuven - University of Leuven, Leuven, Belgium
e-mail: geert.molenberghs@uhasselt.be

K. van Montfort et al. (eds.), *Developments in Statistical Evaluation of Clinical Trials*, DOI 10.1007/978-3-642-55345-5_13

equipment measuring content of nitrogen, potassium and water in the patient's body is required); requires a long follow-up time (e.g., survival in early stage cancers); or requires a large sample size due to a low incidence of the event (e.g., short-term mortality in patients with suspected acute myocardial infarction). A plausible strategy in these circumstances is the use of biomarkers for efficacy. The pursue of this strategy has been further encouraged by recent developments in many medical and biological fields that have considerably increased the number of promising biomarkers for the assessment of efficacy. In addition, a growing 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 [15].

Basically, one would like to replace the problematic true endpoint by a biomarker, which is measured earlier, more conveniently, or more frequently. From a regulatory perspective, a biomarker is not considered an acceptable endpoint for a determination of efficacy of new drugs, unless it has been shown to function as a valid indicator of clinical benefit, i.e., unless it is a valid surrogate marker [5].

Because of the possible benefits for the duration and cost of clinical trials, surrogate markers have been used in medical research for a long time [12, 14]. However, in spite of all its potential advantages, the use of surrogate endpoints in the development of new therapies has always been controversial. This may be due to a number of unfortunate historical instances where treatments showing a highly positive effect on a surrogate endpoint, were ultimately shown to be detrimental to the subjects' clinical outcome. One of such unfortunate events was the approval by the Food and Drug Administration (FDA) in the United States of three antiarrhythmic drugs: encainide, flecainide and moricizine, based on their efficacy to effectively suppress arrhythmias. It was believed that, since arrhythmia is associated with an almost fourfold increase in the rate of cardiac-complication-related death, the drugs would also reduce the death rate. Nonetheless, a clinical trial conducted after the drugs had been approved and introduced into clinical practice showed that, in fact, the death rate among patients treated with encainide and flecainide was more than twice the one among patients treated with placebo [8]. An increase of the risk was also detected for moricizine.

Behind many of these failures in the initial use of surrogate endpoints, was the logical but naive perception that surrogacy could be established by only evaluating the association between the biomarker on the one hand and the corresponding true endpoint on the other hand. Nevertheless, these failed past attempts made clear that the mere existence of an association between a biomarker and the true endpoint is not sufficient for using the former as a surrogate, i.e., a good correlate is not automatically a good surrogate [14]. The recognition of this fact opened an exciting and fruitful debate about the properties that a good surrogate should satisfy. After more than 20 years of research, this debate is far from settled and many questions and practical issues still need to be addressed. This notwithstanding, our level of knowledge has been dramatically increased and plethora statistical methods are now available for the evaluation of surrogate markers.

In Sect. 13.2 some important definitions are given. The single-trial methods and the meta-analytic approach to the validation of surrogate markers are introduced

in Sects. 13.3 and 13.4 respectively. Section 13.5 describes some of the issues that emerge when the true and/or the surrogate endpoints are not normally distributed and in Sects. 13.6 and 13.7 a unified approach based on information theory is introduced. The meta-analytic approach is illustrated using a case study in Sect. 13.8 and the implementation of this method in widely used software packages is addressed in Sect. 13.10. Eventually, some final comments are presented in Sect. 13.11.

13.2 Some General Definitions

The terms "endpoint", "biomarker", and "marker" have often been interchangeably used to refer simply to a 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 frequently suggested as potential surrogates for those that are measured later. The following definitions, introduced by the Biomarker Definitions Working Group, are nowadays widely accepted and adopted in the biomedical literature [4]:

- Clinical endpoint: a characteristic or variable that reflects how a patient feels, functions, or survives;
- Biomarker: a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention;
- Surrogate endpoint: a biomarker that is intended to substitute for a clinical endpoint. A surrogate endpoint is expected to predict clinical benefit (or harm or lack of benefit or harm).

It is important to point out that, although extremely useful, the previous definitions do not include all situations one may encounter in practice. For instance, in our case study we analyze a potential surrogate that is not a biomarker, but an intermediate endpoint that has clinical meaning of its own. This is frequently the case in medical fields like, for instance, oncology, where progression-free survival is often considered as a potential surrogate for survival.

13.3 Single-Trial Methods

All earlier approaches to the validation of surrogate markers were framed in a single-trial setting, i.e., it was assumed that information on both the surrogate (S) and the true endpoint (T) was available from a single clinical trial. Within this setting Prentice introduced in 1989 the first formal definition of surrogacy. Basically, Prentice 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 [21].

Symbolically, Prentice's definition can be written

$$f(S|Z) = f(S) \Leftrightarrow f(T|Z) = f(T)\,, \tag{13.1}$$

where $f(X)$ denotes the probability distribution of random variable X and $f(X|Z)$ denotes the probability distribution of X conditional on the treatment variable Z. Note that this definition involves the triplet (T, S, Z) and, consequently, the endpoint S is a surrogate for T always with respect to the effect of some specific treatment Z. This implies that, at least in principle, if a new treatment is considered, then the validation process would need to be repeated. Prentice and other authors supplemented the previous definition with the following set of operational criteria that has become known as the *Prentice's Criteria*: (1) treatment has a significant impact on the surrogate endpoint $f(S|Z) \neq f(S)$, (2) treatment has a significant impact on the true endpoint $f(T|Z) \neq f(T)$, (3) the surrogate endpoint has a significant impact on the true endpoint $f(T|S) \neq f(T)$, and (4) the full effect of treatment upon the true endpoint is captured by the surrogate $f(T|S, Z) = f(T|S)$ [5].

