Journal of Applied Statistics, Vol. 30, No. 2, 2003, 235–247

Routledge
Taylor & Francis Group

Validation of a longitudinally measured surrogate marker for a time-to-event endpoint

DIDIER RENARD[1], **HELENA GEYS**[1], **GEERT MOLENBERGHS**[1], **TOMASZ BURZYKOWSKI**[1], **MARC BUYSE**[2], **TONY VANGENEUGDEN**[3] **& LUC BIJNENS**[3] [1]*Limburgs Universitair Centrum, Diepenbeek, Belgium,* [2]*International Drug Development Institute, Brussels, Belgium and* [3]*Johnson and Johnson, Pharmaceutical Research and Development, Beerse, Belgium*

ABSTRACT *The objective of this paper is to extend the surrogate endpoint validation methodology proposed by Buyse* et al. *(2000) to the case of a longitudinally measured surrogate marker when the endpoint of interest is time to some key clinical event. A joint model for longitudinal and event time data is required. To this end, the model formulation of Henderson* et al. *(2000) is adopted. The methodology is applied to a set of two randomized clinical trials in advanced prostate cancer to evaluate the usefulness of prostate-specific antigen (PSA) level as a surrogate for survival.*

1 Introduction

In recent years, interest in modelling the relationship between a time-to-event endpoint and longitudinally measured data has developed considerably. This problem occurs naturally in many biomedical or public health studies where participants are followed over time. In such studies, measurements on a number of outcomes can also be obtained at different occasions and times to some clinical events can be observed.

In randomized clinical trials, the main question is often whether a new treatment has some beneficial effect on the time to a certain clinical event, the endpoint of primary interest. The time elapsed between randomization and this event, however, can be very long and it may therefore be desirable to find a surrogate for the

Correspondence: Didier Renard, ELI-LILLY-MSG, Rue Granbompré 11, 1348, Mont-Saint-Guibert, Belgium. E-mail: Renardd@lilly.com

ISSN 0266-4763 print; 1360-0532 online/03/020235-13
DOI: 10.1080/0266476022000023776

clinical outcome of interest that is less distant in time, thereby permitting a trial to be completed sooner and making a potentially useful treatment available earlier to a wider range of patients. A well-known example is in AIDS research where an early proposal of a surrogate marker for clinical outcomes such as disease progression or survival was the number of CD4 T-lymphocytes (see, for example, Tsiatis *et al.*, 1995).

The recent literature on the use of biomarkers as surrogate endpoints has focused on different points of view. Prentice (1989) defines surrogacy in terms of the equivalence of hypothesis tests for treatment effects and proposes operational criteria for his definition. Freedman *et al.* (1992) introduced the *proportion explained* (*PE*) to quantify the proportion of treatment effect on the true endpoint, which is captured by the surrogate endpoint. More recently Buyse *et al.* (2000), building on earlier work by Buyse & Molenberghs (1998), proposed a new definition of surrogacy. They distinguish between *trial-level* surrogacy, which characterizes the quality of prediction of the treatment effects at the trial level, and *individual-level* surrogacy, which measures the strength of association between the surrogate and the endpoint of interest after correction for trial and treatment effects.

The objective of this paper is to extend the methodology of Buyse *et al.* (2000) to the case of a biomarker, measured repeatedly over time, and a time-to-event endpoint. Technically, a joint model for longitudinal measurements and event time data is required. Research on this topic has received substantial attention over recent years and useful references include Pawitan & Self (1993), DeGruttola & Tu (1994), Taylor *et al.* (1994), Faucett & Thomas (1995), Lavalley & DeGruttola (1996), Hogan & Laird (1997), Wulfsohn & Tsiatis (1997), Henderson *et al.* (2000) and Xu & Zeger (2001). In this paper, the model of Henderson *et al.* (2000) will be adopted. Their approach assumes standard models for the longitudinal and event time data and postulates a latent bivariate Gaussian process inducing stochastic dependence between the measurement and event processes.

The paper is organized as follows: Section 2 introduces the motivating example which involves a set of two randomized clinical trials in advanced prostate cancer. Section 3 describes the methodology of Buyse *et al.* (2000) to validate a surrogate endpoint in a meta-analytic setting, while Section 4 shows how it can be adapted to the case of a longitunally measured marker and a time-to-event endpoint. The methodology is applied to the prostate cancer data in Section 5, which is then followed by a concluding discussion.

2 Motivating study

We consider a set of two open-label multicentre clinical trials in which patients with advanced prostate cancer were randomized either to oral liarozole—an experimental retinoic acid metabolism-blocking agent developed by Janssen Research Foundation—or to an antiandrogenic drug: cyproterone acetate (CPA) in the first trial (Debruyne *et al.*, 1998) and flutamide in the second. The two trials accrued 312 and 284 patients in centres spread over nine and ten countries, respectively. All patients were in relapse after first-line endocrine therapy.

