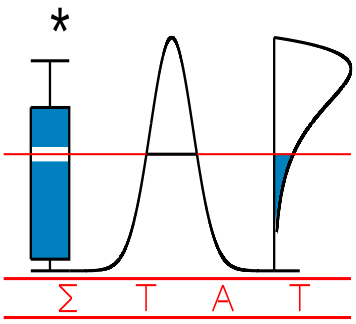
T E C H N I C A L

R E P O R T

0661

FLEXIBLE SURROGATE MARKER EVALUATION FROM SEVERAL RANDOMIZED CLINICAL TRIALS WITH CONTINUOUS ENDPOINTS, USING R and SAS

TILAHUN A., PRYSELEY A., ALONSO A., and G. MOLENBERGHS



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

Flexible Surrogate Marker Evaluation from Several Randomized Clinical Trials with Continuous Endpoints, Using R and SAS

Abel Tilahun **Assam Pryseley** **Ariel Alonso**

Geert Molenberghs

Hasselt University, Center for Statistics, Diepenbeek, Belgium

Abstract

The evaluation of surrogate endpoints is thought to be first studied by Prentice (1989), who presented a definition of a surrogate as well as a set of criteria. Freedman *et al* (2001) supplemented these criteria with the so-called *proportion explained* after notifying some drawbacks in Prentice's approach. Buyse *et al* (2000) framed the evaluation exercise within a meta-analytic setting, thereby overcoming difficulties that necessarily surround evaluation efforts based on a single trial. In this paper, we briefly review the meta-analytic approach for continuous outcomes. Advantages and problems are highlighted by means of two case studies, one in schizophrenia and one in ophthalmology, and a simulation study.

One of the critical issues for the broad adoption of methodology like the one presented here is the availability of flexible implementations in standard statistical software. We have developed generically applicable SAS macros and R functions, at the reader's disposal.

Some Key Words: Adjusted association; Hierarchical model; Meta-analysis; Proportion explained; Random-effects model; Relative effect; Surrogate endpoint.

1 Introduction

Surrogate endpoints come into play in a number of contexts in place of the endpoint of interest, referred commonly to as the true or main endpoint. The use of surrogate endpoints is potentially beneficial, when these endpoints can be measured earlier, leading to a rapid approval of experimental drugs, or can be administered conveniently, which can be equated to less burden on the side of both the experimenter and the patients (Buyse and Molenberghs 1998).

The use of surrogate endpoints in clinical practice is increasing. There are several cases in which there is a need for an accelerated approval of an experimental drug so that its benefit can be witnessed in a shorter time span. This is especially true in the case of chronic diseases with high societal cost.

Ideally, there should be guidelines to declare a marker a useful surrogate for a clinical endpoint. Two possible views are possible when evaluating a marker. The first deals with the individual patient level and is connected with the biological pathway from the surrogate to the true endpoint. This, however, does not necessarily mean a marker is useful to capture the treatment effect in the setting of a clinical trial. Therefore, a second view, focusing on the treatment effect is necessary and possible (Fleming and DeMets 1996). Precisely, this level quantifies the association between the treatment effects on the marker and the clinical endpoint. Buyse *et al* (2000) and Burzykowsky *et al* (2004), among others, have presented a meta-analytic modeling framework, within which both forms of validation can be undertaken. A key stumbling block for the practical use is the availability of flexible implementations within standard and commonly used software packages.

The purpose of this paper is to review the validation framework, to exemplify the methodology in two clinical trial settings, and to present a generic R function and SAS macro. Computational issues, that can be sources of concern, and of which the practitioner should be at current, are discussed in detail and illustrated through a simulation study. These issues center around the choice of unit of analysis, treatment coding schemes and problems with non-positive definite and ill-conditioned matrices.

The rest of this paper is organized as follows. An introduction to the motivating studies is given in Section 2. The different validation methods are outlined in Section 3 with the methods that are based on the meta-analytic approach reviewed first followed by number of simplified computational strategies. Computational issues, arising when using the meta-analytic approach, are discussed in Section 4. Section 5 contains the results of the case studies performed on the datasets introduced in Section 2.

2 Motivating Case Studies

We will present a case study in schizophrenia, followed by one in ophthalmology.

