Biostatistics (2000), **1**, 1, *pp*. 49–67
Printed in Great Britain

The validation of surrogate endpoints in meta-analyses of randomized experiments

M. BUYSE

International Institute for Drug Development, 430 avenue Louise B14, B1050 Brussels, Belgium
mbuyse@id2.be

G. MOLENBERGHS, T. BURZYKOWSKI, D. RENARD AND H. GEYS

Center for Statistics, Limburgs Universitair Centrum, B3590 Diepenbeek, Belgium

SUMMARY

The validation of surrogate endpoints has been studied by Prentice (1989). He presented a definition as well as a set of criteria, which are equivalent only if the surrogate and true endpoints are binary. Freedman *et al*. (1992) supplemented these criteria with the so-called 'proportion explained'. Buyse and Molenberghs (1998) proposed replacing the proportion explained by two quantities: (1) the relative effect linking the effect of treatment on both endpoints and (2) an individual-level measure of agreement between both endpoints. The latter quantity carries over when data are available on several randomized trials, while the former can be extended to be a trial-level measure of agreement between the effects of treatment of both endpoints. This approach suggests a new method for the validation of surrogate endpoints, and naturally leads to the prediction of the effect of treatment upon the true endpoint, given its observed effect upon the surrogate endpoint. These ideas are illustrated using data from two sets of multicenter trials: one comparing chemotherapy regimens for patients with advanced ovarian cancer, the other comparing interferon-α with placebo for patients with age-related macular degeneration.

Keywords: Ovarian cancer; Macular degeneration; Random-effects model; Surrogate endpoint; Two-stage model; Validation.

1. INTRODUCTION

A surrogate endpoint is one which can be used in lieu of the endpoint of primary interest 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). Biological markers of the disease process are often proposed as surrogate endpoints for clinically meaningful endpoints, the hope being that if a treatment showed benefit on the markers, it would ultimately also show benefit upon the clinical endpoints of interest. Before a surrogate endpoint can replace a final endpoint in the evaluation of an experimental treatment, it must be formally 'validated', a process that has caused a number of controversies and has not yet been fully elucidated.

In a landmark paper, Prentice (1989) proposed a formal definition of surrogate endpoints, outlined how they could be validated, and at the same time discussed intrinsic limitations in the surrogate marker validation quest. Much debate ensued, since many authors perceived a formal criteria-based approach as too stringent and not straightforward to verify (Fleming *et al*. 1994). Freedman *et al.* (1992) took

Prentice's approach one step further by introducing the 'proportion explained', which is the proportion of the treatment effect mediated by the surrogate. Buyse and Molenberghs (1998) discussed some problems with the proportion explained and proposed to replace it by two new measures. The first, defined at the population level and termed 'relative effect', is the ratio of the overall treatment effect on the true endpoint over that on the surrogate endpoint. The second is the individual-level association between both endpoints, after accounting for the effect of treatment, and referred to as 'adjusted association'.

In this paper, we extend these concepts to situations in which data are available from several randomized experiments. The individual-level association between the surrogate and final endpoints carries over naturally, the only change required being an additional stratification to account for the presence of multiple experiments. The experimental unit can be the center in a multicenter trial, or the trial in a meta-analysis context. We emphasize the latter situation, because an informative validation of a surrogate endpoint will typically require large numbers of observations coming from several trials. Moreover, meta-analytic data usually carry a degree of heterogeneity not encountered in a single trial, caused by differences in patient population, study design, treatment regimens, etc. We shall argue that these sources of heterogeneity increase one's confidence in the validity of a surrogate endpoint, when the relationship between the effects of treatment on the surrogate and the true endpoints tends to remain constant across such different situations.

The notion of relative effect can then be extended to a trial-level measure of association between the effects of treatment on both endpoints. The two measures of association, one at the individual level, the other at the trial level, are proposed as an alternative way to assess the usefulness of a surrogate endpoint. This approach also naturally yields a prediction for the effect of treatment on the true endpoint, based on the observation of the effect of treatment on the surrogate endpoint.

In Section 2, Prentice's definition and criteria, as well as Freedman's proportion explained, are reviewed. Our notation and motivating examples are presented in Section 3. Some new concepts and an alternative validation strategy are introduced in Section 4. The examples are analysed in Section 5. Fitting of some of the models in Section 4 is computationally not straightforward. Section 6 examines through simulations when numerical problems are likely to occur. Throughout the paper, the emphasis is on normally distributed endpoints, for which standard linear mixed models are appropriate. The mixed-models methodology provides an easy-to-use framework that avoids a lot of complexities encountered with different response types (Laird and Ware, 1982; Verbeke and Molenberghs, 1997). In practice, however, endpoints are seldom normally distributed. Section 7 briefly discusses possible extensions to more general situations where the surrogate and true endpoints are of a different nature, such as the highly relevant situation where the surrogate endpoint is binary and the final endpoint is a survival time, possibly censored (Lin *et al*. 1997). These extensions will be taken up further in separate papers. The paper closes with general comments on the difficulties involved in validating surrogate endpoints for use in future clinical trials.

2. VALIDATION CRITERIA

2.1. Prentice's definition

Prentice (1989) 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' (Prentice, 1989, p. 432). 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. Prentice's definition can be written

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

where $f(X)$ denotes the probability distribution of the random variable X and $f(X|Z)$ denotes the probability distribution of X conditional on the value of Z. As such, this definition is of limited value since a direct verification that a triplet (T, S, Z) fulfills the definition would require a large number of experiments to be available with information on the triplet. Even if many experiments were available, the equivalence of the statistical tests implied in (1) might not be true in all of them because of chance fluctuations and/or lack of statistical power. Operational criteria are therefore needed to check if definition (1) is fulfilled.

2.2. Prentice's criteria

Four operational criteria have been proposed to check if a triplet (T, S, Z) fulfills the definition. The first two verify departures from the null hypotheses implicit in (1):

$$f(S|Z) \neq f(S), \tag{2}$$

$$f(T|Z) \neq f(T). \tag{3}$$

Strictly speaking, (2) and (3) are not criteria since having both $f(T|Z) = f(T)$ and $f(S|Z) = f(S)$ is consistent with definition (1). However, in this case, the validation is practically impossible since one may fail to detect differences due to lack of power. Thus, in practice, the validation requires Z to have an effect on both T and S. Several authors have pointed out that requiring Z to have a statistically significant effect on T may be excessively stringent, for in that case, from the limited perspective of significance testing, there would no longer be a need to establish the surrogacy of S (Fleming *et al.* 1994).

The other two criteria are

$$f(T|S) \neq f(T), \tag{4}$$

$$f(T|S, Z) = f(T|S). \tag{5}$$

Buyse and Molenberghs (1998) reproduce the arguments that establish the sufficiency of conditions (5) and (4) for (1) to hold in the case of binary responses. It is also easy to show that condition (4) is always necessary for (1), and that condition (5) is necessary for binary endpoints but not in general. Indeed, suppose (5) does not hold, then, assuming that $f(S|Z) = f(S)$,

$$f(T|Z) = \int f(T|S, Z)\, f(S)\, dS \tag{6}$$

and

$$f(T) = \int f(T|S)\, f(S)\, dS. \tag{7}$$

However, (6) and (7) are in general not equal to one another, in which case definition (1) is violated. However, it is possible to construct examples where $f(T|Z) = f(T)$, in which case the definition still holds despite the fact that (5) does not hold. Hence (5) is not a necessary condition, except for binary endpoints.