The primary endpoint in each trial was survival time after randomization. Assessments were undertaken before the start of treatment and repeated at two weeks, monthly for six months and every three months thereafter, until patients show clinical progression or develop a serious adverse event. All patients were then followed up until death. The assessments included measurement of prostate-

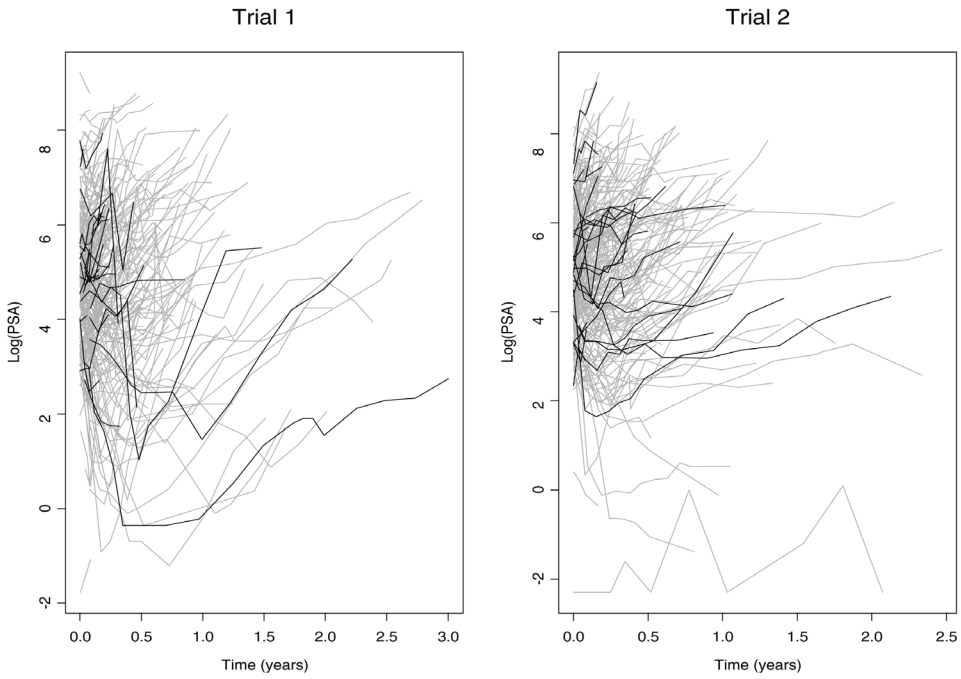

FIG 1. Individual log-transformed PSA profiles for the liarozole trials (30 randomly chosen subjects are plotted using darker lines).

specific antigen (PSA) level. PSA is a glycoprotein that is found almost exclusively in normal and neoplastic prostate cells. Serum PSA usually rise in men who have prostate cancer, but also with some infections of the prostate or non-malignant diseases, such as benign prostatic hyperplasia. As a consequence, changes in PSA often antedate changes in bone scan, and they have been used as a response indicator in patients with androgen-independent prostate cancer (Kelly *et al.*, 1993; Sridhara *et al.*, 1995; Smith *et al.*, 1998). It is therefore of interest to study more formally to which extent a sequence of PSA measurements can be a valuable surrogate for a patient's survival.

Figure 1 shows plots of the individual log-transformed PSA profiles. To avoid overly cluttered plots, profiles were shadowed and 30 randomly chosen subjects are depicted using darker lines. As can be seen, the length of the individual sequences of PSA measurements is highly variable accross patients, with only a few individuals having very long sequences. Figure 2 displays PSA and survival summaries for each trial. The (log-transformed) PSA data were smoothed using the LOESS technique (Cleveland, 1979); the survival curves were obtained using the Kaplan–Meier estimator (Kaplan & Meier, 1958). Notice the scatter of points in the left-hand plots: most of the subjects had their PSA measurements taken within the first few months after treatment randomization.

To further investigate the effect of 'drop-out' induced by patients being taken off the study upon clinical progression, we plotted the mean profiles per drop-out pattern according to visits as they were planned in the protocol (thus, not using the exact date of PSA measurement). This is shown in Figure 3 for the combined data from the two trials, with the label 'control' referring to CPA/flutamide and

