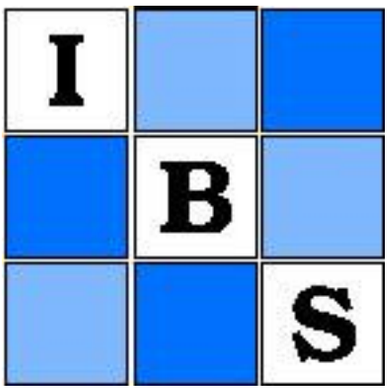

BIOMETRICS 54, 1014–1029
September 1998

Criteria for the Validation of Surrogate Endpoints in Randomized Experiments

Marc Buyse

International Institute for Drug Development,
430 Avenue Louise B14, B 1050 Brussels, Belgium, and
Limburgs Universitair Centrum, Biostatistics, Diepenbeek, Belgium

and

Geert Molenberghs*

Limburgs Universitair Centrum, Biostatistics,
Universitaire Campus, B 3590 Diepenbeek, Belgium

SUMMARY

The validation of surrogate endpoints has been studied by Prentice (1989, *Statistics in Medicine* **8,** 431–440) and Freedman, Graubard, and Schatzkin (1992, *Statistics in Medicine* **11,** 167–178). We extend their proposals in the cases where the surrogate and the final endpoints are both binary or normally distributed. Letting T and S be random variables that denote the true and surrogate endpoint, respectively, and Z be an indicator variable for treatment, Prentice's criteria are fulfilled if Z has a significant effect on T and on S, if S has a significant effect on T, and if Z has no effect on T given S. Freedman relaxed the latter criterion by estimating PE, the proportion of the effect of Z on T that is explained by S, and by requiring that the lower confidence limit of PE be larger than some proportion, say 0.5 or 0.75. This condition can only be verified if the treatment has a massively significant effect on the true endpoint, a rare situation. We argue that two other quantities must be considered in the validation of a surrogate endpoint: RE, the effect of Z on T relative to that of Z on S, and γ_Z, the association between S and T after adjustment for Z. A surrogate is said to be perfect at the individual level when there is perfect association between the surrogate and the final endpoint after adjustment for treatment. A surrogate is said to be perfect at the population level if RE is 1. A perfect surrogate fulfills both conditions, in which case S and T are identical up to a deterministic transformation. Fieller's theorem is used for the estimation of PE, RE, and their respective confidence intervals. Logistic regression models and the global odds ratio model studied by Dale (1986, *Biometrics* **42,** 909–917) are used for binary endpoints. Linear models are employed for continuous endpoints. In order to be of practical value, the validation of surrogate endpoints is shown to require large numbers of observations.

1. Introduction

Surrogate endpoints are loosely referred to as endpoints that can be used in lieu of other endpoints in the evaluation of experimental treatments or other interventions. 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 or final endpoints (Ellenberg and Hamilton, 1989).

The need to evaluate treatment benefits as fast as possible on easily measurable endpoints has always been a preoccupation in clinical research. In most clinical trials, several endpoints are measured over the course of the disease, and treatment benefits can be evaluated on all of them. In general, however, one endpoint is prespecified as being of primary interest and serves to determine the significance of any observed treatment benefit. Ideally, the primary endpoint should be the one

* *Corresponding author's email address:* mbuyse@luc.ac.be

Key words: Fieller's theorem; Perfect surrogate; Surrogate endpoint; Validation.

that is most clinically relevant, but considerations of time and cost may force the investigators to use some other endpoint instead. Examples abound, particularly in chronic diseases in which the duration of survival is the ultimate endpoint that clinicians would like to affect but cannot always afford to observe due to the prolonged period of follow-up needed. Alternative endpoints must then be considered as surrogates for survival, e.g., disease recurrence after surgical removal of early cancers, tumor shrinkage (usually called response) in advanced cancers, progression to AIDS in HIV positive subjects, and lymphocyte T4 (CD4) counts in AIDS patients. The true endpoint may also be a rare event, such as a disease or unexpected side-effect of treatment, that occurs so infrequently as to make a study unrealistically large. Finally, surrogate endpoints may be needed when competing risks and secondary treatments contaminate the impact of an experimental treatment or intervention on the true endpoint (Wittes, Lakatos, and Probstfield, 1989). In some cases, the surrogate endpoint directly affects the patient's condition and is therefore itself of clinical relevance; in other cases, it is merely a biological marker of the disease process leading to the final endpoint. In this case, the term surrogate marker may be preferred, and the endpoint of interest is then referred to as the clinical endpoint.

While the practice of looking at multiple endpoints is by no means recent in clinical research, the validity of using one endpoint as a surrogate for another has been raised and studied only over the last few years. 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 played a major role in focusing attention on the need for a formal definition of surrogate endpoints along with practical methods to validate them. Much applied research on surrogate endpoints has concentrated on evaluating the possible value of changes in CD4 counts as surrogates for time to clinical events in asymptomatic HIV-infected persons and in AIDS patients (Machado, Gail, and Ellenberg, 1990; Lin, Fischl, and Schoenfeld, 1993; De Gruttola et al., 1993; De Gruttola and Tu, 1995). This research has revealed that, if CD4 counts were useful to monitor the disease process, they were only of limited value as a surrogate marker for clinically relevant endpoints (Lagakos and Hoth, 1992). In cardiovascular disease, the unsettling discovery that the two major antiarrhythmic drugs encanaide and flecanaide reduced arrhythmia but cause a more than threefold increase in overall mortality stressed the need for caution in using nonvalidated surrogate markers in the evaluation of the possible clinical benefits of new drugs (Cardiac Arrhythmia Suppression Trial (CAST) Investigators, 1989). The dangers and limitations of using surrogate endpoints have been emphasized (Fleming and DeMets, 1996; De Gruttola et al., 1997).

In a landmark paper, Prentice (1989) 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 too stringent (Fleming et al., 1994) and not straightforward to verify. Freedman, Graubard, and Schatzkin (1992) took Prentice's approach one step further by proposing a formulation of the surrogacy criteria that leads to statistical hypothesis tests.

In this paper, we extend the discussion on operational criteria for the validation of surrogate endpoints with data from randomized trials. All developments are illustrated with data from an actual clinical trial in ophthalmology (Section 2). We first recall Prentice's definition and operational criteria for the validation of surrogate endpoints (Section 3). We show that Prentice's definition and operational criteria are equivalent in the case of binary endpoints, but not in more general situations. We then illustrate these criteria in the simplest case of binary endpoints (Section 4). We recall Freedman's modification of the operational criteria and we discuss the limitations of this modification. We proceed to show that additional measures are required to validate a surrogate that is of practical relevance. We propose two such measures: the first relates the effect of treatment on the true endpoint to that on the surrogate at the population level; the second quantifies the association between the true and the surrogate endpoints after taking treatment into account at the individual level. These concepts extend naturally to normally distributed endpoints (Section 5). We leave the situation of surrogate and true endpoints that are of a different data type (e.g., a continuous surrogate and a time-to-event true endpoint) for a separate paper. We conclude by discussing some general difficulties inherent to the validation of endpoints (Section 6).

2. Motivating Example

We adopt the following notation: T and S are random variables that denote the true and surrogate endpoint, respectively, and Z is an indicator variable for treatment. In the remainder of the paper, we restrict our attention to binary treatments.