Next, assume (5) holds but (4) does not. Then,

$$f(T|Z) = \int f(T|S)\, f(S|Z)\, dS = \int f(T)\, f(S|Z)\, dS = f(T),$$

and hence $f(T|Z) = f(T)$ regardless of the relationship between S and Z. The simplest example is the situation where T is independent of the pair (S, Z). Thus, (4) is necessary to avoid situations where one null hypothesis is true while the other is not. However, criteria (2) and (3) already imply that both null hypotheses must be rejected, and therefore criterion (4) is of no additional value. In fact, criterion (4) indicates that the surrogate endpoint has prognostic relevance for the final endpoint, a condition which will obviously be fulfilled by any sensible surrogate endpoint. Conditions (2)–(5) are informative and will tend to be fulfilled for valid surrogate endpoints, but they should not be regarded as strict criteria. Condition (5) captures the essential notion of surrogacy by requiring that the treatment is irrelevant for predicting the true outcome, given the surrogate. In the next section we discuss how Freedman *et al.* (1992) used this concept in estimation rather than in testing. Our meta-analytic development, laid out in Section 4, also emphasizes estimation and prediction rather than hypothesis testing.

2.3. Freedman's proportion explained

Freedman *et al.* (1992) argued that criterion (5) raises a conceptual difficulty in that it requires the statistical test for treatment effect on the true endpoint to be non-significant after adjustment for the surrogate. The non-significance 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.* (1992) proposed to calculate the proportion of the treatment effect explained by the surrogate. In this paradigm, a good surrogate is one for which this proportion explained (PE) is close to unity (Prentice's criterion (5) would require that $PE = 1$). Buyse and Molenberghs (1998) outlined some conceptual difficulties with the PE, in particular that it is not a proportion: PE can be estimated to be anywhere on the real line, which complicates its interpretation. They argued that PE can advantageously be replaced by two related quantities: the relative effect (RE), which is the ratio of the effects of treatment upon the final and the surrogate endpoint, and the treatment-adjusted association between the surrogate and the true endpoint, ρ_Z. In the remainder of this paper, these proposals are extended using data from several experiments. Motivating examples are introduced in the next paragraph, and the alternative approach in Section 4.

3. Notation and motivating examples

Suppose we have data from N trials, in the ith of which n_i subjects are enrolled. Let T_{ij} and S_{ij} be random variables that denote the true and surrogate endpoints, respectively, for the jth subject in the ith trial, and let Z_{ij} be an indicator variable for treatment. While the main focus of this paper is on binary treatment indicators, the methods proposed generalize without difficulty to multiple category indicators for treatment, as well as to situations where covariate information is used in addition to the treatment indicators.

3.1. An example in cancer

Our methods will first be illustrated using data from a meta-analysis of four randomized multicenter trials in advanced ovarian cancer (Ovarian Cancer Meta-analysis Project, 1991). 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 the logarithm of time to progression, defined as the time (in weeks) from randomization to clinical progression of the disease or death due to the disease, while the final endpoint T_{ij} will be the logarithm of survival, defined as the time (in weeks) 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 (Ovarian Cancer Meta-analysis Project, 1991). The dataset was subsequently updated to include a minimum follow-up of 10 years in all trials (Ovarian Cancer Meta-analysis Project, 1998). After such long follow-up, most patients have had a disease progression or have died (952 of 1194 patients, i.e. 80%), so censoring will be ignored in our analyses. Methods that account for censoring would admittedly be preferable, but we ignore it here for the purposes of illustrating the case where the surrogate and final endpoints are both normally distributed.

The ovarian cancer dataset contains only four trials. This will turn out to be insufficient to apply the meta-analytic methods of Section 4. In the two larger trials, information is also available on the centers in which the patients had been treated. We can then use the center as the unit of analysis for the two larger trials, and the trial as the unit of analysis for the two smaller trials. A total of 50 'units' are thus available for analysis, with the number of individual patients per unit ranging from 2 to 274. To assess sensitivity, all analyses will be performed with and without the two smaller trials in which the center is unknown.

The first three Prentice criteria (2)–(4) are provided by tests of significance of parameters α, β and γ in the following models:

$$S_{ij}|Z_{ij} = \mu_S + \alpha Z_{ij} + \varepsilon_{Sij}, \tag{8}$$

$$T_{ij}|Z_{ij} = \mu_T + \beta Z_{ij} + \varepsilon_{Tij}, \tag{9}$$

$$T_{ij}|S_{ij} = \mu + \gamma S_{ij} + \varepsilon_{ij}, \tag{10}$$

where ε_{Sij}, ε_{Tij}, and ε_{ij} are independent normally distributed errors with mean zero. If the analysis is restricted to the two large trials in which the center is known, $\alpha = 0.228$ (standard error, SE 0.091, $P = 0.013$), $\beta = 0.149$ (SE 0.085, $P = 0.079$), and $\gamma = 0.874$ (SE 0.011, $P < 0.0001$). Strictly speaking, the criteria are not fulfilled because β fails to reach statistical significance. This will often be the case, since a surrogate endpoint is needed when there is no convincing evidence of a treatment effect upon the true endpoint.

As emphasized in the previous paragraph, we cannot strictly show that the last criterion (5) is fulfilled. Instead, we can calculate Freedman's proportion explained,

$$PE = 1 - \frac{\beta_S}{\beta}, \tag{11}$$

where β is the estimate of the effect of Z on T as in (9), and β_S is the estimate of the effect of Z on T after adjustment for S,

$$T_{ij}|Z_{ij}, S_{ij} = \tilde{\mu}_T + \beta_S Z_{ij} + \gamma_Z S_{ij} + \tilde{\varepsilon}_{Tij}. \tag{12}$$

Here, $\beta_S = -0.051$ (SE 0.028), and $PE = 1.34$, (95% delta confidence limits [0.73; 1.95]). The proportion explained is larger than 100%, because the direction of the effect of Z on T is reversed after adjustment for S. Another problem would arise if there were a strong interaction between Z and S, which would require the following model to be fitted instead of (12),

$$T_{ij}|Z_{ij}, S_{ij} = \check{\mu}_T + \check{\beta}_S Z_{ij} + \check{\rho}_Z S_{ij} + \delta Z_{ij} S_{ij} + \check{\varepsilon}_{Tij}. \tag{13}$$

With this model, PE would cease to have a single interpretation and the validation process would have to stop (Freedman *et al.* 1992). In the two large ovarian cancer trials, the interaction term is not statistically significant ($\delta = 0.014$, SE 0.022), and therefore model (12) may be used.

