Statistical Methods in Medical Research 2004; **13**: 1–30

A perspective on surrogate endpoints in controlled clinical trials

Geert Molenberghs, Tomasz Burzykowski, Ariel Alonso Center for Statistics, Limburgs Universitair Centrum, Diepenbeek, Belgium and **Marc Buyse** International Drug Development Institute (IDDI), Brussels, Belgium

The last couple of decades have seen a large amount of activity in the area of surrogate marker and surrogate endpoint validation, both from a clinical and a statistical perspective. Prentice[1] made a pivotal contribution in the context of a single trial. Subsequently, the framework he proposed has been discussed, criticized, and extended. An important class of extensions considers several rather than a single trial. Recently, a lot of work has been done in this so-called hierarchical or meta-analytic framework. In this paper, we review both the single trial and the hierarchical framework. A number of applications, scattered throughout the literature, are brought together. We outline the statistical issues involved in trying to validate surrogate endpoints. Clearly statistical evidence should only be seen as a component in a decision making process that also involves a number of clinical and biological considerations.

1 Introduction

The use of surrogate endpoints in the development of new therapies has always been very controversial. This may be due to a number of unfortunate historical instances where treatments showing a highly positive effect on 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.[2] For example, in cardiovascular disease, the unsettling discovery that the two major antiarrhythmic drugs encanaide and flecanaide reduced arrhythmia but cause a more than three-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.[3] 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,[4] in spite of serious concerns about their limitations as surrogate markers for clinically relevant endpoints.[5]

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

Address for correspondence: Marc Buyse, IDDI (International Drug Development Institute), 430 avenue Louise B14, B1050 Brussels, Belgium. E-mail: marc.buyse@iddi.com

10.1191/0962280204sm362ra

Techset Composition Ltd, Salisbury Doc: {HODDER}smmr/SM362RA.3d Manu No. 0000
Printed: 19/3/04 Page#: 30 page(s) Opp: Task: 3B2 Version: 7.51f/W (Mar 4 2002)ııTechRef:9.01/H (March 20 1997)

definitions reflect the commonly accepted use of various terms in the biomedical literature:[6]

- 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).

In our examples, we also treat situations in which the potential surrogate is not a biomarker, but an intermediate endpoint that has clinical meaning of its own (e.g., progression-free survival as a potential surrogate for survival).

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.[7] There is also 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.[8] 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.[9] Last but not least, 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.

Thus, while many would like to avoid surrogate endpoints altogether, there are situations where surrogates will be the only reasonable alternative, especially when the true endpoint of interest is rare and/or distant in time. It is then best to use *validated* surrogates, but one clearly needs to reflect on the very meaning of validation.[10] As 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 careful using them as substitutes for the true endpoint in pivotal phase III trials versus replacing the true endpoint by the surrogate altogether in all research past a certain point in time. For a biomarker (or intermediate endpoint) to be used as a 'valid' surrogate for a clinical endpoint, 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 (1) the biological plausibility of the relationship, (2) the demonstration in epidemiological studies of the prognostic value of the surrogate for the clinical outcome and (3) evidence from clinical trials that treatment effects on the surrogate correspond to effects on the clinical outcome'.[11] In this chapter, we discuss statistical methods that are useful to address conditions (1) and (3) from this list. Much of the work laid out here is still in progress. The statistical approach proposed has been developed using data from a range of clinically diverse situations, including age-related macular degeneration,[12–14] cardiovascular disease,[13] advanced ovarian cancer,[14,15] chronic schizophrenia,[16,17] advanced prostate cancer[18,19] and advanced colorectal cancer.[12,15,20,21] It is currently being validated in other situations, including advanced breast cancer, early colorectal cancer, early breast cancer, thrombosis and AIDS.

1.1 Types of endpoint

Statistically speaking, the surrogate endpoint and the clinical endpoint are realizations of random variables. As will be clear from the formalisms developed in Section 3, interest needs to focus on the joint distribution of these variables. The easiest situation is where both are Gaussian random variables. This is, however, seldom the case, because the surrogate endpoint and/or the clinical endpoint are often realizations of non-Gaussian random variables. For example, one can encounter:

- binary (dichotomous): biomarker value below or above a certain threshold (e.g., CD4+ counts over 500/mL) or clinical ‘success’ (e.g., tumor shrinkage);
- categorical (polychotomous): biomarker value falling in successive, ordered classes (e.g., cholesterol levels <200 mg/dL, 200–299 mg/dL, 300+ mg/dL) or clinical response (e.g., complete response, partial response, stable disease, progressive disease);
- continuous (Gaussian): biomarker (e.g., log PSA level) or clinical measurement (e.g., diastolic blood pressure);
- censored continuous: time to biomarker below or above a certain threshold (e.g., time to undetectable viral load) or time to clinical event (e.g., time to cardiovascular death);
- longitudinal or repeated measures: biomarker (e.g., CD4+ counts over time) or clinical outcome (e.g., blood pressure over time);
- multivariate longitudinal: several biomarkers (e.g., CD4+ and viral load over time) or several clinical measurements (e.g., dimensions of quality of life over time).

The models used to validate a surrogate for a clinical endpoint will depend on the type of variables observed in the problem at hand. In the present paper, we will illustrate this through the example of advanced prostate cancer treated with either liarozole or antiandrogens. We will analyse the same data in three different ways. The clinical endpoint will be survival in all cases, but the biomarker will consist, respectively, of PSA response (binary variable), time to PSA progression (censored continuous variable), and the PSA pattern over time (longitudinal).

Table 1 shows some further examples of potential surrogate endpoints in various diseases.

1.2 Units of analysis

Prentice,[1] seeing the need for a formal statistical framework for the validation of surrogate endpoints, provided a definition and a set of criteria to be used when data are available on the effect of some intervention on both a surrogate and a clinical endpoint.

Table 1 Examples of possible surrogate endpoints in various diseases

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 infarction	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

AIDS: acquired immune deficiency syndrome.
ARMD: age-related macular degeneration.
HIV: human immunodeficiency virus.

Prentice's approach formed the basis for much subsequent work showing that compliance to a strict definition of surrogacy would impose almost impossible requirements on potential surrogates. Further research showed that while the idea behind Prentice's approach is appealing, a drawback (common to all single trial approaches) is that it rests on strong and unverifiable assumptions. As argued by several authors,[14,22,23] a way out of this problem is the combination of information from several units or trials. Using hierarchical linear models, Buyse *et al.*[23] defined surrogacy in terms of an individual level as well as a trial level measure of surrogacy, both of which are of a coefficient of determination type. Combining ideas from both frame-works, we propose here a unified approach without obviating the obligation to consider biological and clinical plausibility of a surrogate. Q11

The single unit framework, developed by Prentice[1] and his successors, is reviewed in Section 2. The meta-analytic framework is discussed in Section 3 in the context of Gaussian outcomes, and in Section 4 for non-Gaussian outcomes. The use of different models for different non-Gaussian settings implies a strong need for a unifying set of validation measures. This topic is taken up in Section 5. The ideas developed in earlier sections are illustrated, using prostate cancer data, in Section 6. A number of other examples are reviewed and briefly discussed in Section 7.

2 Data from a single unit

In this section, we will discuss the single-unit setting (e.g., a single trial). The notation and modeling concepts introduced are useful to present and discuss critically the key ingredients of the Prentice–Freedman framework. The next section is devoted to the multitrial setting.

Throughout the paper, we will adopt the following notation: T and S are random variables that denote the true and surrogate endpoints, respectively, and Z is an indicator variable for treatment. For ease of exposition, we will assume that S and T

are normally distributed. The effect of treatment on S and T can be modeled as follows:

$$S_j = \mu_S + \alpha Z_j + \varepsilon_{Sj} \tag{1}$$

$$T_j = \mu_T + \beta Z_j + \varepsilon_{Tj} \tag{2}$$

where $j = 1, \ldots, n$ indicates patients, and the error terms have a joint zero-mean normal distribution with covariance matrix

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

In addition, the relationship between S and T can be described by a regression of the form

$$T_j = \mu + \gamma S_j + \varepsilon_j \tag{4}$$

We will assume later that the n patients come from N different experimental units, but for now the simple situation of a single experiment will suffice to explore some fundamental difficulties with the validation of surrogate endpoints.

2.1 Definition and criteria

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'. In terms of our simple model (1)–(2), the definition states that for S to be a valid surrogate for T, parameters α and β must simultaneously be equal to, or different from, zero. This definition is not consistent with the availability of a single experiment only, since it requires a large number of experiments to be available, each with tests of hypothesis on both the surrogate and true endpoints. An important drawback is also that evidence from trials with nonsignificant treatment effects cannot be used, even though such trials may be consistent with a desirable relationship between both endpoints. Finally, it can be shown that this definition generally requires the true endpoint T to be completely determined by knowledge of the surrogate endpoint S. Several authors, including Prentice, pointed out that this surrogacy requirement was too stringent to be fulfilled in real situations.[1,2]

Prentice derived operational criteria that are equivalent to his definition. These criteria require that

- treatment has a significant impact on the surrogate endpoint (parameter α differs significantly from zero in Equation (1));
- treatment has a significant impact on the true endpoint (parameter β differs significantly from zero in Equation (2));
- the surrogate endpoint has a significant impact on the true endpoint (parameter γ differs significantly from zero in Equation (4)); and
- the full effect of treatment upon the true endpoint is captured by the surrogate.

The last criterion is verified through the conditional distribution of the true endpoint, given treatment *and* surrogate endpoint, derived from Equations (1)–(2):

$$T_j = \tilde{\mu}_T + \beta_S Z_j + \gamma_Z S_j + \tilde{\varepsilon}_{Tj} \tag{5}$$

where the treatment effect (corrected for the surrogate S), β_S, and the surrogate effect (corrected for treatment Z), γ_Z, are

$$\beta_S = \beta - \sigma_{TS}\sigma_{SS}^{-1}\alpha \tag{6}$$

$$\gamma_Z = \sigma_{TS}\sigma_{SS}^{-1} \tag{7}$$

and the variance of $\tilde{\varepsilon}_{Tj}$ is given by

$$\sigma_{TT} - \sigma_{TS}^2\sigma_{SS}^{-1} \tag{8}$$

It is usually claimed that the fourth criterion requires parameter β_S to be equal to zero in Equation (5). Buyse and Molenberghs[12] showed that the last two criteria are necessary and sufficient for binary responses, but not in general. In spite of this drawback, the spirit of the fourth criterion is very appealing, especially if it can be considered in the light of an underlying biological mechanism. For example, it is interesting to explore whether the surrogate is part of the causal chain leading from treatment exposure to the final endpoint. While this issue is beyond the scope of the current paper, the connection between statistical validation (with emphasis on association) and biological relevance (with emphasis on causation) deserves further reflection.