2.1 A Meta-analysis of Five clinical Trials in Schizophernia

The data come 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. The treatment indicator for risperidone versus conventional treatment will be denoted by Z. Schizophrenia has long been recognized as a heterogeneous disorder with patients suffering from both 'negative' and 'positive' symptoms. Negative symptoms are characterized by deficits in cognitive, affective and social functions for example poverty of speech, apathy and emotional withdrawal. Positive symptoms entail more florid symptoms such as delusions, hallucinations and disorganized thinking, which are superimposed on mental status (Kay, Fiszbein, and Opler 1987). Several measures can be considered to asses a patient's global condition. Clinician's Global impression (CGI) is generally accepted as an admittedly subjective clinical measure of change. Here, the change of CGI from baseline will be considered as the true endpoint, denoted by T. It is scored on a 7-grade scale used by the treating physician to characterize how well a subject has improved since baseline. Another useful and sufficiently sensitive assessment scales is 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 (Kay, Opler, and Lindenmayer 1988). We will use the change from baseline in PANSS as our surrogate endpoint, denoted by S. The data contains five trials and in all trials, information is available on the investigators that treated the patients. This information is helpful to define group of patients that will become units of analysis.

2.2 Age-related Macular Degeneration Study (ARMD)

This is a clinical trial involving patients with age-related macular degeneration, a condition in which patients progressively lose vision. Overall, 190 patients from 42 centers participated in the trial. Patients' visual acuity was assessed using standardized vision charts displaying lines of five letters of decreasing size, which patients had to read from top to bottom. The visual acuity was measured by the number of letters correctly read. The binary indicator for treatment is set to $Z = -1$ for placebo and $Z = 1$ for interferon-α. The surrogate endpoint S is the change in visual acuity 6 months after starting treatment while the true endpoint T is the change in visual acuity

at 1 year. In the analysis, the centers in which patients were treated will be considered as units of analysis. Six out of 42 centers participating in the trial enrolled patients only to one of the two treatment arms. These centers were excluded from considerations. A total of 36 centers were thus available for analysis, with a number of individual patients per center ranging from 2 to 18 (183 patients overall).

3 Validation Methods

First, we review the meta-analytic framework and since the fitting of models within this framework can be demanding, a number of simplified strategies, as presented by Tibaldi *et al* (2003) will be presented in second instance.

3.1 Review of the Meta-Analytic Approach

Although the single trial based methods are relatively easy in terms of implementation, they are surrounded with difficulty as there evidently is replication at the patient level, but not at the level of the treatment effect. Therefore, several authors, such as Daniels and Hughes (1997), Buyse *et al* (2000), and Gail *et al* (2000), have introduced a meta-analytic approach.

The meta-analytic approach has been formulated originally for two continuous, normally distributed outcomes, and extended in the meantime to a large set of outcome types, ranging from continuous, binary, ordinal, time-to-event, and longitudinally measured outcomes. A review can be found in Burzykowski, Molenberghs, and Buyse (2005). Here, for simplicity, we focus on the continuous case, where the surrogate and true endpoints are jointly normally distributed.

The meta-analytic approach is based on a hierarchical two-level model. Both a fixed-effects and a random-effects view can be taken. Let T_{ij} and S_{ij} be the random variables denoting the true and surrogate endpoint for the jth subject in the ith trial, and let Z_{ij} be the indicator variable for treatment. First, consider the following fixed-effects models:

$$S_{ij} = \mu_{Si} + \alpha_i Z_{ij} + \varepsilon_{Sij}, \quad (1)$$

$$T_{ij} = \mu_{Ti} + \beta_i Z_{ij} + \varepsilon_{Tij}, \quad (2)$$

where μ_{Si} and μ_{Ti} are trial-specific intercepts, α_i and β_i are trial-specific effects of treatment Z_{ij} on the endpoints in trial i, ε_{Si} and ε_{Ti} are correlated error terms, assumed to be zero-mean normally distributed with covariance matrix

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

A classical hierarchical, random-effects modeling strategy can also be adopted in the following manner:

$$S_{ij} = \mu_S + m_{Si} + \alpha Z_{ij} + a_i Z_{ij} + \varepsilon_{Sij}, \tag{4}$$

$$T_{ij} = \mu_T + m_{Ti} + \beta Z_{ij} + b_i Z_{ij} + \varepsilon_{Tij}. \tag{5}$$

Here, μ_S and μ_T are fixed intercepts, α and β are fixed treatment effects, m_{Si} and m_{Ti} are random intercepts, and a_i and b_i are random treatment effects in trial i for the surrogate and true endpoints, respectively. The random effects (m_{Si}, m_{Ti} ,a_i , b_i) are assumed to be mean-zero normally distributed with covariance matrix

$$D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ & d_{TT} & d_{Ta} & d_{Tb} \\ & & d_{aa} & d_{ab} \\ & & & d_{bb} \end{pmatrix}. \tag{6}$$

The error terms ε_{Sij} and ε_{Tij} follow the same assumptions as in the fixed effects models. In addition, following the fixed-effect models (1) and (2), we can specificy

$$\begin{pmatrix} \mu_{Si} \\ \mu_{Ti} \\ \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} m_{Si} \\ m_{Ti} \\ a_i \\ b_i \end{pmatrix}, \tag{7}$$

where the second term on the right hand side of (7) is assumed to follow a zero-mean normal distribution with covariance matrix (6).

