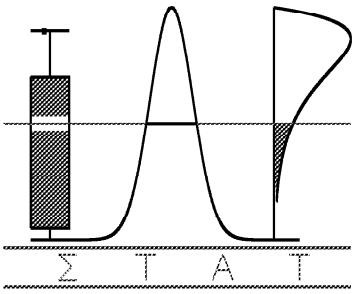
T E C H N I C A L

R E P O R T

0234

Validation of surrogate markers in multiple randomized clinical trials with repeated measures

A. Alonso, H. Geys, M.G. Kenward, G. Molenberghs, and T. Vangeneugden



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Biostatistics (2000), **1**, 1, *pp.* 1–18
Printed in Great Britain

Validation of Surrogate Markers in Multiple Randomized Clinical Trials with Repeated Measurements

By ARIEL ALONSO*, HELENA GEYS, GEERT MOLENBERGHS

Liburgs Universitair Centrum, Center for Statistics,Universitaire Campus B3590, Diepenbeek, Belgium
e-mail: ariel.alonso@luc.ac.be*

MICHAEL G. KENWARD

Medical Statistics Unit London School of Hygiene and Tropical Medicine Keppel Street London WC1E 7HT United Kingdom

TONY VANGENEUGDEN

Johnson Pharmaceutical Research and Development, Turnhoutseweg 30, B-2340 Beerse, Belgium

Summary

Part of the recent literature on the validation of biomarkers as surrogate endpoints proposes to undertake the validation exercise in a multi-trial context which led to a definition of validity in terms of the quality of both trial level and individual level association between the surrogate and the true endpoints (Buyse *et al.*, 2000). These authors concentrated on continuous univariate responses. However, in many randomized clinical studies, repeated measurements are encountered on either or both endpoints. When both the surrogate and true endpoints are measured repeatedly over time, one is confronted with the modeling of bivariate longitudinal data. In this work, we show how such a joint model can be implemented in the context of surrogate marker validation. In addition, another challenge in this setting is the formulation of a simple and meaningful concept of "surrogacy". We propose the use of a new measure, the so-called variance reduction factor, to evaluate surrogacy at the trial and individual level. On the other hand, most of the work published in this area assume that only one potential surrogate is going to be evaluated.We also show that this concept will let us evaluate surrogacy when more than one surrogate variable is available for the analysis. The methodology is illustrated on data from a meta-analysis of five clinical trials comparing antipsychotic agents for the treatment of chronic schizophrenia.

Some key words: Bivariate longitudinal data, Randomized Clinical Trials, Surrogate Marker, Validation

1. Introduction

One of the most important factors influencing the duration and complexity of the process of developing new treatments is the choice of the endpoint, which will be used to assess the efficacy of a treatment. It often happens, however, that the most sensitive and relevant clinical endpoint, the so-called "true" endpoint, is difficult to use in a clinical trial. In that case, the use of the true endpoint might increase the complexity and/or duration of the study. A seemingly attractive solution for this problem is to replace the true endpoint by another one, which may be measured earlier, more conveniently, or more frequently than the endpoints of interest. Such "replacement endpoints" are termed "surrogate" endpoints (Ellenberg and Hamilton 1989).

The dramatic surge of the AIDS epidemic, the pressure for an accelerated evaluation of new therapies, etc., 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. The mere existence of an association between an endpoint and the true endpoint is not sufficient for using the former as a surrogate. What is required is that the effect of the treatment on the surrogate endpoint predicts the effect on the true endpoint. Unfortunately, partly due to the lack of appropriate methodology, this condition was not checked in earlier attempts to use surrogates. As a consequence the use of surrogates has led to misleading and even harmful conclusions. In cardiovascular disease, for example, the unsettling discovery that the two major antiarrhytmic drugs encanaide and flecanaide reduced arrhythmia but caused a more than 3-fold increase in overall mortality stressed the need for caution in using non-validated surrogate markers in the evaluation of the possible clinical benefits of new drugs (The Cardiac Arrhytmia Suppression Trial (CAST) Investigators 1989). This and other examples of unsuccessful replacement of true endpoints led to the scepticism about usefulness of surrogate endpoints, both among statisticians, as well as clinicians.

However, the need to develop new drugs and treatments as quickly as possible is still present in medical research. After all, clinicians and patients want to use effective treatments as soon as possible and shortening the duration of the experiments limits possible problems with non-compliance and missing data and therefore increases effectiveness and reliability of the research. Another reason for shortening the duration of the process of testing new therapies may be related to new discoveries in medicine and biology, which creates a possibility for development of many potentially effective treatments for a particular disease. In that case, there may be a need to cope with a large number of new promising treatments that should be quickly evaluated with respect to their efficacy. Finally, an important area of potential application of surrogate endpoints is the assessment of safety of new treatments. Duration and sample size of clinical trials aimed at development of new drugs are usually insufficient to detect rare or late adverse effects of the treatment (Jones 2001; Dunn and Mann 1999). The use of surrogate endpoints might allow to obtain information about such effects even during the clinical testing phase.

For the above reasons, it is difficult to abandon the idea of using surrogate endpoints altogether. The failed past attempts make clear, however, that before deciding to use a surrogate, it is of the utmost importance to investigate its validity. Recent literature on the validation of biomarkers as surrogate endpoints has focused on different points of view. Prentice (1989) defines surrogacy in terms of the equivalence of hypothesis tests for treatment effects and proposes operational criteria for his definition. Freedman, Graubard and Schatzkin (1992) introduced the *proportion explained* to quantify how much of the treatment effect on the true endpoint is captured by the surrogate endpoint. Buyse and Molenberghs (1998) decomposed the proportion explained into the *relative effect* and *adjusted association* and argued in favor of these quantities instead. These proposals were formulated assuming that the validation of a surrogate is based on data from a

single randomized clinical trial. This leads to problems with untestable assumptions and too low statistical power. To overcome these problems, Albert *et al.* (1998) suggested to combine information from several groups of patients (multi-center trials or meta-analyses). This was implemented by Daniels and Hughes (1997), Gail *et al.* (2000) and Buyse *et al.* (2000). The latter suggested a multi-trial approach that led to a new definition of validity in terms of the quality of both trial level and individual level association between the surrogate and the true endpoint. In their approach, the quality of a surrogate at the trial level is assessed by means of a coefficient of determination R^2_{trial}. At the individual level, the squared correlation R^2_{ind} between the surrogate and true endpoint, after adjustment for both the trial effects and the treatment effects is used. A surrogate will be said to be good when both R^2_{trial} and R^2_{ind} are sufficiently high.