Buyse and Molenberghs (1998) suggested replacing the PE by two quantities: the relative effect,

$$RE = \beta/\alpha, \tag{14}$$

and the association ρ_Z between T and S, adjusted for Z, which can be calculated from jointly modelling (8) and (9). To this end, the error terms of (8) and (9) are assumed to follow a bivariate Gaussian distribution with zero mean and general 2 × 2 covariance matrix. In this case $RE = 0.65$ (95% confidence limits [0.36; 0.95]) and $\rho_Z = 0.944$ (95% confidence limits [0.94; 0.95]). Thus, the adjusted correlation is very close to one and estimated with high precision. The relative effect is determined with reasonable precision, and enables calculation of the predicted effect of treatment upon survival based on the observed effect upon time to progression in a new trial. However, this prediction is based on the strong assumption of a regression through the origin based on a single pair $(\widehat{\alpha}, \widehat{\beta})$.

When the two smaller trials are included in the analysis, the results change very little, providing evidence for the validity of considering each of the smaller trials as a single center. The P values for α, β, and γ become 0.003, 0.054, and < 0.0001, respectively, and $PE = 1.46$ (95% confidence limits [0.80; 2.13]), $RE = 0.60$ (95% confidence limits [0.32; 0.87]), $\rho_Z = 0.942$ (95% confidence limits [0.94; 0.95]). By including both trials, the precision is improved somewhat. However, in this case, the interaction term in model (13) is statistically significant ($\delta = 0.037$, SE 0.018), further complicating the interpretation of PE.

3.2. An example in ophthalmology

Our second example was presented in a previous paper (Buyse and Molenberghs, 1998). It concerns a clinical trial for patients with age-related macular degeneration, a condition in which patients progressively lose vision (Pharmacological Therapy for Macular Degeneration Study Group, 1997). 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. The first three Prentice criteria (2)–(4) are again provided by tests of significance of parameters α, β and γ. Here, $\alpha = -1.90$ (SE 1.87, $P = 0.312$), $\beta = -2.88$ (SE 2.32, $P = 0.216$), and $\gamma = 0.92$ (SE 0.06, $P < 0.001$). Only γ is statistically significant and therefore the validation procedure has to stop inconclusively. Note, however, that the lack of statistical significance of α and β could merely be due to the insufficient number of observations available in this trial. Also note that α and β are negative, indicating a negative effect of interferon-α upon visual acuity. Freedman's proportion explained is calculated as $PE = 0.61$ (95% confidence limits [-0.19; 1.41]). The relative effect is $RE = 1.51$ (95% confidence limits [-0.46; 3.49]), while the adjusted association $\rho_Z = 0.74$ (95% confidence limits [0.68; 0.81]). The adjusted association is determined rather precisely, but the confidence limits of PE and RE are too wide to convey any useful information. Even so, as we will see in Section 5, some conclusions can be reached in this example that are in sharp contrast to those reached in the ovarian cancer example.

4. A META-ANALYTIC APPROACH

We focus on surrogate and true endpoints which are assumed to be jointly normally distributed. Two distinct modelling strategies will be followed, based on a two-stage fixed effects representation on the one hand and random effects on the other hand.

Let us describe the two-stage model first. The first stage is based upon a fixed-effects model,

$$S_{ij}|Z_{ij} = \mu_{Si} + \alpha_i Z_{ij} + \varepsilon_{Sij} \tag{15}$$

and

$$T_{ij}|Z_{ij} = \mu_{Ti} + \beta_i Z_{ij} + \varepsilon_{Tij}, \tag{16}$$

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 normally distributed error terms, assumed to be mean-zero with covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ & \sigma_{TT} \end{pmatrix}. \tag{17}$$

At the second stage, we assume

$$\begin{pmatrix} \mu_{Si} \\ \mu_{Ti} \\ \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} m_{Si} \\ m_{Ti} \\ a_i \\ b_i \end{pmatrix}, \tag{18}$$

where the second term on the right-hand side of (18) is assumed to follow a zero-mean normal distribution with dispersion matrix

$$D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ & d_{TT} & d_{Ta} & d_{Tb} \\ & & d_{aa} & d_{ab} \\ & & & d_{bb} \end{pmatrix}. \tag{19}$$

Next, the random-effects representation is based upon combining both steps, to give

$$S_{ij}|Z_{ij} = \mu_S + m_{Si} + \alpha Z_{ij} + a_i Z_{ij} + \varepsilon_{Sij} \tag{20}$$

and

$$T_{ij}|Z_{ij} = \mu_T + m_{Ti} + \beta Z_{ij} + b_i Z_{ij} + \varepsilon_{Tij}, \tag{21}$$

where now μ_S and μ_T are fixed intercepts, α and β are the fixed effects of treatment Z on the endpoints, m_{Si} and m_{Ti} are random intercepts and a_i and b_i are the random effects of treatment Z on the endpoints in trial i. The vector of random effects $(m_{Si}, m_{Ti}, a_i, b_i)$ is assumed to be mean-zero normally distributed with covariance matrix (19). The error terms ε_{Si} and ε_{Ti} follow the same assumptions as in fixed-effects model (15)–(16), with covariance matrix (17). Appendix B provides sample SAS code (SAS Institute Inc., Cary, NC) to fit the random-effects model.

A lot of debate has been devoted to the relative merits of fixed versus random effects, especially in the context of meta-analysis (Thompson and Pocock, 1991; Fleiss, 1993; Thompson, 1993; Senn, 1998). Although the underlying models rest on different assumptions about the nature of the experiments being analysed, the two approaches yield discrepant results only in pathological situations, or in very small samples where a fixed-effects analysis can yield artificially precise results if the experimental units truly constitute a random sample from a larger population. In our setting both approaches are very similar, and the two-stage procedure can be used to introduce random effects (Laird and Ware, 1982; Verbeke and Molenberghs, 1997). As the data analysis in Section 5 will illustrate, the choice between random and fixed effects can also be guided by pragmatic arguments.

4.1. Trial-level surrogacy

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. It is essential, therefore, to explore the quality of the prediction of the treatment effect on the true endpoint in trial i by (a) information obtained in the validation process based on trials $i = 1, \ldots, N$ and (b) the estimate of the effect of Z on S in a new trial $i = 0$. Fitting either the fixed-effects model (15)–(16) or the mixed-effects model (20)–(21) 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},

$$S_{0j} = \mu_{S0} + \alpha_0 Z_{0j} + \varepsilon_{S0j}. \tag{22}$$

Estimates for m_{S0} and a_0 are

$$\widehat{m}_{S0} = \widehat{\mu}_{S0} - \widehat{\mu}_S$$

and

$$\widehat{a}_0 = \widehat{\alpha}_0 - \widehat{\alpha}.$$

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:

$$E(\beta + b_0 | m_{S0}, a_0) = \beta + \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} \mu_{S0} - \mu_S \\ \alpha_0 - \alpha \end{pmatrix} \tag{23}$$

and

$$\text{var}(\beta + b_0 | m_{S0}, a_0) = d_{bb} - \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}. \tag{24}$$

This suggests calling a surrogate 'perfect at the trial level' if the conditional variance (24) is equal to zero. A measure to assess the quality of the surrogate at the trial level is the coefficient of determination

$$R^2_{\text{trial(f)}} = R^2_{b_i | m_{Si}, a_i} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}}. \tag{25}$$

Coefficient (25) is unitless and ranges in the unit interval if the corresponding variance–covariance matrix is positive definite, two desirable features for its interpretation. Intuition can be gained by considering the special case where the prediction of b_0 can be done independently of the random intercept m_{S0}. Expressions (23) and (24) then reduce to

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

and

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

with corresponding

$$R^2_{\text{trial(r)}} = R^2_{b_i | a_i} = \frac{d_{ab}^2}{d_{aa} d_{bb}}. \tag{26}$$

Now, $R^2_{\text{trial(r)}} = 1$ if the trial-level treatment effects are simply multiples of each other. We will refer to this simplified version as the reduced random-effects model, while the original expression (25) will be said to derive from the full random-effects model.

