Simplified hierarchical linear models for the evaluation of surrogate endpoints

DOI: 10.1080/0094965031000062177
Handle: http://hdl.handle.net/1942/421

Simplified Hierarchical Linear Models for the Evaluation of Surrogate Endpoints

Fabián S. Tibaldi, José Cortiñas Abrahantes, Geert Molenberghs,
Didier Renard, and Tomasz Burzykowski
Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus,
B3590 Diepenbeek, Belgium

Marc Buyse
International Drug Development Institute, 430 avenue Louise, B1050 Brussels

Max Parmar
Cambridge University, Cambridge, United Kingdom

Theo Stijnen
Erasmus Universiteit, Rotterdam, the Netherlands

Russ Wolfinger
SAS Institute, Campus Drive, Cary, North Carolina

Abstract

The linear mixed-effects model (Verbeke and Molenberghs 2000) has become a standard tool for the analysis of continuous hierarchical data such as, for example, repeated measures or data from meta-analyses. However, in certain situations the model does pose insurmountable computational problems. Precisely this has been the experience of Buyse *et al.* (2000a) who proposed an estimation- and prediction-based approach for evaluating surrogate endpoints. Their approach requires fitting linear mixed models to data from several clinical trials. In doing so, these authors built on the earlier, single-trial based, work by Prentice (1989), Freedman *et al.* (1992), and Buyse and Molenberghs (1998). While Buyse *et al.* (2000a) claim their approach has a number of advantages over the classical single-trial methods, a solution needs to be found for the computational complexity of the corresponding linear mixed model. In this paper, we propose and study a number of possible simplifications. This is done by means of a simulation study and by applying the various strategies to data from three clinical studies: Pharmacological Therapy for Macular Degeneration Study Group (1977), Ovarian Cancer Meta-analysis Project (1991) and Corfu-A Study Group (1995).

Some Keywords: Linear mixed model; Macular degeneration; Meta-analytic approach; Oncology; Random effects; Surrogate endpoint.

1 Introduction

Prentice (1989) and Freedman *et al.* (1992) laid the foundations for the evaluation of surrogate endpoints in randomized clinical studies. Precisely, Prentice proposed a definition as well as a set of operational criteria. Freedman *et al.* (1992) supplemented these criteria with a quantity called *proportion explained* (PE). Buyse and Molenberghs (1998) proposed to use the *relative effect* (RE), linking the effect of treatment on both endpoints and an individual-level measure of agreement between both endpoints, after adjusting for the effect of treatment (*adjusted association*), instead of the PE. The adjusted association carries over when data are available on several randomized trials, while the RE can be extended to a trial-level measure of agreement between the effects of treatment of both endpoints. As observed by Molenberghs *et al.* (2002) and Alonso *et al.* (2002) there are serious issues surrounding the Prentice-Freedman framework. Let us briefly expand on this. It has been asserted that the criteria set out by Prentice are too stringent (Fleming *et al.* 1996) and neither necessary nor sufficient for his definition to be fulfilled, except in the special case of binary outcomes (Buyse and Molenberghs). In addition, Freedman, Graubard and Schatzkin Freedman, (1992) showed that these criteria were not straightforward to verify through statistical hypothesis tests. Therefore the PE was suggested but it is surrounded with difficulties, the most dramatic one being that it is not confined to the unit interval (Buyse *et al.*, 2000a). Buyse *et al.* (2000a) argued that some fundamental criticisms towards the process of statistical validation can be overcome by combining 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. In doing so, they carry over the relative effect and adjusted association to a trial-level R^2 and an individual-level R^2, respectively. Similar routes of meta-analytic thinking have been followed by Daniels and Hughes (1997) and Gail *et al.* (2000).

A thorough account on problems related to the Prentice–Freedman framework is given in Molenberghs *et al.* (2002). Of course, the switch to a meta-analytic problem does not solve all problems, surrounding surrogate marker validation, in a definitive way. First, one has to carefully reflect upon the question as to how broad the class of units, to be included in a validation study, can be. Clearly, the issue disappears when the same or similar treatments are considered across units (e.g., in multi-center or multi-investigator studies, or when data are used from a family of related study such as in a single drug development line). In a more loosely connected, meta-analytic setting it is important to ensure that treatment assignments are logically consistent. This is possible, for example, when the same standard treatment is compared to members of a class of experimental therapies.

While the previous issue is relevant, this paper is devoted to a different, very important, computationally-oriented issue. 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.* (2000a) 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. The thrust of their findings is that, when the between-trial variability is sufficiently large, little or no convergence problems occur except when the number of trials is very small.

Given the general importance of linear mixed models, going well beyond the surrogate marker validation case, it is necessary to study convergence properties in more detail, and to contrast the general linear mixed model, such as the one proposed by Buyse *et al.* (2000a), with alternative and/or simplified strategies. A number of such alternative strategies are proposed here and studied in terms of their statistical and numerical properties. To this end, a simulation study is considered, and the various methods are applied to the data studied in Buyse *et al.* (2000a).