The latter two are Prentice's original criteria and it has been proven that the definition and criteria are only equivalent when both the surrogate and the true endpoints are binary [5]. Note that the first two criteria measure the departures from the null hypothesis used in (13.1) and the third criterion implies that the surrogate has a prognostic value for the true endpoint. Finally, the fourth criterion requires S to fully capture the effect of treatment on the true endpoint, that is, there is no effect of treatment on the true endpoint after correcting for the surrogate.

Freedman et al. argued that the last criterion raises conceptual problems, since it requires the statistical test for the treatment effect on the true endpoint to be non-significant after adjustment for the surrogate [16]. In general, the nonsignificance of this test does not prove that the effect of treatment on the true endpoint is totally captured by the surrogate [5, 13]. Freedman further proposed to shift the paradigm from hypothesis testing to estimation and to calculate the so-called proportion of treatment explained (*PTE*). The *PTE* is the proportion of the treatment effect on the true endpoint captured by the surrogate and is defined as $PTE = (\beta - \beta_S)\beta$, where β denotes the effect of the treatment on the true endpoint emanating from $f(T|Z)$ and β_S is the effect of the treatment on the true endpoint after adjusting by the surrogate and can be calculated using $f(T|S, Z)$.

Note that *PTE* is large when β_S is small relative to β, Prentice fourth criterion implies $\beta_S = 0$ and therefore, if this criterion holds, $PTE = 1$. Freedman suggested that a good surrogate is one for which *PTE* is close to one. However, some conceptual problems also surround *PTE*, the most paradoxical one is that it is not a proportion. In fact, *PTE* can take any value on the real line, making its

interpretation problematic [5]. Freedman himself acknowledged that the confidence limits for *PTE* will tend to be rather wide or even unbounded if Fieller's confidence intervals are used.

Frangakis and Rubin strongly criticized the conceptual foundation of Prentice's fourth criterion and the *PTE* [13]. They pointed out that the treatment effect on the true endpoint used in these two procedures is obtained after conditioning on the surrogate, i.e., a post-randomization variable and, consequently, is not a causal effect. Further, they proposed to assess surrogacy using the so-called *principal stratification* which is based on the potential outcomes model often used in causal inference. It has been argued that this method suffers from a similar drawback as the Prentice's definition and criteria, in that it is too stringent and difficult to implement in practice [27]. In addition, the intrinsically unobserved nature of the vector of potential outcomes implies that untestable assumptions are unavoidable.

In a separate line of research, Buyse et al. showed that, for continuous and normally distributed endpoints, *PTE* can be decomposed in three different quantities: the first one merely is the ratio of the surrogate and true endpoint variances and, therefore, it only represents a scale factor, the other two are the so-called relative effect *RE* and the adjusted association ρ_Z [7]. The relative effect is defined as $RE = \beta/\alpha$, where α is the treatment effect on the surrogate emanating from $f(S|Z)$ and β is defined as before. Notice that, unlike Prentice's fourth criterion and the *PTE*, the treatment effects involved in *RE* are not adjusted by post-randomization variables and, hence, have a direct causal interpretation. Indeed, α and β are simply the average causal effects of the treatment on the surrogate and the true endpoint respectively. The adjusted association is the correlation between the surrogate and the true endpoint after adjusting by treatment and is defined as $\rho_Z = \text{Corr}(S, T|Z)$.

The relative effect tries to enable prediction of the treatment effect on the true endpoint based on the treatment effect on the surrogate, but to do so strong and untestable assumptions have to be made. Essentially, in a single trial setting one is confronted with the problem of estimating the relationship between both average causal effects using a single observation, namely the vector of treatment effects (α, β). A way out of the problem is to assume that $E(\beta|\alpha) = RE \times \alpha$, i.e., the average causal effects satisfy the regression through the origin equation $\beta = RE \times \alpha + \varepsilon$. Regression through the origin has often been surrounded by controversy due to the paradoxical results it can produce, like negative coefficients of determination and negative F ratios. Even when there are theoretical reasons to believe that the function relating the two variables of interest does pass through the origin, regression through the origin may be problematic if the relationship between the variables of interest is not linear in a neighborhood of zero. Moreover, if the data at hand lie far from zero, then the assumption of linearity at this point becomes impossible to evaluate. This lack of replication is a fundamental problem of all the previously discussed approaches and it can only be overcome when more than one pair (α, β) is available for the analysis.

13.4 Data from Several Trials: The Meta-analytic Approach

Over the years, it has become clear that the single trial setting is too restrictive for the evaluation of surrogate markers and a general agreement has been growing regarding the need of replication at the trial level as well. A first formal proposal along these lines, using Bayesian methods, was given by Daniels and Hughes [11]. Buyse et al. extended these ideas using the theory of linear mixed-effects models and Gail et al. extended it further using generalized estimating equations methodology [7, 17]. In what follows, we describe the approach as proposed by Buyse et al. under the assumption that both endpoints are normally distributed and in Sects. 13.5–13.7 other types of endpoints will be addressed. To that end let us assume that data from $i = 1, \ldots, N$ trials are available, in the ith of which $j = 1, \ldots, n_i$ subjects are enrolled. Further, let us denote the true and surrogate endpoints for patient j in trial i by T_{ij} and S_{ij}, respectively, and the indicator variable for the new treatment by Z_{ij}. The random treatment allocation in a clinical trial context naturally leads to the following bivariate model

$$\begin{cases} T_{ij} = \mu_{Ti} + \beta_i Z_{ij} + \varepsilon_{Tij} , \\ S_{ij} = \mu_{si} + \alpha_i Z_{ij} + \varepsilon_{sij} , \end{cases} \tag{13.2}$$

where μ_{Ti} and μ_{si} are trial-specific intercepts quantifying the average response in the control group, β_i and α_i are trial-specific average causal effects and ε_{Tij} and ε_{sij} are correlated error terms, assumed to be zero-mean normally distributed with covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{TT} & \sigma_{TS} \\ \sigma_{TS} & \sigma_{SS} \end{pmatrix} , \tag{13.3}$$