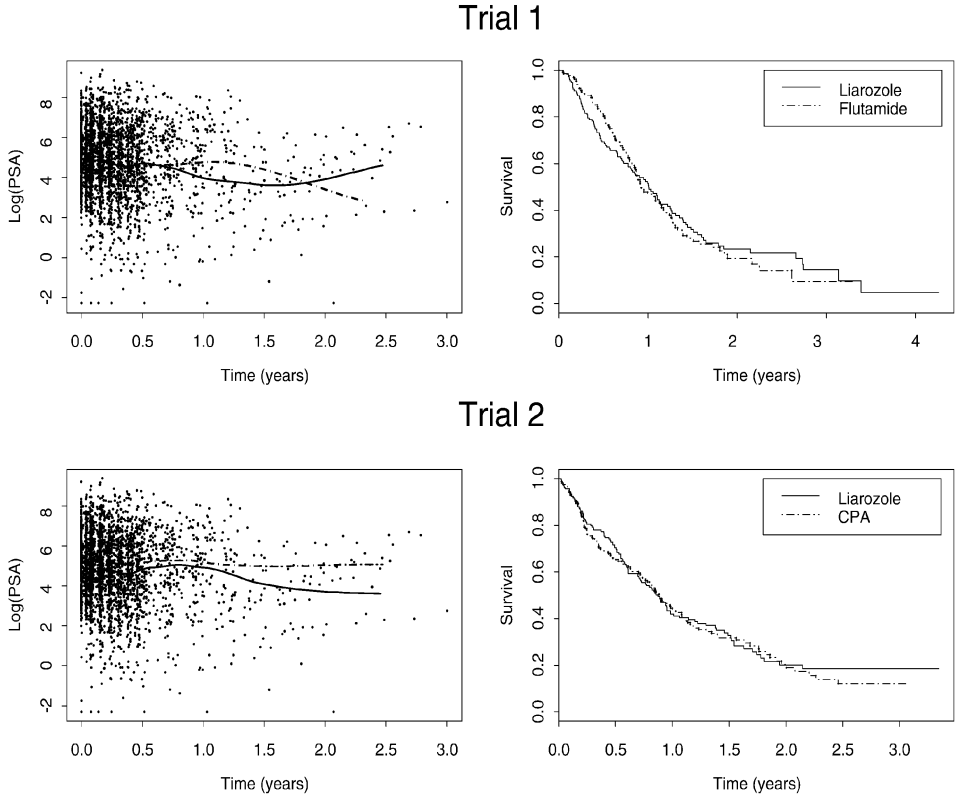

FIG 2. Longitidunal and event time summaries for the liarozole trials (left: smoothed PSA profiles; right: survival curves).

'experimental' to liarozole. Late-dropout patterns are not included in this plot because of the scarcity of data after 1.5 years. Noticeable in the plot is that: patients who progressed early tend to have a higher initial PSA value and do not exhibit an early decline in their PSA level. The mean PSA evolution among subjects who progressed belatedly can also be contrasted with the relatively flat curves displayed in Fig. 2.

3 Validation of a surrogate endpoint

In this section, we describe the two-stage model used by Buyse *et al.* (2000) to validate a surrogate endpoint, when both the surrogate and the true enpoints are assumed to be normally distributed. Refer to this paper for additional details.

The first stage is based upon a joint regression model for S and T:

$$\left.\begin{aligned} S_{ij}|Z_{ij} &= \mu_{S_i} + \alpha_i Z_{ij} + \varepsilon_{S_{ij}} \\ T_{ij}|Z_{ij} &= \mu_{T_i} + \beta_i Z_{ij} + \varepsilon_{T_{ij}} \end{aligned}\right\} \tag{1}$$

where the indices i and j refer to trials and subjects within trials, respectively, μ_{S_i} and μ_{T_i} are trial-specific intercepts, and α_i and β_i are trial-specific effects of treatment

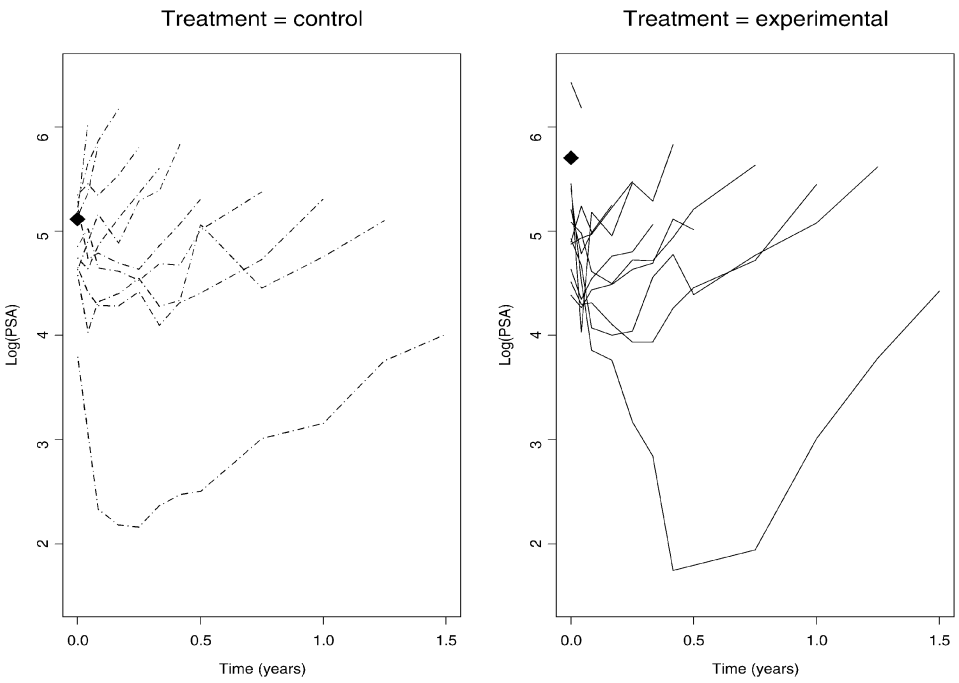

FIG 3. Mean PSA profiles per 'drop-out' patterns (the black diamonds represent the mean PSA level of those patients who only have a baseline measurement).

Z on the two endpoints in trial $i = 1, \ldots, N$. Finally, $\varepsilon_{S_{ij}}$ and $\varepsilon_{T_{ij}}$ are correlated error terms, assumed to be mean-zero normally distributed with covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ \sigma_{ST} & \sigma_{TT} \end{pmatrix} \tag{2}$$

At the second stage, we assume

$$\begin{pmatrix} \mu_{S_i} \\ \mu_{T_i} \\ \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} m_{S_i} \\ m_{T_i} \\ a_i \\ b_i \end{pmatrix} \tag{3}$$

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

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

For the developments that follow, one can start from model (1) and perform calculations in an additional step, or fit the more complex random-effects model obtained by combining the two steps above:

$$\left.\begin{aligned} S_{ij}|Z_{ij} &= \mu_S + m_{S_i} + (\alpha + a_i)Z_{ij} + \varepsilon_{S_{ij}} \\ T_{ij}|Z_{ij} &= \mu_T + m_{T_i} + (\beta + b_i)Z_{ij} + \varepsilon_{T_{ij}} \end{aligned}\right\} \tag{5}$$

Since both the individual- and trial-level associations are of interest here, the surrogate endpoint validation issue is examined at each of these levels. A key motivation for validating a surrogate endpoint is to be able to predict the effect of treatment on the true endpoint, based on the observed effect of treatment on the surrogate endpoint. It is therefore essential to explore the quality of the prediction of the treatment effect on the true endpoint by information obtained in the validation process based on trials $i = 1, \ldots, N$ and by information available on the surrogate endpoint in a new trial, $i = 0$, say.

