Appl. Statist. (2001)
50, *Part* 4, *pp.* 405–422

Validation of surrogate end points in multiple randomized clinical trials with failure time end points

Tomasz Burzykowski and Geert Molenberghs,

Limburgs Universitair Centrum, Diepenbeek, Belgium

Marc Buyse

International Drug Development Institute, Brussels, Belgium

and Helena Geys and Didier Renard

Limburgs Universitair Centrum, Diepenbeek, Belgium

[Received November 1999. Final revision March 2001]

Summary. Before a surrogate end point can replace a final (true) end point in the evaluation of an experimental treatment, it must be formally 'validated'. The validation will typically require large numbers of observations. It is therefore useful to consider situations in which data are available from several randomized experiments. For two normally distributed end points Buyse and co-workers suggested a new definition of validity in terms of the quality of both trial level and individual level associations between the surrogate and true end points. This paper extends this approach to the important case of two failure time end points, using bivariate survival modelling. The method is illustrated by using two actual sets of data from cancer clinical trials.

Keywords: Copula model; Failure time end point; Meta-analysis; Surrogate end point; Validation

1. Introduction

Surrogate end points are referred to as end points that can be used in lieu of other end points in the evaluation of experimental treatments or other interventions. They are useful when they can be measured earlier, more conveniently or more frequently than the end points of interest, which are referred to as the 'true' or 'final' end points (Ellenberg and Hamilton, 1989). Before a surrogate end point can replace a final end point in the evaluation of an experimental treatment, it must be formally 'validated', a process that has caused much controversy and has not been fully elucidated so far.

Prentice (1989) proposed a formal definition of surrogate end points and outlined how potential surrogate end points could be validated. Much debate ensued, for the criteria set out by Prentice are too stringent and are not straightforward to verify (Fleming *et al.*, 1994). Freedman *et al.* (1992) took Prentice's approach one step further by introducing the *proportion explained*, which is the proportion of the treatment effect that is mediated by the surrogate. This proposal is itself surrounded with difficulties, the most important being that it is not confined to the unit interval (Flandre and Saidi, 1999; Molenberghs *et al.*, 2000). Buyse

Address for correspondence: Tomasz Burzykowski, Centre for Statistics, Limburgs Universitair Centrum, Building D, Universitaire Campus, B3590 Diepenbeek, Belgium.
E-mail: tomasz.burzykowski@luc.ac.be

 0035–9254/01/50405

and Molenberghs (1998) proposed to replace the proportion explained by two new measures. The first, defined at the population level and termed the *relative effect*, is the ratio of the overall treatment effect on the true end point over that on the surrogate end point. The second is the individual level association between both end points, after accounting for the effect of treatment, and referred to as *adjusted association*.

To be informative and of practical value, however, the validation of a surrogate end point will typically require large numbers of observations. It is therefore useful to consider situations in which data are available from several randomized experiments, where the experimental unit can be the centre in a multicentric trial or the trial in a meta-analysis of several trials. For two normally distributed end points Buyse *et al.* (2000) suggested a new definition of validity in terms of the quality of both trial level and individual level associations between the surrogate and true end points. From a modelling standpoint, a two-stage model is required that can be fitted by using a fixed or random-effects representation. Standard software for linear mixed models (Verbeke and Molenberghs, 1997) or multilevel models (Goldstein, 1995) can be utilized for this purpose.

As Buyse *et al.* (2000) centred solely on the case of normally distributed end points, it is necessary to explore other settings, which are often more complicated owing to the absence of a unifying framework such as the multivariate normal distribution. In this paper, we concentrate on the important case when both the surrogate and the true end points are failure time variables. Such a setting is commonly encountered, for instance, in oncology, where the time to progression or progression-free survival time is frequently used as a surrogate for the survival time (Chen *et al.*, 1998).

Our notation and motivating studies are presented in Section 2. Section 3 summarizes the method of validation proposed by Buyse *et al.* (2000) for two normally distributed end points. Section 4 describes a proposed extension of the method to two failure time end points. The examples are analysed in Section 5. Section 6 briefly discusses the merits and limitations of the extension proposed and suggests directions for future research.

The programs which are used to analyse the data can be obtained from

`http://www.blackwellpublishers.co.uk/rss/`

2. Notation and motivating studies

Suppose that we have data from $i = 1, \ldots, N$ trials, in the ith of which $j = 1, \ldots, n_i$ subjects are enrolled. Let T_{ij} and S_{ij} be random variables that denote the true and surrogate end points respectively, and let Z_{ij} be an indicator variable for treatment.

2.1. A meta-analysis of advanced ovarian cancer trials

Our methods will first be applied to data from a meta-analysis of four randomized multicentre trials in advanced ovarian cancer (Ovarian Cancer Meta-Analysis Project, 1991). Individual patient data are available in these four trials for the comparison of two treatment modalities: cyclophosphamide plus cisplatin (CP) *versus* cyclophosphamide plus adriamycin plus cisplatin (CAP). The binary indicator for treatment (Z_{ij}) will be set to 0 for treatment CP and to 1 for treatment CAP. The surrogate end point S_{ij} will be the progression-free survival time, defined as the time (in years) from randomization to clinical progression of the disease or death, whereas the final end point T_{ij} will be the survival time, defined as the time (in years) from randomization to death from any cause. The full results of this meta-analysis were published with a minimum follow-up of 5 years in all trials (Ovarian Cancer Meta-Analysis

Project, 1991). The data set was subsequently updated to include a minimum follow-up of 10 years in all trials (Ovarian Cancer Meta-Analysis Project, 1998). After such a long follow-up, disease progression or death has occurred for most patients (952 of 1194 patients, i.e. 80%).

The method proposed can be applied as soon as there is replication not only at the patient level but also at minimally one hierarchically higher level, such as centre within trial or trial within meta-analysis. Technically, the ovarian cancer case is a meta-analysis but it contains only four trials. In the two larger trials, information is also available on the centres in which the patients had been treated. For the two smaller studies, by the Danish Ovarian Cancer Group (DACOVA) and the Gruppo Oncologico Nord-Ovest (GONO), this information is not available; in these studies the investigators argued that the proximity and close co-operation of the centres enable us to consider the enrolled patients as essentially treated in one institution. It is thus natural to use the centre as the unit of analysis for the two larger trials, and the trial as the unit of analysis for the two smaller trials. A total of 50 'units' are then available for analysis, with a number of individual patients per unit ranging from 2 to 274. The replication at the level of the centre is thus sufficient to apply the meta-analytic methods.

2.2. A study of two advanced colorectal cancer trials

As a second case-study, we shall use data from two randomized multicentre trials in advanced colorectal cancer (Corfu-A Study Group, 1995; Greco *et al.*, 1996). In one trial, treatment with fluorouracil (5FU) plus interferon (5FU–IFN) was compared with treatment with 5FU plus folinic acid (5FU–LV) (Corfu-A Study Group, 1995). In the other trial, treatment with 5FU–IFN was compared with treatment with 5FU alone (Greco *et al.*, 1996). The binary indicator for treatment (Z_{ij}) will be set to 0 for 5FU–IFN and to 1 for 5FU–LV or 5FU alone. The surrogate end point S_{ij} will be the progression-free survival time, defined as the time (in years) from randomization to clinical progression of the disease or death, whereas the final end point T_{ij} will be the survival time, defined as the time (in years) from randomization to death from any cause. Disease progression or death had occurred for most patients in the two trials (694 of 736 patients, i.e. 94.3%).

