ELSEVIER

Controlled Clinical Trials 23 (2002) 607–625

Controlled Clinical Trials

Statistical challenges in the evaluation of surrogate endpoints in randomized trials

Geert Molenberghs, Ph.D.[a,*], Marc Buyse, Ph.D.[b], Helena Geys, Ph.D.[a], Didier Renard, Ph.D.[a], Tomasz Burzykowski, Ph.D.[a], Ariel Alonso, M.Sc.[a]

[a]*Limburgs Universitair Centrum, tUL, Center for Statistics, Biostatistics, Diepenbeek, Belgium*
[b]*International Drug Development Institute, Brussels, Belgium*

Manuscript received August 30, 2001; manuscript accepted July 15, 2002

Abstract

The validation of surrogate endpoints has been studied by Prentice, who presented a definition as well as a set of criteria that are equivalent if the surrogate and true endpoints are binary. Freedman et al. supplemented these criteria with the so-called proportion explained. Buyse and Molenberghs proposed to replace the proportion explained by two quantities: (1) the relative effect, linking the effect of treatment on both endpoints, and (2) the adjusted association, an individual-level measure of agreement between both endpoints. In a multiunit setting, these quantities can be generalized to a trial-level measure of surrogacy and an individual-level measure of surrogacy. In this paper, we argue that such a multiunit approach should be adopted because it overcomes difficulties that necessarily surround validation efforts based on a single trial. These difficulties are highlighted.

Keywords: Adjusted association; Meta-analysis; Proportion explained; Random-effects model; Relative effect; Surrogate endpoint; Validation

Introduction

Surrogate endpoints are endpoints that can replace or supplement other endpoints in the evaluation of experimental treatments or other interventions. For example, surrogate endpoints are useful when they can be measured earlier, more conveniently, or more frequently than the endpoints of interest, which are referred to as the "true" endpoints [1].

* Corresponding author: Geert Molenberghs, Limburgs Universitair Centrum, Center for Statistics, Biostatistics, Universitaire Campus, B-3590 Diepenbeek, Belgium. Tel.: +32-11-26-8238; Fax: +32-11-26-8299.
E-mail address: geert.molenberghs@luc.ac.be

0197-2456/02/$—see front matter
PII: S0197-2456(02)00236-2

Prentice [2] proposed a formal definition of surrogate endpoints and outlined how potential surrogate endpoints could be validated. Much debate ensued, for the criteria set out by Prentice are not straightforward to verify [3]. In addition, Prentice's operational criteria are only equivalent to his definition in the case of binary endpoints [4]. Freedman et al. [5] supplemented Prentice's approach by introducing the proportion explained (PE), which is the proportion of the treatment effect mediated by the surrogate. Buyse and Molenberghs [4] proposed to replace it with two new measures. The first one, defined at the population level and termed "relative effect" (RE), is the ratio of the overall treatment effect on the true endpoint over that on the surrogate endpoint. The second one is the individual-level association between both endpoints, after accounting for the effect of treatment, and is referred to as "adjusted association."

In turn, a drawback of the RE is that, when calculated from a single trial, its use depends on strong unverifiable assumptions, the main one being that it should be constant across a class of trials. A way out of this problem is to combine information from several groups of patients (multicenter trials or meta-analyses). Such an approach was suggested by Albert et al. [6] and was implemented by Daniels and Hughes [7] and by Buyse et al. [8]. Gail et al. [9] contrast the work of Daniels and Hughes [7] and Buyse et al. [8] and address several important issues. The latter extended the adjusted association and the RE to an individual-level measure of association and a trial-level measure of association, respectively. They suggest the use of these or similar measures as an alternative way to assess the usefulness of a surrogate endpoint. An important aspect of such measures is that they allow one to quantify the quality of a surrogate. Thus, one is not confined to an "all or nothing" situation where a candidate endpoint is either perfect or no surrogate at all.

A question that then arises naturally is whether, in addition to these new measures, single-trial based quantities such as the PE or the RE still convey useful information. In this paper, we show that these quantities may be misleading.

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. Therefore, this section should not be seen as setting the scene for the rest of the paper. This is reserved for the multiunit case (discussed later).

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

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

Note that this model is introduced because it is a component of the Prentice-Freedman framework. Given that the fourth criterion will involve a dependence on the treatment as well, as in Eq. (5), it is of legitimate concern to doubt whether Eqs. (4) and (5) are simultaneously plausible. Also, the introduction of Eq. (4) should *not* be seen as an implicit or explicit assumption about the absence of treatment effect in the regression relationship, but rather as a model that can be used when the uncorrected association between both endpoints is of interest.

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.

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" [2]. In terms of our simple model, Eqs. (1) and (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. 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 Eq. (1)],
- treatment has a significant impact on the true endpoint [parameter β differs significantly from zero in Eq. (2)],
- the surrogate endpoint has a significant impact on the true endpoint [parameter γ differs significantly from zero in Eq. (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 Eqs. (1) and (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 ε_{Tj} is given by

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

It is usually stated that the fourth criterion requires that the parameter β_S be equal to zero (we return to this notion later). Essentially, this last criterion states that the true endpoint T is completely determined by knowledge of the surrogate endpoint S. Buyse and Molenberghs [4] showed that the last two criteria are necessary and sufficient for binary responses, but not in general. Several authors, including Prentice, pointed out that the criteria are too stringent to be fulfilled in real situations [2,10].