The validation process will be illustrated using data from a simple yet real situation. Our data arise from a randomized clinical trial comparing an experimental treatment (interferon-α) to a

corresponding placebo in the treatment of patients with age-related macular degeneration (ARMD). We focus here on the comparison between placebo and the highest dose of interferon-α (6 million units daily), but the full results of this trial have been reported elsewhere (Pharmacological Therapy for Macular Degeneration Study Group, 1997). Patients with ARMD progressively loose vision. In the trial, a patient's visual acuity was assessed at different times points through the ability to read lines of letters on standardized vision charts. These charts display lines of five letters of decreasing size, which the patient must read from top (largest letters) to bottom (smallest letters). Each line with at least four letters correctly read is called one line of vision. The patient's visual acuity is the total number of letters correctly read. In Section 4, we analyse the primary endpoint of the trial, which was the proportion of patients having lost at least three lines of vision at 1 year, compared to their baseline performance. We examine whether the loss of at least two lines of vision at 6 months can be used as a surrogate for the loss of at least three lines of vision at 1 year with respect to the effect of interferon-α. Hence,

$$Z = \begin{cases} 0 & \text{if patient randomized to placebo,} \\ 1 & \text{if patient randomized to interferon-}\alpha \text{ (6 million units daily),} \end{cases}$$

$$S = \begin{cases} 0 & \text{if patient had lost less than two lines of vision at 6 months,} \\ 1 & \text{otherwise,} \end{cases}$$

$$T = \begin{cases} 0 & \text{if patient had lost less than three lines of vision at 1 year,} \\ 1 & \text{otherwise.} \end{cases}$$

The data are presented in Table 1.

Next we consider the secondary endpoint of the trial, which was the mean visual acuity at 1 year. We assume that visual acuity is a continuous, normally distributed variable. In Section 5, we examine whether visual acuity at 6 months can be used as a surrogate for visual acuity at 1 year with respect to the effect of interferon-α. The data are shown graphically in Figure 1.

3. Validation Criteria

3.1 *Definition*

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

Table 1

Relationship between T (true endpoint: at least three lines of vision lost at 1 year), S (surrogate endpoint: at least two lines of vision lost at 6 months), and Z (treatment: interferon-α or corresponding placebo) in patients with age-related macular degeneration. Cell counts represent numbers of patients.

(a) Three-Way Classification

		Z	
S	T	0	1
0	0	56	31
	1	9	9
1	0	8	9
	1	30	38

(b) Two-Way Classifications

	T		Z	
	0	1	0	1
S				
0	87	18	65	40
1	17	68	38	47
Z				
0	64	39		
1	40	47		

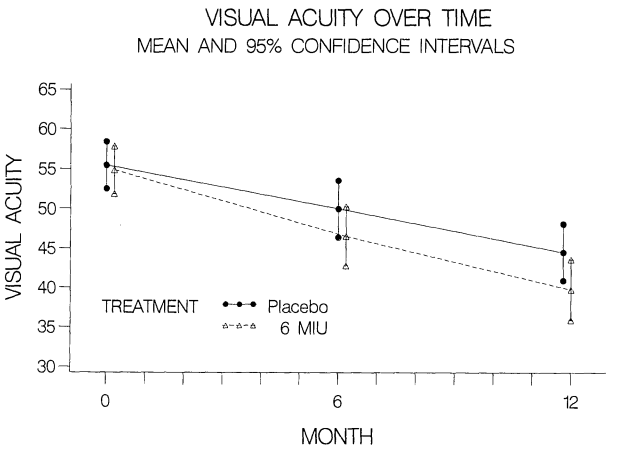

Figure 1. Mean values and confidence intervals of T (true endpoint: visual acuity at 1 year) and S (surrogate endpoint: visual acuity at 6 months) for the two levels of Z (treatment: interferon-α or corresponding placebo) in patients with age-related macular degeneration.

corresponding null hypothesis based on the true endpoint" (Prentice, 1989, p. 432). Prentice's definition can be written

$$f(S \mid Z) = f(S) \Leftrightarrow f(T \mid Z) = f(T), \tag{1}$$

where $f(X)$ denotes the probability distribution of random variable X and $f(X \mid Z)$ denotes the probability distribution of X conditional on the value of Z. Note that this definition involves the triplet (T, S, Z); hence, the endpoint S is a surrogate for T only with respect to the effect of some specific treatment Z except if S were a perfect surrogate for T, i.e., if S and T were the same endpoint up to a deterministic transformation ($S \equiv T$). The endpoints T and S can be discrete or continuous, possibly censored, random variables. Prentice (1989) focuses on the case in which T is a time-to-failure endpoint. Freedman et al. (1992) adapted the arguments to the case in which T is discrete. We follow Fleming (1992) and consider the arguments in the general case without any particular assumption about the nature of either T or S.

3.2 *Prentice's Criteria*

We note, first, that departures from the null hypotheses implicit in (1) are tested through the following two conditions:

$$f(T \mid Z) \neq f(T), \tag{2}$$

$$f(S \mid Z) \neq f(S). \tag{3}$$

Consider next the condition required for ($\Rightarrow$) to hold in (1). By definition, we have

$$f(T \mid Z) = \int f(T, S \mid Z)\, dS = \int f(T \mid S, Z) f(S \mid Z)\, dS. \tag{4}$$

By (1), $f(S \mid Z) = f(S)$ and

$$f(T \mid Z) = \int f(T \mid S, Z) f(S)\, dS. \tag{5}$$

If the condition

$$f(T \mid S, Z) = f(T \mid S) \tag{6}$$

holds, then (5) can be written

$$f(T \mid Z) = \int f(T \mid S) f(S)\, dS = \int f(T, S)\, dS = f(T) \tag{7}$$

and ($\Rightarrow$) holds in (1). Condition (6) implies that the full effect of treatment on the true endpoint is captured by the surrogate.

Consider now the condition required for ($\Leftarrow$) to hold in equation (1). If condition (6) holds, then (4) can be rewritten

$$f(T \mid Z) = \int f(T \mid S, Z) f(S \mid Z)\, dS$$
$$= \int f(T \mid S) f(S \mid Z)\, dS. \tag{8}$$

Similarly,

$$f(T) = \int f(T \mid S) f(S)\, dS. \tag{9}$$

Since $f(T \mid Z) = f(T)$, by substraction of (9) from (8),

$$\int f(T \mid S)[f(S \mid Z) - f(S)]\, dS = 0. \tag{10}$$

For a binary surrogate endpoint $S(0, 1)$, expression (10) reduces to

$$[f(T \mid S = 0) - f(T \mid S = 1)][f(S = 1 \mid Z) - f(S = 1)] = 0. \tag{11}$$

Hence, a sufficient condition for ($\Leftarrow$) to hold in (1) is that $f(T \mid S = 0) \neq f(T \mid S = 1)$, or

$$f(T \mid S) \neq f(T). \tag{12}$$

Condition (12) implies that the surrogate endpoint has prognostic value for the true endpoint. Prentice's criteria consist of the set of conditions (2), (3), (6), and (12). These conditions are necessary and sufficient to establish the validity of binary surrogate endpoints but not of more complex surrogate endpoints. The simplest counterexample is found by considering a multicategorical surrogate endpoint, as illustrated in Table 2.

4. Binary Endpoints

4.1 *Prentice's Criteria*

Operationally, a surrogate endpoint will be validated through tests of significance and estimation of appropriate parameters. Criteria (2), (3), and (12) can be verified using appropriate tests of significance. In general terms, if the effects were assessed through the parameters α, β, and γ as shown schematically on Figure 2, then criteria (2), (3), and (12) could be established through testing $\gamma = 0$, $\beta = 0$, and $\alpha = 0$, respectively. Criterion (6) raises a conceptual difficulty in that it must be established through testing a null hypothesis of the form $\beta_S \neq 0$. We return to this issue in Section 4.2.

