Implementing the meta-analytic approach for the evaluation of surrogate endpoints in SAS and R: a word of caution

Peer-reviewed author version

ONG, Fenny; Wang, Jingzhao; VAN DER ELST, Wim; VERBEKE, Geert; MOLENBERGHS, Geert & ALONSO ABAD, Ariel (2022) Implementing the meta-analytic approach for the evaluation of surrogate endpoints in SAS and R: a word of caution. In: Journal of biopharmaceutical statistics (Print), 32 (5) , p. 705-716.

DOI: 10.1080/10543406.2021.2011903
Handle: http://hdl.handle.net/1942/36511

Implementing the meta-analytic approach for the evaluation of surrogate endpoints in SAS and R: A word of caution

Fenny Ong[a]*, Jingzhao Wang[b], Wim Van der Elst[c], Geert Verbeke[d,a], Geert Molenberghs[a,d] and Ariel Alonso[d]

[a]I-BioStat, Universiteit Hasselt, B-3590 Diepenbeek, Belgium; [b]Center for Drug Evaluation, NMPA, People's Republic of China; [c]The Janssen Pharmaceutical, companies of Johnson & Johnson, Belgium; [d]I-BioStat, KU Leuven, B-3000 Leuven, Belgium

*corresponding author: Fenny Ong, I-BioStat, Universiteit Hasselt, Campus Diepenbeek, Agoralaan gebouw D, BE3590 Diepenbeek, Belgium. E-mail: fenny.ong@uhasselt.be

Implementing the meta-analytic approach for the evaluation of surrogate endpoints in SAS and R: A word of caution

The meta-analytic approach has become the gold-standard methodology for the evaluation of surrogate endpoints and several implementations are currently available in SAS and R. The methodology is based on hierarchical models that are numerically demanding and, when the amount of data is limited, maximum likelihood algorithms may not converge or may converge to an ill-conditioned maximum such as a boundary solution. This may produce misleading conclusions and have negative implications for the evaluation of new drugs. In the present work, we explore the use of two distinct functions in R (*lme* and *lmer*) and the *MIXED* procedure in SAS to assess the validity of putative surrogate endpoints in the meta-analytic framework, via simulations and the analysis of a real case study. We describe some problems found with the *lmer* function in R that led to a poorer performance as compared with the *lme* function and *MIXED* procedure.

Keywords: surrogate markers; lme; lmer; proc mixed; meta-analytic approach

1. Introduction

Surrogate endpoints have helped pharmaceutical companies to carry out faster and more efficient clinical trials. They have also contributed to improve our understanding of some diseases and to identify and track public health concerns. Cholesterol, blood sugar levels, and blood pressure are just some examples of surrogate endpoints that have played a prominent role in medical research and practice, but the use of surrogate endpoints has also raised some controversy (Micheel and Ball 2010). For instance, long-term hormone replacement therapy significantly lowered "bad" cholesterol and raised "good" cholesterol in women, but at the same time, it increased their chances of heart attacks and strokes (Writing Group for the Women's Health Initiative Investigators 2002). In spite of these drawbacks, the potential of surrogate endpoints to speed up the approval of new therapeutics remains appealing, such as in the context of evaluating the efficacy of urgently needed vaccines during the SARS or Covid-19 pandemic.

The first attempts to define and quantify surrogacy took place over 30 years ago in the so-called single trial setting (STS), i.e., the putative surrogate was evaluated using data from a single clinical trial. Methods developed in the STS suffered from many conceptual problems and, at the beginning of the 21st century, a new approach was introduced based on meta-analysis. The so-called meta-analytic approach offered an alternative method to carry out the evaluation of surrogate endpoints and became the gold-standard in this domain. It assumes that the new treatment is evaluated in a sequence of clinical trials that target slightly different populations (owing to differences in the inclusion/exclusion criteria, protocol, and/or differences between countries where the trials take place, among other reasons) and estimate potentially different expected causal treatment effects for the surrogate and true endpoints. The methodology is based on hierarchical models; it is numerically demanding, and convergence issues are often encountered, especially when only a few trials are available.

Nowadays there are several functions or procedures available to fit a variety of hierarchical models. It is of interest in the present work to compare some functions that are implemented in standard software packages and widely used by statisticians in industry and academia to assess the validity of surrogate endpoints. More specifically, we explored the performance of the methodology using the *lmer* and *lme* functions in R as well as the *MIXED* procedure in SAS (Bates et al. 2020; Pinheiro et al. 2020; SAS Institute Inc. 2020).

The rest of the manuscript is organized as follows. In Section 2, the meta-analytic approach is introduced. In Section 3, related software implementation is explained and a case study in schizophrenia is analysed to provide better insight about the application. A simulation study, used to compare the performance of the aforementioned functions, is

described in Section 4. The results of the simulations are presented in Section 5. Finally, some conclusions are given in Section 6.

2. The meta-analytic approach: Lights and shadows

Let us assume that data from $i = 1, \ldots, N$ clinical trials are available, in the ith of which $j = 1,\ldots,n_i$ subjects are enrolled. Further, let us denote the true and surrogate endpoints for patient j in trial i by T_{ij} and S_{ij}, respectively, and the indicator variable for the new treatment by Z_{ij}. The random treatment allocation in a clinical trial context naturally leads to the following bivariate model:

$$\begin{cases} T_{ij} = \mu_{Ti} + \beta_i Z_{ij} + \varepsilon_{Tij} \\ S_{ij} = \mu_{Si} + \alpha_i Z_{ij} + \varepsilon_{Sij} \end{cases} \qquad (1)$$

where μ_{Ti} and μ_{Si} are trial-specific intercepts, β_i and α_i are trial-specific expected causal treatment effects and ε_{Tij} and ε_{Sij} are correlated error terms, assumed to be zero-mean normally distributed with covariance matrix:

$$\Sigma = \begin{pmatrix} \sigma_{TT} & \sigma_{TS} \\ \sigma_{TS} & \sigma_{SS} \end{pmatrix} \qquad (2)$$

