Simplified Modeling Strategies For Surrogate Validation With Multivariate Failure-Time Data

José Cortiñas Abrahantes [a,*] Tomasz Burzykowski [a]

[a] *Interuniversity Institute for Biostatisticsand statistical Bioinformatics (I-BioStat), Center for Statistics, Universiteit Hasselt. Agoralaan, B-3590 Diepenbeek. Belgium*

Abstract

The linear mixed effects model 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 unavoidable computational problems. In the context of surrogate markers, this problem has appeared when using an estimation and prediction-based approach for evaluation of surrogate endpoints. Convergence problems can occur mainly due to small between-trial variability or small number of trials. A number of alternative strategies has been proposed and studied for normally distributed data, but not such study have been conducted for other type of endpoints. The idea is to study if such simplified strategies, which always ignore individual level surrogacy, can also be applied when both surrogate and true endpoints are of failure-time types. It is shown via simulations that the 3 simplified strategies produced biased estimates, especially for the cases in which the strength of individual-level association is different from the strength of trial-level association. For this reason, it is recommended not to use simplified strategies when dealing with failure time data, in contrast to the case of normally distributed data, for which simplified strategies are recommended. Possible reasons for this discrepancy might be that, in this case, ignoring the individual level association influences estimates of the mean structure parameters, what results in distorted estimates of the trial level association.

Key words: Frailty model; Meta-analytic approach; Failure-time Data; Random effects; Surrogate endpoint.

* Corresponding author. Agoralaan 1, Diepenbeek. Belgium. Tel.: +32-11-268286; fax: +32-11-268299

Email address: jose.cortinas@luc.ac.be (José Cortiñas Abrahantes).

1 Introduction

Surrogate endpoints 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). A number of approaches appeared at the end of eighties to deal with this type of problems. Prentice (1989) and Freedman, Graubard and Schatzkin (1992) laid the foundations for the evaluation of surrogate endpoints in randomized clinical studies. Prentice proposed a definition as well as a set of operational criteria. Freedman, Graubard and Schatzkin (1992) supplemented these criteria with a quantity called ***proportion explained*** (PE). Buyse and Molenberghs (1998), on the other hand proposed to replace the PE by 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***). 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.

Molenberghs et al. (2002) and Alonso et al. (2004) pointed out serious issues surrounding the Prentice-Freedman framework. It has been asserted that the criteria set out by Prentice are too stringent (Fleming and DeMets, 1996) and neither necessary nor sufficient for his definition to be fulfilled, except in the special case of binary outcomes (Buyse and Molenberghs, 1998). In addition, Freedman, Graubard and Schatzkin (1992) showed that these criteria were not straightforward to verify through statistical hypothesis tests. Therefore the PE was suggested, but this measure is surrounded with difficulties, the most dramatic being not confined to the unit interval (Buyse et al., 2000). Buyse et al. (2000) 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. Assuming normality, they carried over the relative effect and adjusted association to a trial-level R^2_{trial} and an individual-level R^2_{indiv}, respectively. Similar routes of meta-analytic thinking have been followed by Daniels and Hughes (1997) and Gail et al. (2000).

Of course, the switch to a meta-analytic framework 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 treatments and 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 studies 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.

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 can be considered a basic one from a statistical modeling point of view, fitting such linear mixed effects models turns out to be surprisingly difficult. In order to overcome computational burden Tibaldi et al. (2003) proposed a set of simplified strategies for this particular setting. In other settings, e.g., when both endpoints are failure-times, the use of a mixed effects model is even less straightforward.

The aim of the paper is to study the possibility of also using simplified strategies when both the surrogate and the true endpoints are of the failure-time types. In 2001, Burzykowski in his dissertation already proposed a simplified strategy in this scenario and performed a simulation study, in which a copula model was used to assess the association at the individual level, and a fixed effects model was employed to assess the trial level surrogacy. In this paper we will study simplified modeling strategies that ignore the individual level association to evaluate the trial level surrogacy, following similar ideas as in Tibaldi et al. (2003). A simulation study is carried out to evaluate the performance of the different simplified strategies, in the setting studied by Burzykowski (2001). The idea of the paper is to show that even when simplified strategies performed well in the scenario of normally distributed endpoints, they should be carefully investigated in other scenarios, because they might produce biased results.

This paper is organized as follows. First, we present the original setting proposed by Buyse et al. (2000) and the simplified strategies considered by Tibaldi et al. (2003) (Sections 2 and 3), then extension to the setting for which both endpoints are of failure-time type, and which will be used throughout the paper is presented (Section 4). In Section 5 we describe the strategies that will be used to estimate the parameters of interest. A brief description of the method proposed by Burzykowski et al. (2001) is presented in Section 6, follow by the description of the simulation study considered (Section 7). The results obtained are described in Section 8. Discussion in Section 9 concludes the paper.

2 Setting for Normally Distributed Endpoints

In this section we briefly described the approach proposed by Buyse et al. (2000) in order to study in details the extension for failure-time data. Let $Y_{T_{ij}}$ and $Y_{S_{ij}}$ be random variables denoting the true and the surrogate endpoints for subject $j = 1, \ldots n_i$ in trial $i = 1, \ldots N$. Further, let Z_{ij} denote a binary treatment indicator.

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

$$Y_{S_{ij}} = \mu_S + m_{S_i} + \alpha Z_{ij} + a_i Z_{ij} + \varepsilon_{S_{ij}}, \tag{1}$$
$$Y_{T_{ij}} = \mu_T + m_{T_i} + \beta Z_{ij} + b_i Z_{ij} + \varepsilon_{T_{ij}}, \tag{2}$$

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 $\varepsilon_{T_{ij}}$, which are zero-mean normally distributed with variance-covariance matrix

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

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

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

Buyse et al. (2000) proposed a measure to assess the quality of the surrogate at the trial level, based on the coefficient of determination

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

A good surrogate, *at the trial level*, would have (5) close to 1. Similarly, to measure individual-level surrogacy, Buyse et al. (2000) proposed to use the

coefficient of determination given by

$$R^2_{\text{indiv}} = \frac{\sigma^2_{ST}}{\sigma_{SS}\sigma_{TT}}, \tag{6}$$

where σ_{ST}, σ_{SS} and σ_{TT} are components of variance-covariance matrix (3). The authors also proposed a reduce version to evaluate trial-level surrogacy which is given by the following expression

$$R^2_{\text{trial (r)}} = R^2_{b_i|a_i} = \frac{d^2_{ab}}{d_{aa}d_{bb}}. \tag{7}$$

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. Ample details about the approach can be found in Buyse et al. (2000).