A measure to assess the quality of a surrogate at the trial level is given by the coefficient of determination

$$R^2_{\text{trial}} = R^2_{b_i|m_{S_i},a_i} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}} \tag{6}$$

This coefficient measures how precisely the effect of treatment on the true endpoint can be predicted if the treatment effect on the surrogate endpoint has been observed in a new trial ($i = 0$). It is unitless and ranges in the unit interval if the corresponding variance–covariance matrix $\mathbf{D}$ is positive-definite—two desirable features for its interpretation.

The association between the surrogate and final endpoints after adjustment for the effect of treatment is captured by

$$R^2_{\text{indiv}} = \frac{\sigma^2_{ST}}{\sigma_{SS}\sigma_{TT}} \tag{7}$$

which is simply the squared correlation between S and T, after accounting for trial and treatment effects.

A surrogate endpoint will be said to be 'valid' if it is both trial-level valid ($R^2_{\text{trial}} \approx 1$) and individual-level valid ($R^2_{\text{indiv}} \approx 1$). Guidelines about how close R^2_{trial} and R^2_{indiv} have to be to 1 are hard to formulate in full generality. This will be based, preferably, upon expert opinion, and confidence limits for these coefficients should be examined.

As suggested above, an estimate of R^2_{trial} can practically be obtained in two ways: either by direct use of equation (6) after fitting model (5) or through the linear regression model

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

after fitting model (1). In either case, the association at the individual level (R^2_{indiv}) is a by-product of the joint model. Of course, inference will preferably be based on equation (5) but this model is difficult to fit in practice and does not go without numerous convergence problems (Buyse *et al.*, 2000). Finally, it can be noted that grouping units need not be trials from a meta-analysis but can denote any grouping unit of interest such as centre, investigator or country for example. In the following

we shall continue, however, to refer to the corresponding R^2 surrogacy coefficient as a 'trial'-level measure (R^2_{trial}).

4 Modelling approach

In order to extend the methodology of Buyse *et al.* (2000) to the setting of interest here, a joint model for longitudinal measurements and event time data is required. To that end, we consider the model proposed by Henderson *et al.* (2000). We will follow their notation and thus, we consider a set of N grouping units (trial, centre, etc) with subjects within the ith unit being followed for some time τ_i. The jth subject in unit i provides a set of measurements $\{y_{ijk}: k = 1, \ldots, n_{ij}\}$ at times $\{t_{ijk}: k = 1, \ldots, n_{ij}\}$, together with the realization of a counting process $\{N_{ij}(u): 0 \leqslant u \leqslant \tau_i\}$ for the time-to-event endpoint and a zero-one process $\{H_{ij}(u): 0 \leqslant u \leqslant \tau_i\}$ indicating whether a subject is at risk of experiencing an event at time u.

A central feature of the model is to postulate an unobserved (latent) zero-mean bivariate Gaussian process, $W_{ij}(t) = \{W_{1ij}(t), W_{2ij}(t)\}$, to describe the association between the longitudinal measurement and event processes. The measurement and intensity models are linked as follows:

(1) The sequence of measurements $\{y_{ijk}: k = 1, \ldots, n_{ij}\}$ of a subject is modelled using a standard linear mixed model, possibly allowing for a serially correlated component:

$$Y_{ijk} = \mu_{ij}(t_{ijk}) + W_{1ij}(t_{ijk}) + \varepsilon_{ijk} \tag{8a}$$

where $\mu_{ij}(t_{ijk})$ describes the mean response profile and $\varepsilon_{ijk} \sim N(0, \sigma^2_e)$ is a sequence of mutually independent measurement errors. We will let $\boldsymbol{\alpha}_i$ denote the vector of parameters for the trial-specific treatment effects used in modelling the mean response profile. Examples will be given in what follows.

(2) The event intensity process is modelled using a semi-parametric model

$$\lambda_{ij}(t) = H_{ij}(t)\lambda_0(t)\exp\{\beta_i Z_{ij} + W_{2ij}(t)\} \tag{8b}$$

where the form of $\lambda_0(t)$ is left unspecified. The parameters β_i represent trial-specific treatment effects on the hazard function.

The specification of W_{1ij} and W_{2ij} can take many different forms. As a basic example, suppressing the indices for notational simplicity, one could consider $W_1(t) = U_1 + U_2 t$, with (U_1, U_2) being normally distributed with mean zero and covariance matrix G, to specify a model with random intercept and random slope for the longitudinal marker. The $W_2(t)$ process could then include different effects for the initial value (U_1), the slope (U_2) or the current value ($U_1 + U_2 t$) of the marker according to the assumed model, yielding $W_2(t) = \gamma_1 U_1 + \gamma_2 U_2 + \gamma_3(U_1 + U_2 t)$. Inclusion of a frailty component, orthogonal to the measurement process, is also possible if necessary.