An estimate for $\beta + b_0$ is obtained by replacing the right-hand side of (23) with the corresponding parameter estimates. A confidence interval is obtained by applying the delta method to (23). The covariance matrix of the parameters involved is obtained from the meta-analysis, except for μ_{S0} and α_0, which are obtained from fitting (22) to the data of the new trial. The corresponding prediction interval is found by adding (24) to the variance obtained for the confidence interval. Details are given in Appendix A.

There is a close connection between the prediction approach followed here and empirical Bayes estimation (Verbeke and Molenberghs, 1997, Section 3.11). To see this, consider a similar but non-identical approach where all data are analysed together. This means that a meta-analysis is performed of the surrogate data on trials $i = 0, \ldots, N$ and of the true endpoint data on trials $i = 1, \ldots, N$. The estimate of b_0 will be based only on the surrogate data, since the true endpoint is unknown for trial $i = 0$, and on the parameter estimates. The expression for the empirical Bayes estimate of b_0 is identical to (23), but the numerical value will be slightly different since the parameters of the linear mixed model are determined on a larger set of data. For example, with the MIXED procedure in SAS, obtaining the empirical Bayes estimate of b_0 is immediate, but its conditional variance requires some additional computation (Littell *et al.* 1996).

4.2. Individual-level surrogacy

To validate a surrogate endpoint, Buyse and Molenberghs (1998) suggested using the association between the surrogate and the final endpoints after adjustment for the treatment effect. To this end, we need to construct the conditional distribution of T, given S and Z. From (15)–(16) we derive

$$\begin{aligned} T_{ij}|Z_{ij}, S_{ij} \sim N \Big\{ & \mu_{Ti} - \sigma_{TS}\sigma_{SS}^{-1}\mu_{Si} + (\beta_i - \sigma_{TS}\sigma_{SS}^{-1}\alpha_i)Z_{ij} + \sigma_{TS}\sigma_{SS}^{-1}S_{ij}\ ; \\ & \sigma_{TT} - \sigma_{TS}^2\sigma_{SS}^{-1} \Big\}. \end{aligned} \tag{27}$$

Similarly, the random-effects model (20)–(21) yields

$$\begin{aligned} T_{ij}|Z_{ij}, S_{ij} \sim N \Big\{ & \mu_T + m_{Ti} - \sigma_{TS}\sigma_{SS}^{-1}(\mu_S + m_{Si}) + [\beta + b_i - \sigma_{TS}\sigma_{SS}^{-1}(\alpha + a_i)]Z_{ij} + \sigma_{TS}\sigma_{SS}^{-1}S_{ij}\ ; \\ & \sigma_{TT} - \sigma_{TS}^2\sigma_{SS}^{-1} \Big\}, \end{aligned} \tag{28}$$

where conditioning is also on the random effects. The association between both endpoints after adjustment for the treatment effect is captured in both (27) and (28) by

$$R^2_{\text{indiv}} = R^2_{\varepsilon_{Ti}|\varepsilon_{Si}} = \frac{\sigma_{ST}^2}{\sigma_{SS}\sigma_{TT}}, \tag{29}$$

the squared correlation between S and T after adjustment for both the trial effects and the treatment effect. Note that $R_{\varepsilon_{Ti}|\varepsilon_{Si}}$ generalizes the adjusted association ρ_Z of Section 3 to the case of several trials.

4.3. A new approach to surrogate evaluation

The development in Sections 4.1 and 4.2 suggests calling a surrogate ‘trial-level valid’ if $R^2_{\text{trial(f)}}$ (or $R^2_{\text{trial(r)}}$) is sufficiently close to one, and ‘individual-level valid’ if R^2_{indiv} is sufficiently close to one. Finally, a surrogate is termed ‘valid’ if it is both trial-level and individual-level valid. In order to replace the words ‘valid’ with ‘perfect’, the corresponding R^2 values are required to equal one.

To be useful in practice, a valid surrogate must be able to predict the effect of treatment upon the true endpoint with sufficient precision to distinguish safely between effects that are clinically worthwhile and effects that are not. This requires that both the estimate of $\beta + b_0$ be sufficiently large and that the prediction interval of this quantity be sufficiently narrow.

It should be noted that the validation criteria proposed here do not require the treatment to have a significant effect on either endpoint. In particular, it is possible to have $\alpha \equiv 0$ and yet have a perfect surrogate. Indeed, even though the treatment may not have any effect on the surrogate endpoint as a whole, the fluctuations around zero in individual trials (or other experimental units) can be very strongly predictive of the effect on the true endpoint. However, such a situation is unlikely to occur since the heterogeneity between the trials is generally small compared with that between individual patients.

4.4. Validation in a single trial

If data are available on a single trial (or, more generally, on a single experimental unit), the above developments are only partially possible. While the individual-level reasoning (producing ρ_Z as in (29)) carries over by virtue of the within-trial replication, the trial-level reasoning breaks down and one cannot go beyond the relative effect (RE) as suggested in Buyse and Molenberghs (1998). Recall that the RE is defined as the ratio of the effects of Z on S and T, respectively, as expressed in (14). The confidence limits of RE can be used to assess the uncertainty about the value of β predicted from that of α, but in contrast to the above developments, no prediction interval can be calculated for β.

5. DATA ANALYSIS

5.1. Advanced ovarian cancer

As in Section 3.1, all analyses have been performed with and without the two smaller trials. Excluding the two smaller trials has very little impact on the estimates of interest, and therefore the results reported are those obtained with all four trials. Two-stage fixed-effects models (15)–(16) could be fitted, as well as a reduced version of the mixed-effects model (20)–(21), with random treatment effects but no random intercepts. Point estimates for the two types of model are in close agreement, although standard errors are smaller by roughly 35% in the random-effects model. Figure 1 shows a plot of the treatment effects on the true endpoint (logarithm of survival) by the treatment effects on the surrogate endpoint (logarithm of time to progression). These effects are highly correlated. Similarly to the random-effects situation, we refer to the models with and without the intercept used for determining R^2 as the reduced and full fixed-effects models. The reduced fixed-effects model provides $R^2_{\text{trial(r)}} = 0.939$ (SE 0.017). When the sample sizes of the experimental units are used to weigh the pairs (a_i, b_i), then $R^2_{\text{trial(r)}} = 0.916$ (SE 0.023). The full fixed-effects model yields $R^2_{\text{trial(f)}} = 0.940$ (SE 0.017). In the reduced random-effects model, $R^2_{\text{trial(r)}} = 0.951$ (SE 0.098).