The meta-analytic setting, to be used throughout the paper, is introduced in Section 2. The simplified approaches, organized along three "dimensions", are presented in Section 3. Sections 4–6 are devoted to each of the three dimensions in turn. Case studies are introduced and analyzed in Section 7 and a simulation study is reported in Section 8.

2 Setting

As stated earlier, we will focus on normally distributed endpoints. Let us introduce a set of notation that will be used throughout the paper. Let T_{ij} and S_{ij} be random variables denoting the true and the surrogate endpoints for subject $j = 1, \dots n_i$ in trial $i = 1, \dots N$. Further, let Z_{ij} denote a binary treatment indicator.

The full random-effects model, as introduced by Buyse *et al.* (2000a) is

$$
\begin{aligned}
S_{ij} &= \mu_S + m_{S_i} + \alpha Z_{ij} + a_i Z_{ij} + \varepsilon_{S_{ij}}, &(1)\\
T_{ij} &= \mu_T + m_{T_i} + \beta Z_{ij} + b_i Z_{ij} + \varepsilon_{T_{ij}}, &(2)
\end{aligned}
$$

where μ_S and μ_T are fixed intercepts, m_{S_i} and m_{T_i} are random intercepts for trial i, α and β are fixed treatment effects and a_i and b_i are random treatment effects. The individual-specific error terms are $\varepsilon_{S_{ij}}$ and $\varepsilon_{T_{ij}}$.

The vector of random effects, $(m_{S_i}, m_{T_i}, a_i, b_i)'$, is assumed to be zero-mean normally distributed with covariance matrix

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

The individual-level error terms $(\varepsilon_{S_{ij}}, \varepsilon_{T_{ij}})'$ are also zero-mean normally distributed with covariance matrix

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

Parameter estimation can be based on, for example, maximum likelihood or restricted maximum likelihood (Verbeke and Molenberghs, 2000). Next, suppose we consider a new trial, $i = 0$ say, for which data are available on the surrogate endpoint but not on the true endpoint. We are interested in the estimated effect of Z on T, given the effect of Z on S for this particular trial. Subscript all quantities pertaining to the particular trial under study with 0. It is easy to show (Buyse *et al.* 2000a) that $(\beta + b_0 | m_{S0}, a_0)$ follows a normal distribution with mean and variance:

$$
\begin{aligned}
E(\beta + b_0 | m_{S0}, a_0) &= \beta + \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}' \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} \mu_{S0} - \mu_S \\ \alpha_0 - \alpha \end{pmatrix}, &(3)\\
\mathrm{Var}(\beta + b_0 | m_{S0}, a_0) &= d_{bb} - \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}' \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}. &(4)
\end{aligned}
$$

Related to prediction equations (3)–(4), a measure to assess the quality of the surrogate at the trial level is the coefficient of determination

$$
R^2_{\text{trial (f)}} = R^2_{b_i | m_{Si}, a_i} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}' \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}}. \tag{5}
$$

A good surrogate, *at the trial level*, would have (5) close to 1. Intuition can be gained by considering the simplified case where the prediction of b_0 is done independently of the random intercept m_{S0}. The coefficient (5) then reduces to

$$R^2_{\text{trial }(r)} = R^2_{b_i|a_i} = \frac{d_{ab}^2}{d_{aa}d_{bb}}. \tag{6}$$

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 α_i and β_i. Note that $R^2_{\text{trial }(r)} = 1$ if the trial level treatment effects are simply multiples of each other.

3 Simplified Modelling Strategies

Buyse *et al.* (2000a) showed that fitting random-effects model (1)–(2) can be a surprisingly difficult task in a number of situations. This is particularly true when the number of trials or the number of patients per trial is small. Also, situations with extreme correlations pose problems. It is therefore imperative to explore approximate strategies with better computational properties. These authors studied one alternative approach in the sense that they replaced the random effects by their fixed-effect counterparts. Such a two-stage approach is very similar in spirit to the original proposal of Laird and Ware (1982). We will now embed this ad-hoc strategy in a more formally developed system of model simplifications.

Precisely, we consider three dimensions along which simplifications can be made:

Trial dimension: whether the trial-specific effects are treated as either random or fixed. A full random-effects is then distinguished from a two-stage approach.

Endpoint dimension: whether the surrogate and true endpoints are modelled as a bivariate outcome or two univariate ones. In the latter case the correlation between both endpoints is not incorporated into the modeling strategy, rendering the study of the individual-level surrogacy more involved. However, as stated earlier, throughout this paper the focus is on trial-level surrogacy.

Measurement error dimension: whenever the full random-effects model is abandoned, one is confronted with measurement error since the treatment effects in the various trials are estimated with error. The magnitude of this error is likely to depend on several characteristics, such as trial size, which will vary across trials. We consider three ways to account for measurement error: unadjusted (i.e., no correction at all), adjustment by trial size, and an approach suggested by T. Stijnen and explained in Section 5.

