Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computational Statistics & Data Analysis 47 (2004) 537–563
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Choice of units of analysis and modeling strategies in multilevel hierarchical models

José Cortiñas Abrahantes[a,*], Geert Molenberghs[a], Tomasz Burzykowski[a], Ziv Shkedy[a], Ariel Alonso Abad[a], Didier Renard[b]

[a] *Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, Building D, BB3590 Diepenbeek, Belgium*
[b] *Eli Lilly & Company, Mont Saint Guibert, Belgium*

Received 1 August 2003; accepted 5 December 2003

Abstract

Hierarchical models are common in complex surveys, psychometric applications, as well as agricultural and biomedical applications, to name but a few. The context of interest here is meta-analysis, with emphasis on the use of such an approach in the evaluation of surrogate endpoints in randomized clinical trials. The methodology rests on the ability to replicate the effect of treatment on both the true endpoint, as well as the candidate surrogate endpoint, across a number of trials. However, while a meta-analysis of clinical trials in the same indication seems the natural hierarchical structure, some authors have considered center or country as the unit, either because no meta-analytic data were available or because, even when available, they might not allow for a sufficient level of replication. This leaves us with two important, related questions. First, how sensible is it to replace one level of replication by another one? Second, what are the consequences when a truly three- or higher-level model (e.g., trial, center, patient) is replaced by a coarser two-level structure (either trial and patient or center and patient). The same or similar questions may occur in a number of different settings, as soon as interest is placed on the validity of a conclusion *at a certain level of the hierarchy*, such as in sociological or genetic studies. Using the framework of normally distributed endpoints, these questions will be studied, using both analytic calculation as well as Monte Carlo simulation.

Keywords: Linear mixed model; Meta-analytic approach; Random effects; Surrogate endpoint

* Corresponding author. Tel.: +32-11-26-82-15; fax: +32-11-26-82-99.
E-mail address: jose.cortinas@luc.ac.be (J.C. Abrahantes).

0167-9473/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2003.12.003

1. Introduction

In applied sciences, one is often confronted with the collection of *correlated data*. This generic term embraces a multitude of data structures, such as multivariate observations, clustered data, repeated measurements, longitudinal data, and spatially correlated data. Instances of this type of research can be encountered in virtually every empirical branch of science. Different areas of research will refer to the same or similar concepts with different terminology. For example, multilevel modeling (Goldstein, 1995) is a frequently encountered term in sociological applications, whereas in classical experimental design research one often refers to variance component models (Searle et al., 1992).

A hierarchical structure can consist of more than two levels and examples also abound in practice. Schooling systems, for instance, present an obvious multilevel structure, with pupils grouped into classrooms, which are nested within schools which themselves may be clustered within education authorities. Often in sample surveys, for cost-related reasons or administrative considerations, multistage sampling schemes are adopted. In multistage sampling, the sample is selected in stages, with the sampling units at each stage being sub-sampled from the larger units drawn at the previous stage. Thus, it immediately becomes apparent that a sample obtained by multistage sampling is hierarchical in nature and, therefore, we need to analyze such data using appropriate hierarchical techniques.

Sometimes, not only the design is hierarchical, but in addition, the formulation of the research question involves a particular level of such a hierarchy. There might then be clear dangers associated to misspecifying the hierarchical structure.

In surrogate marker evaluation, one may be interested in association between the true endpoint and the surrogate endpoints at different levels (individual-level and trial-level) but on the one hand data may have more than two levels while on the other hand one may have to resort to an alternative for the level of trial if not enough trial-level replication is at hand. This particular problem has motivated this research.

The aim of this paper is to evaluate the performance of different modelling strategies which allow us to tackle the problems described above. An overview of the methodological steps that have led to the multi-level setup in surrogate marker evaluation is given in Section 2. Two clinical studies in schizophrenia, where different units of analysis can be considered, are introduced in Section 2.1. An alternative area of application is found in survey research. Data from the Belgian Health Interview Survey are described in Section 3. The meta-analytic setting, to be used throughout the paper, is introduced in Section 4. The different analytic approaches are presented in Section 5. A simulation study is reported in Section 6. With the results of the simulation study in mind, the data are analyzed in Section 7.

2. Surrogate marker validation

Surrogate endpoints are referred to as endpoints that can replace or supplement other endpoints in the evaluation of experimental treatments or other interventions.

For example, surrogate endpoints are useful when they can be measured earlier, more conveniently, or more frequently than the endpoints of interest, which are referred to as the "true" endpoints (Ellenberg and Hamilton, 1989). Prentice (1989) proposed a formal definition of surrogate endpoints and outlined how potential surrogate endpoints could be validated.

Buyse et al. (2000) suggested the use of combined evidence from several clinical trials, such as in a meta-analysis, rather than from a single study. To this end, they needed to formulate a bivariate hierarchical model, accommodating the surrogate and true endpoints in a multi-trial setting.

Of course, the switch to a meta-analytic framework does not solve all problems surrounding surrogate marker validation in a definitive way. A result of the change to meta-analysis is that computationally rather involved statistical models have to be used. For the case of surrogates and true endpoints that are both normally distributed, Buyse et al. (2000) employed linear mixed-effects models (Verbeke and Molenberghs, 2000). Even in this case, which from a statistical modeling point of view can be considered a basic one, fitting such linear mixed models turns out to be surprisingly difficult. Tibaldi et al. (2003) carefully studied this computationally-oriented issue. They have investigated several simplified strategies to deal with the computational burden posed by using such hierarchical linear models, primarily in the context of validating surrogate markers. These strategies are ordered following three choices: (1) whether trial-specific parameters are treated as random or fixed, (2) whether the endpoints are treated as correlated or not (bivariate versus univariate) and (3) the method of dealing with measurement error. As a result of their investigation, they recommend simplified computational methods for two main reasons. First, the methods are generally faster and easier to implement with standard software. Second, Tibaldi et al. (2003) showed, through simulations, that the simplified approaches often perform almost as good as the more advanced methods, and moreover enjoy much better convergence properties. As a cautionary remark, it needs to be emphasized that these results were derived in the context of normally distributed endpoints. Different types of outcomes (e.g., of a generalized linear type) may lead to a somewhat different picture (see., for example Xiang et al., 2002; Concordet and Nunez, 2002).

Very important issues arise from the choice of levels within such a hierarchical analysis. Indeed, while perhaps the most natural choice seems to be the one of trial, in practice, other units have been used in the validation process. Indeed, several authors have made alternative choices for reasons of convenience which then leaves us with the question of whether such particular unit will provide us with a good representation of what is happening at the targeted trial level. For example, Buyse et al. (2000), Burzykowski et al. (2001), and Alonso et al. (2002) considered a number of case studies, using either center or country as unit of analysis, rather than trial. Given that the construction of the trial-level surrogacy hinges on the choice of independent unit which preferably is trial, it is necessary to study the impact of deviations of such a choice. There are two broad classes of deviations from such an approach. First, as was done in some of the works cited earlier, some authors change from trial to center, country, or investigator, for reasons of convenience. Indeed, one may have higher number of replication over such alternative units than over trial. Second, the true data

Table 1
Psychiatric study I: number of units with a given number of patients

Patients per unit (n)	Units with n patients	Patients per unit (n)	Units with n patients
2	29	10	2
3	18	11	4
4	23	12	2
5	16	13	3
6	9	15	1
7	12	18	1
8	10	21	1
9	6	30	1

generating mechanism can and will often be hierarchical with more than two levels, such as patients within centers and at the same time centers within trials. The impact of such deviations on the study result are important and will be considered here in terms of their statistical and numerical properties. Analytic considerations are supplemented with results from small sample simulations and illustrated using data from two clinical trials in schizophrenic patients.

2.1. Clinical studies in schizophrenia

The first of the two psychiatric studies in schizophrenic patients is based on a meta-analysis containing only five trials. This is insufficient to apply the meta-analytic methods. In all of the trials, information is also available on the investigators which treated the patients. Thus, we can also use investigator as the unit of analysis. For this case a total of 138 units are available for analysis, with the number of patients per unit ranging from 2 to 30. The true endpoint is Clinician's Global Impression (CGI). This 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 et al., 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. Table 1 shows the frequency of unit-specific sample sizes. Clearly, the majority of units consists of less than 5 patients.

Alternatively, one could also consider the main investigator as unit of analysis. For 4 out of the 5 trials, only one main investigator was used leading to extremely large investigator sites. This leads to a total number of 29 units with the number of patients per unit ranging from 4 to 450, 4 of which represent trials. Another possibility is to consider the countries where patients were treated, which fortunately is also available. Hence, we can also use country within trial as the unit of analysis. In this case a total of 19 units are available, with the number of patients per unit ranging from 9 to 128. The comparison of the three different choices will be used as an empirical assessment as to the importance the choice of unit can have on the results.

In addition, we will use data from an international equivalence trial (INT-10) on schizophrenic patients (Nair and the Risperidone Study Group, 1998). The trial included

206 schizophrenic patients. All patients received an equal daily amount of risperidone during 8 weeks, but 103 patients were randomized to a one-time daily intake (O.D.), while the remaining 103 patients were randomized to receive risperidone twice a day (B.I.D.). The surrogate and true endpoints are again PANNS and CGI, respectively. We will consider the investigator as the unit of analysis. This leads to a total of 34 units available for analysis with the number of patients per unit ranging from 2 to 15.

3. The Belgian Health Interview Survey

In 1997, the second Belgian Health Interview Survey took place (HIS1997). The HIS1997 was conducted to evaluate the usefulness of a periodic health-related survey, with the idea to collect information on the subjective health of the Belgian population, as well as on important predictor variables.

The survey is still ongoing (the second one took place in 2001) and the main goal of the HIS is to give a description of the health status of the overall population in Belgium as well as of the three regional subpopulations (Flemish, Walloon and Brussels region), and in addition of the German community.

The total number of successful interviews for the sample in the HIS1997 was set to 10,000. The sampling of the households and respondents is a combination of several sampling techniques: stratification, multistage sampling and clustering, and differential selection probabilities. The sampling of respondents took place in the following steps: (1) stratification by region and province, (2) selection of the municipalities within each stratum, (3) selection of a cluster of households within each municipality and (4) selection of respondents within a household. Such a multi-stage design with municipalities as primary selection units is a feasible solution.