In spite of these criticisms, the spirit of the fourth criterion is very appealing. This is especially true 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.

The proportion explained

Freedman et al. [5] 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. [5] proposed to calculate the proportion of the treatment effect mediated by the surrogate:

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

with β_S and β obtained respectively from Eq. (5) and Eq. (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 [4,7,11–14]. 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. [5] 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 Eq. (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.

The relative effect

Buyse and Molenberghs [4] suggested calculating 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:

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

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 Eq. (6), $\beta_S=\beta-\rho_Z\lambda\alpha$. As a result, we obtain

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

A similar relationship was derived by Buyse and Molenberghs [4] and by Begg and Leung [15] for standardized surrogate and true endpoints. This relationship will be studied in some detail later in the paper. First, our proposed multiunit framework is introduced.

Data from several units

Using ideas from Buyse et al. [8], 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, Eqs. (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, Eq. (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 Eq. (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 introduce the concept of surrogacy at both the trial level as well as the individual level. We discuss them in turn.

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 $i=1, \ldots, N$ and (2) the estimate of the effect of Z on S in a new trial $i=0$. Fitting models (12) and (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 [16,17].

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 Eqs. (17) and (18), a prediction can be made using Eq. (19), with prediction variance, Eq. (20). Of course, one has to acknowledge properly the uncertainty resulting from the fact that the parameters in Eqs. (17) and (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 (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 Eqs. (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 (θ_0, θ_a, θ_m) follow from Eqs. (14) and (15). In case the surrogate is perfect at the trial level ($R^2_{\text{trial}} = 1$), the error term in Eq. (22) vanishes and the linear relationship becomes deterministic, implying that β_i *equals* the systematic component of Eq. (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, Eq. (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 Eq. (16). In such a case, invariance is obtained only when the fixed effects in Eqs. (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.

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 Eqs. (12) and (13) we derive

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

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

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

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

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

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

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 Eqs. (12) and (13), measurement error in the surrogate is automatically taken into account. When a two-stage approximation (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) is used for such a model [8], this is no longer true. Burzykowski et al. [18] illustrate how measurement error can be incorporated in such a context.

Problems with the proportion explained

In this section, we will discuss issues with Prentice's framework in general and the PE in particular. Of course, we acknowledge that Prentice proposed a paradigm rather than a "take it or leave it" model. Further, his work has laid the foundation for all further thinking. The same holds true, to some extent, for the work of Freedman et al. [5] and Buyse and Molenberghs [4].

Expression (11) allows us to make several useful observations. It is clear from Eq. (11) that the PE is *not* a proportion. Indeed, each of λ and RE can take values over the entire real line. Further, it is hard to interpret PE because it amalgamates three sources of information:

- the adjusted association ρ_Z, which is a measure of association between the surrogate and the true endpoints at the individual level;
- the RE, which expresses the relationship between the treatment effects on the surrogate and the true endpoint at the trial level; and
- the variance ratio λ^2, which is a nuisance parameter, not to be viewed as a useful validation measure.

The fact that the PE is ill defined, except in trivial cases, and the relationship between the three measures introduced above will be studied by means of three hypothetical settings. The first two experiments concentrate on "perfect" conditions, while the last one focuses on general conditions.

Hypothetical setting 1

The PE is obviously equal to one in simple situations of perfect surrogacy, for instance if T is linearly related to S ($T=aS+b$), then Eqs. (1) and (2) can be rewritten as

$$S_j = \mu_S + \alpha Z_j + \varepsilon_{Sj}, \tag{25}$$

$$T_j = b + a\mu_S + a\alpha Z_j + a\varepsilon_{Sj}, \tag{26}$$

and obviously $\rho_Z=1$, $\lambda=a$, and RE$=\alpha$. Other simple situations are discussed by Begg and Leung [15] and Day and Duffy [19].

However, it is possible to construct examples where PE can be chosen to take any arbitrary (positive) value, depending on the values of ρ_Z, λ, and RE. To this end we consider two additional hypothetical settings.

Hypothetical setting 2

Assume $\rho_Z=1$ and RE$=1$, and suppose further that we could reduce (increase) the variance of the surrogate endpoint while keeping all other quantities unaffected, say by improving (deteriorating) the precision of its measurement. Then, Eqs. (1) and (2) would become

$$S_j = \mu_S + \alpha Z_j + \varepsilon_{Sj}, \tag{27}$$

$$T_j = \mu_S + \alpha Z_j + \lambda\varepsilon_{Sj}. \tag{28}$$

λ is arbitrary and hence so is PE, despite the fact that Eqs. (27) and (28) describe a very desirable situation. The key behind this somewhat artificial and counterintuitive hypothetical setting is that the systematic components are kept constant, and the random error terms are in perfect correlation. Then, knowledge about the surrogate endpoint enables exact prediction of the true endpoint: $E[T_j|Z_j, S_j]=T_j$.