Appropriate parameters α, β, β_S, and γ depend on the nature of the endpoints considered. In this section, we consider the simplest case in which the surrogate and the final endpoints are both binary. Section 5 is dedicated to normal endpoints.

A commonly used model for binary endpoints is logistic regression. Let the triplet (T_i, S_i, Z_i) represent the data for subject $i = 1, \ldots, n$. The effect of Z on T might be assessed through the logistic model

Table 2
Relationship between T (true endpoint), S (surrogate endpoint), and Z (treatment) in an artificial set of data for which $f(T \mid S) \neq f(T)$, $f(S \mid Z) \neq f(S)$, and $f(T \mid S, Z) = f(T \mid S)$ yet $f(T \mid Z) = f(T)$. Cell counts represent numbers of patients.

		Z	
S	T	0	1
0	0	40	120
	1	10	30
1	0	150	50
	1	150	50
2	0	30	50
	1	120	200

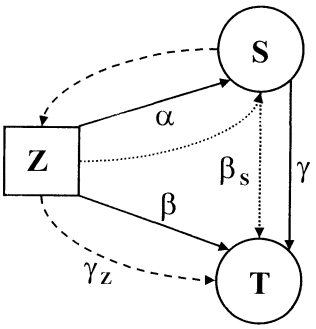

Figure 2. The associations between treatment (Z), a surrogate endpoint (S), and a true endpoint (T) are characterized by the three parameters α, β, and γ. Parameter β_S characterizes the effect of Z on T after adjustment for S. Parameter γ_Z characterizes the effect of S on T after adjustment for Z.

$$\ln\left(\frac{P(T_i = 1 \mid Z_i)}{P(T_i = 0 \mid Z_i)}\right) = \mu_{ZT} + \beta Z_i, \tag{13}$$

and the parameter of interest, β, is then the log odds ratio between Z and T, $\beta = \ln \mathrm{OR}_{ZT}$. Criterion (2) is satisfied if $\beta \neq 0$, i.e., if $\mathrm{OR}_{ZT} \neq 1$. Similarly, criteria (3) and (12) are satisfied if, respectively, $\alpha \neq 0$ ($\mathrm{OR}_{ZS} \neq 1$) and $\gamma \neq 0$ ($\mathrm{OR}_{ST} \neq 1$) in the following logistic models:

$$\ln\left(\frac{P(S_i = 1 \mid Z_i)}{P(S_i = 0 \mid Z_i)}\right) = \mu_{ZS} + \alpha Z_i, \tag{14}$$

$$\ln\left(\frac{P(T_i = 1 \mid S_i)}{P(T_i = 0 \mid S_i)}\right) = \mu_{ST} + \gamma S_i. \tag{15}$$

Criterion (6), on the other hand, is fulfilled if the full effect of Z on T is captured by S, in which case Z has no effect on T after adjustment for S. This criterion would be established by showing that $\beta_S = \delta = 0$ ($\mathrm{OR}_{ZT|S=0} = \mathrm{OR}_{ZT|S=1} = 1$) in the following logistic model:

$$\ln\left(\frac{P(T_i = 1 \mid Z_i, S_i)}{P(T_i = 0 \mid Z_i, S_i)}\right) = \mu_{ZT|S} + \beta_S Z_i + \gamma_Z S_i + \delta Z_i S_i, \tag{16}$$

where β_S is the effect of Z on T adjusted for S, γ_Z is the effect of S on T adjusted for Z, and δ is the three-way interaction. The surrogate is invalidated whenever δ or β_S are significant. Freedman et al. (1992) suggested that one should first assess whether δ can be dropped from the model. In that case, the two-way interaction model

$$\ln\left(\frac{P(T_i = 1 \mid Z_i, S_i)}{P(T_i = 0 \mid Z_i, S_i)}\right) = \mu_{ZT|S} + \beta_S Z_i + \gamma_Z S_i \tag{17}$$

can be used instead.

Note that, in model (17), it is possible to have $\beta_S = 0$ without having $\mathrm{OR}_{ZT|S=0} = \mathrm{OR}_{ZT|S=1} = 1$ and thus without criterion (6) being met. As a counterexample, consider the artificial data in Table 3.

For these data, $\beta_S \simeq 0$ but $\mathrm{OR}_{ZT|S=0} = 3.4$ and $\mathrm{OR}_{ZT|S=1} = 28.4$. Therefore, it is essential to first test the absence of a three-way interaction ($\delta = 0$) prior to testing the adjusted effect of treatment on the true endpoint ($\beta_S = 0$).

While logistic regressions have been used to obtain operational forms of the criteria, it should be noted that they can be obtained from log-linear models instead. For example, the marginal odds ratios are obtained from a log-linear model with only main effects (Z, S, T) (Agresti, 1990), whereas the adjusted odds ratios are obtained from the two-way interaction model (ZS, ZT, ST). Finally, if the interaction between Z and S is included as in (16), then model (ZST) should be fitted. While (17) is to be preferred over (16) because the latter contains the three-way interaction, its main drawback is that no closed-form expressions exist for the maximum likelihood estimators of the parameters, as is well known in the context of the corresponding log-linear model (ZS, ZT, ST).

Table 3
Relationship between T (true endpoint), S (surrogate endpoint), and Z (treatment) in an artificial set of data for which $\beta_S = 0$ yet $f(T \mid S, Z) \neq f(T \mid S)$. Cell counts represent numbers of patients.

		Z	
S	T	0	1
0	0	100	70
	1	200	830
1	0	400	477
	1	250	284

The marginal and conditional odds ratios can be expressed as follows by representing the data in the form of a $2 \times 2 \times 2$ contingency table with counts n_{zst} $(z, s, t = 0, 1)$:

$$\mathrm{OR}_{ST} = \frac{n_{+00}n_{+11}}{n_{+01}n_{+10}} = \exp\gamma, \tag{18}$$

$$\mathrm{OR}_{ZT} = \frac{n_{0+0}n_{1+1}}{n_{1+0}n_{0+1}} = \exp\beta, \tag{19}$$

$$\mathrm{OR}_{SZ} = \frac{n_{00+}n_{11+}}{n_{10+}n_{01+}} = \exp\alpha, \tag{20}$$

$$\mathrm{OR}_{ZT|S=0} = \frac{n_{000}n_{101}}{n_{100}n_{001}} = \exp\beta_S, \tag{21}$$

$$\mathrm{OR}_{ZT|S=1} = \frac{n_{010}n_{111}}{n_{110}n_{011}}, \tag{22}$$

$$\mathrm{OR}_{ZST} = \frac{n_{111}n_{100}n_{010}n_{001}}{n_{110}n_{101}n_{011}n_{000}} = \exp\delta. \tag{23}$$

Example

In order to show that criteria (2), (3), and (12) are fulfilled, the three contingency tables giving the relationships between Z, S, and T need to be considered (Table 1b). The odds ratios in these 2×2 tables are all significantly different from unity at the 5% level: in the (Z, S) table, $\mathrm{OR}_{ZS} = 2.01$ $(\chi^2 = 5.60,\ P = 0.018)$; in the (Z, T) table, $\mathrm{OR}_{ZT} = 1.93$ $(\chi^2 = 4.97,\ P = 0.026)$; and in the (S, T) table, $\mathrm{OR}_{ST} = 19.33$ $(\chi^2 = 74.91,\ P < 0.001)$. Hence, criteria (2), (3), and (12) are all satisfied. Curiously, in this disease, the effect of interferon-α is actually harmful since significantly more patients have lost lines of vision on interferon-α than on placebo, both at 6 months and at 1 year after starting treatment (Pharmacological Therapy for Macular Degeneration Study Group, 1997). The relationship between S and T is very strong (as expected), the odds of losing at least 3 lines of vision by 1 year being almost 20 to 1 when at least 2 lines were lost at 6 months. Note that, for a perfect surrogate, the (S, T) table would have off-diagonal elements equal to zero, so that $\mathrm{OR}_{ST} = \infty$ and $\mathrm{OR}_{ZS} = \mathrm{OR}_{ZT}$. In this example, the (Z, S) table and the (Z, T) table are almost identical so that $\mathrm{OR}_{ZS} \simeq \mathrm{OR}_{ZT}$ even though S is not a perfect surrogate for T.