Second- and third-stage selection units are households within municipalities and individuals within households, respectively. Municipalities are established administrative units. They are stable (in general they do not change during the time the survey is conducted) and they are easy to use in comparison with other specialized sources of data related to the survey. Municipalities are preferred above regions or provinces, because the latter are too large and too few. The large variation in the size of the municipalities is controlled for by systematically sampling within a province with a selection probability proportional to their size.

The three-level hierarchy encountered here, is simpler than the one in surrogate marker evaluation, where we have two endpoints (surrogate and true), and hence a pair of three-level structures. Strictly speaking, the latter situation could therefore be seen as a four-level structure.

4. Model description and setting

In this section, we will introduce a three-level model for normally distributed endpoints. This model will allow us to consider the fully general case of a three-way

hierarchy (e.g., patients within centers and centers within trials), as well as sub-cases that are of a two-level type. The emphasis will be on the surrogate marker situation, where such a model is needed for both the surrogate as well as the true endpoint. For cases such as the Health Interview Survey, the setting simplifies to just one hierarchical model. At the same time, the impact of misspecification by modelling the data as if they arose from a two-way structure, even though they were generated under a three-way model, can be assessed. In addition, the impact of considering the sub-unit effects as fixed, even though they are generated using a random-effects model, is studied. Let T_{ijk} and S_{ijk} be random variables denoting the true and the surrogate endpoints for subject $k=1,\ldots,n_{ij}$ in center $j=1,\ldots,N_i$ within trial $i=1,\ldots,M$. Further, let Z_{ijk} denote a binary treatment indicator. The full three-way random-effects model, as it was introduced by Buyse et al. (2000) for the two-way hierarchy, can the be written as

$$\begin{aligned} S_{ijk} &= \mu_S + m_{S_i} + m_{S_{ij}} + (\alpha + a_i + a_{ij})Z_{ijk} + \varepsilon_{S_{ijk}}, \\ T_{ijk} &= \mu_T + m_{T_i} + m_{T_{ij}} + (\beta + b_i + b_{ij})Z_{ijk} + \varepsilon_{T_{ijk}}, \end{aligned} \tag{1}$$

where μ_S and μ_T are fixed intercepts, m_{S_i} and m_{T_i} are random intercepts for trial i, and $m_{S_{ij}}$ and $m_{T_{ij}}$ are random intercepts for center j in trial i. The parameters α and β are fixed treatment effects, a_i and b_i are random treatment effects associated with trial and a_{ij} and b_{ij} are random treatment effects related to center. The individual-specific error terms are $\varepsilon_{S_{ijk}}$ and $\varepsilon_{T_{ijk}}$. The vector of random effects associated with trial, $(m_{S_i}, m_{T_i}, a_i, b_i)^{\mathrm{T}}$, is assumed to be zero-mean normally distributed with variance–covariance matrix

$$D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ d_{ST} & d_{TT} & d_{Ta} & d_{Tb} \\ d_{Sa} & d_{Ta} & d_{aa} & d_{ab} \\ d_{Sb} & d_{Sa} & d_{ab} & d_{bb} \end{pmatrix}. \tag{2}$$

The vector of random effects associated with center, $(m_{S_{ij}}, m_{T_{ij}}, a_{ij}, b_{ij})^{\mathrm{T}}$, is also assumed to be zero-mean normally distributed with variance–covariance matrix

$$D' = \begin{pmatrix} d'_{SS} & d'_{ST} & d'_{Sa} & d'_{Sb} \\ d'_{ST} & d'_{TT} & d'_{Ta} & d'_{Tb} \\ d'_{Sa} & d'_{Ta} & d'_{aa} & d'_{ab} \\ d'_{Sb} & d'_{Sa} & d'_{ab} & d'_{bb} \end{pmatrix}. \tag{3}$$

Finally, the individual-level error terms $(\varepsilon_{S_{ijk}}, \varepsilon_{T_{ijk}})^{\mathrm{T}}$ are also zero-mean normally distributed with variance–covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ \sigma_{ST} & \sigma_{TT} \end{pmatrix}. \tag{4}$$

Parameter estimation can be based on, for example, maximum likelihood or restricted maximum likelihood (Verbeke and Molenberghs, 2000). Several authors, cited in the introduction, have employed this strategy.

Clearly, (1) is not free from modeling assumptions. For example, one might want to entertain fixed effects rather than random effects. This will be considered in Section 5, where the second strategy would then be very appropriate. Indeed, fitting a random-effects model in such a case might lead to incorrectly attributing components of variability. Further, the joint normality of (1) implies that the regression of T_{ijk} on S_{ijk} is linear, whereas in reality a nonlinear association might apply. In practice, therefore, one may want to carefully assess the fit of the model. For the purpose of this article, model (1) is considered a versatile paradigm.

In the remainder of the section, we will briefly sketch the use of these models in surrogate marker validation. Of course, this is not relevant for the HIS example. The next step considered in the methodology proposed by Buyse et al. (2000) focused on prediction. Precisely, assuming one considers a new trial, $i=0$ say, for which data are available on the surrogate endpoint but not on the true endpoint, the goal is to predict the outcome on the true endpoint. We are interested in the estimated effect of treatment Z on true endpoint T, given the effect of Z on the surrogate S for this particular trial. Let us subscript all quantities pertaining to the particular trial under study with 0.

It is easy to show (Buyse et al., 2000) that $(\beta + b_0|m_{S0}, a_0)$ follows a normal distribution with mean and variance:

$$E(\beta + b_0|m_{S0}, a_0) = \beta + \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} \mu_{S0} - \mu_S \\ \alpha_0 - \alpha \end{pmatrix}, \tag{5}$$

$$\mathrm{Var}(\beta + b_0|m_{S0}, a_0) = d_{bb} - \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}. \tag{6}$$

Related to prediction equations (5)–(6), a measure to assess the quality of the surrogate at the trial level, as it was stated by Buyse et al. (2000), is the coefficient of determination

$$R^2_{\mathrm{trial\,(f)}} = R^2_{b_i|m_{Si},a_i} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}}. \tag{7}$$

Similarly, to measure individual-level surrogacy, Buyse et al. (2000) proposed to use the coefficient of determination given by

$$R^2_{\mathrm{indiv}} = \frac{\sigma^2_{ST}}{\sigma_{SS}\sigma_{TT}}, \tag{8}$$

where σ_{ST}, σ_{SS} and σ_{TT} are components of variance–covariance matrix (4).

As Buyse et al. (2000) posed, $R^2_{\mathrm{trial\,(f)}} = 1$ and $R^2_{\mathrm{indiv}} = 1$ indicate perfect surrogacy at the trial and individual level, respectively. Realistically, one could call a surrogate

'good' at a certain level, if the corresponding R^2 is sufficiently close to one. Of course, this is not just a statistical matter but rather a combination of statistical and substantive considerations.

Intuition can be gained by considering the simplified case where the prediction of $\beta + b_0$ is done independently of the random intercept m_{S0}. Coefficient (7) then reduces to

$$R^2_{\text{trail (r)}} = R^2_{b_i|a_i} = \frac{d^2_{ab}}{d_{aa}d_{bb}}, \tag{9}$$

which is simply the square of the correlation coefficient for a_i and b_i. This formula is useful when the full random-effects model is hard to fit but a reduced version, excluding random intercepts, is easier to reach convergence. It is simply the square of the correlation between a_i and b_i. Note that $R^2_{\text{trail (r)}} = 1$ if the trial-level treatment effects are simply multiples of each other.

In our *three-level* context, the same procedure can be followed for the center level and $R^2_{\text{center (r)}}$ and $R^2_{\text{center (f)}}$ can be computed a way similar to (7) and (9) using matrix (3), providing us with an assessment of the *center-level* surrogacy.

5. Modeling strategies

Tibaldi et al. (2003) showed that, in the two-level hierarchy, fitting random-effects model (1) can be replaced by simplified computational methods. In the remainder of this paper simplified methods will be used to face the computational challenges. In particular, we consider three strategies:

Strategy I: Two-Level Only. This pertains to the case where, in spite of the three-level data generating mechanism, we consider either the trial level or center level for analysis and for validation, but not both. The trial and center-specific effects are treated as fixed.

Strategy II: Three Levels, Fixed Effects. A model in which the full three-level structure of the data is included. Both the trial-specific and the center-specific effects are treated as fixed.

Strategy III: Three Levels, Random Effects. A model in which the full three-level structure of the data is included. Both the trial-specific and center-specific effects are treated as random.

We will now discuss each of these three strategies in turn.

5.1. Strategy I: two-level only

Here again, we will put emphasis on the surrogate marker evaluation context, the Health Interview Survey context being a simplification of this one. As stated before, the parameters of the full random-effects model (1) can be estimated by maximum likelihood or restricted maximum likelihood, using standard linear mixed model software such as the SAS procedure MIXED (Verbeke and Molenberghs, 2000).

5.1.1. Trial level only

In case we only consider the trial level for the validation process, exactly as Tibaldi et al. (2003), we can rewrite and simplify the model as

$$\begin{aligned} S_{ijk} &= \mu_{S_i} + \alpha_i Z_{ijk} + \varepsilon_{S_{ijk}}, \\ T_{ijk} &= \mu_{T_i} + \beta_i Z_{ijk} + \varepsilon_{T_{ijk}}, \end{aligned} \tag{10}$$

where μ_{S_i}, μ_{T_i}, α_i, and β_i are trial-specific intercepts and treatment effects. In addition, the univariate approach is opted for and hence errors $(\varepsilon_{S_{ijk}}, \varepsilon_{T_{ijk}})$ in (10) are assumed independent, rather than correlated. Tibaldi et al. (2003) showed that this approach is computationally advantageous, while resulting in little or no loss of efficiency when the emphasis is on the trial-level surrogacy. Of course, if one is interested in individual-level surrogacy as well, the correlation between the outcomes needs to be accounted for. At the second stage, a regression model is fitted to the treatment effects, estimated at the first stage. For example,

$$\hat{\beta}_i = \lambda_0 + \lambda_1 \hat{\mu}_{S_i} + \lambda_2 \hat{\alpha}_i + \varepsilon_i. \tag{11}$$

As Tibaldi et al. (2003) stated, this model can then be employed to assess the trial-level surrogacy, using the $R^2_{\text{trial (f)}}$ associated with the model. The coefficient is not calculated as in (7), but it rather just is the classical coefficient of determination found by regressing $\hat{\beta}_i$ on $\hat{\mu}_{S_i}$ and $\hat{\alpha}_i$.