Following Henderson *et al.* (2000), the Expectation-Maximization (EM) algorithm can be employed to fit the model. Practically, the coefficients of determination R^2_{trial} and R^2_{indiv} can then be obtained as follows. The inclusion of (fixed) trial-specific coefficients in both the longitudinal measurement and intensity models allows to estimate R^2_{trial}. Unlike the simpler normal setting, which involves solely trial-specific intercepts and treatment effects, the longitudinal measurement model will require, in general, more terms to model the evolution of the marker over time. For practical purposes, we will therefore assume that the mean response

profile within each treatment group can be specified parsimoniously, as a low-order polynomial or as a continuous piecewise linear function of time. Alternatively, fractional polynomials (Royston & Altman, 1994). To illustrate the calculation of R^2_{trial}, suppose that the trajectory of the marker is quadratic over time within each treatment group. Then $\mu_{ij}(t_{ijk})$ can be written

$$\mu_{ij}(t_{ijk}) = \mu_{0i} + \mu_{1i}t_{ijk} + \mu_{2i}t^2_{ijk} + \alpha_{0i}Z_{ijk} + \alpha_{1i}Z_{ijk}t_{ijk} + \alpha_{2i}Z_{ijk}t^2_{ijk}$$

and R^2_{trial} defined as the usual coefficient of determination in the regression of $\hat{\beta}_i$ on $\hat{\alpha}_{0i}$, $\hat{\alpha}_{1i}$ and $\hat{\alpha}_{2i}$:

$$\hat{\beta}_i = \lambda_0 + \lambda_1\hat{\alpha}_{0i} + \lambda_2\hat{\alpha}_{1i} + \lambda_3\hat{\alpha}_{2i} + \varepsilon_i$$

At the individual level, it is natural to consider the association between $W_1(t)$ and $W_2(t)$ in the above model. Therefore, R^2_{indiv} will not refer directly to the association between the two endpoints but rather, to the association between the two components of the bivariate latent process that governs the longitudinal and event processes. This association can no longer be summarized by a single number, however. It will now be a time-dependent measure since the association between the marker and the event process can be defined relative to any time over the course of measurement of the marker. In fact, this could even be extended to the association between the marker, as measured at some time t_1, and the event process defined at a later time $t_2 \geqslant t_1$, thereby yielding a surface to describe the association between the longitudinal and event processes. This feature can be important in selecting an optimal time at which the marker should be evaluated, either to enhance clinical judgement or even further, to predict the event time of interest.

To illustrate the derivation of $R^2_{\text{indiv}}(t)$, we consider the aforementioned example with $W_1(t) = U_1 + U_2t$ and $W_2(t) = \gamma_1U_1 + \gamma_2U_2 + \gamma_3(U_1 + U_2t)$. The correlation between $W_1(t)$ and $W_2(t)$, for any fixed time t, is easily calculated since $W_1(t)$ and $W_2(t)$ have a joint normal distribution. Thus, if $(U_1,U_2) \sim N(0,G)$, we have:

$$\begin{aligned}
\text{var}[W_1(t)] &= G_{11} + 2G_{12}t + G_{22}t^2,\\
\text{var}[W_2(t)] &= (\gamma_1^2 + 2\gamma_1\gamma_3)G_{11} + 2(\gamma_1\gamma_2 + \gamma_1\gamma_3t + \gamma_2\gamma_3)G_{12}\\
&\quad + (\gamma_2^2 + 2\gamma_2\gamma_3t)G_{22} + \gamma_3^2\text{var}[W_1(t)]\\
\text{covar}[W_1(t), W_2(t)] &= \gamma_1G_{11} + (\gamma_2 + \gamma_1t)G_{12} + \gamma_2G_{22}t + \gamma_3\text{var}[W_1(t)]
\end{aligned}$$

from which the (squared) correlation between $W_1(t)$ and $W_2(t)$ can be easily derived by plugging in estimates for γ_1, γ_2, G_{11}, G_{12} and G_{22}. This function, which will be termed 'model-based', is entirely based on the assumptions made in our model. A more heuristic estimate, which we will refer to as 'empirical', could be derived along the same lines of development, except that sample estimators based on the estimated U values obtained at the (final) E-step of the EM algorithm are substituted for the elements of G. Thus, G_{11} is replaced by $\widehat{\text{var}}\{\hat{U}_{1i}\}$, G_{22} by $\widehat{\text{var}}\{\hat{U}_{2i}\}$ and G_{12} by $\widehat{\text{covar}}\{\hat{U}_{1i}, \hat{U}_{2i}\}$.

It should be stressed that the so-obtained curve is still strongly dependent on some aspects of the model. For example, should we assume that $W_2(t) = \gamma W_1(t)$, then $R^2_{\text{indiv}}(t) \equiv 1$. As one departs from this simple model and further terms are added, a finer characterization of the curve is allowed in its admissible forms. Because of this, we recommend including a sufficiently large number of association parameters $\{\gamma_k\}$ in the model.

