J. R. Statist. Soc. A (2004)
167, *Part* 1, *pp.* 103–124

The validation of surrogate end points by using data from randomized clinical trials: a case-study in advanced colorectal cancer

Tomasz Burzykowski and Geert Molenberghs

Limburgs Universitair Centrum, Diepenbeek, Belgium

and Marc Buyse

International Drug Development Institute, Brussels, Belgium

[Received February 2001. Final revision April 2003]

Summary. In many therapeutic areas, the identification and validation of surrogate end points is of prime interest to reduce the duration and/or size of clinical trials. Buyse and co-workers and Burzykowski and co-workers have proposed a validation strategy for end points that are either normally distributed or (possibly censored) failure times. In this paper, we address the problem of validating an ordinal categorical or binary end point as a surrogate for a failure time true end point. In particular, we investigate the validity of tumour response as a surrogate for survival time in evaluating fluoropyrimidine-based experimental therapies for advanced colorectal cancer. Our analysis is performed on data from 28 randomized trials in advanced colorectal cancer, which are available through the Meta-Analysis Group in Cancer.

Keywords: Copula model; Meta-analysis; Plackett copula; Surrogate end point; Survival; Tumour response; Two-stage model

1. Introduction

Surrogate end points are referred to as end points that can replace or supplement other end points in the evaluation of experimental treatments or other interventions. For example, surrogate end points 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' end points (Ellenberg and Hamilton, 1989).

The most meaningful and the most objectively measured end point that is used to evaluate new cancer treatments is the overall survival time. However, it requires a long observation time and so may not be optimal for a fast assessment of therapeutic advances. The Food and Drug Administration has stated in its recommendations for accelerated approval of investigational cancer treatments that

> 'for many cancer therapies it is appropriate to utilize objective evidence of tumour shrinkage as a basis for approval, allowing additional evidence of increased survival and/or improved quality of life associated with that therapy to be demonstrated later'

(Food and Drug Administration, 1996).

Address for correspondence: Tomasz Burzykowski, Center for Statistics, Limburgs Universitair Centrum, Building D, Universitaire Campus, B-3590 Diepenbeek, Belgium.
E-mail: tomasz.burzykowski@luc.ac.be

 0964–1998/04/167103

European legislation allows for granting a marketing authorization under 'exceptional circumstances' where comprehensive data are not available at the time of submission (e.g. because of the rarity of the disease) and provided that the applicant agrees to a further programme of studies that will be the basis for a post-authorizations review of the benefit–risk profile of the drug. Although this primarily refers to situations where randomized clinical trials are lacking, it applies as well to the absence of data on a particular end point. According to the European Agency for the Evaluation of Medicinal Products guidelines for the evaluation of anticancer agents, possible end points for phase III trials in oncology include the response rate (Committee for Proprietary Medicinal Products, 2001). The guidelines state that, if the objective response rate is used as the primary end point, compelling justifications are needed and normally additional supportive evidence of efficacy in terms of, for example, the control of symptoms is necessary. These requirements are close to those specified in the Food and Drug Administration's accelerated approval system.

The shrinkage of tumour mass, also called 'tumour response', has long been the corner-stone of the development of cytotoxic therapies for solid tumours, even though the effect of a tumour response on the patient's survival has often been questioned (Anderson *et al.*, 1983; Oye and Shapiro, 1984; Ellenberg and Hamilton, 1989; Buyse and Piedbois, 1996). In this paper, we take up this issue and study the relationship between the end points of response and survival, as well as between the effects of an investigational treatment on these two end points. More specifically, we study the validity of tumour response as a surrogate for survival in assessing the benefits of various treatment regimens for advanced colorectal cancer. For this, a likelihood model for the joint assessment of survival and ordinal or binary end points needs to be developed.

Prentice (1989) proposed a formal definition of surrogate end points and outlined a set of criteria. Much debate ensued, for the criteria set out by Prentice are not straightforward to verify (Freedman *et al.*, 1992; Fleming *et al.*, 1994). In addition, Prentice's criteria are only equivalent to the definition that he proposed in the case of binary end points (Buyse and Molenberghs, 1998). Freedman *et al.* (1992) supplemented Prentice's approach by introducing the *proportion explained*, aimed at measuring the proportion of the treatment effect that is mediated by the surrogate. This proposal was important in that it shifted the interest in the validation of surrogate end points from significance testing to estimation. However, it is also surrounded with difficulties. Consequently, Buyse and Molenberghs (1998) proposed to replace it 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*. Also these have important drawbacks and therefore Daniels and Hughes (1997) and Buyse *et al.* (2000a) proposed meta-analytic approaches. Some of this discussion is given in Molenberghs *et al.* (2003).

As Buyse *et al.* (2000a) focused solely on the case of normally distributed end points, it is necessary to explore other settings, often more complicated owing to the absence of a unifying framework such as the multivariate normal distribution. Burzykowski *et al.* (2001) extended the approach to the 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 survival time (Chen *et al.*, 1998).

The focus of this paper is twofold. First, using the developments in Burzykowski *et al.* (2001), we further extend the approach that was proposed by Buyse *et al.* (2000a) to the case when the surrogate is an ordinal categorical or a binary variable and the true end point is a failure time variable. For this, motivated by the results developed by Molenberghs *et al.* (2001) in the context

of mixed discrete and continuous data, we propose a novel concept of using copula models to model mixed bivariate categorical or binary variables and survival data jointly. Secondly, we use the new extension to study the validity of response as a surrogate for survival for advanced colorectal cancer. The data that we analyse come from four successive meta-analyses of 28 trials in advanced colorectal cancer (Buyse *et al.*, 2000b). All four meta-analyses compared a standard regimen given as a bolus injection to various experimental regimens. All regimens used fluoropyrimidines in the form of 5-fluorouracil (5FU) or 5-fluoro-2′deoxyuridine (FUDR). In the majority of the trials, individual patient data were available on survival times and tumour responses, which were classified as complete response (CR), partial response (PR), stable disease (SD) or progressive disease (PD).

The rest of the paper is organized as follows. In Section 2 the method of validation that was proposed by Buyse *et al.* (2000a) for the case of two normally distributed end points is summarized and then an extension is proposed to the case of an ordinal categorical surrogate. Our case-study is presented and analysed in Section 3 and some conclusions and amplifications based on this analysis are presented in Section 4. Concluding remarks are formulated in Section 5.

2. Meta-analytic approach to the validation of surrogate end points

Throughout the paper, we adopt the following notation: T and S are random variables denoting the true and surrogate end points respectively and Z is an indicator variable for treatment. We shall expand the notation by using two indices: $i = 1, \ldots, N$ for trial and $j = 1, \ldots, n_i$ for subjects within trial.

2.1. Normally distributed outcomes

The concept of the meta-analytic approach to the validation of surrogate end points has been developed by Buyse *et al.* (2000a) for the case of two normally distributed end points. At the core of this proposal lies a two-level hierarchical model (Laird and Ware, 1982; Goldstein, 1995; Verbeke and Molenberghs, 2000). In practice, such a model can assume that the trial level effects are either random or fixed. In the latter case, a two-stage approach in which models at both levels are fitted separately can be followed (Laird and Ware, 1982). The fixed effects version is also easiest to generalize to the current situation with an ordinal surrogate and a time-to-event true end point.

The first stage is based on a trial-specific model:

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

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

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.

At the second stage, it is assumed that $(\mu_{Si}, \mu_{Ti}, \alpha_i, \beta_i)'$ follows a normal distribution with mean $(\mu_S, \mu_T, \alpha, \beta)'$ and with an unstructured covariance matrix.

A natural quantity to assess the quality of a surrogate at the trial level is the coefficient of determination, R^2_{trial} say, pertaining to the distribution of β_i conditional on μ_{Si} and α_i. This coefficient measures how precisely we may predict the effect of treatment on the true end point on the basis of previous data and the observed treatment effect on the surrogate end point from a new trial. The association between the surrogate and final end points, after adjustment for the effect of treatment, is captured by the coefficient of determination, R^2_{indiv} say, pertaining to the distribution of ε_{Tij} conditional on ε_{Sij}. This coefficient measures how precisely we may

predict the value of T_{ij} for an individual patient on the basis of the observed value of S_{ij} and the treatment assignment.

We may call a surrogate 'valid' when both R^2_{trial} and R^2_{indiv} are sufficiently close to 1, the precise quantification of which will depend on the context.

Technically, the two-stage model described above might be fitted to data by using a mixed effects model representation (Buyse *et al.*, 2000a). The convergence of the Newton–Raphson algorithm yielding maximum likelihood solutions is not guaranteed, however. Simulation results indicate that there should be enough variability at the trial level, and a sufficient number of trials, to obtain convergence of the Newton–Raphson algorithm (Buyse *et al.*, 2000a). Alternatively, we can still rely on the fixed effects representation and obtain estimates of the treatment effects at the trial level and estimates of residuals at the individual level, thus enabling estimation of R^2_{trial} and R^2_{indiv} respectively.

2.2. The case of an ordinal or a binary surrogate end point

We shall now assume that T is a failure time random variable and S is a categorical variable with K ordered categories. For each of $j = 1, \ldots, n_i$ patients from trial i ($i = 1, \ldots, N$) we thus have quadruplets $(X_{ij}, \Delta_{ij}, S_{ij}, Z_{ij})$, where X_{ij} is a possibly censored version of survival time T_{ij} and Δ_{ij} is the censoring indicator asuming the values 1 for observed failures and 0 otherwise.