If trial-specific intercept from the surrogate model (10) is not used, λ_1 is dropped from (11) and an $R^2_{\text{trial (r)}}$ is obtained, similar in spirit to (9).

5.1.2. Center level only

In case we only consider the center level for the validation process, and analogous to the previous case, the model can be rewritten as

$$\begin{aligned} S_{ijk} &= \mu_{S_{ij}} + \alpha_{ij} Z_{ijk} + \varepsilon_{S_{ijk}}, \\ T_{ijk} &= \mu_{T_{ij}} + \beta_{ij} Z_{ijk} + \varepsilon_{T_{ijk}}, \end{aligned} \tag{12}$$

where now $\mu_{S_{ij}}$, $\mu_{T_{ij}}$, α_{ij}, and β_{ij} are center-specific intercepts and treatment effects. As in the previous case, the models are fitted separately and the errors are assumed to be independent. At the second stage, a regression model similar to (11) is fitted to the treatment effects, obtained from the estimation at the first stage:

$$\hat{\beta}_{ij} = \lambda'_0 + \lambda'_1 \hat{\mu}_{S_{ij}} + \lambda'_2 \hat{\alpha}_{ij} + \varepsilon_{ij}. \tag{13}$$

The model can be used to assess the center-level surrogacy, using the $R^2_{\text{center (f)}}$ associated with this regression. In case that center-specific intercept from surrogate model is not used, a reduced $R^2_{\text{center (r)}}$ is obtained.

5.2. Strategy II: three levels, fixed effects

We now include both trial as well as center effects in the first-stage model, but they are considered to be fixed rather than random. The model then reads

$$S_{ijk} = \mu_{S_i} + \mu_{S_{ij}} + (\alpha_i + \alpha_{ij})Z_{ijk} + \varepsilon_{S_{ijk}},$$
$$T_{ijk} = \mu_{T_i} + \mu_{T_{ij}} + (\beta_i + \beta_{ij})Z_{ijk} + \varepsilon_{T_{ijk}}, \tag{14}$$

where both errors $(\varepsilon_{S_{ijk}}, \varepsilon_{T_{ijk}})$ are to be dependent.

At the second stage, an appropriate set of regressions is fitted to the treatment effects, estimated at the first stage:

$$\hat{\beta}_i = \lambda_0 + \lambda_1 \hat{\mu}_{S_i} + \lambda_2 \hat{\alpha}_i + \varepsilon_i, \tag{15}$$

$$\hat{\beta}_{ij} = \lambda_0' + \lambda_1' \hat{\mu}_{S_{ij}} + \lambda_2' \hat{\alpha}_{ij} + \varepsilon_{ij}. \tag{16}$$

Model (15) is used, when the trial-level association is of interest. Model (16) is used, when the focus is on the association at the center level. Both regressions produce an R^2 measure of surrogacy.

5.3. Strategy III: three levels, random effects

Buyse et al. (2000) assumed the availability of individual-patient data and formulated a two-stage model, with the joint distribution $[T, S|Z]$ specified at the first stage and the joint distribution of the treatment effects $[\beta, \alpha]$ specified at the second stage. Shkedy et al. (2003) employed this methodology and developed a Bayesian approach under the assumption that individual data are available (see also Liao, 2002; Browne et al., 2002). We will extend their methodology for model (1).

Generally, consider linear predictors for T and S:

$$E(S_{ijk}|m_{S_i}, m_{S_{ij}}, a_i, a_{ij}) = \mu_S + m_{S_i} + m_{S_{ij}} + (\alpha + a_i + a_{ij})Z_{ijk},$$
$$E(T_{ijk}|m_{T_i}, m_{T_{ij}}, b_i, b_{ij}) = \mu_T + m_{T_i} + m_{T_{ij}} + (\beta + b_i + b_{ij})Z_{ijk}. \tag{17}$$

The coefficients m_{S_i}, m_{T_i}, a_i, b_i, $m_{S_{ij}}$, $m_{T_{ij}}$, a_{ij}, b_{ij} have a similar meaning as those in model (1). Further, the vector of random effects associated to trial, $(m_{S_i}, m_{T_i}, a_i, b_i)^{\mathrm{T}}$, is assumed to be zero-mean normally distributed with covariance matrix (2), while the vector of random effects associated to center, $(m_{S_{ij}}, m_{T_{ij}}, a_{ij}, b_{ij})^{\mathrm{T}}$, is assumed to be zero-mean normally distributed with covariance matrix (3).

Shkedy et al. (2003) proposed to combine (17) and (2)–(3), defining a hierarchical Bayesian model. Thus, at the first stage of the hierarchical model we specify the following joint distribution of T_{ijk} and S_{ijk}:

$$\begin{pmatrix} S_{ijk} \\ T_{ijk} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_S + m_{S_i} + m_{S_{ij}} + (\alpha + a_i + a_{ij})Z_{ijk} \\ \mu_T + m_{T_i} + m_{T_{ij}} + (\beta + b_i + b_{ij})Z_{ijk} \end{pmatrix}, \Sigma \right), \tag{18}$$

where Σ is given by (4).

At the second stage of the model the priors for the 'fixed' effects are specified:

$$\mu_S \sim \mathrm{N}(0, \theta^2_{\mu_S}),$$
$$\mu_T \sim \mathrm{N}(0, \theta^2_{\mu_T}),$$
$$\alpha \sim \mathrm{N}(0, \tau^2_{\alpha}),$$
$$\beta \sim \mathrm{N}(0, \tau^2_{\alpha}). \tag{19}$$

For the precision parameters in (19) (flat) hyperprior models can be specified using Gamma distributions, e.g., $\theta^{-2}_{\mu_S} \sim \text{gamma}(0.001, 0.001)$, etc. As the hyperprior distribution for the covariance matrices D, D' and Σ, a Wishart distribution is assumed:

$$D^{-1} \sim \text{Wishart}(R_D), \quad D'^{-1} \sim \text{Wishart}(R_{D'}), \quad \Sigma^{-1} \sim \text{Wishart}(R_\Sigma). \tag{20}$$

In order to assess the trial-level surrogacy, the coefficient of determination defined by (7) will be used. The center-level surrogacy can be assessed using the coefficient of determination computed from (7) with matrix D', given in (3), in place of matrix D. Finally, to measure individual-level surrogacy, the coefficient of determination given in (8) can be used.

To avoid computational problems, Buyse et al. (2000) proposed a reduced model in which the linear predictors of S and T do not include trial and center specific intercepts. In the hierarchical model, the likelihood at the first stage of the model can be specified by omitting the trial specific random intercepts from (18). This leads to the specification

$$\begin{pmatrix} S_{ij} \\ T_{ij} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_S + (\alpha + a_i + a_{ij}) Z_{ijk} \\ \mu_T + (\beta + b_i + b_{ij}) Z_{ijk} \end{pmatrix}, \Sigma \right). \tag{21}$$

At the second stage of the model, the prior distribution the random effects, $(a_i, b_i)^{\mathrm{T}}$, is assumed to be bivariate normal with mean 0 and covariance matrix D_r. Note that the covariance matrix D_r is the 2×2 lower right submatrix in (2) and is assumed to follow a Wishart distribution, $D_r^{-1} \sim \text{Wishart}(R_{D_r})$. Other prior and hyperprior models remain the same as in the full model. For the reduced model, the coefficient of determination, measuring the trial-level surrogacy, reduces to (9). Similar considerations can be made for $(a_{ij}, b_{ij})^{\mathrm{T}}$, which is assumed normal with zero mean and covariance matrix D'_r, which is the 2×2 right bottom sub matrix of D' defined in (3).

6. A simulation study

We studied the performance of the various strategies in terms of both point estimation, as well as precision, of $R^2_{\text{trial (r)}}$ and of $R^2_{\text{center (r)}}$, by means of a simulation study. A setting, similar to the one used in Buyse et al. (2000) is adopted.

6.1. Simulation settings

6.1.1. Generating Mechanism I

In Mechanism I, data are generated using model (1) with $(m_{S_i}, m_{T_i}, a_i, b_i) \sim \mathrm{N}(0, D)$ and $(m_{S_{ij}}, m_{T_{ij}}, a_{ij}, b_{ij}) \sim \mathrm{N}(0, D')$, where

$$D = \sigma_T^2 \begin{pmatrix} 1 & 0.8 & 0 & 0 \\ 0.8 & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho_T \\ 0 & 0 & \rho_T & 1 \end{pmatrix}, \quad D' = \sigma_C^2 \begin{pmatrix} 1 & 0.8 & 0 & 0 \\ 0.8 & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho_C \\ 0 & 0 & \rho_C & 1 \end{pmatrix}, \tag{22}$$

and $\mu_S = 50,\ \mu_T = 45,\ \alpha = 5,\ \beta = 3$.

Further, the true R^2, following from (7) and (22), is set equal to either 0.5 or 0.9 at the trial or at the center level. Thus, for both ρ_T^2 and ρ_C^2, the values of 0.5 or 0.9 are considered. Parameters σ_T^2 and σ_C^2 are assigned values of 0.1 or 10. Regarding the individual-level variability, $(\varepsilon_{S_{ij}}, \varepsilon_{T_{ij}}) \sim \mathrm{N}(0, \Sigma)$ with

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}.$$

The parameter σ^2 equals either 0.1 or 3.

For every choice of values for ρ_T, ρ_C, σ_T^2, σ_C^2 and σ^2, simulated datasets were obtained assuming 5, 10, 20, or 100 trials, with 10 or 100 centers per trial and with 10 or 100 subjects per center. In total, 250 datasets were simulated for each setting.

6.1.2. Generating Mechanisms II and III

Further, a simulation was performed in which, instead of considering model (1) to generate the data, we used a model in which we have random effects associated to either trial or to center, but not to both of them.