Upon fitting the above models, the surrogate marker evaluation is captured by means of two quantities, the trial-level and individual-level R^2, respectively. The former quantifies the association between the treatment effects on the true and surrogate endpoints at the trial level. The latter

measures the association at the level of the individual patient and after adjustment for the treatment effect. The former is given by:

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

The above quantity is unitless and, at the condition that the corresponding variance-covariance matrix is positive definite, lies within the unit interval.

The models (1) and (2) can be referred to as the full fixed effects models and it is possible to simplify them. The reduced versions of these models are obtained by replacing the fixed trial-specific intercepts, one for each endpoint, common to all trials. The reduced mixed effect models result from removing the random trial-specific intercepts m_{Si} and m_{Ti} from models (4) and (5). The R^2 for the reduced models is then calculated as follows:

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

A surrogate could thus be adopted when R^2_{trial} is sufficiently large. Arguably, rather than using a fixed cutoff above which a surrogate would be adopted, there always will be clinical and other judgment involved in the decision process.

The R^2_{indiv} is based on (3) and takes the following form:

$$R^2_{\text{indiv}} = R^2_{\varepsilon_{Ti}|\varepsilon_{Si}} = \frac{\sigma^2_{ST}}{\sigma_{SS}\sigma_{TT}} \tag{8}$$

Note that, here, trial is considered as experimental unit which can be replaced by center, investigator or any other suitable experimental unit, depending on the nature of the study conducted. The issue of the unit of analysis is discussed in Section 4.1 and has been thoroughly studied by Cortiñas *et al* (2004). Thus, while trial is the optimal unit of analysis from a substantive point of view, practical considerations may render such a choice less than optimal. The need to turn to alternative units may be alleviated by ensuring good access to the widest possible class of trials.

3.2 Review of Simplified Modeling Strategies

Though the above hierarchical modeling is elegant, it often poses a considerable computational challenge (Burzykowski, Molenberghs, and Buyse 2005). To address this problem, Tibaldi *et al*

(2003) suggested several simplifications of the above strategy, briefly outlined here. These authors considered three possible dimensions along which simplifications can be undertaken.

3.2.1 Trial Dimension

This dimension provides a choice between treating the trial-specific effects as fixed or random. If the trial-specific effects are chosen to be fixed, a two-stage approach is adopted. The first-stage model will take the form (1)–(2) and at the second stage, the estimated treatment effect on the true endpoint is regressed on the treatment effect on the surrogate and the intercept associated with the surrogate endpoint as

$$\widehat{\beta}_i = \widehat{\lambda}_0 + \widehat{\lambda}_1 \widehat{\mu}_{Si} + \widehat{\lambda}_2 \widehat{\alpha}_i + \varepsilon_i. \tag{9}$$

The trial-level $R^2_{\text{trial(f)}}$ then is the coefficient of determination obtained by regressing $\widehat{\beta}_i$ on $\widehat{\mu}_{Si}$ and $\widehat{\alpha}_i$, whereas $R^2_{\text{trial(r)}}$ is obtained from the coefficient of determination resulting from regressing $\widehat{\beta}_i$ on $\widehat{\alpha}_i$ only. The individual-level value is calculated as in (8) using the estimates from (3). Note that here (r) and (f) indicate that the trial-level association is obtained based on either the reduced or the full model, respectively.

The second option is to consider the trial-specific effects as random. Depending on the choice made on the endpoint dimension, two directions can be followed. The first one involves a two-stage approach with univariate models (4)–(5) at the first stage. A second stage model consists of a normal regression with the random treatment effect on the true endpoint as response and the random intercept and random treatment effect on the surrogate as covariates. The second direction is based on a full random effects (hierarchical) model as discussed in Section 3.1.

3.2.2 Endpoint Dimension

Though natural to assume the two endpoints to be correlated, this can lead to computational difficulties in fitting the models. The need for the bivariate nature of the outcome is associated with R^2_{indiv}, which is in some cases of secondary importance. In addition, there is also a possibility to estimate it by making use of the correlation between the residuals from two separate univariate models. Thus, further simplification can be achieved by fitting separate models for the true and

surrogate endpoints, the so-called univariate approach.

If in the trial dimension, the trial-specific effects are considered to be fixed, then models (1)–(2) are fitted separately. Similarly, if the trial-specific effects are considered random, then models (4)–(5) are fitted separately, i.e., the corresponding error terms in the two models are assumed to be independent.

