Biometrics 60, 724–728
September 2004

Prentice's Approach and the Meta-Analytic Paradigm: A Reflection on the Role of Statistics in the Evaluation of Surrogate Endpoints

Ariel Alonso,[1*] Geert Molenberghs,[1] Tomasz Burzykowski,[1] Didier Renard,[1] Helena Geys,[1] Ziv Shkedy,[1] Fabián Tibaldi,[1] José Cortiñas Abrahantes,[1] and Marc Buyse[2]

[1]Center for Statistics, Limburgs Universitair Centrum, Diepenbeek, Belgium
[2]International Drug Development Institute (IDDI), Brussels, Belgium
[*]*email:* ariel.alonso@luc.ac.be

Summary. We put a perspective on the strengths and limitations of statistical methods for the evaluation of surrogate endpoints. Whereas using several trials overcomes some of the limitations of a single-trial framework (Prentice, 1989, *Statistics in Medicine* **8,** 431–440), arguably the evaluation of surrogate endpoints can never be done using only statistical evidence but such evidence should be seen as but one component in a decision-making process that involves, among others, a number of clinical and biological considerations. We briefly present a hierarchical framework that incorporates ideas from Prentice's work and is uniformly applicable to different types of surrogate and true clinical outcomes.

Key words: Hierarchical model; Likelihood reduction factor; Prentice criteria; Surrogate endpoints.

1. Introduction

The very mention of surrogate endpoints has always been very controversial. This may be due to a number of well-known unfortunate historical events. For example, in cardiovascular disease, the unsettling discovery that the two major antiarrhythmic drugs encanaide and flecanaide reduced arrhythmia but cause a more than three-fold increase in overall mortality stressed the need for caution in using nonvalidated surrogate markers in the evaluation of the possible clinical benefits of new drugs (CAST Investigators, 1989). On the other hand, the dramatic surge of the AIDS epidemic, the impressive therapeutic results obtained early on with zidovudine, and the pressure for an accelerated evaluation of new therapies have all led to first the use of CD4 blood count and, with the advent of HAART, viral load, as endpoints that replaced time to clinical events and overall survival (DeGruttola and Tu, 1995), in spite of some concerns about their limitations as surrogate markers for clinically relevant endpoints (Lagakos and Hoth, 1992).

Thus, while many would like to avoid surrogate endpoints altogether, sometimes surrogates will be the only reasonable alternative, especially when the true endpoint is rare and/or distant in time. It is then best to use *validated* surrogates, but one clearly needs to reflect on the very meaning of validation. Like in most clinical decisions, statistical arguments will play a major role, but ought to be considered in conjunction with clinical and biological evidence. Further, surrogate endpoints can play different roles in different phases of drug development. While it may be more acceptable in early phases of research, one should be much more careful using them as substitutes for the true endpoint in pivotal phase III trials versus replacing the true endpoint by the surrogate altogether in all research past a certain point. Prentice (1989), aware of this controversy but seeing the need for a formal framework, provided a definition and a set of criteria that have formed the basis for much subsequent work in some of which the connection between statistical and nonstatistical arguments has been lost. It follows from Prentice that a pure and strict statistical position will impose almost impossible requirements on a potential surrogate.

While the idea behind Prentice's definition and main criterion is appealing, a drawback, common to all single-trial approaches, is that it rests on strong, unverifiable assumptions. As argued by Daniels and Hughes (1997), Buyse et al. (2000), and Gail et al. (2000), a way out is the combination of information from several units or trials. Using hierarchical linear models, Buyse et al. (2000) defined surrogacy in terms of an individual-level as well as a trial-level measure, both of which are coefficients of determination. A drawback of this approach is that it hinges on normality and consequently several authors have proposed alternative concepts for other outcome types (e.g., binary or time-to-event). Combining ideas from both frameworks, we will propose a unified approach without obviating the obligation to consider biological and clinical plausibility of a surrogate.