5 Application to the advanced prostate cancer data

In this section, we aim at applying the proposed methodology to the liarozole data introduced in Section 2. We will utilize pooled data from the two trials and will refer to the control and experimental arms as in Fig. 3. Since our methodology requires the estimation of the treatment effects in multiple trials or other meaningful groups of patients, we will use country as a grouping unit within each trial in order to have a sufficient number of patients in each unit. This enables us to define 19 groups containing between 3 and 69 patients per group. For the analysis, however, two of these groups had to be excluded: in one of them ($n = 3$), subjects were accrued in only one treatment arm and no events were observed in the second ($n = 8$).

A first step in the analysis is to specify a parsimonious model that captures the time evolution of the marker within each treatment group. A simplistic attempt could involve second-order polynomials. While this choice may, at first, seem odd after inspection of the average profiles (Fig. 2), this is more in agreement with what Fig. 3 suggested. This was also confirmed by a likelihood ratio test, as the introduction of a quadratic term in the model (as a fixed and random coefficient) yields a large drop in deviance.

As a possible refinement, we can employ fractional polynomials (Royston & Altman, 1994) to characterize the time evolution of the marker more suitably. Starting from the set of powers ranging from -2 to 2 by step of 0.5 and assuming a model with individual-level random effects for the intercept, t and t^2, we select the best-fitting pair of powers in this set, allowing for treatment-specific curves. This approach leads to the selection of terms t and $\sqrt{t}$ and as our final model, we further consider individual-level random effects for t and $\sqrt{t}$ instead of t and t^2 (note: comparison of this final model with the original one also yields a large drop in deviance).

The joint model to be fitted can therefore be written as follows:

$$\begin{aligned} Y_{ijk} = {} & \mu_{0i} + \mu_{1i} t_{ijk} + \mu_{2i} \sqrt{t_{ijk}} + \alpha_{0i} Z_{ij} + \alpha_{1i} Z_{ij} t_{ijk} + \alpha_{2i} Z_{ij} \sqrt{t_{ijk}} \\ & + U_{0j} + U_{1j} t_{ijk} + U_{2j} \sqrt{t_{ijk}} + \varepsilon_{ijk} \end{aligned} \tag{9a}$$

and

$$\lambda_{ij}(t) = \lambda_0(t) \exp\{\beta_i Z_{ij} + \gamma_0 U_{0j} + \gamma_1 U_{1j} + \gamma_2 U_{2j} + \gamma_3 (U_{0j} + U_{1j} t + U_{2j} \sqrt{t})\} \tag{9b}$$

with i denoting country (within trial), j referring to individual patients and k to measurement occasions.

As explained in Section 4, R^2_{trial} can be calculated as the coefficient of determination in the regression of $\{\hat{\beta}_i\}$ on $\hat{\boldsymbol{\alpha}}_i = \{\hat{\alpha}_{0i}, \hat{\alpha}_{1i}, \hat{\alpha}_{2i}\}$, which yields a value of 0.517. This mid-range value is probably too low to permit reliable prediction of treatment effects on survival, having observed the effect of treatment on the marker. Confidence limits on R^2_{trial} can be obtained based on the assumption that $\boldsymbol{\alpha}_i$ and β_i are normally distributed (as in equation (3)), from which the distribution of the coefficient of determination can be derived (Algina, 1999; Ding, 1996). More specifically, a 95% confidence interval can be obtained by finding values of R^2_{trial} for which the corresponding estimates are approximately equal to the 2.5% and 97.5% quantiles of the cumulative distribution function of R^2. In our example, the so-obtained confidence limits for R^2_{trial} are [0.013, 0.748], thus showing that the trial-level association is estimated rather imprecisely.

It is interesting to notice that dependence between the marker and the time-to-event endpoint is a complicating assumption in our framework. If one is interested solely in trial-level surrogacy, a naive computational approach could involve independent models for each endpoint, thereby ignoring dependence between the two endpoints and simplifying model fitting. Tibaldi *et al.* (2001) explore this issue in the situation of Section 3 (normally distributed endpoints) and conclude that this simplified approach seems to perform reasonably well. Obviously, as one departs from the multivariate Gaussian world, it is not at all clear that such a simplistic approach would work effectively well. For comparative purposes, we calculated R^2_{trial} by fitting, separately models, model (9a) and (9b) with $\gamma_0 = \gamma_1 = \gamma_2 = \gamma_3 = 0$. This results in a value of $R^2_{\text{trial}} = 0.291$ that is much lower than the one found above. (Concidence limits should not be overlooked however.) Thus, we see that ignoring dependence between the marker and the survival endpoint might give misleading inference on R^2_{trial} in this setting, although this issue should be further explored.

Figure 4(a) shows the model-based and empirical curves $R^2_{\in d}(t)$ for model (9a)–(9b). Both curves agree fairly well over the time range considered. They start from a relatively low level (~ 0.3), then rise sharply until a value of about 0.9 at year 1 and stabilize at that level thereafter. Although the interpretation of such a plot holds, strictly speaking, at the level of the latent processes $W_1(t)$ and $W_2(t)$, this would suggest that, initially, PSA level bears relatively little information on a patient's future survival, but as information on the marker is gathered over time (first year of treatment), it achieves a better predictive strength for survival, with no further gain in subsequent years. For comparison purposes, the plot in the right-hand panel (Fig. 4(b)) shows the same curves under the model with quadratic