3.2.3 Measurement Error Dimension

When the univariate approach from the endpoint dimension and/or the fixed effects approach from the trial dimension are chosen, there is a need to adjust for the heterogeneity in information content between trial-specific contributions. One way to do so is weighting the contributions according to trial size. This gives rise to a weighted linear regression model (9) in the second stage.

4 Computational Considerations

In this section, we will address a number of computational issues and considerations, such as the choice of the unit for analysis, the effect of treatment coding, the possible occurrence of ill-conditioned and non-positive definite variance-covariance matrices, and the (lack of) availability of standard software.

4.1 Unit of Analysis

A cornerstone of the meta-analytic method is the choice of the unit of analysis such as, for example, trial, center, or investigator. This choice may depend on practical considerations, such as the information available in the data set at hand, experts' considerations about the most suitable unit for a specific problem, the amount of replication at a potential unit's level, and the number of patients per unit. From a technical point of view, the most desirable situation is where the number of units and the number of patients per unit is sufficiently large. This issue has been discussed by Cortiñas *et al* (2004).

4.2 Treatment Coding

Most of the work reported on in Burzykowski, Molenberghs, and Buyse (2005) is for a dichotomous treatment indicator. Two choices need to be made at analysis time. First, the treatment variable can be considered continuous or discrete (a class variable). Second, when a continuous route is chosen, it is relevant to reflect on the actual coding, 0/1 and $-1/+1$ being the most commonly encountered ones. For models with treatment occurring as fixed effect only, these choices are essentially irrelevant, since all choices lead to an equivalent model fit, with parameters from one situation to another connected by simple linear transformations. Note that this is not the case, of course, for more than three treatment arms. However, of more importance for us here is the impact the choices can have on the hierarchical model. Indeed, while the marginal model resulting from (4)–(5) is invariant under such choices, this is not true for the hierarchical aspects of the model, such as, for example, the R^2 measures derived at the trial level. Indeed, a $-1/+1$ coding ensures the same components of variability operate in both arms, whereas a 0/1 coding, for a positive definite D matrix, forces the variability in the experimental arm to be greater than or equal to the variability in the standard arm. To see this, for simplicity, assume that the random terms in the model are independent from each other, then the total variance of the response is the sum of the squared regression coefficients attached to the random terms, multiplied by the variance of these random terms. Thus, for those subjects that have taken a placebo, the coefficient for the random treatment effect is zero and hence their total variance will amount to the sum of the other terms. On the other hand, the total variance of subjects in the treatment group will have an additional term in the form of the variance of the random treatment effect, since here the corresponding parameter equals one. Thus, the treatment coding implicitly makes assumptions about the ordering of the total variability in the treated and placebo groups. Both situations may be relevant, and therefore it is of importance to illicit views on this issue from the study's investigators. While this may require some explanation to the study investigator, this proposal is in line with the view that surrogate marker validation cannot be carried out based solely on statistical grounds and there should be substantive input as well.

4.3 Ill-conditioned and Non-positive Definite Variance-covariance Matrix

When the full bivariate random effect is used, the R^2_{trial} is computed from the variance-covariance matrix (6). It is sometimes possible that this matrix be ill-conditioned and/or non-positive definite. In such cases, the resulting quantities computed based on this matrix might not be trustworthy. One way to asses the ill-conditioning of a matrix is by reporting its condition number, i.e., the ratio of the largest over the smallest eigenvalue. A large condition number is an indication of ill-condioning. The most pathological situation occurs when at least one eigenvalue is equal to zero. This corresponds to a positive semi-definite matrix, which occurs, for example, when a boundary solution is obtained. Thus, in the validation process, it is necessary to check the D matrix for absence of presence of these issues.

4.4 Simulation Results

To asses the impact of using an incorrect treatment coding, a small simulation involving 12 different combinations of trial size and number of individuals per trial has been performed. The data were generated based on the following model:

$$S_{ij} = 45 + m_{Si} + (3 + a_i)Z_{ij} + \varepsilon_{Sij}, \tag{10}$$

$$T_{ij} = 50 + m_{Ti} + (5 + b_i)Z_{ij} + \varepsilon_{Tij}. \tag{11}$$

Here, a_i and b_i are random treatment effects in trial i for the surrogate and true endpoints, respectively. The random effects $(m_{Si}, m_{Ti}, a_i, b_i)$ are assumed to be mean-zero normally distributed with covariance matrix:

$$D = \begin{pmatrix} 3 & 2.4 & 0 & 0 \\ 2.4 & 3 & 0 & 0 \\ 0 & 0 & 3 & 2.7 \\ 0 & 0 & 2.7 & 3 \end{pmatrix}. \tag{12}$$

The error terms ε_{Sij} and ε_{Tij} are assumed to be zero mean random variables with variance-covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} 3 & 2.4 \\ 2.4 & 3 \end{pmatrix}. \tag{13}$$

The number of trials was fixed at either 10, 20, or 50, with each trial involving either 10, 20, 40, or 60 subjects, jointly giving rise to 12 different scenarios. For each combination, 100 datasets

were generated for both treatment codings. The datasets were then analyzed with the correct treatment coding, i.e., the treatment coding with which the data were generated, as well as with the opposite coding. For each case the median condition number and the percentage of positive definite variance-covariance matrices are counted. The results of these simulations are displayed in Tables 1 and 2.