The combination of these three dimensions are graphically represented in Figure 1 and gives rise to twelve strategies. However, some do not have to be considered. For example, when one chooses for a bivariate (endpoint dimension) random-effects (trial dimension) approach, measurement error is automatically accounted for, whence explicit corrections are no longer needed. In the special case when sample size is constant across trials, further simplifications arise (see Section 8).

FIGURE 1, ABOUT HERE.

We will now discuss each of the three simplifying dimensions in turn.

4 The Trial Dimension

As stated before, the parameters of the full random-effects model (1)–(2) can be estimated by maximum likelihood or restricted maximum likelihood, using standard linear mixed model software such as the SAS procedure MIXED.

In case we treat the trial-level parameters as fixed, exactly as Buyse *et al.* (2000a), we can rewrite the model as

$$\begin{aligned} S_{ij} &= \mu_{S_i} + \alpha_i Z_{ij} + \varepsilon_{S_{ij}}, & (7)\\ T_{ij} &= \mu_{T_i} + \beta_i Z_{ij} + \varepsilon_{T_{ij}}, & (8) \end{aligned}$$

where μ_{S_i}, μ_{T_i}, α_i, and β_i are trial-specific intercepts and treatment effects. The assumption about the error terms depends on the choice made on the *endpoint dimension* (Section 6). Indeed, when the univariate approach is opted for, both errors are assumed independent. Otherwise, a bivariate unstructured covariance matrix is considered.

At the second stage, a regression model is fitted to the treatment effects, estimated at the first stage, for example:

$$\widehat{\beta}_i = \lambda_0 + \lambda_1 \widehat{\mu}_{S_i} + \lambda_2 \widehat{\alpha}_i + \varepsilon_i. \tag{9}$$

This model can then be employed to assess trial-level surrogacy, using the$R^2_{\text{trial (f)}}$ associated with this regression. Precisely, this is not calculated as in (5), but is merely the classical coefficient of determination found by regressing $\widehat{\beta}_i$ on $\widehat{\mu}_{S_i}$ and $\widehat{\alpha}_i$.

In case the trial-specific intercept from surrogate model (7) is not used, λ_1 would be dropped and an $R^2_{\text{trial (r)}}$ is obtained, similar in spirit to (6).

5 The Measurement Error Dimension

Recall that this dimension is irrelevant when the full random-effects model is assumed, but is crucial when a fixed-effects approach is selected on the *trial dimension* and/or when a univariate model is chosen on the *endpoint dimension*.

We allow for three possible choices. First, a simple linear model can be assumed to determine the relationship between β_i, α_i, and μ_{S_i}, whereby the errors in (9) are assumed to be zero-mean normally distributed with constant variance σ^2.

Clearly, this approach ignores the fact that the estimated treatment effects α_i and β_i will typically come from trials with large variations in size. One way to address this issue is by weighing the contributions according to trial size, resulting in a weighted linear regression. Such an approach may account for some but not all of the heterogeneity in information content between trial-specific contributions. A nice way to overcome this is T. Stijnen's approach.

To this end, we introduce models for the estimated trial-specific treatment effects $(\widehat{\mu}_{S_i}, \widehat{\alpha}_i, \widehat{\beta}_i)'$, given the true trial-specific treatment effects $(\mu_{S_i}, \alpha_i, \beta_i)'$:

$$\begin{pmatrix} \widehat{\mu}_{S_i} \\ \widehat{\alpha}_i \\ \widehat{\beta}_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{S_i} \\ \alpha_i \\ \beta_i \end{pmatrix}, C_i\right). \tag{10}$$

Here, C_i is the variance-covariance matrix of the estimated treatment effects. In case we assume both treatment-effect estimates to be independent (which would result from a univariate choice on the *endpoint dimension*), C_i would be assumed to be diagonal, even though this may be unrealistic.

Further, we assume a normal model for the true trial-specific treatment effects around the true overall treatment effects:

$$\begin{pmatrix} \mu_{S_i} \\ \alpha_i \\ \beta_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_S \\ \alpha \\ \beta \end{pmatrix}, \Sigma \right). \tag{11}$$

The resulting marginal model, combining (10) and (11), is:

$$\begin{pmatrix} \widehat{\mu}_{S_i} \\ \widehat{\alpha}_i \\ \widehat{\beta}_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_S \\ \alpha \\ \beta \end{pmatrix}, \Sigma + C_i \right). \tag{12}$$

Maximum likelihood estimation for this model can be quite easily carried out by using mixed model software, provided the values for C_i can be input and held fixed, as is the case in the SAS procedure MIXED. An example program is provided in the Appendix.

6 Endpoint Dimension

It seems natural to assume both endpoints to be correlated. However, this assumption will almost always complicate modelling and corresponding parameter estimation. In addition, the bivariate nature of the outcome is related for the better part with individual-level surrogacy whereas our main goal is trial-level surrogacy. This suggests an additional simplification, i.e., by considering separate, independent models for each of the endpoints. It then remains to be seen inhowfar such a simplification hampers estimation of trial-level surrogacy.