However most of the previous work have been focusing on univariate responses for the surrogate and true endpoints. Going from an univariate setting to a multivariate framework represents new challenges. The R^2 measurements proposed by Buyse *et al.* to evaluate surrogacy at the trial and individual level are no longer applicable and new concepts are needed. These authors proposed their methodology based on the simplest case in which both, the surrogate and the true endpoint, are continues and normally distributed. Posteriorly, different applications were implemented in different settings where the true endpoint and/or the surrogate variable were binary, time to event responses, mixture of binary and continues, etc.

Nevertheless, in all the previous cases, it was always assumed that both endpoints could be characterized by a single random variable. It was also assumed that only one potential surrogate was available for the analysis and finally that treatment effect on both responses was constant over time and could be characterized by an univariate parameter.

The previous assumptions can fail when we have repeated measures per patient like it is the case in longitudinal studies. The objective of this paper is to study surrogate and true endpoint that are both longitudinal. To this end, an additional challenge is to summarize "surrogacy" in simple measures. We propose the use of the so-called *variance-reduction factor* (VRF). Technically, a joint model for multivariate repeated measurements is required. Useful references on this topic include Galecki (1994), Sy, Taylor and Cumberland (1997), Jorgensen *et al.* (1996, 1999).

The paper is organised as follows: Section 2 introduces a joint model for bivariate longitudinal data. Section 3 defines the variance reduction factor to evaluate surrogacy when repeatedly measurements for surrogate and true endpoints are taken. Section 4 illustrates the methodology on data from a meta-analysis of randomized clinical trials comparing antipsychotic agents for the treatment of chronic schizophrenia.

2. Model Formulation

In many practical applications, repeated measurements are encountered on either or both endpoints. In analogy to the bivariate normal setting considered by Buyse *et al.* (2000), we will base the calculation of surrogacy measures on a two-stage approach rather than a full random effects approach, to reduce numerical complexity. Technically, we need (1) a model for bivariate longitudinal outcomes, and (2) new measures that let us evaluate surrogacy when longitudinal data is available. In this section we focus on the former issue and introduce a possible joint model for bivariate longitudinal outcomes along the ideas of Galecki (1994). An advantage of this approach is that it can be easily implemented within standardly available software programs. The extension towards more flexible modelling structures for bivariate longitudinal data is the topic of future research.

In the case of univariate longitudinal endpoints one can consider different types of covariance structures, including compound symmetry, autoregressive, banded, factor-analytic, spatial, unstructured, etc. Here, however, we have repeated measurements on two outcome variables, the surrogate and the true endpoint. A possible joint covariance structure can then be based on the Kronecker product of (1) an unstructured covariance structure for the type of outcome and (2) a suitable covariance structure for the repeated measurements on an outcome. While, in the setting of Buyse *et al.* (2000) the error covariance structure could be assumed constant over all trials, this assumption is no longer plausible in most practical longitudinal settings. Measures could be taken at different time points within different trials, the number of measurements could be different in each trial, etc. Therefore, we will allow for different covariance structures over the different trials.

Suppose we have data from $i = 1, \ldots, N$ trials in the ith of which $j = 1, \ldots, n_i$ subjects are enrolled and further suppose that t_{ij} is the time at which subject j in trial i was measured. Let T_{ijt} and S_{ijt} denote the associated true and surrogate endpoints, respectively, and let Z_{ij} be a binary indicator variable for treatment. Following the ideas of Galecki (1994), a possible joint model for both responses can then be written as:

$$\begin{cases} T_{ijt} = \mu_{T_i} + \beta_i Z_{ij} + g_{T_i}(t_{ij}) + \varepsilon_{T_{ijt}} \\ S_{ijt} = \mu_{S_i} + \alpha_i Z_{ij} + g_{S_i}(t_{ij}) + \varepsilon_{S_{ijt}} \end{cases}, \tag{1}$$

where μ_{S_i} and μ_{T_i} are trial-specific intercepts, α_i, β_i are trial-specific effects of treatment Z_{ij} on the two endpoints and g_{T_i} and g_{S_i} are trial-specific time functions in trial $i = 1, \ldots, N$ that could contain random effects. The vectors $\tilde{\varepsilon}_{T_{ij}}$ and $\tilde{\varepsilon}_{S_{ij}}$ are correlated error terms, assumed to be jointly mean-zero multivariate normally distributed with covariance matrix

$$\Sigma_i = \begin{pmatrix} \sigma_{TTi} & \sigma_{STi} \\ \sigma_{STi} & \sigma_{SSi} \end{pmatrix} \otimes R_i. \tag{2}$$

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

$$R_i = \begin{pmatrix} 1 & \rho_i & \ldots & \rho_i^n \\ \vdots & \vdots & \vdots & \vdots \\ \rho_i^n & \rho_i^{n-1} & \ldots & 1 \end{pmatrix}.$$

It is clear from the previous model that the correlation structure for $\varepsilon_{T_{ijt}}$ and $\varepsilon_{S_{ijt}}$ is much more complicated in this setting that the one obtained in the simpler case in which both endpoints were measured at only one time point. As it will be shown later the more complex nature of our data will require a new approach to the problem of evaluating surrogacy at the individual level. Graphically this correlation may be illustrated as:

$$\begin{array}{ccccc} \varepsilon_{T_{ijt}} & \longleftarrow & \rho_i^m & \longrightarrow & \varepsilon_{T_{ijt'}} \\ \uparrow & \nwarrow & & \nearrow & \uparrow \\ \rho_{TSi} & & \rho_{TSi}\rho_i^m & & \rho_{TSi} \\ \downarrow & \swarrow & & \searrow & \downarrow \\ \varepsilon_{S_{ijt}} & \longleftarrow & \rho_i^m & \longrightarrow & \varepsilon_{S_{ijt'}} \end{array} \tag{3}$$

where t and t' are two different time points, $m = t' - t$, and $\rho_{TSi}^2 = \dfrac{\sigma_{TSi}^2}{\sigma_{SSi}\sigma_{TTi}}$.

It should be noted that, if we only have one observation per subject, the variable time will disappear from equation (1) and $R_i = \mathbf{1}$. If it is also assumed that $\Sigma_i = \Sigma$ then our model reduces to the model proposed by Buyse *et al.* (2000).

Due to replication at the trial level, we can impose a distribution on the trial-specific parameters. At the second stage, we therefore assume

$$\begin{pmatrix} \mu_{S_i} \\ \mu_{T_i} \\ \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} m_{S_i} \\ m_{T_i} \\ a_i \\ b_i \end{pmatrix}, \tag{4}$$

where the second term on the right-hand side is assumed to follow a zero-mean normal distribution with covariance matrix D.