i.e., (13.3) denotes the within-trial covariance matrix of T and S after adjusting by treatment and considering the patient the level of analysis. Furthermore, due to replication at the trial level, one can decompose the trial-specific parameters in the following way

$$\begin{pmatrix} \mu_{si} \\ \mu_{Ti} \\ \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \mu_s \\ \mu_T \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} m_{si} \\ m_{Ti} \\ a_i \\ b_i \end{pmatrix} , \tag{13.4}$$

where the second term on the right hand side of (13.4) is assumed to follow a zero-mean normal distribution with covariance matrix

$$\mathbf{D} = \begin{pmatrix} d_{ss} & d_{sT} & d_{sa} & d_{sb} \\ d_{sT} & d_{TT} & d_{Ta} & d_{Tb} \\ d_{sa} & d_{Ta} & d_{aa} & d_{ab} \\ d_{sb} & d_{Tb} & d_{ab} & d_{bb} \end{pmatrix} . \tag{13.5}$$

Essentially, (13.5) denotes the between-trial covariance matrix of intercepts and treatment effects on T and S, considering now trial the level of analysis. Buyse et al. investigated how the treatment effect on the true endpoint can be predicted by the treatment effect on the surrogate [7]. The main idea is to predict the treatment effect on T in a new trial $i = 0$ based on: (a) information obtained in the validation process using trials $i = 1, \ldots, N$, and (b) the estimate of the treatment effect on S in the new trial $i = 0$. To this end, these authors notice that $(\beta + b_0 | m_{S0}, a_0)$ follows a normal distribution with mean and variance

$$E(\beta + b_0 | m_{S0}, a_0) = \beta + \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} \mu_{S0} - \mu_S \\ \alpha_0 - \alpha \end{pmatrix} , \tag{13.6}$$

$$\text{Var}(\beta + b_0 | m_{S0}, a_0) = d_{bb} - \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix} . \tag{13.7}$$

If the treatment effect on the surrogate conveys a lot of information about the treatment effect on the true endpoint, then the conditional variance (13.7) will be close to zero. In that case, there would be an almost deterministic relationship between the treatment effects on the true and surrogate endpoint, and a very accurate prediction of the first one would be possible if the second one has been observed. Based on these ideas Buyse et al. proposed to assess surrogacy at the trial level using the coefficient of determination

$$R^2_{\text{trial}} = R^2_{b_i | m_{Si}, a_i} = \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix} . \tag{13.8}$$

This coefficient measures how precisely the treatment effect on the true endpoint can be predicted, provided that the treatment effect on the surrogate endpoint has been observed in a new trial ($i = 0$). It is unitless and ranges in the unit interval if the corresponding covariance matrix $\mathbf{D}$ is positive-definite, two desirable features for its interpretation.

One special case of the model given in (13.2) is the so-called reduced model, which assumes that the intercepts, i.e. the average responses in the control group, are constant across trials. Under this assumption, expressions (13.6) and (13.7) reduce to

$$E(\beta + b_0 | a_0) = \beta + \frac{d_{ab}}{d_{aa}} (\alpha_0 - \alpha) ,$$

$$\text{Var}(\beta + b_0 | a_0) = d_{bb} - \frac{d^2_{ab}}{d_{aa}} ,$$

with corresponding

$$R^2_{\text{trial}} = R^2_{b_i | a_i} = \frac{d^2_{ab}}{d_{aa} d_{bb}} . \tag{13.9}$$

Similar to the logic in (13.6) and (13.7), the conditional model for β_i given μ_{si} and α_i can be written as

$$\beta_i = \theta_0 + \theta_1 \mu_{si} + \theta_2 \alpha_i + \varepsilon_i \ , \tag{13.10}$$

where expressions for the coefficients $(\theta_0, \theta_1, \theta_2)$ follow from (13.4) and (13.5). In case the surrogate is perfect at the trial level ($R^2_{\text{trial}} = 1$), the error term in (13.10) vanishes and the linear relationship becomes deterministic, implying that β_i *equals* the systematic component of (13.10).

Notice first that, unlike for the *RE*, the regression line (13.10) does not necessarily pass through the origin. Secondly, this new approach avoids the conceptual problems surrounding the *RE*, since the relationship between β_i and α_i is studied across a family of units, rather than in a single unit. By virtue of replication, it is possible to *check* the stated relationship for the treatment effects and, if the posited linear relation does not hold, alternative regression functions can be considered. Nevertheless, one has to be aware of a potentially low power to discriminate between candidate regression functions.

At the individual level, one tries to assess how an individual's surrogate outcome is predictive for the true endpoint outcome. To this end, one needs to construct the conditional distribution of T, given S and Z. From (13.2) we obtain

$$T_{ij}|Z_{ij}, S_{ij} \sim N\left\{\mu_{Ti} - \sigma_{TS}\sigma_{SS}^{-1}\mu_{Si} + (\beta_i - \sigma_{TS}\sigma_{SS}^{-1}\alpha_i)Z_{ij} \right.$$
$$\left. + \sigma_{TS}\sigma_{SS}^{-1}S_{ij}; \sigma_{TT} - \sigma_{TS}^2\sigma_{SS}^{-1}\right\}.$$

The association between both endpoints after adjustment by treatment is captured by the coefficient of determination

$$R^2_{\text{ind}} = \frac{\sigma^2_{ST}}{\sigma_{SS}\sigma_{TT}}. \tag{13.11}$$

Basically, the R^2_{ind} is the squared correlation between both endpoints once we have adjusted for treatment and trial and, therefore, it is a natural extension of the adjusted association. Unlike the trial level surrogacy, the individual level does not depend on the treatment and it can be interpreted as a quantification of the biological plausibility of the surrogate. An endpoint producing a high individual level surrogacy is always a potential surrogate, however, it may fail to be predictive at the trial level for a specific treatment that follows a causal path that completely avoids it.