To propose validation measures, similar to those introduced in the previous section, we shall use copula models (Shih and Louis, 1995; Nelsen, 1999; Burzykowski *et al.*, 2001). Accordingly, we propose to replace model (1)–(2) by a bivariate copula model for the true end point T_{ij} and a latent continuous variable $\tilde{S}_{ij}$ underlying the surrogate end point S_{ij}.

Specifically, to model S_{ij} we propose the proportional odds model

$$\text{logit}\{P(S_{ij} \leqslant k|Z_{ij})\} = \gamma_{ik} + \alpha_i Z_{ij}. \tag{3}$$

It can be interpreted as assuming a logistic distribution for the latent variable $\tilde{S}_{ij}$. The value of the marginal cumulative distribution function of $\tilde{S}_{ij}$, given $Z_{ij} = z$, will be denoted by $F_{\tilde{S}_{ij}}(s; z)$. Note that, in the case of a binary surrogate S_{ij}, model (3) is equivalent to a logistic regression model.

It is worth noting that the estimation of model (3) requires that in each trial all response levels are observed. In practice, it often happens that in some trials not all levels are observed. To adapt model (3) for such a case, we rewrite it as

$$\text{logit}\{P(S_{ij} \leqslant k|Z_{ij})\} = \eta^0_k + \eta_i + \eta_{ik} + \alpha_i Z_{ij}, \tag{4}$$

where for identifiability we might specify that, for example,

$$\eta_1 = \eta_{11} = \ldots = \eta_{1,K-1} = 0.$$

If, for a particular trial, i_0 say, not all levels of S are observed, we might use model (4) with the terms $\eta_{i_0 1}, \ldots, \eta_{i_0, K-1}$ constrained to 0. As a special case, the following model might be considered:

$$\text{logit}\{P(S_{ij} \leqslant k|Z_{ij})\} = \eta^0_k + \eta_i + \alpha_i Z_{ij}. \tag{5}$$

The model assumes a fixed set of cut points $\eta^0_1, \ldots, \eta^0_{K-1}$ but allows for trial-specific shifts η_i of the set.

To model the effect of treatment Z_{ij} on the marginal distribution of T_{ij} we propose, as in Burzykowski *et al.* (2001), to use the proportional hazard model

$$\lambda_{ij}(t|Z_{ij}) = \lambda_i(t) \exp(\beta_i Z_{ij}), \tag{6}$$

where β_i are trial-specific effects of treatment Z and $\lambda_i(t)$ is a base-line hazard function. The marginal cumulative distribution function of T_{ij}, following model (6) with $Z_{ij} = z$, will be denoted by $F_{T_{ij}}(t; z)$.

To specify fully a bivariate model corresponding to equations (1)–(2), let us assume that the joint cumulative distribution of T_{ij} and $\tilde{S}_{ij}$, given $Z_{ij} = z$, is generated by a one-parameter copula function C_θ:

$$F_{T_{ij},\tilde{S}_{ij}}(t, s; z) = C_\theta\{F_{T_{ij}}(t; z), F_{\tilde{S}_{ij}}(s; z), \theta\}. \tag{7}$$

C_θ is a distribution function on $[0, 1]^2$ with $\theta \in \mathbb{R}^1$ (Genest and McKay, 1986; Shih and Louis, 1995; Nelsen, 1999), describing the association between $\tilde{S}_{ij}$ and T_{ij}. An attractive feature of model (7) is that the marginal models (the proportional odds and proportional hazards models in our particular case) and the association model can be selected without constraining each other.

Using the joint distribution function (7) with proportional hazard model (6) and proportional odds model (3) (or its modification) as marginal models, it is possible to construct the likelihood function for the observed data ($X_{ij} = x_{ij}, \Delta_{ij} = \delta_{ij}, S_{ij} = s_{ij}, Z_{ij} = z_{ij}$). The details of the construction are described in Appendix A.

At the first stage we propose to use the likelihood function to obtain an estimate of θ and estimates of trial-specific treatment effects α_i and β_i on the surrogate and the true end point respectively. At the second stage, we propose to use the trial level model

$$\begin{pmatrix} \eta_i \\ \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \eta \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} e_i \\ a_i \\ b_i \end{pmatrix}, \tag{8}$$

with η_i obtained from model (4) or model (5). The second term on the right-hand side of equation (8) is assumed to follow a zero-mean normal distribution with dispersion matrix

$$D = \begin{pmatrix} d_{ee} & d_{ea} & d_{eb} \\ & d_{aa} & d_{ab} \\ & & d_{bb} \end{pmatrix}.$$

The quality of surrogate S at the trial level can be assessed on the basis of the coefficient of determination:

$$R^2_{\text{trial}(\alpha,\eta)} = \frac{\begin{pmatrix} d_{eb} \\ d_{ab} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} d_{ee} & d_{ea} \\ d_{ea} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{eb} \\ d_{ab} \end{pmatrix}}{d_{bb}}. \tag{9}$$

The index 'trial(α, η)' in $R^2_{\text{trial}(\alpha,\eta)}$ indicates that the coefficient pertains to the distribution of β_i conditional on the set of trial-specific parameters including α_i and η_i.

In principle, if the unrestricted marginal model (3) is used at the first stage, we might consider taking into account the information about the cut points $\gamma_{i1}, \ldots, \gamma_{i,K-1}$. A simple solution would be to replace η_i in equation (8) with the vector $(\gamma_{i1}, \ldots, \gamma_{i,K-1})'$. From the formal point of view, however, in this case the assumption of normality would have to be modified to reflect the ordering of the γ_{ij}s.

Alternatively, if the information in the cut points can be ignored, the use of a simple linear regression model could be considered (Daniels and Hughes, 1997; Burzykowski *et al.*, 2001):

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} a_i \\ b_i \end{pmatrix} \tag{10}$$

with dispersion matrix

$$D_\alpha = \begin{pmatrix} d_{aa} & d_{ab} \\ & d_{bb} \end{pmatrix}. \tag{11}$$

In that case the coefficient of determination $R^2_{\text{trial}(\alpha,\eta)}$ reduces to

$$R^2_{\text{trial}(\alpha)} = \frac{d^2_{ab}}{d_{aa} d_{bb}}, \tag{12}$$

the square of the correlation between α_i and β_i. It can be noted here that, using equations (9) and (12), we can write

$$R^2_{\text{trial}(\alpha,\eta)} = \frac{R^2_{\text{trial}(\alpha)}}{1 - \text{corr}^2(\eta_i, \alpha_i)} + \text{corr}(\eta_i, \beta_i) \frac{\text{corr}(\eta_i, \beta_i) - 2\,\text{corr}(\alpha_i, \beta_i)\,\text{corr}(\eta_i, \alpha_i)}{1 - \text{corr}^2(\eta_i, \alpha_i)}. \tag{13}$$

It follows that, formally, $R^2_{\text{trial}(\alpha,\eta)} = R^2_{\text{trial}(\alpha)}$ if $\text{corr}(\eta_i, \alpha_i) = \text{corr}(\eta_i, \beta_i) = 0$. To use $R^2_{\text{trial}(\alpha)}$ and model (10) instead of $R^2_{\text{trial}(\alpha,\eta)}$ and model (8), we would thus require that treatment effects on true and surrogate end points should be uncorrelated with the base-line distribution (for $Z = 0$) of S. The use of $R^2_{\text{trial}(\alpha)}$ might give different results from those of the use of $R^2_{\text{trial}(\alpha,\eta)}$, e.g. in the presence of treatment–surrogate interaction.

To assess the quality at the individual level, a measure of association between S_{ij} and T_{ij} is needed. A natural candidate is θ, since its value modifies the form of the copula function and, consequently, influences the strength of the association between $\tilde{S}_{ij}$ and T_{ij}. A drawback of θ is that, for different copula functions, it may assume values from different domains. To overcome this difficulty the use of Kendall's τ or Spearman's ρ may be considered (Burzykowski *et al.*, 2001). Both measures are transformations of θ and can be interpreted similarly to a correlation coefficient irrespectively of the copula function (Nelsen, 1999). Alternatively, it may be possible to choose a copula such that θ has a meaningful interpretation. This option will be discussed next.

In principle, different copula functions can be used for the bivariate distribution (7). We propose to use the bivariate Plackett copula (Plackett, 1965; Mardia, 1970; Dale, 1986; Nelsen, 1999). This particular choice is motivated by the fact that, for the Plackett copula, the association parameter θ takes the form of a (constant) global odds ratio. Specifically, in our setting (for $k = 1, \ldots, K-1$ and $t > 0$)

$$\begin{aligned} \theta &= \frac{P(T_{ij} > t, S_{ij} > k)\, P(T_{ij} \leqslant t, S_{ij} \leqslant k)}{P(T_{ij} > t, S_{ij} \leqslant k)\, P(T_{ij} \leqslant t, S_{ij} > k)} \\ &= \frac{P(T_{ij} > t | S_{ij} > k)}{P(T_{ij} \leqslant t | S_{ij} > k)} \left\{ \frac{P(T_{ij} > t | S_{ij} \leqslant k)}{P(T_{ij} \leqslant t | S_{ij} \leqslant k)} \right\}^{-1}. \end{aligned} \tag{14}$$

Thus, it is naturally interpreted as the (constant) ratio of the odds for surviving beyond time t given response higher than k to the odds of surviving beyond time t given response at most k. For a binary surrogate, it is just the odds ratio for responders *versus* non-responders (assuming that $k = 2$ indicates the response).

