

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

Ariel Alonso, Helena Geys, Geert Molenberghs

Center for Statistics, Limburgs Universitair Centrum, transnationale Universiteit Limburg  
Universitaire Campus  
B3590 Diepenbeek, Belgium.

Michael G. Kenward

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

## Abstract

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 modelling 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 Keywords:* Bivariate longitudinal data, Randomized Clinical Trials, Surrogate Marker, Validation, Canonical Correlations.

# 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 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.

However some failed attempts in the past, when using surrogates instead of true endpoints, make clear 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 (2000) 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  $R_{\text{trial}}^2$ . At the individual level, the squared correlation  $R_{\text{ind}}^2$  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  $R_{\text{trial}}^2$  and  $R_{\text{ind}}^2$  are sufficiently high. However most of the previous work focuses on univariate responses. Going from a univariate setting to a multivariate framework represents new challenges. The  $R^2$  measures proposed by Buyse *et al.*, are no longer applicable. These authors proposed their methodology based on the simplest cross-sectional case in which both the surrogate and the true endpoint, are continuous and normally distributed. Subsequently, different variations to the theme were implemented for binary responses, times to event, mixtures of binary and continuous endpoints, etc. In all of these cross-sectional cases, one assumed that only one potential surrogate was available and that treatment effect on both responses was constant over time and could be characterized by a single parameter. The previous assumptions can fail when a patient is measured repeatedly over time. Extending the methodology to this setting opens some new conceptual problems.

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.* (1999).

The paper is organized as follows: Section 2 introduces a joint model for bivariate longitudinal data. Section 3 defines the variance reduction factor to evaluate surrogacy when repeated measurements for surrogate and true endpoints are available. 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 cross-sectional 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 matrix 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 variance-covariance matrix 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  $i = 1, \dots, N$  trials in the  $i$ th of which  $j = 1, \dots, n_i$  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 at the first stage for both responses can then be written as

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

where  $\mu_{S_i}$  and  $\mu_{T_i}$  are trial-specific intercepts,  $\alpha_i, \beta_i$  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. Note that, even though in practice  $T_{ij}$  and  $S_{ij}$  are frequently measured at the same time points, model (1) would let us approach situations in which this condition does not hold. 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

$$\Sigma_i = \begin{pmatrix} \sigma_{TT_i} & \sigma_{TS_i} \\ \sigma_{TS_i} & \sigma_{SS_i} \end{pmatrix} \otimes R_i. \quad (2)$$

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)

$$R_i = \begin{pmatrix} 1 & \rho_i & \dots & \rho_i^{p_i} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_i^{p_i} & \rho_i^{p_i-1} & \dots & 1 \end{pmatrix}$$

where  $p_i$  denotes the number of designed time points at trial  $i$ .

As we will argue in what follows, the above model is, of course, not free from assumptions. It is therefore important to check the model assumptions in each specific example. However, the measures of surrogacy we will propose, also hold covariance models, more general than the one defined in (2).

Due to replication at the trial level, we can impose a distribution on the trial-specific para-

meters. At the second stage, we therefore assume

$$\begin{pmatrix} \mu_{S_i} \\ \mu_{T_i} \\ \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} m_{S_i} \\ m_{T_i} \\ a_i \\ b_i \end{pmatrix}, \quad (3)$$

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 treatment effect 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  $i = 1, \dots, N$  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

$$R_{\text{trial}}^2 = \frac{\begin{pmatrix} d_{sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{ss} & d_{sa} \\ d_{sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{sb} \\ d_{ab} \end{pmatrix}}{d_{bb}}. \quad (4)$$

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 endpoint after adjustment for the effect of treatment and trial is captured by

$$R_{\text{ind}}^2 = \frac{\sigma_{ST}^2}{\sigma_{SS}\sigma_{TT}}. \quad (5)$$

Although the inclusion of fixed trial-specific treatment coefficients in our model enables us to estimate  $R_{\text{trial}}^2$  at the trial level, at the individual level the  $R_{\text{ind}}^2$  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

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

$$\Sigma_i = \begin{pmatrix} \Sigma_{TT_i} & \Sigma_{TS_i} \\ \Sigma_{TS_i}^T & \Sigma_{SS_i} \end{pmatrix}$$

where  $\Sigma_{TT_i}$  and  $\Sigma_{SS_i}$  are the variance-covariance matrices associated with the residual vectors  $\tilde{\varepsilon}_{T_{ij}}$  and  $\tilde{\varepsilon}_{S_{ij}}$  respectively and  $\Sigma_{TS_i}$  contains the covariances between the elements of  $\tilde{\varepsilon}_{T_{ij}}$  and the elements of  $\tilde{\varepsilon}_{S_{ij}}$ . 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. Note that, under model (2),  $\Sigma_{TT_i} = \sigma_{TT_i}R_i$ ,  $\Sigma_{SS_i} = \sigma_{SS_i}R_i$ , and  $\Sigma_{TS_i} = \sigma_{TS_i}R_i$ .