In order to show that criterion (6) is also satisfied, the three-way contingency table of Z, T, and S needs to be considered (Table 1a). The three contingency tables shown in Table 1b are the two-way margins of Table 1a. There is no evidence of a three-way interaction (likelihood ratio test statistic of 0.39 on 1 d.f.). Therefore, the choice of model (17) is justified. Using this model, the odds ratio between T and Z, adjusted for S, is not significantly different from unity: $\mathrm{OR}_{ZT|S} = 1.44$ $(\chi^2 = 0.93, P = 0.34)$. The fact that this adjusted χ^2 test is not statistically significant provides evidence that some of the effect of Z on T is mediated through S. However, the nonsignificance of this test fails to provide evidence that the full effect of Z on T is mediated through S. We now discuss this issue further.

4.2 *Freedman's Proportion Explained*

Freedman et al. (1992) argued that criterion (6) raises a conceptual difficulty in that it requires the statistical test for treatment effect on the true endpoint to be nonsignificant after adjustment for the surrogate. Hence, criterion (6) is useful to reject a poor surrogate endpoint (when the statistical test for treatment effect on the true endpoint remains statistically significant after adjustment for

the surrogate), but it is inadequate to validate a good surrogate endpoint, for failing to reject the null hypothesis may be due merely to insufficient power. Note that this observation justifies the use of large numbers of observations for the validation of surrogate endpoints. Even if lack of power were not an issue, the statistical significance of the adjusted and unadjusted tests do not adequately quantify the impact of the surrogate on the analysis of the true endpoint. Since it cannot be proven that the effect of treatment on the true endpoint is fully captured by the surrogate, Freedman et al. (1992) proposed focusing attention on the proportion of the treatment effect explained by the surrogate. A good surrogate is one that explains a large proportion of that effect. Schatzkin et al. (1990), in their discussion of the validation of intermediate endpoints in cancer, observe that a valid surrogate endpoint for screening purposes is one for which the attributable proportion (the proportion of cases with the disease that can be attributed to the intermediate endpoint) is close to one. Freedman's criterion is similar in spirit but concentrates on the proportion of the treatment effect that can be explained by the surrogate. Let $PE(T, S, Z)$ stand for the proportion of the effect of Z on T that can be explained by S, or simply the proportion explained. An estimate of $PE(T, S, Z)$ is as follows:

$$PE(T, S, Z) = \frac{\beta - \beta_S}{\beta} = 1 - \frac{\beta_S}{\beta}, \tag{24}$$

where β and β_S are the estimates of the effect of Z on T without and with adjustment for S (respectively, from models (13) and (17)). For reasons given in Section 4.1, one might want to replace (17) by (16). In both cases, the proportion explained is large if β_S is small in comparison to β. Prentice's criterion (6) requires that $\beta_S = 0$, or equivalently $PE = 1$. A surrogate endpoint for which $PE < 1$ explains only part of the treatment effect on the true endpoint and may therefore be called an incomplete surrogate (Choi et al., 1993).

Note that it is possible for PE to be greater than one if β_S and β have opposite signs, i.e., if the adjustment for S changes the direction of the effect of Z on T. This points to a conceptual problem with the PE, which will be discussed further in Section 4.4.

PE being the ratio of two parameters, its confidence limits can be calculated using Fieller's theorem or the delta method. Using Fieller's theorem, which is generally preferable (Herson, 1975), the $(1 - \alpha)\%$ confidence limits of $PE(T, S, Z)$ are given by

$$1 - \frac{A \pm \sqrt{A^2 - BC}}{B}, \tag{25}$$

where

$$\begin{aligned} A &= \beta\beta_S - Z_\alpha^2 \operatorname{cov}(\beta, \beta_S), \\ B &= \beta^2 - Z_\alpha^2 \operatorname{var} \beta, \\ C &= \beta_S^2 - Z_\alpha^2 \operatorname{var} \beta_S, \end{aligned}$$

and Z_α is the $100(1 - \alpha/2)$ percentile of the normal distribution (or, if n were not large, of the Student's t-distribution with $n - 1$ degrees of freedom). The variances of the parameter estimates ($\operatorname{var} \beta$ and $\operatorname{var} \beta_S$) are easily obtained by fitting the unadjusted and adjusted models (13) and (17), respectively.

To determine the covariance between β and β_S, we follow the suggestion of Freedman et al. (1992). The design for the unadjusted model is the $n \times 2$ matrix $\mathbf{X}$ with ith row $(1, Z_i)$, whereas the design for the adjusted one is the $n \times 3$ matrix $\mathbf{X}_S$ with ith row $(1, Z_i, S_i)$. Denote the predicted probability for subject i from the unadjusted (adjusted) regression π_i (π_{iS}), which are

$$\begin{aligned} \pi_i &= \frac{\exp(\mu_{ZT} + \beta Z_i)}{1 + \exp(\mu_{ZT} + \beta Z_i)}, \\ \pi_{iS} &= \frac{\exp(\mu_{ZT|S} + \beta_S Z_i + \gamma S_i)}{1 + \exp(\mu_{ZT|S} + \beta_S Z_i + \gamma S_i)}. \end{aligned}$$

Denote $\mathbf{V} = \operatorname{diag}\{\pi_i(1 - \pi_i)\}$ and $\mathbf{V}_S = \operatorname{diag}\{\pi_{iS}(1 - \pi_{iS})\}$. The 2×3 covariance matrix between the parameters of both models is then estimated by

$$\hat{\mathbf{W}} = (\mathbf{X}^{\mathrm{T}} \hat{\mathbf{V}} \mathbf{X})^{-1} (\mathbf{X}^{\mathrm{T}} \hat{\mathbf{V}}_S \mathbf{X}_S)(\mathbf{X}_S^{\mathrm{T}} \hat{\mathbf{V}}_S \mathbf{X}_S)^{-1},$$

and thus $\operatorname{cov}(\beta, \beta_s)$ is estimated by the $(2, 2)$ element of $\hat{\mathbf{W}}$.

Freedman et al. (1992) observe that, if the treatment effect on the true endpoint is small and if, in addition, the number of observations is not large (as is the case in most randomized clinical

trials), the confidence interval of PE will be wide, so there will be substantial uncertainty about the proportion of the effect that is truly mediated by the surrogate. This observation justifies the use of large randomized trials or a meta-analysis of many related trials to validate surrogate endpoints. Even when large numbers of observations are available, however, the denominator of the proportion explained (the effect of treatment on the true endpoint) will be estimated with little precision, for otherwise the need for a surrogate endpoint would no longer exist. Therefore, the proportion explained will generally be too poorly estimated to be of much practical value.