2.2 The proportion explained

Freedman *et al.*[24] argued that the last Prentice criterion raises a conceptual difficulty since it requires the statistical test for treatment effect on the true endpoint to be nonsignificant after adjustment for the surrogate. The nonsignificance of this test does not prove that the effect of treatment upon the true endpoint is *fully* captured by the surrogate, and therefore Freedman *et al.*[24] proposed to calculate the proportion of the treatment effect mediated by the surrogate:

$$PE = \frac{\beta - \beta_S}{\beta}$$

with β_S and β obtained respectively from Equations (5) and (2). In this paradigm, a valid surrogate would be one for which the PE is equal to one. In practice, a surrogate would be deemed acceptable if the lower limit of its confidence interval of PE was sufficiently large.

Some difficulties surrounding the PE have been described in the literature.[12,22,25–27] PE will tend to be unstable when β is close to zero, a situation that is likely to occur in practice. As Freedman *et al.*[24] themselves acknowledged, the confidence limits of PE will tend to be rather wide (and sometimes even unbounded if Fieller confidence intervals are used), unless large sample sizes are available or a very strong effect of treatment on the true endpoint is observed. Note that large sample sizes are typically available in epidemiologic studies or in meta-analyses of clinical trials. Another complication arises when Equation (5) is not the correct conditional model, and an interaction term between Z_i and S_i needs to be included. In that case, defining the *PE* becomes problematic.

2.3 The relative effect

Buyse and Molenberghs[12] suggested to calculate another quantity for the validation of a surrogate endpoint: the RE, which is the ratio of the effects of treatment upon the final and the surrogate endpoint. Formally:

$$RE = \frac{\beta}{\alpha} \tag{9}$$

They also considered the treatment adjusted association between the surrogate and the true endpoint, ρ_Z:

$$\rho_Z = \frac{\sigma_{ST}}{\sqrt{\sigma_{SS}\sigma_{TT}}} \tag{10}$$

Now, a simple relationship can be derived between *PE*, *RE*, and ρ_Z. Let us define $\lambda^2 = \sigma_{TT}\sigma_{SS}^{-1}$. It follows that $\lambda\rho_Z = \sigma_{ST}\sigma_{SS}^{-1}$ and, from Equation (6), $\beta_S = \beta - \rho_Z\lambda\alpha$. As a result, we obtain

$$PE = \lambda\rho_Z\frac{\alpha}{\beta} = \lambda\rho_Z\frac{1}{RE} \tag{11}$$

A similar relationship was derived by Buyse and Molenberghs[12] and by Begg and Leung[28] for standardized surrogate and true endpoints. Molenberghs *et al.*[29] provided a detailed study of the disadvantages of the single unit framework, and measures such as *PE* and *RE*. We will therefore introduce a multiunit framework, opening up new opportunities, whilst maintaining the spirit of Prentice's definition and fourth criterion.

3 Data from several units

Using ideas from Buyse *et al.*,[14] we now extend the setting and notation by supposing we have data from $i = 1, \ldots, N$ units (e.g., centers, investigators, trials), in the ith of which $j = 1, \ldots, n_i$ subjects are enrolled. We now denote the true and surrogate endpoints by T_{ij} and S_{ij}, respectively, and by Z_{ij} the indicator variable for treatment.

The linear models (1) and (2) can be rewritten as:

$$S_{ij} = \mu_{Si} + \alpha_i Z_{ij} + \varepsilon_{Sij} \tag{12}$$

$$T_{ij} = \mu_{Ti} + \beta_i Z_{ij} + \varepsilon_{Tij} \tag{13}$$

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 of zero mean and normally distributed with covariance matrix (3), as before. Due to the replication at the trial level, we can impose a distribution on the trial specific parameters:

$$\begin{pmatrix} \mu_{Si} \\ \mu_{Ti} \\ \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} m_{Si} \\ m_{Ti} \\ a_i \\ b_i \end{pmatrix} \tag{14}$$

where the second term on the right hand side of (14) is assumed to follow a zero-mean normal distribution with covariance matrix

$$D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ & d_{TT} & d_{Ta} & d_{Tb} \\ & & d_{aa} & d_{ab} \\ & & & d_{bb} \end{pmatrix} \tag{15}$$

This setting lends itself naturally to introducing the concept of surrogacy at both the trial level as well as the individual level. We discuss them in turn.

3.1 Trial-level surrogacy

As indicated previously, the key motivation for validating a surrogate endpoint is to be able to predict the effect of treatment on the true endpoint based on the observed effect of treatment on the surrogate endpoint *at the trial level*. It is essential, therefore, to explore the quality of the prediction of the treatment effect on the true endpoint in trial i by (1) information obtained in the validation process based on trials $i = 1, \dots, N$ and (2) the estimate of the effect of Z on S in a new trial $i = 0$. Fitting models (12)–(13) to data from a meta-analysis provides estimates for the parameters and the variance components. Suppose then the new trial $i = 0$ is considered for which data are available on the surrogate endpoint but not on the true endpoint. We then fit the following linear model to the surrogate outcomes S_{0j}:

$$S_{0j} = \mu_{S0} + \alpha_0 Z_{0j} + \varepsilon_{S0j} \tag{16}$$

Estimates for m_{S0} and a_0 are

$$\hat{m}_{S0} = \hat{\mu}_{S0} - \hat{\mu}_S \tag{17}$$

$$\hat{a}_0 = \hat{\alpha}_0 - \hat{\alpha} \tag{18}$$

Note that such an approach is closely related to leave-one-out regression diagnostics.[30,31]

We are interested in the estimated effect of Z on T, given the effect of Z on S. To this end, observe that $(\beta + b_0 | m_{S0}, a_0)$ follows a normal distribution with mean and variance:

$$E(\beta + b_0 | m_{S0}, a_0) = \beta + \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} \mu_{S0} - \mu_S \\ \alpha_0 - \alpha \end{pmatrix} \tag{19}$$

$$\mathrm{Var}(\beta + b_0 | m_{S0}, a_0) = d_{bb} - \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix} \tag{20}$$

In practice, these equations can be used as follows. Using Equations (17) and (18), a prediction can be made using Equation (19), with prediction variance Equation (20). Of course, one has to properly acknowledge the uncertainty resulting from the fact that the parameters in Equations (17)–(18) are not known but merely estimated. This follows from a straightforward application of the iterated expectation law.

A surrogate could thus be called *perfect at the trial level* if the conditional variance Equation (20) were equal to zero. A measure to assess the quality of the surrogate at the trial level is the coefficient of determination

$$R^2_{\text{trial}} = R^2_{b_i | m_{Si}, a_i} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}} \tag{21}$$

Similar to the logic in Equations (19) and (20), the conditional model for β_i given μ_{Si} and α_i can be written:

$$\beta_i = \theta_0 + \theta_a \alpha_i + \theta_m \mu_{Si} + \varepsilon_i \tag{22}$$

where expressions for the coefficient $(\theta_0, \theta_a, \theta_m)$ follow from Equations (14) and (15). In case the surrogate is perfect at the trial level ($R^2_{\text{trial}} = 1$), the error term in Equation (22) vanishes and the linear relationship becomes deterministic, implying that β_i *equals* the systematic component of Equation (22).

This approach avoids 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. Even if the posited linear relationships do not hold, it is possible to consider alternative regression functions, although one has to be aware of a potentially low power to discriminate between candidate regression functions. By virtue of replication, it is possible to *check* the stated relationships for the treatment effects. Moreover, the use of a measure of association to assess surrogacy is more in line with the adjusted association suggested in the single trial case.

A key issue when using the proposed meta-analytic framework, and in particular its prediction facility Equation (19), is the coding of the treatment indicators Z_{ij}. While the framework is invariant to coding reversal of *all* treatment indicators at the same time, more caution is needed when the coding of a single trial is considered, such as in Equation (16). In such a case, invariance is obtained only when the fixed effects in Equations (12) and (13) are equal to zero. This issue is intimately linked to the question as to how broad the class of units to be included in a validation study can be. Clearly, the issue disappears when the same or similar treatments are considered across units (e.g., in multicenter or multi-investigator studies, or when data are used from a family of related study such as in a single drug development line). In a more loosely connected, meta-analytic setting it is important to ensure that treatment assignments are logically consistent. This is possible, for example, when the same standard treatment is compared to members of a class of experimental therapies.

Next, we will show that the adjusted association carries over naturally to the multiunit setting as well.

3.2 Individual level surrogacy

We now return to the association between the surrogate and the final endpoints after adjustment for treatment. As described earlier, we need to construct the conditional distribution of T, given S and Z. From Equations (12)–(13) we derive

$$T_{ij}|Z_{ij}, S_{ij} \sim N\left\{\mu_{Ti} - \sigma_{TS}\sigma_{SS}^{-1}\mu_{Si} + (\beta_i - \sigma_{TS}\sigma_{SS}^{-1}\alpha_i)Z_{ij} + \sigma_{TS}\sigma_{SS}^{-1}S_{ij};\ \sigma_{TT} - \sigma_{TS}^{2}\sigma_{SS}^{-1}\right\} \quad (23)$$

Q2

which is an extension of Equation (5). Note that

$$\beta_{Si} = \beta_i - \sigma_{TS}\sigma_{SS}^{-1}\alpha_i \quad (24)$$

The association between both endpoints after adjustment for the treatment effect is captured by

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

the squared correlation between S and T after adjustment for both the trial effects and the treatment effect. R^2_{indiv} generalizes ρ^2_Z as described earlier by adjusting the association both for treatment and for trial. We call a surrogate *perfect at the individual level* if $R^2_{\text{indiv}} = \rho^2_Z = 1$.

Taken together, the R^2 measures allow one to quantify the properties of a candidate surrogate endpoint. In addition, by using a hierarchical model such as Equations (12)–(15), measurement error in the surrogate is automatically taken into account. When a two stage approximation is used (i.e., fitting a separate model to each unit in the first stage and fitting a regression on the resulting treatment effect parameters in the second stage), this is no longer true. Burzykowski *et al.*[15] illustrate how measurement error can be incorporated in such a context.