Now, we would like to call the situation described by Eqs. (27) and (28) "perfect," even though PE may not be equal to one, nor β_S equal to zero. This casts doubts on the fourth Prentice criterion, which states that the full effect of treatment should be captured by the surrogate, even though this criterion has much intuitive appeal. In the above example, the conditional distribution of the true endpoint, given treatment and surrogate endpoint, is

$$T_j = \tilde{\mu}_T + \alpha(1-\lambda)Z_j + \lambda S_j, \tag{29}$$

which shows that the true endpoint does depend on treatment, although the residual unexplained variability in the true endpoint has been eliminated. In other words, in this perfect situation (at the individual level), Eq. (8) vanishes, which is equivalent to stating that $\rho_Z=1$. This suggests focusing on the adjusted association rather than on the adjusted treatment effect upon the true endpoint. Note that perfection in this context has no implication for the surrogate across units. To study the latter very important quality it is necessary to turn to RE or even to our multiunit setting.

It should be very clear that the thought experiment conducted here differs fundamentally from rescaling the true and surrogate endpoints. In such a case, one would divide the true endpoint by $\sqrt{\sigma_{TT}}$ and the surrogate endpoint by $\sqrt{\sigma_{SS}}$. While this, at first sight, looks like a useful calibration strategy, it is easy to see that the PE is unaffected by such a transformation, and so is the adjusted association. The only effect is that the RE is divided by the vari-

ance ratio and $\lambda = 1$. However, there is still no reason to believe that PE will be inside the unit interval.

Hypothetical setting 3

We will now switch to general conditions and consider two transformations of the surrogate endpoint:

$$S_j^{(1)} = \phi S_j + \psi = (\phi\mu_S + \psi) + \phi\alpha Z_j + \phi\varepsilon_{Sj}, \tag{30}$$

$$S_j^{(2)} = \mu_S + \alpha Z_j + \phi\varepsilon_{Sj}. \tag{31}$$

It is important to realize that the second transformation cannot be conducted through a simple transformation of a dataset variable. It might refer, for example, to a situation in a sequence of trials where at some point the precision changed due to a change in instrument or method used to measure the surrogate.

Transformation Eq. (30) operates on the fixed and random parts of the surrogate endpoint alike whereas transformation Eq. (31) operates on the random part only. The second transformation is similar to the one in the second hypothetical setting, except that we now consider the general rather than the perfect situation. It is easy to show that the following relationships hold between the validation measures:

$$\mathrm{RE}^{(1)} = \mathrm{RE}/\phi, \quad \rho_Z^{(1)} = \rho_Z, \quad \lambda^{(1)} = \lambda/\phi, \quad \mathrm{PE}^{(1)} = \mathrm{PE},$$

$$\mathrm{RE}^{(2)} = \mathrm{RE}, \quad \rho_Z^{(2)} = \rho_Z, \quad \lambda^{(2)} = \lambda/\phi, \quad \mathrm{PE}^{(2)} = \mathrm{PE}/\phi,$$

with obvious notation. Thus, for transformation Eq. (30), there is no impact on the PE, but under Eq. (31), PE is rescaled with an arbitrary amount.

There are also problems with the RE. Indeed, while the adjusted association expresses agreement between both endpoints at the individual level, the trialist will want to know how the trial-specific treatment effect on T can be predicted from the treatment effect on S. RE serves this purpose, but it is typically based on information from only one trial. It might not be constant for all trials testing the therapeutic question under consideration. The constancy of RE implies that the relation between α and β is linear through the origin. This assumption may be untenable in practice, and it cannot be verified from a single trial. Therefore, it is useful to switch to the multiunit situation. In such a context, the regression line may or may not go through the origin, without affecting the usefulness of the framework. Indeed, the regression line is concerned with the fixed effects, while the validation measures are based on variance components. Let us explore the problems with the single-unit validation measures further, within the broader context of the multiunit setting.

The PE can be calculated for each unit i:

$$\mathrm{PE}_i = \lambda\rho_Z\frac{1}{\mathrm{RE}_i}, \tag{32}$$

where $\mathrm{RE}_i = \beta_i/\alpha_i$.

Let us now examine how the PE_i behaves relative to the R^2 measures. To make the point clearly, it is useful to concentrate on a "perfect" surrogate, i.e., one for which $R^2_{\text{trial}} = 1$ and $R^2_{\text{indiv}} = \rho_Z^2 = 1$.

Perfect surrogate at the trial level

Let us first assume that the surrogate is perfect at the trial level (i.e., $R^2_{\text{trial}} = 1$). Then the relationship between α_i and β_i, expressed by Eq. (22), is deterministic, and Eq. (32) becomes

$$PE_i = \rho_Z \lambda \frac{\alpha_i}{\theta_0 + \theta_a \alpha_i + \theta_m \mu_{Si}}. \tag{33}$$

Thus, even if the important condition $R^2_{\text{trial}} = 1$ is satisfied and one can predict the treatment effect on the true endpoint without error from the treatment effect on the surrogate endpoint, PE_i cannot be constant across units, and consequently would not be equal to unity in all of them. Note also that RE_i is not constant across units. The reason is that for RE_i to be constant the relationship between α_i and β_i must be multiplicative rather than merely linear.

Perfect surrogate at the individual level

Let us now make the additional assumption that the surrogate is also perfect at the individual level, i.e., $\rho_Z = 1$. In this case, Eq. (33) becomes

$$PE_i = \lambda \frac{\alpha_i}{\theta_0 + \theta_a \alpha_i + \theta_m \mu_{Si}} \tag{34}$$

and the property of nonconstant PE_i and RE_i persists, again due to the linear but nonmultiplicative relationship between α_i and β_i.