Example (Continued)

In our example (Table 1), $\text{OR}_{ZT} = 1.93$ and $\text{OR}_{ZT|S} = 1.44$. Hence, $PE = (0.66 - 0.36)/0.66 = 0.45$ so that less than half of the effect of interferon-α on the loss of vision at 1 year is explained by the loss of vision at 6 months. This may seem surprisingly little given the strength of the relationship between S and T and considering the fact that $\text{OR}_{ZS} \simeq \text{OR}_{ZT}$, which might be interpreted (wrongly) as suggesting that S and T are interchangeable. The 95% confidence interval of PE is given by $[-0.30, 4.35]$, and thus the true PE might be anywhere from zero to well over 100%, an interval far too wide to be of practical relevance. Had all the numbers in Table 1 been multiplied by a factor of 10 (for a total of 1900 patients instead of 190), the 95% confidence interval of PE would have been $[0.22, 0.75]$, which provides a much more precise estimate of the proportion of the treatment effect truly explained by the surrogate. In particular, in this hypothetical situation, it would be highly likely that S explains less than three quarters of the effect of Z on T, and we might thus conclude that the loss of two lines of vision at 6 months is an incomplete surrogate for the loss of three lines of vision at 1 year.

It is interesting to note that the point estimate and the confidence interval of PE depend on the data only through the two-way classifications (the pairs (Z, T), (Z, S), and (T, S)). In other words, all three-way tables compatible with the three two-way tables will yield exactly the same proportion explained. This property is due to the fact that the logistic regression model for T (17) involves no interaction term between Z and S.

4.3 *Relative Effect*

For a surrogate endpoint to be useful in practice, the investigators must be able to predict the effect of treatment on the true endpoint based on the observed effect of treatment on the surrogate. Thus, we need to relate the magnitude of the treatment effects on the true and surrogate endpoints (Boissel et al., 1992). A new treatment could then be tested through its effect on the surrogate endpoint and declared efficacious if its predicted effect on the true endpoint were sufficiently large to be of clinical interest (Ellenberg, 1991).

Let $RE(T, S, Z)$ stand for the effect of Z on T relative to that of Z on S, or simply relative effect. An intuitively appealing way of defining $RE(T, S, Z)$ is as follows:

$$RE(T, S, Z) = \frac{\beta}{\alpha}, \tag{26}$$

where β and α are estimated from (13) and (14), respectively. Intuitively, RE is the slope of a regression line between $\ln(\text{OR}_{ZT})$ and $\ln(\text{OR}_{ZS})$, which has been suggested by other authors (A'Hern, Ebbs, and Baum, 1988). Given that $\ln(\text{OR}) \simeq 1 - OR$, RE is also in first approximation equal to the ratio of the odds reduction on the true endpoint over the odds reduction on the surrogate endpoint. RE will be equal to one if the effects of Z on T and on S are of identical magnitude (in terms of the odds ratio). For reasons that will become clearer in Section 5, we call a surrogate for which $RE = 1$ perfect at the population level. In practice, RE will tend to be less than one if the true endpoint is more difficult to affect than the surrogate endpoint. A surrogate endpoint for which the relative effect is small will result in small changes in the true endpoint even if treatment achieves large changes in the surrogate endpoint. Such a surrogate may nonetheless be quite useful if it predicts clinically worthwhile effects on the true endpoint. The precision with which RE can be estimated will also be relevant to predict the effect of Z on T based on an observed effect of Z on S.

In order to estimate $RE(T, S, Z)$, the parameters β and α can be estimated from the two logistic regressions (13) and (14). For the estimation of $\text{cov}(\beta, \alpha)$, we suggest considering a bivariate model in which the effects of Z on T and S are estimated jointly. Such a model can be built from supplementing the marginal logistic regressions (13) and (14) with a measure for the association between S and T (given Z). A common choice with binary data is the odds ratio

$$\psi_i = \frac{P(S_i = 0, T_i = 0 \mid Z_i)P(S_i = 1, T_i = 1 \mid Z_i)}{P(S_i = 0, T_i = 1 \mid Z_i)P(S_i = 1, T_i = 0 \mid Z_i)}. \tag{27}$$

For convenience, denote the probabilities by $\pi_{ijk} = P(S_i = j, T_i = k \mid Z_i)$ $(j, k = 0, 1)$. Summing over a variable is indicated by replacing the corresponding subscript with $+$. Then the model can be written in terms of three link functions for each individual as

$$\eta_{i1} = \ln\left(\frac{\pi_{i+1}}{1-\pi_{i+1}}\right) = \mu_{ZT} + \beta Z_i, \tag{28}$$

$$\eta_{i2} = \ln\left(\frac{\pi_{i1+}}{1-\pi_{i1+}}\right) = \mu_{ZS} + \alpha Z_i, \tag{29}$$

$$\eta_{i3} = \ln\psi_i = \ln\left(\frac{\pi_{i11}(1-\pi_{i1+}-\pi_{i+1}+\pi_{i11})}{(\pi_{i1+}-\pi_{i11})(\pi_{i+1}-\pi_{i11})}\right) = \mu_{STZ} + \rho Z_i. \tag{30}$$

For notational convenience, group probabilities and links into vectors $\boldsymbol{\pi}_i = (\pi_{i1+}, \pi_{i+1}, \pi_{i11})^{\mathrm{T}}$ and $\boldsymbol{\eta}_i = (\eta_{i1}, \eta_{i2}, \eta_{i3})^{\mathrm{T}}$, respectively. Other choices for the marginal regression and association link functions are possible (Zhao and Prentice, 1990), but the particular choice presented here has been studied in detail by Dale (1986) and Molenberghs and Lesaffre (1994, 1997). It has logistic margins and extends naturally to ordinal outcomes.

Maximum likelihood estimates for the parameter vector $\boldsymbol{\gamma} = (\mu, \beta, \alpha, \rho)^{\mathrm{T}}$ are found by solving the score equations

$$\mathbf{U}(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \mathbf{X}_i^{\mathrm{T}} \mathbf{D}_i^{\mathrm{T}} \mathbf{V}_i^{-1} \mathbf{E}_i = \mathbf{0}, \tag{31}$$

where

$$\begin{aligned}
\mathbf{X}_i &= \left(\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\gamma}}\right) = (1,1,1) \otimes (1, Z_i), \\
\mathbf{D}_i &= \left(\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\pi}_i}\right)^{-1}, \\
\mathbf{V}_i &= \mathrm{cov}(S_i, T_i, S_i T_i), \\
\mathbf{E}_i &= \begin{pmatrix} S_i - \pi_{i1+} \\ T_i - \pi_{i+1} \\ S_i T_i - \pi_{i11} \end{pmatrix}.
\end{aligned}$$

All quantities needed to evaluate $U(\boldsymbol{\gamma})$ are determined from $\boldsymbol{\gamma}$. Indeed, $\boldsymbol{\gamma}$ determines $\boldsymbol{\eta}_i$,

$$\begin{aligned}
\pi_{i1+} &= \frac{\exp(\eta_{i1})}{1+\exp(\eta_{i1})}, \\
\pi_{i+1} &= \frac{\exp(\eta_{i2})}{1+\exp(\eta_{i2})},
\end{aligned}$$

and

$$\pi_{i11} = \begin{cases} \dfrac{1 + (p_{i1+} + p_{i+1})(\psi_i - 1) - S(p_{i1+}, p_{i+1}, \psi_i)}{2(\psi_i - 1)} & \text{if } \psi_i \neq 1, \\ p_{i1+} p_{i+1} & \text{if } \psi_i = 1, \end{cases}$$

with

$$S(q_1, q_2, \psi) = \sqrt{[1 + (q_1 + q_2)(\psi - 1)]^2 + 4\psi(1-\psi)q_1 q_2}.$$

The above expression was studied by Plackett (1965) and Mardia (1970). In order to estimate the covariance matrix of $\boldsymbol{\gamma}$, one calculates the matrix of second derivatives of the log-likelihood, i.e., the derivative of $\mathbf{U}(\boldsymbol{\gamma})$, and replaces $\boldsymbol{\gamma}$ by its maximum likelihood estimate $\hat{\boldsymbol{\gamma}}$. Details can be found in Molenberghs and Lesaffre (1994).