4 Non-Gaussian endpoints

In this section, 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.*[16] have shown that extension to this situation is easily done using a latent variable formulation. That is, we posit the existence of a pair of continuously distributed latent variable responses $(\tilde{S}_{ij}, \tilde{T}_{ij})$ 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 Equations (12)–(13) 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.*[16] used pseudo-likelihood methods to estimate the model parameters.

4.2 Two failure time endpoints

Assume now that S_{ij} and T_{ij} are failure time endpoints. Model (12)–(13) is replaced by a model for two correlated failure time random variables. Burzykowski[15] used copulas to this end.[32,33] Precisely, one assumes the joint survivor function of (S_{ij}, T_{ij}) is written as:

$$F(s, t) = P(S_{ij} \geq s, T_{ij} \geq t) = C_\delta\{F_{Sij}(s), F_{Tij}(t)\}, \quad s, t \geq 0 \tag{25}$$

where (F_{Sij}, F_{Tij}) denote marginal survivor functions and C_δ is a distribution function on $[0, 1]^2$ with $\delta \in R^1$. C_δ is called a *copula function*.[34]

When the hazard functions are specified, estimates of the parameters for the joint model can be obtained using maximum likelihood. Shih and Louis[35] discuss alternative estimation methods. The association parameter is hard to interpret. However, it can be shown[34] there is a link with Kendall's τ: Q11

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

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

We will now 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.*[21] also used bivariate copulas, thereby combining ideas of Geys[36] and Burzykowski *et al.*[15] One marginal distribution is a proportional odds logistic regression, while the other is a proportional hazards model. The Plackett copula[37] was chosen to capture the association between both endpoints. The advantage of this choice is that the association is expressed as a global odds ratio, which is relatively easy to interpret.

4.4 Longitudinal endpoints

Repeated measurements are often encountered on either or both endpoints. However it is clear that methods taking into account the longitudinal structure of the data will yield much more complex statistical modelling strategies and will require further extensions in the surrogate marker evaluation methodology. In analogy to the cross-sectional setting considered by Buyse *et al.*,[14] we will base the calculation of surrogacy measures on a two stage approach rather than a full random effects approach, to reduce numerical complexity.

In their work Alonso *et al.*[17] 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_{ij}. If T_{ijt} and S_{ijt} denote the associated true and surrogate endpoints, respectively, and Z_{ij} is a binary indicator variable for treatment then along the ideas of Galecki,[38] they propose the following joint model, at the first stage, for both responses

$$\begin{cases} T_{ijt} = \mu_{T_i} + \beta_i Z_{ij} + g_{T_i}(t_{ij}) + \varepsilon_{T_{ijt}} \\ S_{ijt} = \mu_{S_i} + \alpha_i Z_{ij} + g_{S_i}(t_{ij}) + \varepsilon_{S_{ijt}} \end{cases} \tag{26}$$

where μ_{S_i} and μ_{T_i} are trial specific intercepts, α_i, β_i are trial specific effects of treatment Z_{ij} on the two endpoints and g_{T_i} and g_{S_i} are trial specific time functions. Note that, even though in practice T_{ij} and S_{ij} are frequently measured at the same time points, model (26) would let us approach situations in which this condition does not hold. The vectors $\tilde{\varepsilon}_{T_{ij}}$ and $\tilde{\varepsilon}_{S_{ij}}$ are correlated error terms, assumed to be jointly mean-zero multivariate normally distributed with covariance matrix

$$\Sigma_i = \begin{pmatrix} \Sigma_{TTi} & \Sigma_{TSi} \\ \Sigma'_{TSi} & \Sigma_{SSi} \end{pmatrix} = \begin{pmatrix} \sigma_{TTi} & \sigma_{TSi} \\ \sigma_{TSi} & \sigma_{SSi} \end{pmatrix} \otimes R_i \tag{27}$$

In the aforementioned formulation, R_i reflects a general correlation matrix for the repeated measurements of the responses. A frequent choice in practice would be the first order autoregressive structure (in case measures are equally spaced, otherwise a spatial type structure is better).

Even though Buyse *et al.*[14] assumed, in the special case of a single measurement per response, that the error covariance structure was constant over all trials, this assumption is no longer tenable for most longitudinal settings. Measures could be taken at different time points within different trials, the number of measurements could be different in each trial, etc. Therefore, the covariance structure should be allowed to vary over trials. If treatment effect can be assumed constant over time then the R^2_{trial} measured proposed by Buyse *et al.*[14] can still be useful to evaluate surrogacy at the trial level. However at the individual level the situation is totally different, there the R^2_{ind} is no longer applicable and new concepts are needed.

Using multivariate ideas, Alonso *et al.*[17] proposed the *variance reduction factor* (*VRF*). Essentially, they summarized the variability of the repeated measurements on the true endpoint within trial by the trace of its variance covariance matrix and sum this over all trials. In a similar way they summarized the conditional variability of the true endpoint measurements, given the surrogate by the trace of the conditional variance covariance matrix, summed over trials and they quantified the relative reduction in the true endpoint variance after adjustment by the surrogate as

$$VRF_{\text{ind}} = \frac{\sum_i \{\text{tr}(\Sigma_{TTi}) - \text{tr}(\Sigma_{(T|S)i})\}}{\sum_i \text{tr}(\Sigma_{TTi})} \tag{28}$$

where $\Sigma_{(T|S)_i}$ denotes the conditional variance of $\tilde{\varepsilon}_{T_{ij}}$ given $\tilde{\varepsilon}_{S_{ij}}$: $\Sigma_{(T|S)i} = \Sigma_{TTi} - \Sigma_{TSi}\Sigma_{SSi}^{-1}\Sigma'_{TSi}$, 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. An advantage of the proposed expression is that data from trials with unequal length vectors of repeated measures can easily be handled. Intuitively, Equation (28) tries to quantify how much of the total variability around the repeated measurements on the true endpoint is explained by adjusting for the treatment effects Z_{ij} and the repeated measurements on the surrogate endpoint.

Further, they proved that the VRF_{ind} ranges between zero and one, that it equals zero if and only if the error terms of the true and surrogate endpoints are independent within each trial, that the VRF_{ind} equals one if and only if there exists a deterministic relationship between the error terms of the true and surrogate endpoints within each trial and finally they proved that the VRF_{ind} reduces to the R^2_{ind} when the endpoints are measured only once.

In addition, they showed that the VRF can be incorporated into a much more general framework that allows interpretation in terms of the canonical correlations of the error vectors. Indeed, if at trial i we have p_i time points then we will also have $t = 1, \ldots, p_i$ canonical correlations ρ^2_{it} for $(\tilde{\varepsilon}_{T_{ij}}, \tilde{\varepsilon}_{S_{ij}})$ such that $\rho^2_{i1} \geq \rho^2_{i2} \geq \ldots \geq \rho^2_{ip_i}$ and ρ^2_{it} are the Q3
eigenvalues of $\Sigma_{TTi}^{-1/2}\Sigma_{TSi}\Sigma_{SSi}^{-1}\Sigma_{TSi}^{T}\Sigma_{TTi}^{-1/2}$. Based on these canonical correlations they defined a family of parameters to study surrogacy at the individual level. The family was conceived in such a way that the properties, based on which R^2_{ind} was considered a useful measure of surrogacy by Buyse *et al.*,[14] are preserved:

$$\Omega = \left\{\theta\colon \theta = \sum_i \sum_k \alpha_{ik}\rho^2_{ki}, \quad \text{where:} \quad \alpha_{ik} > 0 \quad \forall(i, k), \quad \sum_i \sum_k \alpha_{ik} = 1\right\}$$

Here, $i = 1, \dots, N$ denotes the trial and $k = 1, \dots, p_i$ denotes the designed time points. They also proved that previous definitions like the VRF or the R^2 measurements are special members of Ω.

Given that Ω can be seen as a *family* of measures to study individual level surrogacy, they evaluated the operational characteristics of some of its members, including the VRF, that one might want to consider in practice using extensive simulations. Based on the results of the simulations as well as on interpretational and mathematical arguments they suggested that a very plausible choice in several practical situations could be θ_p defined as

$$\theta_p = \sum_i \frac{1}{Np_i} \mathrm{tr}\{(\Sigma_{TTi} - \Sigma_{(T|S)i})\Sigma_{TTi}^{-1}\} \tag{29}$$

It is important to notice that structurally, both *VRF* and θ_p are similar, the difference being the reversal of summing the trace and calculating the ratio. Moreover, θ_p has the appealing property of coinciding with Pillai's trace statistic, well-known from classical multivariate analysis. In spite of this strong structural similarity, these parameters have fundamental differences. First, the VRF is not symmetric in S and T. Second, it is only invariant with respect to linear orthogonal transformations. In contrast, θ_p is both symmetric and invariant with respect to the broader class of linear bijective transformations. Q4 Finally and based on all these considerations these authors suggested that θ_p seems to be a more preferable choice in the analysis of real problems.

One serious drawback of the previous approach is that it is strongly based on the normality assumption and an extension to nonnormal settings seems to be difficult. In the present work we consider an alternative methodology that offers some practical and conceptual advantages with respect to the previous one and it also allows a straightforward extension to nonnormal settings.

We propose a new parameter, 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

$$R^2_\Lambda = \frac{1}{N} \sum_i (1 - \Lambda_i) \tag{30}$$

where $\Lambda_i = |\Sigma_i| / |\Sigma_{TTi}||\Sigma_{SSi}|$.

First one should notice that R^2_Λ is defined based on the Wilks' Lambda statistic used in multivariate analysis and it involves the determinants of the variance covariance matrices. Therefore all the elements of the covariance structure are used when calculating Equation (30). On the other hand Equations (28) and (29) only use the information in the diagonal of the matrices that defined association between both endpoints what can make them less informative.

It is possible to prove that R^2_Λ is symmetric and invariant with respect to linear bijective transformations, R^2_Λ ranges between zero and one, $R^2_\Lambda = 0$ if and only if, for all Q5 i, $(\tilde{\varepsilon}_{T_i}, \tilde{\varepsilon}_{S_i})$ are independent, $R^2_\Lambda = 1$ if and only if, for all i, there exist vectors a_i and b_i such that $a_i'\tilde{\varepsilon}_{T_i} = b_i'\tilde{\varepsilon}_{S_i}$ with probability one and finally in the cross-sectional case $R^2_\Lambda = R^2_{\mathrm{ind}}$.