Although elegant, the above hierarchical model often poses a considerable computational challenge [5]. To address this problem, Tibaldi et al. suggested several simplifications, like treating the trial-specific parameters in (13.2) as fixed effects in a two-stage approach [25]. The first-stage model will take the form (13.2)

and at the second stage, the estimated treatment effect on the true endpoint is regressed on the estimated treatment effect on the surrogate and the intercept associated with the surrogate endpoint as

$$\hat{\beta}_i = \lambda_0 + \lambda_1 \hat{\mu}_{si} + \lambda_2 \hat{\alpha}_i + \varepsilon_i \ . \tag{13.12}$$

Essentially, the trial-level surrogacy R^2_{trial} is assessed by regressing $\hat{\beta}_i$ on $(\hat{\mu}_{si}, \hat{\alpha}_i)$ and the individual-level value is calculated as before, using the estimates from (13.3). Notice that, when the fixed-effects approach is 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 (13.12) in the second stage.

Another cornerstone of the meta-analytic method is the choice of unit of analysis such as, for example, trial, center, or country. 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. Of course, after choosing a specific unit for the analysis, one always 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. This issue has been covered at large by Cortiñas et al. and we refer the interested reader to this work for more details [9].

13.5 Other Types of Endpoints

In the previous section, the formalism developed by Buyse et al. was introduced using the *simplest* setting where both endpoints are Gaussian random variables measured cross-sectionally. However, this is not always the case, for example, one can encounter:

- Binary (dichotomous): the surrogate and/or true endpoints are binary, for instance, biomarker value below or above a certain threshold (e.g., viral load in HIV+ patients below detection limit) or clinical "success" (e.g., tumor shrinkage).
- Categorical (polychotomous): the surrogate and/or true endpoints are categorical, for instance, biomarker value falling in successive, ordered classes (e.g., cholesterol levels <200, 200–299, 300+ mg/dl) or clinical response (e.g., complete response, partial response, stable disease, progressive disease).

- Longitudinal or repeated measures: the surrogate and/or true endpoints are longitudinally measured, for instance, biomarker (e.g., CD4+ counts over time) or clinical outcome (e.g., blood pressure over time).
- Multivariate longitudinal: the surrogate and/or true endpoints are multivariate outcomes measured longitudinally, for instance, several biomarkers (e.g., CD4+ and viral load over time) or several clinical measurements (e.g., dimensions of quality of life over time).
- Time to event: the surrogate and/or true endpoints are failure-time random variables, for instance, time to cancer recurrence as a surrogate marker for survival.

Assessing surrogacy in these more complex scenarios raises a number of difficult challenges. Firstly, one now needs to deal with highly complicated hierarchical models. These models frequently bring severe numerical issues and the use of alternative, simplified approaches like the ones proposed by Tibaldi et al., becomes unavoidable. Secondly, based on the outputs of these models, one needs to define meaningful measures to quantify surrogacy at both the trial and individual level.

If one is ready to only consider linear models to study the relationship between the treatment effect on the surrogate and the true endpoint, then the methodology previously described can be applied in a straightforward fashion to quantify trial level surrogacy. At the individual level, however, abandoning the realm of normality has much deeper implications. Indeed, based on this meta-analytic paradigm, several individual-level measures have been proposed. For instance, in the binary-binary setting Renard et al. assumed that the observed dichotomic outcomes emerge from two latent and normally distributed variables $(\tilde{S}, \tilde{T})$. Essentially, it is assumed that the surrogate (true endpoint) takes value one when corresponding latent variable exceeds a threshold value, i.e., when $\tilde{S} > \eta_S$ ($\tilde{T} > \eta_T$) and zero otherwise. In this framework, using a bivariate probit model, these authors defined individual-level surrogacy as $R^2_{\text{ind}} = \rho^2_{\tilde{S}\tilde{T}}$, which is the correlation at the latent level. Alternatively, they also defined $R^2_{\text{ind}} = \psi$, the global odds ratio between both binary endpoints estimated from a so-called bivariate Plackett-Dale model [22].

When the true endpoint is a survival time and the surrogate is a longitudinal sequence, Renard et al., using Henderson's model, proposed to study the individual level based on a time function defined as $R^2_{\text{ind}}(t) = \text{corr}[W_1(t), W_2(t)]^2$, where $(W_1(t), W_2(t))$ is a latent bivariate Gaussian process [23]. Burzykowski et al. approached the case of two failure-time endpoints based on copula models and quantified the individual level surrogacy using Kendall's τ [6].

Using multivariate ideas, the so-called R^2_Λ has been proposed to evaluate surrogacy when both responses are measured longitudinally [1]. The R^2_Λ coefficient quantifies the association between both longitudinal sequences and is defined using the covariance matrices emanating from a hierarchical model that characterized the joint distribution of both endpoints. Furthermore, the 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, one can define a family of individual-level parameters [1].

All these examples underscore 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, in all cases, a joint and often non-standard model for both endpoints is needed, frequently representing a serious computational burden. In the next section, a unified approach to the validation of surrogate markers based on information theory will be introduced. Furthermore, it will be argued that this approach may help to overcome some of the aforementioned problems.

13.6 An Information-Theoretic Unification

Information theory, originated as a rigorous science in the 1940s, deals with the study of problems concerning complex systems, and has been applied in a variety of fields such as modern communication theory. In spirit and concepts, information theory has its mathematical roots connected with the idea of disorder or entropy used in thermodynamics and statistical mechanics. An early attempt to formalize the theory was made by Nyquist in 1924 who recognized the logarithmic nature of information [19]. Another major contribution in this area came in 1948 when Shannon published a remarkable paper on the properties of information sources and communication channels [24].

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 in 1951 studied another statistical information measure, involving two probability distributions associated with the same experiment [18].