Having estimated β, α, $\mathrm{var}(\beta)$, $\mathrm{var}(\alpha)$, and $\mathrm{cov}(\beta, \alpha)$, RE can be estimated along with its confidence limits using Fieller's theorem (see Section 4.2).

Example (Continued)

In our example, the effect of interferon-α on the loss of two lines of vision at 6 months is almost identical to the effect on the loss of three lines of vision at 1 year. Numerically, $RE(T, S, Z) = \ln(1.93)/\ln(2.01) = 0.94$. Thus, the effect of Z on T is only slightly less than that of Z on S, so

that a reduction in the odds of a loss of two lines of vision at 6 months will translate into a similar reduction in the odds of a loss of three lines of vision at 1 year. However, the 95% confidence interval of RE is given by $[0.20, 3.15]$, and thus the true effect of Z on T might be much smaller, or indeed much larger, than that of Z on S. The confidence limits of RE will be wide when estimated from a typical (medium-sized) clinical trial. Had all the numbers in Table 1 been multiplied by a factor of 10, the 95% confidence interval of RE would have been $[0.73, 1.20]$, which is narrow enough to permit a useful prediction of the treatment effect that may be expected on the true endpoint. Just as for PE, the point estimate of RE and its confidence limits depend only on the two-way classifications. The reason is similar: the marginal logistic regressions describe the effects of Z on T and on S, while the odds ratio accounts for the association between S and T. Should we let the association depend on Z, then a three-way effect would be included. The point estimate for the RE would still be equal to 0.94, but the confidence interval would depend (mildly) on the three-way classification.

4.4 *Adjusted Association*

The marginal association between S and T is assessed by means of the γ parameter in model (15). It is of interest to also derive the association between S and T after adjustment for the treatment Z. This can be done in three identical ways. First, the ST interaction in the two-way log-linear model (ZS, ZT, ST) can be considered. Secondly, exactly the same quantity follows as the log odds ratio $\ln \psi_i$ from (27). Finally, this number is also equal to the coefficient γ_Z in model (17). When γ_Z is large (infinity), it means that the surrogate and true endpoints are very similar (the same), possibly up to a deterministic transformation. This transformation is a function of the respective treatment effects α and β, and equality obtains only if $\gamma_Z \to \infty$ and $\alpha = \beta$, i.e., $RE = 1$. This suggests the following definition. We call a surrogate for which $\gamma_Z = \infty$ perfect at the individual level. When $\gamma_Z = \infty$, the logit in (17) degenerates to a step function so that at least one of the following condition holds: $S = 0$ implies $T = 0$ or $S = 1$ implies $T = 1$ with probability one (at least one off-diagonal cell in the cross-classification of S and T is equal to zero). Both conditions will hold simultaneously if and only if $RE = 1$ as well (the two off-diagonal cells are equal to zero).

The pair (γ_Z, RE) usefully complements the PE. Indeed, γ_Z describes the subject-specific association between the surrogate and true endpoints, while RE links them at the population-averaged level. A perfect surrogate is one for which $\gamma_Z = \infty$ (the surrogate is perfect at the individual level) and $RE = 1$ (the surrogate is perfect at the population level). The case of normal endpoints, discussed in the next section, will shed further light on the nature of the proportion explained and will reinforce the conclusions reached here.

Example (Continued)

For the visual acuity data, $\gamma_Z = \ln \psi = 2.92$, leading to an odds ratio of 18.53 with confidence interval $[8.86, 38.77]$, clearly pointing to a strong dependence between the surrogate and the true endpoint after adjustment for treatment.

5. Normally Distributed Endpoints

We now treat the case where both the surrogate and the true endpoint are continuous and jointly normally distributed. Specifically, we assume the following model in the standardized endpoints:

$$S_i = \mu_S + \alpha Z_i + \varepsilon_{Si}, \tag{32}$$

$$T_i = \mu_T + \beta Z_i + \varepsilon_{Ti}, \tag{33}$$

where μ_S and μ_T are fixed intercepts and α and β are the fixed effects of the treatment Z on the surrogate and true endpoints, respectively. Further, ε_{Si} and ε_{Ti} are correlated error terms assumed to satisfy

$$\begin{pmatrix} \varepsilon_{Si} \\ \varepsilon_{Ti} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right). \tag{34}$$

We can verify Prentice's first two criteria and calculate PE and RE from (32)–(34). In order to verify Prentice's third criterion, which concerns the marginal association between T and S, we need the regression of T on S

$$\mathrm{E}(T_i) = \mu_{ST} + \gamma S_i.$$

For the proportion explained, the regression of T on Z, after correction for S, is necessary. This regression follows from (34) by constructing the appropriate conditional distribution as

$$(T_i \mid Z_i, S_i) \sim \mathrm{N}[(\mu_T - \rho\mu_S) + (\beta - \rho\alpha)Z_i + \rho S_i, 1 - \rho^2]. \tag{35}$$

Thus, the adjusted treatment effect equals $\beta_S = \beta - \rho\alpha$ and then the adjusted association $\gamma_Z = \rho$, implying

$$PE = \frac{\alpha}{\beta}\rho = \frac{\gamma_Z}{RE}. \tag{36}$$

It follows immediately that PE is a mixture of two aspects of the model: fixed effects (through RE) and random components (through γ_Z). The interpretation of PE may therefore be problematic. For example, PE vanishes either if $\alpha = 0$ (the treatment has no effect on the surrogate) or $\rho = 0$ (given treatment, the surrogate and the true endpoint are uncorrelated). This suggests that PE, being a composite quantity, is less attractive to assess surrogacy than the couple RE and γ_Z instead. RE connects the treatment effects at the population-averaged level (fixed-effects model terms), while γ_Z connects them at the individual specific level (random-effects model terms). γ_Z captures the association between the surrogate and the true endpoint after correction for the treatment effect. Thus, in addition to Prentice's third criterion, which states that there should be a strong unadjusted association between S and T, it appears to be fruitful to ascertain a strong adjusted association γ_Z.

We can now propose similar definitions as in the binary case and call a surrogate for which $RE = 1$ perfect at the population level and one for which $\gamma_Z = \rho = 1$ perfect at the individual level. RE would be useful in future trials to predict the treatment effect on the true endpoint from the corresponding effect on the surrogate. On the other hand, even if either α or β were small, a surrogate for which γ_Z is close to 1 would be useful to predict the outcome at the individual level, not only in trial conditions but also in clinical practice.

We conclude this section by returning to some difficulties with PE. First, unlike γ_Z, PE does not necessarily lie between zero and one. Second, even in the case of a perfect correlation ($\gamma_Z = 1$), PE is not necessarily equal to one. Indeed, if $\gamma_Z = \rho = 1$, $\beta_S = \beta - \alpha$ and $PE = 1$ if and only if $\beta = \alpha$ (i.e., if $RE = 1$). The case where $PE = 1$ is special only insofar that then $T_i = S_i$, whereas they are only so up to a deterministic, treatment-dependent term when $PE \neq 1$. Finally, since the PE is a composite quantity it is likely to produce wide confidence intervals for the PE but smaller ones for the other quantities. RE is obtained from a joint model for S and T given Z and γ_Z is obtained from the regression of T on Z and S simultaneously. For the PE on the other hand, two models for T are required, one on Z only and one on both Z and S simultaneously, implying an *ad hoc* derivation of the confidence interval of PE.

