STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:2175–2197 (DOI: 10.1002/sim.1203)

Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects

Patrick Royston*,† and Mahesh K. B. Parmar

Cancer Division, MRC Clinical Trials Unit, 222 Euston Road, London NW1 2DA, UK

SUMMARY

Modelling of censored survival data is almost always done by Cox proportional-hazards regression. However, use of parametric models for such data may have some advantages. For example, non-proportional hazards, a potential difficulty with Cox models, may sometimes be handled in a simple way, and visualization of the hazard function is much easier. Extensions of the Weibull and log-logistic models are proposed in which natural cubic splines are used to smooth the baseline log cumulative hazard and log cumulative odds of failure functions. Further extensions to allow non-proportional effects of some or all of the covariates are introduced. A hypothesis test of the appropriateness of the scale chosen for covariate effects (such as of treatment) is proposed. The new models are applied to two data sets in cancer. The results throw interesting light on the behaviour of both the hazard function and the hazard ratio over time. The tools described here may be a step towards providing greater insight into the natural history of the disease and into possible underlying causes of clinical events. We illustrate these aspects by using the two examples in cancer. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: survival analysis; proportional hazards; proportional odds; splines; parametric models

1. INTRODUCTION

Modelling of censored survival data by medical researchers and practical biostatisticians is these days almost always done by using Cox regression. About a decade ago, in a mathematically detailed survey of inference in parametric survival models, Hjort [1] pointed out:

> 'The success of Cox regression has perhaps had the unintended side-effect that practitioners too seldomly invest efforts in studying the baseline hazard... A parametric version [of the Cox model], ... if found to be adequate, would lead to more precise estimation of survival probabilities and ... concurrently contribute to a better understanding of the phenomenon under study.'

*Correspondence to: Patrick Royston, Cancer Division, MRC Clinical Trials Unit, 222 Euston Road, London NW1 2DA, UK.
†E-mail: patrick.royston@ctu.mrc.ac.uk

Received January 2001
Accepted January 2002

This apparently little heeded statement neatly summarizes our main motivation in the present paper. For example, visualization of the survival function for censored survival data is easily done by using the Kaplan–Meier plot. In the Cox model, however, the baseline hazard function, which may be estimated by the method of Kalbfleisch and Prentice (reference [2], equation (4.23), p. 85) among others, is treated as a high-dimensional nuisance parameter and is highly erratic. However, the behaviour of the hazard function is of potential medical interest because it is directly related to the time-course of an illness. To estimate it informatively (that is, smoothly), some type of parametric model may be appropriate. For example, Gelfand *et al.* [3] proposed a parametric method of estimating hazard functions within a proportional hazards framework based on a mixture of Weibull distributions. They illustrated its use in a small data set of times to first hospitalization following a diagnosis of AIDS. Other recent contributions to 'hazard regression' include Kooperberg *et al.* [4] and Rosenberg [5].

A second issue is how to deal with non-proportional hazards which may occur, for example, when modelling prognostic factors in studies with medium or long follow-up times. An interesting example in ovarian cancer is given by Valsecchi *et al.* [6]. Although the Cox model may be extended to allow for non-proportional hazards, for example by incorporating time-varying regression coefficients (for example, Hastie and Tibshirani [7] and Hess [8]), there is no 'natural', widely accepted approach, and obtaining a satisfactory model can be complicated. There are further concerns about the complexity involved in the practical interpretation of the coefficients and in the robustness of such models.

A third issue, mentioned also by Gelfand *et al.* [3], arises in the validation of a survival model. Since the baseline hazard function in a Cox model may be regarded as a high-dimensional, very overfitted estimate, the fitted model is too closely adapted to the data at hand. It is unclear how to validate such models in an independent data set. A useable probabilistic specification of the data set requires a parsimonious parametric approximation to the data-generating process.

The proportional hazards model is well known, but the proportional odds model for survival data, potentially a competitor to the Cox model, also has a fairly long history. It was first described in a semi-parametric framework by Bennett [9] and was further developed by several authors, including importantly by Yang and Prentice [10]. Rossini and Tsiatis [11] adapted the semi-parametric framework for modelling current status (that is, interval censored) data.