A more detailed treatment of the Plackett copula is given in Appendix B.

3. Case-study

The two-stage approach described above was applied to the advanced colorectal cancer data. Four-category tumour response was considered a surrogate for survival time and contrasted with a binary response.

3.1. Description of the data

We shall use data from 28 advanced colorectal cancer trials (Advanced Colorectal Cancer Meta-Analysis Project, 1992, 1994; Meta-Analysis Group in Cancer, 1996, 1998). The individual patient data were collected by the Meta-Analysis Group in Cancer between 1990 and 1996 to obtain an overall quantitative assessment of the value of several experimental treatments in advanced colorectal cancer. In the four meta-analyses, the comparison was between an experimental treatment and a control treatment. The control treatments, referred to hereafter as 'FU bolus', were similar across the four meta-analyses and consisted of fluoropyrimidines (5FU or FUDR) given as a bolus intravenous injection. The experimental treatments, referred to hereunder as 'experimental FU', differed across the four meta-analyses and consisted of 5FU modulated by leucovorin (Advanced Colorectal Cancer Meta-Analysis Project, 1992), of 5FU modulated by methotrexate (Advanced Colorectal Cancer Meta-Analysis Project, 1994), of 5FU given in continuous infusion (Meta-Analysis Group in Cancer, 1998) and of hepatic arterial infusion of FUDR for patients with metastases confined to the liver (Meta-Analysis Group in Cancer, 1996). As noted by Daniels and Hughes (1997), the use of an 'experimental' treatment that varies among the trials can be defended on the grounds of generalizability of the results of the validation process to future clinical trials and treatments. The experimental treatments in our example might be considered as representatives of 'the modifications of the standard fluoropyrimidine-based regimen' in advanced colorectal cancer.

Several of the 28 trials were multiarmed. In total, 33 randomized comparisons were considered in the four meta-analyses. Individual patient data were available for 27 of the comparisons (in 24 studies). From now on, we shall refer to each of the comparisons as a separate 'trial'.

Table 1 presents summary data for the trials included in the analysis. In particular, for each trial and each treatment arm Table 1 contains the median survival time (in months) and the distribution of the four tumour response categories CR, PR, SD and PD (World Health Organization, 1979). Also, the observed percentage for the binary response CR + PR is given. The first column of Table 1 contains the labels that are used to identify the trials in Advanced Colorectal Cancer Meta-Analysis Project (1992, 1994) and Meta-Analysis Group in Cancer (1996, 1998) describing the four meta-analyses; we refer to these for additional details regarding the original publications of results of the trials.

From Table 1 it can be seen that the trials varied quite considerably in sample size. The total size ranged from 15 ('City of Hope, HAI *versus* ST') to 382 ('GITSG') patients. The last two rows of Table 1 indicate that, overall, CR was rarely observed. Nevertheless, CR and PR were observed more frequently for experimental FU (3.2% and 19.2% respectively) than for FU bolus (2.1% and 9.6% respectively). Consequently, the response rate, i.e. the combined percentage of CR and PR, was higher for experimental FU (22.4% compared with 11.7% for FU bolus). This conclusion applies also to all except three ('NCOG', 'GOIRC' and 'RPCI, 5FU + M') individual trials. Similarly, the median survival time was slightly longer for experimental FU (9.8 months) than for FU bolus (8.9 months). This pattern can be consistently seen for all except eight individual trials.

Table 2 presents estimates of odds for binary response (CR + PR *versus* SD + PD) and the relative mortality hazard for experimental FU *versus* FU bolus. Overall, the odds were approximately double for the experimental treatment, with a simultaneous 10% reduction in the risk of death.

Fig. 1 shows survival curves by treatment within tumour response categories. There is no statistically significant difference between experimental FU and bolus FU in any tumour response category (CR, $p = 0.544$; PR, $p = 0.791$; SD, $p = 0.525$; PD, $p = 0.059$ for a log-rank test stratified by trial; three patients with unknown responses were treated as 'progressions'), which

Table 1. Summary data for 27 analysed trials†

Trial	*Treatment*	*N*	*Tumour reponse (%)*					*Median survival (months)*
			CR	*PR*	*SD*	*PD*	*CR + PR*	
Advanced Colorectal Cancer Meta-Analysis Project (1992)								
GITSG	5FU + L	269	1.5	20.1	0.0	78.4	21.6	11.3
	ST	113	0.0	10.6	0.0	89.4	10.6	10.7
NCOG	5FU + L	107	5.6	12.1	62.6	19.6	17.7	10.5
	ST	55	9.1	9.1	65.4	16.4	18.2	11.4
GOIRC	5FU + L	91	3.3	9.9	36.3	50.5	13.2	12.4
	ST	90	6.7	8.9	31.1	53.3	15.6	14.5
GISCAD	5FU + L	91	5.5	15.4	31.9	47.2	19.9	13.0
	ST	89	3.4	6.7	31.5	58.4	10.1	13.0
Genova	5FU+L	75	6.7	14.7	36.0	42.7	21.4	11.0
	ST	73	2.7	5.5	52.0	39.7	8.2	11.0
Toronto	5FU+L	66	0.0	31.8	0.0	68.2	31.8	12.0
	ST	64	0.0	6.2	0.0	93.7	6.2	9.6
City of Hope	5FU+L	39	2.6	35.9	35.9	25.6	38.7	14.2
	ST	40	0.0	12.5	47.5	40.0	12.5	12.7
RPCI	5FU + L	30	3.3	36.7	23.3	36.7	40.2	11.0
	ST	23	0.0	8.7	4.3	87.0	8.7	11.1
Bologna	5FU + L	34	0.0	26.5	32.3	41.2	26.5	10.1
	ST	30	0.0	3.3	56.7	40.0	3.3	7.5
Advanced Colorectal Cancer Meta-Analysis Project (1994)								
EORTC	5FU + M	152	2.6	15.1	38.2	44.1	17.7	12.1
	ST	154	2.6	9.1	31.2	57.1	11.7	8.9
RPCI	5FU + M	23	0.0	4.3	13.0	82.6	4.3	10.3
	ST	23	0.0	8.7	4.3	87.0	8.7	11.1
NGTAG	5FU + M + L	122	2.5	13.9	39.3	44.3	16.4	8.1
	ST	127	0.0	2.4	43.4	54.3	2.4	6.0
AIO	5FU + M + L	86	4.6	18.6	33.7	43.0	23.2	10.7
	ST	78	2.6	14.1	46.1	37.2	16.7	13.7
NCOG	5FU + M + L	103	5.8	12.6	65.0	16.5	18.4	12.3
	ST	55	9.1	9.1	65.4	16.4	18.2	11.4
GOCS	5FU + M + L	64	1.6	25.0	32.8	40.6	26.6	11.9
	ST	61	0.0	11.5	22.9	65.6	11.5	8.9
Mar del Plata	5FU + M + L	28	3.6	14.3	7.1	75.0	17.9	0.7
	ST	33	0.0	0.0	57.6	42.4	0.0	1.0
Spain	5FU + M + L	26	3.8	19.2	53.8	23.1	23.0	13.2
	ST	33	3.0	12.1	51.5	33.3	15.1	8.6
Meta-Analysis Group in Cancer (1996)								
MSKCC	HAI	43	0.0	48.8	37.2	13.9	48.8	18.3
	ST	48	0.0	16.7	33.3	50.0	16.7	14.5
NCCTG	HAI	39	2.6	38.5	33.3	25.6	41.1	12.8
	ST	35	0.0	17.1	57.1	25.7	17.1	11.0
NCI	HAI	32	3.1	37.5	3.1	56.2	40.6	16.9
	ST	32	3.1	12.5	0.0	84.4	15.6	11.6
City of Hope	HAI	9	0.0	77.8	0.0	22.2	77.8	22.9
	ST	6	0.0	50.0	0.0	50.0	50.0	23.0

(*continued*)

Table 1 (*continued*)

Trial	*Treatment*	*N*	*Tumour reponse (%)*					*Median survival (months)*
			CR	*PR*	*SD*	*PD*	*CR+PR*	
Meta-Analysis Group in Cancer (1998)								
SWOG	CII	174	2.9	10.3	19.5	67.2	13.2	15.0
	ST	182	2.7	9.9	30.2	57.1	12.6	13.9
ECOG	CII	162	4.9	22.8	8.6	63.6	27.7	13.0
	ST	162	3.1	14.2	5.6	77.2	17.3	10.5
NCIC	CII	95	1.0	10.5	36.8	51.6	11.5	10.1
	ST	90	1.1	5.6	32.2	61.1	6.7	9.3
France	CII	77	3.9	22.1	41.6	32.5	26.0	8.5
	ST	78	0.0	12.8	39.7	47.4	12.8	9.8
MAOP	CII	88	4.5	25.0	69.3	1.1	29.5	10.6
	ST	85	0.0	9.4	89.4	1.2	9.4	11.2
Jerusalem	CII	11	0.0	9.1	18.2	72.7	9.1	8.6
	ST	15	0.0	6.7	60.0	33.3	6.7	12.0
Total	EX	2136	3.2	19.2	29.9	47.7	22.4	9.8
	ST	1874	2.1	9.6	34.1	54.2	11.7	8.9

†ST, control treatment (bolus 5FU or FUDR); EX, experimental treatment (M, methotrexate; L, leucovorin; HAI, FUDR by hepatic arterial infusion; CII, 5FU by continuous intravenous infusion); *N*, sample size. The median survival times are estimated from the Kaplan–Meier survival curve.