The simulation has revealed that, for a small number of analysis units and/or a small number of subjects per analysis unit, the wrong treatment coding could result in a high degree of uncertainty in the resulting variance-covariance matrix. For the 0/1 coding, the effect is noticed even when the correct coding was followed to do the analysis, i.e. there was high degree of uncertainty even when the data were analyzed with the correct 0/1 coding for small sample sizes. The effect, however, seems to vanish with increasing repetition of the unit of analysis and number of subjects per unit of analysis. If we consider a median condition number of 100 as an arbitrary cutoff value, we notice that we require a minimum of 20 trials to achieve a condition number less than 100 for 0/1 coding. This number, however, reduces to only 10 trials to reach a condition number less than 100 for $-1/+1$ coding. With respect to the positive-definitness of the variance-covariance matrix, the percentage of positive-definite matrices increases with increase in the sample size for both treatment coding schemes. However, the $-1/+1$ produced relatively a higher percentage of positive definite matrices even for small samples as compared to the 0/1 coding where the percentage of positive definite matrices is low even for moderately higher sample sizes. Based on the results of this simulation, it seems reasonable to consider the $-1/+1$ treatment coding and chose a reasonable unit of analysis to avoid the numerical problems and achieve positive definiteness in the variance-covariance matrix.

4.5 Software

R functions and SAS macros have been developed to implement the methods discussed in the previous sections. In the appendix, we outline how both tools operate. Further, these tools can be downloaded from `http://www.censtat.be/research/software.asp`. Turning the R functions into a full-fledged R library is work in progress, which will be posted on the same web pages in due course.

Table 1: *Simulation results for* $-1/1$ *treatment coding.*

	simulation strategy		**% positive-definite**		**median condition number**	
simulation #	# trials	# subjects	correct	incorrect	correct	incorrect
1	10	10	42	41	3.44E+16	3.71E+17
2	10	20	66	65	178.00	403.10
3	10	40	91	91	78.36	172.86
4	10	60	98	98	81.23	158.39
5	20	10	90	90	52.43	138.62
6	20	20	97	98	43.33	102.34
7	20	40	100	100	34.87	101.55
8	20	60	100	100	32.97	84.41
9	50	10	100	100	27.55	84.56
10	50	20	100	100	26.54	80.64
11	50	40	100	100	24.28	75.01
12	50	60	100	100	24.92	72.86

5 Application to the Case Studies

The two case studies, introduced in Section 2, are analyzed here. Let us start with the schizophrenia study. Here, trial seems the natural unit of analysis. Unfortunately, the number of trials is not sufficient to apply the full meta-analytic approach. The use of trial as unit of analysis for the simplified methods might also entail problems. The second stage involves a regression model based on only five points, which might give overly optimistic or at least unreliable R^2 values. The other possible unit of analysis for this study is 'investigator'. There were 176 investigators who each treated between 2 and 60 patients. The use of investigator as unit of analysis is also surrounded with problems. Although a large number of investigators is convenient to explain the between investigator variability, because there are few patients per investigators for some investigators, the resulting within-unit variability might not be estimated correctly.

The basic meta-analytic approach and the corresponding simplified strategies have been applied to this data set. The results are displayed in Table 3. Both investigator and trial were used as unit of analysis. However, as there were only five trials, it became difficult to base the analysis on

Table 2: *Simulation results for 0/1 treatment coding.*

	simulation strategy		**% positive-definite**		**median condition number**	
simulation #	# trials	# subjects	correct	incorrect	correct	incorrect
1	10	10	10	10	5.44E+16	3.71E+17
2	10	20	25	25	4.09E+16	9.03E+16
3	10	40	57	58	304.05	1184.91
4	10	60	68	68	196.44	436.48
5	20	10	38	38	2.79E+16	6.6E+16
6	20	20	62	62	136.94	560.39
7	20	40	89	89	51.17	186.94
8	20	60	97	97	38.32	166.40
9	50	10	70	71	67.83	225.77
10	50	20	93	93	34.18	158.24
11	50	40	100	100	27.31	134.00
12	50	60	100	100	25.56	127.24