We need to make a distinction between two cases, according to the corresponding choice on the *trial dimension*. In the random-effects approach, this simplification would lead to a pair of *univariate* hierarchical models, one for each endpoint. In the fixed-effects approach, one would fit a separate linear regression model per endpoint and per trial. It is easy to show that the parameter estimates as well as the estimated variances are identical to the ones obtained from fitting a fixed-effects *bivariate* model to each trial separately. This follows from standard multivariate normal theory (Johnson and Wichern 1992).

7 Case Studies

We consider three case studies. Since they were considered by Buyse *et al.* (2000), we are able to compare their results with those obtained from a full set of computational approaches. Further, they cover three important but different therapeutic areas. Finally, by considering three case studies, we avoid the risk of running into results that are interesting but too specialized to a particular situation.

The first one, the Age Related Macular Degeneration Study, is an ophtalmologic study. The other two are from advanced colorectal and advanced ovarian cancer. These examples have been studied in Buyse *et al.* (2000a, 2000b). We will compare their results to the ones from the simplified approaches proposed in this paper. Results are summarized in Table 1, following the three dimensions of Figure 1. The focus is on trial-level surrogacy, captured by R^2_{trial}. While, of course, the individual-level surrogacy is of interest when the focus is on predicting a particular patient's behavior and, in some contexts, can even be of primary interest (Alonso *et al.* 2001), it is fair to say that the clinical trialist will primarily be interest in this quantity. Further, since the inclusion of the individual-level surrogacy forces the models to have a bivariate nature, the study thereof comes at a computational cost.

In addition, we distinguish between "full" models where the trial level surrogacy $R^2_{\text{trial (f)}}$ is calculated as in (5), and "reduced" models, where no random intercepts are included and hence $R^2_{\text{trial (r)}}$ as in (6) is used. Combining all possibilities on three dimensions and furthermore distinguishing between full and reduced models would, in principle, lead to 24 different approaches. However, the three bivariate random-effects approaches coincide. The columns for the full approaches are numbered for reference in the simulation study (Section 8).

Table 1, About Here.

7.1 Age Related Macular Degeneration Study (ARMD)

These data arose from a randomized clinical trial comparing an experimental treatment (interferon-α) to placebo in the treatment of patients with age-related macular degeneration. The aim of the study was to compare placebo and the highest dose of interferon-α. The treatment indicator is

$Z_{ij} = 1$ for treatment and 0 for placebo. Since we have a single multi-centric trial, i refers to *center* and j to patient within center. The true endpoint in this study was the change in visual acuity at 12 months after starting the treatment. The surrogate endpoint considered is visual acuity at 6 months. Results from assessing the surrogate in terms of the Prentice-Freedman framework were reported in Buyse *et al.* (2000a) and are not repeated here.

Buyse *et al.* (2000a) experienced problems in fitting the full random-effects models, irrespective of whether standard statistical software or user-developed alternatives were used. Therefore, they entertained a (unweighted) fixed-effects approach instead. This produced a moderate trial-level surrogacy: $R^2_{\text{trial (f)}} = 0.692$ (*s.e.* 0.087). The standard error has been calculated by means of a straightforward application of the delta method. Let us now compare their result to the ones obtained from the approaches described in Section 3.

As mentioned earlier, for the fixed-effects approaches, univariate and bivariate results values are equal. Of course, the univariate approach prohibits the assessment of individual-level surrogacy but, as mentioned earlier, in many trials the main interest is on trial-level surrogacy.

For the $R^2_{\text{trial (f)}}$, Stijnen's approach is more difficult to fit in the sense that the random-effects values cannot be obtained.

The reduced-model values are generally higher than the full-model values, suggesting that the trial-specific intercept terms for the surrogate model does convey information and, if possible, full models should be used. Within the reduced-model approach, Stijnen's univariate random-effects approach yields a low value. This is in line with intuition, since it corrects for measurement error present in the estimated treatment effects. Simulations will have to weigh costs and benefits from this approach. In general computational terms, a choice for univariate models and/or fixed-effects approaches is less expensive.

7.2 Advanced Colorectal Cancer

We consider data from two randomized multicenter trials in colorectal cancer. These constitute the largest source of randomized data available in advanced colorectal cancer. All data were collected and checked by the Meta-Analysis Group In Cancer between 1990 and 1996 (Corfu-A Group, 1995; Greco *et al.* 1996) to confirm the benefits of experimental fluoropyrimidine treatments with 5-fluorouracil (5FU) in advanced colorectal cancer. The principal investigators of all trials provided data for every patient, whether eligible or not, and whether properly followed-up or not. Previous publications provide full details on the trials included the treatments tested, the patient characteristics, and the therapeutic results (Burzykowski *et al.* 2001).

In this example, we will use $Z_{ij} = 0$ to denote 5FU plus interferon and for 5FU alone. The final endpoint T_{ij} will be survival time in years. The surrogate endpoint S_{ij} will be progression-free survival time, i.e., the years between the randomization to clinical progression of the disease or death. In agreement with previous anlyses, only centers with at least 3 patients on each treatment arm are considered. The data include 48 centers, with a total sample size of 642 patients.