3 Simplified Modelling Strategies

Buyse et al. (2000) showed that fitting mixed effects model (1)–(2) can be a surprisingly difficult task in a number of situations. Especially when the number of trials or the number of patients per trial is small. Also, situations with extreme correlations pose problems. It is therefore that approximate strategies with better computational properties have been studied for this particular setting in which both endpoints are normally distributed. Tibaldi et al. (2003) consider three dimensions along which simplifications that can be made in order to deal with computational burden in such setting.

Trial dimension: for which the trial-specific effects can be treated as either random or fixed.

Endpoint dimension: the surrogate and true endpoints are modelled as a bivariate outcome or two univariate ones. In the latter case the individual-level surrogacy is not incorporated into the modeling strategy. However, throughout this paper the focus is on trial-level surrogacy.

Measurement error dimension: whenever the full mixed effects model is abandoned, measurement error arises. The authors consider three ways to account for measurement error: unadjusted (i.e., no correction at all), adjustment by trial size, and an approach suggested by Van Houwelingen, Arends and Stijnen (2001).

4 Extension to Failure-Time Endpoints

In order to extend the meta-analytic approach used in the case of two normally-distributed endpoints described in Section (2), one could consider replacing model (1)–(2) by the following mixed effects proportional hazards model:

$$\lambda_{ijS}(y_{S_{ij}}|m_{S_i};\alpha;a_i;\varepsilon_{ij}) = \lambda_S(y_{S_{ij}})\varepsilon_{ij}e^{m_{S_i}+\alpha Z_{ij}+a_i Z_{ij}} \ , \tag{8}$$

$$\lambda_{ijT}(y_{T_{ij}}|m_{T_i};\beta;b_i;\varepsilon_{ij}) = \lambda_T(y_{T_{ij}})\varepsilon_{ij}e^{m_{T_i}+\beta Z_{ij}+b_i Z_{ij}} \ , \tag{9}$$

where i indexes the trials, j indexes the subjects and, as in Section (2), 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 trial specific random treatment effects. The vector of random effects, $(m_{S_i}, m_{T_i}, a_i, b_i)^T$, is assumed to be zero-mean normally distributed with variance-covariance matrix

$$D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ d_{ST} & d_{TT} & d_{Ta} & d_{Tb} \\ d_{Sa} & d_{Ta} & d_{aa} & d_{ab} \\ d_{Sb} & d_{Sa} & d_{ab} & d_{bb} \end{pmatrix}. \tag{10}$$

The random effects ε_{ij} are chosen to induce individual-level association between $Y_{S_{ij}}$ and $Y_{T_{ij}}$. The major stumbling block in the use of model (8)–(9) is the presence of individual random effects ε_{ij}. Given the current methodology and available software, fitting the model is practically impossible. Therefore, we will focus on strategies that ignore the association at the individual level and concentrate on the evaluation of trial level surrogacy. In the next section we will briefly describe the simplified strategies considered which focus on endpoint and trial dimension previously presented.

5 Modelling Strategies

We will consider the following three modelling strategies, in which the individual-level association is ignored.

5.1 Marginal PH Model with Trial-Specific Fixed Effects (MFE).

A Cox proportional hazards model will be fitted separately for each trial and also for each endpoint. More specifically, the model can be written as

$$\lambda_{ijS}(y_{S_{ij}}|\alpha_i) = \lambda_{Si}(y_{S_{ij}})e^{\alpha_i Z_{ij}} \ , \tag{11}$$
$$\lambda_{ijT}(y_{T_{ij}}|\beta_i) = \lambda_{Ti}(y_{T_{ij}})e^{\beta_i Z_{ij}} \ , \tag{12}$$

where α_i, and β_i are trial-specific treatment effects. Similar to the approach proposed by Buyse et al. (2000) for the case of two normally distributed endpoints, at the second stage we compute the determination coefficient ($R^2_{\text{trial (r)}}$) from the regression of β_i on α_i. To compute its variance we first use the delta method and treat the determination coefficient as a function of the correlation $R_{\text{trial (r)}}$

$$\text{Var}(R^2_{\text{trial (r)}}) \approx 4R^2_{\text{trial (r)}}\text{Var}(R_{\text{trial (r)}}). \tag{13}$$

Then we use the fact that the variance of Fisher's transformation $Z = \frac{1}{2}\ln\left(\frac{1+R_{\text{trial (r)}}}{1-R_{\text{trial (r)}}}\right)$ is equal to $\frac{1}{N-3}$ (Anderson 1958, p.78). Now, $R_{\text{trial (r)}}$ can be rewritten as a function of Z

$$R_{\text{trial (r)}} = \frac{e^{2Z}-1}{e^{2Z}+1},$$

so, using the delta method and treating $R_{\text{trial (r)}}$ as a function of Z, we get

$$\text{Var}(R_{\text{trial (r)}}) \approx \frac{(1-R^2_{\text{trial (r)}})^2}{N-3}. \tag{14}$$

Combining (13) and (14) leads to

$$\text{Var}(R^2_{\text{trial (r)}}) \approx \frac{4R^2_{\text{trial (r)}}(1-R^2_{\text{trial (r)}})^2}{N-3}.$$

5.2 A Stratified PH Model with Random Trial-Specific Treatment Effects (SRTE).

In this model we use stratified baseline hazards to account for the between-trial variability in baseline hazards, and bivariate random effects associated to the treatment. The model can be written as follows:

$$\lambda_{ijS}(y_{S_{ij}}|\alpha; a_i) = \lambda_{Si}(y_{S_{ij}})e^{(\alpha+a_i)Z_{ij}} \ , \tag{15}$$
$$\lambda_{ijT}(y_{T_{ij}}|\beta; b_i) = \lambda_{Ti}(y_{T_{ij}})e^{(\beta+b_i)Z_{ij}} \ , \tag{16}$$

where α and β are fixed treatment effects. The trial specific random effects a_i and b_i are assumed to be zero-mean normally distributed with variance-

covariance matrix

$$\begin{pmatrix} d_{aa} & d_{ab} \\ d_{ab} & d_{bb} \end{pmatrix}. \tag{17}$$

The trial-level validity of a surrogate is evaluated using the square of the correlation coefficient based on the estimated variance-covariance matrix (17). Its variance is estimated using equation (13), with $\mathrm{Var}(R_{\mathrm{trial\ (r)}})$ estimated by applying the delta method to $R_{\mathrm{trial\ (r)}}$ treated as a function of the estimated elements of matrix (17).