Constant relative effect

Let us make the final assumption that a simple multiplicative relationship holds between α_i and β_i, i.e., $\theta_0 = \theta_m = 0$ and hence $RE_i = \theta_a$. Thus,

$$PE = PE_i = \frac{\lambda}{\theta_a}. \tag{35}$$

Now, RE_i is constant and so is PE_i, but the latter is still a function of two quantities:

- the multiplicative factor θ_a linking the treatment effects in each trial, and
- the multiplicative factor λ linking the two error terms in each patient.

Clearly, under the three assumptions made above, the surrogate and true endpoints are identical, up to scaling factors that translate the treatment effects within a trial and the subject-specific deviations within each patient. Yet, depending on the values of θ_a and λ, the PE can assume any positive real value.

Thus, the single-unit validation measures can be used, at best, only with the greatest caution and it seems more desirable to turn to multiunit validation measures. Regarding the choice of units, one can consider trials, but also centers, investigators, countries, etc.

Case studies

We will present results from five studies, three of which have been reported earlier: advanced macular degeneration, advanced ovarian cancer, and advanced colorectal cancer. The other two studies are in schizophrenia, with the second one of the equivalence trial type. Single- as well as multiunit results will be reported. The emphasis will be on validation measures and on predicting the treatment effect on the true endpoint in a new trial.

Description of the data

The first set of data concerns a clinical trial for patients with age-related macular degeneration (ARMD), a condition in which patients progressively lose vision [20]. Overall, 190 patients from 42 centers participated in the trial. Patients' visual acuity was assessed using standardized vision charts displaying lines of five letters of decreasing size, which patients had to read from top (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_{ij}) is set to 0 for placebo and to 1 for interferon-α. The surrogate endpoint S_{ij} is the change in the visual acuity (which we assume to be normally distributed) at 6 months after starting treatment, while the final endpoint T_{ij} is the change in the visual acuity at 1 year. In the multiunit analyses the centers in which the patients were treated will be considered as the units of analysis. Six of 42 centers participating in the trial enrolled patients only to one of the two treatment arms. These centers were excluded from consideration. A total of 36 centers were thus available for analysis, with the number of individual patients per center ranging from 2 to 18 (183 patients overall).

The second set of data comes from a meta-analysis of four randomized multicenter trials in advanced ovarian cancer [21]. Individual patient data are available in these four trials for the comparison of two treatment modalities: cyclophosphamide plus cisplatin (CP) versus cyclophosphamide plus adriamycin plus cisplatin (CAP). The binary indicator for treatment (Z_{ij}) will be set to 0 for CP and to 1 for CAP. The surrogate endpoint S_{ij} will be progression-free survival time, defined as the time (in years) from randomization to clinical progression of the disease or death, while the final endpoint T_{ij} will be survival time, defined as the time (in years) from randomization to death from any cause. The full results of this meta-analysis were published with a minimum follow-up of 5 years in all trials [21]. The dataset was subsequently updated to include a minimum follow-up of 10 years in all trials [22]. After such long follow-up, most patients had a disease progression or had died (980 of 1194 patients, 81.8%). In the majority of cases, death was clearly related to the disease (850 of 952 deaths, 89.2%). The ovarian cancer dataset contains four trials. In the two larger trials of the Gynecologic Oncology Group ($n=412$ patients) and the Gruppo Interegionale Cooperativo Oncologico Ginecologia ($n=383$ patients), information is also available on the centers in which the patients had been treated. For the two smaller trials of the Danish Ovarian Cancer Group (DACOVA, $n=274$ patients) and the Gruppo Oncologico Nord-Ovest (GONO, $n=125$ patients), the information is not available. According to the clinical investigators, the close collaboration of the members of the corresponding research groups allows the patients treated in these trials to be considered as a homogenous group. In the analyses we will then use center as the unit of analysis for the two larger trials, and the trial as the unit of analysis for the two

smaller trials. Two centers enrolled only one patient each and were excluded from considerations. A total of 50 "units" are thus available for analysis, with the number of individual patients per unit ranging from 2 to 274.

The third set contains data from two randomized multicenter trials in advanced colorectal cancer [23,24]. In one trial, treatment with 5FU plus interferon (5FU/IFN) was compared to treatment with 5FU plus folinic acid (5FU/LV) [23]. In the other trial, treatment with 5FU/IFN was compared to treatment with 5FU alone [24]. The binary indicator for treatment (Z_{ij}) will be set to 0 for 5FU/IFN and to 1 for 5FU/LV or 5FU alone. The surrogate endpoint S_{ij} will be progression-free survival time, defined as the time (in years) from randomization to clinical progression of the disease or death, while the final endpoint T_{ij} will be survival time, defined as the time (in years) from randomization to death from any cause. Most patients in the two trials had a disease progression or died (694 of 736 patients, 94.3%). Similarly to the previous example, we will consider center as the unit of analysis. However, in eight centers there were no patients accrued to one of the treatment arms. These eight centers were therefore excluded from the analysis. As a result, a total of 68 units were thus available for analysis, with the number of individual patients per unit ranging from 2 to 38 (642 patients overall). An analysis exploiting the survival nature of the endpoints in the latter two studies has been done in Burzykowski et al. [18].