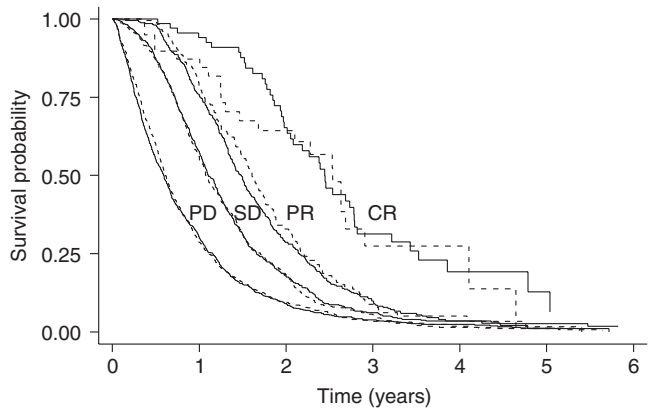

Fig. 1. Survival curves by treatment within tumour response categories (———, experimental; -------, control): CR, *N* = 108; PR, *N* = 590; SD, *N* = 1276; PD, *N* = 2036)

confirms that the overall survival benefit in favour of experimental FU is due to the higher tumour response rates that are obtained with experimental FU compared with bolus FU. This observation suggests that tumour response might be a valid surrogate for survival according to Prentice's (1989) definition.

The true end point T is the survival time, defined as the time from randomization to death from any cause. In our set of data, most patients have died (3591 out of 4010 patients, i.e. 89.5%). The surrogate end point S is tumour response, defined either as a binary variable with $S = 2$ for CR or PR and $S = 1$ for SD or PD, or as a categorical variable with $S = 4, 3, 2, 1$ for CR, PR,

Table 2. Summary results for binary tumour response and survival for the 27 analysed trials†

Trial	*Odds ratio [95% confidence interval]*	*Hazard ratio [95% confidence interval]*
Advanced Colorectal Cancer Meta-Analysis Project (1992)		
GITSG	2.31 [1.19, 4.50]	0.88 [0.70, 1.12]
NCOG	0.97 [0.42, 2.26]	1.22 [0.86, 1.72]
GOIRC	0.82 [0.36, 1.90]	1.23 [0.88, 1.72]
GISCAD	2.34 [1.00, 5.51]	1.09 [0.76, 1.56]
Genova	3.03 [1.11, 8.24]	0.90 [0.65, 1.25]
Toronto	7.00 [2.24, 21.82]	0.78 [0.54, 1.13]
City of Hope	4.37 [1.40, 13.65]	0.78 [0.50, 1.23]
RPCI	7.00 [1.38, 35.51]	1.13 [0.65, 1.98]
Bologna	10.44 [1.23, 88.21]	0.74 [0.43, 1.28]
Advanced Colorectal Cancer Meta-Analysis Project (1994)		
EORTC	1.63 [0.86, 3.11]	0.79 [0.62, 1.02]
RPCI	0.48 [0.04, 5.66]	1.28 [0.71, 2.30]
NGTAG	8.10 [2.34, 28.05]	0.76 [0.59, 0.98]
AIO	1.51 [0.70, 3.30]	1.03 [0.75, 1.40]
NCOG	1.02 [0.44, 2.37]	0.89 [0.63, 1.26]
GOCS	2.79 [1.06, 7.31]	0.78 [0.54, 1.12]
Mar del Plata	15.68 [0.83, 297.4]	0.98 [0.58, 1.67]
Spain	1.68 [0.45, 6.28]	1.17 [0.62, 2.24]
Meta-Analysis Group in Cancer (1996)		
MSKCC	4.77 [1.81, 12.54]	0.77 [0.51, 1.17]
NCCTG	3.36 [1.13, 9.96]	0.95 [0.60, 1.50]
NCI	3.69 [1.13, 12.10]	0.81 [0.46, 1.40]
City of Hope	3.50 [0.37, 32.97]	0.91 [0.31, 2.66]
Meta-Analysis Group in Cancer (1998)		
SWOG	1.05 [0.57, 1.96]	0.93 [0.75, 1.15]
ECOG	1.84 [1.08, 3.14]	0.89 [0.71, 1.12]
NCIC	1.83 [0.65, 5.18]	0.80 [0.59, 1.07]
France	2.39 [1.03, 5.51]	0.86 [0.62, 1.19]
MAOP	4.04 [1.71, 9.54]	0.83 [0.58, 1.20]
Jerusalem	1.40 [0.08, 25.14]	1.29 [0.57, 2.91]
Overall	2.19 [1.84, 2.61]	0.90 [0.84, 0.96]

†Observed odds ratios for response for experimental FU *versus* 5FU bolus, with 95% confidence intervals based on the Mantel–Haenszel test (except for the odds ratio for Mar del Plata, which used Gart's (1966) logit estimate with 0.5 correction for zero cells) and hazard ratios for experimental FU *versus* 5FU bolus estimated by using a proportional hazard model, with 95% confidence intervals based on Wald's test. The overall odds ratio was estimated by using a trial-adjusted Mantel–Haenszel estimator. The overall hazard ratio was estimated by using a trial-stratified proportional hazard model.

SD and PD respectively. The binary indicator for treatment (Z) is set to 0 for FU bolus and to 1 for experimental FU.

3.2. Analysis of four-category tumour response

The bivariate model (7) was defined by using the Plackett copula. For survival, proportional hazards model (6) was used, with Weibull trial-specific base-line hazard functions. Tumour response was modelled by using a constrained version of proportional odds model (4). More

specifically, for those trials, for which not all levels of tumour response were observed (see Table 1), all coefficients η_{ik} were constrained to 0.

Under these assumptions, the likelihood function for the observed data (see equation (15) in Appendix A) is fully specified. Maximum likelihood parameter estimates can be obtained by using the Newton–Raphson algorithm. In our example, the algorithm with numerical second-order derivatives, as implemented in SAS-IML, version 6.12 (and higher versions), in the form of a standard routine `NLPNRR` (SAS Institute, 1995), was used.

It should be noted that θ, as defined by equation (14), involves a comparison of survival times of patients classified according to tumour response. It is well known that such a comparison is likely to be length biased, because a response to treatment is not observed instantaneously. As a result, patients who enjoy long survival times are more likely to be responders than non-responders, and therefore the survival of responders is likely to be biased upwards compared with that of non-responders. One method of correcting for length bias in such a comparison is a landmark analysis (Anderson *et al.*, 1983).

In a landmark analysis, only patients who are alive at an arbitrary, prespecified, landmark time are considered, and their response status is assessed at the landmark time. In this way, response is no longer time dependent, and no bias affects the comparison of responders and non-responders. Unfortunately, when the data were originally collected, the time to response was not reported. Hence, it was not possible to reclassify patients' responses at the landmark time. As an approximate solution in this case we excluded patients who died before the landmark time and assumed that all recorded responses had occurred before the landmark.

Theoretically, we might consider using the (trial-specific) planned time of response analysis to reclassify patients' responses at the landmark. Unfortunately, for several trials the original publication of results did not provide appropriate information about the time. Consequently, the planned time of response analysis could not be used in the landmark analysis.

By way of a sensitivity analysis, we conducted the analysis based on the bivariate model (3)–(7) for landmark times ranging from 0 (no correction) to 6 months. Of most interest, however, is the range between 3 and 6 months. That is because tumour response is usually assessed 3–6 months after the beginning of chemotherapy. In fact, this was the case for most of the trials analysed, for which information on the response assessment scheme could be obtained from the original publication of results.

Fig. 2 shows a plot of estimates of θ for different landmark times up to 12 months for the four-category tumour response. Here, the zero value corresponds to the analysis without any correction for length bias. As expected, the estimates decrease, approaching a value of 2 around 10–12 months. The dependence of θ on the landmark time clearly illustrates the need for the correction of the analysis for the length bias. It should be underlined that this need is not due to the particular choice of the method of analysis, but to the nature of the end points considered. In fact, a correction for length bias would most probably have to be considered in any analysis of the validity of tumour response as a surrogate for survival.

Importantly, lower 95% confidence limits for θ at all landmark times in Fig. 2 are greater than 1.7. It might therefore be concluded that length bias, if any, does not induce an association but rather affects the magnitude of an existing association. Moreover, as already mentioned, the landmark times between 3 and 6 months are of most interest. For these time points, estimates of θ remain between 3 and 4.6, with lower 95% confidence interval limits above 2.5. They indicate that the odds for surviving beyond time t for, for example, responders (PR or CR) were at least 2.5 times higher than the odds for non-responders (patients with SD or PD). This suggests that, even after taking into account possible length bias, there remains a considerable association between tumour response and survival time at the level of individual.