In the special case of a single measurement per response, Buyse *et al.* (2000) examined the validity question at each of these two levels. They argue that a 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 and that it is therefore essential to explore the quality of the prediction of the treatment effect on the true endpoint by information obtained in the validation process based on trials $i = 1, \ldots, N$ and by information available on the surrogate endpoint in a new trial $i = 0$, say. A measure to assess the quality of a surrogate at the trial level is then calculated based on some of the elements of D. It is given by the coefficient of determination

$$R^2_{\text{trial}} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}}. \tag{5}$$

This coefficient measures how precisely the effect of treatment on the true endpoint can be predicted, provided that the treatment effect on the surrogate endpoint has been observed in a new trial ($i = 0$). It is unitless and ranges in the unit interval if the corresponding variance-covariance matrix D is positive-definite, two desirable features for its interpretation. The association between the surrogate and final endpoints after adjustment for the effect of treatment is captured by

$$R^2_{\text{ind}} = \frac{\sigma_{ST}^2}{\sigma_{SS}\sigma_{TT}}, \tag{6}$$

which is simply the squared correlation between S and T, after accounting for trial and treatment effects.

Although the inclusion of fixed trial-specific treatment coefficients in our model enables us to estimate R^2_{trial} at the trial level, at the individual level the R^2_{ind} proposed by Buyse *et al* (2000) is no longer applicable and new proposals are needed. Even at the trial level extensions may be necessary for more complicated models where treatment effects may vary over time. Hence, there is a clear need for alternative approaches to summarize "surrogacy" in simple yet meaningful measures. In the next section, we propose the use of the so-called variance reduction factor (VRF) to this effect.

3. Variance Reduction Factor

In this section, we will first define a new measure of validity at the individual level. Later, it will be shown how this can be easily translated into a validity measure at the trial level.

From Section 2 we know that, in general, the error vector $\widetilde{\varepsilon}_{T_{ij}}$ and $\widetilde{\varepsilon}_{S_{ij}}$ follow a multivariate normal distribution with variance-covariance matrix:

$$\Sigma_i = \begin{pmatrix} \Sigma_{TTi} & \Sigma_{TSi} \\ \Sigma_{TSi}^T & \Sigma_{SSi} \end{pmatrix}$$

where Σ_{TTi} and Σ_{SSi} are the variance-covariance matrices associated with the residual vectors $\widetilde{\varepsilon}_{T_{ij}}$ and $\widetilde{\varepsilon}_{S_{ij}}$ respectively and Σ_{TSi} contains the covariances between the elements of $\widetilde{\varepsilon}_{T_{ij}}$ and the elements of $\widetilde{\varepsilon}_{S_{ij}}$. Hence, we allow for a different covariance structure in each clinical trial, thus leaving the possibility to tackle very general problems for which the assumption of homogeneous covariance structures over trials would be overly restrictive.

To validate a surrogate endpoint at the individual level in an univariate setting, Buyse *et al.*(2000) suggested to look at the correlation between the surrogate and the true endpoint after adjustment for trial and treatment effects. Instead, we propose a new concept, named the *Variance Reduction Factor* (VRF). Essentially, we summarize the variability of the repeated measurements on the true endpoint by the trace of its variance-covariance matrix and sum this over all trials. In a similar way, we summarize the conditional variability of the true endpoint measurements, given the surrogate by the trace of the conditional variance-covariance matrix summed once more over trials. Following these ideas the relative reduction in the true endpoint variance after adjusting by the surrogate can be quantified as:

$$VRF_{\text{ind}} = \frac{\sum_i \{\text{tr}(\Sigma_{TTi}) - \text{tr}(\Sigma_{(T|S)i})\}}{\sum_i \text{tr}(\Sigma_{TTi})}, \tag{7}$$

where $\Sigma_{(T|S)_i}$ denotes the conditional variance-covariance matrix of $\widetilde{\varepsilon}_{T_{ij}}$ given $\widetilde{\varepsilon}_{S_{ij}}$: $\Sigma_{(T|S)i} = \Sigma_{TTi} - \Sigma_{TSi}\Sigma_{SSi}^{-1}\Sigma_{TSi}^T$. Intuitively, expression (7) tries to quantify how much of the total variability around the repeated measurements on the true endpoint is explained by adjusting for the treatment effects and the repeated measurements on the surrogate endpoints. In that respect, expression (7) fits into the general definition of the "proportion of variation of a dependent variable, Y, explained by a vector of covariates X" (PVE) in general regression models:

$$PVE = \frac{\sum\{D(Y_i) - D(Y_i|X_i)\}}{\sum D(Y_i)},$$

where $D(Y_i)$ denotes a measure of distance of Y_i from a central location parameter of the estimated marginal distribution of Y and $D(Y_i|X_i)$ denotes the same measure using distributions of Y

conditional on a given model and on the covariate vector for the ith observation (Schemper and Stare 1996).

Further one can show (i) that the VRF_{ind} ranges between zero and one, (ii) that the VRF_{ind} equals zero if and only if the error terms of the true and surrogate endpoints are independent within each trial, (iii) that the VRF_{ind} equals one if and only if there exists a deterministic relationship between the error terms of the true and surrogate endpoints within each trial and finally (iv) that the VRF_{ind} reduces to the R^2_{ind} when the endpoints are measured only once. The proofs of these properties are deferred to the appendix.

Next, suppose that p_i denotes the number of designed time points at trial i and consider the covariance structure (2), then we have:

$$\begin{aligned} \text{tr}(\Sigma_{TSi}\Sigma_{SSi}^{-1}\Sigma_{TSi}^{T}) &= \frac{\sigma_{TSi}^2}{\sigma_{SSi}}p_i, \\ \text{tr}(\Sigma_{TTi}) &= \sigma_{TTi}p_i. \end{aligned}$$

Thus, the VRF_{ind} can be rewritten in terms of the correlations (ρ_{TSi}) between surrogate and true endpoints at each time point at the different trials $i = 1, \ldots, N$:

$$VRF_{\text{ind}} = \sum_i \left(\frac{p_i\sigma_{TTi}}{\sum_i p_i\sigma_{TTi}} \right) \rho_{TSi}^2$$

The latter expression yields an appealing interpretation of the VRF. Indeed, the VRF is just a sum of different trial contributions, in which each contribution is just the product of the squared correlation between the surrogate and the true endpoint at each time point in that trial with the proportion of the total true endpoint variance that is accounted for by that trial.