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

$$VRF_{\text{ind}} = \frac{\sum_i \{\text{tr}(\Sigma_{TTi}) - \text{tr}(\Sigma_{(T|S)i})\}}{\sum_i \text{tr}(\Sigma_{TTi})}, \quad (6)$$

where  $\Sigma_{(T|S)i}$  denotes the conditional variance-covariance matrix of  $\tilde{\varepsilon}_{Tij}$  given  $\tilde{\varepsilon}_{Sij}$ :  $\Sigma_{(T|S)i} = \Sigma_{TTi} - \Sigma_{TSi}\Sigma_{SSi}^{-1}\Sigma_{TSi}'$ . Intuitively, expression (6) 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 (6) 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

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

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  $i$ th observation (Schemper and Stare 1996).

Further one can show (i) that the  $VRF_{\text{ind}}$  ranges between zero and one, (ii) that the  $VRF_{\text{ind}}$  equals zero if and only if the error terms of the true and surrogate endpoints are independent within each trial, (iii) that the  $VRF_{\text{ind}}$  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  $VRF_{\text{ind}}$  reduces to the  $R_{\text{ind}}^2$  when the endpoints are measured only once.

If model (1) is considered then the  $VRF_{\text{ind}}$  can be rewritten in terms of the squared correlations  $\left(\rho_{TSi}^2 = \frac{\sigma_{TSi}}{\sigma_{TTi}\sigma_{SSi}}\right)$  between surrogate and true endpoints at each time point at the



different trials  $i = 1, \dots, N$

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

The latter expression yields an appealing interpretation of the VRF. Indeed, the VRF is just a sum of different trial contributions, where each contribution is 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  $t = 1, \dots, p_i$  canonical correlations  $\rho_t^{i^2}$  for  $(\tilde{\varepsilon}_{\mathbf{T}_{ij}}, \tilde{\varepsilon}_{\mathbf{S}_{ij}})$  such that  $\rho_1^{i^2} \geq \rho_2^{i^2} \geq \dots \geq \rho_{p_i}^{i^2}$  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  $i$ th trial could not be characterized by the scalars  $\beta_i$  and  $\alpha_i$  but by the  $p_i$  dimensional vectors  $\tilde{\beta}_i$  and  $\tilde{\alpha}_i$ , Verbyla (1999).

For reasons explained earlier it would then be unrealistic to assume that the variance-covariance matrix  $D$  is constant over the trials. We can then define the Variance Reduction

Factor at the trial level ( $VRF_{trial}$ ). Suppose that  $(\tilde{\beta}_i, \tilde{\alpha}_i) \sim N((\bar{\beta}_i, \bar{\alpha}_i), D_i)$ , 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  $i$ th trial. Then we can define, similarly to the individual level and with straightforward notations,  $VRF_{trial}$  as

$$VRF_{trial} = \frac{\sum_i \{\text{tr}(D_{\beta\beta i}) - \text{tr}(D_{(\beta|\alpha)i})\}}{\sum_i \text{tr}(D_{\beta\beta i})} \quad (7)$$

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  $VRF_{trial} = R_{trial}^2$ .

The scope of this methodology is not limited to the longitudinal framework, there are other settings in which the use of these tools can be appealing. Most of the work published in this area assumes 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 potential surrogates are now available. At the first stage the following multivariate regression model is assumed

$$\begin{cases} T_{ij} = \mu_{T_i} + \beta_i Z_{ij} + \varepsilon_{T_{ij}} \\ S_{1ij} = \mu_{S_{1i}} + \alpha_{1i} Z_{ij} + \varepsilon_{S_{1ij}} \\ S_{2ij} = \mu_{S_{2i}} + \alpha_{2i} Z_{ij} + \varepsilon_{S_{2ij}} \end{cases} \quad (8)$$

where:  $(\varepsilon_{T_{ij}}, \varepsilon_{S_{1ij}}, \varepsilon_{S_{2ij}}) \sim N(0, \Sigma)$ . At the second stage we will by way of illustration assume that  $(\beta_i, \alpha_{1i}, \alpha_{2i}) \sim N((\beta, \alpha_1, \alpha_2), D)$  where