Notice that now the evaluation exercise is carried out across different populations and one can decompose the trial-specific parameters in the following way:

$$\begin{pmatrix} \mu_{Si} \\ \mu_{Ti} \\ \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} m_{Si} \\ m_{Ti} \\ a_i \\ b_i \end{pmatrix} \qquad (3)$$

where the second term on the right-hand side of (3) is assumed to follow a zero-mean normal distribution with covariance matrix:

$$D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ d_{ST} & d_{TT} & d_{Ta} & d_{Tb} \\ d_{Sa} & d_{Ta} & d_{aa} & d_{ab} \\ d_{Sb} & d_{Tb} & d_{ab} & d_{bb} \end{pmatrix} \quad (4)$$

Based on these ideas, Buyse et al. (2000) proposed to assess surrogacy at two different levels, the so-called trial and individual level. At the trial-level these authors quantify surrogacy using the coefficient of determination:

$$\hat{R}^2_{trial} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}} \quad (5)$$

This metric always lies in the unit interval, it takes value zero if and only if β_i is independent of (μ_{Si}, α_i) and value one when the former is deterministically related to the later. Clearly, any value in between will give different evidence about the validity of the surrogate at this level. A special case arises where the prediction of the treatment effect on the true endpoint can be done independently of the random intercept associated with the surrogate. In that case, the coefficient of determination reduces to:

$$R^2_{trial} = \frac{d^2_{ab}}{d_{aa} d_{bb}} \quad (6)$$

At the individual-level, surrogacy is defined as the association between both endpoints after adjustment by trial and treatment and it is captured by the coefficient of determination:

$$R^2_{ind} = \frac{\sigma^2_{TS}}{\sigma_{SS}\sigma_{TT}} \quad (7)$$

This metric has a similar interpretation but, unlike trial-level surrogacy, the individual-level does not depend on the treatment and it can be interpreted as a quantification of the biological plausibility of the surrogate. An endpoint producing a high individual-level

surrogacy is always a potential surrogate. However, it may fail to be predictive at the trial-level for a specific treatment that follows a causal path that completely avoids it. As expressed in Equation (7), the individual-level surrogacy is based on the variance-covariance matrix of the residual Σ, and since the replication at this level is usually large, there are typically no issues with the estimation of the R^2_{ind} (Van der Elst et al. 2015). In addition, the main goal of surrogate endpoints is to predict the treatment effect on the true endpoint and, consequently, trial-level surrogacy is commonly the most relevant dimension. Therefore, the focus of this study will be limited to the trial-level surrogacy only.

The meta-analytic approach overcame many of the conceptual problems that previous methods had, but it also created a number of serious practical challenges. For instance, fitting the above hierarchical model often implies a considerable computational burden. Tibaldi et al. (2003) suggested several simplifications to tackle this problem, like treating the trial-specific parameters in (1) as fixed effects in a two-stage approach. At the first stage, the bivariate regression model (1) is fitted for each trial separately and the trial-specific parameters are treated as fixed effects. Then, at the second stage, the estimated treatment effects on the true endpoint are regressed on the estimated treatment effects on the surrogate endpoints and the intercepts associated with the surrogate endpoints as:

$$\hat{\beta}_i = \gamma_0 + \gamma_1 \hat{\mu}_{Si} + \gamma_2 \hat{\alpha}_i + \epsilon_i \qquad (8)$$

Essentially, the trial-level surrogacy metric R^2_{trial} is estimated using the coefficient of determination obtained by regressing $\hat{\beta}_i$ on $(\hat{\mu}_{Si}, \hat{\alpha}_i)$. Another important limitation of the methodology is that it requires a large amount of data. In fact, one needs to have several clinical trials where the expected causal effects of the treatment on both

the surrogate and true endpoints are assessed. The availability of such data may be a serious problem, especially at the early stages of the drug development process, when surrogate endpoints are needed most. A possible workaround to this problem is to use other units of analysis. The choice of an alternative unit of analysis, e.g., centre or country, may depend on practical considerations such as the information available in the data, 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 when the number of units and the number of patients per unit are both sufficiently large. Of course, after choosing a specific unit for the analysis, one always has to reflect carefully on the status of the results obtained. Arguably, they may not be as reliable as one might hope for, and one should undertake every effort possible to increase the amount of information available. This issue has been covered at large by Cortiñas Abrahantes et al. (2004) and we refer the interested reader to this work for more details.

3. Software implementation and case study

In this section, we introduce a commonly used software implementation to assess the validity of surrogate endpoints within the meta-analytic framework. We then exemplify the methodology using a case study in schizophrenia to provide better insight about the application. The interested reader can find the details about the SAS and R codes in the Supplementary Material accompanying the manuscript, while more generic code and the illustration about how it relates with the notations in Section 2 are given in Section 3.1.

3.1 Software implementation

We compared four functions in this study: *lme* with description of the within-group correlation and heteroscedasticity structure, indicated by the additional *correlation* and

weights statement in the model (denoted as *lme1* this code allows for heteroscedastic and correlated error terms), *lme* without description of the within-group correlation and heteroscedasticity structure (denoted as *lme2* this code considers homoscedastic and uncorrelated error terms), *lmer*, and *proc MIXED* with unstructured (UN) variance-covariance parameterization for both the random-effects and residual matrices, respectively. The *lme* function is provided in the R package *nlme* (Version 3.1-149 was used in this study) whereas *lmer* can be found in the R package *lme4* (Version 1.1-25 was used in this study). As for the *MIXED* procedure, it is incorporated within SAS/STAT software (Version 15.2 was used in this study).