In addition, the VRF can be incorporated into a much more general framework that allows interpretation in terms of the canonical correlations of the error term vectors. Indeed, if at trial i we have p_i time points then we will also have $t = 1, \ldots, p_i$ canonical correlations $\rho_t^{i^2}$ for $(\widetilde{\varepsilon}_{\mathbf{Tij}}, \widetilde{\varepsilon}_{\mathbf{Sij}})$ such that:

$$\rho_1^{i^2} \geq \rho_2^{i^2} \geq \ldots \geq \rho_{p_i}^{i^2}$$

and $\rho_t^{i^2}$ are the eigenvalues of $\Sigma_{TTi}^{-1/2}\Sigma_{TSi}\Sigma_{SSi}^{-1}\Sigma_{TSi}^{T}\Sigma_{TTi}^{-1/2}$. Now, one can show that the VRF can be written as a linear combination over all trials and over all timepoints within a trial of the canonical correlations of the error terms. The coefficients in this linear combination need to be positive and sum to 1. The investigation of advantages and disadvantages of this canonical correlation framework as well as the potential extension to non-normal data will be a topic of further research.

As mentioned before, as soon as the treatment effect cannot be assumed to be constant over time, the classical multi trial approach becomes inapplicable as well at the trial level and other approaches are needed. In this case the treatment effect at the *ith* trial could not be characterized by the scalars β_i and α_i but by the p_i dimensional vectors $\widetilde{\beta}_i$ and $\widetilde{\alpha}_i$, Verbyla (1999).

For reasons explained earlier it would then be unrealistic to assume that the variance-covariance matrix D is constant over the trials. In that case we define the Variance Reduction Factor at the trial level, (VRF_{trial}) as follows, suppose that

$$\begin{pmatrix} \widetilde{\beta}_i \\ \widetilde{\alpha}_i \end{pmatrix} \sim N\left(\begin{pmatrix} \bar{\beta}_i \\ \bar{\alpha}_i \end{pmatrix}, D_i \right),$$

with

$$D_i = \begin{pmatrix} D_{\beta\beta i} & D_{\beta\alpha i} \\ D_{\beta\alpha i}^T & D_{\alpha\alpha i} \end{pmatrix},$$

Here $(\bar{\beta}_i, \bar{\alpha}_i)$ is the $2p_i$ dimensional mean treatment effect vector at the ith trial. Then we can define, similarly to the individual level and with straightforward notations, VRF_{trial} as:

$$VRF_{trial} = \frac{\sum_i \left\{ \text{tr}(D_{\beta\beta i}) - \text{tr}(D_{(\beta|\alpha)i}) \right\}}{\sum_i \text{tr}(D_{\beta\beta i})} \tag{8}$$

The properties stated above can now be easily extended for the trial level and in case of a single normally distributed endpoint it can be shown that $VRF_{\text{trial}} = R^2_{\text{trial}}$.

The scope of this methodology is not limited to the longitudinal framework, there are other settings in which the used of these tools can be appealing. Most of the work published in this area assume that only one potential surrogate is going to be evaluated. However in many practical situations the analyst has to study surrogacy in a multivariate framework, for instance, it is plausible to think that a treatment can affect a medical condition in a very complex way acting at the same time on different factors. Therefore it would be sensible to presume that prediction of the treatment effect on the true endpoint can be substantially improved if we use the information about the treatment effect not only on a single surrogate but on a whole set of possibly relevant variables.

Let us consider again the setting used by Buyse *et al.*(2000) to introduce their R^2 measurements but assuming that two potencial surrogates are now available. At the first stage the following multivariate regression model is assumed:

$$\begin{cases} T_{ij} = \mu_{T_i} + \beta_i Z_{ij} + \varepsilon_{T_{ij}} \\ S_{1ij} = \mu_{S_{1i}} + \alpha_{1i} Z_{ij} + \varepsilon_{S_{1ij}} \\ S_{2ij} = \mu_{S_{2i}} + \alpha_{2i} Z_{ij} + \varepsilon_{S_{2ij}} \end{cases}, \tag{9}$$

where

$$\begin{pmatrix} \varepsilon_{T_{ij}} \\ \varepsilon_{S_{1ij}} \\ \varepsilon_{S_{2ij}} \end{pmatrix} \sim N(0, \Sigma).$$

At the second stage we will assume that

$$\begin{pmatrix} \beta_i \\ \alpha_{1i} \\ \alpha_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} \beta \\ \alpha_1 \\ \alpha_2 \end{pmatrix}, D \right), \tag{10}$$

where

$$D = \begin{pmatrix} 2\sigma + \theta & \sigma & \sigma \\ \sigma & \sigma & 0 \\ \sigma & 0 & \sigma \end{pmatrix}$$

If we now applied the methodology proposed by Buyse *et al.*(2000) using both surrogates independently then it is not difficult to show that

$$R^2_{1\text{trial}} = R^2_{2\text{trial}} = \frac{\sigma}{2\sigma + \theta}$$

whereas if both of them are jointly used and we also use the VRF concept proposed in the present work to evaluate surrogacy then we obtain

$$VRF_{\text{trial}} = \frac{2\sigma}{2\sigma + \theta}$$

Finally if we also notice that $V(\beta_i|\alpha_{1i}, \alpha_{2i}) = \theta$ then it is clear that for small values of θ there is almost a deterministic relationship between β_i and $(\alpha_{1i}, \alpha_{2i})$. This will imply that we should be able to predict the treatment effect on the true endpoint with a high precision if the treatment effect on both surrogates S_1 and S_2 is known. However, these surrogates would poorly predict the treatment effect on the true endpoint if they were considered independtly as can be concluded from the expressions

$$\lim_{\theta \to 0} R^2_{1\text{trial}}(\theta) = \lim_{\theta \to 0} R^2_{2\text{trial}}(\theta) = 0{\cdot}5$$

On the other hand, the VRF_{trial} clearly reflects that, in this setting, a very accurate prediction for the true endpoint treatment effect can be obtained if both endpoints are jointly used.

$$\lim_{\theta \to 0} VRF_{\text{trial}}(\theta) = 1$$

The previous example illustrates that a lot can be gained in some practical situations if more that one single surrogate is used. The methodology proposed here will let us approach the surrogacy problem from a new point of view. In principle, any number of potential surrogates could be studied and even severals endpoints and several surrogates could be analized in a multivariate framework what could considerably improve our prediction's capabilities.