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

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

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

whereas if both of them are considered jointly with the VRF concept to evaluate surrogacy then we obtain

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

This leads us to a very interesting point about the new concept. Reasoning at the population level, note that  $\text{Var}(\beta_i | \alpha_{1i}, \alpha_{2i}) = \theta$  and hence it is clear that for small values of  $\theta$  there is an almost deterministic relationship between  $\beta_i$  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 effects on both surrogates  $S_1$  and  $S_2$  are known. However, these surrogates would poorly predict the treatment effect on the true endpoint if they were considered independently as can be concluded from the expressions

$$\lim_{\theta \rightarrow 0} R_{1\text{trial}}^2(\theta) = \lim_{\theta \rightarrow 0} R_{2\text{trial}}^2(\theta) = 0.5.$$

On the other hand, the  $VRF_{\text{trial}}$  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:

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

Of course, while in practice effects like this would be milder, because we would have to take measurement error due to finite sampling into account (Gail *et al* 2000), the previous

example does illustrate that a lot can be gained if more than a single surrogate is used. In principle, any number of potential surrogates could be studied and even several endpoints and several surrogates could be analyzed in a multivariate framework.

## 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, in line with findings reported in Buyse *et al* (2000), where it is shown that a sufficient amount of replication at all levels is necessary to identify all of the variance components, preferably with a decent amount of precision. 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. The choice of the unit is an important issue and it is

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

not free of controversy. It can depend on practical considerations, such as the information available in the data set at hand and also on experts' 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 the data and so on. Ideally, both the number of units and the number of patients per unit should be sufficiently large to avoid numerical problems (Buyse *et al* 2000). For the specific context of schizophrenia, Molenberghs *et al* (2002) reported a particular instance where choice of units (investigator versus main investigator) has a mild impact only. These authors also compare results from two different trials. Of course, this is only evidence from a particular, though important, example. Cortiñas *et al* (2003) study a three-level hierarchy (e.g., country, trial, and patient) and the impact on the assessment of surrogacy when either all three levels are used for analysis or when one of the levels is ignored instead. For a number of situations, these authors give explicit formulas for the remaining variance components and hence  $R^2$  measures, in case a level is ignored, as well as guidelines regarding the estimation strategy to obtain the best possible estimate in such cases.

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. Another 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  $\log(\text{CGI})$  and  $\log(\text{PANSS})$  instead of the original variables to stabilize the variances. Figure 1 shows the individual profiles for  $\log(\text{CGI})$  and  $\log(\text{PANSS})$  by treatment groups. In all the panels a linear time trend seems plausible.

FIGURE (1) – ABOUT HERE

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 models. 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 term  $\text{lin}_{ki}(t)$  denotes the linear effect of  $t_{ij}$  and contributes a single regression parameter (slope) and  $\text{spl}_{ki}(t)$  denotes the corresponding random-spline component.

The AIC criterion was then used to select the best model in each trial and model (1) had the best performance in all them. 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  $\log(\text{CGI})$  variance components ( $\hat{\sigma}_{TTi}$ ), the estimated  $\log(\text{PANSS})$  variance components ( $\hat{\sigma}_{SSi}$ ), the  $\log(\text{CGI})-\log(\text{PANSS})$  correlation as well as  $\rho_i$  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.

FIGURE (2) – ABOUT HERE

If we now want to study the relationship between the  $\log(\text{CGI})$  and the  $\log(\text{PANSS})$ , then it is clear that the  $R_{\text{ind}}^2$  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  $VRF_{\text{ind}}$  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: [0.36; 0.41]). Note that we could modify our confidence interval estimation to include measurement error present in the estimation

of the variance components. While this issue is of importance (Gail *et al* 2000), results by Tibaldi *et al* (2003) have shown that in cases similar to the one considered here, the improvement was minor but the computational burden increased considerably.

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

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

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

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, with values of the VRF smaller than 0.6 in almost all the cases.

FIGURE (3) – ABOUT HERE

At the trial level the results are much more encouraging. Since treatment is assumed not to vary with time, the  $R_{\text{trial}}^2$  as introduced by Buyse *et al.* (2000) can still be calculated. We find a value of  $R_{\text{trial}}^2$  of 0.85. The resulting correlation between treatment effects on  $\log(\text{CGI})$  and  $\log(\text{PANSS})$  equals 92% suggesting that a reliable prediction can be made of the treatment effect on  $\log(\text{CGI})$  having observed the treatment effects on  $\log(\text{PANSS})$ . Graphically this is represented in the second panel of figure 3 which plots the treatment effects on  $\log(\text{CGI})$



by the treatment effects on  $\log(\text{PANSS})$ . The size of each point is proportional to the number of patients within a unit.

