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

Ariel Alonso[1], Helena Geys[1], Michael G. Kenward[2], Geert Molenberghs[1] and Tony Vangeneugden[3]

[1] Liburgs Universitair Centrum, Center for Statistics, Universitaire Campus B3590, Diepenbeek, Belgium
[2] Inst. Medical Statistics Unit London School of Hygiene and Tropical Medicine Keppel Street London WC1E 7HT United Kingdom
[3] Janssen Research Foundation, Turnhoutseweg 30, B-2340 Beerse,, Belgium

Abstract: While the practice of looking at multiple endpoints is by no means recent in clinical research, the validity of using one endpoint as a surrogate for another one has been raised and studied only over the last decade or so. Past 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 endpoint (Buyse *et al*, 2000). These authors concentrated on continuous responses. 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, a further challenge consists of summarizing the concept of "surrogacy" in simple yet meaningful measures. We propose the use of the so-called variance reduction factor. The methodology is illustrated on data from a meta-analysis of five clinical trials comparing antipsychotic agents for the treatment of chronic schizophrenia.

Keywords: Bivariate longitudinal data, Randomized Clinical Trials, Surrogate Marker, Validation

1 Introduction

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. More recently, Buyse et al. (2000), building on earlier work of Buyse and Molenberghs (1998), 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_{indiv} 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 valid when it is both trial-level valid ($R^2_{trial} \approx 1$) and individual-level valid ($R^2_{indiv} \approx 1$).

Buyse et al. (2000) centered solely on normally distributed surrogate and true endpoints. However, in many practical applications, repeated measurements are encountered on either or both endpoints. Methods that take into account the longitudinal structure of the data yield much more complex statistical modelling strategies and require further extensions in the surrogate marker evaluation methodology. In analogy to the bivariate normal setting considered by Buyse *et al.* (2000) the calculation of these measures should be based on a two-stage approach rather than a full random effects approach, in order to reduce the numerical complexity.

Technically, we need (1) a model for bivariate longitudinal outcomes, and (2) an extension of the R^2 measures towards longitudinal data. In the case of univariate longitudinal endpoints one can consider different types of covariance structures, including compound symmetry, autoregressive, bounded, factor, linear, Toeplitz, spatial, unstructured etc. Now 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 first order autoregressive structure for the repeated measurements on an outcome. While, in the setting of Buyse *et al.* 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 allow for different covariance structures over the different trials.

Hence, suppose that 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:

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

where μ_{S_i} and μ_{T_i} are trial-specific intercepts, α_i and β_i are trial-specific effects of treatment Z_{ij} on the two endpoints and θ_{S_i} and θ_{T_i} are fixed trial-specific time effects in trial $i = 1, \ldots, N$. The vectors $\widetilde{\varepsilon}_{S_{ij}}$ and $\widetilde{\varepsilon}_{T_{ij}}$ are correlated error terms, assumed to be mean-zero bivariate 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 structure 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 should be noticed that if we only have one observation per subject the variable time will disappear from equation (1) and $R_i = \boldsymbol{I}$. If it is also assumed that $\Sigma_i = \Sigma$ then our model is reduced to the model proposed by Buyse *et al.* (2000).

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

$$\begin{pmatrix} \mu_{S_i} \\ \mu_{T_i} \\ \alpha_i \\ \beta_i \\ \theta_{S_i} \\ \theta_{T_i} \end{pmatrix} = \begin{pmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \\ \theta_S \\ \theta_T \end{pmatrix} + \begin{pmatrix} m_{S_i} \\ m_{T_i} \\ a_i \\ b_i \\ \tau_{S_i} \\ \tau_{T_i} \end{pmatrix}, \tag{3}$$

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. 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

$$R^2_{\text{trial}} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}}. \tag{4}$$

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 endpoints after adjustment for the effect of treatment is captured by

$$R^2_{\text{indiv}} = \frac{\sigma^2_{ST}}{\sigma_{SS}\sigma_{TT}}, \tag{5}$$

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

Problems occur however when trying to adopt the above mentioned validation criteria to the specific case of bivariate longitudinal endpoints. In that case the concept of R^2_{indiv} has to be extended because R^2_{indiv} is limited to a single measurement where it represents the squared correlation after correction for the trial effect. Although the inclusion of fixed trial-specific treatment coefficients in our model enables us to estimate R^2_{trial} at the trial level, extensions may be needed 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.

2 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.

We already know that the error terms $\widetilde{\varepsilon}_{T_{ij}}$ and $\widetilde{\varepsilon}_{S_{ij}}$ follow a multivariate normal distribution with variance-covariance matrix :

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

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.

Essentially, we summarize the variability of the repeated measurements on the true endpoint by the trace of its variance-covariance matrix and summing 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 and summing once more over trials. Following these ideas the relative reduction in the true endpoint variance after adjusting by the surrogate can be quantified as:

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

where $\Sigma_{(T|S)_i}$ denotes the conditional variance 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 (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 Z_{ij} 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 (Schemper and Stare 1996).

The VRF_{indiv}, as defined here, is a very natural extension of the R^2_{indiv} validation measure to multivariate longitudinal data. Indeed, 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_{indiv} 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 (3), then we have:

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

Thus, the VRF_{ind} can be rewritten in terms of the correlations (ρ_{TSi})

between surrogate and true endpoints at the different trials $i = 1, \ldots, N$:

$$VRF_{ind} = \frac{\sum_i p_i \sigma_{TTi} \rho^2_{TSi}}{\sum_i \sigma_{TTi} p_i}$$

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 correlation between the surrogate and the true endpoint in that trial with the proportion of the total true endpoint variance that is accounted for by that trial.