The *lme1* model is implemented using the code:

lme(response ~ -1 + endpoint + endpoint:treat, random = ~ -1 + endpoint + endpoint:treat | trial, data = <data>, correlation = corSymm(form = ~ 1 | trial/subject), weights = varIdent(form = ~ 1 | endpoint)),

where the *response* vector contains the outcomes of the surrogate and true endpoints for the patient given by the variable *subject*, the *endpoint* variable indicates if the value of the response vector corresponds to the surrogate or true endpoint, and the variables *treat* and *trial* refer to the treatment received by the patients and the trial he/she belongs to, respectively. The first part of the code (*response ~ -1 + endpoint + endpoint:treat*) allows us to estimate the average intercepts and treatment effects across trials given by the first vector on the right side of equation (3). In addition, the random statement (*random = ~ -1 + endpoint + endpoint:treat | trial*) specifies the random effects of the model given by the second vector on the right hand of equation (3) and allows to estimate the variance-covariance matrix D in equation (4). Finally, the statement *correlation = corSymm(form = ~ 1 | trial/subject)* and *weights = varIdent(form = ~ 1 | endpoint)* specify the variance-covariance matrix for the error structure Σ in equation (2), with *varIdent(form = ~ 1 |*

endpoint) indicating that the errors are heteroscedastic. The *lme2* model differs from the *lme1* only in the absence of the *correlation* and *weights* arguments.

The following code is used to implement the *lmer* model:

lmer(response ~ -1 + endpoint + endpoint:treat + (-1 + endpoint + endpoint:treat | trial), data = <data>),

where, similarly to *lme1*, the part of the code (*response ~ -1 + endpoint + endpoint:treat*) allows us to estimate the average intercepts and treatment effects across trials while the code (*-1 + endpoint + endpoint:treat | trial*) specifies the random effects of the model. The error structure assumes independence and equal variances.

Finally, in the SAS implementation we used this code:

PROC MIXED data = <data>;

CLASS endpoint subject trial;

*MODEL response = endpoint endpoint*treat;*

*RANDOM endpoint endpoint*treat / subject = trial;*

REPEATED endpoint / subject = subject(trial);

RUN;

where all variables (*response*, *endpoint*, *subject*, *treat*, and *trial*) have been explained previously. The *CLASS* statement indicates the categorical variables used in the model. The formula in the *MODEL* statement allows us to estimate the average intercepts and treatment effects across trials (first term on the right-hand side of equation 3), while the *RANDOM* statement specifies the random effects of the model given by the second vector on the right-hand of equation (3). Lastly, the *REPEATED* statement indicates the variance-covariance matrix for the error structure Σ.

The default settings of each function were used in the present work. More information about the algorithm, optimization, and convergence criteria adopted by each

function can be found in the Supplementary Material. However, a detailed comparison between different settings is beyond the scope of this manuscript. We believe that the findings presented here will still be generally valid despite different convergence criteria between functions. Other than that, it is our intention to minimize any restrictions in the programming step to remain as close as possible to common practice, where typical users primarily stick to the default setting in their analysis.

3.2 Schizophrenia study

In this subsection, we introduce the motivating case study. 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 data set can be accessed in the R library *Surrogate* (Van der Elst et al. 2020) and has previously been used in the surrogate evaluation literature (Alonso et al. 2017; Burzykowski et al. 2005). Schizophrenia is one of the most disabling and emotionally devastating illnesses affecting humans and it is characterized by a constellation of distinctive and predictable symptoms. The symptoms that are most commonly associated with the disease are called positive symptoms, that denote the presence of grossly abnormal behaviour. Less obvious than the positive symptoms but equally serious are the deficit or negative symptoms that represent the absence of normal behaviour. These include flat or blunted affect (i.e. lack of emotional expression), apathy, and social withdrawal. Several measures exist to assess a patient's global condition. Two sensitive psychiatric scales are the Positive and Negative Syndrome Scale (PANSS) and the Brief Psychiatric Rating Scale (BPRS). PANSS provides an operationalized, drug-sensitive instrument, which is highly useful for both typological and dimensional assessment of schizophrenia while BPRS is a sub-scale of PANSS. Interest is in knowing

to which extent a simpler and easier to administered scale like BPRS can be used as a substitute for a more reliable and complex scale like PANSS, when assessing the efficacy of these or similar drugs.

There were a total of 2128 patients in the complete data set, 537 and 1591 of which were in the active control and experimental treatment group, respectively. Given the insufficient number of trials, the treating physician has often been used as the clustering variable when analysing these data in the meta-analytic framework. Following Equation (1) and (3) in Section 2, each trial has its own intercept and treatment effect for the true (PANSS) and surrogate (BPRS) endpoints, respectively. The trial-specific parameters were then decomposed into the fixed- and random-effects. The random-effects are assumed to follow a zero-mean normal distribution with covariance matrix D as expressed in Equation (4). More explanation about the practical implementation of the model and the software code is detailed in the Supplementary Material.

A summary of the results is shown in Table 1. ***[Table 1 near here]*** A special issue catches immediately the eye, the *proc MIXED* procedure with unstructured parameterization resulted in a negative $\hat{R}^2_{trial}$. This undesirable problem is the direct consequence of a non-positive-definite $\hat{D}$ matrix as it is clearly seen in the presence of a negative minimum eigenvalue. Oddly enough, the unstructured parameterization in *proc MIXED* forces the diagonal of $\hat{D}$ to be positive but the complete matrix may still not be positive-definite. To tackle this issue, another parameterization for the *D* matrix, the so-called non-diagonal factor-analytic structure with 4 factors (*FA0(4)*), was considered. This parameterization applies a log-Cholesky decomposition to the *D* matrix to address positive-definiteness constraints and results in substantial simplification of the optimization problem (Pinheiro and Bates 1996; West et al. 2015). This analysis produced

a moderate value of trial-level surrogacy but here again an almost singular matrix $\hat{D}$ was obtained.