5.3 PH Model with Trial-Specific Random Intercepts and Treatment Effects (RITE).

In this model we consider both random intercepts and treatment effects. The model is:

$$\lambda_{ijS}(y_{S_{ij}}|\alpha; a_i; \varepsilon_{S_{ij}}) = \lambda_S(y_{S_{ij}})e^{m_{S_i}+\alpha Z_{ij}+a_i Z_{ij}}\ , \tag{18}$$

$$\lambda_{ijT}(y_{T_{ij}}|\beta; b_i; \varepsilon_{T_{ij}}) = \lambda_T(y_{T_{ij}})e^{m_{T_i}+\beta Z_{ij}+b_i Z_{ij}}\ , \tag{19}$$

where $(m_{S_i}, m_{T_i}, a_i, b_i)^T$ is a vector of random effects, assumed to be zero-mean normally distributed with variance-covariance matrix (10). The association at the trial level is evaluated using the determination coefficient computed using equation (7). The variance of the coefficient is estimated using the same procedure as in strategy SRTE.

The models were fitted using the SAS procedure PHREG for the MFE approach and the modified version of the EM algorithm (as described in Cortiñas and Burzykowski (2005), we only considered this approach given that in a comparison performed by Cortiñas et al. (2007) the estimated values for fixed and random effects were comparable), implemented using SAS-IML v8.2, for the SRTE and RITE approaches. We will compare the performance of the simplified strategies with the approach developed by Burzykowski et al. (2001), which we describe in the next Section.

6 A Copula Modelling Approach

Burzykowski et al. (2001) assumed that the variables $Y_{T_{ij}}$ (the true endpoint) and $Y_{S_{ij}}$ (the surrogate endpoint) were distributed according to a bivariate

distribution with the joint survival function

$$P(Y_{S_{ij}} \geq y_{S_{ij}}, Y_{T_{ij}} \geq y_{T_{ij}}) = C_\theta \left\{ F_{Y_{S_{ij}}}(y_{S_{ij}}), F_{Y_{T_{ij}}}(y_{T_{ij}}) \right\}, \tag{20}$$

where $C_\theta\{.,.\}$ is a copula function. In particular, they considered the use of the Clayton copula

$$C_\theta(u, v) = \left(u^{1-\theta} + v^{1-\theta} - 1 \right)^{\frac{1}{1-\theta}}, \ \theta > 1. \tag{21}$$

The marginal survivor functions $F_{Y_{S_{ij}}}(y_{S_{ij}})$ and $F_{Y_{T_{ij}}}(y_{T_{ij}})$ were modeled using PH models. The use of the copula allows to assess the individual-level surrogacy by Kendall's τ, which for the Clayton copula (21) can be computed as a simple function of θ:

$$\tau = \frac{\theta - 1}{\theta + 1}.$$

Furthermore, $Y_{T_{ij}}$ and $Y_{S_{ij}}$ were assumed to be exponentially distributed, with the marginal survivor functions $F_{Y_{T_{ij}}}$ and $F_{Y_{S_{ij}}}$ defined as

$$F_{Y_{S_{ij}}}(y_{S_{ij}}) = e^{-y_{S_{ij}} \lambda_S \cdot e^{\{m_{S_i} + (\alpha + a_i) Z_{ij}\}}}, \tag{22}$$

$$F_{Y_{T_{ij}}}(y_{T_{ij}}) = e^{-y_{T_{ij}} \lambda_T \cdot e^{\{m_{T_i} + (\beta + b_i) Z_{ij}\}}}, \tag{23}$$

where i indexes the trials, j indexes the subjects, 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 trial specific random treatment effects. The vector of random effects, $(m_{S_i}, m_{T_i}, a_i, b_i)^T$, was also assumed to be zero-mean normally distributed with variance-covariance matrix (10).

Variables $Y_{T_{ij}}$ and $Y_{S_{ij}}$ are assumed exponentially distributed with marginal survival functions given by (22)–(23) and the joint survival function (20)–(21), were generated using the conditional distribution method (Nelsen , 1999). Note, that the marginal survivor functions are also conditional on the random effects m_{S_i}, m_{T_i}, a_i and b_i.

Burzykowski (2001) proposed to use the maximum likelihood estimates of the parameters of model (20)–(21), assuming the fixed effects representation of (22)–(23)

$$F_{Y_{S_{ij}}}(y_{S_{ij}}) = e^{-y_{S_{ij}} \lambda_{S_i} \cdot e^{\alpha_i Z_{ij}}}, \tag{24}$$

$$F_{Y_{T_{ij}}}(y_{T_{ij}}) = e^{-y_{T_{ij}} \lambda_{T_i} \cdot e^{\beta_i Z_{ij}}}, \tag{25}$$

In (20), trial-specific treatment effects were estimated as fixed effects α_i and β_i in the marginal PH models for $F_{Y_{S_{ij}}}(y_{S_{ij}})$ and $F_{Y_{T_{ij}}}(y_{T_{ij}})$, respectively. The trial level surrogacy was evaluated using the determination coefficient from the linear regression of β_i on α_i.

The main advantage of the approach proposed by Burzykowski et al. (2001) is that it does allow for the evaluation of the individual level association. However, in the approach treatment effects are modelled as fixed rather than random effects. The strategies proposed in Section 5 (SRTE and RITE) use random effects at the trial level, but ignore individual level association. The question is, how much does this influence the performance of the simplified strategies. To answer this question, a simulation study is conducted.