Here, we will develop flexible parametric models based initially on the assumption of either proportional hazards or proportional odds scaling of covariate effects. Generically, the class of such models is based on transformation of the survival function by a link function $g(.)$

$$g[S(t;\mathbf{z})] = g[S_0(t)] + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{z} \tag{1}$$

where $S_0(t) = S(t;\mathbf{0})$ is the baseline survival function and $\boldsymbol{\beta}$ is a vector of parameters to be estimated for covariates $\mathbf{z}$. Within this framework, Younes and Lachin [12] used the parameterized link function of Aranda-Ordaz [13]

$$g(x;\theta) = \log \frac{x^{-\theta} - 1}{\theta}$$

where $\theta = 1$ corresponds to the proportional odds model and $\theta \to 0$ to the proportional hazards model. Younes and Lachin [12] tackled the estimation problem by using B-splines to estimate

the baseline hazard function and determined $S_0(t)$ by integration. A related and computationally complex proposal is that of Shen [14], in which so-called 'sieve' maximum likelihood and monotone splines with variable orders and knots are used to estimate rather general proportional odds models.

Our approach will be to use natural cubic splines to model $g[S_0(t)]$ within the Aranda-Ordaz family of link functions. We chose to work only with the odds ($\theta = 1$) and hazards ($\theta \to 0$) scaling, rather than with more general values of θ for which the interpretation of covariate effects is obscure. Models with a probit link function which extend the log-normal distribution are also possible, but are not pursued in detail, relevant formulae being given in Appendix A. We smooth the transformed survival function rather than the hazard function [12] because we anticipated that 'end effects', artefacts in the fitted spline functions at the extremes of the time range, would be more severe for the hazard function. When smoothing of $g[S_0(t)]$ is implemented on the log time scale, the fitted function is typically gently curved or nearly linear, and is usually very smooth. The smoothness tends to reduce the chance of artefacts in the estimated hazard function. The estimate of $g[S_0(t)]$ must theoretically be monotone in t, whereas natural cubic splines, which are constrained to be linear beyond certain extreme observations, are not globally monotone. However, the linearity constraint imposes monotonicity in the tail regions where the observed data are sparse, whereas in regions where data are dense (and provided the sample size is not too small), monotonicity is effectively imposed by the data themselves. As a result, the use of natural cubic splines appears virtually to guarantee monotonicity in data sets of any reasonable size (for example, 50 uncensored observations). The use of monotone splines, as in Shen [14], makes the computational problem much more difficult and awkward, a cost we do not feel is in general justified.

When the relationship between the baseline log cumulative hazard or log cumulative odds of failure and log time is modelled as linear rather than by using splines, our approach reduces to fitting Weibull or log-logistic distributions by maximum likelihood. The Weibull is of course familiar as a model for lifetimes. The (generalized) log-logistic distribution has been used by Mackenzie for survival modelling [15] and for modelling recurrent event data [16].

We also describe extensions of our basic models to include models with non-proportional scaling for some subset of the covariates. The covariates may be of any type (binary, categoric or continuous). Non-proportionality is induced by multiplicative interactions between the covariates and the spline basis functions. One should be aware that as with the extended Cox model mentioned above, such models are more complex and require extra care in construction, evaluation of appropriateness and interpretation. The proportionately scaled models are much simpler and may be more robust. Even under statistically significant but quantitatively mild departures from proportionality, they may give a description of the data adequate for practical purposes. Based on the extensions, a hypothesis test of the appropriateness of the scale chosen for the covariate effects is proposed.

The methods are illustrated in two cancer data sets: a study of lymph node-positive primary breast cancer in which prognostic factors are of interest, and a randomized controlled trial of two chemotherapy regimens in advanced bladder cancer, in which modelling the treatment effect is considered. These data sets are introduced first. In Section 3 we describe the methodology, including the models, parameter estimation and model selection. Section 4 introduces the extensions and the test of scale. Section 5 presents detailed analyses of the two data sets, focusing on estimation of the hazard function and hazard ratio. Section 6 is a discussion.