The first of the two psychiatric studies is based on a meta-analysis containing only five trials. This is insufficient to apply the meta-analytic methods. In all of the trials, information is also available on the investigators who treated the patients. Thus, we can also use investigator as the unit of analysis. A total of 138 units are thus available for analysis, with the number of patients per unit ranging from 2 to 30. The true endpoint is Clinician's Global Impression (CGI). This is a seven-grade scale used by the treating physician to characterize how well a subject is doing. As a surrogate measure, we consider the Positive and Negative Syndrome Scale (PANSS) [25]. The PANSS consists of 30 items that provide an operationalized drug-sensitive instrument, which is highly useful for both typological and dimensional assessment of schizophrenia. Table 1 shows the frequency of unit-specific sample sizes. Clearly, the majority of units consists of less than five patients. Alternatively, one could also consider the main investigator as unit of analysis. For four of the five trials, only one main investigator was used, leading to extremely large investigator sites. This leads to a total number of 29 units with the number of patients per unit ranging from 4 to 450, four of which represent tri-

Table 1. Psychiatric study I. Frequency table of the number of units with a given number of patients

Patients per unit (n)	Units with n patients	Patients per unit (n)	Units with n patients
2	29	10	2
3	18	11	4
4	23	12	2
5	16	13	3
6	9	15	1
7	12	18	1
8	10	21	1
9	6	30	1

als. The comparison of both choices will be used as an empirical assessment as to the importance the choice of unit can have on the results.

Finally, we will use data from an international equivalence trial on schizophrenic patients [26]. The trial includes 206 schizophrenic patients. All patients received an equal daily amount of risperidone during 8 weeks, but 103 patients were randomized to a one-time daily intake while the remaining 103 patients were randomized to receive risperidone twice a day. The surrogate and true endpoints are again PANNS and CGI, respectively. A total of 34 units were thus available for analysis with the number of patients per unit ranging from 2 to 15. The importance of this example lies in the fact that, due to the equivalence nature, the Prentice-Freedman framework should be expected to break down, since it is based on hypothesis testing in a superiority-trial setting.

Analysis of case studies

Table 2 presents validation results for the studies described earlier; some of them have been reported elsewhere [8,18]. The first three Prentice criteria fail in four of five examples. Note in addition that for psychiatric study I, the p value for the fourth criterion is 0.513, which is not inconsistent with the framework, but cannot be seen as definitive evidence to the equivalence testing problem that surrounds the fourth criterion.

The point estimate of PE lies outside the unit interval in two cases (once greater than 1, once smaller than 0). All confidence intervals are wide and exceed the unit interval; in one case it is infinitely large. Hence, the Prentice criteria and the PE fail to convey any useful information on the quality of the proposed surrogates. Other single-trial measures are descriptively useful but also fail to permit a full qualification of the surrogates. All RE estimates exhibit extremely large confidence intervals with, again, one of them infinitely large. In all cases, the adjusted association is close to the square root of the individual-level R^2. Similarly in all cases, the confidence intervals of the adjusted association are sufficiently narrow to be of use.

The meta-analytic validation measures allow us to qualify the potential of the surrogates much better. The highest values of R^2s are seen in advanced ovarian cancer, where progres-

Table 2. Single- and multitrial validation measures for various diseases and endpoints

	Age-related macular degeneration	Advanced ovarian cancer	Advanced colorectal cancer	Psychiatric study I (138 units)	Psychiatric study I (29 units)	Psychiatric study II
Surrogate	Visual acuity (6 months)	Progression-free survival	Progression-free survival	PANSS	PANSS	PANSS
True	Visual acuity (1 year)	Overall survival	Overall survival	CGI	CGI	CGI
Prentice criteria 1–3 (p = value)						
Association (Z, S)	0.31	0.013	0.90	0.016		0.835
Association (Z, T)	0.22	0.08	0.86	0.007		0.792
Association (S, T)	<0.001	<0.001	<0.001	<0.001		<0.001
Single-unit validation measures (estimate and 95% CI)						
Proportion explained	0.61[−0.19; 1.41]	1.34[0.73; 1.95]	0.51[−4.97; 5.99]	0.81[0.46; 1.67]		−0.94[∞]
Relative effect	1.51[−0.46; 3.49]	0.65[0.36; 0.95]	1.59[−15.49; 18.67]	0.055[0.01; 0.16]		−0.03[∞]
Adjusted association	0.74[0.68; 0.81]	0.94[0.94; 0.95]	0.73[0.70, 0.76]	0.72[0.69;0.75]		0.74[0.69; 0.79]
Multiunit validation measures (estimate and 95% CI)						
R^2_{trial}	0.69[0.52; 0.86]	0.94[0.91; 0.97]	0.57[0.41, 0.72]	0.56[0.43 ;0.68]	0.58[0.45; 0.71]	0.70[0.44; 0.96]
R^2_{indiv}	0.48[0.38; 0.59]	0.89[0.87; 0.90]	0.57[0.52, 0.62]	0.51[0.47; 0.55]	0.52[0.48; 0.56]	0.55[0.47; 0.62]

sion-free survival has good potential as a surrogate for survival. Note, however, that progression-free survival would not be practically attractive in this disease since disease progression may take several years to develop and is typically followed by death within a few months. In ARMD, the loss of vision at 6 months shows little correlation with the loss of vision at 1 year, although the association between the treatment effects is stronger. This particular analysis is based upon a small dataset, but it nevertheless suggests that 6-month measurements cannot satisfactorily replace 1-year measurements in this disease. A weak individual-level surrogacy, combined with a stronger trial-level surrogacy, might be due to the impact of measurement error. Likewise, in advanced colorectal cancer, progression-free survival does not appear to be a good surrogate for survival.