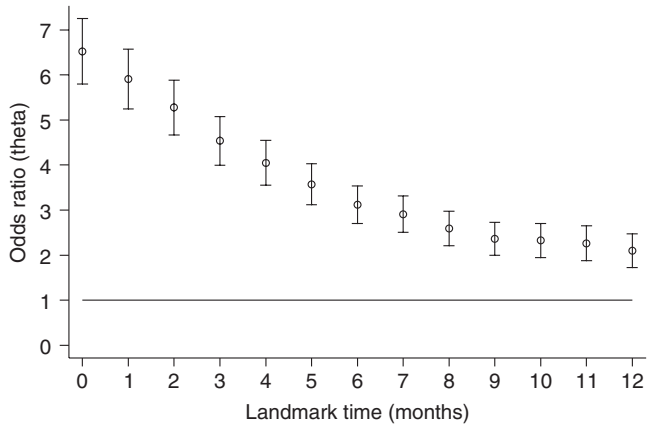

Fig. 2. Estimated individual level association parameter θ, with 95% confidence interval limits, by landmark time

Table 3. Four-category tumour response: individual level association θ and trial level associations $R^2_{\text{trial}(\alpha,\eta)}$ and $R^2_{\text{trial}(\alpha)}$, for various landmark times†

Landmark (months)	*Individual level association* θ	*Trial level associations*	
		$R^2_{\text{trial}(\alpha,\eta)}$	$R^2_{\text{trial}(\alpha)}$
Without adjustment for performance status			
0	6.78 [6.01, 7.55]	0.16 [0, 0.42]	0.16 [0, 0.42]
3	4.59 [4.04, 5.15]	0.15 [0, 0.41]	0.15 [0, 0.41]
4	4.07 [3.56, 4.57]	0.10 [0, 0.34]	0.10 [0, 0.34]
5	3.56 [3.10, 4.03]	0.06 [0, 0.28]	0.05 [0, 0.26]
6	3.09 [2.67, 3.51]	0.08 [0, 0.31]	0.06 [0, 0.28]
With adjustment for performance status			
0	6.50 [5.75, 7.25]	0.22 [0, 0.49]	0.20 [0, 0.49]
3	4.52 [3.96, 5.07]	0.16 [0, 0.42]	0.16 [0, 0.45]
4	4.00 [3.49, 4.51]	0.11 [0, 0.35]	0.11 [0, 0.39]
5	3.50 [3.04, 3.96]	0.07 [0, 0.29]	0.06 [0, 0.32]
6	3.03 [2.62, 3.45]	0.08 [0, 0.31]	0.06 [0, 0.33]

†95% confidence intervals are given in square brackets.

The upper part of Table 3 ('without adjustment for performance status' (PS)) presents estimates of θ, $R^2_{\text{trial}(\alpha,\eta)}$ and $R^2_{\text{trial}(\alpha)}$ for the analysis with no adjustment for length bias (landmark 0) and for landmark times between 3 and 6 months. The estimates of $R^2_{\text{trial}(\alpha,\eta)}$ and $R^2_{\text{trial}(\alpha)}$ were obtained by using models (8) and (10) respectively. The 95% confidence intervals for $R^2_{\text{trial}(\alpha,\eta)}$ and $R^2_{\text{trial}(\alpha)}$ were obtained by finding such values of these parameters, for which the corresponding estimates were equal to the 2.5% and 97.5% quantiles of the cumulative distribution function of R^2 (Fisher, 1928; Algina, 1999). The distribution function was computed by using the algorithm proposed by Ding (1996).

The estimates of $R^2_{\text{trial}(\alpha,\eta)}$ that are presented in the upper part of Table 3 are only slightly higher than those of $R^2_{\text{trial}(\alpha)}$. Thus, not much would be gained in the precision of the prediction

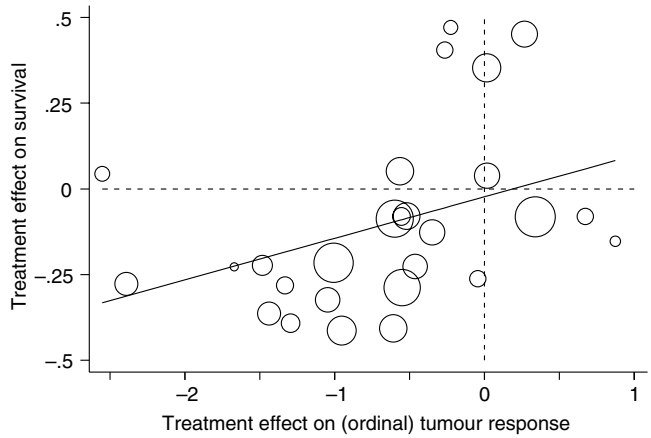

Fig. 3. Estimated trial-specific treatment effects on survival *versus* treatment effects on four-category tumour response

if instead of model (10) the more complex model (8) were used to predict the treatment effect on survival.

Overall, however, the estimates are low and do not exceed 20%. The low association between the estimated trial-specific treatment effects for survival and tumour response can be observed in Fig. 3, which presents the plot of the effects for the analysis using the landmark time of 3 months. The size of each point is proportional to the number of patients in the corresponding trial. The straight line is the prediction from model (10). The estimated slope of the regression line is 0.12 (standard error 0.06). (According to the parameterization that is used in model (3)–(7), $\beta_i > 0$ and $\alpha_i > 0$ indicate increases in the hazard of death and in the odds of non-response respectively, for the experimental treatment.) The line passes very close to the origin. In fact, the estimated intercept is -0.02 (standard error 0.06) and is not significantly different from 0. This suggests a simple multiplicative association between treatment effects for survival and tumour response. Daniels and Hughes (1997) considered this as one of the conditions for a good surrogate. Buyse and Molenberghs (1998) required it for prediction based on the relative effect estimated from a single trial.

The estimates of $R^2_{\text{trial}(\alpha,\eta)}$ and $R^2_{\text{trial}(\alpha)}$ from the upper part of Table 3 suggest that a four-category tumour response is a weak surrogate for survival at the trial level, in that it does not permit reliable predictions of treatment effects on survival. In contrast, the estimates of θ indicate a strong association between tumour response and survival time for individual patients, after adjusting for treatment effects.

One might ask whether taking into account information about prognostic factors would influence the estimates of the trial level R^2 that are shown in the upper part of Table 3. The data collected for the patients included in the four meta-analyses of advanced colorectal cancer trials contained information about PS at randomization. Overall, 41.3% of patients had PS 0, 43.5% had PS 1 and 13.7% had PS 2 (1.5% had missing information on PS). To investigate the extent to which taking into account the information about PS would change the estimates shown in the upper part of Table 3, the two-stage analysis was repeated with PS included as a continuous covariate in the marginal models (4) and (6). The patients with missing PS status were excluded from the analysis. The results are shown in the lower part of Table 3. The 95% confidence intervals for $R^2_{\text{trial}(\alpha)}$ were computed in the same way as those in Table 3. It can be seen that, compared with the upper part of Table 3, the individual level association remains

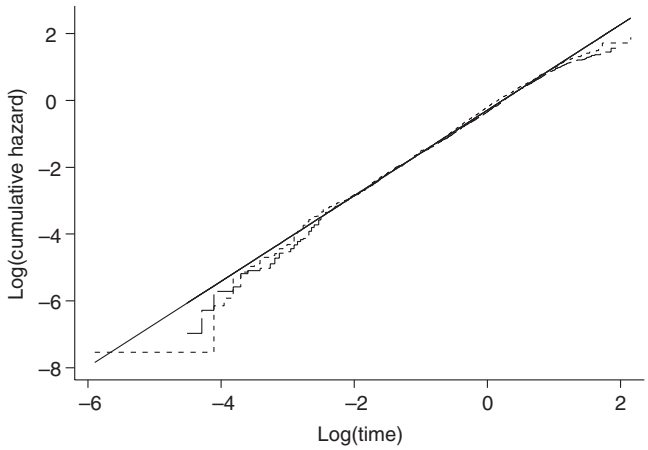

Fig. 4. Estimated (⌐¬, experimental; ⌐¬, control) and predicted (———, experimental; - - - - - - -, control) cumulative hazard functions by treatment group

essentially unchanged, whereas trial level estimates of R^2 increase only slightly. Altogether, the results from Table 3 indicate no substantial increase in the individual or trial level association after adjusting for PS.

It is advisable to conduct model checking. Fig. 4 shows logarithms of Nelson–Aalen (Nelson, 1972; Aalen, 1978) estimates of the cumulative hazard for the experimental and control treatment groups with predictions based on a simple linear regression model. The plots look reasonably linear, justifying the choice of the Weibull distribution for survival.

Additionally, the assumed bivariate Plackett copula model was fitted using a separate association parameter θ for each trial. The analysis was performed for the landmark time of 3 months. It led to the log-likelihood of -6759.15. The log-likelihod for the model corresponding to the second row of the upper part of Table 3 was equal to -6781.86. The resulting difference in deviances is $-2 \times (-22.71) = 45.42$ on 26 degrees of freedom. It suggests ($p = 0.010$) that there might be somewhat more variability in individual level association between the trials than allowed in the model that was used to obtain the results that are presented in Table 3.