2. EXAMPLE DATA SETS

2.1. Node-positive primary breast cancer

From July 1984 to December 1989, the German Breast Cancer Study Group recruited 720 patients with primary node positive breast cancer into a 'comprehensive cohort study' (Schmoor *et al.* [17]). The recurrence-free survival time of 686 patients with complete data for the standard prognostic factors age (X_1), menopausal status, tumour size, tumour grade ($X_{4a} = 1$ if grade $\geqslant 2$, 0 otherwise), number of positive lymph nodes (X_5), progesterone receptor concentration (X_6) and oestrogen receptor concentration is analysed in this paper. (For clarity, we have used Sauerbrei and Royston's [18] variable names.) A total of 299 events accrued during the follow-up period (median nearly 5 years). Sauerbrei and Royston [18] proposed three multivariable prognostic models derived from different approaches, all estimated within the Cox proportional-hazards framework and all adjusted for hormonal treatment ($X_8 = 1$ if treated, 0 otherwise). Here we consider only model III, in which the functional form of the strongest predictor (X_5) was constrained to be monotonic with an asymptote. The estimated linear predictor (log hazard ratio) in model III was

$$1.79(X_1/50)^{-2} - 8.02(X_1/50)^{-0.5} + 0.50X_{4a} - 1.98\exp(-0.12X_5) - 0.058(X_6+1)^{0.5} - 0.394X_8$$

(Sauerbrei and Royston, reference [18], Table 4). For present purposes, we created three equal-sized prognostic groups (representing good, medium and poor outcome) according to cutpoints placed at the tertiles of this linear predictor. The models examined here will be based only on two dummy variables representing these three groups, the reference group being patients with a good outcome. The original prognostic factors will not be considered further.

2.2. Advanced bladder cancer

Transitional cell carcinomas usually occur in the bladder and are a source of much morbidity and mortality. Patients with metastatic disease have a particularly poor prognosis, with less than 5 per cent surviving beyond 5 years. The British Medical Research Council (MRC) conducted a UK-based randomized controlled clinical trial comparing methotrexate and vinblastine (MV) with cisplatin, methotrexate and vinblastine (CMV) in advanced transitional cell carcimoma (Mead *et al.* [19]). From April 1991 to June 1995, 214 patients were entered by 16 centres, 108 randomized to CMV and 106 to MV. The hazard ratio of death was reported to be 0.68 (CI 0.51–0.90) in favour of CMV. Taking into account follow-up to May 2000, 205 patients (95.8 per cent) have died. Mead *et al.* [19] found that WHO performance status and extent of disease were significant pre-treatment prognostic factors, but for present purposes we shall consider modelling only the relative effects of the two treatments, the hazard function and the hazard ratio within a parametric framework.

3. BASICS OF METHODOLOGY

3.1. Proportional odds and proportional hazards models

The general proportional hazards ('Cox') model for survival data with covariate vector $\mathbf{z}$ is defined through the hazard function $h(t;\mathbf{z})$ as

$$h(t;\mathbf{z}) = h_0(t)\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z})$$

where $h_0(t) = h(t; \mathbf{0})$ is the baseline hazard function. The model may be written in integrated form as

$$H(t; \mathbf{z}) = \left(\int_0^t h_0(u)\, \mathrm{d}u \right) \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}) = H_0(t) \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z})$$

where $H(t; \mathbf{z})$ is the cumulative hazard function. By analogy, the general proportional (cumulative) odds model with covariate vector $\mathbf{z}$ (Bennett [9]) may be defined as

$$O(t; \mathbf{z}) = \frac{1 - S(t; \mathbf{z})}{S(t; \mathbf{z})} = O_0(t) \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z})$$

where $O_0(t) = O(t; \mathbf{0})$ and $O(t; \mathbf{z})$ is the odds of an event occurring in $(0, t)$ for an individual with covariate vector $\mathbf{z}$. Some simple algebra shows that the derivative $\frac{\mathrm{d}O(t;\mathbf{z})}{\mathrm{d}t}$ equals $\frac{h(t;\mathbf{z})}{S(t;\mathbf{z})}$, the ratio of the hazard and survival functions. Covariates in the model act multiplicatively on the odds of an event, as with the more familiar logistic regression model for a binary outcome Y, where for example $O_0(t)$ is replaced with the baseline odds $\Pr(Y = 1 \mid \mathbf{z} = 0)/\Pr(Y = 0 \mid \mathbf{z} = 0)$.