7 Simulation Study

To investigate the performance of the various strategies presented in Section 5, a simulation study was undertaken. The strategies were compared against each other and against the copula approach developed by Burzykowski et al. (2001). In the simulations, data for N randomized clinical trials with n_i observations (subjects) within a trial were generated. A single binary covariate Z_{ij} was considered, corresponding to a treatment randomization within each trial. The data were generated using model (8)–(9), exponential marginal models were considered and the baseline hazards were assumed constant, with $\lambda_S(y_{S_{ij}}) = 0.69$ and $\lambda_T(y_{T_{ij}}) = 1.39$, resulting in median times equal to 1 and 0.5 respectively. Two settings for treatment effects were considered: $\alpha = \beta = 0$ (no treatment effect) or $\alpha = \beta = -0.4$ (corresponding to 33% reduction in failure rate). The variance-covariance matrix of the normally distributed random effects $(m_{S_i}, m_{T_i}, a_i, b_i)^T$, was assumed to be equal to

$$D = \sigma^2 \begin{pmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho \\ 0 & 0 & \rho & 1 \end{pmatrix}. \tag{26}$$

The value of σ^2 was set to equal 0.2, while ρ was set to $\sqrt{0.5}$ or $\sqrt{0.9}$, resulting in the trial-level $R^2_{\text{trial (r)}}$ of 0.5 and 0.9, respectively. The individual random effects ε_{ij} were considered gamma distributed with density function

$$f(x) = \frac{x^{\frac{2-\theta}{\theta-1}} e^{-x}}{\Gamma(\frac{1}{\theta-1})}. \tag{27}$$

In this way, the data generated using model (8)–(9) were comparable to those generated by Burzykowski (2001) using the Clayton copula (21). Two hundred and fifty datasets were simulated for every setting, with $N = 10$ or $N = 20$ trials each. The number n_i of observations (subjects) per trial was assumed to equal 50, 100 and 200. The copula parameter θ was assumed equal to 3 or 19, resulting in Kendall's τ of 0.5 and 0.9, respectively, for the association between $Y_{T_{ij}}$ and $Y_{S_{ij}}$ (conditional on the random effects m_{S_i}, m_{T_i}, a_i and b_i). We considered no censoring or homogeneous censoring. In the latter set-up, $Y_{T_{ij}}$ and $Y_{S_{ij}}$ were assumed to be simultaneously censored by an independent variable C_{ij} uniformly distributed on the interval $[0, \delta_c]$. The parameter δ_c was chosen to be 2.3 for $\alpha = \beta = 0$, resulting in 50 % of the observations of $Y_{T_{ij}}$ (the true endpoint) and 30 % of the observations of $Y_{S_{ij}}$ (the surrogate endpoint) censored, similar to the censoring schemes considered by Burzykowski (2001).

8 Results of the Simulation Study

The results of the simulations are shown in Tables 1–4, note that column quoted as Burzykowski means that we displayed the results obtained by Burzykowski (2001) for the case of unadjusted R^2; MFE refers to marginal PH model with trial-specific fixed effects; SRTE to stratified PH model with random trial-specific treatment effects and RITE to PH model with trial-specific random intercepts and treatment effects results. The %NC represents the % of samples with non-convergence and %bias the bias relative to the true value of the parameter (in %). The tables show the relative bias of $R^2_{\text{trial (r)}}$ for different values of τ, $R^2_{\text{trial (r)}}$ $(= \rho^2)$, α, β, and different censoring schemes. It is worth noting that the results displayed for the method proposed by Burzykowski et al. (2001) are based on 500 simulations. For the latter method and $\tau = 0.9$ the percentage of samples with non-convergence is also shown (for $\tau = 0.5$ there were no convergence problems). The simplified strategies proposed in Section 5 do not suffer from convergence problems.

Table 1

The mean estimates of trial level $R^2_{trial\ (r)}$ for the method proposed by Burzykowski (2001) and the simplified strategies when $\tau = 0.5$ and $R^2_{trial\ (r)} = \rho^2 = 0.5$. In parentheses: the mean model-based and empirical (first and second number) standard error.

		Burzykowski	MFE	SRTE	RITE
N	n_i	%bias	%bias	%bias	%bias
No censoring, no treatment effect ($\alpha = \beta = 0$)					
10	50	-0.2(0.225;0.228)	-0.5(0.236;0.220)	-3.4(0.226;0.219)	0.6(0.230;0.218)
	100	1.4(0.224;0.226)	-7.6(0.241;0.223)	-1.5(0.212;0.209)	-1.5(0.218;0.214)
	200	1.7(0.225;0.220)	-9.7(0.240;0.226)	-0.3(0.211;0.210)	-0.3(0.214;0.210)
20	50	5.5(0.154;0.164)	-1.0(0.161;0.166)	-2.7(0.169;0.156)	-1.7(0.162;0.157)
	100	2.6(0.158;0.156)	-8.7(0.165;0.169)	-0.6(0.151;0.145)	-0.6(0.149;0.146)
	200	0.9(0.158;0.168)	-12.5(0.167;0.170)	-0.3(0.141;0.135)	-0.2(0.137;0.134)
No censoring, 33% reduction in the failure rate ($\alpha = \beta = -0.4$)					
10	50	4.4(0.227;0.215)	-0.4(0.236;0.221)	-1.6(0.233;0.218)	0.7(0.229;0.216)
	100	4.3(0.225;0.220)	-7.9(0.242;0.222)	-1.5(0.224;0.214)	-1.4(0.226;0.215)
	200	0.5(0.228;0.221)	-9.8(0.240;0.226)	-0.3(0.218;0.212)	-0.3(0.214;0.210)
20	50	2.9(0.157;0.161)	-1.2(0.162;0.166)	-2.3(0.168;0.154)	-1.8(0.162;0.157)
	100	4.7(0.157;0.154)	-8.8(0.165;0.169)	-0.7(0.153;0.147)	-0.6(0.149;0.146)
	200	0.3(0.159;0.163)	-12.6(0.167;0.170)	-0.2(0.141;0.136)	-0.1(0.137;0.133)
Homogeneous Censoring (50%/30%), no treatment effect					
10	50	-12.6(0.228;0.233)	-4.7(0.235;0.235)	-4.9(0.247;0.239)	-4.3(0.256;0.243)
	100	-7.2(0.228;0.235)	-4.1(0.239;0.227)	-3.3(0.229;0.222)	-2.1(0.228;0.221)
	200	-1.9(0.228;0.225)	-10.4(0.240;0.222)	-4.1(0.229;0.223)	-2.6(0.225;0.221)
20	50	-19.1(0.167;0.167)	-4.6(0.164;0.161)	-5.7(0.161;0.154)	-3.7(0.164;0.157)
	100	-9.7(0.163;0.169)	-6.0(0.164;0.166)	-2.9(0.154;0.148)	-1.8(0.151;0.147)
	200	-5.4(0.163;0.162)	-12.1(0.167;0.163)	-2.5(0.147;0.142)	-1.7(0.144;0.141)