Our ideas are illustrated using data from a clinical trial in patients with age-related macular degeneration (ARMD), a condition in which patients progressively lose vision

(Pharmacological Therapy for Macular Degeneration, 1997); 194 patients from 43 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. The treatment indicator is $Z = 0$ for placebo and $Z = 1$ for interferon-α. The final (surrogate) endpoint $T = 1$ ($S = 1$) if the lines of vision lost at 1 year (6 months) is greater than or equal to 3 (2) and 0 otherwise. The final endpoint is that used by the FDA for the approval of new drugs. The surrogate endpoint would be attractive since it is observed 6 months earlier. Five out of 42 participating centers enrolled patients only to one of the two treatment arms. These centers were excluded from consideration. Thus, 37 centers were available for analysis, with the number of individual patients per center ranging from 2 to 18 (189 patients overall).

In Section 2, we show how the key ingredients of Prentice's framework can be seen as integrated into a meta-analytic one. Unified and appealing measures to quantify trial-level and individual-level surrogacy are presented in Section 3 and their application to the case study is given in Section 4.

2. Prentice's Approach Versus a Meta-Analytic Paradigm: Some Considerations

Prentice's definition and the meta-analytic methodology have been perceived as competing to evaluate surrogate markers. However, it is possible to show that, in spite of obvious differences, the original definition given by Prentice (1989) and the individual-level surrogacy defined by Buyse et al. (2000) are strongly related.

Let S and T denote the surrogate and true endpoints, respectively, and let Z be an indicator variable for treatment. At this point, restricting attention to the single-trial case, we will conveniently assume that S and T are normally distributed. We will further assume that the following bivariate regression model holds

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

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

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

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

Prentice defined 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" (Prentice, 1989). From models (1) and (2), we can calculate the maximum-likelihood estimators $\hat{\alpha}$ and $\hat{\beta}$ for the treatment effects on the surrogate and true endpoints. Based on these we define

$$\begin{pmatrix} \tilde{\alpha} \\ \tilde{\beta} \end{pmatrix} = \begin{pmatrix} \dfrac{\hat{\alpha}}{\sqrt{\left(\dfrac{1}{n_0} + \dfrac{1}{n_1}\right)\sigma_{SS}}} \\ \dfrac{\hat{\beta}}{\sqrt{\left(\dfrac{1}{n_0} + \dfrac{1}{n_1}\right)\sigma_{TT}}} \end{pmatrix}$$

$$\sim N\left[\begin{pmatrix} \dfrac{\alpha}{\sqrt{\left(\dfrac{1}{n_0} + \dfrac{1}{n_1}\right)\sigma_{SS}}} \\ \dfrac{\beta}{\sqrt{\left(\dfrac{1}{n_0} + \dfrac{1}{n_1}\right)\sigma_{TT}}} \end{pmatrix}, \tilde{\Sigma}\right], \tag{4}$$

where n_0 and n_1 denote the number of observations with $Z_j = 0$ and $Z_j = 1$, respectively, and

$$\tilde{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

If we now apply Prentice's definition using the marginal distributions associated with (4), it is possible to prove that the definition holds if and only if $\rho^2 = 1$.

Observe that the meta-analytic paradigm encompasses Prentice's definition in the association at the individual level. In practice, in the meta-analytic framework (to be introduced later in this section), we will estimate the association at the individual-level $R^2_{\text{indiv}} = \rho^2$, which quantifies the strength of the relationship between the surrogate and the true endpoint, after removing the effects of treatment. Since this association allows us to evaluate how far the endpoints are from fulfilling Prentice's definition, a yes/no statement has been replaced by a quantification.

Prentice (1989, for survival times), followed by Freedman, Graubard, and Schatzkin (1992, for binary outcomes) outlined a set of criteria. In summary, they read (a) treatment has an impact on the surrogate endpoint, $\alpha \neq 0$ in (1); (b) treatment has an impact on the true endpoint, $\beta \neq 0$ in (2); (c) the surrogate endpoint has an impact on the true endpoint, $\gamma \neq 0$ in the regression relationship of T on S, captured in, for example, $T_j = \mu + \gamma S_j + \varepsilon_j$; and (d) the full effect of treatment upon the true endpoint is captured by the surrogate. The latter two are Prentice's original criteria. As Prentice himself later acknowledged, he proposed the framework "not to encourage their adoption in any particular setting (...) but rather to reinforce that it is only in very special circumstances that treatment information on an early surrogate end point will convey direct information concerning a treatment effect on a true later end point" (Prentice, 2000).