Essentially, these are the same properties satisfied by the VRF, θ_p and all the members of the Ω family. However the fourth property makes an important difference between the new proposal and the previous ones. Whereas the elements of Ω take the value 1 only when there is a deterministic relationship between both endpoints, R_Λ^2 is 1 whenever there is a deterministic relationship between two linear combinations of both endpoints letting us detect strong associations in cases in which the VRF or θ_p would fail to do so.

Here again, using canonical correlation ideas, it is possible to define a whole family of parameters to study surrogacy at the individual level so that R_Λ^2 is just a special member of that family

$$\Omega_\Lambda = \left\{ \theta_\Lambda \colon \theta_\Lambda = 1 - \sum_{i=1}^{N} \alpha_i \prod_{k=1}^{p_i} (1 - \rho_{ik}^2), \quad \text{where:} \quad \alpha_i > 0 \quad \forall i, \quad \sum_i \alpha_i = 1 \right\}$$

Its use towards unification of the various proposals that have been made for the various settings, will be discussed in the next section.

5 A unified theory

While the meta-analytic framework clearly has advantages, there are some downsides as well. First, the modelling exercise increases in complication, since the need arises for a joint, hierarchical model for the surrogate and true endpoints. In addition, a different model is needed depending on the type of outcome. As a consequence, while the trial level surrogacy is typically expressed by means of a R^2 measure, the individual level surrogacy is expressed by a model specific quantity. This calls for further unification.

5.1 Relationship between $\mathbf{R}_\Lambda^2$ and θ_P

In Section 4.4 two 'parallel' approaches have been described to evaluate surrogacy at the individual level in a repeated measurement framework. Even though each definition for one of these proposals seems to be translatable into a similar definition for the other one, no clear connection between these two approaches has been made.

Let us first consider the special case defined by Galecki's model. Under this model the variance covariance matrix of the error vectors is 'decomposed' in two basic components describing the association between sequences respectively and the association within sequences and these two components are pulled together using the Kronecker product. It is easy to show that under this assumption of separability for the covariance structure

$$\theta_P = \frac{1}{N} \sum_i \rho_{TSi}^2 \quad \text{and} \quad R_\Lambda^2 = 1 - \frac{1}{N} \sum_i (1 - \rho_{TSi}^2)^{p_i}$$

where $\rho_{TSi}^2 = \sigma_{TS}/\sigma_{TT}\sigma_{SS}$. Taking into account that

$$(1 - \rho_{TSi}^2)^{p_i} = (1 - \rho_{TSi}^2) + (1 - \rho_{TSi}^2)\left\{(1 - \rho_{TSi}^2)^{p_i - 1} - 1\right\}$$

we finally obtain

$$R_\Lambda^2 = \theta_P + \frac{1}{N}\sum_i (1-\rho_{TSi}^2)\left\{1-(1-\rho_{TSi}^2)^{p_i-1}\right\} \tag{31}$$

Formula (31) clearly shows that θ_P can be seen as an approximation for R_Λ^2 when the second part of the sum is negligible.

If we also take into account that

$$\frac{1}{N}\sum_i (1-\rho_{TSi}^2)\left\{1-(1-\rho_{TSi}^2)^{p_i-1}\right\} \geq 0$$

then we have that

$$\theta_P \leq R_\Lambda^2 \tag{32}$$

The equality is obtained for some special interesting cases:

- $p_i = 1$, for all i, univariate setting and both proposals reduce to the R_{ind}^2;
- $\rho_{TSi}^2 = 0$, for all i if and only if $(\tilde{\varepsilon}_{T_i}, \tilde{\varepsilon}_{S_i})$ are independent and $R_\Lambda^2 = \theta_p = 0$;
- $\rho_{TSi}^2 = 1$ for all i if and only if there is a deterministic relationship between $\tilde{\varepsilon}_{T_i}$ and $\tilde{\varepsilon}_{S_i}$; in this case $R_\Lambda^2 = \theta_p = 1$.

If we now switch to a totally general framework where the separability assumption does not necessarily hold, then it is easy to see that, for all $\theta_\Lambda \in \Omega_\Lambda$ we have

$$\begin{aligned}\theta_\Lambda &= \sum_{i=1}^{N}\sum_{h=1}^{p_i}\frac{\alpha_i}{p_i}\rho_{ih}^2 + \sum_{i=1}^{N}\sum_{k=1}^{p_i}\frac{\alpha_i}{p_i}(1-\rho_{ik}^2)\left(1-\prod_{h\neq k}(1-\rho_{ih}^2)\right)\\ &= \theta + \sum_{i=1}^{N}\sum_{k=1}^{p_i}\frac{\alpha_i}{p_i}(1-\rho_{ik}^2)\left(1-\prod_{h\neq k}(1-\rho_{ih}^2)\right)\end{aligned} \tag{33}$$

Equation (33) shows that for all $\theta_\Lambda \in \Omega_\Lambda$, there exists a $\theta \in \Omega$ so that θ_p can be considered an approximation of θ_Λ if the last term in the sum (33) is negligible.

Here, exactly as before,

$$\sum_{i=1}^{N}\sum_{k=1}^{p_i}\frac{\alpha_i}{p_i}(1-\rho_{ik}^2)\left(1-\prod_{h\neq k}(1-\rho_{ih}^2)\right) \geq 0$$

Therefore, the previous expression also shows that for all $\theta_\Lambda \in \Omega_\Lambda$, there exists a $\theta \in \Omega$ so that

$$\theta \leq \theta_\Lambda \tag{34}$$

leading to the conclusion that Equation (32) holds in a totally general setting and not only under the assumptions of separability defined by Galecki's model.

5.2 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 very clear how we can quantify the association between both endpoints after adjusting for treatment and trial effect. Hence, a number of different parameters have been proposed. This, of course, calls for further unification.

Here, we offer a general procedure that will allow us to evaluate surrogacy at the individual level in very general settings. To develop this idea, let us assume that for the data from trial *i* the following generalized linear models hold

$$g_T(T_{ij}) = \mu_{T_i} + \beta_i Z_{ij}, \tag{35}$$

$$g_T(T_{ij}) = \theta_{0_i} + \theta_{1i} Z_{ij} + \theta_{2i} S_{ij} \tag{36}$$

If necessary, we can assume that repeated measurements on the same patient have been taken. For instance, in a longitudinal study, different functions of time can be included in Equations (35) and (36). In general, other more complex settings could be analysed in a very similar way using the methodology that will be described below. We could even construct a model that takes into account a nonlinear relationship between *S* and *T*, for instance,

$$g_T(T_{ij}) = \theta_{0_i} + \theta_{1i} Z_{ij} + f(S_{ij})$$

Note that Equation (36) is just the model proposed by Prentice in his fourth criterion. In the present approach, it comes back to play a key role in the definition of a unifying procedure to quantify surrogacy at the individual level. One of the most appealing characteristics of this proposal is that we can avoid the joint fitting of complicated models for the surrogate and the true endpoint. In general, models like Equation (35) and (36) can usually be fitted using standard commercial software.

To this end, we can consider G_i^2 as the log-likelihood ratio test statistics to compare Equation (35) and (36) in trial *i*. We propose to quantify the association between both endpoints at the individual level using a scale likelihood reduction factor (LRF)

$$\mathrm{LRF} = 1 - \frac{1}{N} \sum_i \exp\left(- \frac{G_i^2}{n_i} \right) \tag{37}$$

Following the ideas of Kent,[39] we can think of Equation (37) as a sample estimate of a general measure of association between both endpoints, based on the information gain about the true endpoint by using the surrogate, it is also possible to see that: 1) the LRF always lies between 0 and 1; 2) the LRF is zero if the surrogate and the true endpoint are independent in each trial; 3) as LRF approaches 1 for Gaussian outcomes, Q6 there is usually some degeneracy appearing in the true joint distribution of S and T in each trial; often $\phi_{1i}(S) = \phi_{2i}(T)$ implying a deterministic relationship between both variables, in the longitudinal-longitudinal case LRF becomes the R^2_{Λ} defined earlier and finally in the cross-sectional normal-normal case, the LRF reduces to the R^2_{ind}.

6 Prostate-specific antigen in advanced prostate cancer

6.1 Two liarozole trials

We illustrate the statistical approach based on the individual level and trial level associations using two trials in patients with advanced (metastatic) prostate cancer. These trials compared oral liarozole, an experimental retinoic acid metabolism blocking agent developed by the Janssen Research Foundation, with two antiandrogenic drugs: cyproterone acetate (CPA) in the first trial and flutamide in the second. In both trials, patients were in relapse after first-line endocrine therapy.[40] The trials accrued 312 and 284 patients, respectively. Each trial was multinational and multicentric. Since our analyses require the estimation of the effect of treatment in multiple trials or other meaningful groups of patients, we grouped the patients by trial and by country. This allowed us to define 19 groups containing between four and 69 patients per group.

The primary endpoint of the trials was overall survival from the start of treatment. Assessments were undertaken before the start of treatment, at two weeks, monthly for six months, at three-month intervals until the second year, and at six-month intervals until treatment discontinuation or death. The assessments included measurement of the prostate specific antigen (PSA) level. PSA is a glycoprotein that is found almost exclusively in normal and neoplastic prostate cells. Changes in PSA often antedate changes in bone scan, and they have been used as an indicator of response in patients with androgen independent prostate cancer.[41–43] Figure 1 shows the structure of the validation problem.

We consider successively, PSA response, time to PSA progression (TPP), and the full longitudinal PSA profile of each patient as potential surrogates for survival in this disease.[44]

6.2 PSA response as surrogate for survival

The best PSA outcome was determined for each patient, and hierarchically ordered as:[45]

- complete response (CR) if the PSA level was at least 20 ng/mL at baseline, returned to normal (<4 ng/mL) at any time, and remained normal for at least 28 days;
- partial response (PR) if the PSA level was at least 20 ng/mL at baseline, decreased by at least 50% from the baseline level, and remained under 50% of the baseline level for at least 28 days;
- no change (NC) if the PSA level was at least 20 ng/mL at baseline, and fluctuated between 50% below and 50% above the baseline level for at least 28 days;
- progressive disease (PD) if no other response category applied, and if PSA was at least equal to 10 ng/mL
- not evaluable (NE), if none of the above applied.