Predictions of the effect of treatment on log(survival), based on the observed effect of treatment on log(time to progression), are of interest. Table 1 reports prediction intervals for several experimental units: six centers taken at random from the two large trials, and the two small trials in which the center is unknown. Note that none of the predictions is significantly different from zero. The predicted values for $\beta + b_0$ agree reasonably well with the effects estimated from the data. The ratio $\widehat{\beta}_0/\widehat{\alpha}_0$ ranges from 0.69 to 0.73, which is close to the RE estimated in Section 3.1.

At the individual level, $R^2_{\text{indiv}} = 0.886$ (SE 0.006) in the fixed-effects model, and $R^2_{\text{indiv}} = 0.888$ (SE 0.006) in the reduced random-effects model. The square roots of these quantities are, respectively, 0.941 and 0.942, very close to the value of ρ_Z estimated in Section 3.1.

Thus, we conclude that time to progression can be used as a surrogate for survival in advanced ovarian cancer. The effect of treatment can be observed earlier if time to progression is used instead of survival,

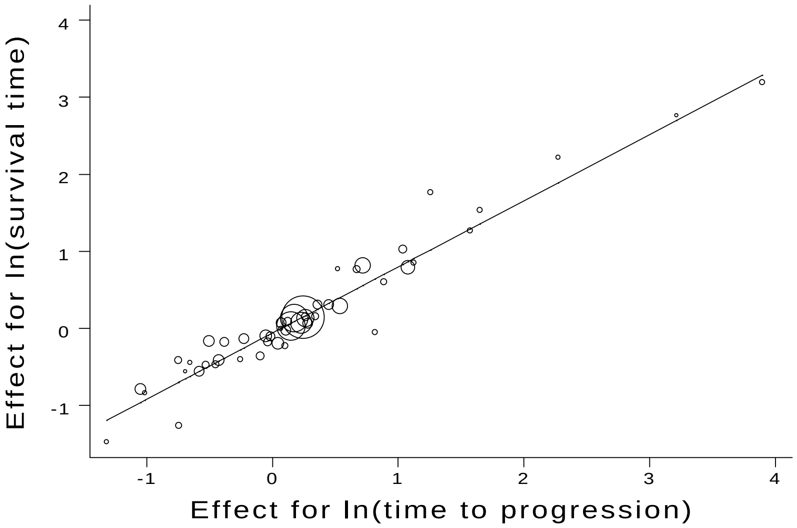

Fig. 1. Ovarian cancer trials: treatment effects. Treatment effects on the true endpoint (logarithm of survival time) versus treatment effects on the surrogate endpoint (logarithm of time to progression) for all units of analysis. The size of each point is proportional to the number of patients in the corresponding unit.

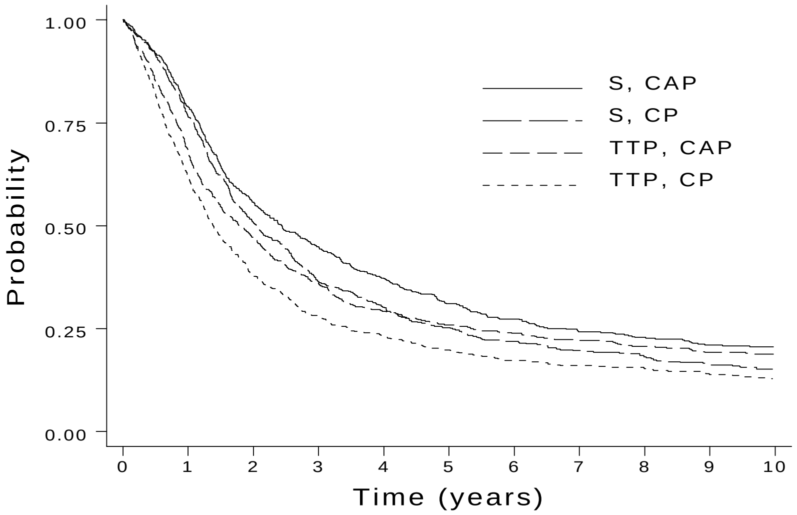

Fig. 2. Ovarian cancer trials: survival curves. Kaplan–Meier estimates of survival (S) and time to progression (TTP) for the two treatment groups: cyclophosphamide plus cisplatin (CP) and cyclophosphamide plus adriamycin plus cisplatin (CAP).

and it is also somewhat more pronounced as shown by the overall Kaplan–Meier estimates of Figure 2. Hence, a trial that used time to progression would require less follow-up time and less patients to establish the statistical significance of a truly superior treatment than a trial that used survival (Chen *et al.* 1998). The gains, however, would be modest because progression is followed by death within 1 year for most patients.

Table 1. *Predictions for the advanced ovarian cancer data (Ovarian Cancer Meta-analysis Project, 1991)*

Unit	# Patients	# Trials	$\widehat{\alpha}_0$ (SE)	$E(\beta + b_0 \mid a_0)$ (SE)	$\widehat{\beta}_0$ (SE)
Center 6	17	2	−0.58(0.33)	−0.45(0.29)	−0.56(0.32)
		4		−0.45(0.29)	
Center 8	10	2	0.67(0.76)	0.49(0.57)	0.76(0.39)
		4		0.47(0.56)	
Center 37	12	2	1.02(0.61)	0.76(0.54)	1.04(0.70)
		4		0.73(0.53)	
Center 49	40	2	0.54(0.34)	0.39(0.26)	0.28(0.28)
		4		0.37(0.25)	
Center 55	31	2	1.08(0.56)	0.80(0.44)	0.79(0.45)
		4		0.77(0.44)	
Center BB	21	2	−1.05(0.55)	−0.80(0.46)	−0.79(0.51)
		4		−0.79(0.46)	
Trial DACOVA	274	2	0.25(0.15)	0.17(0.13)	0.14(0.14)
Trial GONO	125	2	0.15(0.25)	0.10(0.20)	0.03(0.22)

Note: The number of patients is reported for each unit, as well as which sample is used for the estimation (only two trials or all four). $\widehat{\alpha}_0$ and $\widehat{\beta}_0$ are values estimated from the data; $E(\beta + b_0 \mid a_0)$ is the predicted effect of treatment on survival (β_0), given its effect upon time to progression ($\widehat{\alpha}_0$). The DACOVA and GONO trials are the two smaller studies, for which predictions are based on parameter estimates from the centers in the two larger studies.

The results derived here are considerably more useful than the conclusions in Section 3.1. Indeed, the first three Prentice criteria provide only marginal evidence and PE cannot be estimated on the full dataset, since there is a three-way interaction between Z, S, and T. RE is meaningful and estimated with precision, but it is derived from a regression through the origin based on a single data point. In contrast, the approach used here combines evidence from several experimental units and allows prediction intervals to be calculated for the effect of treatment on the true endpoint.