In the first psychiatric study, there is remarkably little difference between the versions with 138 and 29 units. This similarity supports the use of the multiunit approach. Let us turn attention to the second psychiatric case study, where data from an equivalence trial are used. Obviously, the equivalence nature of the study renders the use of the Prentice-Freedman framework impossible since the p-values merely reflect the absence of treatment effect. Nevertheless, the meta-analytic measures are fairly precise and, importantly, the results from studies I and II are very close to each other. The latter observation supports that we have been able to quantify reasonably and accurately the surrogacy of PANSS for CGI in the context of certain compounds for schizophrenia. Of course, the R^2 values are not terribly high, so that a mere replacement of CGI by PANSS may be questionable. An interesting topic of research would be the combination of several surrogates, with a view on better overall surrogacy.

Another important advantage of the multiunit framework is the ability to predict the treatment effect on the true endpoint, given the effect on the surrogate endpoint. Results are presented for both the ovarian study (Table 3) as well as the first psychiatric study (Table 4). In all cases, the predictions were done on a leave-one-out basis to avoid overly optimistic predictions. In these tables, $\hat{\alpha}_0$ and $\widehat{\beta + b_0}$ are values estimated from the data; $E(\beta + b_0)$ is the predicted treatment effect on the true endpoint, given its effect on the surrogate endpoint. In the ovarian case, the agreement is very strong, while the agreement is reasonable in the psychiatric study; these findings are in line with the strength of the trial-level surrogacy in these examples. In the ovarian case, DACOVA and GONO refer to two large "centers," for which no subunit information was possible. It is seen that the prediction is good for both these large centers, as well as for randomly selected small centers.

Table 3. Ovarian cancer. Comparison of estimated and predicted treatment effects on true endpoint

Unit	Patients (n)	$\hat{\alpha}_0$	$E(\beta + b_0 \mid a_0)$	$\widehat{\beta + b_0}$
6	17	−0.58 (0.33)	−0.45 (0.29)	−0.56 (0.32)
8	10	0.67 (0.76)	0.49 (0.57)	0.76 (0.39)
55	31	1.08 (0.56)	0.80 (0.44)	0.79 (0.45)
DACOVA	274	0.25 (0.15)	0.17 (0.13)	0.14 (0.14)
GON	125	0.15 (0.25)	0.10 (0.20)	0.03 (0.22)

Standard errors are shown in parentheses.

Table 4. Psychiatric study I. Comparison of estimated and predicted treatment effects on true endpoint

Unit	Patients (n)	$\hat{\alpha}_0$	$E(\beta+b_0\|a_0)$	$\widehat{\beta+b_0}$
1	8	14.00 (16.35)	0.53 (0.63)	0.50 (1.26)
2	6	−43.33 (29.02)	−1.99 (0.63)	−2.33 (1.25)
3	9	−13.50 (12.75)	−0.75 (0.60)	0.30 (1.18)
4	4	7.50 (35.28)	0.08 (0.58)	1.50 (1.80)
5	9	−7.60 (7.65)	−0.45 (0.63)	−0.40 (0.99)
6	8	−42.00 (18.93)	−1.88 (0.63)	−2.50 (1.04)
7	7	−39.58 (18.71)	−2.07 (0.61)	−1.00 (1.18)
8	6	−13.33 (13.79)	−0.69 (0.62)	−1.33 (1.56)
9	6	−7.33 (23.35)	−0.44 (0.63)	−0.33 (1.33)
10	4	−2.00 (18.06)	−0.18 (0.63)	−0.50 (1.80)
11	68	−4.84 (4.46)	−0.32 (0.63)	−0.47 (0.36)
12	8	−14.25 (30.53)	−0.72 (0.62)	−1.50 (0.89)
13	7	−6.33 (11.24)	−0.37 (0.63)	−0.83 (0.95)
14	4	−36.5 (14.77)	−1.96 (0.58)	−0.50 (0.50)
15	5	−13.00 (26.93)	−0.66 (0.61)	−1.66 (1.72)
16	8	−22.75 (10.45)	−1.13 (0.63)	−1.25 (0.63)
17	8	−9.00 (10.93)	−0.52 (0.63)	−0.50 (0.65)
18	450	−3.57 (2.13)	−0.28 (0.63)	−0.15 (0.13)
19	7	−23.5 (12.02)	−1.16 (0.63)	−1.25 (0.74)
20	5	−5.33 (13.52)	−0.33 (0.63)	−0.83 (0.57)
21	70	2.75 (5.79)	−0.00 (0.63)	0.21 (0.38)
22	7	−7.50 (16.13)	−0.46 (0.63)	−0.25 (1.40)
23	7	−20.66 (15.39)	−1.00 (0.62)	−1.83 (1.06)
24	9	−4.00 (11.06)	−0.31 (0.63)	0.05 (0.93)
25	5	−7.83 (11.16)	−0.43 (0.61)	−1.33 (0.86)
26	45	−20.15 (9.68)	−1.01 (0.63)	−1.18 (0.50)
27	9	1.14 (19.19)	−0.06 (0.63)	0.00 (0.95)
28	5	−10.50 (10.96)	−0.63 (0.59)	0.66 (0.86)
29	8	−3.25 (10.71)	−0.24 (0.63)	−0.49 (0.79)

Standard errors are shown in parentheses.