Table 2

The mean estimates of trial level $R^2_{trial\ (r)}$ for the method proposed by Burzykowski (2001) and the simplified strategies when $\tau = 0.9$ and $R^2_{trial\ (r)} = \rho^2 = 0.9$. In parentheses: the mean model-based and empirical (first and second number) standard error.

			Burzykowski	MFE	SRTE	RITE
N	n_i	%NC	%bias	%bias	%bias	%bias
No censoring, no treatment effect ($\alpha = \beta = 0$)						
10	50	0.4	0.7(0.065;0.065)	-1.8(0.094;0.085)	2.1(0.091;0.076)	-1.3(0.076;0.064)
	100	0.2	-0.7(0.073;0.082)	-1.7(0.092;0.079)	-0.7(0.078;0.068)	-0.7(0.074;0.067)
	200	1.0	-0.2(0.071;0.068)	-4.1(0.105;0.086)	-0.6(0.069;0.065)	-0.2(0.073;0.069)
20	50	0.2	1.0(0.042;0.045)	-1.1(0.054;0.054)	-0.4(0.056;0.043)	-0.2(0.049;0.044)
	100	0.8	0.3(0.045;0.044)	-1.4(0.055;0.049)	-0.7(0.046;0.037)	-0.6(0.043;0.039)
	200	0.2	-0.2(0.046;0.049)	-3.4(0.063;0.060)	-0.2(0.041;0.039)	-0.1(0.042;0.039)
No censoring, 33% reduction in the failure rate ($\alpha = \beta = -0.4$)						
10	50	0.6	-0.1(0.068;0.089)	-2.7(0.099;0.090)	-2.2(0.083;0.077)	-1.3(0.076;0.062)
	100	0.6	0.7(0.065;0.066)	-2.4(0.096;0.082)	-0.9(0.085;0.075)	-0.7(0.074;0.067)
	200	0.4	0.0(0.070;0.068)	-4.8(0.109;0.091)	-0.5(0.074;0.068)	-0.1(0.073;0.067)
20	50	0.2	1.4(0.040;0.040)	-1.8(0.057;0.057)	-0.6(0.053;0.043)	-0.2(0.049;0.044)
	100	0.2	0.1(0.045;0.052)	-2.0(0.057;0.052)	-0.8(0.049;0.041)	-0.5(0.043;0.038)
	200	0.0	0.1(0.045;0.044)	-4.0(0.065;0.061)	-0.3(0.045;0.040)	-0.1(0.043;0.039)
Homogeneous Censoring (50%/30%), no treatment effect						
10	50	0.6	-3.1(0.086;0.096)	-10.2(0.139;0.128)	-8.5(0.123;0.106)	-7.9(0.113;0.101)
	100	0.2	-1.4(0.077;0.079)	-9.3(0.135;0.120)	-8.6(0.116;0.108)	-7.6(0.111;0.104)
	200	0.8	-1.1(0.075;0.086)	-10.5(0.140;0.139)	-8.4(0.140;0.134)	-7.3(0.137;0.133)
20	50	0.2	-2.8(0.054;0.089)	-9.6(0.085;0.086)	-8.8(0.083;0.069)	-6.8(0.076;0.068)
	100	0.8	-1.0(0.049;0.053)	-9.1(0.084;0.083)	-7.8(0.078;0.070)	-6.4(0.073;0.068)
	200	0.6	-1.0(0.049;0.050)	-8.9(0.083;0.080)	-7.3(0.080;0.075)	-6.1(0.079;0.075)

Table 3

The mean estimates of trial level $R^2_{trial\ (r)}$ for the method proposed by Burzykowski (2001) and the simplified strategies when $\tau = 0.5$ and $R^2_{trial\ (r)} = \rho^2 = 0.9$. In parentheses: the mean model-based and empirical (first and second number) standard error.