The concept of entropy lies at the center of information theory and it can be interpreted as a measure of the randomness or uncertainty associated with a random variable. If Y is a discrete random variable taking values $\{k_1, k_2, \ldots, k_m\}$ with probability function $P(Y = k_i) = p_i$, then the entropy of Y is defined as

$$H(Y) = -E[\log P(Y)] = -\sum_i p_i \log p_i \ .$$

$H(Y)$ can be interpreted as the average uncertainty associated with P. The joint and conditional entropies are defined in an analogous fashion. Entropy is always non-negative and satisfies $H(Y|X) \leq H(Y)$ for any pair of random variables (X, Y), with equality holding under independence. Basically, the previous inequality states that uncertainty about Y can only decrease if additional information (X) becomes available. Furthermore, entropy is invariant under a bijective transformation [10].

Similarly, the so-called differential entropy $h_d(Y)$ of a continuous random variable Y with density $f_Y(y)$ and support S_{f_Y} is defined as

$$h_d(Y) = -E[\log f_Y(Y)] = -\int_{S_{f_Y}} f_Y(y) \log f_Y(y)\, dy\ .$$

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 $W = \upsilon(Y)$, it follows that $h_d(W) = h_d(Y) - \mathrm{E}_W\left(\log\left|\frac{d\upsilon^{-1}}{dw}(W)\right|\right)$.

One can now quantify the amount of uncertainty in Y, expected to be removed if the value of X were known, by $I(X,Y) = h(Y) - h(Y|X)$, the so-called *mutual information*, where $h = H$ in the discrete case and $h = h_d$ for continuous random variables. 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 $I(X,X) = h(X)$.

Additionally, if $\boldsymbol{Y}$ is a n-dimensional random vector, then the entropy-power of $\boldsymbol{Y}$ can be defined as

$$\mathrm{EP}(\boldsymbol{Y}) = \frac{1}{(2\pi e)^n} e^{2h(\boldsymbol{Y})}\ .$$

The differential entropy of a continuous normal random variable is given by $h(Y) = \frac{1}{2}\log\left(2\pi e\sigma^2\right)$, a simple function of the variance and, therefore, on the natural logarithmic scale $\mathrm{EP}(Y) = \sigma^2$, i.e., for the normal distribution variability and information are equivalent concepts. However, this equivalence does not hold in the general case. Indeed, in general, $\mathrm{EP}(Y) \leq \mathrm{Var}(Y)$ with equality if and only if Y is normally distributed.

We can now define an information-theoretic measure of association as

$$R_h^2 = \frac{\mathrm{EP}(\boldsymbol{Y}) - \mathrm{EP}(\boldsymbol{Y}|\boldsymbol{X})}{\mathrm{EP}(\boldsymbol{Y})}\ , \tag{13.13}$$

which ranges in the unit interval, equals zero if and only if $(\boldsymbol{X}, \boldsymbol{Y})$ are independent, is symmetric, is invariant under bijective transformation of $\boldsymbol{X}$ and $\boldsymbol{Y}$, and, when $R_h^2 \to 1$ for continuous models, there is usually some degeneracy appearing in the distribution of $(\boldsymbol{X}, \boldsymbol{Y})$; often $\boldsymbol{Y} = \phi(\boldsymbol{X})$ with probability one for some nontrivial function ϕ. This means that there exists a deterministic relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$. There is a direct link between R_h^2 and the mutual information: $R_h^2 = 1 - e^{-2I(\boldsymbol{X},\boldsymbol{Y})}$. For $\boldsymbol{Y}$ discrete: $R_h^2 \leq 1 - e^{-2H(\boldsymbol{Y})}$, implying that R_h^2 has an upper bound smaller than 1; in this setting it is better to consider

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

reaching 1 when both endpoints are deterministically related.

Surrogacy can now be redefined preserving previous proposals as special cases. It is important to point out that, although the focus will be on the individual-level surrogacy, all results apply to the trial level as well. Let $Y = T$ and $X = S$ be the true and surrogate endpoints, respectively. S would be considered 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. This definition, in spite of being based on formal concepts rooted in information theory, is simple and intuitive, since the idea behind surrogacy is to reduce our lack of knowledge about a true endpoint through the use of a surrogate alternative. At the trial level, the situation is similar: we want to gain information about the unobserved treatment effect on the true endpoint using the known treatment effect on the surrogate.

The R_h^2 coefficient is a valuable tool to evaluate surrogacy in practice. $R_h^2 \approx 1$ 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, $R_h^2 \approx 0$ evidences a poor surrogate, unable to reduce our uncertainty about the true endpoint.

For the cross-sectional normal-normal case, Alonso and Molenberghs have shown that $R_h^2 = R_{\text{ind}}^2$ [1]. The same holds for R_A^2, defined in a longitudinal context. Finally, when the true and surrogate endpoints have distributions in the exponential family, then LRF $\xrightarrow{P}$ R_h^2 when the number of subjects per trial goes to infinity, where LRF denotes the likelihood reduction factor introduced by Alonso et al. [3]. These authors also showed that (13.13) can be estimated based on $f(T|Z,S)$ and $f(T|Z)$, i.e., two univariate models that can often be easily fitted using standard software packages, in contrast to the original meta-analytic approach that requires the fitting of the complex joint hierarchical model $f(T,S|Z,\alpha,\beta)$.

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

Fano's inequality relates prediction accuracy with different information-theoretic concepts and, when applied to the evaluation of surrogate endpoints, this inequality sets a limit for our capacity to successfully predict the true endpoint using the surrogate [3, 10]. For continuous endpoints it can be written as

$$\mathrm{E}\left[(T - g(S))^2\right] \geq \mathrm{EP}(T)(1 - R_h^2)\ . \tag{13.14}$$

Note that nothing has been assumed about the distribution of both the surrogate and true endpoint and no specific form has been considered for the prediction function g.