The first of these, termed Mechanism II and where only trial-level random effects are considered, is given by

$$\begin{aligned} S_{ijk} &= \mu_S + m_{S_i} + (\alpha + a_i) Z_{ijk} + \varepsilon_{S_{ijk}}, \\ T_{ijk} &= \mu_T + m_{T_i} + (\beta + b_i) Z_{ijk} + \varepsilon_{T_{ijk}}. \end{aligned} \tag{23}$$

Alternatively, when only random-effects at the center level are present (Mechanism III), (1) simplifies to

$$\begin{aligned} S_{ijk} &= \mu_S + m_{S_{ij}} + (\alpha + a_{ij}) Z_{ijk} + \varepsilon_{S_{ijk}}, \\ T_{ijk} &= \mu_T + m_{T_{ij}} + (\beta + b_{ij}) Z_{ijk} + \varepsilon_{T_{ijk}}. \end{aligned} \tag{24}$$

The random vectors associated to trial and center were considered, as in Mechanism I, to follow mean-zero normal distributions: $(m_{S_i}, m_{T_i}, a_i, b_i) \sim \mathrm{N}(0, D)$, $(m_{S_{ij}}, m_{T_{ij}}, a_{ij}, b_{ij}) \sim \mathrm{N}(0, D')$.

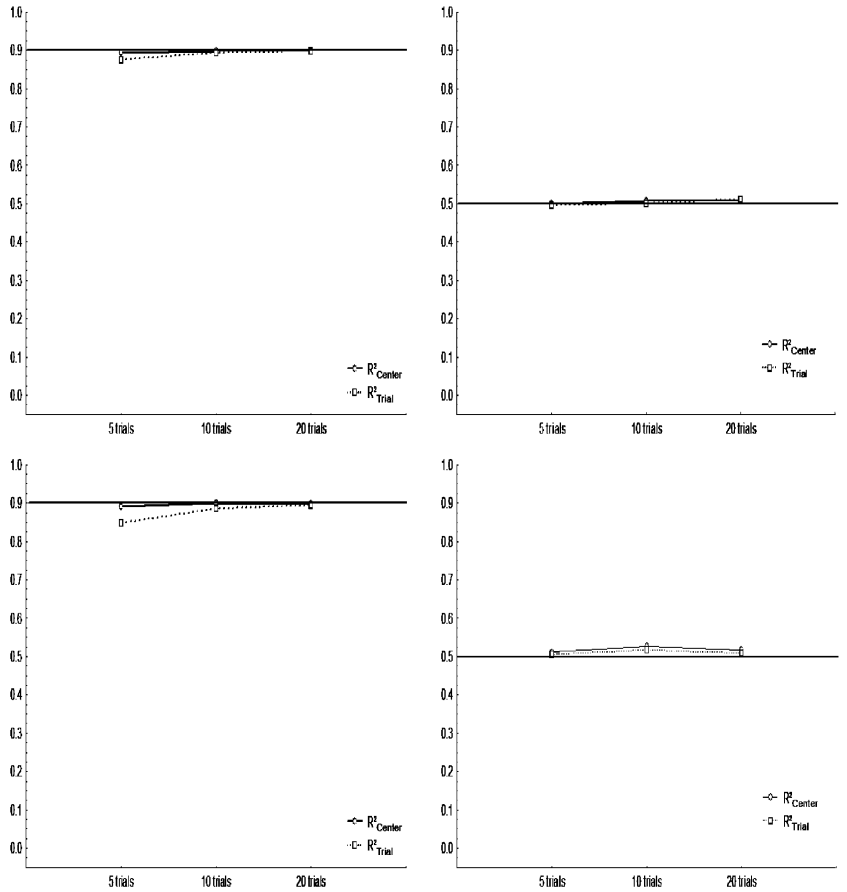

Fig. 1. Simulation study. The estimation of R^2 and its precision for $R^2_{\text{trial (r)}} = R^2_{\text{center (r)}}$ and $\sigma_T^2 = \sigma_C^2 = 10$. Data were generated using Mechanism I. Left column: Strategy I (Two-level only); right column: Strategy II (Three-levels, fixed effects). Top row: estimation of $R^2 = 0.5$; bottom row: estimation of $R^2 = 0.9$.

A setting of simulation parameters similar to the one used for Mechanism I was considered, i.e., 5, 10, 20 or 100 trials, with 10 or 100 centers and 10 or 100 subjects per center; σ_T^2 and σ_C^2 equal to 10 or 0.1; $\sigma^2 = 3$ or 0.1; $\rho_T^2 = 0.5$ or 0.9 and $\rho_C^2 = 0.5$ or 0.9.

6.2. Simulation results, equal trial- and center-level association

6.2.1. Generating Mechanism I

The results of the simulations for Mechanism I, assuming $\rho_T^2 = \rho_C^2 = 0.5$ or 0.9, $\sigma_T^2 = \sigma_C^2 = 10$, and $\sigma^2 = 3$ for 5, 10 or 20 trials with 10 centers per trial and 10 subjects per center are shown in Fig. 1. Results for other settings of the parameters are similar.

Fig. 1 shows the results obtained when Strategies I and II were used. In particular, the use of Strategy I means, that the association at the trial level was evaluated using a model without the center level (see (10) in Section 5.1.1), while the association at the center level was assessed using a model without the trial level (see (12) in Section 5.1.2).

Fig. 1 indicates that both strategies give comparable results. It can be observed that Strategy II has larger bias in the estimation than Strategy I. It is important to point out that when $\rho_T^2 = \rho_C^2 = 0.5$, both methods tend to overestimate the strength of the association, while if $\rho_T^2 = \rho_C^2 = 0.9$, the strategies underestimate it.

6.2.2. Generating Mechanisms II and III

When only one level of association is present in the data generating mechanism, we can try to estimate the effects at this particular level using Strategy I, with either the correct or the incorrect level included in the model. That is, if Mechanism II was used, which involved only the trial-level association, we could try to capture this association using center as the unit of analysis. A similar approach could be used for Mechanism III, but in this case the center-level association could be evaluated using trial as the unit of analysis. This would correspond to realistic situations where our interest lies at another level than at which data are available from. For example, the first scenario (Mechanism II with center as the unit of analysis) is of practical interest when there are too few trials available and, to assess the trial-level surrogacy, data for centers is used instead. The results for 5, 10 or 20 trials with 10 centers per trial and 10 subjects per center and $\rho_T^2 = \rho_C^2 = 0.5$ or 0.9, $\sigma_T^2 = \sigma_C^2 = 10$, $\sigma^2 = 3$ are shown in Fig. 2. Results for other settings of the parameters are similar.

From Fig. 2, it can be seen that when the data were generated using Mechanism II, the strategies proposed in Sections 5.1.1 and 5.1.2 led to very similar results. That is, the estimated strength of the (trial-level) association was similar irrespectively of whether trial (correctly) or center (incorrectly) was used as the unit of analysis. On the other hand, when Mechanism III was used to generate the data, the method described in Section 5.1.2 (i.e., using, correctly, center as the unit of analysis) performed much better than the method of Section 5.1.1. To be precise, for the analysis based on centers the estimates were closer to the true parameter. As in Section 6.2.1, when ρ_T^2 and ρ_C^2 were equal to 0.5, Strategy I tended to overestimate the strength of the association, while ρ_T^2 and ρ_C^2 were equal to 0.9, it was generally underestimated.

In addition, Strategy II was also applied to the simulated datasets. In this case, first three-level fixed-effects model (14) was fitted to the data, and then models (15) and (16) were used to compute the determination coefficients assessing the strength of association at the trial and center level, respectively. The results for 5, 10 or 20 trials with 10 centers per trial and 10 subjects per center and $\rho_T^2 = \rho_C^2 = 0.5$ or $\rho_T^2 = \rho_C^2 = 0.9$, $\sigma_T^2 = \sigma_C^2 = 10$, $\sigma^2 = 3$ are shown in Fig. 3. Results for other settings of the parameters are similar.

From Fig. 3 it is clear that, when Mechanism II was used, Strategy II with model (15) at the second stage (based on trial-specific estimates) was giving satisfactory results in terms of the bias of the estimation. On the other hand, for model (16),

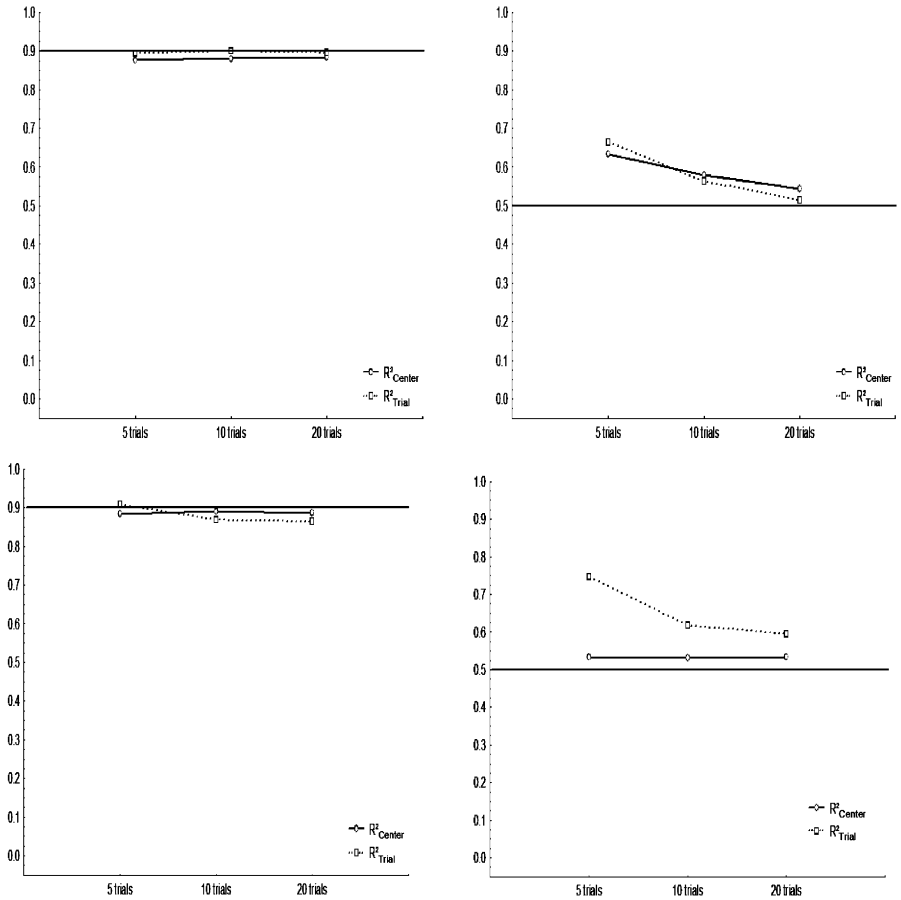

Fig. 2. Simulation study. The estimation of R^2 and its precision for Strategy I (Two-level only) when $R^2_{\text{trial (r)}} = R^2_{\text{center (r)}}$ and $\sigma^2_T = \sigma^2_C = 10$. Left column: *generating Mechanism II*; right column: *generation Mechanism III*. Top row: estimation of $R^2 = 0.5$; bottom row: estimation of $R^2 = 0.9$.

based on center-specific estimates, the results were poor. Fig. 3 also shows that when Mechanism III was used to generate the data, Strategy II gave similar results in terms of bias irrespectively of the model used at the second stage.