Similarly to the ovarian cancer example, we shall use the centre as the unit of analysis. A total of 76 'units' are thus available for analysis. However, in eight centres one of the treatment arms accrued no patients. These eight centres were therefore excluded from the analysis. As a result, the data used for illustration contained 68 units, with the number of individual patients per unit ranging from 2 to 38.

3. Two normally distributed end points

3.1. The two-stage model

In this section, we describe the two-stage model which is the core of the method proposed by Buyse *et al.* (2000) for two normally distributed end points.

The first stage is based on a fixed effects model:

$$S_{ij}|Z_{ij} = \mu_{Si} + \alpha_i Z_{ij} + \epsilon_{Sij}, \tag{1}$$

$$T_{ij}|Z_{ij} = \mu_{Ti} + \beta_i Z_{ij} + \epsilon_{Tij}, \tag{2}$$

where μ_{Si} and μ_{Ti} are trial-specific intercepts and α_i and β_i are trial-specific effects of treatment Z on the end points in trial i. Finally, ϵ_{Sij} and ϵ_{Tij} are correlated error terms, assumed to be mean 0 normally distributed with covariance matrix

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

At the second stage, it is assumed that

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

where the second term on the right-hand side of equation (3) follows a zero-mean normal distribution with dispersion matrix

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

3.2. Trial level surrogacy

Suppose then that the new trial $i = 0$ is considered for which data are available on the surrogate end point but not on the true end point. We are interested in the estimated effect of Z_{0j} on T_{0j}, given the effect of Z_{0j} on S_{0j}. It can be shown that $(\beta + b_0 | m_{S0}, a_0)$ follows a normal distribution with mean and variance

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

Consequently, a measure to assess the quality of the surrogate at the trial level is the coefficient of determination

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

A surrogate could be called *perfect at the trial level* if the coefficient of determination (5) were equal to 1. Intuition can be gained by considering the special case where the prediction of b_0 can be done independently of the random intercept m_{S0}. The coefficient (5) then reduces to

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

It is simply the square of the correlation between α_i and β_i. Now, $R^2_{\mathrm{trial(r)}} = 1$ if the trial level treatment effects are simply multiples of each other. We shall refer to this simplified version as the *reduced* random-effects model, whereas the original expression (5) will be said to derive from the *full* random-effects model.

An estimate for $\beta + b_0$ is obtained by replacing the right-hand side of equation (4) with the corresponding parameter estimates. A confidence interval is obtained by applying the delta method to equation (4). The covariance matrix of the parameters involved is obtained from

the meta-analysis, except for μ_{S0} and α_0, which are obtained from fitting model (1) to the data for the surrogate end point in the new trial. The corresponding prediction interval is found by adding equation (5) to the variance obtained for the confidence interval. Although it can be of interest to study the performance of delta-type intervals and to propose alternatives if necessary, this is outside the scope of the current paper.

3.3. Individual level surrogacy

To validate a surrogate end point, Buyse *et al.* (2000) suggested that we consider the association between the surrogate and the final end point after adjusting the marginal models (1) and (2) for the treatment effect. From model (1)–(2) it follows that the conditional distribution of T_{ij}, given S_{ij} and Z_{ij}, is

$$T_{ij}|Z_{ij},\ S_{ij} \sim N\{\mu_{Ti} - \sigma_{TS}\sigma_{SS}^{-1}\mu_{Si} + (\beta_i - \sigma_{TS}\sigma_{SS}^{-1}\alpha_i)Z_{ij} + \sigma_{TS}\sigma_{SS}^{-1}S_{ij};\ \sigma_{TT} - \sigma_{TS}^2\sigma_{SS}^{-1}\}. \quad (7)$$

The association between both end points after adjustment for both the trial effects and the treatment effect in distribution (7) is captured by

$$R^2_{\text{indiv}} = R^2_{\epsilon_{Tij}|\epsilon_{Sij}} = \sigma_{ST}^2/\sigma_{SS}\sigma_{TT}, \quad (8)$$

the squared correlation between 'adjusted' variables $S_{ij} - (\mu_{Si} + \alpha_i Z_{ij})$ and $T_{ij} - (\mu_{Ti} + \beta_i Z_{ij})$.

On the basis of the development in Sections 3.2 and 3.3, Buyse *et al.* (2000) suggested that we term a surrogate *trial level valid* if $R^2_{\text{trial(f)}}$ (or $R^2_{\text{trial(r)}}$) is sufficiently close to 1 and call it *individual level valid* if R^2_{indiv} is sufficiently close to 1. A surrogate might be termed *valid* if it were both trial level and individual level valid. The notion of 'sufficiently close to 1' will be discussed in the examples.

4. Two failure time end points

4.1. The two-stage model

Assume now that S_{ij} and T_{ij} are failure time end points. To extend the approach used in the case of two normally distributed end points that was described in Section 3, model (1)–(2) might be replaced by a model for two correlated failure time random variables. An important requirement is that the model should provide a measurement of association between the two failure time variables. There are several classes of model that might be considered: copula models (Genest and McKay, 1986; Shih and Louis, 1995; Nelsen, 1999); univariate (Hougaard, 1995; Anderson, 1995) or bivariate (Xue and Brookmeyer, 1996) proportional frailty models; scale change models (Anderson, 1995); marginal models estimated using generalized estimation equations (Prentice and Hsu, 1997). We propose to use copula models as they offer greater flexibility than the other models (with the exception of the bivariate frailty model proposed by Xue and Brookmeyer (1996), the use of which will be discussed in Section 6). In particular, they include univariate proportional frailty models as a subclass (Oakes, 1989). Moreover, univariate proportional frailty models and scale change models generally induce a non-negative association between the two failure time variables (Anderson, 1995), whereas copula models in principle do not suffer from this limitation. Finally, though based on the generalized estimating equations approach, the models proposed by Prentice and Hsu (1997) require the use of a parametric bivariate survivor function and in that paper they in fact used copula models.

Thus, to replace model (1)–(2), we assume that the joint survivor function of $(S_{ij},\ T_{ij})$ can be written as

$$F(s, t) = P(S_{ij} \geqslant s, T_{ij} \geqslant t) = C_\delta\{F_{Sij}(s), F_{Tij}(t)\}, \qquad s, t \geqslant 0, \tag{9}$$

where (F_{Sij}, F_{Tij}) denotes marginal survivor functions and C_δ is a distribution function on $[0, 1]^2$ with $\delta \in R^1$. C_δ is called a *copula function* (Genest and McKay, 1986; Shih and Louis, 1995; Nelsen, 1999). It describes association between S_{ij} and T_{ij}. An attractive feature of model (9) is that the margins do not depend on the choice of the copula function.

To model the effect of treatment on the marginal distributions of S_{ij} and T_{ij} in equation (9) we propose to use the proportional hazard model:

$$F_{Sij}(s) = \exp\left\{-\int_0^s \lambda_{Si}(x)\, \exp(\alpha_i Z_{ij})\, \mathrm{d}x\right\}, \tag{10}$$

$$F_{Tij}(t) = \exp\left\{-\int_0^t \lambda_{Ti}(x)\, \exp(\beta_i Z_{ij})\, \mathrm{d}x\right\}, \tag{11}$$

where λ_{Si} and λ_{Ti} are trial-specific marginal base-line hazard functions and α_i and β_i are trial-specific effects of treatment Z on the end points in trial i. A version of model (10)–(11) with common (across trials) base-line hazard functions can also be considered. The hazard functions can be specified parametrically or can be left unspecified as in the classical model proposed by Cox (1972). When the hazard functions are specified, estimates of the parameters for the joint model (9) and (10)–(11) can be obtained by using the maximum likelihood method. Alternatively, the two-stage parametric procedure proposed by Shih and Louis (1995) can be used, in which parameters of the marginal survivor functions F_{Sij} and F_{Tij} are estimated first (assuming independence), and then δ is estimated conditionally on the estimated values of the marginal parameters. When the hazard functions are left unspecified, a two-stage semiparametric procedure of Shih and Louis (1995), similar to the parametric version described above, can be applied.