The implementation of *lmer* led to the matrix $\hat{D}$ with the highest condition number and a degenerated value for $\hat{R}^2_{trial}$. A more reasonable estimate was obtained with *lme1*. In fact, this implementation produced the estimated $\hat{D}$ with the smallest condition number and a large but acceptable estimate for trial-level surrogacy. Given that BPRS is a sub-scale of PANSS such a large value of $\hat{R}^2_{trial}$ may not be completely unexpected. However, the $\hat{D}$ matrix obtained from *lme1* was still nearly singular and, hence, all these results should be interpreted with extreme caution. Finally, the two-stage approach introduced in Section 2 was also applied and its results were in close agreement with those obtained from *lme1*. The numerical issues encountered in this case study are not uncommon and in the next section they are studied in more detail via simulation.

4. Simulation study

The simulations aim at mimicking the case study, i.e., the scenario in which an alternative unit of analysis is used to assess trial-level surrogacy in a meta-analytic framework. The random-effects model based upon combining model (1) and (3) was used to generate the data. The treatment allocation (Z_{ij}) was coded as -1 and 1 for the control and experimental group, respectively, to ensure the same components of variability in both treatment groups (Burzykowksi et al. 2005). The variable Z_{ij} was generated using a Bernoulli distribution with probability $\pi = 0.5$. In all simulations the mean structure parameters were fixed at $\mu_S = 450$, $\mu_T = 500$, $\alpha = 300$, and $\beta = 500$, while the between-trial heterogeneity was defined as:

$$D = \gamma \begin{pmatrix} 1000 & 400 & 0 & 0 \\ 400 & 1000 & 0 & 0 \\ 0 & 0 & 1000 & 707.107 \\ 0 & 0 & 707.107 & 1000 \end{pmatrix}$$

The block-diagonal structure of the previous matrix implies that the trial-specific intercepts in model (1) are independent of the trial-specific treatment effects, and hence trial-level surrogacy can be calculated as the correlation between α_i and β_i. More specifically, $R^2_{trial} = corr(a_i, b_i)^2 = 0.5$.

Several conditions were varied. First, the number of clusters $N = \{5, 10, 20\}$ were used to evaluate the performance of the meta-analytic approach in situations where it is implemented in a small, moderate, or large number of units. The term "cluster" was used interchangeably with "trial" and "unit of analysis", with the latter serving as the most general term. The between-cluster variability (D), was either larger ($\gamma = 1$) or smaller ($\gamma = 0.1$) than the within-cluster variability. Furthermore, for each combination of the number of clusters and the value of γ, six simulation settings were defined based on the values of other parameters and a complete description of the settings considered are given in Table 2. ***[Table 2 near here]*** For instance, while the individual-level surrogacy was kept constant at $R^2_{ind} = corr(\varepsilon_{Sij}, \varepsilon_{Tij})^2 = 0.5$, the within-cluster (residual) variability was differentiated into two conditions, homoscedastic ($\sigma_{SS} = \sigma_{TT}$ in Simulation 1) and heteroscedastic ($\sigma_{SS} \neq \sigma_{TT}$ in Simulations 2 – 6), as were given by:

$$\Sigma_{homo} = \begin{pmatrix} 300 & 212.132 \\ 212.132 & 300 \end{pmatrix}$$

$$\Sigma_{hetero} = \begin{pmatrix} 500 & 158.114 \\ 158.114 & 100 \end{pmatrix}$$

Balanced and unbalanced cluster sizes were considered. In the balanced scenario, all cluster sizes were equal, and three settings were evaluated: $n_i = 20$ in Simulations 1 –

2, n_i = 100 in Simulation 3 and n_i = 500 in Simulation 4. The smaller cluster size is, on average, the number of observations one often encounters when units such as centres are used as cluster variable, while larger cluster sizes might be found when units like countries or trials are used (Alonso et al. 2017; Burzykowski et al. 2005). In the unbalanced scenario (Simulations 5 – 6) cluster sizes were determined based on a draw from a normal distribution with mean $\mu = n_i = \{20, 500\}$ and rounded to the nearest integer. The standard deviation of the cluster size was determined as a fraction of the mean, i.e. $\sigma_n = 0.25\ n_i$.

Simulations 1 and 2 were intended to explore the performance of the R and SAS implementations of hierarchical models in two practically relevant scenarios when evaluating surrogate endpoints. Indeed, when an alternative unit of analysis such as centre is used, small cluster sizes with homoscedastic or heteroscedastic variances are often encountered. Given that the *lmer* function, in its current implementation, does not allow modelling heteroscedastic and correlated error terms, Simulations 3 – 4 were considered to evaluate the impact of such misspecification on the results at moderate and large cluster sizes. Finally, the potential effect of unbalanced cluster sizes was explored in Simulations 5 – 6 with a small and large average cluster size, respectively.

A total of 500 data sets were generated for each scenario and the simulated data sets were then analysed based on the meta-analytic approach using the *lmer* and *lme* functions in R as well as the *MIXED* procedure in SAS. Similar parameter values were used in the simulation study by Van der Elst et al. (2015) and Flórez et al. (2019), though they had different objectives. During our simulation study, we compared results across several software procedures to evaluate the surrogate endpoint using the meta-analytic approach. In addition, two simplified model-fitting strategies, i.e. treating the trial-specific parameters as fixed-effects in a two-stage approach and simplifying the random-

effects structure by assuming that there is no heterogeneity in the random intercepts for the surrogate and true endpoints, were also explored in the study. The outcome of interest was the relative bias in the estimate of trial-level surrogacy, defined as the ratio between the difference of the mean $\hat{R}^2_{trial}$ and the target R^2_{trial} (which is 0.5) over the target R^2_{trial}. The proper convergence rate, i.e., when the model converged and the variance-covariance matrix of the random-effects (D) was positive-definite, was also evaluated.