trial as unit of analysis in the case of the full bivariate random-effects approach. The results have shown a remarkable difference in the two cases. Consistently, in all of the different simplifications, the R^2_{trial} values were found to be higher when trial was used as unit of analysis. This is to be expected, since the second-stage model involved a simple linear regression based on only five data points. Furthermore, note that, when investigator is used as unit of analysis, the R^2_{trial} values are higher when the reduced model is used as compared to the the case where the full model used. This indicates that the investigator-specific intercept terms for the surrogate model does convey information and generally the full model is to be preferred. The opposite result obtains when trials are used as unit of analysis. This result can, and is, explained in the same fashion. The bivariate full random effects model does not converge when trial is used as the unit of analysis. This might be due to lack of sufficient information to compute all sources of variability. The reduced bivariate random effects model converged for both cases, but the resulting variance-covariance matrices were not positive-definite and were ill conditioned, as can be seen from the very large value of the condition number.

Consequently, the results of the bivariate random effects model should be treated with caution as

there might be high uncertainty attached to the results obtained based upon these ill-conditioned matrices. If we concentrate on the results based on investigator as unit of analysis, we observe a low level of surrogacy of PANSS for CGI, with R^2_{trial} ranging roughly between 0.5 and 0.68 for the different simplified models. This result, however, has to be coupled with other findings based on expert opinion to fully guarantee the validation of PANSS as possible surrogate for the CGI. Turning to R^2_{indiv}, it ranges between 0.4904 and 0.5230, depending on the method of analysis, which is relatively low. To conclude, based on the investigators as unit of analysis, PANSS does not seem a good surrogate for the CGI.

For the ARMD study, the only available unit of analysis was center. There were 36 centers which treated between 2 and 18 patients. Note that these data has been analyzed by Buyse *et al* (2000) with a treatment coding of 0 and 1 for the placebo and treatment arms, respectively. Here, the $-1/+1$ coding was used and thus slightly different results obtain. The basic meta-analytic approach and the corresponding simplified modeling strategies have also been applied to this dataset and the results are displayed in Table 4 for the $-1/+1$ coding and in Table 5 for the 0/1 coding.

For the ARMD study, the R^2_{trial} ranges roughly between 0.64 and 0.8, except for the full bivariate random effects models where we find $\widehat{R}^2_{\text{trial}} = 0.9999$. However, the corresponding variance-covariance matrices were non-positive definite and have very large condition number, a sign of high uncertainty surrounding the latter estimate. Hence, it cannot be trusted. Based on the findings, it is possible to say that assessment of change in visual acuity at 6 months does not seem to be a very strong surrogate for the same assessment at 1 year.

6 Discussion

In this paper we reviewed the meta-analytic strategy for validating a surrogate endpoint. The choice of unit of analysis and corresponding computational issues that need to be given due attention have also been adressed. The choice of unit of analysis in applying the meta-analytic approach is a very important issue to be considered. There might be a large difference in the findings depending on the unit of analysis chosen. The optimal unit of analysis is the one for which there is a sufficient number of repetition and each unit has sufficiently large number of individuals within it. Ideally,

Table 3: *Schizophrenia study. Results of the trial-level (R^2_{trial}) surrogacy analysis.*

	Fixed effects		**Random effects**	
Unit of analysis	Unweighted	Weighted	Unweighted	Weighted
Full Model				
Univariate approach				
Investigator	0.5887	0.5608	0.5488	0.5447
Trial	0.9641	0.9636	0.9849	0.9909
Bivariate approach				
Investigator	0.5887	0.5608	0.9898∗	
Trial	0.9641	0.9636	—	
Reduced Model				
Univariate approach				
Investigator	0.6707	0.5927	0.5392	0.5354
Trial	0.8910	0.8519	0.7778	0.8487
Bivariate approach				
Investigator	0.6707	0.5927	0.9999∗	
Trial	0.7418	0.8367	0.9999∗	

∗: *The variance-covariance matrix is ill-conditioned; in particular, at least one eigenvalue is very close to zero.The condition numbers for the three models with ill-condition matrices, from top to bottom are 3.415E+18, 2.384E+18 and 1.563E+18 respectively.*

the choice of unit of analysis should be based on both statistical and subject-matter considerations. The treatment coding also needs to be given serious consideration, in consultation with experts who may be able to formulate an opinion on the possible variability of the two treatment arms. A small simulation study and analysis of two real sets of data supported these points.

Absence of standard software has been one of the limiting factors hampering the use of the meta-analytic approach. We have developed an R library and a SAS macro which can be used to conduct these analyses for continuous outcomes. Efforts are under way to incorporate more R functions and SAS macros for different types and combinations of endpoints.