If the copula function in equation (9) can be represented as

$$C_\delta(u, v) = \phi_\delta\{\phi_\delta^{-1}(u) + \phi_\delta^{-1}(v)\}, \qquad 0 \leqslant u, v \leqslant 1$$

where ϕ_δ is a Laplace transform of some distribution, then model (9) reduces to a proportional frailty model (Oakes, 1989). In the case-studies, the following special cases of a proportional frailty model are considered: the model proposed by Clayton (1978) and the model proposed by Hougaard (1986).

In Clayton's model the copula function has the form

$$C_\delta(u, v) = (u^{1-\delta} + v^{1-\delta} - 1)^{1/(1-\delta)}, \qquad \delta > 1. \tag{12}$$

It is generated by the Laplace transform $\phi_\delta(x) = (1 + x)^{1/(1-\delta)}$ of a gamma distribution with density

$$f(x) = \frac{x^{1/(\delta-1)-1}\, \exp(-x)}{\Gamma\{1/(\delta - 1)\}}.$$

S_{ij} and T_{ij} are positively associated when $\delta > 1$ and are independent when $\delta \to 1$.

In Hougaard's model the copula function has the form

$$C_\delta(u, v) = \exp(-[\{-\ln(u)\}^{1/\delta} + \{-\ln(v)\}^{1/\delta}]^\delta), \qquad 0 < \delta < 1. \tag{13}$$

It is generated by the Laplace transform $\phi_\delta(x) = \exp(-x^\delta)$ of a positive stable distribution with density (Hougaard, 1986)

$$f(x) = -\frac{1}{\pi x}\sum_{k=1}^{\infty}\frac{\Gamma(k\delta+1)}{k!}(-x^{-\delta})^k \sin(\delta k\pi).$$

S_{ij} and T_{ij} are positively associated when δ is small and are independent when $\delta \to 1$.

At the second stage, we propose to use the reduced random-effects model

$$\begin{pmatrix}\alpha_i\\ \beta_i\end{pmatrix} = \begin{pmatrix}\alpha\\ \beta\end{pmatrix} + \begin{pmatrix}a_i\\ b_i\end{pmatrix} \tag{14}$$

where the second term on the right-hand side of equation (14) is assumed to follow a zero-mean normal distribution with dispersion matrix

$$D = \begin{pmatrix}d_{aa} & d_{ab}\\ d_{ab} & d_{bb}\end{pmatrix}. \tag{15}$$

4.2. *Validation criteria*

Since the reduced random-effects model (14) is used at the second stage of the two-stage model, the quality of surrogate S at the trial level will be assessed on the basis of the coefficient of determination (6).

From the considerations presented in Section 3.3 it follows that, to assess the quality of the surrogate at the individual level, a measure of association between S_{ij} and T_{ij}, calculated while adjusting the marginal distributions of the two end points for both the trial effects and the treatment effect, is needed. For two normally distributed end points, the natural measure was the correlation coefficient $R_{\epsilon_{Tij}|\epsilon_{Sij}}$ (8). It is important to note that the coefficient remains constant after specifying trial-specific intercepts and treatment effects in model (1)–(2).

For the two failure time end points the situation is different. First, non-linear associations between the end points are more likely. Second, the correlation between S_{ij} and T_{ij} depends on the shape of the marginal base-line hazard functions. It follows that if the general form of model (10)–(11) is assumed there will be a separate correlation coefficient for each trial. Consequently, the correlation is not a good candidate for the required measure of the association between S_{ij} and T_{ij}.

However, for a particular copula model the strength of the association between S_{ij} and T_{ij}, after adjusting their marginal distributions for the trial and the treatment effects, of course depends on δ. Thus, δ may be considered a natural candidate for the measure of association that is needed. Its drawback is that it is difficult to interpret and cannot be directly compared for different models. It would be easier to work with a transformation of δ that would, for example, have the interpretational properties of a correlation coefficient. To a large extent, such a measure is Kendall's τ. It can be shown that for the copula models (9) the following relationship between δ and Kendall's τ holds (Genest and MacKay, 1986):

$$\tau = 4\int_0^1\int_0^1 C_\delta(u, v)\, C_\delta(\mathrm{d}u, \mathrm{d}v) - 1. \tag{16}$$

Kendall's τ is the difference between the probability of concordance and the probability of discordance of two realizations of (S_{ij}, T_{ij}). It belongs to the interval $[-1, 1]$ and assumes a zero value when S_{ij} and T_{ij} are independent. From formula (16) it follows that τ depends only on the copula function C_δ. It is therefore independent of the marginal distributions of S_{ij} and

T_{ij} (Schweizer and Wolff, 1981) and measures the association between the two end points remaining after adjustment, through the marginal models (10)–(11), for trial and treatment effects. Thus, it is a transformation of δ that is easier to interpret and may be used as the required measure of the association.

The relationship between δ and Kendall's τ is particularly simple in Clayton's and Hougaard's models. For Clayton's model $\tau = (\delta - 1)/(\delta + 1)$, whereas for Hougaard's model $\tau = 1 - \delta$. These relationships allow for constructing a maximum likelihood estimate $\hat{\tau}$ of τ, given a maximum likelihood estimate $\hat{\delta}$ of δ. Furthermore, using the delta method it can be shown that for Clayton's model

$$\text{var}(\hat{\tau}) \approx \frac{4\,\text{var}(\hat{\delta})}{(\hat{\delta} + 1)^4}, \tag{17}$$

whereas for Hougaard's model

$$\text{var}(\hat{\tau}) = \text{var}(\hat{\delta}). \tag{18}$$

It is perhaps also worth mentioning that in Clayton's model the association parameter δ can be identified from the marginal distributions if a proportional hazard model with a set of covariates is used for modelling conditional (on frailty) hazard functions. Thus, as Hougaard (1987) pointed out, in the Clayton model case, δ measures something besides dependence. This problem does not appear for Hougaard's model.

5. Data analysis

In this section the two-stage approach proposed is applied to the two case-studies introduced in Section 2.

To construct the bivariate model at the first stage, the base-line hazard functions in model (10)–(11) were assumed to arise from a Weibull distribution. For both data sets the models of Clayton and Hougaard were considered. Consequently, the following two forms of the bivariate joint survivor function were assumed:

$$F(s,\, t) = [\exp\{-(1-\delta)(\lambda_{Si}s)^{r_{Si}}\,\exp(\alpha_i Z_{ij})\} + \exp\{-(1-\delta)(\lambda_{Ti}t)^{r_{Ti}}\,\exp(\beta_i Z_{ij})\} - 1]^{1/(1-\delta)}, \tag{19}$$

corresponding to Clayton's model with the copula given by equation (12), and

$$F(s,\, t) = \exp[\{(\lambda_{Si}s)^{r_{Si}}\,\exp(\alpha_i Z_{ij})\}^{1/\delta} + \{(\lambda_{Ti}t)^{r_{Ti}}\,\exp(\beta_i Z_{ij})\}^{1/\delta}]^{\delta}, \tag{20}$$

corresponding to Hougaard's model with the copula given by equation (13). In formulae (19) and (20), λ_{Si} and r_{Si} denote respectively the scale and shape parameter of the (trial-specific) marginal Weibull distribution of the surrogate end point, and λ_{Ti} and r_{Ti} denote the corresponding parameters for the true end point. In the analysis, both trial-specific and common base-line hazard ($\lambda_{Si} = \lambda_S$, $\lambda_{Ti} = \lambda_T$, $r_{Si} = r_S$, $r_{Ti} = r_T$, for all i) versions of model (10)–(11) were applied.