Discussion

In this paper, we have argued that a classical approach to surrogate marker validation, based on the Prentice criteria and measures derived therefrom, such as the PE and the RE, is surrounded with difficulties. The PE attempts to capture the concept that the treatment effect on the true endpoint is fully explained by the surrogate. In doing so, it focuses on the conditional regression coefficient of the treatment indicator (β_S) and requires that $\beta_S=0$, or equivalently that PE=1. We have discussed cases in which this approach fails because it does not appropriately distinguish between different sources of variability. PE is in fact an amalgamation of three quantities: the trial-level relative effect, the individual-level adjusted association, and a nuisance factor related to the ratio of variances of the true and surrogate endpoints. This conceptual difficulty seems to us more worrisome than the confidence interval of PE, which, as pointed out by many authors, tends to be too wide to be useful unless trial sizes are very large or the treatment effect on the true endpoint is very strong [5]. We have argued that it is more meaningful to view the problem from the multiunit point of view. At

the individual level, we have focused on the residual variability of the conditional regression of T on S and Z, which is captured by the individual-level adjusted association between the surrogate and true endpoints. If that residual variability vanishes, then knowledge of the surrogate endpoint and treatment indicator allows one to predict the true endpoint without error, which we consider to be a perfect situation (at the individual level).

At the trial level, we have focused on the prediction of the effect of treatment on the true endpoint given its effect on the surrogate endpoint. We have called this quantity RE, the effect of treatment on the true endpoint relative to that on the surrogate endpoint. When only one trial is available, an estimate of RE is based on the strong assumption that the relationship between the treatment effects on the surrogate and true endpoints is multiplicative, an assumption that may be too strong to hold and is unverifiable. Again, this difficulty is more fundamental than the limited precision of RE, which will typically be obtained in trials of small or moderate size [4]. A more convincing approach to the problem can be worked out when multiple units are available. The adjusted association generalizes naturally to an R^2 measure of individual-level association, and the RE can be supplemented by a corresponding trial-level measure of association. If the association is perfect, then knowledge of the treatment effect upon the surrogate allows one to predict its effect upon the true endpoint, again a situation that we would consider perfect (at the trial level). For both levels, there is replication in the data and hence the posited models can be checked.

In summary, our approach differs from Prentice's criterion of full capture in that we propose to use two quantities to measure the quality of a surrogate endpoint. Prediction of the treatment effect at the trial level is undoubtedly central to the problem of surrogate validation, and some approaches are in fact based exclusively on trial-level information [7]. Prediction of the true endpoint at the individual level is only incidental to the validation problem, even though the correlation between the surrogate and the true endpoints is one of the elements to consider in the evaluation process [27,28]. Although a good correlate may not be a good surrogate [9], a poor correlate is even less likely to be one.

The conditions of perfect prediction can only be verified if data are available at both the trial and the individual levels, which implies a multiunit approach based on patient (rather than summary) data. Note that the "trial" level can be defined by any other sensible grouping of individual patients (e.g., by treatment type, by hospital, etc.). Note also that the homogeneity that is often sought in meta-analyses (same treatment regimens, same patient mix, etc.) may not be a desirable feature for the purposes of evaluating surrogate endpoints, since the trial-level association must be investigated over a sufficiently wide range of treatment effects. This consideration may in turn drive the choice of appropriate groupings of patients at the "trial" level.

In our evaluation, a subjective assessment is required as to what values of R^2 are close enough to one for the candidate surrogate to be deemed acceptable. Such subjectivity seems inescapable but will be less of an issue if several endpoints are evaluated simultaneously as candidate surrogates for the same true endpoint. In the two psychiatric case studies, for examples, it was found that in spite of two different choices for unit in the first study and with the second case study of the equivalence-trial type, the trial-level surrogacy for PANSS on CGI is around 0.60 and the individual-level surrogacy is around 0.50. This provides evidence for the conclusion that PANSS is a surrogate of a moderately weak type. Perhaps it has po-

tential to be used in conjunction with other surrogates, but this would require additional research.

A methodological issue is that the choice of an individual-level measure of agreement, such as the R^2, is not universal. In this paper, we have concentrated on the situation where the true and surrogate endpoints are both normally distributed (in which case the individual-level R^2 follows naturally as the coefficient of determination of the adjusted regression). In practice, endpoints will often be binary, time-dependent, or repeatedly measured over time, and so different association measures will have to be used depending on the problem at hand. Fortunately, in most settings it is possible to retain an R^2 measure for the trial-level surrogacy. For the individual-level surrogacy, it depends on the type of joint model for the surrogate and true outcome that is used. A bivariate probit model for binary data [29] would produce a tetrachoric correlation, while a Dale model produces odds ratios [30]. For survival endpoints [18] copula-based models have been used, of which the natural association parameters may be quite difficult to interpret. Fortunately, they can often be transformed into Kendall's tau or Spearman's rank correlation.