Using the bivariate unweighted fixed-effects approach model proposed by Buyse *et al.* (2000a) we obtain $R^2_{\text{trial (f)}} = 0.473$ (*s.e.* 0.108), which is, of course, too low to be useful.

Results of fitting the various approaches and reported in Table 1 largely confirm the results from the ARMD study in terms of ease of convergence for the univariate and/or fixed-effects approaches. All coefficients are relatively close to each other, although the reduced versions tend to be a bit higher than the full versions.

7.3 Advanced Ovarian Cancer

These data arose from a meta-analysis of ovarian cancer (Ovarian Cancer Meta-Analysis Project, 1991). The comparison of two treatments was the principal aim of this study. We use $Z_{ij} = 0$ when cyclosphosphamide was applied and $Z_{ij} = 1$ when cyclosphosphamide plus cisplatin was applied. We considered survival time in years as final endpoint T_{ij}. The surrogate endpoint S_{ij} is progression-free survival time. We used center as the unit of analysis given that the number of trials is insufficient to applied meta-analytic methods. The number of patients distributed over a total of 50 units varies from 2 to 254.

The bivariate fixed-effects approach used by Buyse *et al.* (2000a) produces $R^2_{\text{trial (f)}} = 0.917$ (*s.e.* 0.017), which is much higher than in the colorectal cancer case. Arguably, this is due to the relatively short time span that typically elapses between both endpoints. The difference between this result and those from the other approaches is even smaller than in the other two case studies. Further, the relative computational complexity, suggested by the other case studies, is confirmed here as well.

8 A Simulation Study

We studied performance of the various approaches, in terms of estimation (point and interval) of $R^2_{\text{trial (f)}}$, and in terms of convergence through a simulation study. To make our results comparable with those from Buyse *et al.* (2000a), the same configuration setting is adopted.

Precisely, model (1)–(2) is considered with $(m_{S_i}, m_{T_i}, a_i, b_i) \sim N(0, D)$, $\mu_S = 50$, $\mu_T = 45$, $m_{S_i} = 5$, $m_{T_i} = 3$,

$$D = \sigma^2 \begin{pmatrix} 1 & 0.8 & 0 & 0 \\ 0.8 & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho \\ 0 & 0 & \rho & 1 \end{pmatrix}, \tag{13}$$

with $\rho^2 = 0.5$ or $\rho^2 = 0.9$, and $(\varepsilon_{S_{ij}}, \varepsilon_{T_{ij}}) \sim N(0, \Sigma)$ with

$$\Sigma = 3 \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}.$$

The parameter σ^2 was chosen to be either 3 or 10. Five hundred runs were completed for every setting, consisting of 25 trials each. The true R^2, following from (5) and (13) is set equal to either 0.5 or 0.9.

Results are presented in Tables 2–3. In all settings, convergence was 100%, which is slightly different from the analysis of the examples.

Tables 2–3, About Here.

Stijnen's approach exhibits a small amount of bias. In case $R^2 = 0.9$ and $\sigma^2 = 3$, there is a hint of underestimation in column 3, 6, and somehow also 9. The situation is more dramatic in the case of $R^2 = 0.5$, where indeed we observe now overestimation in all but one columns, the exception being the full model (columns 10–12).

9 Concluding Remarks

In this paper, we have investigated several strategies to deal with the computational burden posed by using 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 approach) and (3) the method of dealing with measurement error.

As a result of this, we recommend simplified computational methods for two main reasons. First, they are generally faster and easier to implement with standard software. Second, we showed, through simulations, that the simplified approaches often perform almost as good as the more advanced methods, and moreover enjoy much better convergence properties. In particular, opting for a fixed-effects approach over a full random-effects approach is very beneficial since there is at most a minor loss in statistical efficiency, the method has extremely good convergence properties, and is usually more than 10 times faster than the full approach.

We re-analyzed the three case studies considered by Buyse *et al.* (2000), from three therapeutic areas: ophtalmology, advanced colorectal cancer, and advanced ovarian cancer. In agreement with the simulation study, the fixed-effects approaches have good convergence properties, but there are problems with the random-effects approaches. In particular, none of the fully bivariate random-effects models converged, while there were also problems with their univariate and/or reduced counterparts. While there are twelve versions of each fixed-effects approach, the results are generally very similar across these, except that there is a noticeable but not a dramatic difference between the full and reduced versions. Therefore, it is recommendable to use the full model version since, in doing so, full information is used towards estimation of the trial-level surrogacy.