5. Simulation results

5.1 Convergence rate

Table 3 displays the rate of properly-converged data sets after being analysed by each function. ***[Table 3 near here]*** Proper convergence refers to the condition where the model converges and the variance-covariance matrix of the random-effects (D) is positive-definite. Since it is important to guarantee that the value of R^2_{trial} lies in the unit interval, the condition where the D matrix is positive-definite is of utmost interest. In Simulation 1 where the variances of the residual terms were equal, the convergence rate obtained from *lme1* and the *MIXED* procedure was less than 15% when the number of clusters was small (N = 5) and the between-cluster variability was smaller (γ = 0.1) than the within-cluster variability. On the other hand, the *lmer* function seemed to produce a higher convergence rate (around 65%) in this strenuous setting. However, as we will see later, the $\hat{R}^2_{trial}$ emanating from this function tended to overestimate the trial-level surrogacy due to convergence to an ill-conditioned maximum. Given that they are essentially fitting the same model, the rather different convergence rates found between *lme2* and *lmer* in this and other settings is eye catching. The convergence rate was substantially improved when the number of clusters was increased and/or the between-cluster variability was larger (γ = 1) than the within-cluster variability. For *lme1*, *lme2*, and *MIXED*, we observed

improved convergence with larger between-cluster variability. However, this is slightly reversed with *lmer*.

The results obtained in the heteroscedastic setting studied in Simulation 2, were similar to those of Simulation 1 for each function. Moreover, as can be seen in Simulation 3 (n_i = 100) and Simulation 4 (n_i = 500), when the cluster size was increased, the convergence rate was also improved, prominently when the *lme1* or *proc MIXED* procedure were used. In agreement with the study by Van der Elst et al. (2015), we found that the impact of imbalance in cluster size on the proper convergence rate was small. This can be observed by comparing the result from Simulation 2 and 6 when the mean cluster size was small (n_i = 20) or Simulation 4 and 5 when the mean cluster size was larger (n_i = 500). Finally, as expected for the simplified model-fitting strategies, the convergence rates were significantly improved in all simulation conditions.

5.2 Trial-level surrogacy: Relative bias

Table 4 summarizes the results obtained when proper convergence was achieved, i.e., the estimates were calculated based on the data sets where the corresponding function (*lme*, *lmer*, or *proc MIXED*) converged to a maximum with a positive-definite *D* matrix. ***[Table 4 near here]*** Let us start by discussing the results obtained in the setting in which all functions fitted the correct model, i.e., Simulation 1. Interestingly, the $\hat{R}^2_{trial}$ emanating from the *lmer* and *lme2* functions exhibited a large positive relative bias for all values of N when γ = 0.1. This is a rather unexpected result given that both implementations describe the correct data generating mechanism. More in line with expectations, *proc MIXED* and *lme1* exhibited much smaller relative bias that decreased with the number of clusters. Furthermore, when the between-cluster variability was larger than the within-cluster variability (γ = 1) all functions delivered similar results.

In Simulations 2 – 4 *lme2* and *lmer* are fitting misspecified models, whereas *lme1* and *proc MIXED* are correctly describing the data generating mechanism. When the within-cluster variability is larger than the between cluster variability the impact of the misspecification is substantial for small cluster sizes (Simulation 2) and noticeable for moderate sizes (Simulation 3). However, for large cluster sizes (Simulation 4) all implementations lead to comparable results, i.e., the impact of the misspecification becomes negligible. As it can be clearly seen in Simulations 5 – 6, a similar behaviour is observed for unbalanced cluster sizes, i.e., the misspecification has a large impact when the average cluster size is small (Simulation 6) and it becomes negligible when the average cluster size is large (Simulation 5). In general, the two-stage approach produced smaller relative bias in all simulations compared to the results from the other functions, except in the smallest number of cluster ($N = 5$) setting. Meanwhile, the simplified model generated the smallest relative bias amongst all the other functions, which is sensible considering the simulation data-generating mechanism was based on this simplified model. Finally, the Monte Carlo standard errors are also reported in Table 5. They do not vary between functions within the same simulation and setting condition. As predicted, the Monte Carlo standard errors decrease when the number of clusters are higher.

6. Conclusions

Nowadays there are several functions and procedures available to fit a variety of hierarchical models. Therefore, it is important to know the constraints and limitations of different implementations when tackling the evaluation of surrogate candidates. To our knowledge, until now there have not been studies comparing the performance of different software tools to assess the validity of surrogate endpoints within the meta-analytic framework. The present work aims to tackle this gap by comparing the performance of three very popular implementations of hierarchical models, when assessing the validity

of surrogate markers in the meta-analytic framework. Rather surprisingly, some substantial differences were actually found.

The performance of *lme* with description of the within-group correlation and heteroscedasticity structure, indicated by the additional *correlation* and *weights* statement in the model (denoted as *lme1*), *lme* without description of the within-group correlation and heteroscedasticity structure (denoted as *lme2*), *lmer*, and *proc MIXED* with unstructured (UN) variance-covariance parameterization for both the random-effects and residual matrices was investigated in a case study in schizophrenia as well as the simulation study.

From the analysis of the case study in schizophrenia, some important messages can be drawn. Without the necessity for an in-depth understanding, one may need to be aware of different limitations and constraints in some software procedures and their impact on the results. In the evaluation of a candidate surrogate endpoint, the estimation of the covariance parameters of the random effects and residuals (D and Σ matrices, respectively) is crucial. The proper estimation of these matrices requires the numerical optimization of the log-likelihood functions, subject to constraints imposed on the parameters to ensure the positive-definiteness of the D and Σ matrices (West et al. 2015). It can sometimes happen that the iterative estimation routines converge to a value that lies very close to or outside the boundary of the parameter space, leading to the violation of positive-definiteness.