Maximum likelihood parameter estimates were obtained by using the Newton–Raphson procedure with numerical second-order derivatives implemented in SAS-IML 6.12 as routine `NLPNRR` (SAS Institute, 1995). Standard errors of the parameters were calculated by using the inverse of the observed matrix of second derivatives. The standard error of τ was computed by using formulae (17) and (18).

At the second stage the reduced random-effects model (14)–(15) was used. Effectively, it implied ignoring the information about λ_{Si}, λ_{Ti}, r_{Si} and r_{Ti} in modelling the relationship between α_i and β_i. Note that we used centres as the analysed units (rather than trials); thus the term 'trial specific' should be understood as meaning 'centre specific' in the remainder of this paper.

Under the reduced random-effects model (14)–(15), $R^2_{\text{trial(r)}}$ can be estimated by the square of the correlation coefficient between treatment effects α_i and β_i. However, in practice only estimates $\hat{\alpha}_i$ and $\hat{\beta}_i$, obtained from the first-stage copula model, are available. The estimate of $R^2_{\text{trial(r)}}$, obtained by calculating the square of the correlation coefficient between $\hat{\alpha}_i$ and $\hat{\beta}_i$, are likely to be biased. To see this more formally, assume that the estimated treatment effects $\hat{\alpha}_i$ and $\hat{\beta}_i$ follow the model

$$\begin{pmatrix} \hat{\alpha}_i \\ \hat{\beta}_i \end{pmatrix} = \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} + \begin{pmatrix} \epsilon_{ai} \\ \epsilon_{bi} \end{pmatrix} \tag{21}$$

where the estimation errors ϵ_{ai} and ϵ_{bi} are normally distributed with means 0 and covariance matrix

$$\Omega_i = \begin{pmatrix} \sigma_{aa,i} & \sigma_{ab,i} \\ \sigma_{ab,i} & \sigma_{bb,i} \end{pmatrix}, \tag{22}$$

and $(\alpha_i, \beta_i)^{\mathrm{T}}$ follows the reduced random-effects model (14) with the dispersion matrix D given by equation (15). Consequently, $(\hat{\alpha}_i, \hat{\beta}_i)^{\mathrm{T}}$ follows a normal distribution with mean $(\alpha, \beta)^{\mathrm{T}}$ and dispersion matrix $D + \Omega_i$.

For illustration, let us assume for the time being that $\Omega_i = \Omega$ (this assumption will be relaxed in what follows), with

$$\Omega = \begin{pmatrix} \sigma_{aa} & \sigma_{ab} \\ \sigma_{ab} & \sigma_{bb} \end{pmatrix},$$

and denote by ρ the correlation based on Ω. The correlation between $\hat{\alpha}_i$ and $\hat{\beta}_i$ can then be written as

$$\text{corr}(\hat{\alpha}_i, \hat{\beta}_i) = \text{corr}(\alpha_i, \beta_i)\{(1+\kappa_a)(1+\kappa_b)\}^{-1/2} + \rho\{(1+\kappa_a^{-1})(1+\kappa_b^{-1})\}^{-1/2}, \tag{23}$$

where $\kappa_a = \sigma_{aa}/d_{aa}$ and $\kappa_b = \sigma_{bb}/d_{bb}$ denote the reliability ratios for $\hat{\alpha}_i$ and $\hat{\beta}_i$. From equation (23) it follows that in the presence of independent estimation errors $R^2_{\text{trial(r)}}$ will be underestimated, whereas, for $\rho \neq 0$, $R^2_{\text{trial(r)}}$ may be either underestimated or overestimated. Additional insight might be gained under the assumption that $\kappa_a = \kappa_b = \kappa$. Then, equation (23) can be written as

$$\text{corr}(\hat{\alpha}_i, \hat{\beta}_i) = \text{corr}(\alpha_i, \beta_i) + \frac{\kappa}{1+\kappa}\{\rho - \text{corr}(\alpha_i, \beta_i)\}.$$

It follows that, if $\rho > \text{corr}(\alpha_i, \beta_i)$, then the correlation coefficient $\text{corr}(\hat{\alpha}_i, \hat{\beta}_i)$ will overestimate $\text{corr}(\alpha_i, \beta_i)$. Conversely, if $\rho < \text{corr}(\alpha_i, \beta_i)$, then $\text{corr}(\hat{\alpha}_i, \hat{\beta}_i)$ will underestimate $\text{corr}(\alpha_i, \beta_i)$.

To adjust for the possible bias in the estimation of $R^2_{\text{trial(r)}}$, we used an approach based on the developments by van Houwelingen *et al.* (2000). More specifically, we estimated the dispersion matrix D, defined by equation (14), by fitting the model resulting from equations (21)–(22) and (14)–(15) to the estimated pairs $(\hat{\alpha}_i, \hat{\beta}_i)$. To fit the model, the covariance matrices Ω_i, defined by equation (22), were assumed known and equal to their estimates obtained from the bivariate copula model (19) or (20). Computations were performed using

procedure MIXED implemented in SAS 6.12 (SAS Institute, 1997). An estimate $\hat{R}^2_{\text{trial(r)}}$ of $R^2_{\text{trial(r)}}$ was then obtained from the resulting estimate $\hat{D}$ of D by means of formula (6). The standard error of $\hat{R}^2_{\text{trial(r)}}$ was calculated from the estimated covariance matrix of the elements of $\hat{D}$ by using the delta method. In principle, the use of the delta method can lead to confidence limits violating the [0, 1] constraints on a coefficient of determination. To restrict the limits to the range [0, 1], a profile-likelihood-based approach for $R^2_{\text{trial(r)}}$ might be considered instead (Barndorff-Nielsen and Cox, 1994). This approach is numerically more involved and, as its results would not materially change the conclusions of the paper, it was not applied here.

5.1. Advanced ovarian cancer

The analysis was restricted to centres with at least three patients on each treatment arm. This constraint was adopted to ensure estimability of models (19) and (20), as they require the estimation of six marginal parameters (λ_{Si}, λ_{Ti}, r_{Si}, r_{Ti}, α_i, β_i) for each trial i. (In general, the minimum for the estimability of the marginal parameters would require at least three patients per centre, with at least one observed failure and at least one patient in each treatment group.) As a result, data for 39 centres (including the two smaller trials) were used, with a total sample size of 1153 patients. For comparability, the common marginal hazard functions version of model (10)–(11) was applied to the same data set.

Table 1 presents the results of the analysis. For all four models two values of $R^2_{\text{trial(r)}}$ are given: unadjusted and adjusted. The former was not adjusted for the measurement error in $\hat{\alpha}_i$ and $\hat{\beta}_i$ and was obtained by calculating the correlation coefficient for pairs $(\hat{\alpha}_i, \hat{\beta}_i)$. The latter was adjusted for the measurement error through fitting the model resulting from equations (21)–(22) and (14)–(15) to the estimated pairs $(\hat{\alpha}_i, \hat{\beta}_i)$, as described earlier.