Example

The analyses were carried out using the SAS procedure MIXED (Littell et al., 1996). The relevant regression coefficients (and their standard errors) are $\hat{\alpha} = 2.83$ (S.E. $= 1.86, P = 0.13$), $\hat{\beta} = 4.12$ (S.E. $= 2.32, P = 0.078$), and $\hat{\gamma} = 0.95$ (S.E. $= 0.06, P < 0.0001$). Thus, there is little evidence for an effect of Z on either endpoint but overwhelming evidence that the surrogate is strongly correlated with the true endpoint. We find a proportion explained of 0.65 with confidence interval $[-0.22, 1.51]$ and a relative effect of 1.45 with confidence interval $[-0.48, 3.39]$. It is noteworthy that, while the confidence intervals for PE and RE are hopelessly wide, $\gamma_Z = 0.75$ with confidence interval $[0.69, 0.82]$, which implies that a very large part of the variability of the surrogate is shared with the true endpoint.

6. Discussion

Several quantities can be considered in the validation of a candidate surrogate endpoint (S) with respect to a true endpoint (T) for a specific treatment (Z) (Table 4):

- The first three quantities are the parameters characterizing the pairwise associations: between Z and S (α), between Z and T (β), and between S and T (γ). The first three operational criteria of Prentice require that these three parameters be significantly different from zero. The magnitude of these associations is not of direct relevance in Prentice's validation process. However, the number of observations available must clearly be sufficiently large for each of these tests to have good power to detect the corresponding effect. In practice, the numbers of observations available from individual clinical trials is often quite sufficient to establish the effect of S on T but insufficient to establish the effect of Z on S or of Z on T (particularly if T is a far distant endpoint, a situation not covered in the present paper). It may in fact be

Table 4
The quantities of interest for the validation of a surrogate endpoint (T, true endpoint; S, surrogate endpoint; Z, treatment; $f(\cdot)$, density function; PE, proportion explained; RE, relative effect.) See text for details.

Quantity of interest	Estimate	Test	See Section
Effect of treatment on true endpoint	β	H_0: $f(T \mid Z) = f(T)$	4.1
Effect of treatment on surrogate endpoint	α	H_0: $f(S \mid Z) = f(S)$	4.1
Effect of surrogate on true endpoint	γ	H_0: $f(T \mid S) = f(T)$	4.1
Proportion of treatment effect on true endpoint explained by surrogate	$PE = \frac{\beta - \beta_S}{\beta}$		4.2
Effect of treatment on true endpoint relative to that on surrogate endpoint	$RE = \frac{\beta}{\alpha}$		4.3
Adjusted effect of surrogate on true endpoint	γ_Z		4.4

desirable to relax the requirement of the significance of the effect of Z on T in the validation process, for the very motivation of surrogate endpoints is to temporarily replace the test of the effect of Z on T by that of Z on S! However, it is clearly not sufficient to establish the significance of the effect of S on T in order to validate S as a surrogate for T (Boissel et al., 1992). For instance, in patients with AIDS, it has been shown repeatedly that a drop in CD4 counts negatively affects survival (Jacobson, Bacchetti, and Kolokathis, 1991), and yet such a drop does not qualify as a valid surrogate for survival (Ellenberg, 1991).

- Prentice's fourth criterion states that Z has no further effect on T after accounting for S. Freedman et al. (1992) argued that this criterion cannot be verified as it implies proving the null hypothesis and proposed instead to estimate the proportion explained (PE), i.e., the proportion of the effect of Z on T that is explained by S, and its confidence interval. A surrogate is acceptable, according to Freedman et al. (1992), if the lower confidence limit of PE is larger than some proportion, say 0.5 or 0.75. In this case, a large proportion of the effect of treatment on the true endpoint is explained by the surrogate. However, the confidence limits around PE will generally be wide because this quantity is a ratio, the denominator of which is estimated with substantial error. Freedman et al. (1992) argues that the effect of Z on T should exceed four standard errors for PE to be estimated with reasonable accuracy. As pointed out above, the need for a surrogate may be questioned when such highly significant treatment effects have already been observed on the true endpoint.
- The fifth quantity proposed in the present paper is the relative effect (RE), i.e., the effect of Z on T relative to that of Z on S. A surrogate for which $RE = 1$ is said to be perfect at the population level, as any effect of treatment on the surrogate predicts exactly the same effect on the true endpoint. To be of practical value, RE must be estimated with good precision, which implies that the number of observations should be large. Clearly, in order to be meaningful, the validation process will have to be based on large-scale randomized evidence. Such evidence is not always available from individual trials, and therefore meta-analyses based on individual patient data from several randomized trials will often be the best way to validate a surrogate endpoint.
- The sixth quantity proposed in the present paper, γ_Z, is the association between S and T after adjustment for Z. A surrogate is said to be perfect at the individual level when this association is perfect, which corresponds to $\gamma_Z = \infty$ for binary endpoints and $\gamma_Z = 1$ for normal endpoints. A perfect surrogate is perfect both at the individual and at the population levels, in which case S and T are identical up to a deterministic transformation ($S \equiv T$).

The validation approach outlined in this paper departs from the definition of surrogate endpoints set forth by Prentice (1989). Strictly speaking, the Prentice definition would require a large number of independent experiments to be available. A surrogate would be valid if and only if the tests for the effect of treatment simultaneously rejected or failed to reject the null hypothesis for both the surrogate and the true endpoint in all experiments. Such an approach places undue emphasis on hypothesis tests. Moreover, the validation criteria suggested by Prentice are equivalent to his definition only in the case of binary endpoints. The approach presented here requires estimation of several quantities, assuming that individual patient data are available from one or several

randomized trials testing the effect of a treatment on both the surrogate and the final endpoints. In practice, the validation we suggest will typically be carried out on data pertaining to some well-known treatments in the hope of using the surrogate endpoint thus validated to test some other (new) treatment. Doing so implicitly assumes that the relationships between the treatment-dependent parameters (α, β, and β_S) will be essentially the same for the new treatment as for the treatments on which data are available. This may or may not be the case, and therefore even a surrogate endpoint that has been formally validated may be questioned if used to test the effects of new treatments, particularly if the biological mode of action of these new treatments is substantially different from that of their predecessors.

If the use of a surrogate is accepted, then it is essential that investigators be able to predict the effect of a treatment on the true endpoint based on its effect on the surrogate as well as to quantify the statistical uncertainty of this prediction (letting alone its biological uncertainty). Our proposal is to estimate the relative effect (RE) in order to perform such a prediction. The two major limitations of RE are that its confidence limits may be too wide to permit clinically useful predictions and that its value may depend on the value of α. In other words, since RE is the slope of a regression line between α and β, the linearity of this regression may be questioned. RE might change with, e.g., the strength of the association between Z and the outcomes themselves. Arguably, verification of these assumptions from a sufficient number of independent trials would be necessary.