A separate issue is the verification of the assumed form of the copula function. For this, some method allowing for a comparison of the goodness of fit of models based on different copula functions, including the Plackett copula, would be needed. At present, however, no such method is known.

3.3. Analysis of binary tumour response

In clinical practice, tumour response is very often used as a binary variable, with patients with CR or PR considered responders and patients with SD or PD considered non-responders. It is therefore of interest to investigate the validity of binary tumour response as a surrogate for survival. The methodology developed can be applied in this case as well. Table 4 presents the corresponding estimates of θ and $R^2_{\text{trial}(\alpha)}$ by landmark time for the analysis without and with the adjustment for PS. The 95% confidence intervals for $R^2_{\text{trial}(\alpha)}$ were computed in the same way as those in Table 3. Note that, for a binary response, proportional odds models (3)–(5) are equivalent to a logistic regression model. In the computations, model (4) was used. In one of the smallest trials, 'Mar del Plata' (see Table 1), no tumour responses in the control arm

Table 4. Binary tumour response: individual level association θ and trial level associations $R^2_{\text{trial}(\alpha,\eta)}$ and $R^2_{\text{trial}(\alpha)}$, for various landmark times†

Landmark (months)	*Individual level association* θ	*Trial level associations*	
		$R^2_{\text{trial}(\alpha,\eta)}$	$R^2_{\text{trial}(\alpha)}$
Without adjustment for performance status			
0	4.91 [4.16, 5.67]	0.46 [0.12, 0.69]	0.44 [0.13, 0.69]
3	3.62 [3.07, 4.17]	0.47 [0.13, 0.70]	0.44 [0.13, 0.69]
4	3.29 [2.78, 3.80]	0.41 [0.08, 0.65]	0.37 [0.08, 0.64]
5	3.01 [2.54, 3.48]	0.36 [0.04, 0.61]	0.32 [0.05, 0.60]
6	2.71 [2.28, 3.14]	0.31 [0.02, 0.58]	0.29 [0.03, 0.57]
With adjustment for performance status			
0	4.78 [4.04, 5.53]	0.47 [0.13, 0.70]	0.46 [0.15, 0.71]
3	3.57 [3.02, 4.13]	0.49 [0.15, 0.71]	0.46 [0.15, 0.71]
4	3.25 [2.74, 3.76]	0.44 [0.10, 0.67]	0.41 [0.11, 0.67]
5	2.97 [2.50, 3.45]	0.39 [0.06, 0.64]	0.35 [0.07, 0.63]
6	2.68 [2.25, 3.12]	0.35 [0.04, 0.61]	0.32 [0.05, 0.60]

†95% confidence intervals are given in square brackets.

were observed at all. This precluded the estimation of the trial-specific treatment effect on the surrogate. Therefore, this trial has been removed from the analysis that is presented in Table 4. The influence of the exclusion on the results of the analysis will be discussed later.

The estimates of $R^2_{\text{trial}(\alpha,\eta)}$ and $R^2_{\text{trial}(\alpha)}$ that are presented in Table 4 do not exceed 50%, irrespectively of the landmark time and the adjustment for the information about PS. They suggest that not more than 50% of the variability in treatment effect on survival could be explained through a treatment effect on the binary tumour response. Consequently, though somewhat better than the four-category response, the binary tumour response would be a poor surrogate for survival at the trial level, in that it would not permit reliable predictions of treatment effects on survival.

The weak association between the estimated trial-specific treatment effects for survival and binary tumour response can be observed in Fig. 5, which presents the plot of the effects for the analysis with the landmark time set to 3 months and without adjustment for PS. The estimated intercept and slope of the straight line, containing the predictions from model (10), are respectively 0.10 (standard error 0.06) and 0.22 (standard error 0.05). It follows that, similar to the case of the four-category response, a simple multiplicative association between treatment effects for survival and binary tumour response can be inferred.

The estimates of θ that are presented in Table 4 indicate a considerable association between binary tumour response and survival time for individual patients, after adjusting for treatment effects.

We used a simple method to provide evidence that the assumed parametric form, applied within the bivariate Plackett copula model, was appropriate. For this, we first fitted a model separately for each treatment arm in each trial, adjusting for length bias by using a landmark time of 3 months. Each model used four parameters—one for association (θ), one for the intercept in the marginal logistic regression for tumour response and two for the Weibull model for survival. This led to a log-likelihood of -4973.0. The log-likelihod for the model corresponding to the second row in the upper part of Table 4 was equal to -5019.0. Consequently, the difference in deviances was $-2 \times (-46.0) = 92.0$ on $208 - 131 = 77$ degrees of freedom and

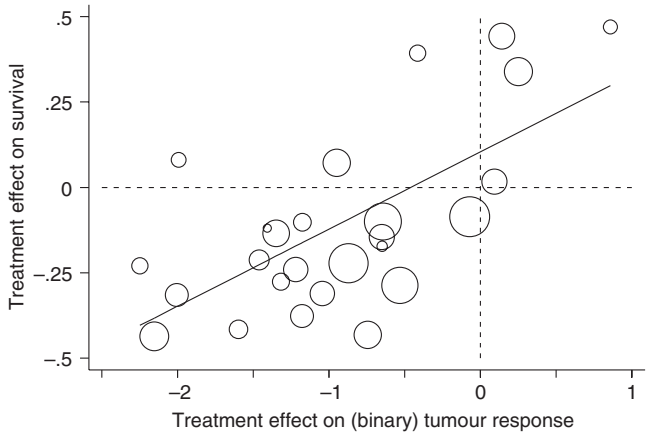

Fig. 5. Estimated trial-specific treatment effects on survival *versus* treatment effects on binary tumour response

it was not significant ($p = 0.117$). The use of the reduced model (assuming a common copula parameter for all trials) to obtain the results presented in Table 4 thus seemed justified.

It is worth noting here that the results of the Mar del Plata trial, which was excluded from the analysis, indicated a large effect of the experimental treatment on the surrogate with virtually no effect on the true end point. Fig. 5 allows us to infer that adding a point corresponding to the raw treatment estimates for the excluded trial (based on the odds ratio and hazard ratio from Table 1) might decrease, rather than increase, the value of $R^2_{\text{trial}(\alpha)}$ presented in the first row of Table 4. The bias resulting from the exclusion of the data for the Mar del Plata trial from the analysis, if any, is thus most probably positive. Consequently, the weak association that is observed at the trial level for the binary response model might be in fact overestimated.

To include the Mar del Plata trial in the analysis of binary tumour response, we might consider using the EM algorithm (Dempster *et al.*, 1977) or multiple imputation (Schafer, 1997). Adapting these methods to the models that are considered in this paper is not straightforward, though. As an alternative, we have considered including the data for the Mar del Plata trial in the analysis (unadjusted for length bias and PS), with an assumed fixed value of the treatment effect on tumour response. The following values of the effect, in terms of the logarithm of the odds ratio of response in favour of the 'experimental' treatment, were considered: $-6, -3, -1, 0, 1, 3, 6$. (Note that the value of 3 is close to the logarithm (2.75) of the crude estimate of the odds ratio (15.68) that is presented in Table 1 for the Mar del Plata trial.) As a result, the following estimates of $R^2_{\text{trial}(\alpha)}$ were obtained: 0.28, 0.38, 0.44, 0.45, 0.44, 0.36 and 0.20 respectively. The differences observed between the coefficients of determination were entirely due to the changes in treatment estimates for the Mar del Plata trial: the estimates for the remaining trials did not essentially change. These results indicate that the exclusion of the trial from the analysis that is presented in Table 4 leads to an overestimation of the trial level R^2, as conjectured in the previous paragraph.

4. Discussion of case-study

The increase in the strength of the trial level association for binary tumour response might raise a question whether using a different dichotomization of the response categories might yield an even bigger increase. To verify this possibility, we performed two additional analyses (without adjusting for length bias or PS). In the first analysis, the tumour response was defined as CR,

Table 5. Tumour response, under various definitions: individual level association θ and trial level associations $R^2_{\text{trial}(\alpha)}$†

Response	θ	$R^2_{\text{trial}(\alpha)}$
Four category	6.78 [6.01, 7.55]	0.16 [0.00, 0.42]
Binary		
CR *versus* PR + SD + PD	7.59 [4.71, 10.48]	0.08 [0.00, 0.51]
CR + PR *versus* SD + PD	4.91 [4.16, 5.67]	0.44 [0.13, 0.69]
CR + PR + SD *versus* PD	8.32 [7.17, 9.47]	0.04 [0.00, 0.28]

†95% confidence intervals are given in square brackets.

with PR, SD or PD regarded a failure (CR *versus* PR + SD + PD; it should be noted that, owing to a small number of CRs, the analysis was based on 12 trials only). In the second analysis, the response was defined as CR, PR or SD, with PD treated as a failure (CR + PR + SD *versus* PD). Table 5 presents the results of the analyses, along with the corresponding result from Table 4 for the conventional dichotomization (CR or PR *versus* SD or PD). It can be seen that the estimates of $R^2_{\text{trial}(\alpha)}$ for the two alternative dichotomizations are much lower than the estimate that is obtained for the conventional binary tumour response.