Fig. 1 shows a plot of the treatment effects on the true end point (survival) by the treatment effects on the surrogate end point (progression-free survival), corresponding to the four models considered in the analysis. The effects are strongly correlated. The results shown in Table 1 confirm this conclusion. For the models with base-line hazards common to all centres, estimates of $R^2_{\text{trial(r)}}$ adjusted for the measurement error are equal to 0.95. They are higher than the corresponding unadjusted estimates and their 95% confidence intervals are wider. Because of convergence problems, the adjusted estimates for the trial-specific versions of models (19) and (20) could not be obtained. The unadjusted estimates suggest values of

Table 1. Results of the trial and individual level surrogacy analysis for the advanced ovarian cancer data (Ovarian Cancer Meta-Analysis Project, 1991)†

Parameter	*Results for Clayton's model with the following marginal hazards:*		*Results for Hougaard's model with the following marginal hazards:*	
	Common	*Trial specific*	*Common*	*Trial specific*
Trial level $R^2_{\text{trial(r)}}$				
Adjusted	0.95 [0.76, 1.14]	‡	0.95 [0.82, 1.07]	‡
Unadjusted	0.86 [0.77, 0.94]	0.87 [0.80, 0.95]	0.94 [0.90, 0.98]	0.88 [0.81, 0.95]
Individual level				
δ	13.03 [11.87, 14.31]	14.52 [13.20, 15.97]	0.16 [0.15, 0.17]	0.15 [0.14, 0.16]
τ	0.857 [0.845, 0.870]	0.871 [0.860, 0.883]	0.839 [0.828, 0.850]	0.853 [0.842, 0.863]

†95% confidence intervals are given in brackets.
‡Adjusted estimates of $R^2_{\text{trial(r)}}$ could not be obtained owing to numerical problems.

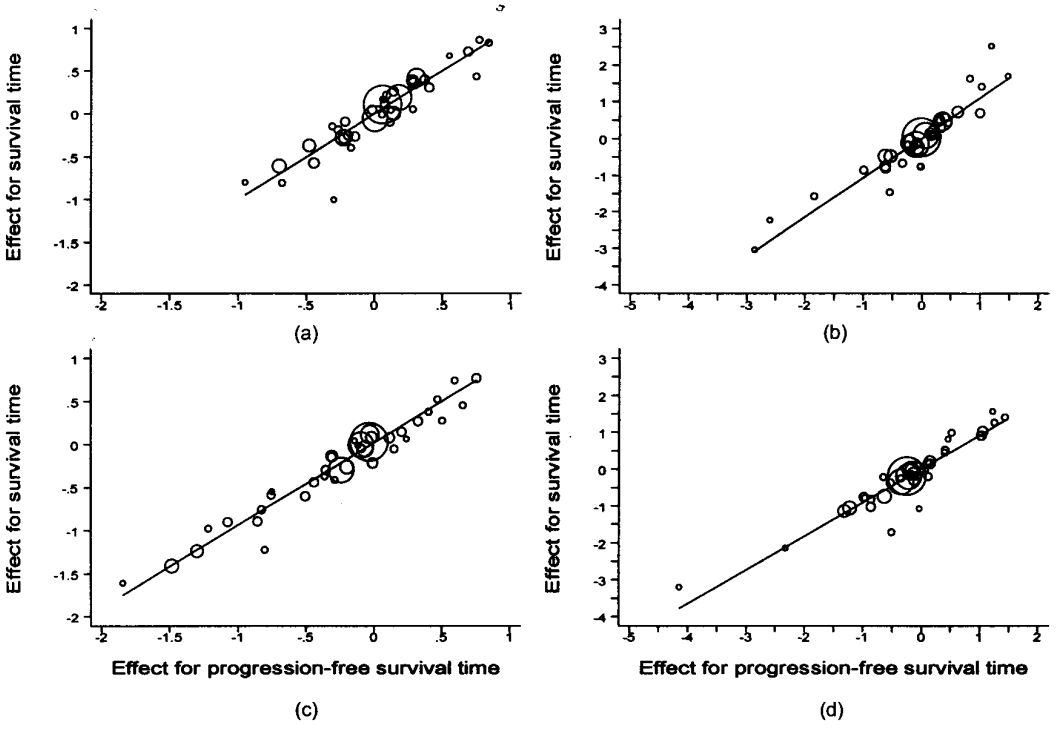

Fig. 1. Advanced ovarian cancer data — treatment effects on the true end point (survival time) *versus* treatment effects on the surrogate end point (progression-free survival time) for all units of analysis (the size of each point is proportional to the number of patients in the corresponding unit; ———, predictions from a (weighted by sample size) simple linear regression model): (a) Clayton's model with common base-line hazards; (b) Clayton's model with centre-specific base-line hazards; (c) Hougaard's model with common base-line hazards; (d) Hougaard's model with centre-specific base-line hazards

$R^2_{\text{trial(r)}}$ around 0.88. It might be conjectured that, on the basis of the results obtained for the models with the common base-line hazards, the unadjusted estimates are likely to be underestimating $R^2_{\text{trial(r)}}$.

It may be of interest to compare these results with those obtained by Buyse *et al.* (2000) by ignoring censoring and assuming a normal distribution for the logarithm of both end points. Their results were based on data for 1192 patients included in the meta-analysis (excluding two individuals lost to follow-up after randomization). In the analysis of the trial level surrogacy, they obtained unadjusted $R^2_{\text{trial(r)}} = 0.94$ (standard error 0.02). This value is somewhat higher than the unadjusted estimates presented in Table 1 (with the exception of Hougaard's model with common marginal hazard functions).

The values of Kendall's τ shown in Table 1 are close to 0.85 for all the models. They are slightly higher for the models generated by Clayton's family of distributions.

Although an interpretation of the value of the coefficients of determination is subjective, on the basis of the results presented in Table 1 it seems plausible to conclude that progression-free survival is a valid surrogate for survival in advanced ovarian cancer for treatments of the type used in the trials analysed. The effect of treatment can be observed earlier if progression-free survival is used instead of survival, although in this particular example the difference is small (Fig. 2). Hence, a trial that used progression-free survival would require less follow-up time and, possibly, fewer patients to conclude to the statistical

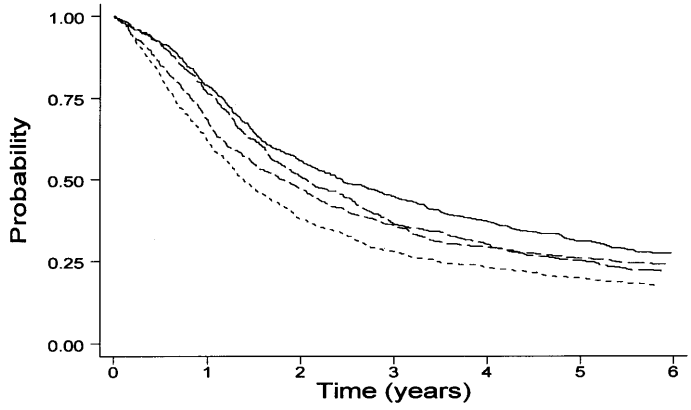

Fig. 2. Advanced ovarian cancer data — Kaplan–Meier estimates of survival (OS) and progression-free survival time (PFS) for the two treatment groups CP and CAP: ———, OS, CAP; — —, OS, CP; – – –, PFS, CAP; - - - - -, PFS, CP

Table 2. Predictions of the treatment effect on survival based on the estimated effect on progression-free survival for the advanced ovarian cancer data (Ovarian Cancer Meta-Analysis Project, 1991)†