Table 4: *ARMD data. Results of the trial-level (R^2_{trial}) surrogacy analysis $-1/+1$ coding.*

	Fixed effects		**Random effects**	
Unit of analysis	Unweighted	Weighted	Unweighted	Weighted
		Full Model		
		Univariate approach		
Center	0.6922	0.6963	0.6605	0.7959
		Bivariate approach		
Center	0.6922	0.6963	0.9999∗	
		Reduced Model		
		Univariate approach		
Center	0.6409	0.6562	0.6772	0.7929
		Bivariate approach		
Center	0.6409	0.6562	0.9999∗	

∗: *The variance-covariance matrix is ill-conditioned; in particular, at least one eigenvalue is very close to zero.The condition numbers for Full and Reduced Bivariate random effects models are 1.109E+17 and 1.965E+18 respectively*

A Implementations

In this appendix, we will briefly outline the use of our SAS macro and R function, respectively.

A.1 SAS Macro

The SURCONCON macro can be used to perform the above analysis involving surrogate marker validation using the meta-analytic approach. The macrocan be invoked as follows:

```
%surconcon(yvar=,endpoint=,trial=,id=,data=,trt=,adj=,
           red=,boot=,type=,bootnum=,dmat=,outf=,plot=,drive=,file=,solutionf=)
```

where

yvar: Name of the response variable.

endpoint: Name of the endpoint indicator (-1=surrogate endpoint, 1=true endpoint).

Table 5: *ARMD data. Results of the trial-level (R^2_{trial}) surrogacy analysis* 0/1 *coding.*

	Fixed effects		**Random effects**	
Unit of analysis	Unweighted	Weighted	Unweighted	Weighted
		Full Model		
		Univariate approach		
Center	0.692	0.693	0.664	0.801
		Bivariate approach		
Center	0.692	0.693	—	
		Reduced Model		
		Univariate approach		
Center	0.776	0.758	0.659	0.786
		Bivariate approach		
Center	0.776	0.758	—	

trial: Name of the unit of analysis (center, trial,. . .)

id: Name of the variable indicating the unique subject identification number.

data: Name of input dataset. See the macro description on data formatting and layout.

trt: Name of the treatment indicator variable.

adj: A choice for using weighted (adj=1)or unweighted(adj=0) regression in the second stage.

red: A choice for using reduced (red=1) or full (red=0) model.

type: A choice for using the different modeling approaches (1–4).

1. Univariate fixed effect
2. Bivariate fixed effect
3. Univariate random effect
4. Bivariate random effect

boot: A choice for using different bootstrapping approaches(0-4).This option is required only when "type" is set to 1 or 3.

0. Simple percentile confidence interval

1. Improved normal confidence interval
2. The studentized confidence interval
3. The Basic confidence interval
4. Percentile confidence interval

bootnum: Number of bootstrap samples required.

dmat: A choice for printing the matrix of the random terms to the output file (1=yes,0=no).

solutionf: A choice for printing the solution for fixed effect to the output file (1=yes,0=no).

outf: A choice for printing the trial specific random(for type=3 or 4) or fixed (for type=1 or 2) effects to the output file (1=yes,0=no).

plot: A choice for printing the plot of the raw outcomes, residuals and treatment effects of the main endpoint against those of the surrogate endpoint (plot=1).

drive: The drive on your computer where you want to save the output file (like A, C, D).

file: The name of the output file containing the macro results.

rescale: The option to control the size of the bubble plots when the treatment effects on the true endpoint are plotted against those of the surrogate endpoints. It can take integer values or fractions depending on the size of the plots. It is important that the endpoint indicator be coded as $-1/+1$, with 1 indicating the true endpoint.

Example: Consider part of the data arranged as the format in Table 6. Once the data are arranged in this format and saved as a SAS dataset, the macro can be invoked as

```
%surconcon(yvar=outcome,endpoint=endpoint,trial=trial,id=subject,
data=data,trt=treatment,adj=1,red=0,boot=2,type=1,bootnum=1000,
plot=1,drive=c, file=output1,solutionf=1)
```

This call produces the following results:

Table 6: *Data layout for surrogacy analysis.*

subject	*trial*	*outcome*	*endpoint*	treatment
1	1	0	1	1
1	1	-10	-1	1
2	1	-3	1	-1
2	1	1	-1	-1
3	2	-6	1	1
3	2	-17	-1	1
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

- A full (red=0) univariate fixed effect (type=1) model with a weighted regression (adj=1) in the second stage.
- The standard error of the individual-level R-square is computed based on 1000 bootstrap samples(bootnum=1000).
- The confidence interval for the individual-level R-square will be computed using "the studentized confidence interval" (boot=2).
- The out put will be saved on the c-drive (drive=c) with a pdf file output1.pdf (file=output1).
- The output contains the trial and individual level R-square (default), the plot of the outcome, treatment effect and residual of the true endpoint against those of the surrogate endpoint (plot=1).
- The solution for the fixed effects will be printed (solutionf=1).