		Burzykowski	MFE	SRTE	RITE
N	n_i	%bias	%bias	%bias	%bias
No censoring, no treatment effect ($\alpha = \beta = 0$)					
10	50	-8.5(0.114;0.119)	-30.7(0.211;0.201)	-30.9(0.187;0.179)	-28.1(0.181;0.171)
	100	-4.8(0.096;0.094)	-27.5(0.201;0.186)	-25.4(0.173;0.159)	-23.3(0.168;0.159)
	200	-3.7(0.088;0.105)	-24.4(0.191;0.186)	-17.6(0.122;0.121)	-15.2(0.126;0.119)
20	50	-8.0(0.074;0.078)	-30.3(0.140;0.143)	-31.5(0.135;0.125)	-31.1(0.131;0.125)
	100	-4.8(0.063;0.065)	-27.3(0.133;0.135)	-23.2(0.113;0.105)	-21.7(0.106;0.103)
	200	-2.6(0.055;0.057)	-24.5(0.128;0.120)	-16.1(0.076;0.071)	-14.4(0.074;0.069)
No censoring, 33% reduction in the failure rate ($\alpha = \beta = -0.4$)					
10	50	-9.4(0.118;0.125)	-30.7(0.211;0.201)	-30.2(0.188;0.179)	-27.8(0.181;0.171)
	100	-5.9(0.102;0.101)	-27.7(0.202;0.185)	-25.3(0.171;0.163)	-23.5(0.167;0.158)
	200	-2.5(0.084;0.080)	-24.5(0.191;0.186)	-17.8(0.123;0.121)	-15.2(0.126;0.120)
20	50	-9.0(0.077;0.083)	-30.4(0.140;0.143)	-31.5(0.141;0.130)	-31.0(0.131;0.125)
	100	-4.3(0.061;0.067)	-27.4(0.133;0.135)	-23.1(0.113;0.105)	-21.8(0.106;0.103)
	200	-2.0(0.053;0.057)	-24.5(0.128;0.120)	-16.2(0.076;0.072)	-14.4(0.074;0.069)
Homogeneous Censoring (50%/30%), no treatment effect					
10	50	-32.7(0.202;0.206)	-36.9(0.224;0.217)	-36.6(0.215;0.200)	-34.5(0.213;0.202)
	100	-20.8(0.166;0.173)	-31.1(0.213;0.199)	-31.0(0.187;0.178)	-30.7(0.184;0.175)
	200	-12.2(0.131;0.140)	-28.3(0.203;0.191)	-27.3(0.167;0.158)	-25.5(0.162;0.154)
20	50	-33.4(0.141;0.151)	-36.2(0.151;0.148)	-36.4(0.148;0.135)	-34.2(0.140;0.134)
	100	-20.0(0.111;0.107)	-31.0(0.141;0.140)	-30.4(0.122;0.115)	-29.0(0.119;0.114)
	200	-12.4(0.089;0.092)	-28.3(0.136;0.124)	-26.3(0.099;0.093)	-24.6(0.098;0.092)

Table 4

The mean estimates of trial level $R^2_{trial\ (r)}$ for the method proposed by Burzykowski (2001) and the simplified strategies when $\tau = 0.9$ and $R^2_{trial\ (r)} = \rho^2 = 0.5$. In parentheses: the mean model-based and empirical (first and second number) standard error.

			Burzykowski	MFE	SRTE	RITE
N	n_i	%NC	%bias	%bias	%bias	%bias
No censoring, no treatment effect ($\alpha = \beta = 0$)						
10	50	0.6	16.2(0.211;0.209)	71.7(0.110;0.096)	72.8(0.086;0.079)	71.2(0.084;0.071)
	100	0.2	6.9(0.225;0.217)	66.7(0.123;0.108)	72.5(0.107;0.089)	71.0(0.096;0.087)
	200	0.4	3.9(0.224;0.217)	55.7(0.153;0.125)	60.6(0.091;0.087)	59.4(0.095;0.091)
20	50	0.4	10.2(0.150;0.161)	73.0(0.064;0.061)	75.4(0.055;0.047)	73.5(0.055;0.049)
	100	0.8	5.3(0.156;0.155)	67.9(0.074;0.068)	68.9(0.059;0.049)	67.8(0.053;0.050)
	200	0.2	0.8(0.159;0.153)	57.0(0.095;0.089)	57.1(0.058;0.055)	56.7(0.060;0.057)
No censoring, 33% reduction in the failure rate ($\alpha = \beta = -0.4$)						
10	50	0.6	13.6(0.209;0.217)	70.1(0.115;0.100)	72.8(0.085;0.079)	70.9(0.084;0.070)
	100	0.0	6.8(0.218;0.225)	65.3(0.127;0.111)	72.5(0.104;0.091)	70.7(0.096;0.088)
	200	0.0	7.0(0.220;0.223)	53.8(0.157;0.130)	59.4(0.097;0.092)	58.6(0.094;0.092)
20	50	0.0	14.5(148;0.146)	71.5(0.068;0.065)	75.9(0.055;0.046)	73.6(0.055;0.049)
	100	0.6	8.6(0.153;0.156)	66.4(0.077;0.071)	68.9(0.057;0.050)	67.5(0.053;0.050)
	200	0.0	1.7(0.158;0.161)	55.4(0.098;0.092)	56.7(0.061;0.057)	56.2(0.060;0.057)
Homogeneous Censoring (50%/30%), no treatment effect						
10	50	0.6	16.6(0.209;0.215)	60.8(0.141;0.127)	58.3(0.123;0.106)	55.8(0.116;0.101)
	100	1.4	10.5(0.215;0.220)	61.5(0.139;0.125)	57.5(0.126;0.111)	55.2(0.115;0.108)
	200	0.4	5.1(0.220;0.225)	56.8(0.150;0.146)	48.9(0.149;0.140)	47.3(0.145;0.139)
20	50	0.2	16.1(0.147;0.138)	61.7(0.087;0.084)	60.2(0.081;0.069)	58.4(0.075;0.068)
	100	0.4	10.1(0.151;0.156)	61.5(0.088;0.090)	53.6(0.080;0.071)	50.2(0.077;0.071)
	200	1.2	4.6(0.156;0.159)	59.9(0.090;0.088)	49.1(0.085;0.079)	44.9(0.084;0.079)

A few general observations can be made. First, we can note that the presence of a treatment effect does not have much influence on the relative bias of the estimation of $R^2_{\text{trial (r)}}$. Also, under no censoring the relative bias in absolute value is smaller than when censoring is considered. One can also observe that for MFE, model-based estimates of the standard error of $R^2_{\text{trial (r)}}$ overestimate the empirical standard errors. The other approaches yield comparable model-based and empirical standard errors of the estimates of $R^2_{\text{trial (r)}}$.