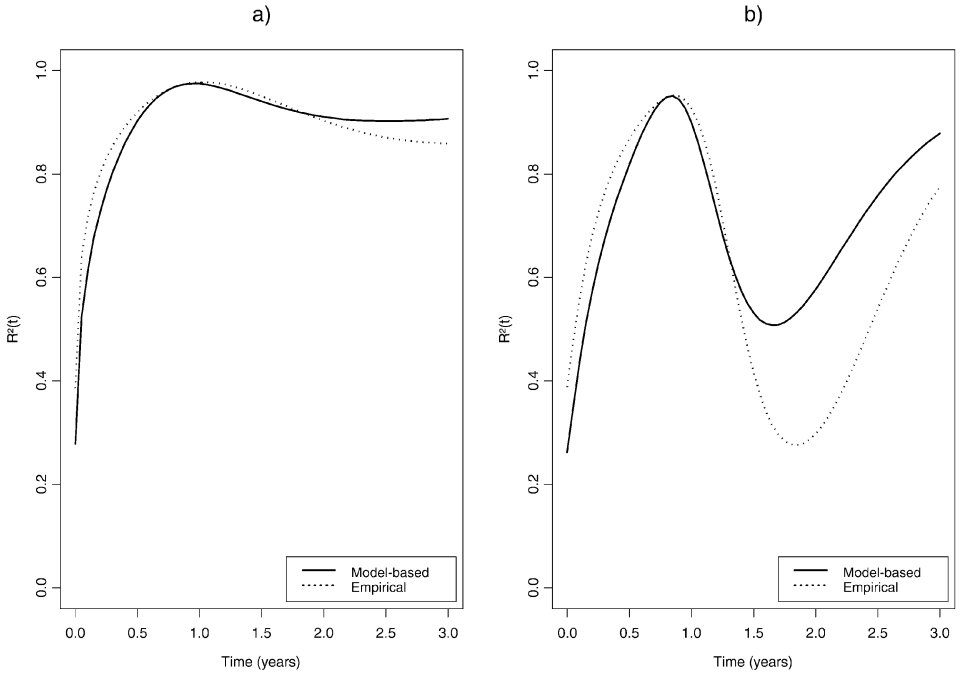

FIG 4. Plots of the model-based and empirical $R^2_{\text{indiv}}(t)$ curves. Left panel: final model (int., t, $\sqrt{t}$). Right panel: original model (int., t, t^2).

time evolution and individual-level random effects for the intercept, t and t^2. The two curves show a similar behaviour within the first year, but then a dip can be observed. In addition, it can be noticed that the two curves do not coincide so well. It is not clear whether this is caused by the inferior fit of the model, or by constraints imposed by the model itself, but this calls for caution when interpreting such curves. We do believe that they might shed some light on the basic intricacies between the marker and the survival endpoint of interest, but these curves should not be over-interpreted as they may be strongly model-dependent.

6 Discussion

An extension of the surrogate endpoint validation methodology of Buyse *et al.* (2000) was proposed for the case where a longitudinally measured biomarker is a potential surrogate for a survival endpoint. To that end, the formulation of Henderson *et al.* (2000) was adopted for the joint model relating the marker and the survival time. A limiting feature arises from the inherent complexity of the joint modelling of longitudinal measurements and event time data, which is most noticeable in the computational aspect of this approach. In particular, intensive computing times can be expected in the type of applications covered by the present paper because of the typically high sample size of the meta-analytic data sets required for our validation exercise. In addition, use of the EM algorithm to fit the model fails to provide precision estimates for parameters. In their paper, Henderson *et al.* obtained standard errors by a Monte-Carlo method refitting the model to simulated data sets generated using parameter values taken from the original analysis. Clearly, this procedure may be overly time-consuming here, unless one has a powerful computer at one's disposal.

At the trial level, which will be mostly the level of interest in practice, the surrogacy measure R^2_{trial} can be easily derived by considering extra terms needed to characterize the longitudinal evolution of the marker, and our method provides point estimates and uncertainty measures for this parameter. In addition, the individual-level surrogacy can be explored through the function $R^2{}_{\text{indiv}}(t)$, which captures the association induced by the two underlying Gaussian processes, $W_1(t)$ and $W_2(t)$, used in our joint model. Since the latter quantity is primarily of interest for exploratory purposes, and since calculating precision estimates within the joint model is cumbersome, we do not attach uncertainty measures on $R^2_{\text{indiv}}(t)$ here. Note that when such a step is done, it could also help to incorporate the measurement error introduced by the fact that estimates of the $\boldsymbol{\alpha}_i$s and the β_is are effectively employed when estimating R^2_{trial}.

Finally, it would be desirable to investigate model adequacy better (with an application to $R^2_{\text{indiv}}(t)$ in mind, for example) or the diagnostic assessment of fitted models. Unfortunately, such tools are currently lacking and this is an area for further research, as pointed out by Henderson *et al.* (2000).

As for the clinical interpretation of our work, we saw that PSA level and survival seem, as expected, to be strongly related, at least when a sufficiently large amount of information has been gathered on the marker. While bearing in mind that R^2_{trial} was estimated with large uncertainty, the value that was found stands in the mid-range of the unit interval and would prevent us from formulating any firm conclusion had it been estimated more precisely. This points to an issue, not of the methodology, but rather of the biological nature of the marker. We may tentatively say, however, that PSA level has some value as a surrogate marker for

survival (for the class of treatments considered in the two trials at least) but probably is not a very good one. Obviously, these results should be taken with caution since this study involved only a couple of clinical trials with a relatively limited number of subjects. The issue of validiting a surrogate marker will, ideally, be based on a much more extended set of randomized trials and will cover different classes of therapies that are commonly used for treating patients with the disease in question.