West et al. (2015) proposed some alternative approaches for fitting a model when problems in the estimation of the covariance occur. When convergence fails because of few and highly unbalanced trials, Van der Elst et al. (2015) used multiple imputation to reduce model convergence problems. A non-iterative unbiased estimator based on the so-called split-sample methodology and pseudo-likelihood may also be considered as

another option to alleviate the computational difficulties in the surrogacy evaluation based on this meta-analytic approach (Flórez et al. 2019).

It is important to note that at the time of this writing, the *lmer* function does not allow to fit models with heterogeneous and correlated residual variance structure. Among other possible reasons, this limitation might explain the large relative bias observed on the estimated trial-level surrogacy obtained with the function. However, the bias seems to get smaller when the cluster size is moderate or large and the between cluster variability is larger than the within-cluster variability. Interestingly, even when the residuals are homoscedastic and independent, the results obtained with *lme* are often less biased than those obtained with *lmer*.

More generally, the simulation study seems to indicate that *lme* and *proc MIXED* produced lower relative bias for R^2_{trial} compared to *lmer*. In general, one should interpret the parameter estimates with caution, especially when there are indications that the estimate of the D matrix may not be positive-definite. In such a situation, alternative approaches like the two-stage approach or the simplified model may be of value. These two strategies exhibited rather good results concerning bias and they offer a greater computational stability. Finally, we want to point out that caution is needed when the results of a simulation study are extrapolated beyond the settings used to generate the data. For instance, in our simulations the random intercepts and treatment effects are independent and the error terms are normally distributed. Further studies will be needed to explore if similar results are also obtained when the random intercepts and treatment effects are correlated or when the surrogate and/or true endpoint are not normal.

References

Alonso, A., T. Bigirumurame, T. Burzykowski, M. Buyse, G. Molenberghs, L. Muchene, N. J. Perualila, Z. Shkedy, and W. Van der Elst. 2017. *Applied surrogate endpoint evaluation methods with SAS and R*. Boca Raton, FL: Chapman & Hall/CRC.

Bates, D., M. Maechler, B. Bolker, S. Walker, R. H. B. Christensen, H. Singmann, B. Dai, F. Scheipl, G. Grothendieck, P. Green, J. Fox, A. Bauer, and P. N. Krivitsky. lme4: Linear mixed-effects models using 'eigen' and S4 (R Package Version 1.1-25). Last Modified October 23, 2020. https://cran.r-project.org/web/packages/lme4/index.html.

Burzykowski, T., G. Molenberghs, and M. Buyse. 2005. *The evaluation of surrogate endpoints*. New York: Springer-Verlag.

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

Cortiñas Abrahantes, J., G. Molenberghs, T. Burzykowski, Z. Shkedy, and D. Renard. 2004. Choice of units of analysis and modeling strategies in multilevel hierarchical models. *Computational Statistics and Data Analysis* 47(3):537-63.

Flórez, A. J., G. Molenberghs, G. Verbeke, and A. Alonso. 2019. A closed-form estimator for meta-analysis and surrogate markers evaluation. *Journal of Biopharmaceutical Statistics* 29(2):318-32.

Micheel, C. M. and J. R. Ball. 2010. *Evaluation of biomarkers and surrogate endpoints in chronic disease*. Washington, DC: National Academies Press.

Pinheiro, J. C. and D. M. Bates. 1996. Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing* 6:289-96.

Pinheiro, J., D. Bates, S. DebRoy, D. Sarkar, EISPACK authors, S. Heisterkamp, B. Van Willigen, and R-core. nlme: Linear and nonlinear mixed effects models (R Package Version 3.1-149). Last Modified August 23, 2020. https://cran.r-project.org/web/packages/nlme/index.html.

SAS Institute Inc. 2020. *SAS/STAT®15.2 User's Guide*. Cary, NC: SAS Institute Inc.

Tibaldi, F. S., J. Cortiñas Abrahantes, G. Molenberghs, D. Renard, T. Burzykowski, M. Buyse, M. Parmar, T. Stijnen, and R. Wolfinger. 2003. Simplified hierarchical linear models for the evaluation of surrogate endpoints. *Journal of Statistical Computation and Simulation* 73:643-58.

Van der Elst, W., L. Hermans, G. Verbeke, M. G. Kenward, V. Nassiri, and G. Molenberghs. 2015. Unbalanced cluster sizes and rates of convergence in mixed-effects models for clustered data. *Journal of Statistical Computation and Simulation* 86:2123-39.

Van der Elst, W., P. Meyvisch, A. J. Flórez, A. Alonso, H. M. Ensor, C. J. Weir, and G. Molenberghs. Surrogate: Evaluation of surrogate endpoints in clinical trials (R Package Version 1.7). Last Modified December 13, 2020. https://cran.r-project.org/web/packages/Surrogate/index.html.

West, B. T., K. B. Welch, and A. T. Galecki. 2015. *Linear mixed models: A practical guide using statistical software*. Boca Raton, FL: CRC Press Taylor & Francis Group.

Writing Group for the Women's Health Initiative Investigators. 2002. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: Principal results from the women's health initiative randomized controlled trial. *The Journal of the American Medical Association* 288(3):321-33.

Table 1. R^2 trial and condition number for Schizophrenia data: Results across software procedures.