Unit	*N*	$\hat{\alpha}_0$	$\hat{E}(\beta + b_0 \mid a_0)$ *for the following models:*				$\widehat{\beta + b_0}$
			Cc	*Cs*	*Hc*	*Hs*	
Centre 6	17	1.40	1.40	1.59	1.27	1.26	1.14
		(0.64)	[0.08, 2.71]	[0.02, 3.16]	[0.09, 2.45]	[−0.04, 2.57]	(0.738)
Centre 8	10	−1.00	−1.01	−1.04	−0.90	−0.85	−1.43
		(0.93)	[−2.86, 0.83]	[−3.13, 1.04]	[−2.55, 0.75]	[−2.56, 0.85]	(1.06)
Centre 37	12	−0.82	−0.85	−0.89	−0.76	−0.73	−0.55
		(0.68)	[−2.24, 0.53]	[−2.53, 0.75]	[−2.01, 0.49]	[−2.10, 0.63]	(0.78)
Centre 49	40	−1.14	−1.18	−1.23	−1.04	−1.02	−1.06
		(0.46)	[−2.17, −0.19]	[−2.46, 0.00]	[−1.91, −0.18]	[−2.06, 0.03]	(0.48)
Centre 55	31	−1.13	−1.17	−1.22	−1.04	−1.01	−1.13
		(0.47)	[−2.18, −0.16]	[−2.48, 0.03]	[−1.93, −0.15]	[−2.08, 0.05]	(0.49)
Centre BB	21	1.24	1.22	1.38	1.13	1.12	0.92
		(0.64)	[−0.09, 2.54]	[−0.18, 2.95]	[−0.06, 2.32]	[−0.20, 2.43]	(0.78)
DACOVA	274	−0.26	−0.29	−0.27	−0.24	−0.23	−0.21
		(0.13)	[−0.71, 0.13]	[−1.05, 0.52]	[−0.60, 0.11]	[−0.92, 0.46]	(0.14)
GONO	125	−0.24	−0.27	−0.24	−0.23	−0.21	−0.16
		(0.20)	[−0.79, 0.25]	[−1.08, 0.60]	[−0.69, 0.22]	[−0.95, 0.53]	(0.23)

† N is the number of patients per unit. $\hat{\alpha}_0$ and $\widehat{\beta + b_0}$ are treatment effects on progression-free survival and survival respectively, estimated from the data; $\hat{E}(\beta + b_0 \mid a_0)$ is the predicted effect of treatment on survival, given its effect on progression-free survival. Standard errors are given in parentheses, 95% prediction intervals in brackets; Cc, Clayton's model with common base-line hazards; Cs, Clayton's model with trial-specific base-line hazards; Hc, Hougaard's model with common base-line hazards; Hs, Hougaard's model with trial-specific base-line hazards.

significance of a truly superior treatment than a trial that used survival (Chen *et al.*, 1998).

Predictions of the effect of treatment on the survival time, based on the observed effect of treatment on progression-free survival time, are obviously of interest. Table 2 reports the predicted treatment effects for several centres selected randomly from the two large trials, as well as from the two small trials (DACOVA and GONO), in which the centre is unknown. The predictions for each unit were calculated on the basis of model (14)–(15). In each case,

the data for the unit for which the prediction was computed were excluded from fitting the model. In most cases the values for $\beta + b_0$ predicted under Hougaard's model are closer to the values estimated from the data than those predicted under Clayton's model. The former agree reasonably well with the effects estimated from the data, although in certain cases (for centre 8 or trial GONO, for instance) they are underestimated or overestimated by approximately 50%. As the differences between point estimates and predictions are expected, the prediction intervals are of more interest. Despite a high value of $R^2_{\text{trial(r)}}$, the intervals are wide. This is due to the error in the estimation of $\hat{E}(\beta + b_0|a_0)$, which remains substantial, in spite of a relatively large amount of data.

5.2. Colorectal cancer

Fig. 3 shows overall Kaplan–Meier estimates of the probability of survival and the probability of progression-free survival for the two advanced colorectal cancer trials. The estimates are based on the pooled data that are available for both trials (736 patients in total). The time gap between survival and progression-free survival is similar (around 6 months) to the gap observed in the previous example.

Similarly to the advanced ovarian cancer example that was presented in the previous section, in the analysis of the advanced colorectal cancer data only centres with at least three patients on each treatment arm were considered. As a result, data for 48 centres were used, with a total sample size of 642 patients. For comparability, the common marginal hazard functions version of model (10)–(11) was applied to the same data set.

Table 3 shows results obtained from the analysis. Fig. 4 shows a plot of the treatment effects on the true end point (survival time) by the treatment effects on the surrogate end point (progression-free survival time), corresponding to the four models considered in the analysis. The picture is very much different from that for the ovarian cancer example. For all four models the association of the trial-specfic treatment effects is low. The unadjusted estimates of $R^2_{\text{trial(r)}}$ are around 0.50. The adjusted estimates for the common base-line hazards versions of models (19) and (20) are equal to 0.24 and 0.33 respectively. The adjusted estimates for the trial-specific versions of models (19) and (20) could not be obtained owing to convergence problems. The estimates obtained for the common base-line hazards version suggest, however, that the unadjusted estimates are likely to be overestimating $R^2_{\text{trial(r)}}$.

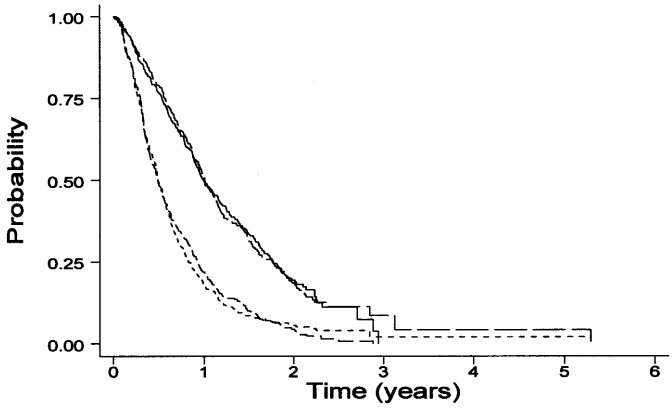

Fig. 3. Advanced colorectal cancer data—Kaplan–Meier estimates of survival (OS) and progression-free survival time (PFS) based on the pooled data for the Corfu trial (Corfu-A Study Group, 1996) and the trial by Greco *et al.* (1996): ———, OS, 5FU–IFN; — —, OS, 5FU–LV; – – –, PFS, 5FU–IFN; - - - - -, PFS, 5FU–LV

Table 3. Results of the trial and individual level surrogacy analysis for the advanced colorectal cancer data (Greco *et al.*, 1996; Corfu-A Study Group, 1995)†

Parameter	*Results for Clayton's model with the following marginal hazards:*		*Results for Hougaard's model with the following marginal hazards:*	
	Common	*Trial specific*	*Common*	*Trial specific*
Trial level $R^2_{\text{trial(r)}}$				
Adjusted	0.24 [−0.40, 0.89]	‡	0.33 [−0.69, 1.36]	‡
Unadjusted	0.45 [0.24, 0.66]	0.46 [0.26, 0.67]	0.50 [0.31, 0.70]	0.53 [0.34, 0.72]
Individual level				
δ	3.02 [2.68, 3.42]	4.04 [3.54, 4.64]	0.42 [0.38, 0.45]	0.37 [0.33, 0.40]
τ	0.502 [0.457, 0.548]	0.603 [0.560, 0.646]	0.583 [0.548, 0.619]	0.632 [0.597, 0.667]

†95% confidence intervals are given in brackets.
‡Adjusted estimates of $R^2_{\text{trial(r)}}$ could not be obtained owing to numerical problems.