Table 5 includes also the corresponding result from Table 3 for the four-category response. The strength of the trial level association for the conventionally defined binary tumour response (CR + PR *versus* SD + PD) is remarkably higher than the strength for the other two binary responses or for the four-category response. This is an interesting observation from a practical (clinical) point of view. It is not straightforward to explain this difference. A possible reason might be that, for example, the categorizations other than CR + PR *versus* SD + PD are clinically more difficult to establish and lead to more complicated, than proportional odds, models. (This might also be why for the four-category response some inadequacies of the constant association model were observed, whereas for the conventional binary response the model seemed satisfactory.) This is a topic for more detailed future research.

All of the aformentioned analyses might be subject to bias from another source, though. The results that are presented in Tables 3 and 4 were calculated under the assumption that the true trial-specific treatment effects α_i and β_i were equal to their estimates, $\hat{\alpha}_i$ and $\hat{\beta}_i$ say, obtained from the first-stage copula model (7). Thus, the estimation error that is present in the estimates was ignored. However, it is known that ignoring the measurement error may lead to bias (Fuller, 1987; Carroll *et al.*, 1995). To address this issue, Burzykowski *et al.* (2001), using the developments of van Houwelingen *et al.* (2002), proposed a method of estimation of $R^2_{\text{trial}(\alpha,\eta)}$ and $R^2_{\text{trial}(\alpha)}$ that takes the estimation error $\hat{\alpha}_i$ and $\hat{\beta}_i$ into account. However, the method is numerically involved and its convergence is not guaranteed. Non-convergence can happen, e.g. when there is a substantial estimation error in $\hat{\alpha}_i$ and $\hat{\beta}_i$ (Tibaldi *et al.*, 2003). Unfortunately, in none of the cases considered in Tables 3 and 4 could we obtain adjusted estimates of $R^2_{\text{trial}(\alpha,\eta)}$ or $R^2_{\text{trial}(\alpha)}$. In all cases, the algorithm converged to the boundary of the parameter space. This is probably because several of the trials included in the analysis had small sizes (see Table 1) and, consequently, the estimation error for their trial-specific treatment effects was quite large.

It is worth noting here that the meta-analytic approach to the validation of surrogate end points, as any meta-analysis, simply uses the data from previously organized clinical trials. We might expect that the trials will be powered for the true end point. Of course, the resulting sample sizes—and the treatment estimation errors—will vary, reflecting different assumptions

made at the trials' design stage. Our case-study illustrates this point very well. However, it is difficult to quantify in general the extent of the sample size variability that will still allow for the adjustment of the estimation of trial level R^2 for the error in the estimates of treatment effects. More investigation into this issue is needed.

In this analysis, meta-analytic data were used to investigate the validity of tumour response as a surrogate for survival. As is often the case with real life data, we were faced with several practical complications, like zero frequencies in contingency tables or incomplete information (e.g. about the time of tumour response assessment). The effect of these problems on the final conclusions was assessed by means of a sensitivity analysis. In particular, owing to the nature of the end points considered (tumour response and survival), an adjustment of the analysis for length bias had to be considered. For this, we used a form of a landmark analysis. We found that the strength of the individual level and trial level association depended on the landmark time; irrespectively of the landmark time, however, the individual level association remained substantial, whereas the trial level association was low. It is worth stressing that the dependence should not be seen as a problem with the method of the analysis, but as a feature related to the question asked (about the association between tumour response and survival).

Using the meta-analytic approach we found that tumour response is a poor surrogate for a prediction of the treatment effect *at the trial level*, even though the response is highly prognostic of survival, after adjusting for treatment, *at the individual level*. These results suggest that using tumour response as a surrogate for survival in trials investigating the effect of treatment involving 5FU or FUDR in advanced colorectal cancer may lead to invalid results. This conclusion casts a doubt, at least for the type of treatments that are considered for advanced colorectal cancer, on the Food and Drug Administration's guidelines for accelerated approval of investigational cancer treatments, mentioned in Section 1. Further research is required, however, before a more decisive statement can be reached. We believe that our results can contribute to the process of arriving at the conclusion.

5. Concluding remarks

In this paper, we have studied the validity of tumour response as a surrogate for survival in trials investigating the effect of treatment involving 5FU or FUDR in advanced colorectal cancer. We suggest that the prediction of the treatment effect at the trial level, and the assessment of the quality of such predictions, is central to the problem of surrogate marker validation; in fact, some approaches are based exclusively on trial level information (Daniels and Hughes, 1997). A similar postulate was considered by, for example, Fleming and DeMets (1996). Recently, Begg and Leung (2000) formulated two principles for guiding the choice of a surrogate end point. Their principle 2 requires that the results that are obtained in a trial using a surrogate should be 'concordant' with the results that are obtained by using the true end point. Begg and Leung (2000) did not point to any particular measure for assessing the 'concordance'; in an illustrative example, they considered the probability that treatment effects on surrogate and true end points have the same sign. It is worth noting that, from the point of view of the approach that is used in this paper, a requirement concerning the desired level of the probability of concordance might be translated into a requirement regarding the value of the trial level coefficient of determination R^2 derived under model (10), for instance. Hence, the concept of validating a surrogate based on the assessment of the precision of the prediction of a treatment effect on the true end point at the trial level can be linked to Begg and Leung's (2000) proposal.

From a methodological point of view, this paper proposes a novel concept of using copula models to model mixed bivariate categorical or binary and survival data jointly. It is motivated

by the results that were developed by Molenberghs *et al.* (2001) in the context of mixed discrete and continuous data. One of the models that they considered is equivalent to using the Plackett copula to construct the joint distribution for the continuous outcome and for a latent continuous variable underlying the (observed) discrete outcome. However, Molenberghs *et al.* (2001) did not formulate their results in terms of copula models. The explicit use of copulas, as proposed in this paper, substantially broadens the range of possible models that can be formulated. For instance, it is possible to choose various association structures through the choice of various forms of copula functions. Moreover, the chosen copula can be combined with various models for the marginal distributions for the categorical or binary and survival variables, including the proportional hazards and proportional odds models that are considered in this paper.

The method of validating surrogate end points, described in this paper, extends meta-analytic ideas proposed by Buyse *et al.* (2000a) to the case of an ordinal categorical surrogate end point and a failure time true end point. It offers considerable flexibility. In the analysis of the advanced colorectal cancer trials the Plackett copula was used. This choice was motivated by the natural interpretation of the association parameter θ for this copula. Generally, other copulas can be considered (Oakes, 1989; Shih and Louis, 1995; Nelsen, 1999). In principle, the choice might be guided by the adequacy of fit of the bivariate model (7), using a particular copula that fits the data at hand best. For this, an adaptation of the method of checking the goodness of fit of Archimedean copulas to bivariate survival data, proposed recently by Wang and Wells (2000), might be developed. This is an important topic for future research.

It is also worth noting that, although in the application that was considered in this paper the true end point was assumed to have a Weibull distribution, it is possible to use other distributional assumptions, or even to use a semiparametric approach with unspecified base-line hazard functions (Shih and Louis, 1995), while maintaining the copula value.

In our evaluation, a subjective assessment was required about what values of R^2 or θ are sufficiently 'high' for the candidate surrogate to be deemed acceptable. On purely theoretical grounds it is difficult to propose 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 that was proposed by Buyse *et al.* (2000a). For obvious reasons such an experience is thus far very limited. Taking the above into account, observed values of R^2_{trial} below 0.5 have been judged 'not close to 1'. Such subjectivity will be less of an issue if several end points are evaluated simultaneously as candidate surrogates for the same true end point. However, the possibility of an assessment of strength of evidence for validity of a surrogate can be seen as an advantage of the method proposed by Buyse *et al.* (2000a), especially when compared, for example, with the rigid 'yes' or 'no' decision rule that is implied by the Prentice's (1989) definition.

At first sight, it is striking to see that the individual level and trial level surrogacies can be dramatically different. Two comments are appropriate. First, it is easily seen from the normal outcomes hierarchical model (Section 2.1 and Buyse *et al.* (2000a)) that these surrogacies are based on different components of variability. The individual level surrogacy indicates how a subject's two measurements covary, whereas the trial level surrogacy is directed towards the joint behaviour of the subjects within a trial, based on their treatment allocation. For example, if one group of treatment arm subjects responds on the surrogate, whereas another does so on the true end point, and such behaviour is seen across trials, then we might have a substantial trial level surrogacy but a small individual level surrogacy. The clinical trialist will primarily be interested in the trial level surrogacy. When we are interested, for example, in predicting the behaviour of a given patient (e.g. for prophylactic reasons), then the individual level surrogacy will be the more relevant quantity.

From the point of view of assessing trial level validity, especially when the reduced model (10) is used at the second stage, the meta-analytic approach, as described in this paper, can be seen as corresponding to the model that was proposed by Daniels and Hughes (1997). An important difference is that our approach permits a simultaneous study of the quality of the surrogate at the trial level and at the individual level. The latter aspect of surrogacy was not considered by Daniels and Hughes (1997) at all. If we are less interested in the individual level surrogacy, however, our approach can be simplified, e.g. by using only marginal models for S and T at the first stage (Tibaldi *et al.*, 2003), which would amount to following the proposal of Daniels and Hughes (1997).

Acknowledgements

The authors are grateful to the Meta-Analysis Group in Cancer for permission to use their individual patient data. The first author gratefully acknowledges support from Bijzonder Onderzoeksfonds Limburgs Universitaire Centrum.