Procedure	PANSS and BPRS		
	R^2 trial	ME	CN
SAS: two-stage approach	0.919	NA	NA
SAS: *proc MIXED* (UN)	-1.564	-0.289	962.451
SAS: *proc MIXED* (FA0)	0.685	1.107e-05	8.859e+06
R: *lme*	0.938	1.518e-05	39503.3
R: *lmer*	1	1.453e-08	9.806e+09

UN = unstructured; FA0 = factor-analytic; ME = minimum eigenvalue; CN = condition number; NA = not available

Table 2. Simulation parameter.

Parameter	Simulation					
	1	2	3	4	5	6
Cluster size (n_i)	20	20	100	500	500	20
Balance status	Balanced				Unbalanced	
σ_{SS}	300	100	100	100	100	100
σ_{TT}	300	500	500	500	500	500
σ_{ST}	212.132			158.114		
$\mu_S, \mu_T, \alpha, \beta$	450, 500, 300, 500					
$d_{11} = d_{22} = d_{33} = d_{44}$	1000					
$d_{12} = d_{21}$	400					
$d_{34} = d_{43}$	707.107					
$d_{13} = d_{31} = d_{14} = d_{41} = d_{23} =$ $d_{32} = d_{24} = d_{42}$	0					
R^2 trial target	0.5					
R^2 individual target	0.5					

Table 3. Proper convergence rate.

		Simulation 1		Simulation 2		Simulation 3		Simulation 4		Simulation 5		Simulation 6	
		γ		γ		γ		γ		γ		γ	
Function	Number of clusters	0.1	1	0.1	1	0.1	1	0.1	1	0.1	1	0.1	1
lme1	5	0.108	0.606	0.076	0.540	0.378	0.758	0.888	0.952	0.852	0.956	0.080	0.500
	10	0.804	0.998	0.704	0.986	0.984	0.998	1	1	1	0.998	0.654	0.996
	20	1	1	0.986	0.950	1	0.672	1	0.818	1	0.802	0.980	0.944
lme2	5	0.042	0.476	0.050	0.480	0.320	0.746	0.630	0.806	0.610	0.806	0.038	0.468
	10	0.484	0.966	0.520	0.968	0.916	0.926	0.856	0.854	0.846	0.862	0.494	0.976
	20	0.846	0.944	0.858	0.950	0.904	0.882	0.840	0.872	0.844	0.830	0.846	0.960
lmer	5	0.654	0.738	0.634	0.748	0.754	0.708	0.714	0.634	0.648	0.656	0.602	0.694
	10	0.846	0.824	0.850	0.830	0.924	0.816	0.878	0.812	0.900	0.842	0.868	0.848
	20	0.946	0.858	0.924	0.866	0.936	0.848	0.910	0.804	0.926	0.864	0.952	0.882
proc MIXED	5	0.110	0.614	0.094	0.564	0.418	0.806	0.724	0.914	0.716	0.906	0.086	0.548
	10	0.820	0.998	0.744	1	0.994	1	1	1	1	1	0.704	1
	20	1	1	0.996	1	1	1	1	1	1	1	0.988	1
Two-stage	5	1	1	0.998	0.998	1	1	1	1	1	1	1	1
approach	10	1	1	1	1	1	1	1	1	1	1	1	1
	20	1	1	1	1	1	1	1	1	1	1	1	1
Simplified	5	0.736	0.912	0.720	0.900	0.938	0.988	0.996	1	0.994	0.998	0.722	0.930
model	10	0.972	0.986	0.930	0.992	1	1	1	1	1	1	0.942	0.998
	20	0.998	1	0.998	1	1	1	1	1	1	1	0.992	1

Note: γ = between-cluster variability; *lme1* = *lme* with *correlation* and *weights* statement; *lme2* = *lme* without *correlation* and *weights* statement;

Simulation 1: $\sigma_{SS} = \sigma_{TT}$, balanced $n_i = 20$; Simulation 2: $\sigma_{SS} \neq \sigma_{TT}$, balanced $n_i = 20$; Simulation 3: $\sigma_{SS} \neq \sigma_{TT}$, balanced $n_i = 100$; Simulation 4: $\sigma_{SS} \neq \sigma_{TT}$, balanced $n_i = 500$;

Simulation 5: $\sigma_{SS} \neq \sigma_{TT}$, unbalanced n_i ($\mu = 500$); Simulation 6: $\sigma_{SS} \neq \sigma_{TT}$, unbalanced n_i ($\mu = 20$)

Table 4. Mean relative bias of R^2 trial obtained from each function when proper convergence occurred, i.e. when D was positive-definite.

		Simulation 1		Simulation 2		Simulation 3		Simulation 4		Simulation 5		Simulation 6	
		γ		γ		γ		γ		γ		γ	
Function	Number of clusters	0.1	1	0.1	1	0.1	1	0.1	1	0.1	1	0.1	1
lme1	5	0.183	0.385	0.194	0.351	0.340	0.390	0.384	0.412	0.394	0.404	0.285	0.344
	10	0.160	0.131	0.125	0.126	0.127	0.118	0.122	0.116	0.119	0.116	0.145	0.119
	20	0.076	0.049	0.078	0.031	0.067	0.022	0.051	0.045	0.050	0.025	0.062	0.020
lme2	5	0.030	0.360	0.155	0.353	0.321	0.396	0.395	0.407	0.369	0.409	0.146	0.360
	10	0.284	0.163	0.249	0.156	0.164	0.123	0.132	0.110	0.136	0.119	0.328	0.157
	20	0.372	0.082	0.313	0.073	0.114	0.058	0.060	0.037	0.066	0.037	0.301	0.073
lmer	5	0.687	0.437	0.649	0.435	0.479	0.378	0.387	0.374	0.396	0.343	0.705	0.432
	10	0.462	0.146	0.410	0.146	0.175	0.093	0.106	0.092	0.113	0.094	0.453	0.128
	20	0.419	0.068	0.348	0.071	0.099	0.046	0.051	0.040	0.055	0.041	0.353	0.066
proc MIXED	5	0.210	0.388	0.237	0.370	0.378	0.406	0.374	0.410	0.375	0.404	0.339	0.376
	10	0.166	0.131	0.142	0.132	0.135	0.118	0.122	0.116	0.119	0.117	0.182	0.119
	20	0.076	0.049	0.085	0.051	0.067	0.047	0.051	0.050	0.050	0.048	0.066	0.040
Two-stage	5	0.387	0.404	0.352	0.403	0.410	0.406	0.405	0.411	0.390	0.395	0.390	0.393
approach	10	0.108	0.123	0.069	0.116	0.103	0.115	0.115	0.116	0.094	0.093	0.065	0.081
	20	0.048	0.047	-0.006	0.040	0.047	0.045	0.047	0.049	0.013	0.012	-0.047	-0.005
Simplified	5	0.091	0.123	0.118	0.116	0.134	0.113	0.111	0.111	0.108	0.111	0.060	0.106
model	10	0.050	0.048	0.031	0.051	-0.001	-0.002	-0.009	-0.010	-0.015	-0.008	0.030	0.002
	20	0.041	0.035	0.053	0.035	0.009	-0.005	-0.009	-0.010	-0.011	-0.015	-0.027	-0.007