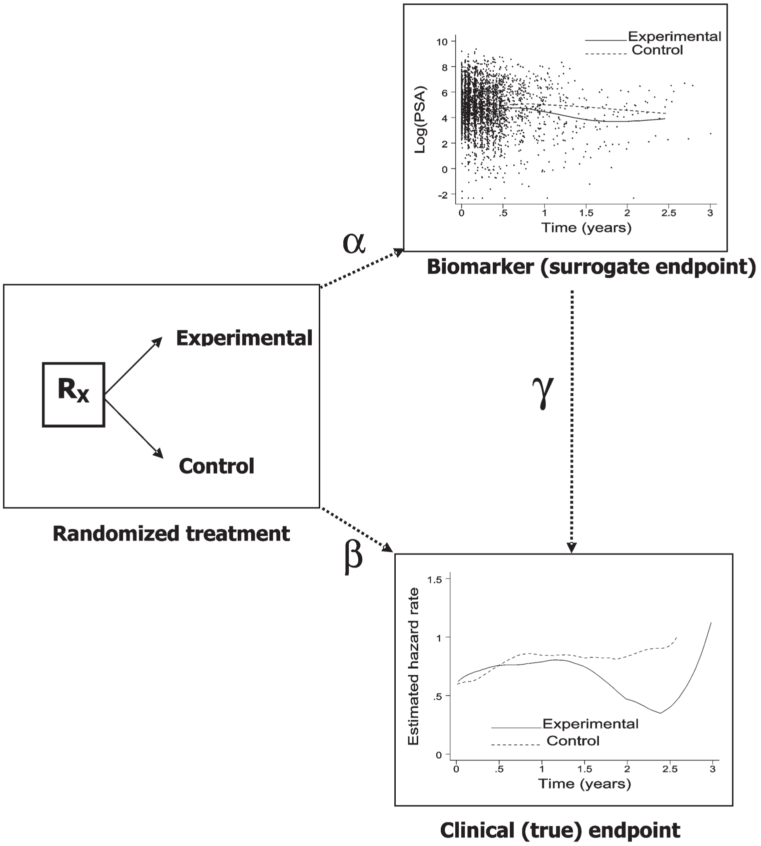

Figure 1 Validation of a biomarker (or intermediate endpoint) as a surrogate for a clinical endpoint (or true endpoint) with respect to the effect of a randomized treatment. In advanced prostate cancer, the biomarker could be the level of PSA over time, and the clinical endpoint could be the time to death. The figure shows individual PSA values and their mean over time by treatment group, and the hazard rate over time by treatment group.

A patient was defined as having a PSA response if his best PSA outcome was either PR or CR. Hence the biomarker is binary here, and the clinical endpoint is a (possibly censored) survival time.

At the individual level, PSA response was a very strong predictor of survival (Figure 2a). Because PSA response is binary and survival is censored, the normal theory coefficient of determination (R^2) discussed in Section 3 does not apply, and another measure of association between PSA response and survival is needed. One way to express the impact of PSA response on survival is as follows (8): consider the odds of Q7 surviving beyond time t for PSA responders and for nonresponders; the ratio of these odds is a survival odds ratio. Although the odds of surviving beyond time t decrease in time for both responders and nonresponders, in our model the ratio of these odds is assumed constant. This survival odds ratio is equal to 5.5 (95% confidence interval [2.7,8.2]), which means that at any point in time the odds of surviving beyond that time

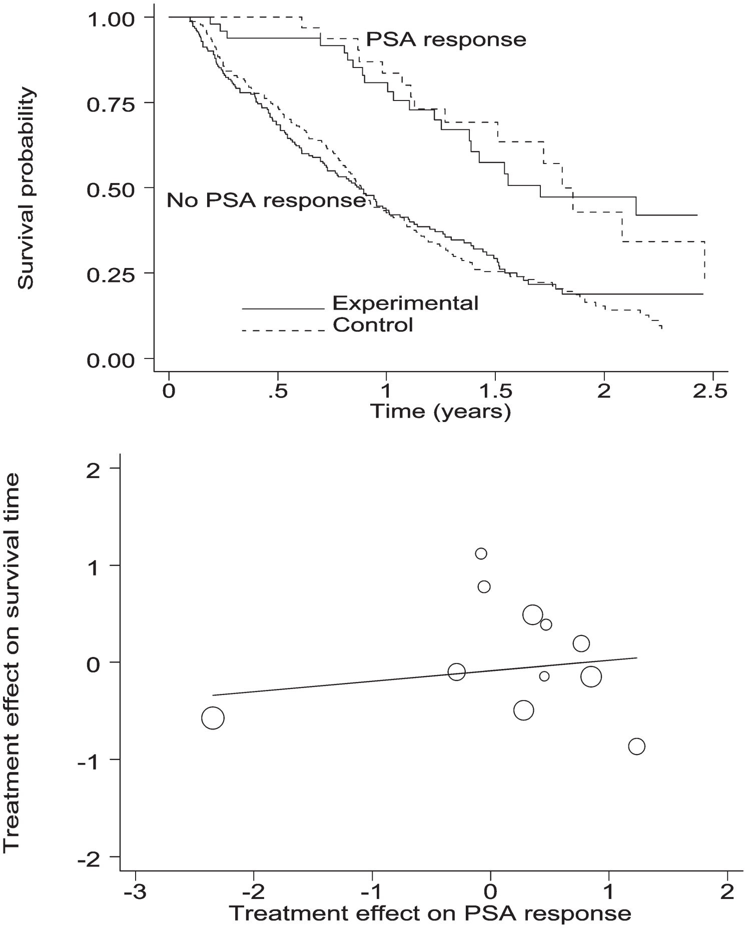

Figure 2 (a) The survival of patients with a PSA response differs substantially from that of patients without a PSA response. At any point in time the odds of surviving beyond that time are more than five times higher for patients with a PSA response as compared to patients without such a response. (b) The treatment effects on survival and on PSA response show no correlation in advanced prostate cancer ($R^2_{\text{trial}} = 0.05$).

are more than five times higher for patients with a PSA response as compared to patients without such a response. The strong prognostic impact of PSA response can be explained in at least three plausible ways:

- PSA response and survival are largely determined by a common set of prognostic factors, so that patients who are likely to have a response are also those who are potentially long survivors.
- Patients who survive a long time are more likely to have a PSA response because of length biased sampling.[46]
- There is a true causal relationship between the achievement of a PSA response and a prolongation of survival.

The first and second explanation are amenable, at least in part, to statistical investigations, the first through adjustments of the comparison of responders and nonresponders for all known prognostic factors, and the second through a landmark analysis.[47] When these investigations fail to explain a large portion of the prognostic impact of PSA response, then there is indirect evidence that PSA response truly results in a survival improvement.[20]

At the group level, the effects of liarozole on PSA response and on survival were poorly correlated, with a coefficient of determination $R^2_{\text{trial}} = 0.05$ (standard error $= 0.13$) (Figure 2b).

There was no overall significant benefit of liarozole over control for either PSA response or survival: the PSA response rate was 16% and 11%, respectively, for liarozole and control ($P = 0.11$), while median survival was 11.3 and 10.9 months, respectively, for liarozole and control ($P = 0.71$).

6.3 Time to PSA progression as surrogate for survival

The time to PSA progression (TPP) was determined on the basis of a moving average of three consecutive values of PSA. Progression was defined as an increase in PSA equal to, or larger than, 50 above the lowest prior moving average. This increase had to be either the last determination in the patient's follow-up, or had to be maintained for at least 28 days.

At the individual level, PSA progression occurred much earlier than the patients' death. PSA progression occurred within six months for half of the patients, while about half of the patients were still alive at one year (Figure 3a). Here again, because TPP and survival may both be censored, the normal theory coefficient of determination (R^2) discussed in Section 3 does not apply, and a possible measure of association between TPP and survival is a generalization of that proposed above:[15] consider the odds of surviving beyond time t for patients who have not yet had a PSA progression, and for those who have; the ratio of these odds is a survival odds ratio. Although the odds of surviving beyond time t decrease in time for both patients with and without PSA progression, in our model the ratio of these odds is assumed constant.

This odds ratio is equal to 6.3 (95% confidence interval [4.4,8.2]), which means that at any point in time the odds of surviving beyond that time are more than six times higher for patients who have not yet had a PSA progression as compared to patients who have already had such a progression. Thus, here again, there is a strong individual level association between TPP and survival.

At the group level, the effects of liarozole on TPP and on survival were poorly correlated, with a coefficient of determination $R^2_{\text{trial}} = 0.22$ (standard error 0.18) (Figure 3b). There was a significant benefit of liarozole over control in terms of time to PSA progression, with a median time of 4.9 months for liarozole and 3.7 months for control ($P = 0.001$).

6.4 Longitudinal measurements of PSA as surrogate for survival

Since PSA levels were measured repeatedly over time, it seems natural to make use of all these measurements, rather than to define a single PSA response or time to PSA progression for each patient. The statistical models required to take the longitudinal nature of the measurements into account are more complex, and the analyses

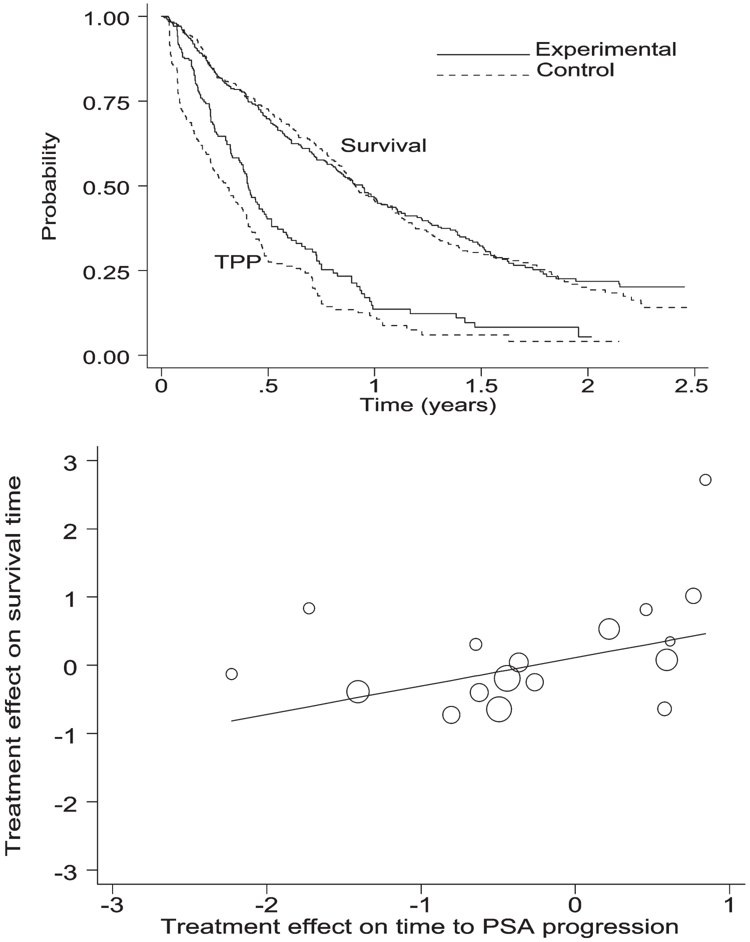

Figure 3 (a) PSA progression is a strong predictor of death in advanced prostate cancer. At any point in time the odds of surviving beyond that time are more than six times higher for patients who have not yet had a PSA progression as compared to patients who have already had such a progression. (b) The treatment effects on survival and on time to PSA progression show very little correlation in advanced prostate cancer ($R^2_{\text{trial}} = 0.22$).