Our two examples, coming from the same clinical trial, are limited in scope as they only deal with situations in which the two endpoints are of the same nature (binary or continuous). They are somewhat contrived as they actually represent repeated measures of the same variable over time. However, they illustrate the need for large numbers of observations for the validation procedure to be informative about the treatment-related parameters (particularly RE and PE). These simple situations also shed light on the concepts of relative effect (RE) and adjusted association (γ_Z) in addition to that of proportion explained (PE). We have outlined some problems of interpretation with PE as well as some practical limitations that seriously limit the interest of PE in practice. One advantage of PE is that it can easily be defined regardless of the nature of the endpoints considered (Lin, Fleming, and De Gruttola, 1997). In contrast, the interpretation of RE and γ_Z is model dependent, which may be considered a drawback. This difficulty, however, reveals the complexity of the problem, since the notion of a perfect surrogate is hard to define when the surrogate and the true endpoints are not of the same nature. If, e.g., the surrogate endpoint was binary and the true endpoint continuous, the surrogate could never be a perfect surrogate for the true endpoint, except in degenerate cases, because it would not account for all the variability in the true endpoint.

The approach proposed here does not cover the important cases of time-related endpoints, which will be the subject of a separate paper. More complex situations in which the endpoints are repeatedly measured over time also require further work. The concepts discussed in this paper, once extended to these more general situations, could provide decision makers with the statistical tools they need, in addition to all relevant biological and medical factors, in order to properly assess surrogate endpoints in clinical research (Temple, 1995).

Acknowledgements

The authors are grateful to the Pharmacological Therapy for Macular Degeneration Study Group and to F. Hoffman-LaRoche for permission to use their data and to Dr Terry Neeman and to anonymous referees for very insightful comments. The software used to perform maximum likelihood estimation for binary endpoints was written in GAUSS and is freely available from the second author on request.

Résumé

Prentice (Statistics in Medicine, 1989), ainsi que Freedman et al. (Statistics in Medicine, 1992), ont étudié la validation des critères de substitution (surrogate endpoints). Nous généralisons leurs propositions dans les cas où le critère de substitution et le critère réel sont soit tous deux des variables binaires, soit tous deux des variables gaussiennes. On notera T et S les variables aléatoires correspondant respectivement au critère réel et au critère réel et au critère de substitution et Z la variable indicatrice du groupe de traitement. Les conditions de Prentice sont satisfaites si on a simultanément: Z a un effet significatif sur T et S; S a un effet significatif sur T; Z n'a pas d'effet sur T à S fixé. Freedman a assoupli la dernière condition en estimant PE, la proportion de l'effet de Z sur T expliquée par S et en demandant que la borne inférieure de l'intervalle de confiance de PE soit plus grande qu'un certain seuil (par exemple 0.5 ou 0.75). Cette dernière condition ne peut

être établie que dans le cas rare où le traitement a un effet massif sur le critère réel. Nous soutenons que deux autres paramètres doivent être pris en compte lors de la validation d'un critère de substitution: *RE,* qui est l'effet de Z sur T relativement à celui de Z sur S; γ_Z, l'association entre S et T après ajustement pour Z. Un critère de substitution est considéré comme parfait au niveau individuel quand il y a une association parfaite entre lui et le critère réel après ajustement pour le traitement ($\gamma_Z = \infty$ dans le cas binaire; $\gamma_Z = 1$ dans le cas gaussien). Un critère de substitution est considéré comme parfait au niveau de la population si $RE = 1$. Un critère de substitution parfait remplit ces deux conditions, auquel cas S et T sont identiques à une transformation près. Nous utilisons le théorème de Fieller pour estimer *PE, RE* et leurs intervalles de confiance respectifs. Pour le cas binaire, nous utilisons des modèles de régression logistique et le modèle global d'odds-ratios selon Dale (Biometrics, 1986). Pour le cas gaussien, nous utilisons des modèles linéaires mixtes. Nous montrons finalement que la mise en œuvre de la validation de critères de substitution nécessite de grands effectifs.

References

Agresti, A. (1990). *Categorical Data Analysis.* New York: Wiley.

A'Hern, R. P., Ebbs, S. R., and Baum, M. B. (1998). Does chemotherapy improve survival in advanced breast cancer? A statistical overview. *British Journal of Cancer* **57,** 615–618.

Boissel, J. P., Collet, J. P., Moleur, P., and Haugh, M. (1992). Surrogate endpoints: A basis for a rational approach. *European Journal of Clinical Pharmacology* **43,** 235–244.

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

Choi, S., Lagakos, S., Schooley, R. T., and Volberding, P. A. (1993). CD4+ lymphocytes are an incomplete surrogate marker for clinical progression in persons with asymptomatic HIV infection taking zidovudine. *Annals of Internal Medicine* **118,** 674–680.

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

De Gruttola, V. and Tu, X. M. (1995). Modelling progression of CD-4 lymphocyte count and its relationship to survival time. *Biometrics* **50,** 1003–1014.

De Gruttola, V., Wulfsohn, M., Fischl, M. A., and Tsiatis, A. (1993). Modelling the relationship between survival and CD4 lymphocytes in patients with AIDS and AIDS-related complex. *Journal of Acquired Immune Deficiency Syndromes* **6,** 359–365.

De Gruttola, V., Fleming, T. R., Lin, D. Y., and Coombs, R. (1997). Validating surrogate markers—Are we being naive? *Journal of Infectious Diseases* **175,** 237–246.

Ellenberg, S. S. (1991). Surrogate endpoints in clinical trials: Getting closer to identifying markers for survival in AIDS. *British Medical Journal* **302,** 63–64.

Ellenberg, S. S. and Hamilton, J. M. (1989). Surrogate endpoints in clinical trials: Cancer. *Statistics in Medicine* **8,** 405–413.

Fleming, T. R. (1992). Evaluating therapeutic interventions: Some issues and experiences (with discussion). *Statistical Science* **7,** 428–456.

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

Fleming, T. R., Prentice, R. L., Pepe, M. S., and Glidden, D. (1994). Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. *Statistics in Medicine* **13,** 955–968.

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

Herson, J. (1975). Fieller's theorem versus the delta method for significance intervals for ratios. *Journal of Statistical Computing and Simulation* **3,** 265–274.

Jacobson, M. A., Bacchetti, P., and Kolokathis, A. (1991). Surrogate markers for survival in patients with AIDS and AIDS related complex treated with zidovudine. *British Medical Journal* **302,** 73–78.

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

Lin, D. Y., Fischl, M. A., and Schoenfeld, D. A. (1993). Evaluating the role of CD4-lymphocyte change as a surrogate endpoint in HIV clinical trials. *Statistics in Medicine* **12,** 835–842.

Lin, D. Y., Fleming, T. R., and De Gruttola, V. (1997). Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in Medicine* **16,** 1515–1527.

Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. (1996). *SAS System for Mixed Models.* Cary, North Carolina: SAS Institute.

Machado, S. G., Gail, M. H., and Ellenberg, S. S. (1990). 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* **3,** 1065–1073.

Mardia, K. V. (1970). *Families of Bivariate Distributions.* London: Griffin.

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

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

Plackett, R. L. (1965). A class of bivariate distributions. *Journal of the American Statistical Association* **60,** 516–522.

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

Schatzkin, A., Freedman, L. S., Schiffman, M. H., and Dawsey, S. M. (1990). Validation of intermediate end points in cancer research. *Journal of the National Cancer Institute* **82,** 1746–1752.

Temple, R. J. (1995). A regulatory authority's opinion about surrogate endpoints. In *Clinical Measurement in Drug Evaluation,* W. Nimmo and G. Ticker (eds), 3–22. Chichester: John Wiley.

Wittes, J., Lakatos, E., and Probstfield, J. (1989). Surrogate endpoints in clinical trials: Cardiovascular diseases. *Statistics in Medicine* **8,** 415–425.

Zhao, L. P. and Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika* **77,** 642–648.

Received August 1996; *revised August* 1997; *accepted October* 1997.