Essentially, Fano's inequality states a lower bound for the prediction error and this lower bound can be decomposed in two different elements. The second element on the right side of (13.14) depends on the surrogate through the value of R_h^2, the first element, however, is an intrinsic characteristic of the true endpoint and it is independent of the surrogate. It is clear from (13.14) that the prediction error increases with EP(T) and, consequently, if the true endpoint has a large entropy-power then a surrogate should produce a close to one R_h^2 to have some predictive value. In other words, the surrogate would need to be almost deterministically related to the true endpoint to have some predictive power. Essentially, this inequality hints on the fact that, for some true endpoints, the search for a good surrogate may be a dead end street.

13.8 An Age-Related Macular Degeneration (ARMD) Trial

In what follows, the use of the meta-analytic approach will be illustrated using a clinical trial involving patients suffering from age-related macular degeneration (ARMD), a condition in which patients progressively lose vision [20]. Overall, 240 patients from 43 centers participated in the trial. Patients' visual acuity was assessed using standardized vision charts (see Fig. 13.1) displaying lines of five letters of decreasing size, which patients had to read from top (largest letters) to bottom (smallest letters).

The visual acuity was measured by the total number of letters correctly read. In this example, the binary indicator for treatment (Z) is set to -1 for placebo and

Fig. 13.1 Visual acuity study. Visual chart

to 1 for treatment with interferon-α. The surrogate endpoint S is the change in the visual acuity at 6 months after starting treatment, while the true endpoint T is the change in the visual acuity at 1 year. In the meta-analytic approach the centers in which the patients were treated will be considered the units of analysis. Two out of 43 centers participating in the trial enrolled patients only to one of the two treatment arms. These centers were excluded from considerations. A total of 41 centers were thus available for analysis.

13.9 Analysis of the ARMD Trial

In this section, the data from the age-related macular degeneration trial, described in Sect. 13.8, are used to evaluate visual acuity at 6 months as a surrogate endpoint for visual acuity at 1 year. Primarily, one would like to assess, for a given patient, how much information his visual acuity at 6 months provides on his visual acuity at 1 year and, similarly, one would also like to assess how much information the treatment effect at 6 months conveys about the treatment effect at 1 year. These are the questions addressed by the individual- and trial-level surrogacy. Notice that the individual level may be especially relevant for a treating physician who, having observed a particular outcome for a patient with a treatment at 6 months, wants to know what this means for the status of the patient at 1 year. On the other hand, the trial level may be more relevant for a data analyst that wants to know if the follow up period of a new trial might be shorten by 6 months in order to reduce cost.

Figure 13.2 shows the scatterplot of the two endpoints for all patients included in the trial. Clearly, there is a correlation between both variables. Indeed, the estimated Pearson correlation coefficient equals 0.757 and the 95 % confidence interval is $CI_{95\,\%} = (0.688, 0.812)$. We have learned in previous sections that, although appealing, the existence of correlation does not imply that visual acuity at 6 months

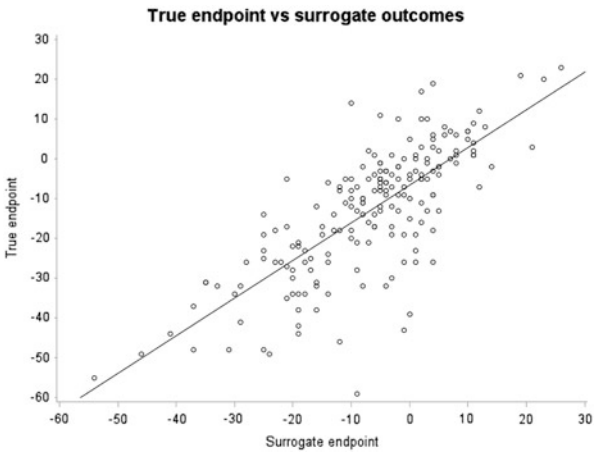

Fig. 13.2 Age-related macular degeneration trial. True endpoint (change in visual acuity at 1 year) versus surrogate endpoint (change in visual acuity at 6 months) for all individual patients, raw data

is a valid surrogate and further analyses are needed. In the present section we will follow the multi-units paradigm introduced in Sect. 13.4.

Using similar data, Buyse et al. experienced problems when fitting the full random-effects model, irrespective of whether standard statistical software or user developed alternatives were employed [7]. Similarly, our attempt to fit the complete hierarchical model given in (13.2) produced an infinite likelihood and the resulting **D** matrix was ill-conditioned with a condition number equal to 5.852−E15.

It is important to point out that when the full bivariate random-effects model is used, severe numerical issues are often encountered, especially if the surrogate and/or the true endpoint are not normally distributed. This numerical issues may have a huge impact on the assessment of surrogacy, particularly at the trial level. Indeed, the R^2_{trial} is computed based on the covariance matrix **D** and it is possible that this matrix becomes ill-conditioned and/or non-positive definite due to numerical problems. In such cases, the resulting quantities computed based on this matrix might not be trustworthy. For example, in our case study, the estimated **D** matrix produced a $R^2_{\text{trial}} = 0.972$ with a 95 % confidence interval $(0.955, 0.989)$. Although possible, such a large value for the trial level surrogacy inevitably raises some doubts. Obviously, this result emanates from an ill-conditioned matrix and is probably misleading. 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 the maximization procedure used to calculate the maximum likelihood estimators converges to a boundary solution. Thus, when using the full hierarchical model in the validation process, it is always necessary to check the D matrix to evaluate the presence of these issues.