Appendix A: The construction of the likelihood function

Assume that the joint cumulative distribution of T_{ij} and $\tilde{S}_{ij}$, given $Z_{ij} = z$, can be described by the copula model (7), with the marginal cumulative distribution functions given by marginal proportional hazard model (6) and proportional odds model (3). The bivariate density $g_{ij}(t, k; z)$ for T_{ij} and S_{ij}, given $Z_{ij} = z$, can then be specified by taking

$$g_{ij}(t,k;z) = \frac{\partial F_{T_{ij},\tilde{S}_{ij}}(t,\gamma_{ik};z)}{\partial t} - \frac{\partial F_{T_{ij},\tilde{S}_{ij}}(t,\gamma_{i(k-1)};z)}{\partial t}.$$

Consequently, we can define

$$\begin{aligned} G_{ij}(t,k;z) &\equiv P(T_{ij} \geqslant t, S_{ij} = k | Z_{ij} = z) \\ &= F_{\tilde{S}_{ij}}(\gamma_{ik};z) - F_{\tilde{S}_{ij}}(\gamma_{i(k-1)};z) - \{F_{T_{ij},\tilde{S}_{ij}}(t,\gamma_{ik};z) - F_{T_{ij},\tilde{S}_{ij}}(t,\gamma_{i(k-1)};z)\}. \end{aligned}$$

As a result, for the observed data ($X_{ij} = x_{ij}, \Delta_{ij} = \delta_{ij}, S_{ij} = s_{ij}, Z_{ij} = z_{ij}$), the log-likelihood can be expressed as

$$\sum_{i,j} [\delta_{ij} \log\{g_{ij}(x_{ij}, s_{ij}; z_{ij})\} + (1 - \delta_{ij}) \log\{G_{ij}(x_{ij}, s_{ij}; z_{ij})\}]. \tag{15}$$

Appendix B: The bivariate Plackett copula

Let $Y = (Y_1, Y_2)$ be a bivariate random variable with joint distribution function $F(y_1, y_2)$ and marginal distributions $F_j(y_j)$ $(j = 1, 2)$. The global cross-ratio function $\theta(y_1, y_2)$ is defined by

$$\theta(y_1, y_2) = \frac{p_{11}p_{22}}{p_{12}p_{21}} = \frac{F(1 - F_1 - F_2 + F)}{(F_1 - F)(F_2 - F)} \tag{16}$$

with $F_j \equiv F_j(y_j)$ $(j = 1, 2)$ and $F \equiv F(y_1, y_2)$. The cross-ratio satisfies $0 \leqslant \theta(y_1, y_2) \leqslant \infty$ when $F(y_1, y_2)$ satisifes the Fréchet (1951) bounds. The components p_{ij} in equation (16) are the quadrant probabilities in $\mathbb{R}^2$ with vertex at (y_1, y_2). The Plackett distribution is obtained for constant cross-ratio $\theta(y_1, y_2) \equiv \theta$ (Plackett, 1965; Mardia, 1970). Equation (16) is the defining equation for F, given that F_1, F_2 and θ are known.

Given the marginal distribution function F_1 and F_2 and the cross-ratio θ the values of the Plackett distribution functions are found as one of the two solutions of the second-degree polynomial equation

$$\theta(F - a_1)(F - a_2) - (F - b_1)(F - b_2) = 0,$$

where $a_1 = F_1$, $a_2 = F_2$, $b_1 = 0$ and $b_2 = F_1 + F_2 - 1$. Its solution is given explicitely by for example Dale (1986) and Mardia (1970):

$$F(y_1, y_2) = C_\theta\{F_1(y_1), F_2(y_2), \theta\},$$

where

$$C_\theta(u, v, \theta) = \begin{cases} \dfrac{1 + (u + v)(\theta - 1) - S_\theta(u, v)}{2(\theta - 1)} & \text{if } \theta \neq 1, \\ uv & \text{otherwise} \end{cases} \tag{17}$$

and

$$S_\theta(u, v) = \sqrt{[\{1 + (\theta - 1)(u + v)\}^2 + 4\theta(1 - \theta)uv]}. \tag{18}$$

It immediately follows from Mardia (1970), chapter 8, that $C_\theta(u, v)$ always is a 2-copula, with θ in $[0, \infty]$.

A detailed discussion of the bivariate Plackett distribution can be found in Mardia (1970), where it is called a 'contingency-type distribution'. A thorough discussion of copulas is presented in Nelsen (1999).

References

Aalen, O. O. (1978) Nonparametric inference for a family of counting processes. *Ann. Statist.*, **6**, 701–726.

Advanced Colorectal Cancer Meta-Analysis Project (1992) Modulation of 5-fluorouracil by leucovorin in patients with advanced colorectal cancer: evidence in terms of response rate. *J. Clin. Oncol.*, **10**, 896–903.

Advanced Colorectal Cancer Meta-Analysis Project (1994) Meta-analysis of randomized trials testing the biochemical modulation of 5-fluorouracil by methotrexate in metatstatic colorectal cancer. *J. Clin. Oncol.*, **12**, 960–969.

Algina, J. (1999) A comparison of methods for constructing confidence intervals for the squared multiple correlation coefficient. *Multiv. Behav. Res.*, **34**, 494–504.

Anderson, J. R., Cain, K. C. and Gelber, R. D. (1983) Analysis of survival by tumour response. *J. Clin. Oncol.*, **1**, 710–719.

Begg, C. B. and Leung, D. H. Y. (2000) On the use of surrogate end points in randomized trials. *J. R. Statist. Soc.* A, **163**, 15–24.

Burzykowski, T., Molenberghs, G., Buyse, M., Geys, H. and Renard, D. (2001) Validation of surrogate end points in multiple randomized clinical trials with failure time end points. *Appl. Statist.*, **50**, 405–422.

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. (2000a) The validation of surrogate endpoints in meta-analysis of randomized experiments. *Biostatistics*, **1**, 49–68.

Buyse, M. and Piedbois, P. (1996) On the relationship between response to treatment and survival. *Statist. Med.*, **15**, 2797–2812.

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

Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995) *Measurement Error in Nonlinear Models*. London: Chapman and Hall.

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.

Committee for Proprietary Medicinal Products (2001) *Note for Guidance on Evaluation of Anticancer Medicinal Products in Man*. London: European Agency for the Evaluation of Medicinal Products.

Dale, J. R. (1986) Global cross ratio models for bivariate, discrete, ordered responses. *Biometrics*, **42**, 909–917.

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

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1997) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1–38.

Ding, C. G. (1996) On the computation of the distribution of the square of the sample multiple correlation coefficient. *Comput. Statist. Data Anal.*, **22**, 345–350.

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

Fisher, R. A. (1928) The general sampling distribution of the multiple correlation coefficient. *Proc. R. Soc.*, **121**, 654–673.

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

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.

Food and Drug Administration (1996) Reinventing the regulation of cancer drugs—accelerating approval and expanding access. *National Performance Review, 1996404883/41014*. US Government Printing Office, Washington DC.

Fréchet, M. (1951) Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon* A, ser. 3, **14**, 53–77.

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

Fuller, W. A. (1987) *Measurement Error Models*. New York: Wiley.

Gart, J. J. (1966) Alternative analyses of contingency tables. *J. R. Statist. Soc.* B, **28**, 164–179.

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.

van Houwelingen, H. C., Arends, L. R. and Stijnen, T. (2002) Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statist. Med.*, **21**, 589–624.

Laird, N. M. and Ware, J. H. (1982) Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Mardia, K. V. (1970) *Families of Bivariate Distributions*. London: Griffin.

Meta-Analysis Group in Cancer (1996) Reappraisal of hepatic arterial infusion in the treatment of nonresectable liver metastases from colorectal cancer. *J. Natn Cancer Inst.*, **88**, 252–258.

Meta-Analysis Group in Cancer (1998) Efficacy of intravenous continuous infusion of 5-fluorouracil compared with bolus administration in patients with advanced colorectal cancer. *J. Clin. Oncol.*, **16**, 301–308.

Molenberghs, G., Buyse, M., Geys, H., Renard, D. and Burzykowski, T. (2003) Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Contr. Clin. Trials*, **23**, 607–625.

Molenberghs, G., Geys, H. and Buyse, M. (2001) Evaluation of surrogate endpoints in multiple randomized clinical trials with mixed discrete and continuous outcomes. *Statist. Med.*, **20**, 3023–3038.

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

Nelson, W. (1972) Theory and applications of hazard plotting for censored failure data. *Technometrics*, **14**, 945–965.

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

Oye, R. and Shapiro, M. F. (1984) Does response make a difference in patient survival? *J. Am. Med. Ass.*, **252**, 2722–2725.

Plackett, R. L. (1965) A class of bivariate distributions. *J. Am. Statist. Ass.*, **60**, 516–522.

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

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

Schafer, J. L. (1997) *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.

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

Tibaldi, F. S., Cortiñas Abrahantes, J., Molenberghs, G., Renard, D., Burzykowski, T., Buyse, M., Parmar, M., Stijnen, T. and Wolfinger, R. (2003) Simplified hierarchical linear models for the evaluation of surrogate endpoints. *J. Statist. Computn Simuln*, **73**, 643–658.

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

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

World Health Organization (1979) WHO handbook for reporting results of cancer treatment. *WHO Offset Publication 48*. World Health Organization, Geneva.