Appendix

```
/* First stage: bivariate fixed-effects model */

proc mixed data=mydata method=reml;
  class trial subj endpoint;
  model outcome=endpoint*trial endpoint*trial*treat
                / noint s covb ddfm=bw;
  repeated endpoint / subject=subj type=un r rcorr;
  make ’SolutionF’ out=effects;
  make ’CovParms’ out=covparms;
  make ’covb’ out=covar;
run;

/*
** Assembling trial-specific covariance matrices of estimated
** fixed effects. There is one line per trial, each such line
** corresponding to a matrix.
*/

data cov0;
set covar;
 drop _row_ _effect_ trial endpoint;
run;

proc iml;
 use cov0;
 ntrial=25;
 read all into tempdat;
 dummy=j(ntrial,7,0);
 do i=1 to ntrial;
   dummy[i,1]=tempdat[2*i-1,2*ntrial+(2*i-1)];
   dummy[i,2]=i;
   dummy[i,3]=tempdat[2*i-1,2*ntrial+2*i];
   dummy[i,4]=tempdat[2*ntrial+2*i,2*ntrial+(2*i-1)];
   dummy[i,5]=tempdat[2*i-1,2*i-1];
   dummy[i,6]=tempdat[2*ntrial+(2*i-1),2*ntrial+(2*i-1)];
   dummy[i,7]=tempdat[2*ntrial+2*i,2*ntrial+2*i];
 end;
 nms={"cmsal","trial","cmsbe","calbe","varms","varal","varbe"};
 create cova0 from dummy [colname=nms];
 append from dummy;
quit;

data effects;
set effects;
 keep _EFFECT_ _EST_ _se_ trial endpoint order int surro main;
 int=0;
 surro=0;
 main=0;
 if _effect_=’TRIAL*ENDPOINT’ then do;
      if endpoint=1 then delete;
      if endpoint=0 then do;
```

```
         order=3;
         int=1;
      end;
 end;
 if _effect_='TREAT*TRIAL*ENDPOINT' then do;
      if endpoint=0 then do;
         order=1;
         surro=1;
      end;
      if endpoint=1 then do;
         order=2;
         main=1;
      end;
 end;
run;

proc sort data=effects;
 by trial order;
run;

data stijnen;
set effects;
 drop _est_;
 est=_est_;
run;

proc sort data=stijnen;
 by trial order;
run;

data row1;
set cova0;
 keep row col value trial;
 col=trial;
 row=1;
 value=varal;
run;
...

data row6;
set cova0;
 keep row col value trial;
 col=trial;
 row=6;
 value=varms;
run;

data matrix;
 set row1 row2 row3 row4 row5 row6;
run;

proc sort data=matrix;
 by col row;
run;
```

```
/* Second stage: Stijnen's regression */

proc mixed data=stijnen order=data method=reml asycov scoring=2;
  class trial order;
  model est = order / solution noint ddfm=bw;
  random order / subject=trial group=trial type=un gdata=matrix;
  repeated order / subject=trial type=un;
  make 'CovParms' out=covparms noprint;
  make 'AsyCov' out=asycov noprint;
run;
```

Acknowledgments

The first, second, fourth, and fifth authors wish to thank "Bijzonder Onderzoeksfonds" of the Limburgs Universitair Centrum.

References

Alonso, A., Geys, H., Molenberghs, G., and Vangeneugden, T. (2001). Investigating the criterion validity of psychiatric symptom scales using surrogate marker validation methodology. *Submitted for publication.*

Arnold, B.C. and Strauss, D. (1991). Pseudolikelihood estimation: some examples. *Sankhya B* **53**, 233–243.

Boissel, J.P., Collet, J.P., Moleur, P., and Haugh M. (1992). Surrogate endpoints: a basis for a rational approach. *European Journal of Clinical Pharmacology* **43**, 235–244.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Society* **88**, 9–25.

Burzykowski, T., Molenberghs, G, and Buyse, M. (2001). The validation of surrogate endpoints using data from randomized clinical trials: A case study in advanced colorectal cancer. *Submitted for publication.*

Burzykowski, T., Molenberghs, G., Buyse, B., Geys, H., and Renard, D. (2001). Validation of surrogate endpoints in multiple randomized clinical trials with failure-time endpoints. *Applied Statistic* **50**, 405–422.

Buyse, M. and Molenberghs, G. (1998). The validation of surrogate endpoints in randomized experiments. *Biometrics* **54**, 186–201.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. (2000a). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* **1**, 1–19.

Buyse, M., Thirion, P., Carlson, R.W., Burzykowski, T., Molenberghs, G., and Piedbois, P. (2000b). Tumour response to first line chemotherapy improves the survival of patients with advanced colorectal cancer. *Lancet*, **356**, 373–378.

Chuang-Stein C. and DeMasi R. (1998). Surrogate endpoints in AIDS drug development: current status (with discussion). *Drug Information Journal* **32**, 439–459.

Corfu-A Study Group (1995). Phase III randomized study of two fluorouracil combinations with either interferon alfa-2a or leucovorin for advanced colorectal cancer. *Journal of Clinical Oncology* **13**, 921–928.

Daniels, M.J. and Hughes, M.D. (1997). Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine*, **16**, 1965–1982.

De Gruttola, V., Fleming, T.R., Lin, D.Y. and Coombs, R. (1997). Validating surrogate markers - are we being naive ? *Journal of Infectious Diseases* **175**, 237–246.