A.2 R functions

`bivs` is an R function that performs surrogate endpoint evaluation through the full hierarchical models; both the two-stage (fixed) and full random effect models can be considered. `unis` is an R function that performs surrogate endpoint evaluation through the univariate approach, by fitting univariate fixed and mixed effects models based on the simplified strategies of Tibaldi *et al* (2003).

The calls are

```
bivs(outcomes, endpoints, treatment, trialunit, subject,
     mixed=FALSE, method=1, reduced=FALSE, weighted=FALSE)
```

```
unis(outcomes, endpoints, treatment, trial.unit, mixed=TRUE,
     reduced=FALSE, weighted=FALSE, alpha=0.05, Type=2, sample.size=1000)
```

The arguments are as follows:

alpha: Significant level used to construct bootstrap confidence intervals for individual level R-square. The default value is 0.05.

endpoints: Endpoint indicator, which should be coded as: Surrogate endpoint= -1 and True endpoint= 1.

mixed: If TRUE univariate linear mixed effects models are used for both the endpoints. If FALSE univarite fixed effect models are used. Default is TRUE.

outcomes: The response variable holding the values for both the True endpoint and the Surrogate endpoint, for each case/subject in the data. It should be preferably be sorted by endpoints (see below).

reduced: Logical value indicating whether a reduced(=TRUE) or a full (=FALSE) model is fitted. Default is FALSE.

sample.size: Integer value specifying the number of bootstrap samples.

subjects: Variable holding the identification of the various subjects/patients in the data.

treatment: Treatment indicator, should be coded as: Standard treatment is coded as -1 and new treatment as 1.

trial.unit: The unit of analysis, e.g., trial, center, investigator,...

type: Indicates the type of bootstrap confidence interval to estimate for R^2_{indiv}:

1. Normal confidence interval
2. Improved normal confidence interval
3. Basic bootstrap confidence interval
4. Percentile confidence interval
5. Studentized interval

weighted: If TRUE weighted normal regression is performed for the estimation of trial level R-square. If FALSE, normal regression is performed. Default is FALSE.

Details: It is necessary that treatment and endpoint be coded as 1 and -1. A consequence of the 0/1 coding is that the group coded as 0 is assumed to have smaller variance than the group coded as 1. On the other hand the $-1/+1$ coding leads to equal variance in both groups.

Value: An object of class bivs or unis representing the surrogate endpoint evaluation through the hierarchical approach or univariate approach, respectively. Generic functions such as print, plot and summary have methods to show the results of the fit. The functions residuals and coef can be used to extract some of the fitted objects components.

The implementation example performed in the description of the SAS macro section can also be done using the *unis* function in R as follows. Arrange the data as given in Table 6 and then run the function as:

```
unis(outcome, endpoint, treatment, trial, mixed=FALSE,
     reduced=FALSE, weighted=TRUE, alpha=0.05, Type=5, sample.size=1000)
```

Acknowledgement

We gratefully acknowledge support from Belgian IUAP/PAI network "Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data".

References

Burzykowski, T., Molenberghs, G., and Buyse, M. (2005). *The Evaluation of Surrogate Endpoints.* New York: Springer.

Burzykowski, T., Molenberghs, G., and Buyse, M. (2004). The validation of surrogate endpoints using data from randomized clinical trials: a case-study in advanced colorectal cancer. *Journal of the Royal Statistical Society, Series A*, **167**, 103–124.

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

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

Cortiñas Abrahantes, J., Molenberghs, G., Burzykowski, T., Shkedy, Z., and Renard, D. (2004). Choice of units of analysis and modeling strategies in multilevel hierarchical models. *Computational Statistics and Data Analysis*, **47**, 537–563.

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

Fleming, T.R. and DeMets, D.L. (1996). Surrogate endpoints in clinical trials: are we being misled? *Annals of Internal Medicine*, **125**, 605–613.

Freedman, L.S. (2001). Confidence intervals and statistical power of the 'Validation' ratio for surrogate or intermediate endpoints. *Journal of Statistical Planning and Inference* **96**, 143–153.

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

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

Kay, S.R., Fiszbein, A., and Opler, L.A. (1987) The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin* **13**, 261–276.

Kay, S.R., Opler, L.A., and Lindenmayer, J.P. (1988) Reliability and validity of the Positive and Negative Syndrome Scale for schizophrenics. *Psychiatric Research* **23**, 99-110.

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

Tibaldi, F.S, 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*, **73**, 643–658.