6.3. Simulation results, unequal trial- and center-level association

The results of simulations presented in Section 6.2 allow to conclude that both Strategy I and Strategy II performed reasonably well when the association at the trial and at the center levels were equal. In this section we present the case in which the associations at both levels differ.

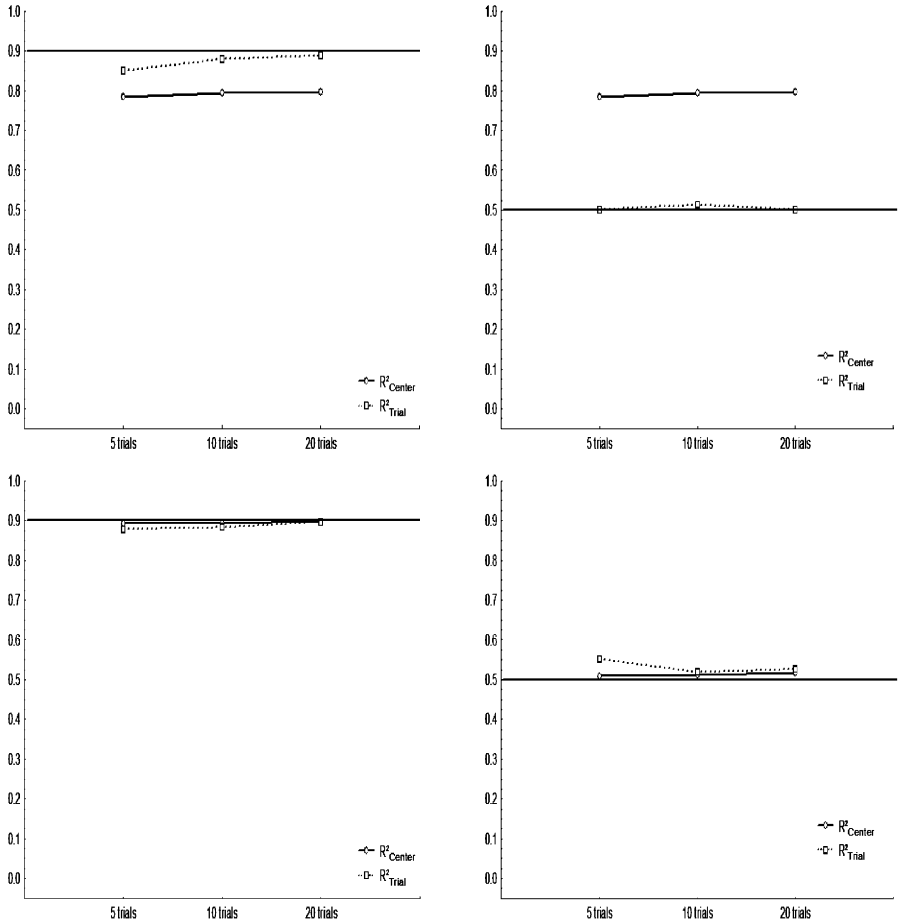

Fig. 3. Simulation study. The estimation of R^2 and its precision for Strategy II (Three-levels, fixed effects) when $R^2_{\text{trial (r)}} = R^2_{\text{center (r)}}$ and $\sigma_T^2 = \sigma_C^2 = 10$. Left column: *generating Mechanism II*; right column: *generating Mechanism III*. Top row: estimation of $R^2 = 0.5$; bottom row: estimation of $R^2 = 0.9$.

6.3.1. Performance of Strategies I and II

To further study the performance of Strategies I and II, we simulated data using Mechanism I, with $\rho_T^2 \neq \rho_C^2$. In particular, we considered $\rho_T^2 = 0.5$ with $\rho_C^2 = 0.9$ and $\rho_T^2 = 0.9$ with $\rho_C^2 = 0.5$. The values for the other parameters were similar to those used for the simulations presented in Section 6.2.1. The results for 5, 10 or 20 trials with 10 centers per trial and 10 subjects per center and $\sigma_T^2 = \sigma_C^2 = 10$ and $\sigma^2 = 3$ are shown in Table 2.

In terms of bias, the results from Table 2 are reasonable for the estimation of the trial-level association when Strategy I was applied (i.e., using trial as the unit of analysis at both stages) and of the center-level association when Strategy II was applied (i.e.,

Table 2
Simulation study

ρ_T^2	ρ_C^2	No. of trials	Modeling strategies			
			Strategy I		Strategy II	
			Trial as unit[a]	Center as unit[b]	Trial as unit[a]	Center as unit[b]
0.5	0.9	5	0.521(0.309,0.317)	0.706(0.158,0.169)	0.623(0.296,0.301)	0.891(0.050,0.054)
0.5	0.9	10	0.528(0.220,0.226)	0.700(0.116,0.121)	0.676(0.182,0.183)	0.900(0.034,0.040)
0.5	0.9	20	0.540(0.147,0.151)	0.698(0.077,0.079)	0.681(0.122,0.121)	0.898(0.025,0.027)
0.9	0.5	5	0.830(0.179,0.186)	0.655(0.113,0.118)	0.663(0.268,0.273)	0.511(0.139,0.145)
0.9	0.5	10	0.851(0.098,0.099)	0.676(0.085,0.088)	0.685(0.190,0.196)	0.527(0.119,0.122)
0.9	0.5	20	0.856(0.064,0.065)	0.681(0.059,0.058)	0.686(0.123,0.124)	0.518(0.092,0.093)

Results for Strategies I and II for $\sigma_T^2 = \sigma_C^2 = 10$, with 10 patients per center and 10 centers per trial. Mean estimates of ρ_T^2 and ρ_C^2 with model-based and empirical standard errors (in parentheses).
[a]Gives estimates of ρ_T^2.
[b]Gives estimates of ρ_C^2.

Table 3
Simulation study

ρ_T^2	ρ_C^2	No. of trials	Modeling strategies			
			Strategy I		Strategy II	
			Trial as unit[a]	Center as unit[b]	Trial as unit[a]	Center as unit[b]
0.5	0.9	5	0.535(0.305,0.315)	0.537(0.294,0.312)	0.526(0.312,0.320)	0.819(0.075,0.079)
0.5	0.9	10	0.504(0.228,0.235)	0.516(0.220,0.231)	0.508(0.231,0.238)	0.822(0.060,0.062)
0.5	0.9	20	0.507(0.151,0.157)	0.519(0.145,0.153)	0.513(0.156,0.161)	0.822(0.044,0.045)
0.9	0.5	5	0.894(0.122,0.131)	0.880(0.123,0.134)	0.870(0.151,0.154)	0.722(0.109,0.111)
0.9	0.5	10	0.891(0.094,0.102)	0.884(0.087,0.092)	0.882(0.088,0.090)	0.730(0.087,0.089)
0.9	0.5	20	0.897(0.043,0.047)	0.890(0.042,0.046)	0.888(0.048,0.047)	0.731(0.068,0.070)

Results for Strategies I and II for $\sigma_T^2 = 10$ and $\sigma_C^2 = 0.1$ and 10 patients per center and 10 centers per trial. Mean estimates of ρ_T^2 and ρ_C^2 with model-based and empirical standard errors (in parentheses).
[a]Gives estimates of ρ_T^2.
[b]Gives estimates of ρ_C^2.

a three-level fixed-effects model at the first stage with center-specific effects analyzed at the second stage).

The above conclusions were drawn for the case when $\sigma_T^2 = \sigma_C^2 = 10$. It is also of interest to study what would happen if σ_C^2 were much smaller than σ_T^2. From a practical point of view this situation is desirable, since a large variance for the center level means existence of a strong center-specific treatment effect, what makes difficult to draw general conclusions. Table 3 presents results for Strategies I and II for the case of $\sigma_T^2 = 10$ and $\sigma_C^2 = 0.1$.

Table 3 indicates that, when the variability at the center level was much smaller than at the trial level, the estimates obtained using either Strategy I or Strategy II for the trial-level association were close to the true value of the parameter of interest. On the other hand, for the center-level association, reasonable results were obtained only for Strategy II when $\rho_C^2 = 0.9$. For other cases using center as the unit of the analysis, either at both stages (Strategy I) or only at the second one (Strategy II), was producing results that, on average, were close to the value of the coefficient of determination related to the trial-level association.