Table 1 shows the simulation results for $\tau = 0.5$ and $\rho^2 = 0.5$. In terms of point estimation it can be seen that, in general, RITE approach yields the smallest relative bias in absolute value, followed by SRTE approach, while

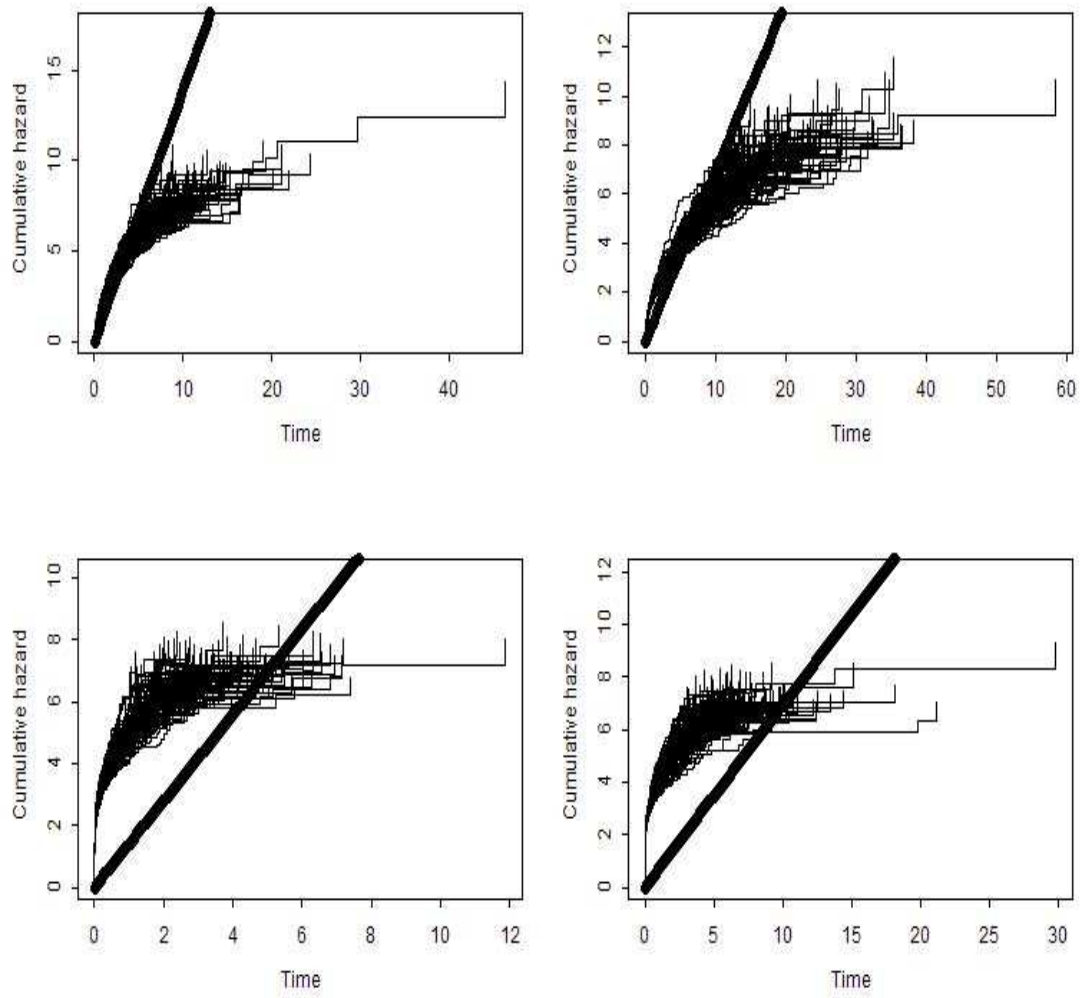
the largest bias is observed for MFE approach. It can also be noted that the RITE and SRTE approaches are subject to similar empirical variability as in the method proposed by Burzykowski et al. (2001), with the MFE approach showing larger variability. For the other settings (Tables 2–4), in general, the smallest relative bias in absolute value is observed for the method proposed by Burzykowski et al. (2001), while the MFE approach yields estimates with the largest relative bias in absolute value. It is worth noting that the three simplified strategies produce substantially biased estimates of $R^2_{\text{trial (r)}}$ when the association at the individual level (measured by τ) and the association at the trial level (measured by ρ^2) are different. It can be also observed that if $\rho^2 = 0.9$, the Burzykowski et al. (2001) approach produces estimates with the smallest empirical standard error, while the MFE approach produces estimates with the largest standard errors. If $\rho^2 = 0.5$, the RITE approach yields estimates with the smallest variability, while the approach proposed by Burzykowski et al. (2001) produces estimates with the largest variability. It is important to note that for different values of ρ^2 when the value of τ is kept the same, only moderate changes in the magnitude of the variability of the estimates produced by the simplified strategies are observed. Thus, if the individual level is ignored in the fitting process, the value of τ may determine the magnitude of the variability of the estimates. From the evaluation of the relative bias one can conclude that the use of the simplified strategies does not yield reasonable results. This is in contrast to the case of normally distributed data considered by Tibaldi et al. (2003). In order to investigate the cause of the large relative bias in absolute value produced by the simplified strategies, we evaluated the estimated cumulative baseline hazards when the RITE approach was used for two particular settings under no censoring (Setting I: $N = 20$, $n_i = 50$, $\tau = 0.5$, $\rho^2 = 0.9$; Setting II: $N = 20$, $n_i = 50$, $\tau = 0.9$, $\rho^2 = 0.5$). Figure 1 shows the estimated cumulative baseline hazard functions (step curves) versus the observed time for both endpoints and settings I and II. From this figure, it can be concluded that the estimated cumulative baseline hazards do not correspond to the true cumulative hazard function (thick solid line) used to generate the data. This suggests that ignoring individual level association results in a strong modification of the baseline hazard. Consequently, the estimates of the mean structure parameters are affected, which in turn influences the trial-specific treatment effects and the estimation of the trial-level association in MFE approach. It is also important to note that Cortiñas et al. (2004) have also observed similar problems, when they fitted a hierarchical model ignoring a level, and the variance associated to the level was of the same magnitude or larger than the variance at the higher level, the association at the higher level was affected. Here a similar pattern can be observed. When ε_{ij} is gamma-distributed with density function (27), the logarithm of ε_{ij} will have variance $\psi'(\frac{1}{\theta-1})$, where $\psi'()$ is the trigamma function (Johnson and Kotz, 1970, p. 181). For $\theta = 3$, $\psi'(\frac{1}{\theta-1}) = 4.9$, while for $\theta = 19$, $\psi'(\frac{1}{\theta-1}) = 325.5$. It is clear that the variance at the individual

level is much larger than the variance at the trial level. Thus, if the individual level is ignored, and by analogy to the results obtained for the normallly distributed data case, the trial-level variability is probably contaminated with $\frac{\psi'(\frac{1}{\theta-1})}{N} \times 100\%$ of the variance at the individual level, what results in bias observed for SRTE and RITE strategies.