potentially more sensitive to model assumptions, than for singly measured endpoints. Such models have been used extensively to study the relationship between CD4 lymphocytes and survival in patients with AIDS and AIDS related complex.[4,48–51]

In our example, the mean PSA levels over time shown in the upper right hand panel of Figure 1 are not fully informative, because these means were not calculated on the same patients over time. Indeed, patients who had a PSA progression left the study, and no longer contributed to the mean PSA after that time point, thus creating a selection bias in the calculation of the mean. A more informative way of looking at mean PSA levels over time is to consider cohorts of patients defined by the time they leave the study (for any reason). Figure 4a shows four such cohorts, split by treatment group: patients leaving the study within six months, between six and 12 months, between 12

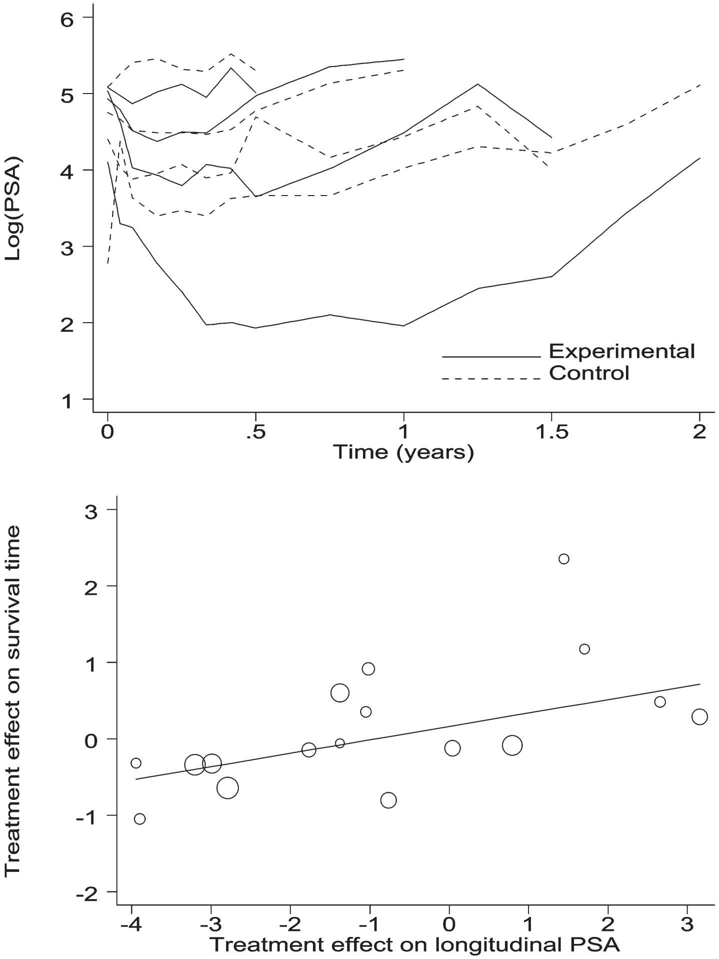

Figure 4 (a) The mean PSA profiles for cohorts of patients with similar follow-up times show a tendency for PSA to go down initially (PSA response), and to come up again after a while (PSA progression). The longitudinal PSA profiles are strongly correlated with the hazard of death ($R^2_{\text{indiv}} < 0.84$ at any point in time). (b) The treatment effects on survival and on longitudinal PSA show a weak correlation in advanced prostate cancer ($R^2_{\text{trial}} = 0.42$).

and 18 months, and between 18 and 24 months (PSA data became too scarce to calculate meaningful means after 24 months). The patterns exhibited by these cohort specific means show a tendency for PSA to go down initially (PSA response), and to come up again after a while (PSA progression).

At the individual level, the PSA longitudinal process was correlated with the hazard rate, which is the risk of dying at a certain time for a patient who has survived up until that time. The coefficient of determination between the PSA process and the hazard rate (R^2_{indiv}) here is a function of time that cannot be easily summarized into a single measure.[16] Suffice to say that R^2_{indiv} was greater than 0.84 at all times to indicate that there was again a strong association, at the individual patient level, between the evolution of PSA and the hazard of dying.

At the group level, the effects of liarozole on longitudinal PSA and on survival were moderately correlated, with a coefficient of determination $R^2_{\text{trial}} = 0.45$ (standard error 0.18) (Figure 4b). There was a significant benefit of liarozole in terms of longitudinal PSA ($P = 0.01$); in other words, the profiles shown on Figure 4a were significantly different between liarozole and control.

7 Further examples

We have illustrated, through an actual example, statistical approaches that may be useful to study the complex relationships between a biomarker, a clinical endpoint, and the effects of a treatment on both the biomarker and the clinical endpoint. Our analyses emphasize the importance of distinguishing between two types of association: one between the biomarker and the clinical endpoint at the individual level, the other between the effects of treatment on the biomarker and on the clinical endpoint at the trial, or some other group, level.

For instance, in the prostate cancer example outlined above, only two trials were available for analysis, and thus we considered country in each trial as the grouping unit of interest. Table 2 summarizes the clinical situation in this and in other examples, while Table 3 shows the corresponding individual level and the trial level measures of association. The first three rows of Table 2 (labelled 1a, 1b and 1c) show the surrogates

Table 2 Description of clinical situations

Disease	Z Treatment comparison	S Surrogate endpoint	T True endpoint
1a Advanced prostate cancer[18,19]	Liarozole vs antiandrogens	PSA response	Survival
1b		Time to PSA progression	Survival
1c		Repeated measures PSA	Survival
2a Advanced colorectal cancer[20,21]	Experimental vs bolus 5FU	Tumor response (Yes/No)	Survival
2b		Tumor response (CR/PR/SD/PD)	Survival
3 Advanced colorectal cancer[15,30,51]	Interferon-α + 5FU vs 5FU	Progression-free survival	Survival
4 Advanced ovarian cancer[14,15,30,51]	CAP vs CP	Progression-free survival	Survival
5 Schizophrenia[30]	Risperidone vs active control	PANSS response	CGI response
6 Schizophrenia[16,30]	Risperidone o.d. vs b.i.d.	PANSS response	CGI response
7 Age-related macular degeneration[14,15,30,51]	Interferon-α vs placebo	Visual acuity at 6 months	Visual acuity at 12 months

PSA: prostate-specific antigen.
5FU: 5-fluorouracil.
CR: complete response, PR: partial response, SD: stable disease, PD: progressive disease.
CAP: cyclophosphamide, adriamycin, platinum, CP: cyclophosphamide, platinum.
PANSS: positive and negative symptom scale, CGI: Clinician's global impression.
o.d.: once daily, b.i.d.: twice daily.

considered successively in advanced prostate cancer: response to PSA, time to PSA progression, and longitudinal PSA. The first three rows of Table 3 indicate that PSA does not qualify as an acceptable surrogate at this stage of the disease, regardless of how it is analysed, in spite of its strong association with survival at the individual level. The trial level associations are all low, and even when the full PSA pattern is taken into account in a longitudinal analysis, R^2_{trial} is less than 50%, a value too low to permit reliable prediction of the effect of treatment on the clinical endpoint, having observed the effect of treatment on the biomarker.[18,19]

It is also clear from Table 3 that the trial level associations are estimated very imprecisely, because of the relatively small number of units (centers) available to estimate treatment effects. In general, the individual level associations can be estimated far more precisely than the trial level associations, because of the large number of patients available.[12–17,20,21] Unfortunately, the individual level associations between biomarkers and clinical endpoints are usually of secondary interest in validation studies, because these associations are strong and often well documented in previous studies. The main focus of validation studies will thus typically be on trial level associations between the effects of some treatment(s) on the biomarker and the clinical endpoint.[14,22,23]

Similar comments can be made about the second example illustrated in Tables 2 and 3 (rows labelled 2a and 2b). This example concerns advanced colorectal cancer, the clinical endpoint of interest is survival as in the example above, and the potential surrogate is response to treatment, defined either as a binary variable (complete or partial tumor shrinkage versus no tumor shrinkage), or as an ordinal variable (complete response, partial response, stable disease, progressive disease). Response is strongly associated to survival at the individual level, but the effects of treatment on survival are again poorly predicted by the effects on response.[20,21] Ordinal response is more strongly associated to survival than binary response, but the opposite is true, perhaps somewhat surprisingly, at the trial level. This example nicely illustrates that the two levels of association are independent of each other.

Table 3 Association statistics

Surrogate endpoint	True endpoint	Individual level		Trial level	
		Number of individuals	Association measure (95% CI)	Number of units	R^2_{trial} (95% CI)
1a Binary	Failure time	596	OR = 5.5([2.7,8.2])	19	0.05([0.00,0.31])
1b Failure time	Failure time		OR = 6.3([4.4,8.2])		0.22([0.00,0.58])
1c Longitudinal process	Failure time		$R^2_{\text{indiv}}(t) > 0.84$		0.45([0.09,0.81])
2a Binary	Failure time	3791	OR = 4.7([4.0,5.5])	25	0.38([0.09,0.68])
2b Categorical	Failure time		OR = 6.3([5.6,7.0])		0.12([0.00,0.42])
3 Failure time	Failure time	736	$R^2_{\text{indiv}} = 0.57$([0.52,0.62])	68	0.57([0.41,0.72])
4 Failure time	Failure time	1194	$R^2_{\text{indiv}} = 0.89$([0.87,0.90])	50	0.94([0.91,0.97])
5 Binary	Binary	805	$R^2_{\text{indiv}} = 0.56$([0.43,0.68])	138	0.51([0.47,0.55])
6 Binary	Binary	206	$R^2_{\text{indiv}} = 0.70$([0.44,0.96])	34	0.55([0.47,0.62])
7 Continuous	Continuous	190	$R^2_{\text{indiv}} = 0.48$([0.38,0.59])	42	0.69([0.52,0.86])

OR: odds survival ratio.
CI: confidence interval.