Let us return to the fact that definition and criteria, as detailed in Freedman et al. (1992), are equivalent only when both the surrogate and the true endpoints are binary (Buyse and Molenberghs, 1998). They are easily seen not to hold in the normal situation, by concentrating on the fourth criterion, which is verified through the conditional distribution of the true endpoint, given treatment *and* surrogate endpoint, derived from (1) and (2)

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

where $\beta_S = \beta - \sigma_{ST}\sigma_{SS}^{-1}\alpha$, $\gamma_Z = \sigma_{ST}\sigma_{SS}^{-1}$, and $V(\tilde{\varepsilon}_{Tj}) = \sigma_{TT}(1-\rho^2)$. It is usually stated that the fourth criterion requires that the parameter β_S be equal to zero, which leads to the condition on the mean structure $\beta = \sigma_{ST}\sigma_{SS}^{-1}\alpha$. Prentice's definition would simply require $\rho^2 = 1$ or equivalently $V(\tilde{\varepsilon}_{Tj}) = 0$, an intuitive condition in terms of the error structure of (5), rather than its systematic part.

We have assumed so far that the endpoints are normally distributed, which implies no mean-variance link, and the above considerations obviously do not apply as such to other types of endpoints (e.g., binary endpoints as in our example). The key point is, however, that Prentice's definition attributes a central role to ρ, whereas Prentice's fourth criterion attributes a central role to the relationship between β and α. To meaningfully study the latter relationship, replication at the trial level is required, motivating a hierarchical framework (Buyse et al., 2000). Suppose we have data from $i = 1, \ldots, N$ trials, in the ith of which $j = 1, \ldots, n_i$ subjects are enrolled. We now denote the true and surrogate endpoints by T_{ij} and S_{ij}, respectively, and by Z_{ij} the indicator variable for treatment. The linear models (1) and (2) can be rewritten as

$$S_{ij} = \mu_{Si} + \alpha_i Z_{ij} + \varepsilon_{Sij}, \tag{6}$$

$$T_{ij} = \mu_{Ti} + \beta_i Z_{ij} + \varepsilon_{Tij}, \tag{7}$$

where μ_{Si} and μ_{Ti} are trial-specific intercepts, α_i and β_i are trial-specific effects of treatment Z on the endpoints in trial i, and ε_{Si} and ε_{Ti} are correlated error terms, assumed to be mean-zero normally distributed with covariance matrix (3). The assumption of a constant covariance matrix may be unrealistic in certain settings but it should be noted that generalization to trial-specific covariance matrices is straightforward and the framework sketched in the next section fully allows for that.

Within this framework and as stated in the introduction, Buyse et al. (2000) defined surrogacy in terms of an individual-level as well as a trial-level measure, denoted by R^2_{indiv} (capturing the squared correlation between ε_{Sij} and ε_{Tij}) and R^2_{trial} (measuring how well β_i can be predicted by μ_{Si} and α_i), respectively. It has been suggested by some authors to extend Prentice's definition to this setting by applying it within each single trial involved in a meta-analysis. However, the foregoing discussion proved that this procedure would be, at least, very inefficient. Under models (6) and (7) the correlation between both endpoints is constant over trials, i.e., $\rho_i = \rho$ for all i. Applying Prentice's definition to each trial would be, for the same reasons as described above, equivalent to testing the hypotheses $H_0 : \rho_i^2 = 1$ versus $H_1 : \rho_i^2 \neq 1$. In contrast, the meta-analytic approach combines all of the information coming from the different trials by estimating $R^2_{\text{indiv}} = \rho^2$. This approach makes more efficient use of the data at hand, in addition to being more informative than several tests of hypothesis.