4. Case Study: a Meta-analysis of Trials in Schizophrenic Subjects

In this section we apply the proposed definition to individual patient data from a meta-analysis of five double-blind randomized clinical trials, comparing the effects of risperidone to conventional antipsychotic agents for the treatment of chronic schizophrenia. Only subjects who received doses of risperidone (4-6 mg/day) or an active control (haloperidol, perphenazine, zuclopenthixol) were included in the analysis. Depending on the trial, treatment was administered for a duration of 4 to 8 weeks.

Our meta-analysis contains five trials. This is insufficient to apply the meta-analytic methods described in previous sections. Fortunately, in all the trials information is also available on the countries where patients were treated. Hence, we can use country within trial as unit of analysis. A total of 20 units are thus available for analysis, with the number of patients ranging from 9 to 128. The number of patients per country is tabulated in Table 1.

Table 1. *Number of Patients per Country-unit*

Country Id	1	2	3	4	5	6	7	8	9	10
# Patients	31	29	26	44	44	9	37	32	68	49
Country Id	11	12	13	14	15	16	17	18	19	20
# Patients	43	21	25	39	36	17	33	69	30	128

The choice of the unit is an important issue and it is not free of certain controversy. It can depend on practical reasons, such as the information available in the data set at hand and also on expert's considerations about the most suitable unit for a specific problem.

In general, the choice of the unit should be made considering different aspects like physician's opinion, statistical ideas, information available in our data and so on. Ideally, both the number of units and the number of patients per units should be sufficiently large to avoid numerical problems.

Several measures can be considered to assess a patient's global condition. The Clinician's Global Impression (CGI) is generally accepted by practitioners as a reliable clinical measure of patient's status. This is a 7-grade scale used by the treating physician to characterize how well a subject has improved.

Other useful and sufficiently sensitive assessment scale is the Positive and Negative Syndrome Scale (PANSS). PANSS consists of 30 items that provide an operationalized, drug-sensitive instrument, which is highly useful for both typological and dimensional assessment of schizophrenia.

Even though this is not a standard situation for surrogate validation due to the lack of a clear "gold" standard, we consider as our primary measure (true endpoint) the Clinician's Global Impression scale which is the one that has the clearest clinical interpretation.

It is important to notice that even though in this case a clear "gold" standard is not available, our analysis will let us address some very important issues. At the trial level it will allow a flexible assessment of a common question among practitioners, i.e. how a treatment effect on PANSS can be translated into a treatment effect on CGI which is easier to interpret clinically. On the other hand, at the individual level a VRF equal to one will imply that the variability of CGI conditional on PANSS and the treatment effect is equal to zero. In other words, it would mean that CGI could be estimated without error from PANSS. Other values of the VRF will give us different levels of evidence about how strong the association between both scales is.

In our model we use log(CGI) and log(PANSS) instead of the original variables to stabilize the variances. Figure 1 shows the individual's profiles for log(CGI) and log(PANSS) by treatment groups. In all the panels a linear time trend seems plausible.

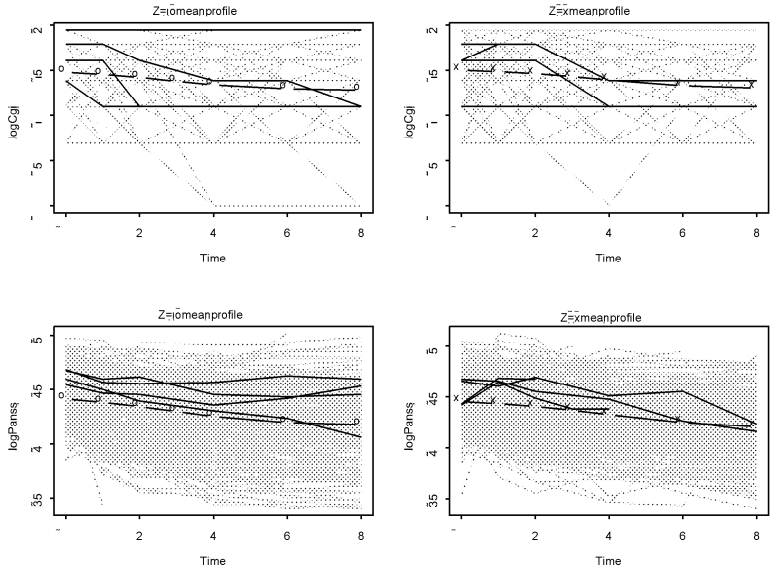


Fig. 1. log(CGI) and log(PANSS): Mean profiles.

We applied the two-stage approach introduced in Section 2 to these data. At the first stage different choices of g_{Ti} and g_{Si} can be considered, each of them leading to different bivariate joint model. Four different models were fitted, here $k = 1, 2$ denote the true endpoint (CGI) and the surrogate scale (PANSS) respectively

1. Linear trend over time within each trial: $g_{ki}(t) = \theta_{ki}t$
2. Random intercept model: This model assumes a linear trend over time and independent random intercepts are considered for each scale within each trial, $g_{ki}(t) = \theta_{ki}t + b_{ki}$
3. Random intercept and slope model: A linear trend over time and independent random intercepts and slopes are considered for each scale within each trial , $g_{ki}(t) = \theta_{ki}t + b_{k0i} + b_{k1i}t$
4. General trend over time modeled using splines via random effects as proposed by Verbyla *et al*(1999), $g_{ki}(t) = \text{lin}_{ki}(t) + \text{spl}_{ki}(t)$

The AIC criteria was then used to select the best model in each trial. Model (1) showed to have the best performance in all the trials. The comparison of model (1) with the bivariate cubic smoothing splines model showed that, for the data at hand, a linear trend over time seems to be a good model for the mean structure of both scales in all the trials which is in total agreement with the profiles displayed in figure 1.

The estimated log(CIG) variance components ($\hat{\sigma}_{TTi}$), the estimated log(PANSS) variance components ($\hat{\sigma}_{SSi}$), the log(CGI) – log(PANSS) correlation as well as ρ_i parameter, separately for

each unit were obtained. All these variance components are plotted in Figure 2, which clearly shows that the assumption of a constant covariance structure over all trials is not really plausible, as already suggested before.