The third and fourth examples of Tables 2 and 3 (rows labelled 3 and 4) concern two situations in advanced cancer where progression free survival (the time to tumor progression or death from any cause) is contemplated as a possible surrogate for survival. In the case of advanced colorectal cancer, moderate associations exist both at the individual level ($R^2_{\text{indiv}} = 0.57$) and at the trial level ($R^2_{\text{trial}} = 0.57$).[15,29,52] In the case of advanced ovarian cancer, the associations are much stronger both at the individual level ($R^2_{\text{indiv}} = 0.89$) and at the trial level ($R^2_{\text{trial}} = 0.94$).[14,15,29,52] Thus we could claim that progression free survival is a better surrogate for survival in advanced ovarian cancer than in advanced colorectal cancer. This may be due to the fact that advanced ovarian cancer is a more slowly progressing disease than advanced colorectal cancer, though once progression is noted, the remaining time to death is about identical in both diseases.

The fifth and sixth examples of Tables 2 and 3 (rows labelled 5 and 6) concern the very different clinical situation of schizophrenia, a disease in which several scales exist to measure the functional status of the patient. In fact, none of these scales can be viewed as a gold standard, but we posit here, for illustrative purposes, that the PANSS (positive and negative symptom scale) is proposed as a surrogate scale for the CGI (clinician's global impression). In two successive trials, the measures of associations were rather close to each other both at the individual level ($R^2_{\text{indiv}} = 0.56$ and 0.70, respectively) and at the trial level ($R^2_{\text{trial}} = 0.51$ and 0.55, respectively), yet the treatment comparisons were rather different in these two trials: it was risperidone versus active control in the first one[29] and two modes of administration of risperidone in the second one.[16,29] The first trial had more patients and more units (centers) than the second one, which resulted in tighter confidence limits of the association statistics at both levels.

The last example of Tables 2 and 3 (row labelled 7) illustrate an interesting situation in ophthalmology, where the surrogate and the true endpoints are in fact the same continuous variable (the visual acuity score) measured at two different time points (six months and 12 months).[14,15,29,52] It is well known that in the disease considered, age-related macular degeneration, measures of visual acuity taken on the same patient at different time points are only poorly correlated ($R^2_{\text{indiv}} = 0.48$). Yet, at the trial level, the effect of the drug Interferon-α on the six-month time point predicts the effect of the drug on the 12-month time point reasonably well ($R^2_{\text{trial}} = 0.69$, with 95% confidence limits [0.52,0.86]). If regulatory agencies found this trial level association sufficiently convincing to grant market authorization based on the six-month data, half a year could thus be gained in the drug approval process (note that the final endpoint at 12 months would be available later anyway). In other ophthalmic conditions such as glaucoma, where the loss of vision is much slower than in age-related macular degeneration, having an acceptable surrogate would potentially yield far greater gains in the time required to approve active new drugs.

8 Concluding comments

The approach outlined above essentially combines the concepts initially formalized by Prentice[1] with a hierarchical view. The intuitive appeal of this approach is the requirement of a good surrogate to satisfy two distinct but similar looking properties. First, the surrogate endpoint must predict the true endpoint in individual patients. This condition is in fact strongly related to Prentice's definition. Second, the effect of a treatment on the surrogate endpoint must predict the effect of that treatment on the true endpoint. While different, at face value, from Prentice's fourth criterion, this condition is consistent with its spirit, properly exploiting replication across trials.

The strength of the associations can be quantified in a straightforward manner through R^2 statistics in the case of normally distributed endpoints. These statistics do not readily generalize to settings with nonnormal outcomes, but the likelihood reduction factor applies to a wide variety of settings (normal, binary, categorical, survival, and longitudinal outcomes) and reduces to the R^2 measure for normally distributed endpoints. It generalizes both Prentice[1] and Buyse *et al.*[14]

The hierarchical framework applies when several units of analysis are available. In general, this framework requires that a multicentric trial, or preferably several distinct trials, be available for analysis. For the R^2 statistics to be practically useful, they must be estimated with sufficient precision, which in turn requires large amounts of data (in terms of both patients and units of analysis). This should not be seen as a disadvantage of the approach, but as a necessary requirement before a surrogate endpoint is used in lieu of a clinical endpoint. It is clear that this requirement is not sufficient. Surrogate marker validation cannot rely solely on statistical findings, as important clinical and biological considerations will always need to be factored into the decision.

Acknowledgements

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

References

1 Prentice RL. Surrogate endpoints in clinical trials: definitions and operational criteria. *Statistics in Medicine* 1989; **8**: 431–40.

2 Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Annals of Internal Medicine* 1996; **125**: 605–13.

3 The 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* 1989; **321**: 406–12.

4 DeGruttola V, Tu XM. Modelling progression of CD-4 lymphocyte count and its relationship to survival time. *Biometrics* 1995; **50**: 1003–14.

5 Lagakos SW, Hoth DF. Surrogate markers in AIDS: Where are we? Where are we going? *Annals of Internal Medicine* 1992; **116**: 599–601.

6 Biomarkers Definition Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical Pharmacology and Therapeutics* 2001; **69**: 89–95.

7 Ferentz AE. Integrating pharmacogenomics into drug development. *Pharmacogenomics* 2002; **3**: 453–67.

8 Lesko LJ, Atkinson AJ. Use of biomarkers and surrogate endpoints in drug development and regulatory decision making: criteria, validation, strategies. *Annual Review of*

Pharmacololgy and Toxicology 2001; **41**: 347–66.

9 Royston P, Parmar MKB, Qian W. Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer. *Statistics in Medicine* 2003; **22**: 2239–56.

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

11 International Conference on harmonisation of technical requirements for registration of pharmaceuticals for human use. ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials. (http://www.ich.org/pdfICH/e9.pdf), Federal Register 1998, **63**, No. 179, 49583.

12 Buyse M, Molenberghs G. Criteria for the validation of surrogate end-points in randomized experiments. *Biometrics* 1998; **54**: 1014–29.

13 Molenberghs G, Geys H, Buyse M. Evaluation of surrogate end-points in randomized experiments with mixed discrete and continuous outcomes. *Statistics in Medicine* 2001; **20**: 3023–38.

14 Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* 2000; **1**: 49–68.

15 Burzykowski T, Molenberghs G, Buyse M, Geys H, Renard D. Validation of surrogate endpoints in multiple randomized clinical trials with failure-time endpoints. *Applied Statistics* 2001; **50**: 405–22.

16 Renard D, Geys H, Molenberghs G, Burzykowski T, Buyse M. Validation of surrogate endpoints in multiple randomized clinical trials with discrete outcomes. *Biometrical Journal* 2002; **44**: 921–35.

17 Alonso A, Geys H, Molenberghs G, Vangeneugden T. Investigating the criterion Q8 validity of psychiatric symptom scales using surrogate marker validation methodology. (Submitted), 2003.

18 Renard D, Geys H, Molenberghs G, Burzykowski T, Buyse M, Vangeneugden T and Bijnens L. Validation of a longitudinally measured surrogate marker for a time-to-event endpoint. *Journal of Applied Statistics* 2002; **30**: 235–47.

19 Buyse M, Vangeneugden T, Bijnens L, Geys H, Renard D, Burzykowski T, Molenberghs G. Validation of Biomarkers as Surrogates for Clinical Endpoints. In Bloom JC and Dean RA, eds *Biomarkers in Clinical Drug Development*. Marcel Dekker: New York, 2003.

20 Buyse M, Thirion P, Carlson RW, Burzykowski T, Molenberghs G, Piedbois P, for the Meta-Analysis Group In Cancer. Relation between tumour response to first-line chemotherapy and survival in advanced colorectal cancer: a meta-analysis. *Lancet* 2000; **356**: 373–78.

21 Burzykowski T, Molenberghs G, Buyse M, Geys H, Renard D. The validation of surrogate endpoints using data from randomized clinical trials: a case study in advanced colorectal cancer. *Journal of the* Q9 *Royal Statistical Society, Series A* 2003; **00**: 000–000. Q9

22 Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine* 1997; **16**: 1515–27.

23 Gail MH, Pfeiffer R, van Houwelingen HC, Carroll RJ. On meta-analytic assessment of surrogate outcomes. *Biostatistics* 2000; **1**: 231–46.

24 Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine* 1992; **11**: 167–78.

25 Choi S, Lagakos S, Schooley RT, Volberding PA. CD4+ lymphocytes are an incomplete surrogate marker for clinical progression in persons with asymptomatic HIV infection taking zidovudine. *Annals of Internal Medicine* 1993; **118**: 674–80.

26 Lin DY, Fleming TR, De Gruttola V. Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in Medicine* 1997; **16**: 1515–27.

27 Flandre P, Saidi Y. Letter to the editor: estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in Medicine* 1999; **18**: 107–15.

28 Begg C, Leung D. On the use of surrogate endpoints in randomized trials. *Journal of the Royal Statistal Society Series A* 2000; **163**: 26–27. Q11

29 Molenberghs G, Buyse M, Geys H, Renard D, Burzykowski T. Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Controlled Clinical Trials* 2002; **23**: 607–25.

30 Cook RD, Weisberg S. Residuals and influence in regression. London: Chapman & Hall, 1982.

31 Chatterjee S, Hadi AS. *Sensitivity analysis in linear regression*. New York: John Wiley & Sons, 1988.

32 Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 1978; **65**: 141–51.

33 Hougaard P. Survival models for heterogeneous populations derived from stable distributions. *Biometrika* 1986; **73**, 387–96.

34 Genest C, McKay J. The joy of copulas: bivariate distributions with uniform marginals. *American Statistician* 1986; **40**, 280–83.

35 Shih JH, Louis, TA. Inferences on association parameter in copula models for bivariate survival data. *Biometrics* 1995; **51**, 1384–99.