A 95% confidence interval for  $R_{\text{trial}}^2$  was obtained using bootstrap. Precisely, patients within trials were resampled, with the models refitted on the bootstrap samples. The  $\alpha_i$  and  $\beta_i$ , estimated from these samples were then used to calculate the  $R_{\text{trial}}^2$  measures, as well as the confidence intervals. The so-obtained confidence limits for  $R_{\text{trial}}^2$  are [0.68; 0.95].

## 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; Lin, Fleming and De Gruttola (1997); Buyse *et al.* (2000); Molenberghs *et al.* (2002).

However most of the previous work focused on univariate responses for the surrogate and true endpoints. Going from an univariate setting to a multivariate framework presents new challenges. In this paper, we proposed a new concept to validate surrogate endpoints within the meta-analytic framework but in more complicated contexts. Even though in many practical situations the analyst has to study surrogacy in a multivariate framework up to now most of the research developed in this area assume that only one potential surrogate is going to be evaluated. The example constructed in Section 3 clearly shows that a lot can be lost if we limit ourselves to the analysis of single surrogates.

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 can be studied.

In some practical situations there is no clear idea about which variable (or variables) could be the best possible surrogate for certain endpoint of interest. The VRF allows us to explore which subset of potential surrogates would be optimal. Another important limitation in the current surrogacy literature is that most of the techniques are designed for two treatments only. However, the use of three or more treatments in clinical trials is common practice in some medical fields. The tools introduced in the present work allow to study surrogacy in this setting as well.

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

On the other hand fitting a joint model to analyze mixtures of discrete and continuous responses in a longitudinal framework is a challenging task. Most research so far have concentrated on simultaneous analyses of binary and continuous responses. Further extensions of our methodology using models for different types of responses are necessary and will be the objective of future work.

## ACKNOWLEDGMENTS

The first author gratefully acknowledges support from an LUC Bijzonder Onderzoeksfonds grant. The second author was supported as postdoctoral research fellow from the “Fonds voor wetenschappelijk onderzoek (FWO)-Vlaanderen”, 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.
- 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.
- CORTIÑAS ABRAHANTES, J., MOLENBERGHS, G., BURZYKOWSKI, T., SHKEDY, Z., ALONSO, A., and RENARD, D. (2003). Choice of units and modelling strategies in multilevel hierarchical models, with application to surrogate endpoint evaluation and complex surveys. *Submitted for publication*.
- DANIELS, M.J. and HUGHES, M.D. (1998). Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine* **15**, 1515–1527.
- 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.

- JORGENSEN, B., LUNDBYE-CHRISTENSEN, S., SONG, P. and SUN, L. (1999). A state space model for multivariate longitudinal count data. *Biometrika* **86**, 169–181.
- 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.
- MOLENBERGHS, G., BUYSE, M., GEYS, H., RENARD, D., BURZYKOWSKI, T., and ALONSO, A. (2002). Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Controlled Clinical Trials*, **23**, 607–625.
- 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.
- SY, J., TAYLOR, J. and CUMBERLAND, W. (1997). A stochastic model for the analysis of bivariate longitudinal AIDS data. *Biometrics*, **53**, 542–555.
- TIBALDI, F., CORTIÑAS ABRAHANTES, J., MOLENBERGHS, G., RENARD, D., BURZYKOWSKI, T., BUYSE, M., PARMAR, M., STIJNEN, T., and WOLFINGER, R. (2003) Simplified hierarchical linear models for the evaluation of surrogate endpoints. *Journal of Statistical Computation and Simulation*, **00**, 000–000.
- VERBYLA, A., CULLIS, B., KENWARD, M., WELHAM, S. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines. *Appl. Statist.*, **48**, 269–311.