Due to the numerical problems found with the ARMD data when fitting the complete hierarchical model, simplifying strategies along the lines introduced by Tibaldi et al. were called for and a two-stage approach was adopted [25]. At a first stage, the bivariate regression model given in (13.2) was fitted considering the trial-specific parameters as fixed effects. Within the two-stage approach, Tibaldi et al. explored two plausible strategies for fitting the model in (13.2), the so-called univariate and bivariate strategies, taking into account whether the surrogate and true endpoints are modeled as a bivariate outcome or rather as two univariate ones. In the latter case, the correlation between both endpoints is not incorporated into the model, rendering the study of the individual-level surrogacy more involved. However, it is important to point out that, if the trial-level surrogacy is of most interest and the investigation of the individual-level surrogacy is only of secondary importance, then the adoption of the univariate strategy can largely ease the computational burden in some scenarios. For the ARMD trial, the bivariate strategy was feasible and, hence, always adopted. In addition, the reduced model that assumes constant intercepts across units was also employed. Finally, at the second stage, one can consider weighted and unweighted versions of the model given

Table 13.1 Results of the trial and individual level surrogacy: R^2_{trial}, R_{ind} and 95 % confidence intervals (CI) obtained using the Delta method for the ARMD trial

Full model		
	Unweighted	Weighted
R^2_{trial}	0.381	0.437
R^2_{trial} CI	(0.138, 0.6234)	(0.200, 0.674)
R^2_{ind}&CI	0.512, CI= (0.422, 0.601)	
Reduced model		
	Unweighted	Weighted
R^2_{trial}	0.601	0.517
R^2_{trial} CI	(0.404, 0.797)	(0.297, 0.738)
R^2_{ind}&CI	0.581, CI= (0.499, 0.662)	

in (13.12) to estimate the trial level surrogacy. A summary of all these analyses is given in Table 13.1.

Note firstly that the individual-level surrogacy is estimated at the first stage and, consequently, it is not affected by the strategies followed to fit the second-stage model (weighted/unweighted). Secondly, the R^2_{ind} produced very similar results for both the reduced and full models. However, the AIC associated with the reduced and full model were 2668.3 and 2185.4 respectively, indicating that the assumption of equal intercepts across units produced a poorer fit to the data.

At the trial level, the results are much more variable, with the estimates R^2_{trial} varying from 0.38 to 0.60 across different settings. Because the full model seems to produce a better description of the data in what follows we will focus on the results displayed at the top panel of Table 13.1.

Taking into account that the sample size greatly varied across centers, one may consider a weighted analysis a more reliable option in this case. Nonetheless, the point estimate of R^2_{trial} was similar when the weighted or unweighted strategy was used and the confidence intervals largely overlapped in both scenarios. The general conclusion is that the trial level surrogacy seems to be rather weak, with the upper bound of the confidence intervals never exceeding 0.7.

Figure 13.3 displays the results obtained with the full-weighted model approach. Figure 13.3a shows a plot of the treatment effects on the true endpoint by the treatment effects on the surrogate endpoint and the size of the points are proportional to the sample size of each center. These effects are weakly correlated. Figure 13.3b shows a certain degree of correlation between the measurements at 6 months and at 1 year, after correction for treatment effect and center. Based on the previous findings, even with the limited data available, one may conclude that the assessment of visual acuity at 6 months seems to be a poor surrogate for the same assessment at 1 year.

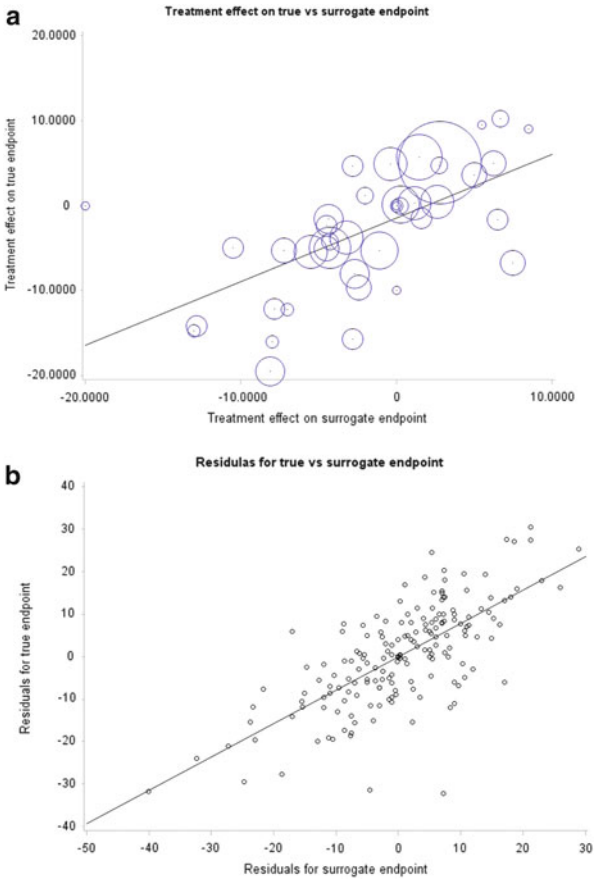

Fig. 13.3 Age-related macular degeneration trial. (**a**) Treatment effects on the true endpoint versus treatment effects on the surrogate endpoint in all centers. The size of each point is proportional to the number of patients in the corresponding center. (**b**) True endpoint versus surrogate endpoint for all individual patients, after correction for treatment effect

13.10 Software Packages

R functions and SAS macros have been developed to implement the methods discussed in the previous sections [26]. The ARMD trial was analyzed using the macro SURCONCON in SAS 9.3. The macro is a slight modification of the one that can be downloaded from http://www.ibiostat.be/software/surrogate.asp. The SAS code to carry out the analysis, the modified version of the macro and the data set will be available from the book's website. A detailed account of the macro can also be found in [26].

13.11 Conclusion

The initial enthusiasm that accompanied the use of surrogate markers, was followed by concern and skepticism after some dramatic failures. However, these failures opened a fruitful and stimulating scientific debate that has resulted in the development of different approaches and schools of thoughts for the validation of surrogate markers [2]. It is now clear that surrogate markers are a powerful tool that can play an important role in the drug development process. But it has also transpired that they need to be properly evaluated. Consequently, the initial enthusiasm and subsequent skepticism have been substituted by a more scientific and objective comprehension of their potentials and limitations.