6.3.2. Insights in the performance of Strategy I

The bad performance of Strategy I, especially for the center level, can be explained by the fact that ignoring a level can lead to overestimation of the variability at the levels surrounding the level being ignored. To this aim, we will use the results obtained by Hutchison and Healy (2001). For example, consider the following model:

$$S_{ijk} = \mu_S + m_{S_i} + m_{S_{ij}} + (\alpha + \alpha_i + \alpha_{ij})Z_{ijk} + \varepsilon_{S_{ijk}}.$$

It is similar to model (1), but it contains only three random effects: random intercepts m_{S_i} and $m_{S_{ij}}$ associated to trial and center, respectively, and the random error $\varepsilon_{S_{ijk}}$. Assume that the data are balanced ($N_i \equiv N, n_{ij} \equiv n$) and the variances of the random effects corresponding to the trial, center and individual level are equal to σ_T^2, σ_C^2 and σ^2, respectively. It can be then shown that the two variance components of the model in which the center level is ignored are:

$$\tilde{\sigma}_T^2 = \sigma_T^2 + \frac{n-1}{Nn-1}\,\sigma_C^2 \approx \sigma_T^2 + \frac{1}{N}\,\sigma_C^2, \tag{25}$$

$$\tilde{\sigma}^2 = \sigma^2 + \frac{n(N-1)}{Nn-1}\,\sigma_C^2 \approx \sigma^2 + \frac{N-1}{N}\,\sigma_C^2. \tag{26}$$

Thus, they can be seen as the true variance, plus a certain fraction of the variance of the random effect associated to the level that has been ignored. For this particular case not much variability is added to the variance corresponding to the level above the one ignored (trial), since most of the information is sent to the level below. This is the reason why in Tables 2 and 3 the trial-level association is generally well estimated when Strategy I is used. On the other hand, if the trial level is ignored, the center-level variance becomes

$$\tilde{\sigma}_C^2 = \sigma_C^2 + \frac{N(M-1)}{MN-1}\,\sigma_T^2 \approx \sigma_C^2 + \frac{M-1}{M}\,\sigma_T^2. \tag{27}$$

The individual-level variability remains unchanged. Thus, most of the variability contained in the trial level is sent to the center level, which affects the estimation of the association at the center level. This is the reason why in Tables 2 and 3 the center-level association is poorly estimated when Strategy I is used.

6.3.3. Performance of strategy II in a large dataset

In order to further explore the behavior of Strategy II observed in Tables 2 and 3, an additional simulation was conducted. Table 4 shows results for several different combinations of the values of parameters σ_T^2, σ_C^2, σ^2, ρ_T^2, and ρ_C^2, for 100 trials with

Table 4
Simulation study

σ_T^2	σ_C^2	σ^2	ρ_T^2	ρ_C^2	Trial as unit[a]	Center as unit[b]
10	10	3	0.5	0.9	0.685(0.030,0.033)	0.900(0.004,0.009)
10	10	3	0.9	0.5	0.684(0.031,0.035)	0.501(0.014,0.021)
10	10	0.1	0.5	0.9	0.685(0.030,0.033)	0.900(0.004,0.009)
10	10	0.1	0.9	0.5	0.683(0.031,0.035)	0.499(0.014,0.020)
10	0.1	3	0.5	0.9	0.508(0.028,0.030)	0.877(0.010,0.012)
10	0.1	3	0.9	0.5	0.896(0.010,0.013)	0.565(0.024,0.027)
10	0.1	0.1	0.5	0.9	0.506(0.028,0.031)	0.899(0.005,0.007)
10	0.1	0.1	0.9	0.5	0.896(0.010,0.013)	0.503(0.015,0.017)
0.1	0.1	0.1	0.5	0.9	0.686(0.030,0.032)	0.899(0.005,0.008)
0.1	0.1	0.1	0.9	0.5	0.684(0.031,0.033)	0.503(0.015,0.0017)

Results for Strategy II for different values of variance components associated to trial and center random effects, with 100 subjects per center, 100 centers per trial and 100 trials. Mean estimates of ρ_T^2 and ρ_C^2 with model-based and empirical standard errors (in parentheses).

[a]Gives estimates of ρ_T^2.

[b]Gives estimates of ρ_C^2.

100 centers per trial and 100 subjects per center. The idea is to investigate the behavior of the strategy in a large dataset.

Results presented in Table 4 indicate that, the center-level association was in general estimated reasonably well. It is worth noting that the bias, observed in Table 3 for the combination of $\rho_T^2 = 0.9$ and $\rho_C^2 = 0.5$, was greatly reduced when $\sigma^2 = 3$, and essentially disappeared when $\sigma^2 = 0.1$. This suggests that, for Strategy II, the bias in the estimation of the center-level surrogacy may be negligible as long as the variability at the level of center is at least as large as the variability at the lower (individual) level.

On the other hand, from Table 4 one can see that, when the variability at the trial and center level was of the same magnitude, the trial-level association was estimated poorly, even though the sizes of the units were large. The bias generally disappeared when the variability at the center level became much smaller than that at the level of trial. This suggests that, as for the center-level association, bias in the assessment of the trial-level association for Strategy II may be negligible as long as the variability at the lower (center) level is smaller.

6.3.4. Comparison of strategies II and III

Finally, we attempted to compare Strategy II with Strategy III. Since using a maximum-likelihood approach to implement Strategy III was numerically too complex, we considered the use of a Bayesian approach. Unfortunately, performing an extensive simulation using the latter approach turned out to be too time-consuming. Therefore, the simulation study was limited to the random generation of only one dataset for different parameter settings, and the comparison of the results obtained for Strategy II

Table 5
Simulation study

σ_C	No. of trials	ρ_T^2	ρ_C^2	R^2	Actual value	Strategy II	Strategy III		
							Mean	StDev	Median
10	5	0.5	0.9	R^2_{trail}	0.750	0.840	0.653	0.2420	0.7127
10	5	0.5	0.9	R^2_{center}	0.927	0.914	0.934	0.0210	0.9381
10	5	0.9	0.5	R^2_{trail}	0.916	0.822	0.917	0.0856	0.9430
10	5	0.9	0.5	R^2_{center}	0.443	0.539	0.497	0.1079	0.5012
10	10	0.5	0.9	R^2_{trail}	0.263	0.501	0.260	0.2128	0.2234
10	10	0.5	0.9	R^2_{center}	0.930	0.951	0.929	0.0154	0.9311
10	10	0.9	0.5	R^2_{trail}	0.872	0.725	0.826	0.1207	0.8572
10	10	0.9	0.5	R^2_{center}	0.431	0.454	0.399	0.0837	0.3999
10	20	0.5	0.9	R^2_{trail}	0.358	0.719	0.425	0.1697	0.4329
10	20	0.5	0.9	R^2_{center}	0.912	0.938	0.901	0.0153	0.9018
10	20	0.9	0.5	R^2_{trail}	0.915	0.747	0.894	0.0667	0.9109
10	20	0.9	0.5	R^2_{center}	0.502	0.557	0.524	0.0532	0.5250
0.1	5	0.5	0.9	R^2_{trail}	0.777	0.760	0.777	0.1871	0.8355
0.1	5	0.5	0.9	R^2_{center}	0.914	0.810	0.907	0.1112	0.9482
0.1	5	0.9	0.5	R^2_{trail}	0.941	0.948	0.960	0.0504	0.9751
0.1	5	0.9	0.5	R^2_{center}	0.533	0.635	0.534	0.1889	0.5572
0.1	10	0.5	0.9	R^2_{trail}	0.444	0.421	0.447	0.2169	0.4628
0.1	10	0.5	0.9	R^2_{center}	0.932	0.760	0.892	0.1082	0.9265
0.1	10	0.9	0.5	R^2_{trail}	0.795	0.776	0.792	0.1276	0.8217
0.1	10	0.9	0.5	R^2_{center}	0.488	0.551	0.494	0.1513	0.5054
0.1	20	0.5	0.9	R^2_{trail}	0.292	0.288	0.311	0.1652	0.3063
0.1	20	0.5	0.9	R^2_{center}	0.915	0.819	0.951	0.0468	0.9668
0.1	20	0.9	0.5	R^2_{trail}	0.950	0.933	0.952	0.0250	0.9574
0.1	20	0.9	0.5	R^2_{center}	0.466	0.691	0.465	0.1325	0.4698

Results for Strategy II and Strategy III for a simulated sets of data with 10 subjects per center and 10 centers per trial and $\sigma_T = 10$ (Mean=posterior mean; StDev=posterior standard deviation; Median=posterior median).

and Strategy III to the true values of the parameters used for simulations. The results are shown in Table 5.

By comparing the estimates of the coefficients of determination to their actual values (i.e., the values computed from the actual, simulated random effects) in Table 5 we can observe that, when the variability at the center and trial level was of the same magnitude, Strategy II did not estimate the trial-level association well, in contrary to Strategy III. Even when the variability at the center level was smaller than that at the level of trial, estimates obtained for Strategy III were closer to the actual values than the estimates produced by Strategy II. One can conclude, admittedly based on the anecdotal evidence obtained by generating a single dataset under each setting, that

Strategy III gives better results, which is reasonable if we take into account that the other two strategies are ignoring levels and are using fixed effects as representation of random effects.

7. Analysis of case studies

7.1. Studies in schizophrenia

In the first psychiatric study, several options were studied considering the units available. The first row in Table 6 shows the results obtained when Strategy I was applied. That is, the coefficient of determination associated to a particular level was estimated using a (two-stage) model including only this level and individual variability. We observe that, in general, there is relatively little difference between the estimates obtained.

Strategy II, using a fixed-effects model with all three levels included at the first stage, was fitted as well. In this model the estimate of the magnitude of the association at the highest level (country) is close to that obtained using Strategy I. For the other two levels more substantial differences can be observed.

Finally, a random-effects (Strategy III) analysis, based on the Bayesian approach, was performed. The results are shown in the last row of Table 6. One can see that the estimates of the magnitude of the association for the two highest levels (main investigator and country) are lower than those obtained for the two other strategies. As for Strategy I, there is relatively little difference in the estimates obtained for different levels.

Let us turn attention to the second psychiatric case study, where data from an equivalence trial are used. The result for the investigator level ($R^2 = 0.70$, bootstrap-based 95% C.I. [0.44; 0.96]), obtained using Strategy I, is within the range of the estimates observed for the first study (see Table 6). This observation supports the claim that might have been able to reasonably accurately quantify the surrogacy of PANSS for CGI in the context of certain compounds for schizophrenia. Of course, the R^2 values are not terribly high, so that a mere replacement of CGI by PANSS may be questionable.

Table 6
R^2 values (with 95% confidence/credible intervals) at different levels for the first psychiatric study, using different modelling strategies

	Unit of analysis		
	Investigator (138 units)	Main investigator (29 units)	Country (19 units)
Strategy I	0.56[0.43; 0.68][a]	0.69[0.41; 0.86][a]	0.62[0.25, 0.88][a]
Strategy II	0.42[0.30, 0.55][a]	0.77[0.49, 0.89][a]	0.56[0.15; 0.86][a]
Strategy III	0.52[0.24, 0.74][b]	0.66[0.31, 0.88][b]	0.51[0.11; 0.83][b]

[a]Bootstrap confidence interval.
[b]Credible set.