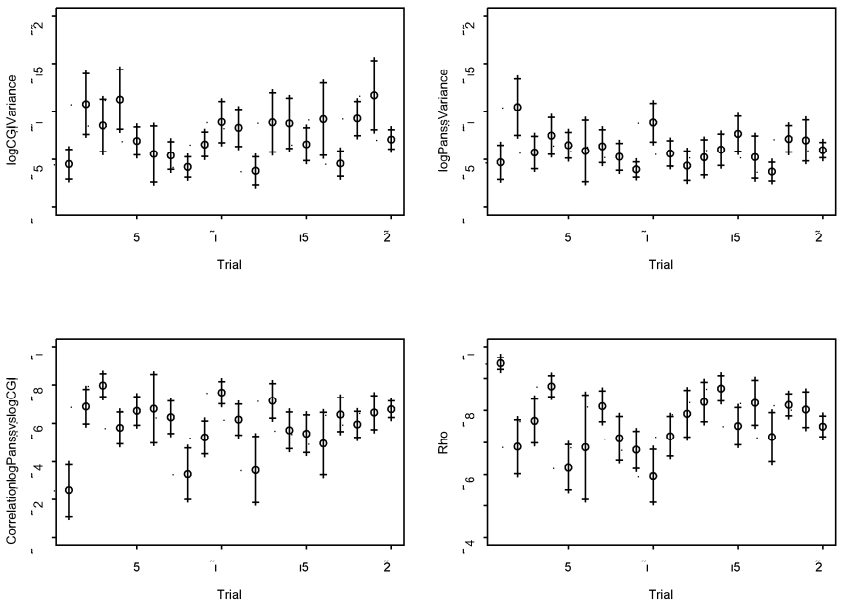


Fig. 2. Variance Components.

If we now want to study the relationship between the log(CGI) and the log(PANSS), then it is clear that the R^2_{ind} measure proposed by Buyse *et al.* (2000) is no longer useful in such a general situation with a complex variance-covariance structure for the bivariate longitudinal data which cannot be assumed to be constant over trial. In contrast, the VRF_{ind} that we proposed in Section 3 does provide an adequate summary measure for the validation at the individual level. By applying the two-stage approach based on model 1 we obtained an estimate for VRF of 0·39 (95% confidence interval: $]0{\cdot}36; 0{\cdot}41[$).

This shows that after adjusting by the surrogate log(PANSS) there is a relative reduction in the marginal variance of log(CGI) of 39 percent. Of course, this should be interpreted as an "average" reduction due to the fact that we are summing over trials. Hence, log(PANSS) seems to be a rather poor surrogate for log(CGI) at the individual level.

Our procedure also allows to estimate the contribution of each trial to the meta-analytic VRF. Within each unit we can define

$$VRF^i_{\text{ind}} = \frac{\text{tr}(\Sigma_{TTi}) - \text{tr}(\Sigma_{(T|S)i})}{\text{tr}(\Sigma_{TTi})},$$

The first panel of figure 3 shows the different trial contributions as well as the VRF meta-analytic value. From the graph it is clear that in most of the trials there was a relative weak

association between the surrogate and the true endpoint, which values of the VRF smaller than 0·6 in almost all the cases.

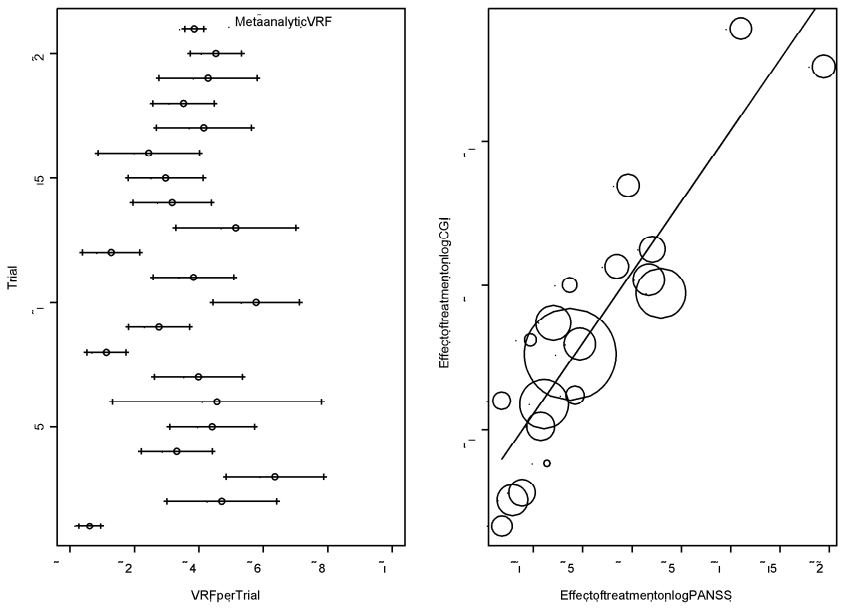


Fig. 3. First panel: VRF trials contributions (VRF^2_{ind}) and Meta-analytic VRF. Second panel: Treatment effect for BPRS vs treatment effect for PANSS

At the trial level the results are much more encouraging. Since treatment is assumed not to vary with time, in this case the R^2_{trial} as introduced by Buyse *et al.* (2000) can still be calculated. We thus find a value of R^2_{trial} of 0·85. The resulting correlation between treatment effects on $\log(CGI)$ and $\log(PANSS)$ equals 92% suggesting that a reliable prediction can be made of the treatment effect on $\log(CGI)$ having observed the treatment effects on $\log(PANSS)$. Graphically this is represented in the second panel of figure 3 which plots the treatment effects on log(CGI) by the treatment effects on log(PANSS). The size of each point is proportional to the number of patients within a unit.

A 95% confidence interval for R^2_{trial} was obtained using bootstrap. The so-obtained confidence limits for R^2_{trial} are $[0{\cdot}68; 0{\cdot}95]$, which shows that the trial-level association is estimated rather precisely.

5. Concluding Remarks

In the past decade, research on the use of surrogate endpoints concentrated mainly on the development of criteria and methods of validation for surrogate endpoints. The use of a meta-analysis approach, as introduced by Daniels and Hughes (1997), Gail *et al.* (2000) and Buyse *et al.* (2000) was a promising way forward compared to the single-trial approaches that were proposed previously and that coped with serious conceptual problems (Choi *et al.* 1993; Lin, Fleming and De Gruttola 1997; Flandre and Saidi 1998; Buyse *et al.* 2000; Molenberghs *et al.* 2001).

However most of the previous work have been focusing on univariate responses for the surrogate and true endpoints. Going from an univariate setting to a multivariate framework represents new challenges. In this paper, we proposed a new concept to validate surrogate endpoints within the meta-analytic framework but in more complicated contexts. Up to now most of the research developed in this area assume that only one potential surrogate is going to be evaluated. However in many practical situations the analyst has to study surrogacy in a multivariate framework. The example constructed in Section 3 clearly showed that a lot can be lost if we limited ourself to analyze surrogacy univariately.