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

Flandre, P. and Saidi, Y. (1999). Letters to the editor: Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in Medicine* **18**, 107–115.

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

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

Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data, *Biometrika* **73**, 43–56.

Goldstein, H. (1995). *Multilevel Statistical Models*, 2nd edition. London: Edward Arnold.

Goldstein, H. and Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society A* **159**, 505-513.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M. *et al.* (1998). A User's Guide to MLwiN. London: Institute of Education.

Greco, F.A., Figlin, R., York, M., Einhorn, L., Schilsky, R., Marshall, E.M., *et al.* (1996). Phase III randomized study to compare interferon alfa-2a in combination with fluorouracil versus fluorouracil alone in patients with advanced colorectal cancer. *Journal of Clinical Oncology* **14**, 2674–2681.

Johnson, R.A. and Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis.* 3rd ed. Englewood Cliffs: Prentice-Hall.

Le Cessie, S. and Van Houwelingen, J.C. (1994). Logistic regression for correlated binary data. *Applied Statistics* **43**, 95-108.

Lehmann, E.L. (1983). *Theory of Point Estimation.* Wiley: New-York.

Molenberghs, G., Buyse, M., Geys, H., Renard, D., Burzykowski, T., and Alonso, A. (2002) Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Controlled Clinical Trials*, **00**, 000–000.

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

Rodríguez, G. and Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the American Statistical Society* **158**, 73-89.

Van Houwelingen, J.C., Arends, L.A., Stijnen, T. (2001). Advanced methods for meta-analysis. *Statistics in Medicine*, to appear.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Model for Longitudinal Data.* New York: Springer-Verlag.

Wolfinger, R., and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* **48**, 233–243.

Table 1: Results of the trial-level surrogacy analysis for the three examples R^2_{trial} (a $-$ symbol indicates non-convergence).

	Full Model					
	Univariate Approach					
	Fixed-effects approach			Random-effects approach		
	Unweighted	Weighted	Stijnen	Unweighted	Weighted	Stijnen
Study	1	2	3	4	5	6
ARMD	0.692	0.693	0.689	0.664	0.801	-
Colorectal	0.473	0.488	0.466	-	-	-
Ovarian	0.939	0.917	0.937	0.911	0.905	-
	Bivariate Approach					
	Fixed-effects approach			Random-effects approach		
	Unweighted	Weighted	Stijnen			
Study	7	8	9	10–12		
ARMD	0.692	0.693	0.698	-		
Colorectal	0.473	0.488	0.472	-		
Ovarian	0.939	0.917	0.938	-		
	Reduced Model					
	Univariate Approach					
	Fixed-effects approach			Random-effects approach		
Study	Unweighted	Weighted	Stijnen	Unweighted	Weighted	Stijnen
ARMD	0.776	0.758	0.775	0.659	0.786	0.623
Colorectal	0.527	0.497	0.596	-	-	-
Ovarian	0.928	0.909	0.925	0.911	0.905	0.900
	Bivariate Approach					
	Fixed-effects approach			Random-effects approach		
Study	Unweighted	Weighted	Stijnen			
ARMD	0.776	0.758	0.719	-		
Colorectal	0.527	0.497	0.471	-		
Ovarian	0.928	0.909	0.938	0.951		

Table 2: Means of the estimated trial-level surrogacy and 95% simulation-based confidence intervals for $R^2 = 0.90$. Column numbers refer to the columns of Table 1.

# Sub	1, 2, 7, 8	3	4, 5	6	9	10–12
			Variance 10			
50	0.898 (0.894;0.902)	0.895 (0.890;0.900)	0.898 (0.895;0.902)	0.894 (0.890;0.898)	0.898 (0.894;0.902)	0.896 (0.892;0.900)
60	0.900 (0.897;0.904)	0.899 (0.896;0.903)	0.901 (0.897;0.904)	0.897 (0.893;0.900)	0.900 (0.896;0.903)	0.897 (0.894;0.901)
70	0.898 (0.894;0.902)	0.896 (0.892;0.901)	0.898 (0.894;0.902)	0.894 (0.890;0.899)	0.897 (0.893;0.902)	0.895 (0.891;0.900)
80	0.899 (0.895;0.903)	0.898 (0.894;0.902)	0.899 (0.895;0.903)	0.895 (0.891;0.899)	0.898 (0.894;0.902)	0.896 (0.892;0.900)
90	0.900 (0.896;0.903)	0.899 (0.895;0.902)	0.900 (0.896;0.903)	0.896 (0.892;0.899)	0.899 (0.896;0.903)	0.897 (0.893;0.901)
100	0.901 (0.898;0.905)	0.901 (0.897;0.904)	0.901 (0.898;0.905)	0.897 (0.894;0.901)	0.901 (0.897;0.904)	0.898 (0.895;0.902)
			Variance 3			
50	0.893 (0.889;0.897)	0.889 (0.885;0.894)	0.894 (0.890;0.898)	0.892 (0.888;0.896)	0.892 (0.888;0.896)	0.896 (0.891;0.900)
60	0.896 (0.893;0.900)	0.893 (0.889;0.897)	0.897 (0.893;0.901)	0.896 (0.892;0.899)	0.895 (0.892;0.899)	0.897 (0.893;0.901)
70	0.894 (0.890;0.898)	0.890 (0.886;0.895)	0.894 (0.890;0.898)	0.891 (0.887;0.896)	0.893 (0.889;0.897)	0.895 (0.890;0.899)
80	0.895 (0.891;0.899)	0.892 (0.888;0.896)	0.896 (0.892;0.900)	0.894 (0.890;0.898)	0.895 (0.891;0.899)	0.896 (0.892;0.900)
90	0.897 (0.893;0.900)	0.894 (0.890;0.898)	0.897 (0.894;0.901)	0.893 (0.889;0.897)	0.896 (0.893;0.900)	0.897 (0.893;0.901)
100	0.898 (0.895;0.902)	0.896 (0.892;0.899)	0.899 (0.895;0.902)	0.895 (0.891;0.899)	0.898 (0.894;0.901)	0.898 (0.894;0.902)