5.2. Age-related macular degeneration

The age-related macular degeneration data come from a single multicenter trial. Therefore, it is natural to consider the center in which the patients were treated as the unit of analysis. A total of 36 centers were thus available for analysis, with the number of individual patients per center ranging from 2 to 18.

Figure 3(a) shows a plot of the raw data (true endpoint versus surrogate endpoint for all individual patients). Irrespective of the software used, the random effects are difficult to obtain. Therefore, we report only the result of a two-stage fixed-effects model and explore the computational issues further in Section 6. Figure 3(b) shows a plot of the treatment effects on the true endpoint by the treatment effects on the surrogate endpoint. These effects are moderately correlated, with $R^2_{\text{trial(f)}} = 0.692$ (SE 0.087). The estimates based on the reduced model are virtually identical. At the individual level, $R^2_{\text{indiv}} = 0.483$ (SE 0.053). Note that $R_{\text{indiv}} = 0.69$ is close to $\rho_Z = 0.74$ as estimated in Section 3.2. The coefficients of determination $R^2_{\text{trial(r)}}$ and R^2_{indiv} are both too low to make visual acuity at 6 months a reliable surrogate for visual acuity at 12 months. Figure 3(c) shows that the correlation of the measurements at 6 months and at 1 year is indeed rather poor at the individual level. Therefore, even with the limited data available, it is clear that the assessment of visual acuity at 6 months is not a good surrogate for the same assessment at 1 year. This is in contrast with the inconclusive analysis in Section 3.2.

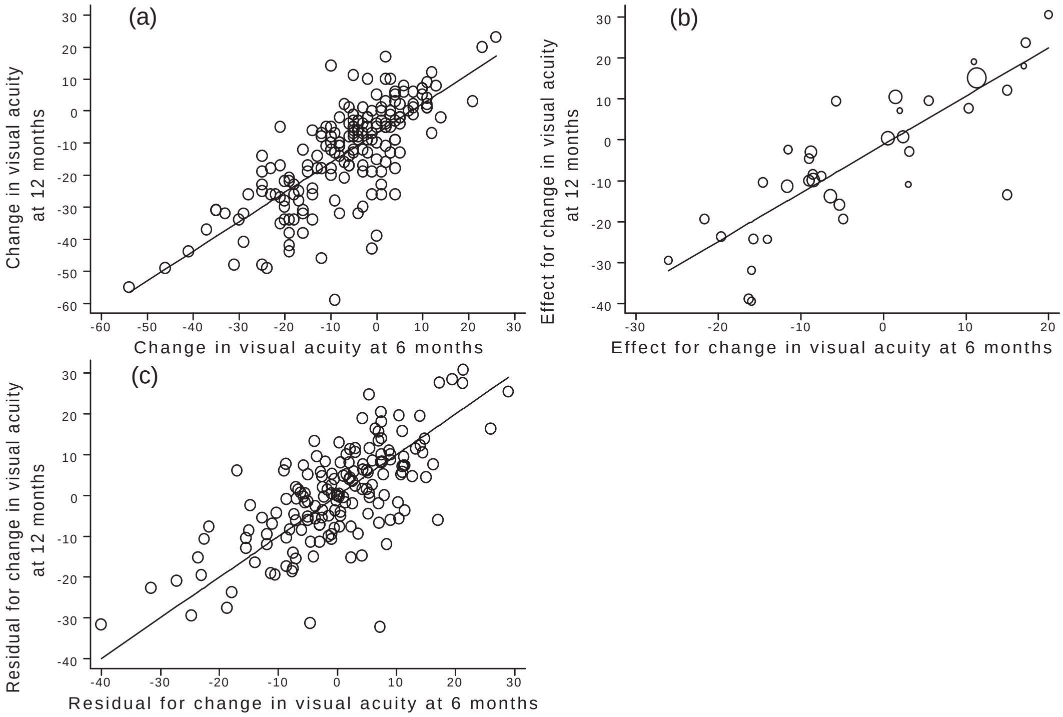

Fig. 3. Age-related macular degeneration trial. (a) True endpoint (change in visual acuity at 1 year) versus surrogate endpoint (change in visual acuity at 6 months) for all individual patients, raw data. (b) Treatment effects on the true endpoint versus treatment effects on the surrogate endpoint in all centers. The size of each point is proportional to the number of patients in the corresponding center. (c) True endpoint versus surrogate endpoint for all individual patients, after correction for treatment effect.

6. Computational issues

In this paragraph, we investigate convergence properties of the random-effects approach as proposed in Section 4. The need for such an investigation arises from the observation that in many practical instances, convergence of the Newton–Raphson algorithm yielding (restricted) maximum likelihood solutions could hardly be achieved. Therefore, it is worth knowing what features of the problem at hand may be of influence in easing convergence of the algorithm, since this may be an additional factor to decide between a two-stage or a random-effects model.

We explored the following factors: number of trials, size of the between-trial variability (compared with residual variability), number of patients per trial, normality assumption, and strength of the correlation

Table 2. *Number and percentage of runs out of 500 for which convergence was achieved within 20 iterations*

	Number of trials		
σ^2	50	20	10
1	500 (100%)	498 (100%)	412 (82%)
0.1	491 (98%)	417 (83%)	218 (44%)

between random treatment effects. Since only the first two factors were found significantly to affect convergence of the algorithm, we do not report on the others in the remainder of this section.

Table 2 shows the number of runs for which convergence could be achieved within 20 iterations. In each case, 500 runs were performed, assuming the following model,

$$S_{ij} \mid Z_{ij} = 45 + m_{S_i} + (3 + a_i)Z_{ij} + \varepsilon_{Sij}$$

and

$$T_{ij} \mid Z_{ij} = 50 + m_{Ti} + (5 + b_i)Z_{ij} + \varepsilon_{Tij},$$

where $(m_{Si}, m_{Ti}, a_i, b_i) \sim N(0, D)$ with

$$D = \sigma^2 \begin{pmatrix} 1 & 0.8 & 0 & 0 \\ & 1 & 0 & 0 \\ & & 1 & 0.9 \\ & & & 1 \end{pmatrix},$$

and $(\varepsilon_{Sij}, \varepsilon_{Tij}) \sim N(0, \Sigma)$ with

$$\Sigma = 3 \begin{pmatrix} 1 & 0.8 \\ & 1 \end{pmatrix}.$$

The number of trials was fixed to either 10, 20 or 50, each trial involving 10 subjects randomly assigned to treatment groups. The σ^2 parameter was set to 0.1 or 1.

From Table 2, we see that when the between-trial variability is large ($\sigma^2 = 1$), no convergence problems occur, except when the number of trials is very small. As the between-trial variability gets smaller, convergence problems do arise and worsen as the number of trials decreases.

These simulation results indicate that there should be enough variability at the trial level, and a sufficient number of trials, to obtain convergence of the Newton–Raphson algorithm for fitting mixed-effects models. When these requirements are not fulfilled, one must rely on simpler fixed-effects models, or mixed-effects models with random treatment effects but no random intercepts.