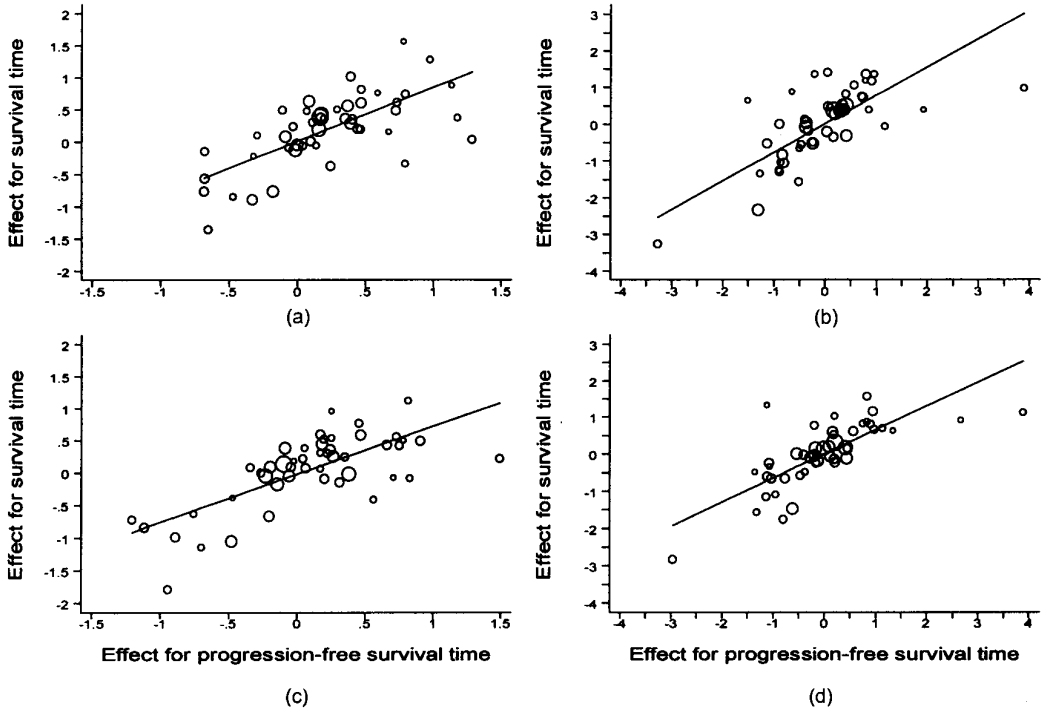

Fig. 4. Advanced colorectal cancer data—treatment effects on the true end point (survival time) *versus* treatment effects on the surrogate end point (progression-free survival time) for all units of analysis (the size of each point is proportional to the number of patients in the corresponding unit; ———, predictions from a (weighted by sample size) simple linear regression model): (a) Clayton's model with common base-line hazards; (b) Clayton's model with centre-specific base-line hazards; (c) Hougaard's model with common base-line hazards; (d) Hougaard's model with centre-specific base-line hazards

These results suggest the conclusion that the progression-free survival time is neither trial level nor individual level valid. Hence, it should probably not be used as a surrogate for survival in colorectal cancer for treatments of the type used in the trials analysed.

The marked difference between this example in colorectal cancer and the previous example in ovarian cancer underscores the difficulty of making general claims about surrogate end

points. In both examples, the average time between progression and death is about 6 months (see Figs 2 and 3), yet in colorectal cancer progression-free survival is not nearly as good a surrogate for survival as in ovarian cancer. This may be because, in advanced colorectal cancer, progression occurs early (a median time to progression of about 6 months) and is often followed by aggressive second-line therapies that may themselves have an influence on survival. In the presence of effective second-line therapies, progression-free survival might be expected to be a poor surrogate for survival because of the 'dilution' of the effect of first-line therapy on the final end point (Prentice, 1989). The examples analysed illustrate that generally the validity of a particular end point as a surrogate may depend on both the treatment and the disease under consideration.

6. Discussion

From a practical point of view it is unrealistic to expect perfect surrogacy. Thus, the application of the method developed by Buyse *et al.* (2000) requires the specification of a threshold allowing for an assessment of the proximity to 1 of the value of association measures such as Kendall's τ or the coefficients of determination $R^2_{\text{trial(r)}}$ and R^2_{indiv}. On purely theoretical grounds, however, it is difficult to propose such a threshold. Any other choice is necessarily subjective. Preferably, it should be guided by practical experience in using the definition of validity of a surrogate proposed by Buyse *et al.* (2000). For obvious reasons such an experience is thus far very limited. Taking this into account, observed values of $R^2_{\text{trial(r)}}$ around 0.9 have been judged as 'sufficiently close to 1', whereas those around 0.5 as 'not close to 1'.

One might argue whether the estimates and intervals for $R^2_{\text{trial(r)}}$ that are presented in Table 1 constitute enough evidence to consider progression-free survival a valid surrogate for survival in advanced ovarian cancer. However, even if it is judged insufficient, from Table 2 it is clear that for advanced colorectal cancer there is even less evidence. This possibility of assessing the strength of evidence for validity of a surrogate can be seen as an advantage of the method proposed by Buyse *et al.* (2000), especially when compared, for example, with the rigid 'yes' or 'no' decision rule that is implied by Prentice's definition (Prentice, 1989).

An important issue that is related to the assessment of the observed value of $R^2_{\text{trial(r)}}$ is the possibility of bias induced by using the two-stage model and estimation of treatment effects. To account for the bias, the mixed effects model (21)–(22) was fitted to the estimated treatment effects while adjusting for the (estimated) measurement error in the estimates. From this point of view it would be of interest to construct a full random-effects model with random intercepts and random treatment effects, e.g. by using multilevel modelling methodology (Goldstein, 1995). Such a model might replace the two-stage model (9)–(14) and allow for a full generalization of the method proposed by Buyse *et al.* (2000). However, Buyse *et al.* (2000) indicated that the conclusions coming from the somewhat *ad hoc* two-stage approach followed here compare very well with the more elegant full random-effects model. The drawback of such a random-effects model is its increased computational complexity. More work is being carried out in this area.

The approach proposed in this paper allows the method of validation of surrogate end points developed by Buyse *et al.* (2000) to be extended to the important case of two failure time end points. In the case-studies S_{ij} and T_{ij} were assumed to have Weibull marginal distributions. In general other distributional assumptions can be made. It is also possible to use a semiparametric approach with unspecified base-line hazard functions (Shih and Louis,

1995). If deemed necessary, the marginal models for the surrogate and the true end point can include important prognostic factors.

In the case-studies two copula models equivalent to proportional frailty models were used. In general, other models, which are not necessarily equivalent to proportional frailty models, can be applied (Oakes, 1989; Shih and Louis, 1995). The choice of a particular form of the copula function is of course important. It is possible to assess the goodness of fit of the chosen copula model by comparing a nonparametric estimate of a time-dependent correlation coefficient of martingale residuals for S_{ij} and T_{ij} with an estimate based on the model (Shih and Louis, 1995). Alternatively, for Archimedean copulas the method recently proposed by Wang and Wells (2000) might be used. Using these methods different copula models can be fitted to the data and the choice can be made on the basis of the evaluation of their fit.

The copula models are marginal models but, as has been already mentioned, in particular cases (Clayton, 1978; Hougaard, 1986; Oakes, 1989; Shih and Louis, 1995) they can be also seen as proportional frailty models. Several limitations of such models have been reported (Lindeboom and Van Den Berg, 1994; Xue and Brookmeyer, 1996). First, the unobserved frailty is assumed to be the same for both end points, which in general may not be reasonable. Second, in most cases the univariate frailty will induce only a positive association.