Table 3: Means of the estimated trial-level surrogacy and 95% simulation-based confidence intervals for $R^2 = 0.50$. Column numbers refer to the columns of Table 1.

# Sub	1, 2, 7, 8	3	4, 5	6	9	10–12
			Variance 10			
50	0.527 (0.515;0.539)	0.526 (0.514;0.538)	0.528 (0.516;0.540)	0.523 (0.511;0.535)	0.526 (0.514;0.538)	0.498 (0.485;0.510)
60	0.532 (0.520;0.544)	0.531 (0.519;0.543)	0.533 (0.521;0.544)	0.529 (0.517;0.540)	0.531 (0.519;0.543)	0.502 (0.490;0.515)
70	0.525 (0.513;0.538)	0.524 (0.512;0.537)	0.526 (0.513;0.538)	0.522 (0.509;0.535)	0.525 (0.512;0.537)	0.500 (0.487;0.513)
80	0.522 (0.509;0.536)	0.522 (0.509;0.535)	0.523 (0.510;0.536)	0.520 (0.506;0.533)	0.522 (0.509;0.535)	0.498 (0.484;0.511)
90	0.524 (0.512;0.535)	0.523 (0.511;0.535)	0.524 (0.512;0.536)	0.520 (0.509;0.532)	0.523 (0.511;0.535)	0.501 (0.488;0.513)
100	0.526 (0.514;0.538)	0.525 (0.513;0.538)	0.527 (0.514;0.539)	0.523 (0.510;0.535)	0.525 (0.513;0.538)	0.503 (0.490;0.516)
			Variance 3			
50	0.539 (0.527;0.551)	0.535 (0.523;0.547)	0.542 (0.530;0.554)	0.534 (0.522;0.546)	0.538 (0.526;0.550)	0.496 (0.483;0.510)
60	0.542 (0.531;0.554)	0.539 (0.527;0.551)	0.545 (0.534;0.557)	0.538 (0.526;0.550)	0.542 (0.530;0.553)	0.501 (0.488;0.514)
70	0.533 (0.521;0.546)	0.530 (0.518;0.543)	0.535 (0.522;0.547)	0.528 (0.516;0.541)	0.532 (0.520;0.545)	0.497 (0.484;0.511)
80	0.531 (0.517;0.544)	0.529 (0.516;0.542)	0.533 (0.519;0.546)	0.527 (0.514;0.540)	0.530 (0.517;0.543)	0.497 (0.483;0.511)
90	0.531 (0.519;0.542)	0.529 (0.517;0.540)	0.532 (0.520;0.544)	0.527 (0.515;0.538)	0.530 (0.518;0.542)	0.500 (0.487;0.512)
100	0.531 (0.519;0.544)	0.530 (0.518;0.542)	0.534 (0.521;0.546)	0.528 (0.516;0.541)	0.531 (0.519;0.543)	0.502 (0.489;0.515)

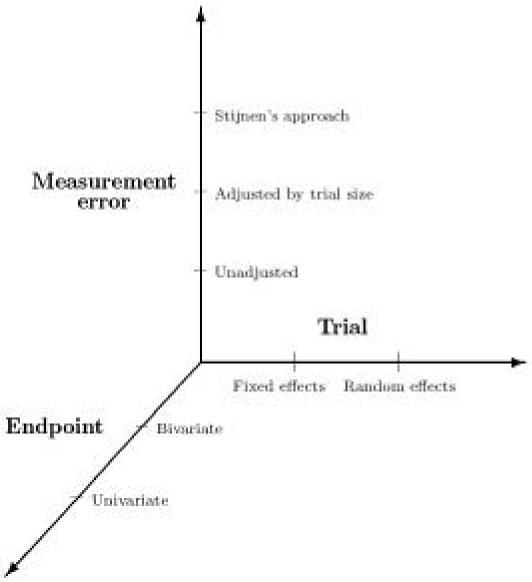

Figure 1: Graphical representation of the different approaches of this paper.