The VRF concept introduced here to evaluate surrogacy when repeated measurements are present in both endpoints gives us the possibility of approaching the surrogacy problem from a new point of view. In principle, any number of potential surrogates could be studied and even severals endpoints and several surrogates could be canalized in a multivariate framework what could considerably improved our prediction's capabilities.

In some cases there is not totally clear idea about which variable or variables could be the best possible surrogate for certain endpoint of interest. The VRF could let us explore which subset of our potential surrogates would be the most suitable one. Another important limitation found in the currently literature about this topic is that most of the technics are designed for the special case in which only two treatments are considered. However, in some medical field the use of three or more treatments in clinical trials is a common practice. The tools introduced here will let us study surrogacy also in this setting.

The implementation of this methodology implies that bivariate longitudinal models should be fitted, within each unit, at the first stage of our analysis. Statistical methods and estimation techniques are well developed for repeated measures on a univariate normal variable, and lately much research has been dedicated to repeated observations on a binary variable and more generally on variables with distributions in the exponential family.

However, for multivariate longitudinal responses less has been done. General models for this situation are necessarily complex as two types of correlations must be taken into account: correlations between measurements on different variables at each occasion and correlations between measurements at different occasions. Matsuyama and Ohashi (1997) and Heitjan and Sharma (1997) considered models for normally distributed responses. Rochon (1996) demonstrates how generalized estimating equations can be used to fit extended marginal models for bivariate repeated measures of discrete or continues outcomes. Rochon's approach is very general and allows for a large class of response distributions. However not that many diagnostics tools are available yet in this setting. In the present work some univariate plots of the residuals did not show problems of lack of fit for our final model.

Finally it is important to notice that in our specific example PANSS can be considered continues given the large number of items. However more debate surrounds the CGI scale. Although many researches might argue that a 7-itemed scale can be considered continues, others might find this an unrealistic assumption. In the present work we have followed historical papers in which CGI has been treated as a continues scale and the results obtained seem to be biologically plausible.

On the other hand fitting a joint model to analyze mixtures of discrete and continues responses in a longitudinal framework is a challenging task. Difficulties in joint modeling of responses of different types arise because of the need of a specify multivariate joint distribution for the outcome variables. Most research so far have concentrated on simultaneous analyzes of binary and continues responses. Further extensions of our methodology using models for different types of responses are necessary and will be the objective of future work.

Appendix

Properties of VRF_{ind}

(a) $0 \leq VRF_{\text{ind}} \leq 1$

Proof

(a.1) $VRF_{\text{ind}} \geq 0$

$$VRF_{\text{ind}} = \frac{\sum_i \{\text{tr}(\Sigma_{TTi}) - \text{tr}(\Sigma_{(T|S)i})\}}{\sum_i \text{tr}(\Sigma_{TTi})} = 1 - \frac{\sum_i \text{tr}(\Sigma_{(T|S)i})}{\sum_i \text{tr}(\Sigma_{TTi})} \tag{11}$$

and after some transformation it is possible to obtain

$$VRF_{\text{ind}} = \frac{\sum_i \text{tr}(\Sigma_{TSi}\Sigma_{SSi}^{-1}\Sigma'_{TSi})}{\sum_i \text{tr}(\Sigma_{TTi})} \tag{12}$$

so $VRF_{\text{ind}} \geq 0$ if and only if

$$\sum_i \text{tr}(\Sigma_{TSi}\Sigma_{SSi}^{-1}\Sigma'_{TSi}) \geq 0 \tag{13}$$

It is easy to see that $diag(\Sigma_{TSi}\Sigma_{SSi}^{-1}\Sigma'_{TSi}) = (f_{i_1}\Sigma_{SSi}^{-1}f'_{i_1}, \ldots, f_{i_{m_i}}\Sigma_{SSi}^{-1}f'_{i_{m_i}})$ where f_{i_s} is the s row of Σ_{TSi} and $diag(A)$ is the diagonal of A. The previous statements imply

$$\sum_i \text{tr}(\Sigma_{TSi}\Sigma_{SSi}^{-1}\Sigma'_{TSi}) = \sum_i \sum_s f_{i_s}\Sigma_{SSi}^{-1}f'_{i_s} \tag{14}$$

and (13) follows from the fact that $f_{i_s}\Sigma_{SSi}^{-1}f_{i_s} \geq 0 \quad \forall(i,s)$

(a.2) $VRF_{\text{ind}} \leq 1$ follows from (11) and the inequalities $\text{tr}(\Sigma_{(T|S)i}), \text{tr}(\Sigma_{TTi}) \geq 0$.

(b) $VRF_{\text{ind}} = 0 \Leftrightarrow \forall i \quad \widetilde{\varepsilon}_{T_{ij}} \quad and \quad \widetilde{\varepsilon}_{S_{ij}}$ are independent.

Proof

Let's first notice that $(\widetilde{\varepsilon}_{T_{ij}}, \widetilde{\varepsilon}_{S_{ij}})$ are independent if and only if $\Sigma_{TSi} = \mathbf{0}$. Combining (12) and (14) we get

$$VRF_{\text{ind}} = \frac{\sum_i \sum_s f_{i_s}\Sigma_{SSi}^{-1}f'_{i_s}}{\sum_i \text{tr}(\Sigma_{TTi})} \tag{15}$$

so $VRF_{\text{ind}} = 0 \quad \Leftrightarrow \quad \sum_i \sum_s f_{i_s}\Sigma_{SSi}^{-1}f'_{i_s} = 0 \quad \Leftrightarrow \quad f_{i_s} = 0 \quad \forall(i,s) \Leftrightarrow \quad \Sigma_{TSi} = \mathbf{0}$

(c) $VRF_{\text{ind}} = 1 \Leftrightarrow \forall i$ there is a deterministic relationship between $\widetilde{\varepsilon}_{T_{ij}}$ and $\widetilde{\varepsilon}_{S_{ij}}$.

Proof

$$VRF_{\text{ind}} = 1 \Leftrightarrow \sum_i \text{tr}(\Sigma_{(T|S)i}) = 0 \quad \forall i \tag{16}$$

$$\leftrightarrow V(\widetilde{\varepsilon}_{T_{ij}}|\widetilde{\varepsilon}_{S_{ij}}) - 0 \leftrightarrow (\widetilde{\varepsilon}_{T_{ij}}|\widetilde{\varepsilon}_{S_{ij}}) - \mu_{(T|S)i} \leftrightarrow \widetilde{\varepsilon}_{T_{ij}} - \Sigma_{TSi}\Sigma_{SSi}^{-1}\widetilde{\varepsilon}_{S_{ij}} \quad \forall i \tag{17}$$

(d) In the single endpoint case $VRF_{\text{ind}} = R^2_{\text{ind}}$.