3. The Likelihood Reduction Factor

Based on this meta-analytic paradigm several parameters have been proposed to quantify the association between the surrogate S and the true endpoint T, at individual level, depending on the type of response variable used for S and T. In the cross-sectional normal–normal case, Buyse et al. (2000), using a bivariate normal regression model, defined the R^2_{indiv} as the squared correlation between the surrogate and the true endpoint after adjusting for treatment and trial effects. In the binary–binary setting, Geys (1994) used a bivariate conditional probit model and defined $R^2_{\text{indiv}} = \rho^2_{\tilde{S}\tilde{T}}$, which is the squared correlation between two latent variables $(\tilde{S},\tilde{T})$. Alternatively, they also proposed to define $R^2_{\text{indiv}} = \psi$, the global odds ratio between both endpoints estimated from a bivariate Plackett–Dale model. When the true endpoint is a survival time and the surrogate is a longitudinal sequence, Renard et al. (2001), using Henderson's model, proposed to study the individual level based on a time function defined as $R^2_{\text{indiv}}(t) = \text{corr}\{W_1(t), W_2(t)\}^2$, where $\{W_1(t), W_2(t)\}$ is a latent bivariate Gaussian process. Other proposals have been suggested in other settings. Estimating individual-level surrogacy, as the previous examples clearly showed, has frequently been based on a variance–covariance matrix coming from the distribution of the residuals. However, if we move away from the normal distribution it is not always very clear how we can quantify the association between both endpoints after adjusting for treatment and trial effects; therefore, several different parameters have been proposed showing a clear lack of a unified approach.

Here, we offer a general procedure that will allow us to evaluate surrogacy at the individual level in very general settings. Let us consider two generalized linear models for trial i:

$$g_T\{E(T_{ij} \mid Z_{ij})\} = \mu_{T_i} + \beta_i Z_{ij}, \tag{8}$$

$$g_T\{E(T_{ij} \mid Z_{ij}, S_{ij})\} = \theta_{0_i} + \theta_{1i} Z_{ij} + \theta_{2i} S_{ij}. \tag{9}$$

Longitudinal data are easily incorporated by including functions of time in (8) and (9). Other extensions readily follow, such as nonlinear relationship between S and $E(T)$. For example, $g_T\{E(T_{ij} \mid Z_{ij}, S_{ij})\} = \theta_{0_i} + \theta_{1i} Z_{ij} + f_i(S_{ij})$.

Denote the log-likelihood ratio test statistics to compare (8) with (9) within trial i by G_i^2. We then propose to quantify the association between both endpoints at the individual level using a likelihood reduction factor (LRF)

$$\text{LRF} = 1 - \frac{1}{N} \sum_i \exp\left(-\frac{G_i^2}{n_i}\right). \tag{10}$$

Following the ideas of Kent (1983), we can think of (10) as a sample estimate of a general measure of association between both endpoints based on the information gain about the true endpoint by using the surrogate. It follows that (1) LRF is always between 0 and 1, (2) LRF = 0 if the surrogate and the true endpoint are independent in each trial, and (3) as LRF → 1 for continuous models, there is usually some degeneracy appearing in the true joint distribution of S and T in each trial; often $\phi_{1i}(S) = \phi_{2i}(T)$, implying a deterministic relationship between both variables and finally in the cross-sectional normal–normal case the LRF reduces to the R^2_{indiv}.

Note that (9) is just the model proposed by Prentice in his fourth criterion. Thus, Prentice's fourth criterion features in the unifying procedure to quantify surrogacy at the individual level. One of the most appealing characteristics of this proposal is its avoidance of complicated hierarchical joint models for S and T. Models like (8) and (9) can usually be fitted using standard commercial software. The formulation helps to bridge the gap between the Prentice (1989) and Buyse et al. (2000) paradigms.