Note: γ = between-cluster variability; *lme1* = *lme* with *correlation* and *weights* statement; *lme2* = *lme* without *correlation* and *weights* statement;

Simulation 1: $\sigma_{SS} = \sigma_{TT}$, balanced $n_i = 20$; Simulation 2: $\sigma_{SS} \neq \sigma_{TT}$, balanced $n_i = 20$; Simulation 3: $\sigma_{SS} \neq \sigma_{TT}$, balanced $n_i = 100$; Simulation 4: $\sigma_{SS} \neq \sigma_{TT}$, balanced $n_i = 500$;

Simulation 5: $\sigma_{SS} \neq \sigma_{TT}$, unbalanced n_i ($\mu = 500$); Simulation 6: $\sigma_{SS} \neq \sigma_{TT}$, unbalanced n_i ($\mu = 20$)

Table 5. Monte Carlo standard errors of R^2 trial obtained from each function when proper convergence occurred, i.e. when D was positive-definite.

		Simulation 1		Simulation 2		Simulation 3		Simulation 4		Simulation 5		Simulation 6	
		γ		γ		γ		γ		γ		γ	
Function	Number of clusters	0.1	1	0.1	1	0.1	1	0.1	1	0.1	1	0.1	1
lme1	5	0.293	0.247	0.258	0.243	0.235	0.242	0.243	0.249	0.246	0.250	0.258	0.252
	10	0.240	0.220	0.238	0.220	0.230	0.218	0.224	0.217	0.222	0.217	0.232	0.220
	20	0.180	0.161	0.189	0.158	0.165	0.165	0.161	0.161	0.160	0.165	0.197	0.162
lme2	5	0.296	0.250	0.287	0.245	0.243	0.245	0.242	0.247	0.249	0.250	0.312	0.253
	10	0.236	0.224	0.245	0.223	0.234	0.218	0.224	0.220	0.219	0.219	0.240	0.219
	20	0.167	0.164	0.180	0.163	0.165	0.159	0.158	0.160	0.162	0.162	0.191	0.164
lmer	5	0.237	0.254	0.250	0.257	0.241	0.252	0.244	0.246	0.247	0.240	0.222	0.258
	10	0.240	0.223	0.247	0.225	0.232	0.217	0.222	0.223	0.221	0.217	0.258	0.224
	20	0.180	0.161	0.194	0.164	0.160	0.161	0.159	0.161	0.157	0.162	0.203	0.164
proc MIXED	5	0.295	0.247	0.278	0.243	0.234	0.245	0.238	0.247	0.244	0.249	0.241	0.253
	10	0.241	0.220	0.241	0.221	0.232	0.218	0.224	0.217	0.222	0.217	0.240	0.220
	20	0.180	0.161	0.192	0.162	0.165	0.161	0.161	0.160	0.160	0.160	0.198	0.165
Two-stage	5	0.261	0.252	0.266	0.253	0.245	0.251	0.247	0.250	0.250	0.256	0.253	0.260
approach	10	0.219	0.218	0.220	0.219	0.226	0.217	0.223	0.217	0.222	0.219	0.223	0.219
	20	0.155	0.159	0.162	0.159	0.159	0.161	0.160	0.160	0.162	0.164	0.167	0.164
Simplified	5	0.319	0.307	0.318	0.304	0.298	0.304	0.303	0.302	0.302	0.302	0.315	0.306
model	10	0.277	0.254	0.280	0.257	0.249	0.240	0.240	0.237	0.237	0.238	0.266	0.247
	20	0.200	0.179	0.214	0.179	0.175	0.169	0.168	0.167	0.168	0.167	0.216	0.181

Note: γ = between-cluster variability; *lme1* = *lme* with *correlation* and *weights* statement; *lme2* = *lme* without *correlation* and *weights* statement;

Simulation 1: $\sigma_{SS} = \sigma_{TT}$, balanced $n_i = 20$; Simulation 2: $\sigma_{SS} \neq \sigma_{TT}$, balanced $n_i = 20$; Simulation 3: $\sigma_{SS} \neq \sigma_{TT}$, balanced $n_i = 100$; Simulation 4: $\sigma_{SS} \neq \sigma_{TT}$, balanced $n_i = 500$; Simulation 5: $\sigma_{SS} \neq \sigma_{TT}$, unbalanced n_i ($\mu = 500$); Simulation 6: $\sigma_{SS} \neq \sigma_{TT}$, unbalanced n_i ($\mu = 20$)