Proof

In the single endpoint case we have:

$$\begin{array}{l} \widetilde{\varepsilon}_{T_{ij}} = \varepsilon_{T_{ij}} \\ \widetilde{\varepsilon}_{S_{ij}} = \varepsilon_{S_{ij}} \end{array} \Rightarrow \begin{pmatrix} \varepsilon_{T_{ij}} \\ \varepsilon_{S_{ij}} \end{pmatrix} \sim N(0, \Sigma_i) \tag{18}$$

where

$$\Sigma_i = \begin{pmatrix} \sigma_{TTi} & \sigma_{TSi} \\ \sigma_{TSi} & \sigma_{SSi} \end{pmatrix} \tag{19}$$

There we also assumed

$$\Sigma_i = \Sigma = \begin{pmatrix} \sigma_{TT} & \sigma_{TS} \\ \sigma_{TS} & \sigma_{SS} \end{pmatrix} \tag{20}$$

$$\begin{array}{l} \varepsilon_{T_{ij}} \sim N(0, \sigma_{TT}) \\ \varepsilon_{S_{ij}} \sim N(0, \sigma_{SS}) \end{array} \Rightarrow (\varepsilon_{T_{ij}}|\varepsilon_{S_{ij}}) \sim N(\mu_{(T|S)}, \Sigma_{(T|S)}) \tag{21}$$

where

$$\begin{array}{rcl} \mu_{(T|S)} & = & \sigma_{TS}\sigma_{SS}^{-1}\varepsilon_{S_{ij}} \\ \sigma_{(T|S)i} & = & \sigma_{TT} - \sigma_{TS}^2\sigma_{SS}^{-1} \end{array} \tag{22}$$

$$VRF_{\text{ind}} = 1 - \frac{\sum_i \text{tr}(\Sigma_{(T|S)i})}{\sum_i \text{tr}(\Sigma_{TTi})} = 1 - \frac{\text{tr}(\Sigma_{(T|S)})}{\text{tr}(\Sigma_{TT})} = 1 - \frac{\sigma_{TT} - \sigma_{TS}^2\sigma_{SS}^{-1}}{\sigma_{TT}} = 1 - (1 - R^2_{\text{ind}}) = R^2_{\text{ind}} \tag{23}$$

completing the proof

Acknowledgements

The first author gratefully acknowledges support from an LUC Bijzonder Onderzoeksfonds grant. The second author was supported by the Institute for the Promotion of Innovation by Science and Technology (IWT) in Flanders, Belgium. The authors are also grateful to the Johnson Pharmaceutical Research and Development for kind permission to use their data.

References

Albert, A., Ioannidis, J.P.A., Reichelderfer, P., Conway, B., Coombs, R.W., Crane, L., Demasi, R., Dixon, D.O., Flandre, P., Hughes, M.D., Kalish, L.A., Larntz, K., Lin, D., Marschner, I.C., Munoz, A., Murray, J., Neaton, J., Pettinelli, C., Rida, W., Taylor, J.M.G., and Welles, S.L. (1998). Statistical issues for HIV surrogate endpoints: point/counterpoint. *Statistics in Medicine* **17**, 2435–2462.

Alonso, A., Geys, H., Molenberghs, G., Vangeneugden, T. (2002).Investigating the Criterion Validity of Psychiatric Symptom Scales using Surrogate Marker Validation Methodology. *Journal of Biopharmaceutical Statistics* **12**, 161–179.

Burzykowski, T., Molenberghs, G., Buyse, M., Geys, H. and Renard, D. (2001).Validation of surrogate endpoints in multiple randomized clinical trials with failure-time endpoints. *Applied Statistics* **50**, 405–422.

Buyse, M., and Molenberghs, G. (1998).Criteria for 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-analyses of Randomized Experiments. *Biostatistics* **1**, 49-67.

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.

Daniels, M.J. and Hughes, M.D. (1998).Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine* **15**, 1515–1527.

Dunn, N., and Mann, R.D. (1999).Prescription-event and other forms of epidemiological monitoring of side-effects in the UK. *Clinical and Experimental Allergy* **29**, 217–239.

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

Flandre, P. and Saidi, Y. (1998).Letters to the Editor: Estimating the Proportion of Treatment Effect Explained by a Surrogate Marker.*Statistics in Medicine* **18**, 107–115.

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

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

Galecki. (1994).General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics: theory and methods* **23**, 3105–3119.

Heitjan, D.F. and Sharma, D. (1997). Modelling repeated-series longitudinal data. *Statistics in Medice*, **16**, 347–365.

Henderson, R., Diggle, P. and Dobson, A. (2000). Joint Modelling of longitudinal measurements and event time data. *Biostatistics* **1**, 465–480.

Jones, T.C. (2001).Call for a new approach to the process of clinical trials and drug registration. *British Medical Journal* **322**, 920–923.

Jorgensen, B., Lundbye-Christensen, S., Song, P. and Sun, L. (1996).State-space models for multivariate longitudinal data of mixed types. *The Canadian Journal of Statistics* **24**, 385–402.

Jorgensen, B., Lundbye-Christensen, S., Song, P. and Sun, L. (1999).A state space model for multivariate longitudinal count data. *Biometrika* **86**, 169–181.

Kay, S.R., Opler, L.A. Lindenmayer, J.P. (1988). Reliability and validity of the Positive and Negative Syndrome Scale for Schizophrenics. *Psychiat. Res* **23**, 99–110.

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

Matsuyama, Y. and Ohashi, Y. (1997). Mixed models for bivariate response repeated measures data using Gibbs sampling. *Statistics in Medice*, **16**, 1587–1601.

Molenberghs, G., Buyse, M., Geys, H., Renard, D. and Burzykowski, T. (2001). *Statistical Challenges in the Evaluation of Surrogate Endpoints in Randomized Trials* (submitted).

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

Schemper, M. and Stare, J. (1996).Explained Variation in Survival Analysis. *Statistics in Medicine* **1996**, 1999-2012.

Renard, D., Geys, H., Molenberghs,G., Burzykowski, T., Buyse, M. (2001).Validation of surrogate endpoints in randomized clinical trials with discrete endpoints. *Statistics in Medicine* (submitted).

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* (submitted).

Rochon, J.(1996). Analyzing bivariate repeated measures for discrete and continuous outcome variables. *Biometrics* **52**, 740–750.

Sy, J., Taylor, J. and Cumberland, W. (1997).A stochastic model for the analysis of bivariate longitudinal AIDS data. *Biometrics,* **53**, 542–555.

[*Received*]