7. EXTENSIONS

In Section 4 we focused on the methodologically appealing case of normally distributed endpoints. In practice, situations abound with binary and time-to-event endpoints, and more generally with surrogate and final endpoints of a different type (Molenberghs, Geys and Buyse, unpublished report). Whereas the linear mixed model (Verbeke and Molenberghs, 1997) provides a unified and flexible framework to analyse multivariate and/or repeated measurements that are normally distributed, similar tools for non-normal outcomes are unfortunately less well developed. For binary outcomes, there are both marginal models such as generalized estimating equations (Liang and Zeger, 1986) or full likelihood approaches (Fitzmaurice and Laird, 1993; Lang and Agresti, 1994; Molenberghs and Lesaffre, 1994; Glonek and McCullagh, 1995) and random-effects models (Stiratelli *et al.*, 1984; Zeger *et al.*, 1988; Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993; Lee and Nelder, 1996). Reviews are given in Diggle *et al.* (1994) and Fahrmeir and Tutz (1995).

Since our developments focus not only on main-effect parameters, such as treatment effects, but prominently on association (random-effects structure and residual covariance structure), standard generalized estimating equations are less relevant. Possible approaches are second-order generalized estimating equations (Liang *et al.*, 1992; Molenberghs and Ritter, 1996) and random-effects models. Since the latter are computationally involved, the likelihood-based approaches need to be supplemented with alternative methods of estimation such as quasi-likelihood. All these issues need to be taken up further in separate papers.

8. Discussion

The validation of surrogate endpoints is a controversial issue. Difficulties have arisen on several fronts: firstly, some endpoints used as surrogates have been shown to provide wholly misleading predictions of the treatment effect upon the important clinical endpoints: the case of encainide and flecainide, two harmful drugs that were approved by the Food and Drug Administration based on their anti-arrhythmic effects, will remain a painful illustration of such an unfortunate circumstance (Fleming, 1992). Secondly, some endpoints that have not been so catastrophically misleading have still failed to explain the totality of the treatment effect upon the final endpoints: the case of the CD4+ lymphocyte counts in patients with AIDS is an example (Choi *et al.*, 1993; De Gruttola *et al.* 1993; Lin *et al.* 1993; De Gruttola *et al.*, 1995). Many of these problems were mentioned in Prentice (1989). All these reasons have led some authors to express reservations about attempts to validate surrogate endpoints statistically (Fleming and DeMets, 1996; De Gruttola *et al.* 1997). Their reservations rest to a large extent on biological considerations: a good surrogate must be shown to be causally linked to the true endpoint, and even so, it is implausible that the surrogate will ever capture the whole effect of treatment upon the true endpoint. These reservations are well taken, but biologically complex situations lend themselves to statistical evaluations that may shed light on the underlying mechanisms involved (Chuang-Stein and DeMasi, 1998). The approach proposed in this paper indirectly addresses these issues: a large individual-level coefficient of determination (R^2_{indiv} close to 1) indicates that the endpoints are likely to be causally linked to each other, while a large trial-level coefficent of determination ($R^2_{\text{trial(f)}}$ close to 1) indicates that a large proportion of the treatment effect is captured by the surrogate.

The approach proposed in this paper provides a quantitative assessment of the value of a surrogate, as well as predictions of the expected effect of treatment upon the true endpoint (Boissel *et al.* 1992; Chen *et al.* 1998). It evaluates the 'validity' of a surrogate in terms of coefficients of determination, which are intuitively appealing quantities in the unit interval. Such an approach is more informative than a mere dichotomization of surrogate endpoints as being 'valid' or 'invalid'. Moreover, the validation procedure no longer requires statistical tests to be statistically significant: for instance, an endpoint with a low individual-level coefficent of determination ($R^2_{\text{indiv}} \ll 1$) is unlikely to be a good surrogate (even if $R^2_{\text{trial(f)}} = 1$), a conclusion that may be reached with a limited number of observations.

The need for validated surrogate endpoints is as acute as ever, particularly in diseases where an accelerated approval process is deemed necessary (Cocchetto and Jones, 1998; Weihrauch and Demol, 1998). Some surrogate endpoints or combinations of endpoints, such as viral load measures combined with CD4+ lymphocyte counts, have in fact already replaced assessment of clinical outcomes in AIDS clinical trials (O'Brien *et al.* 1996; Mellors *et al.* 1997). The approach presented in this paper may offer a better understanding of the worth of a surrogate endpoint, provided that large enough sets of data from multiple randomized experiments are available to estimate the required parameters (Daniels and Hughes, 1997). Large numbers of observations are needed for the estimates to be sufficiently precise, while multiple studies are needed to distinguish individual-level from trial-level associations between the endpoints and effects of interest. However, it has to be emphasized that, even if the results of a surrogate evaluation seem encouraging based on several trials, applying these results to a new trial requires a certain amount of extrapolation that may or may not be deemed acceptable. In particular, when a new treatment is under investigation, is it reasonable to assume that the quantitative relationship between its effects on the surrogate and true endpoints will be the same as with other treatments? The leap of faith involved in making that assumption rests primarily on biological considerations, although the type of statistical information presented above may provide essential supporting evidence.

ACKNOWLEDGEMENT

The authors are grateful to Dr L. Freedman for useful discussions on the validation of surrogate endpoints through meta-analysis, to the Ovarian Cancer Meta-analysis Project and the Pharmacological Therapy for Macular Degeneration Study Group for permission to use their individual patient data, and to Dr R. Wolfinger for his kind assistance with the use of the SAS procedure MIXED.

APPENDIX A

Prediction intervals

Denote $f = E(\beta + b_0 | m_{S0}, a_0) = \beta + D_1 D_2^{-1} D_3$ where D_1, D_2, and D_3 refer to the corresponding matrices in (23). Let f_d be the derivate of f w.r.t. the parameter vector

$$(\beta, \mu_S, \alpha, d_{Sb}, d_{ab}, d_{SS}, d_{Sa}, d_{aa}, \mu_{S0}, \alpha_0)^T.$$

The components of f_d are

$$\begin{aligned}
\frac{\partial f}{\partial \beta} &= 1,\\
\frac{\partial f}{\partial \mu_{S0}} &= -\frac{\partial f}{\partial \mu_S} = D_1 D_2^{-1} \begin{pmatrix} 1 \\ 0 \end{pmatrix},\\
\frac{\partial f}{\partial \alpha_0} &= -\frac{\partial f}{\partial \alpha} = D_1 D_2^{-1} \begin{pmatrix} 0 \\ 1 \end{pmatrix},\\
\frac{\partial f}{\partial d_{Sb}} &= \begin{pmatrix} 1 \\ 0 \end{pmatrix}^T D_2^{-1} D_3,\\
\frac{\partial f}{\partial d_{ab}} &= \begin{pmatrix} 0 \\ 1 \end{pmatrix}^T D_2^{-1} D_3,
\end{aligned}$$

$$\begin{aligned}
\frac{\partial f}{\partial d_{SS}} &= -D_1 D_2^{-1} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} D_2^{-1} D_3,\\
\frac{\partial f}{\partial d_{Sa}} &= -D_1 D_2^{-1} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} D_2^{-1} D_3
\end{aligned}$$

and

$$\frac{\partial f}{\partial d_{aa}} = -D_1 D_2^{-1} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} D_2^{-1} D_3.$$

Denoting the asymptotic covariance matrix of the estimated parameter vector by V, the asymptotic variance of f is given by $f_d V f_d$, producing a confidence interval in the usual way. For a prediction interval, the variance to be used is $f_d V f_d + \mathrm{var}(\beta + b_0 | m_{S0}, a_0)$.