A few remarks are in place. First, in our developments, we have focused on the individual-level surrogacy. It ought to be understood that a statement at the individual level does not imply anything about trial-level surrogacy. For example, a surrogate can be poor at the individual level, but still very promising at the trial level when R^2_{trial} is sufficiently large. Second, and related to the previous point, one has to reflect carefully on which of the two levels is of most interest. On the one hand, one could restrict attention to the trial level only, in which case an individual-level measure of surrogacy, such as the LRF, would not need to be calculated. This could occur, for example, when one is merely interested in the prediction of the treatment effect in a new trial, given evidence from a meta-analysis. It then ought to be clear that Prentice's definition is less relevant, since it connects to the individual level and not to the trial level. On the other hand, if the prediction of a patient's outcome (e.g., survival) is of interest, based on one or several surrogate measurements (e.g., a longitudinal profile), the individual-level surrogacy is of primary interest.

4. Analysis of Case Study

We illustrate our method, using the ARMD data, for the binary outcomes: visual acuity at 6 months (S) and acuity at 1 year (T). The following three models should be independently fitted

$$\operatorname{logit}\left(\pi^T_{ij}\right) = \mu_{T_i} + \beta_i Z_{ij}, \tag{11}$$

$$\operatorname{logit}\left(\pi^{T|S}_{ij}\right) = \mu^S_{T_i} + \beta^S_i Z_{ij} + \gamma_{ij} \text{ vis } 6_{ij}, \tag{12}$$

$$\operatorname{logit}\left(\pi^S_{ij}\right) = \mu_{S_i} + \alpha_i Z_{ij}, \tag{13}$$

here $S_{ij} = \text{vis } 6_{ij}$ and $T_{ij} = \text{vis } 12_{ij}$ are the dichotomized visual acuity, for the jth patient in the ith trial, at 6 months and 1 year, respectively. We also use the notation $\pi^T_{ij} = \mathrm{E}(\text{vis } 12_{ij})$, $\pi^{T|S}_{ij} = \mathrm{E}(\text{vis } 12_{ij}|\text{ vis } 6_{ij})$, and $\pi^S_{ij} = \mathrm{E}(\text{vis } 6_{ij})$.

Even though, at the trial level, surrogacy can still be evaluated using the estimated values of $(\mu_{\mathrm{S}_i}, \alpha_i, \beta_i)$, obtained from models (11)–(13) and applying the coefficient of determination (R^2_{trial}) proposed by Buyse et al. (2000), at the individual level, the R^2_{indiv} proposed in the continuous case can no longer be used. However, the LRF can be used instead. Assuming that the association between both variables is constant over trials, (11) and (12) can be used to compute the LRF

$$\text{LRF} = 1 - \exp\left(-\frac{G^2}{n}\right), \tag{14}$$

where G^2 is the log-likelihood ratio statistics to compare models (11) and (12) and $n = \sum n_i$ is the total number of patients. Note that (14) results from (10) by considering all data as coming from a single study (so that $N = 1$). By way of sensitivity analysis, the assumption of a constant covariance structure was relaxed. The results obtained were virtually identical and therefore omitted.

Table 1

Age-related macular degeneration trial. Binary true and surrogate endpoint. Estimates (95% confidence intervals) calculated using the bootstrap.

Parameter	Estimate (95% CI)
R^2_{trial}	0.3845 (0.1494; 0.6144)
LRF	0.2648 (0.2213; 0.3705)
LRF_{adj}	0.4955 (0.3252; 0.6044)

Note that, as pointed out by Kent (1983), if the true endpoint has a fixed discrete distribution and if the true endpoint given the surrogate is modeled by a family of discrete distributions, then the conditional information gain is bounded above and hence the LRF is bounded by a number strictly less than one. Therefore, we will also report here the value of $\text{LRF}_{\text{adj}} = \text{LRF}/\max(\text{LRF})$ which can always reach one and hence is more meaningful.

Table 1 shows the results at both levels. All of the estimated values are too low to make visual acuity at 6 months a reliable surrogate for visual acuity at 12 months. At the trial level, an R^2_{trial} of 0.38 clearly shows that an accurate prediction of treatment effect at 1 year based on the treatment effect observed at 6 months does not seem to be possible. Of course, it is clear that when the outcome at 6 months is sufficiently large, then the prediction of the month 12 outcome, together with its prediction limits, may contain useful information. While this would hold for every R^2 greater than zero, the closer it is to zero, the larger and hence the more unrealistic will the surrogate endpoint value have to be. Switching to LRF_{adj}, we do obtain some evidence of a weak association at the individual level. These results are similar to the ones reported by Geys (1999), who used a joint bivariate probit model based on latent variables. She reported a smaller association at the trial level ($R^2_{\text{trial}} = 0.22$) and a stronger relationship at the individual level ($R^2_{\text{indiv}} = 0.64$). Nevertheless, these coefficients describe the association at an unobservable latent scale, rendering their interpretation more awkward than is the case with our proposal.