As mentioned before, we need an extension of R^2_{trial} as soon as the treatment effect cannot be assumed to be constant over time. For reasons explained earlier it would then be unrealistic to assume that the variance-covariance matrix D would be constant over trial. In that case we can define the Variance Reduction Factor at the trial level, (VRF_{trial}). 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

$$D_i = \begin{pmatrix} D_{\beta\beta i} & D_{\beta\alpha i} \\ D'_{\beta\alpha i} & D_{\alpha\alpha i} \end{pmatrix},$$

then we can define, similarly to the individual level and with straightforward notations, VRF_{trial} as:

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

In case of a single normally distributed endpoint this reduces to R^2_{trial}.

3 Example: a Meta-analysis of Trials in Schizophrenic Subjects

Now 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 only 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.

Even though this is not a standard situation for surrogate validation due to the lack of a "gold" standard, we consider as our primary measure (true endpoint) the Clinician's Global Impression (CGI).

CGI is a 7-grade scale used by the treating physician to characterize how well a subject is doing. As a surrogate measure we consider the Positive and Negative Syndrome Scale (PANSS) (Kay, Opler and Lindenmayer, 1988). The PANSS consists of 30 items that provide an operationalized, drug-sensitive instrument, which is highly useful for both typological and dimensional assessment of schizophrenia. In our model we considered $\log(CGI)$ and $\log(PANSS)$ instead of the original variables to stabilized the variances. Figures 1 and 2 show the mean profiles for $\log(CGI)$ and $\log(PANSS)$ by treatment groups. Clearly, both figures show more or less linear time trends.

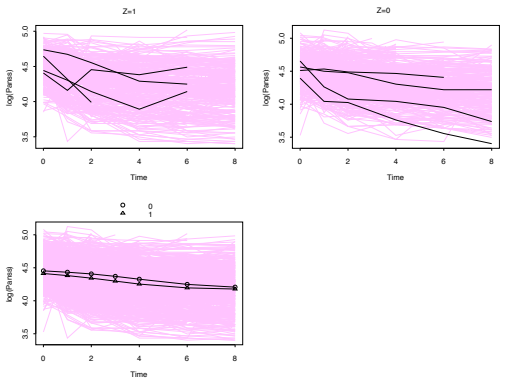

FIGURE 1. $\log(Panss)$: mean profiles

By applying the two-stage approach introduced before with model (1) at the first stage to these data, one can obtain the estimated $\log(CGI)$ 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. All these variance components are plotted in Figure 3, which clearly shows that the assumption of a constant covariance structure over all trials is indeed not really plausible, as already suggested

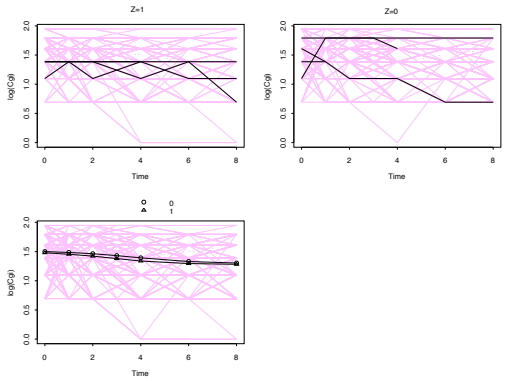

FIGURE 2. log(CGI): mean profiles

before.

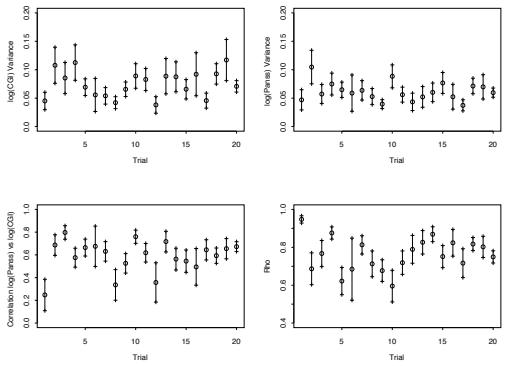

FIGURE 3. Variance Components

If we now want to study the relationship between the log($PANSS$) scale and the log(CGI), then it is clear that the R^2_{ind} measure proposed by Buyse *et al.* is inappropriate for 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 2 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: $[0.38; 0.39]$). This shows that after adjusting by the surrogate log($Panss$) there is a relative reduction in the marginal variance of log(CGI) of 39 per-

cent. 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.

At the trial level the results are much more encouraging. We find a value of R^2_{trial} of 0.83. The resulting correlation between treatment effects on $log(CGI)$ and $log(PANSS)$ equals 91% 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 Figure 2 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} wass

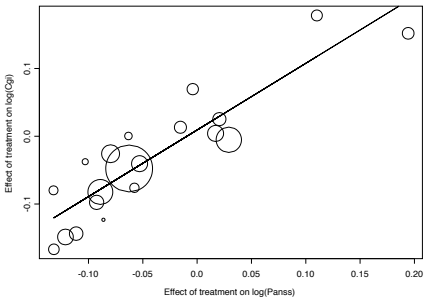

FIGURE 4. Treatment effects on CGI by treatments effects on PANSS. The size of each point is proportional to the number of patients within a unit.

obtained using bootstrap. The so-obtained confidence limits for R^2_{trial} are $[0.67; 0.95]$, which shows that the trial-level association is estimated rather precisely.

References

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.

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

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.

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

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.

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.