7.2. Belgian Health Interview Survey

We focus on log-transformed body mass index (BMI) as a normally distributed outcome. In an attempt to find a parsimonious model for these data, the following covariates were examined: sex, age (eight categories), education (five categories), household income (5 categories), and smoking behavior. Note that the question about smoking behavior was addressed only to persons aged 15 or more, thus reducing the effective sample size from 10,221 to 8560. In addition to the aforementioned covariates, information about the sample design can be taken into consideration, such as: stratification variables (quarter and provinces); size variables (province, municipality, household); number of groups to be interviewed within a municipality; and interviewee status (indicating whether he/she is the reference person or his/her partner).

Due to unit- and item non-response, 7422 out of 8560 (87%) observations were available with complete information on the selected covariates and BMI. For illustration purpose the analysis will be restricted to the province of Limburg, where 324 observations were available for the modelling procedure. The following model was fitted to the data:

$$y_{ijk} = x_{ijk}^{\mathrm{T}}\boldsymbol{\beta} + a_i + a_{ij} + \varepsilon_{ijk}, \tag{28}$$

with y_{ijk} being the log of BMI for subject k in household j from municipality i, $a_i \sim \mathrm{N}(0, \sigma^2_{\mathrm{MUN}})$, $a_{ij} \sim \mathrm{N}(0, \sigma^2_{\mathrm{HH}})$, and $\varepsilon_{ijk} \sim \mathrm{N}(0, \sigma^2)$. Thus, the total variation in (log) BMI can be decomposed into that between individuals within each household (σ^2), that between households within municipalities (σ^2_{HH}), and that between municipalities (σ^2_{MUN}). Among covariates listed above, only sex, age, education, and smoking behavior were found to have a significant effect and were included in the model. Second-order interaction terms of sex with age and smoking behavior, education with smoking behavior and age, and smoking behavior with age were also included. Among sampling-related variables, only interviewee status was retained.

The variance components σ^2_{MUN}, σ^2_{HH}, and σ^2 can be interpreted in terms of intra-unit correlation. Thus, the apparent intra-municipality correlation is defined as

$$\rho_{\mathrm{MUN}} = \frac{\sigma^2_{\mathrm{MUN}}}{\sigma^2_{\mathrm{MUN}} + \sigma^2_{\mathrm{HH}} + \sigma^2}, \tag{29}$$

while the intra-household correlation is equal to

$$\rho_{\mathrm{HH}} = \frac{\sigma^2_{\mathrm{MUN}} + \sigma^2_{\mathrm{HH}}}{\sigma^2_{\mathrm{MUN}} + \sigma^2_{\mathrm{HH}} + \sigma^2}. \tag{30}$$

The intra-unit correlation reflects the proportion of the total variability in the outcome variable that is attributable to the clustering effect at a certain level and, as such, is a measure of within-group homogeneity.

Table 7 shows the estimated variance and intra-class correlation coefficients associated with the household and municipality levels, obtained for model (28) fitted to the data using different modelling approaches.

Table 7
Belgian Health Interview Survey

Strategies		Endpoint		
		Fixed	Random	Bayesian
Ignoring	$\hat{\sigma}^2_{HH}$	0.0225(0.0030)[a]	0.0036(0.0014)[b]	0.0037(0.0019)[d]
municipality	$\hat{\sigma}^2$	0.0156(0.0014)[a]	0.0131(0.0014)[b]	0.0167(0.0020)[d]
level	$\hat{\rho}_{HH}$	0.5905(NA)	0.2156(0.0775)[c]	0.1821(0.0854)[d]
Ignoring	$\hat{\sigma}^2_{MUN}$	0.0011(0.0003)[a]	0.0003(0.0004)[b]	0.0009(0.0010)[d]
household	$\hat{\sigma}^2$	0.0189(0.0015)[a]	0.0164(0.0013)[b]	0.0198(0.0017)[d]
level	$\hat{\rho}_{MUN}$	0.0550(NA)	0.0199(0.0248)[c]	0.0409(0.0396)[d]
	$\hat{\sigma}^2_{MUN}$	0.0385(0.0187)[a]	0.0001(0.0004)[b]	0.0007(0.0009)[d]
Considering	$\hat{\sigma}^2_{HH}$	0.0535(0.0070)[a]	0.0035(0.0015)[b]	0.0033(0.0018)[d]
both	$\hat{\sigma}^2$	0.0156(0.0015)[a]	0.0131(0.0014)[b]	0.0168(0.0020)[d]
levels	$\hat{\rho}_{MUN}$	0.3578(NA)	0.0052(0.0221)[c]	0.0348(0.0361)[d]
	$\hat{\rho}_{HH}$	0.8550(NA)	0.2162(0.0778)[c]	0.1913(0.0858)[d]

Multilevel linear regression model on log(*BMI*). Estimated variance for the random effects for different modelling strategies.

[a]Standard errors were calculated using bootstrap.

[b]Likelihood-based standard errors.

[c]Standard errors were calculated using the delta method.

[d]Posterior standard errors.

We will first concentrate on the fixed-effects approach, i.e., treating random-effects as fixed. The variance component corresponding to the random effects at a particular level is computed as the sample variance of the estimated fixed-effects at that level. The results obtained for models with one of the levels ignored show a strong disagreement with those obtained for the model with both levels included. For example, the intra-household correlation is estimated to equal 0.855 and 0.590 when municipality level is excluded or included (in addition to the household level) in the model, respectively. Clearly, such a difference can lead to completely different conclusions.

For a likelihood-based random-effects model estimation approach (such as, for example, the one implemented in the SAS procedure MIXED), the estimated variance components are smaller (column "Random" in Table 7). No big differences are observed in the estimated intra-municipality and intra-household correlation when both levels are included or when one of them is ignored.

Using reasoning similar to the one leading to Eqs. (25)–(26), one can show that if we ignore the household level, the municipality and the residual variances for a balanced design become:

$$\tilde{\sigma}^2_{MUN} = \sigma^2_{MUN} + \frac{N-1}{N_{HH}N-1}\,\sigma^2_{HH},$$

$$\tilde{\sigma}^2 = \sigma^2 + \frac{N(N_{HH}-1)}{N_{HH}N-1}\,\sigma^2_{HH},$$

where N is the number of subjects per household and N_{HH} is the number of households.

If we now replace σ^2_{MUN} by $\tilde{\sigma}^2_{\text{MUN}}$ and σ^2 by $\tilde{\sigma}^2$ in (29), the intra-municipality correlation becomes

$$\tilde{\rho}_{\text{MUN}} = \frac{\sigma^2_{\text{MUN}} + \frac{N-1}{N_{\text{HH}}N-1}\,\sigma^2_{\text{HH}}}{\sigma^2_{\text{MUN}} + \frac{N-1}{N_{\text{HH}}N-1}\,\sigma^2_{\text{HH}} + \sigma^2 + \frac{N(N_{\text{HH}}-1)}{N_{\text{HH}}N-1}\,\sigma^2_{\text{HH}}} = \frac{\sigma^2_{\text{MUN}} + \frac{N-1}{N_{\text{HH}}N-1}\,\sigma^2_{\text{HH}}}{\sigma^2_{\text{MUN}} + \sigma^2_{\text{HH}} + \sigma^2}.$$

This result indicates that if $N > 1$, the resulting intra-municipality correlation will be larger than the true one. This is exactly what we observe: the intra-municipality correlation is bigger in the case where we ignore the household level.

The same can be done for the model in which municipality is ignored. In this case the residual variance remains the same, but the household variance becomes equal to

$$\tilde{\sigma}^2_{\text{HH}} = \sigma^2_{\text{HH}} + \frac{N_{\text{HH}}(N_{\text{MUN}} - 1)}{N_{\text{MUN}}N_{\text{HH}} - 1}\,\sigma^2_{\text{MUN}}, \tag{31}$$

where N_{MUN} is the number of municipalities.

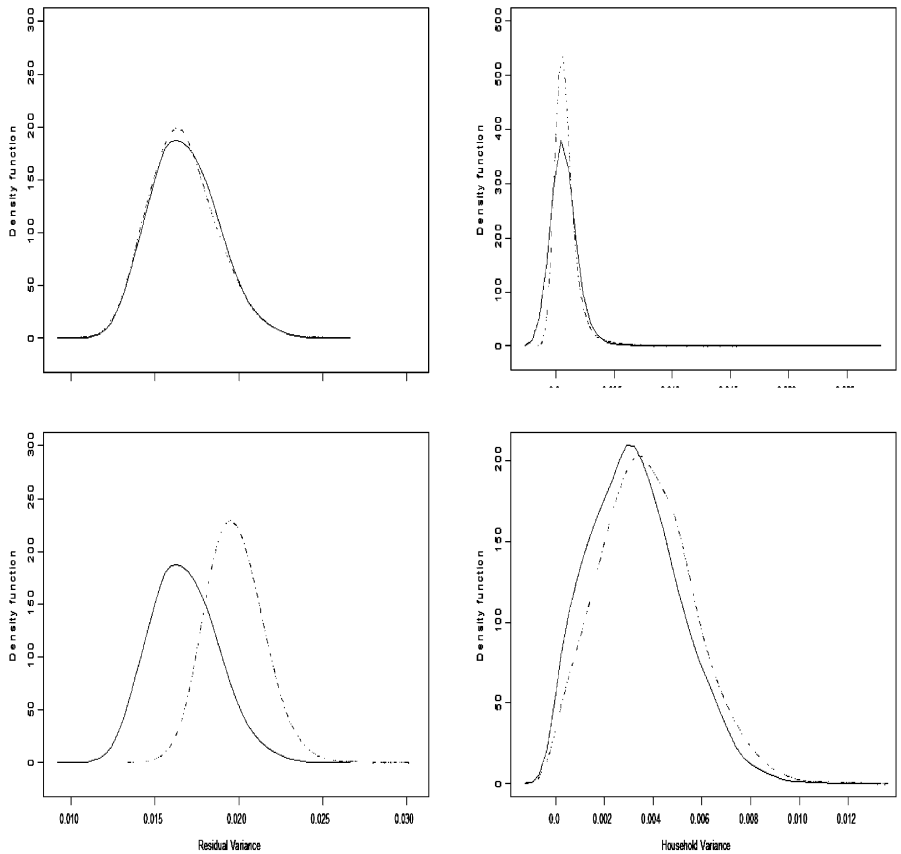

Fig. 4. Belgian Health Interview Survey. Density function for variance components for the full model (solid line) and for models when one of the levels is ignored (dashed line). Left column: *household level was ignored*; right column: *municipality level was ignored.*