Acknowledgments

Dr. Geys is supported by a grant from IWT — Vlaanderen. Drs. Burzykowski and Alonso are supported by an LUC Bijzonder Onderzoeksfonds grant.

References

[1] Ellenberg SS, Hamilton JM. Surrogate endpoints in clinical trials: cancer. Stat Med 1989;8:405–413.
[2] Prentice RL. Surrogate endpoints in clinical trials: definitions and operational criteria. Stat Med 1989;8:431–440.
[3] Fleming TR, Prentice RL, Pepe MS, Glidden D. Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. Stat Med 1994;13:955–968.
[4] Buyse M, Molenberghs G. The validation of surrogate endpoints in randomized experiments. Biometrics 1998;54:1014–1029.
[5] Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. Stat Med 1992;11:167–178.
[6] Albert JM, Ioannidis JPA, Reichelderfer P, et al. Statistical issues for HIV surrogate endpoints: point/counterpoint. Stat Med 1998;17:2435–2462.
[7] Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. Stat Med 1997;16:1515–1527.
[8] Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. Biostatistics 2000;1:49–67.
[9] Gail MH, Pfeiffer R, van Houwelingen HC, Carroll RJ. On meta-analytic assessment of surrogate outcomes. Biostatistics 2000;1:231–246.
[10] Fleming TR, DeMets DL. Surrogate endpoints in clinical trials: are we being misled? Ann Intern Med 1996;125:605–613.
[11] Volberding PA, Lagakos SW, Koch MA, et al. Zidovudine in asymptomatic human immunodeficiency virus infection: a controlled trial in persons with fewer than 500 CD4-positive cells per cubic millimeter. N Engl J Med 1990;322:941–949.
[12] 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. Ann Intern Med 1993;118:674–680.
[13] Lin DY, Fleming TR, DeGruttola V. Estimating the proportion of treatment effect explained by a surrogate marker. Stat Med 1997;16:1515–1527.
[14] Flandre P, Saidi Y. Letters to the editor: estimating the proportion of treatment effect explained by a surrogate marker. Stat Med 1999;18:107–115.
[15] Begg CB, Leung DHY. On the use of surrogate endpoints in randomized trials (with discussion). J R Stat Soc A 2000;163:15–28.
[16] Cook RD, Weisberg S. Residuals and influence in regression. London: Chapman & Hall, 1982.
[17] Chatterjee S, Hadi AS. Sensitivity analysis in linear regression. New York: John Wiley & Sons, 1988.

[18] Burzykowski T, Molenberghs G, Buyse M, Geys H, Renard D. Validation of surrogate endpoints in multiple randomized clinical trials with failure-time endpoints. Appl Stat 2001;50:405–422.
[19] Day NE, Duffy SW. Trial design based on surrogate end points–application to comparison of different breast screening frequencies. J R Stat Soc A 1996;159:49–60.
[20] Pharmacological Therapy for Macular Degeneration Study Group. Interferon a-IIA is ineffective for patients with choroidal neovascularization secondary to age-related macular degeneration. Results of a prospective randomized placebo-controlled clinical trial. Arch Ophthalmol 1997;115:865–872.
[21] Ovarian Cancer Meta-Analysis Project. Cyclophosphamide plus cisplatin versus cyclophosphamide, doxorubicin, and cisplatin chemotherapy of ovarian carcinoma: a meta-analysis. J Clin Oncol 1991;9:1668–1674.
[22] Ovarian Cancer Meta-Analysis Project. Cyclophosphamide plus cisplatin versus cyclophosphamide, doxorubicin, and cisplatin chemotherapy of ovarian carcinoma: a meta-analysis. Classic Papers and Current Comments 1998;3:237–243.
[23] Corfu-A Study Group. Phase III randomized study of two fluorouracil combinations with either interferon alfa-2a or leucovorin for advanced colorectal cancer. J Clin Oncol 1995;13:921–928.
[24] Greco FA, Figlin R, York M, et al. Phase III randomized study to compare interferon alfa-2a in combination with fluorouracil versus fluorouracil alone in patients with advanced colorectal cancer. J Clin Oncol 1996;4:2674–2681.
[25] Kay SR, Opler LA, Lindenmayer JP. Reliability and validity of the positive and negative syndrome scale for schizophrenics. Psychiatry Res 1988;23:99–110.
[26] Nair NPV and the Risperidone Study Group. Therapeutic equivalence of risperidone given once daily and twice daily in patients with schizophrenia. J Clin Psychopharmacol 1998;18:103–110.
[27] Jacobson MA, Bacchetti P, Kolokathis A. Surrogate markers for survival in patients with AIDS and AIDS related complex treated with zidovudine. BMJ 1991;302:73–78.
[28] O'Brien WA, Hartigan PM, Martin D, et al. Changes in plasma HIV-1 RNA and CD4+ lymphocyte counts and the risk of progression to AIDS. N Engl J Med 1996;334:426–431.
[29] Dale JR. Global cross ratio models for bivariate, discrete, ordered responses. Biometrics 1986;42:909–917.
[30] Renard D, Geys H, Molenberghs G, Burzykowski T, Buyse M. Validation of surrogate endpoints in multiple randomized clinical trials with discrete outcomes. Biometrical J 2002;44:1–15.