# APPENDIX B

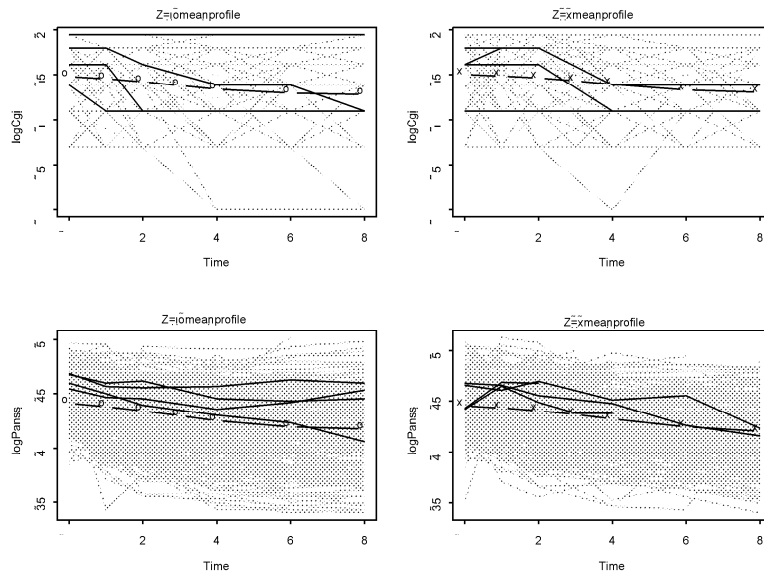


Figure 1:  $\log(\text{CGI})$  and  $\log(\text{PANSS})$ : Mean profiles.

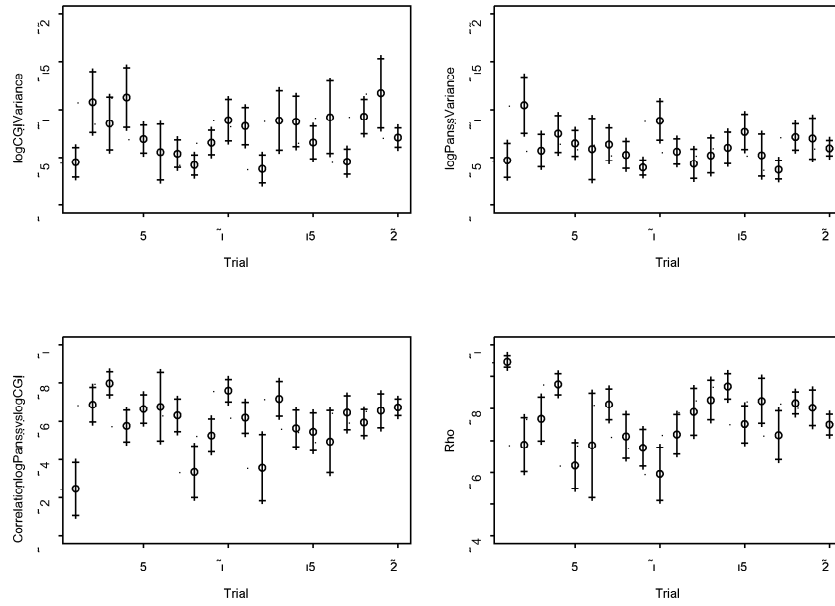


Figure 2: Variance Components.

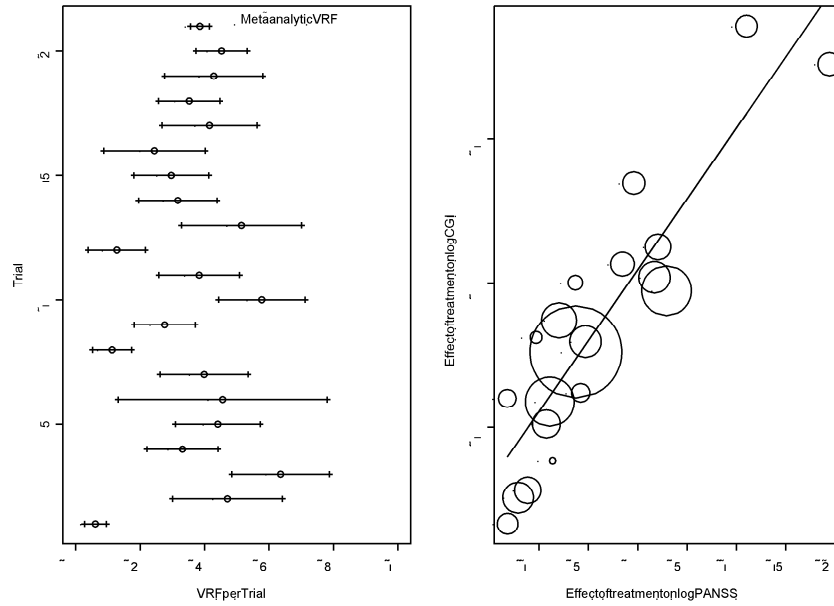


Figure 3: First panel: VRF trials contributions ( $VRF_{ind}^i$ ) and Meta-analytic VRF. Second panel: Treatment effect for BPRS vs treatment effect for PANSS