36 Geys H, Regan M, Catalano P, Molenberghs G. Two latent variable risk assessment approaches for combined continuous and discrete outcomes from developmental toxicity data. *Journal of Agricultural, Biological, and Environmental Statistics* 2001; **6**, 340–55.

37 Dale JR. Global cross ratio models for bivariate, discrete, ordered responses. *Biometrics* 1986; **42**: 909–17.

38 Galecki A. General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics: Theory and Methods* 1994; **23**: 3105–19.

39 Kent, J. Information gain and a general measure of correlation. *Biometrika* 1983, **70**, 163–73.

40 Debruyne FJM, Murray R, Fradet Y, Johansson JE, Tyrrell C, Boccardo F, Denis L, Marberger JM, Brune D, Rassweiler J, Vangeneugden T, Bruynseels J, Janssens M, de Porre P for the Liarozole Study Group. Liarozole – a novel treatment approach for advanced prostate cancer: results of a large randomized trial versus cyproterone acetate. *Urology* 1998; **52**: 72–81.

41 Sridhara R, Eisenberger MA, Sinibaldi VJ, *et al.* Evaluation of prostate-specific antigen as a surrogate marker for response of hormone-refractory prostate cancer to suramin therapy. *Journal of Clinical Oncology* 1995; **13**: 2944–53.

42 Smith DC, Dunn RL, Strawderman MS, *et al.* Change in serum prostate-specific antigen as a marker of response to cytotoxic therapy for hormone-refractory prostate cancer. *Journal of Clinical Oncology* 1998; **16**: 1835–43.

43 Kelly WK, Scher HI, Mazumdar M, *et al.* Prostate-specific antigen as a measure of disease outcome in metastatic hormone-refractory prostate cancer. *Journal of Clinical Oncology* 1993; **11**: 607–15.

44 Scher HI, Kelly WK, Zhang ZF, *et al.* Post-therapy serum prostate-specific antigen level and survival in patients with androgen-independent prostate cancer. *Journal of the National Cancer Institute* 1999; **91**: 244–51.

45 Bubley GJ, Carducci M, Dahut W, Dawson N, Daliani D, Eisenberger M, Fidd WD, Freidlin B, Halabi S, Hudes G, Hussain M, Kaplan R, Myers C, Oh W, Petrylak DP, Reed E, Roth B, Sartor O, Scher H, Simons J, Sinibaldi V, Small EJ, Smith MR, Trump DL, Vollmer R, Wilding G. Eligibility and response guidelines for phase II clinical trials in androgen-independent prostate cancer: recommendations from the prostate-specific antigen working group. *Journal of Clinical Oncology* 1999; **17**: 3461–67.

46 Buyse M, Piedbois P. On the relationship between response to treatment and survival. *Statistics in Medicine* 1996; **15**: 2797–812. **Q10**

47 Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumor response. *Journal of Clinical Oncology* 1983; **1**: 710–19.

48 DeGruttola V, Wulfsohn M, Fischl MA, Tsiatis A. Modelling the relationship between survival and CD4 lymphocytes in patients with AIDS and AIDS-related complex. *Journal of Acquired Immune Deficiency Syndromes* 1993; **6**: 359–65.

49 Diagnostic and therapeutic technology assessment (DATTA). Surrogate markers of progressive HIV disease. *Journal of the American Medical Association* 1992; **267**: 2948–52.

50 Ellenberg SS. Surrogate endpoints in clinical trials: getting closer to identifying markers for survival in AIDS. *British Medical Journal* 1991; **302**: 63–64.

51 Machado SG, Gail MH, Ellenberg SS. On the use of laboratory markers as surrogates for clinical endpoints in the evaluation of

treatment for HIV infection. *Journal of Acquired Immune Deficiency Syndromes* 1990; **3**: 1065–73.

Q11 52 Tibaldi FS, Abrahantes JC, Molenberghs G, Renard D, Burzykowski T, Buyse M, Parmar M, Stijnen T, Wolfinger R. Simplified hierarchical linear models for the evaluation of surrogate endpoints. *Journal of Statistical Computation and Simulation* 2003; **00**: 000–000.

53 Alonso A, Molenberghs G, Burzykowski T, Renard D, Geys H, Shkedy Z, Tibaldi F, Abrahantes JC, Buyse M. Prentice's approach and the meta-analytic paradigm: a reflection on the role of statistics in the evaluation of surrogate endpoints. (Submitted), 2003. Q8

JOURNAL TITLE: Statistical Methods in Medical Research PUBLISHER: Arnold

FIRST AUTHOR: Molenberghs PAPER NO.: SM362RA

PAGE (AUTHOR'S MANUSCRIPT)	QUERY NO.	QUERY	RESPONSE
2	1	Is this the correct spelling for the drugs? ?Flecainide	
12 Equation	2	Is the semi-colon part of the equation or not?	
16	3	Bold changed to italic ?OK	
17	4	bijective ?OK	
17	5	bijective ?OK	
21	6	This whole paragraph is one sentence. Please repunctuate.	
23	7	What is (8)?	
31	8	Both Alanso et al. references. More details please – are they now accepted? Also second reference (now 53) does not seem to be mentioned in text.	
31	9	Volume and page number please	
35	10	Volume and page numbers please	
5 14	11	Buyse et al. (previously 30, now 23) but this ref is Gail et al. Also please check Shih and Louis (ref 35)?	
References have been renumbered as style, please check very carefully			

Journals Offprint Order Form

ARNOLD

A member of the
*Hodder Headline
Group*

Arnold Journals
338 Euston Road
London NW1 3BH
Tel: +44 (0)20 7873 6000
Fax: +44 (0)20 7873 6376

This form should be returned at once to the above address

- Title of Journal:
- 1st Author:

Statistical Methods in Medical Research 13(3) [provisional]

A. FREE OFFPRINTS - 25 offprints of your article will be supplied free of charge

Please indicate opposite, the name and full postal address to whom they should be sent. In the case of multi-author articles, free offprints are only sent to the corresponding author.

B. PURCHASE OF ADDITIONAL OFFPRINTS

Please note that if an article is by more than one author, only one offprint form is sent and all offprints should be ordered on that form in consultation with the co-authors.

Offprint Price List (£ sterling UK and Europe; US$ Rest of World)

	25		50		100		150		200	
	£	$	£	$	£	$	£	$	£	$
1-4 pages	76	122	99	158	149	238	227	363	273	437
5-8 pages	100	160	131	210	200	320	276	442	359	574
9-16 pages	131	210	149	238	227	363	301	482	399	638
17-24 pages	149	238	173	277	250	400	357	571	454	726
Extra 8 pages	15	24	22	35	26	42	32	51	45	72

For larger quantities contact the publisher for a quotation.

Add 100% for any offprints including colour reproduction.

I wish to purchase ....additional offprints

ADDRESS FOR DELIVERY

(please print in capitals)

IMPORTANT

1. Cheques drawn on a UK or US bank should be made payable to Hodder Headline Group. We are unable to accept credit or debit card payments.

2. Orders will not normally be mailed until the publisher is in receipt of either the appropriate payment or an official purchase order.

3. The above are prepublication prices and apply only to orders received before the publication goes to press.

4. All despatches are by surface mail, normally within four weeks of publication.

5. Claims cannot be considered more than three months after despatch.

VAT will be added to UK invoices. Members of the EU will be required to pay VAT unless a VAT number is provided with order.

VAT Number: ……………………………………………………..

ADDRESS FOR INVOICE (please print in capitals)

□ Payment enclosed

□ Please invoice

□ Official order follows

□ Official order attached

Order no. ....

Signed..

Date........................./................./..................

TRANSFER OF COPYRIGHT

ARNOLD
A member of the
Hodder Headline
PLC Group

Please return completed form to:

Arnold Journals
Statistical Methods in Medical Research
338 Euston Road
London NW1 3BH
UK.

STATISTICAL METHODS IN MEDICAL RESEARCH

In consideration of the publication in the above journal *Statistical Methods in Medical Research* the contribution entitled:..

.... ("the contribution")

by (all authors' names): ....

1. **To be filled in if copyright belongs to you**
I/we warrant that I am/we are the sole owner/s of the complete copyright in the Contribution and I/we hereby assign to Arnold (Publishers) Limited the complete copyright in the Contribution in all formats and media.

2. **To be filled in if copyright does not belong to you**

a) Name and address of copyright holder..
....
....

b) I/we warrant that I am/we are the sole owner/s of the complete copyright in the Contribution and I/we hereby grant Arnold (Publishers) Limited the non-exclusive right to publish the Contribution throughout the world in all formats and media and to deal with requests from third parties in the manner specified in paragraphs 2 and 4 overleaf.

3. **To be filled in if US Government exemption applies**
I/we certify that the Contribution was written in the course of employment by the United States Government, and therefore copyright protection is not available.

4. I/we warrant that I/we have full power to enter into this Agreement, and that the Contribution does not infringe any existing copyright, or contain any scandalous, defamatory, libellous or unlawful matter.

Signed as (tick one)
□ the sole author(s) of the Contribution
□ one author authorised to execute this transfer on behalf of all the authors of the Contribution
□ the copyright holder or authorised agent of the copyright holder of the Contribution

Name (block letters) ....

Address ....

Signature .... Date ....

(Additional authors should provide this information on a separate sheet please)

Notes for Contributors

1. The Journal's policy is to acquire copyright in all contributions. There are two reasons for this: (a) ownership of copyright by one central organisation tends to make it easier to maintain effective international protection against unauthorised use; (b) it also allows for requests from third parties to reprint or reproduce a contribution, or part of it, to be handled in accordance with a general policy which is sensitive both to any relevant changes in international copyright legislation and to the general desirability of encouraging the dissemination of knowledge.

2. Arnold co-operates in various licensing schemes which allow organisations to copy material within agreed restrains (e.g. the CLA in the UK and the CCC in the USA).

3. All contributors retain the rights to reproduce their paper for their own purposes provided no sale is involved, and to reprint their paper in any volume of which they are editor or author. Permission will automatically be given to the publisher of such a volume, subject to the normal acknowledgement.

4. It is understood that in some cases copyrights will be held by the contributor's employer. If so, Arnold requires non-exclusive permission to deal with requests from third parties, on the understanding that any requests it receives will be handled in accordance with paragraph 3.

5. Arnold will provide each contributor with a complimentary copy of the issue of the Journal in which the Contribution appears.