At the same time, regulatory agencies around the globe, in particular in the United States and in Europe, have developed new policies and methods to accelerate the approval of certain types of drugs through the use of surrogate endpoints. In the United States, accelerated approval, sometimes referred as "conditional approval" or subpart H, refers to an acceleration of the overall development plan by allowing submission of an application, and if approved, marketing of a drug based on the evidence obtained, for instance, using a surrogate endpoint while further studies demonstrating direct patient benefit are underway. In the same way, the European regulatory agency has developed a set of regulations that are converging to an accelerated approval system like in the United States, perhaps with more flexibility [5].

As the previous sections illustrate, the scientific debate and research on surrogate markers, initiated more than 20 years ago, is still thriving and we believe this work together with the clear regulations established by leading regulatory agencies in the world will arguably allow, in the near future, a more rational and efficient use of this powerful tool.

References

1. Alonso, A., Geys, H., and Molenberghs, G.: A unifying approach for surrogate marker validation based on Prentice's criteria. *Statistics in Medicine* **25**, 205–221 (2006)
2. Alonso, A. and Molenberghs, G.: Surrogate endpoints: Hopes and Perils. *Pharmacoeconomics and Outcomes Research*, **8(3)** 255–259 (2008)
3. Alonso, A. and Molenberghs, G.: Surrogate marker evaluation from an information theoretic perspective. *Biometrics*, **63**, 180–186 (2007)
4. Biomarkers Definition Working Group: Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther*, **69** 89–95 (2001)
5. Burzykowski, T., Molenberghs, G. and Buyse, M. (Eds).: The Evaluation of Surrogate Endpoints. New York: Springer-Verlag (2005)
6. Burzykowski, T., Molenberghs, G., Buyse, M., Renard, D., and Geys, H. Validation of surrogate endpoints in multiple randomized clinical trials with failure-time endpoints. *Applied Statistics* **50**, 405–422 (2001).

7. 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 (2000)
8. Cardiac Arrhythmia Suppression Trial (CAST) Investigators: 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 (1989).
9. Cortiñas Abrahantes, J., Molenberghs, G., Burzykowski, T., Shkedy, Z., and Renard, D.: Choice of units of analysis and modeling strategies in multilevel hierarchical models. *Computational Statistics and Data Analysis*, **47**, 537–563 (2004).
10. Cover, T. and Tomas, J.: *Elements of Information Theory.* New York: Wiley (1991)
11. Daniels, M.J. and Hughes, M.D.: Meta-analysis for the evaluation of potential surrogate markers. *Stat Med* **16**, 1515–1527 (1997)
12. Ellenberg SS and Hamilton JM.: Surrogate endpoints in clinical trials: cancer. *Stat Med* **8**, 405–413 (1989)
13. Frangakis CE, Rubin DB.: Principal stratification in causal inference. *Biometrics* **58**, 21–29 (2002)
14. Fleming TR and DeMets DL.: Surrogate endpoints in clinical trials: are we being misled? *Annals of Internal Medicine* **125**, 605–613 (1996)
15. Ferentz AE.: Integrating pharmacogenomics into drug development. *Pharmacogenomics* **3**, 453–467 (2002)
16. Freedman, L., Graubard, B. and Schatzkin, A.: Statistical validation of intermediate endpoints for chronic diseases. *Stat Med* **11**, 167–178 (1992)
17. Gail, M.H., Pfeiffer, R., van Houwelingen, H.C., and Carroll, R.J.: On meta-analytic assessment of surrogate outcomes. *Biostatistics* **1**, 231–246 (2000)
18. Kullback, S. and Leibler, R.: On information and sufficiency. *Annals of Mathematical Statistics*, **22**, 79–86 (1951)
19. Nyquist, H.: Certain factors affecting telegraph speed. *Bell System Technical Journal*, **3**, 324–346 (1924)
20. Pharmacological Therapy for Macular Degeneration Study Group. Interferon α-IIA is ineffective for patients with choroidal neovascularization secondary to age-related macular degeneration. Results of a prospective randomized placebo-controlled clinical trial. *Archives of Ophthalmology* **115**, 865–872 (1997)
21. Prentice, R.L.: Surrogate endpoints in clinical trials: definitions and operational criteria. *Stat Med* **8**, 431–440 (1989)
22. Renard, Geys, H., Molenberghs, G., Burzykowski, T., Buyse, M., Bijnens, L. and Vangeneugden, T.: Validation of a longitudinally measured surrogate marker for time-to-event endpoint. *Journal of Applied Statistics* **29**, 000–000 (2002)
23. Renard, D., Geys, H., Molenberghs, G., Burzykowski, T., and Buyse, M.: Validation of surrogate endpoints in randomized trials with discrete outcomes. *Biometrical Journal* **44**, 1–15 (2002).
24. Shannon, C.: A mathematical theory of communication. *Bell System Technical Journal* **27**, 379–423 and 623–656 (1948)
25. Tibaldi, F.S, Cortiñas Abrahantes, J., Molenberghs, G., Renard, D., Burzykowski, T., Buyse, M., Parmar, M., Stijnen, T., and Wolfinger, R.: Simplified hierarchical linear models for the evaluation of surrogate endpoints. *Journal of Statistical Computation and Simulation*, **73**, 643–658 (2003).
26. Tilahun, A., Pryseley, A., Alonso, A., and Molenberghs, G.: Flexible Surrogate Marker Evaluation from Several Randomized Clinical Trials with Continuous Endpoints, Using R and SAS. *Computational Statistics and Data Analysis.* **51**, 4152–4163 (2007).
27. Weir, C.J., Walley, R.J.: Statistical evaluation of biomarkers as surrogate endpoints: a literature review. *Stat Med* **25** 183–203 (2006)