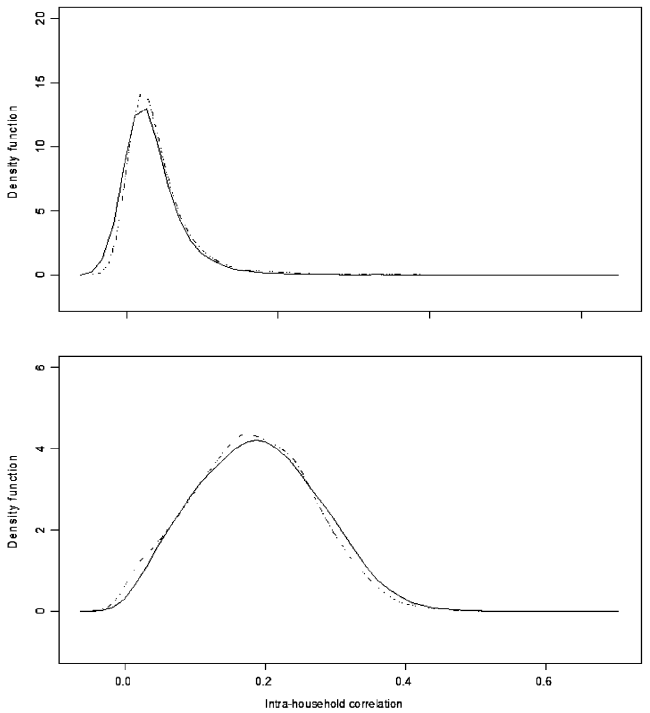

Fig. 5. Belgian Health Interview Survey. Density function for correlation coefficients for the full model (solid line) and for models one of the levels ignored (dash line). Upper panel: *household level was ignored*; Lower panel: *municipality level was ignored.*

If we replace σ^2_{HH} by $\tilde{\sigma}^2_{\mathrm{HH}}$ in (30), we get

$$\tilde{\rho}_{\mathrm{HH}} = \frac{\frac{N_{\mathrm{HH}}(N_{\mathrm{MUN}}-1)}{N_{\mathrm{MUN}}N_{\mathrm{HH}}-1}\,\sigma^2_{\mathrm{MUN}} + \sigma^2_{\mathrm{HH}}}{\frac{N_{\mathrm{HH}}(N_{\mathrm{MUN}}-1)}{N_{\mathrm{MUN}}N_{\mathrm{HH}}-1}\,\sigma^2_{\mathrm{MUN}} + \sigma^2_{\mathrm{HH}} + \sigma^2} = \frac{\left(1+\frac{N_{\mathrm{MUN}}-1+N^{-1}_{\mathrm{HH}}}{N_{\mathrm{MUN}}-N^{-1}_{\mathrm{HH}}}\right)\sigma^2_{\mathrm{MUN}} + \sigma^2_{\mathrm{HH}}}{\left(1+\frac{N_{\mathrm{MUN}}-1+N^{-1}_{\mathrm{HH}}}{N_{\mathrm{MUN}}-N^{-1}_{\mathrm{HH}}}\right)\sigma^2_{\mathrm{MUN}} + \sigma^2_{\mathrm{HH}} + \sigma^2}.$$

It follows that

$$\tilde{\rho}_{\mathrm{HH}} < \frac{\sigma^2_{\mathrm{MUN}} + \sigma^2_{\mathrm{HH}}}{\sigma^2_{\mathrm{MUN}} + \sigma^2_{\mathrm{HH}} + \sigma^2}.$$

Again, this is in accordance with our observation: the intra-household correlation is smaller when municipality is ignored. In our application the difference is negligible, though, due to the fact that the variability at the municipality level is very small.

Finally, similar results for the intra-class correlation coefficients can be observed for the random-effects Bayesian approach (column "Bayesian" in Table 7). In this case the results can be illustrated using the (posterior) density functions of the estimated variance components. From Fig. 4, we can observe that, in general, when a level is ignored the densities are shifted to the right, except for the case of the residual variance when municipality was ignored. When the household level is ignored the shift to the

right is bigger, and it is due to the fact that the household variance is six times larger than the municipality variance. In Fig. 5 the (posterior) densities for the correlations are plotted. It can be seen that there is almost no impact of the exclusion of a level in the estimation model on the density functions. This is due to the fact that the municipality variance is six times smaller than the household variation.

8. Concluding remarks

In this paper, we have investigated several strategies to deal with hierarchical linear models. We have been interested primarily in the estimation of the strength of the association between random effects at different levels. This interest has been motivated in the context of validating surrogate markers.

Three different strategies have been considered in the paper: (1) applying fixed-effects models with only the trial level or the center level used in the validation process (Strategy I); (2) including both levels in a fixed-effects model at the first stage (Strategy II); and (3) including both levels in a random-effects model at the first stage (Strategy III). The strategies differ in the complexity of the models. Consequently, they also differ in the ease of their practical implementation.

In general terms, the results indicate that the performance of the strategies depends on the sample sizes, as well as on the variablity present at different levels. The latter dependence, especially for Strategy I, can be explained using theoretical results on the effect of ignoring levels when fitting multi-level models presented in a recent article by Hutchison and Healy (2001).

In particular, from the conducted simulations we could conclude that, when data were generated according to a model with random effects present at both levels, and when the strength of association between the random effects was the same at both levels, all the strategies produced reasonable results. When the association was different, Strategy I, with trials as the units of analysis, produced satisfactory estimates of the trial-level association. On the other hand, using centers as the units of analysis resulted in biased estimates of the center-level association. The estimates were, in fact, close to the true value of the measure of the strength of the trial-level association, when the variability of center-specific random effects was smaller than the variability of trial-specific effects. This observation gives some justification to the use of, e.g., centers instead of trials as the units of analysis in practical applications of the meta-analytic approach to the validation of surrogate endpoints.

On the other hand, to obtain plausible estimates of the strength of the association at a particular level for Strategy II, the variability at the level below the one of interest had to be smaller.

A limited investigation of the performance of Strategy III suggested that it was able to correctly identify different sources of variability and association. The estimates obtained under Strategy III were closer to the actual values than, e.g., those for Strategy II. In view of the structure of the model used in Strategy III, these conclusions were not surprising. However, an important problem associated with the practical use of this strategy is its numerical complexity. From this point of view, a possibility to use, e.g., Strategy I might be very advantageous.

Acknowledgements

We gratefully acknowledge support from FWO-Vlaanderen Research Project "Sensitivity Analysis for Incomplete and Coarse Data" and Belgian IUAP/PAI network "Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data".

References

Alonso, A., Geys, H., Molenberghs, G., Vangeneugden, T., 2002. Investigating the criterion validity of psychiatric symptom scales using surrogate marker validation methodology. J. Biopharma. Statist. 12, 161–179.

Browne, W.J., Draper, D., Goldstein, H., Rasbash, J., 2002. Bayesian and likelihood methods for fitting multilevel models with complex level-1 variation. Comput. Statist. Data Anal. 39, 203–225.

Burzykowski, T., Molenberghs, G., Buyse, M., Geys, H., Renard, D., 2001. Validation of surrogate endpoints in multiple randomized clinical trials with failure-time endpoints. J. Roy. Statist. Soc. C (Appl. Statist.) 50, 405–422.

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

Concordet, D., Nunez, O.G., 2002. A simulated pseudo-maximum likelihood estimator for nonlinear mixed models. Comput. Statist. Data Anal. 39, 187–201.

Ellenberg, S.S., Hamilton, J.M., 1989. Surrogate endpoints in clinical trials: cancer. Statist. Med. 8, 405–413.

Goldstein, H., 1995. Multilevel Statistical Models. In: Kendall's Libary of Statistics, Vol. 3. Arnold, London.

Hutchison, D., Healy, M., 2001. The effect of variance component estimates of ignoring a level in a multilevel model. Multilevel Modelling Newsletter 13, 4–5.

Kay, S.R., Opler, L.A., Lindenmayer, J.P., 1988. Reliability and validity of the positive and negative syndrome scale of shizophrenics. Psychiatry Res. 23, 99–110.

Liao, T.F., 2002. Bayesian model comparison in generalized linear models across multiple groups. Comput. Statist. Data Anal. 39, 311–327.

Nair, N.P.V., the Risperidone Study Group, 1998. Therapeutic equivalence of risperidone given once daily of twice daily in patients with scizophrenia. J. Clin. Psychopharmacol. 18, 103–110.

Prentice, R.L., 1989. Surrogate endpoints in clinical trials: definitions and operational criteria. Statist. Med. 8, 431–440.

Searle, S.R., Casella, G., McCulloch, C.E., 1992. Variance Components. Wiley, New York.

Shkedy, Z., Torres Barbosa, F., Burzykowski, T., Molenberghs, G., 2003. A hierarchical bayesian approach for the evaluation of surrogate endpoints in multiple randomized clinical trials, in preparation.

Tibaldi, F.S., Cortinas Abrahantes, J., Molenberghs, G., Renard, D., Burzykowski, T., Buyse, M., Parmar, M., Stijnen, T., Wolfinger, R., 2003. Simplified hierarchical linear models for the evaluation of surrogate endpoints. J. Statist. Comput. Simul. 73, 643–658.

Verbeke, G., Molenberghs, G., 2000. Linear Mixed Model for Longitudinal Data. Springer, New York.

Xiang, L., Tse, S.-K., Lee, A.H., 2002. Influence diagnostics for generalized linear mixed models: applications to clustered data. Comput. Statist. Data Anal. 40, 759–774.