These limitations can be overcome by using a bivariate frailty model. Such a model has been proposed by Xue and Brookmeyer (1996). It can be seen as a random-intercepts proportional hazards model for S_{ij} and T_{ij}. The implementation of the model is numerically complex, however, as it requires extensive bidimensional numerical integration. Recently, Xue (1998) has proposed fitting the model by using quasi-likelihood equations, which dramatically reduces the numerical complexity. However, in the proposed form, the variances and correlation of the bivariate frailty distribution are treated as nuisance parameters and estimated by the method of moments. As a consequence, the estimator for the correlation is not restricted to lie in the interval $[-1, 1]$. In the (desirable) situation of high correlation between the true and the surrogate end points, this may practically preclude convergence of the estimation procedure for the model. This was observed when the method proposed by Xue (1998) was applied to the examples analysed in Section 5. A possible remedy would be to supplement the current procedure with an estimating equation for the parameters of the bivariate frailty distribution. With such a remedy, the model might become a better candidate than the copula model (9) and (10)–(11) for the first stage of the two-stage model.

A common limitation of the copula models and the model proposed by Xue and Brookmeyer (1996) is that the two end points are treated as exchangeable. In general, this need not be so, as is clear from the examples analysed (the progression-free survival time cannot be longer than the survival time). Thus, from a practical point of view it would be of interest to develop an approach allowing for a non-symmetrical treatment of the end points, e.g. using a conditional survival type of model (Arnold, 1995). Alternatively, the method of estimation of copula models when one of the failure time variables might be censored by the other, recently proposed by Wang (2000), might be considered. This is an important topic for future research.

In several cases, the confidence intervals are based on the normal approximation. This assumption, although motivated by large sample theory and in line with common practice, may require additional considerations, e.g. by means of a simulation study.

From a practical point of view, the extension of the method of validation of surrogate end points that was developed by Buyse *et al.* (2000) to the case of two failure time end points considerably broadens its applicability. Further extensions, e.g. to the case of a binary surrogate and a failure time true end point, are being developed.

Acknowledgements

The authors are grateful to the Ovarian Cancer Meta-Analysis Project and the Meta-Analysis Group in Cancer for permission to use their individual patient data, and to Dr Russ Wolfinger, Dr Theo Stijnen and Dr Weijing Wang for their assistance and helpful discussions. The Joint Editor's and referees' comments, which considerably improved the paper, are gratefully appreciated. The first and the fifth authors gratefully acknowledge support from Bijzonder Onderzoeksfonds Limburgs Universitair Centrum. The fourth author gratefully acknowledges support from Vlaams Instituut voor de Bevordering van het Wetenschapelijk-Technologisch Onderzoek in Industrie.

References

Anderson, J. E. (1995) Multivariate survival analysis using random effect models. In *Recent Advances in Life-testing and Reliability* (ed. N. Balakrishnan), pp. 603–622. Boca Raton: CRC.

Arnold, B. C. (1995) Conditional survival models. In *Recent Advances in Life-testing and Reliability* (ed. N. Balakrishnan), pp. 589–601. Boca Raton: CRC.

Barndorff-Nielsen, O. E. and Cox, D. R. (1994) *Inference and Asymptotics*, 1st edn. London: Chapman and Hall.

Buyse, M. and Molenberghs, G. (1998) The validation of surrogate endpoints in randomized experiments. *Biometrics*, **54**, 1014–1029.

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

Chen, T. T., Simon, R. M., Korn, E. L., Anderson, S. J., Lindblad, A. D., Wieand, H. S., Douglass, Jr, H. O., Fisher, B., Hamilton, J. M. and Friedman, M. A. (1998) Investigation of disease-free survival as a surrogate endpoint for survival in cancer clinical trials. *Communs Statist. Theory Meth.*, **27**, 1363–1378.

Clayton, D. G. (1978) A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, **65**, 141–151.

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

Cox, D. R. (1972) Regression models and life-tables. *J. R. Statist. Soc.* B, **34**, 187–202.

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

Flandre, P. and Saidi, Y. (1999) Estimating the proportion of treatment effect explained by a surrogate marker. *Statist. Med.*, **18**, 107–115.

Fleming, T. R., Prentice, R. L., Pepe, M. S. and Glidden, D. (1994) Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. *Statist. Med.*, **13**, 955–968.

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

Genest, C. and McKay, J. (1986) The joy of copulas: bivariate distributions with uniform marginals. *Am. Statistn*, **40**, 280–283.

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

Greco, F. A., Figlin, R., York, M., Einhorn, L., Schilsky, R., Marshall, E. M., Buys, S. S., Froimtchuk, M. J., Schuller, J., Buyse, M., Ritter, L., Man, A. and Yap, A. K. L. (1996) Phase III randomized study to compare interferon alfa-2a in combination with fluorouracil versus fluorouracil alone in patients with advanced colorectal cancer. *J. Clin. Oncol.*, **14**, 2674–2681.

Hougaard, P. (1986) Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, **73**, 387–396.

———(1987) Modelling multivariate survival. *Scand. J. Statist.*, **14**, 291–304.

———(1995) Frailty models for survival data. *Lifetime Data Anal.*, **1**, 255–274.

van Houwelingen, H. C., Arends, L. R. and Stijnen, T. (2000) Advanced methods in meta-analysis: multivariate approach and meta-regression. To be published.

Lindeboom, M. and Van Den Berg, G. J. (1994) Heterogeneity in models for bivariate survival: the importance of the mixing distribution. *J. R. Statist. Soc.* B, **56**, 49–60.

Molenberghs, G., Buyse, M., Geys, H., Renard, D. and Burzykowski, T. (2000) Statistical challenges in the evaluation of surrogate endpoints in randomized trials. To be published.

Nelsen, R. G. (1999) An introduction to copulas. *Lect. Notes Statist.*, **139**.

Oakes, D. (1989) Bivariate survival models induced by frailties. *J. Am. Statist. Ass.*, **84**, 487–493.

Ovarian Cancer Meta-Analysis Project (1991) Cyclophosphamide plus cisplatin versus cyclophosphamide, doxorubicin, and cisplatin chemotherapy of ovarian carcinoma: a meta-analysis. *J. Clin. Oncol.*, **9**, 1668–1674.

———(1998) Cyclophosphamide plus cisplatin versus cyclophosphamide, doxorubicin, and cisplatin chemotherapy of ovarian carcinoma: a meta-analysis. *Class. Pap. Curr. Comments*, **3**, 237–243.

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

Prentice, R. L. and Hsu, L. (1997) Regression on hazard ratios and cross ratios in multi-variate failure time analysis. *Biometrika*, **84**, 349–363

SAS Institute (1995) *SAS/IML Software: Changes and Enhancements through Release 6.11*. Cary: SAS Institute.

———(1997) *SAS/STAT Software: Changes and Enhancements through Release 6.12*. Cary: SAS Institute.

Schweizer, B. and Wolff, E. F. (1981) On nonparametric measures of dependence for random variables. *Ann. Statist.*, **9**, 879–885.

Shih, J. H. and Louis, T. A. (1995) Inferences on association parameter in copula models for bivariate survival data. *Biometrics*, **51**, 1384–1399.

Verbeke, G. and Molenberghs, G. (1997) Linear mixed models in practice: a SAS-oriented approach. *Lect. Notes Statist.*, **126**.

Wang, W. (2000) Semi-parametric inference for evaluating surrogate marker's predictive performance under dependent censoring. To be published.

Wang, W. and Wells, M. T. (2000) Model selection and semiparametric inference for bivariate failure-time data. *J. Am. Statist. Ass.*, **95**, 62–76.

Xue, X. (1998) Multivariate survival data under bivariate frailty: an estimating equation approach. *Biometrics*, **54**, 1631–1637.

Xue, X. and Brookmeyer, R. (1996) Bivariate frailty model for the analysis of multivariate survival time. *Lifetime Data Anal.*, **2**, 277–289.