Fig. 1. *Graphical representation of the estimated cumulative baseline hazard (step curves) versus time. The thick solid line corresponds to the cumulative hazard used to generate the data.*

Left column: surrogate endpoint; right column: true endpoint.

Top row: $\tau = 0.5$ and $\rho^2 = 0.9$; bottom row: $\tau = 0.9$ and $\rho^2 = 0.5$.



9 Conclusions

In this paper we have investigated the use of several strategies for the evaluation of the trial level surrogacy in the case of two failure-time endpoints. In particular, we have considered three strategies which may reduce the computational burden and the complexity of the model. The first one (marginal PH model with trial-specific fixed effects) is a very simple procedure, which can easily be implemented with standard software. This model is comparable to the simplest model used in Tibaldi et al. (2003), in which only fixed effects are used, every endpoint is modeled separately, and we do not adjust for measurement errors. In the second strategy, random effects are considered, but stratified baseline hazards are used in order to capture the between-trial variability in baseline hazards (SRTE). The third strategy is a PH model with trial-specific random intercepts and treatment effects (RITE). It is shown via simulations that the three simplified strategies produce highly biased estimates, especially for the cases in which the strength of the individual-level association is different from the strength of the trial-level association. For this reason, the simplified strategies should not be used in the case of two failure time endpoints. This is in contrast to the case of normally distributed data (see Tibaldi et al. (2003), for more details). In this scenario simplified modeling strategies produce untrustful results, thus they should not be employed to evaluate surrogacy. For the MFE approach the possible reason for this discrepancy is that, in the case of failure-time data, ignoring the individual level association influences estimates of the mean structure parameters, what results in distorted estimates of the trial level association. It is important to note that, in this paper we did not study methods that correct for measurement error, as it has been done in Tibaldi et al. (2003). In this respect, a GEE approach might be of interest. This is a topic for further research. Another way of handling the complexities in the model is to use a bayesian modeling framework, which could handle the hierarchical structure on the hazard scale very easily, but this is out of the scope of the paper, since we would like to be able to use easy to implement procedures in standard software, but unfortunately in this setting with two failure time endpoints simplified strategies do not provide promising results.

Acknowledgement

The authors gratefully acknowledge support from the fund of Scientific Research (FWO, Research Grant G.0151.05) and Belgian IUAP/PAI network P6/03 “Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data” of the Belgian Government (Belgian Science Policy).

References

Alonso, A., Geys, H., Molenberghs, G., and Vangeneugden, T., (2002). Investigating the criterion validity of psychiatric symptom scales using surrogate marker validation methodology. Journal of Biopharmaceutical Statistics 12, 161–179.

Alonso, A., Molenberghs, G., Burzykowski, T., Renard, D., Geys, H., Shkedy, Z., Tibaldi, F., Cortiñas Abrahantes, J., and Buyse, M., 2004. Prentices Approach and the Meta-analytic paradigm: A reflection on the role of statistics in the evaluation of surrogate endpoints. Biometrics, 60(3), 724–728.

Burzykowski, T., Molenberghs, G., Buyse, M., Geys, H., and Renard, D., 2001. Validation of surrogate endpoints in multiple randomized clinical trials with failure-time endpoints.Journal of the Royal Statistical Society C (Applied Statistics) 50, 405–422.

Burzykowski, T., 2001. *Validation of Surrogate Endpoints From Multiple Randomized Clinical Trials With a Failure-Time True Endpoint.* Unpublished Ph.D. dissertation, Limburgs Universitair Centrum, Dept. of Mathematics

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., 2000. The validation of surrogate endpoints in meta-analyses of randomized experiments. Biostatistics 1, 49–67.

Cortinas Abrahantes, J., Molenberghs, G., Burzykowski, T., Shkedy, Z., Alonso, A., Renard, D., 2004. Choice of Units of Analysis and Modeling Strategies in Multilevel Hierarchical Models. Computational Statistics and Data Analysis, 47, 537–563.

Cortiñas Abrahantes, J. and Burzykowski, T., 2005. A version of the EM algorithm for proportional hazards model with random effects. Biometrical Journal, 47, 847–862.

Cortiñas Abrahantes, J., Legrand, C., Burzykowski, T., Janssen, P., Ducrocq, V. and Duchateau, L., 2007. Comparison of different estimation procedures for proportional hazards model with random effects. Computational Statistics and Data Analysis, 51, 3913–3930.

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

Ellenberg S.S., and Hamilton J.M., 1989. Surrogate endpoints in clinical trials: cancer. Statistics in Medicine 8, 405–413.

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

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

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

Goldstein, H., 1995. Multilevel Statistical Models. Kendall's Libary of Statis-

tics 3. London: Arnold.

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

Johnson, N.L., and Kotz, S. (1970) *Distribution in Statistics: Continuous univariate distributions.* New York: John Wiley & Sons.

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

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 23, 607–625.

Nair, N.P.V. and 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.

Nelsen, R.G., 1999. An introduction to copulas. *Lecture Notes in Statistics*, **139**. New York: Springer-Verlag.

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

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

Searle, S.R., Casella, G., and McCulloch, C.E., 1992. Variance Components. New York: John Wiley & Sons.

Shkedy, Z., Torres Barbosa, F., Burzykowski, T., and Molenberghs, G., 2003. A hierarchical bayesian approach for the evaluation of surrogate endpoints in multiple randomized clinical trials. Verbeke, G., Molenberghs, G., Aerts, M. & Fieuws, S. (Ed.) Proceedings of the 18th International Workshop on Statistical Modelling, Belgium : Leuven, p. 403-407

Van Houwelingen, J.C., Arends, L.A., and Stijnen, T., 2001. Advanced methods for meta-analysis. Statistics in Medicine, 4, 589–624.

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