Acknowledgement

The first and fourth authors gratefully acknowledge support from an LUC Bijzonder Onderzoeksfonds grant. The second author was supported by the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), Belgium. The authors are grateful to Johnson and Johnson (Pharmaceutical Research and Development) for the kind permission to use their data. The authors would also like to thank Angela Dobson for her kind assistance and help in implementing the model used in the paper.

REFERENCES

ALGINA, J. (1999) A comparison of methods for constructing confidence intervals for the squared multiple correlation coefficient, *Multivariate Behavioral Research*, 34, pp. 494–504.

BUYSE, M. & MOLENBERGHS, G. (1998) CRITERIA FOR THE VALIDATION OF SURROGATE ENDPOINTS IN RANDOMIZED EXPERIMENTS, *Biometrics*, 54, pp. 1014–1029.

BUYSE, M., MOLENBERGHS, G., BURZYKOWSKI, T., RENARD, D. & GEYS, H. (2000) The validation of surrogate endpoints in meta-analyses of randomized experiments, *Biostatistics*, 1, pp. 49–67.

CLEVELAND, W. S. (1979) Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association*, 74, pp. 829–836.

DEBRUYNE, F. J. M., MURRAY, R., FRADET, Y., JOHANSSON, J. E., TYRRELL, C., BOCCARDO, F. *et al.* (1998) Liarozole—a novel treatment approach for advanced prostate cancer: results of a large randomized trial versus cyproterone acetate, *Urology*, 52, pp. 72–81.

DEGRUTTOLA, V. & TU, X. M. (1994) Modelling progression of CD4-lymphocyte count and its relationship to survival time, *Biometrics*, 50, pp. 1003–1014.

DING, C. G. (1996) On the computation of the distribution of the squared multiple correlation coefficient, *Computational Statistics and Data Analysis*, 22, pp. 345–350.

FAUCETT, C. L. & THOMAS, D. C. (1995) Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach, *Statistics in Medicine*, 15, pp. 1663–1685.

FREEDMAN, L. S., GRAUBARD, B. I. & SCHATZKIN, A. (1992) Statistical validation of intermediate endpoints for chronic diseases, *Statistics in Medicine*, 11, pp. 167–178.

HENDERSON R., DIGGLE, P. & DOBSON, A. (2000) Joint modelling of longitudinal measurements and event time data, *Biostatistics*, 1, pp. 465–480.

HOGAN, J. W. & LAIRD, N. H. (1997) Mixture models for the joint distribution of related measures and event times, *Statistics in Medicine*, 16, pp. 239–257.

KAPLAN, E. L. & MEIER, P. (1958) Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, 53, pp. 457–481

KELLY, W. K., SCHER, H. I., MAZUMDAR, M. *et al.* (1993) Prostate-specific antigen as a measure of disease outcome in metastatic hormone-refractory prostate cancer, *Journal of Clinical Oncology*, 11, pp. 607–615.

LAVALLEY, M. P. & DEGRUTTOLA, V. (1996) Models for empirical Bayes estimators of longitudinal CD4 counts, *Statistics in Medicine*, 15, pp. 2289–2305.

PAWITAN, Y. & SELF, S. (1993) Modelling disease marker processes in AIDS, *Journal of the American Statistical Association*, 88, pp. 719–726

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

RAMLAU-HANSEN, H. (1983) Smoothing counting process intensities by means of kernel functions, *Annals of Statistics*, 11, pp. 453–466

ROYSTON, P. & ALTMAN, D. G. (1994) Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling, *Applied Statistics*, 43, pp. 429–467.

SMITH, D., DUNN, R. L., STAWDERMAN, M. S. & PIENTA, K. J. (1998) Change in serum prostate-specific antigen as a marker of response to cytotoxic therapy for hormone-refractory prostate cancer, *Journal of Clinical Oncology*, 16, pp. 1835–1843.

SRIDHARA, R., EISENBERGER, M. A., SINIBALDI, V. J. *et al.* (1995) Evaluation of prostate-specific antigen as a surrogate marker for response of hormone-refractory prostate cancer to suramin therapy, *Journal of Clinical Oncology*, 13, pp. 2944–2953.

TAYLOR, J. M. G., CUMBERLAND, W. G. & SY, J. P. (1994) A stochastic model for analysis of longitudinal AIDS data, *Journal of the American Statistical Association*, 89, pp. 727–736.

TIBALDI, F., CORTIÑAS ABRAHANTES, J., MOLENBERGHS, G., RENARD, D., BURZYKOWSKI, T., BUYSE, M., PARMAR, M., STIJNEN, T. & WOLFINGER, R. (2001) Computational approaches to the evaluation of surrogate endpoints.

TSIATIS, A. A., DEGRUTTOLA, V. & WULFSOHN, M. S. (1995) Modelling the relationship of survival to longitudinal data measured with error, *Journal of the American Statistical Association*, 90, pp. 27–37.

WULFSOHN, M. S. & TSIATIS, A. A. (1997) A joint model for survival and longitudinal data measured with error, *Biometrics*, 53, pp. 330–339.

XU, J. & ZEGER, S. (2001) Joint analysis of longitudinal data comprising repeated measures and times to events, *Applied Statistics*, 50, pp. 375–387.