APPENDIX B

SAS code for random-effects model

We describe how to use the SAS statistical software package to fit the random-effects model proposed in Section 4. Note that other packages such as MLwiN are also well suited for fitting this type of multivariate multilevel model and could therefore be utilized instead.

The SAS code to fit model (20)–(21) may be written as follows:

```
PROC MIXED DATA=DATASET COVTEST;
 CLASS ENDPOINT SUBJECT TRIAL;
 MODEL OUTCOME = ENDPOINT ENDPOINT*TREAT / S NOINT;
 RANDOM ENDPOINT ENDPOINT*TREAT / SUB=TRIAL TYPE=UN;
 REPEATED ENDPOINT / SUB=SUBJECT(TRIAL) TYPE=UN;
RUN;
```

The above syntax presumes that there are two records per subject in the input dataset, one corresponding to the surrogate endpoint and the other to the true endpoint. The variable `ENDPOINT` is an indicator for the kind of endpoint (coded 0 for surrogate and 1 for true endpoint) and the variable `OUTCOME` contains measurements obtained from each endpoint. The variable `TREAT` is also assumed to be 0–1 coded.

The `RANDOM` statement defines the covariance matrix D in (19) of random effects at the trial level, while the `REPEATED` statement builds up the residual covariance matrix Σ in (17). Note that the nesting notation in the `SUB=` option is necessary for SAS to recognize the nested structure of the data (subjects are clustered within trials). Acknowledgement of the hierarchical nature of the data enables SAS to build a block-diagonal covariance matrix, with diagonal blocks corresponding to the different trials. This speeds up computations considerably.

REFERENCES

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.

BRESLOW, N. E. AND CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

BUYSE, M. AND MOLENBERGHS, G. (1998). The validation of surrogate endpoints in randomized experiments. *Biometrics* **54**, 1014–1029.

CHEN, T. T. *et al.* (1998). Investigation of disease-free survival as a surrogate endpoint for survival in cancer clinical trials. *Communications in Statistics,* Series A **27**, 1363–1378.

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.

CHUANG-STEIN, C. AND DEMASI, R. (1998). Surrogate endpoints in AIDS drug development: current status (with Discussion). *Drug Information Journal* **32**, 439–448.

COCCHETTO, D. M. AND JONES, D. R. (1998). Faster access to drugs for serious or life-threatening illnesses through use of the accelerated approval regulation in the United States. *Drug Information Journal* **32**, 27–35.

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

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 Syndrome* **6**, 359–365.

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., FLEMING, T. R., LIN, D. Y. AND COOMBS, R. (1997). Validating surrogate markers—are we being naive? *Journal of Infecious Diseases* **175**, 237–246.

DIGGLE, P. J., LIANG, K.-Y. AND ZEGER, S. L. (1994). *Analysis of Longitudinal Data*. Oxford: Clarendon Press.

ELLENBERG, S. S. AND HAMILTON, J. M. (1989). Surrogate endpoints in clinical trials: cancer. *Statistics in Medicine* **8**, 405–413.

FAHRMEIR, L. AND TUTZ, G. (1995). *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer-Verlag.

FITZMAURICE, G. M. AND LAIRD, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika* **80**, 141–151.

FLEISS, J. L. (1993). The statistical basis of meta-analysis. *Statistical Methods in Medical Research* **2**, 121–145.

FLEMING, T. R. (1992). Evaluating therapeutic interventions: some issues and experiences (with discussion). *Statistical Sciences* **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.

GLONEK, G. F. V. AND MCCULLAGH, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society,* Series B **81**, 477–482.

LAIRD, N. M. AND WARE, J. H. (1982). Random effects models for longitudinal data. *Biometrics* **38**, 963–974.

LANG, J. B. AND AGRESTI, A. (1994). Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association* **89**, 625–632.

LEE, Y. AND NELDER, J. A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society,* Series B **58**, 619–678.

LIANG, K.-Y. AND ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

LIANG, K.-Y., ZEGER, S. L. AND QAQISH, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society,* Series B **54**, 3–40.

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, NC: SAS Institute Inc.

MELLORS J. W. *et al.* (1997). Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection. *Annals of Internal Medicine* **126**, 946–954.

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.

MOLENBERGHS, G. AND RITTER, L. (1996). Likelihood and quasi-likelihood based methods for analysing multivariate categorical data, with the association between outcomes of interest. *Biometrics* **52**, 1121–1133.

O'BRIEN W. A., HARTIGAN P. M., MARTIN D., EISNHART J., HILL A., BENOIT S., RUBIN M., SIMBERKOFF M. S. AND HAMILTON J. D. (1996). Changes in plasma HIV-1 RNA and CD4+ lymphocyte counts and the risk of progression to AIDS. *New England Journal of Medicine* **334**, 426–431.

OVARIAN CANCER META-ANALYSIS PROJECT (1991). Cyclophosphamide plus cisplatin versus cyclophosphamide, doxorubicin, and cisplatin chemotherapy of ovarian carcinoma: a meta-analysis. *Journal of Clinical Oncology* **9**, 1668–1674.

OVARIAN CANCER META-ANALYSIS PROJECT (1998). Cyclophosphamide plus cisplatin versus cyclophosphamide, doxorubicin, and cisplatin chemotherapy of ovarian carcinoma: a meta-analysis. *Classic Papers and Current Comments* **3**, 237–43.

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.

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

SENN, S. (1998). Some controversies in planning and analysing multi-centre trials. *Statistics in Medicine* **17**, 1753–1765.

STIRATELLI, R., LAIRD, N. AND WARE, J. (1984). Random effects models for serial observations with dichotomous response. *Biometrics* **40**, 961–972.

THOMPSON, S. G. (1993). Controversies in meta-analysis: the case of the trials of serum cholesterol reduction. *Statistical Methods in Medical Research* **2**, 173–192.

THOMPSON, S. G. AND POCOCK, S. J. (1991). Can meta-analyses be trusted? *Lancet* **338**, 1127–1130.

VERBEKE, G. AND MOLENBERGHS, G. (1997). *Linear Mixed Models in Practice: A SAS-oriented Approach*. Lecture Notes in Statistics 126. New York: Springer-Verlag.

WEIHRAUCH, T. R. AND DEMOL, P. (1998). The value of surrogate endpoints for evaluation of therapeutic efficacy. *Drug Information Journal* **32**, 737–43.

WOLFINGER, R. AND O'CONNELL, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* **48**, 233–243.

ZEGER, S. C., LIANG, K.-Y. AND ALBERT, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**, 1049–1060.

[*Received June 11, 1999. Revised August 23, 1999*]