5. Discussion

Following the above, Buyse et al. (2000) proposed both R^2 measures as well as prediction equations to allow quantification of the properties of a candidate surrogate endpoint, thereby combining the essence of Prentice's framework with a hierarchical view. The intuitive appeal of this is the requirement of a good surrogate to satisfy two distinct but similarly looking properties. First, the surrogate endpoint must predict the true endpoint in individual patients. This condition is strongly related to Prentice's definition. Second, the effect of a treatment on the surrogate endpoint must predict the effect of that treatment on the true endpoint. While different, at face value, from Prentice's fourth criterion, it is consistent with its spirit, properly exploiting replication across trials.

While the R^2 measures do not readily generalize to settings with nonnormal outcomes, the likelihood reduction factor applies to a wide variety of settings (normal, binary, categorical, survival, and longitudinal outcomes) and reduces to the R^2 measure for normally distributed endpoints. It generalizes both Prentice (1989) and Buyse et al. (2000). A more formal study of the LRF in nonnormal settings will be reported elsewhere.

It could be argued that a disadvantage of the meta-analytic framework discussed above is that it requires several trials (or other experimental units) with reasonably large amounts of data in each. This should not be seen as a disadvantage, but as a necessary requirement in preparation of a potentially serious decision. In spite of the appeal of these methodological developments, it is clear however that surrogate marker validation cannot rely solely on a statistical approach, since important clinical and biological considerations need to be factored into the decision.

Résumé

Nous présentons une perspective sur les forces et les limites des méthodes statistiques dans l'évaluation des points finaux de substitution. Alors que l'utilisation de plusieurs essais dépasse certaines limitations posées par le cadre d'un essai unique (Prentice, 1989), on peut soutenir que l'évaluation de points finaux de substitution ne peut jamais résulter de la seule évidence statistique, mais que cette évidence doit être envisagée comme une composante dans un processus de décision impliquant, parmi d'autres, nombre de considérations biologiques et cliniques. Nous présentons brièvement un cadre hiérarchique incorporant les idées du travail de Prentice et qui est entièrement applicable aux différents types d'issues cliniques réelles ou substitutives.

References

Buyse, M. and Molenberghs, G. (1998). The validation of surrogate endpoints in randomized experiments. *Biometrics* **54,** 1014–1029.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. (2000). The validation of surrogate endpoints in meta-analysis of randomized experiments. *Biostatistics* **1,** 49–67.

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

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

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

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

Gail, M. H., Pfeiffer, R., van Houwelingen, H. C., and Carroll, R. J. (2000). On meta-analytic assessment of surrogate outcomes. *Biostatistics* **1,** 231–246.

Geys, H. (1994). Pseudo-likelihood methods and generalized estimating equations: Efficient estimation techniques for the analysis of correlated multivariate data. Ph.D. Thesis, Limburgs Universitair Centrum, Belgium.

Kent, J. (1983). Information gain and a general measure of correlation. *Biometrika* **70,** 163–173.

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.

Pharmacological Therapy for Macular Degeneration Study Groups (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.

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

Prentice, R. L. (2000). Comment on the paper by Begg and Leung. *Journal of the Royal Statistical Society, Series A* **163,** 26–27.

Renard, D., Geys, H., Molenberghs, G., Burzykowski, T., Buyse, M., Vangeneugden, T., and Bijnens, L. (2001). Validation of a longitudinally measured surrogate marker for a time-to-event endpoint. *Journal of Applied Statistics* **30,** 235–247.

Received December 2002. *Revised February* 2004.
Accepted February 2004.