3.2. Survival, density and hazard functions

As mentioned in the Introduction, we propose to estimate the hazard, density and survival functions by smoothing either the cumulative odds function or the cumulative hazard function. With the notation of the previous section, but for the time being suppressing $\mathbf{z}$, suppose that T is a random variable having a log-logistic distribution with location parameter μ and scale parameter σ. Let $x = \ln t$. The survival function for T is

$$S(t) = \left[1 + \exp\left(\frac{x - \ln \mu}{\sigma} \right) \right]^{-1}$$

Therefore

$$\ln O(t) = \ln \frac{1 - S(t)}{S(t)} = \frac{x - \ln \mu}{\sigma} = \gamma_0 + \gamma_1 x$$

where $\gamma_0 = -(\ln \mu)/\sigma$, $\gamma_1 = 1/\sigma$, is linearly related to x. In general, the distribution of T will not of course be log-logistic, and the log cumulative odds function will then be related to x by some non-linear function $s \equiv s(x; \boldsymbol{\gamma})$, having adjustable parameter vector $\boldsymbol{\gamma}$. The survival, density and hazard functions are then

$$S(t) = (1 + \exp s)^{-1}$$

$$f(t) = \frac{\mathrm{d}s}{\mathrm{d}t} \exp(s)(1 + \exp s)^{-2}$$

$$h(t) = \frac{\mathrm{d}s}{\mathrm{d}t} \exp(s)(1 + \exp s)^{-1}$$

Suppose now that T has a Weibull distribution with 'characteristic life' μ and shape parameter p (or scale parameter $\sigma = p^{-1}$). Let the cumulative hazard function be $H(t) = -\ln S(t)$. Then

$$\ln H(t) = \ln \left[\left(\frac{t}{\mu} \right)^p \right] = px - p \ln \mu = \frac{x - \ln \mu}{\sigma} = \gamma_0 + \gamma_1 x$$

which is linear in x. If the distribution of T departs from a Weibull then $\ln H(t)$ will be related to x by a non-linear function $s \equiv s(x;\boldsymbol{\gamma})$. The survival, density and hazard functions are

$$S(t) = \exp(-\exp s)$$

$$f(t) = \frac{\mathrm{d}s}{\mathrm{d}t}\exp(s - \exp s)$$

$$h(t) = \frac{\mathrm{d}s}{\mathrm{d}t}\exp(s)$$

3.3. *Spline-based parametric survival models*

Since the distribution of survival times may be neither log-logistic nor Weibull, more flexible models are needed. The approach taken here is to model the logarithm of the baseline cumulative odds or hazard function as a 'natural' cubic spline function of log time. Therefore, the general function $s(x;\boldsymbol{\gamma})$ of Section 3.2 is to be approximated by a spline. Following the notation of transformation models given in the Introduction, for the PH spline model with fixed covariate vector $\mathbf{z}$ we have

$$g[S(t;\mathbf{z})] = \ln[-\ln S(t;\mathbf{z})] = \ln H(t;\mathbf{z}) = \ln H_0(t) + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{z} = s(x;\boldsymbol{\gamma}) + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}$$

whereas for the PO spline model

$$g[S(t;\mathbf{z})] = \ln[S(t;\mathbf{z})^{-1} - 1] = \ln O(t;\mathbf{z}) = \ln O_0(t) + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{z} = s(x;\boldsymbol{\gamma}) + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}$$

Therefore in summary

$$\left.\begin{array}{ll}\text{PH spline model:} & \ln H(t;\mathbf{z}) \\ \text{PO spline model:} & \ln O(t;\mathbf{z})\end{array}\right\} = s(x;\boldsymbol{\gamma}) + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}$$

Note that since $\ln[-\ln S(t;\mathbf{z})] = \ln H(t;\mathbf{z})$, a PH model may also be regarded as one in which the covariates act on the complementary log-log probability of an event in $(0,t)$. Also, due to the non-linear transformation of the log time-scale, the metric for covariate effects with these models is the log (cumulative) hazard, complementary log-log probability or log-odds scale. The accelerated failure time interpretation is not available for the general spline models, since the Weibull (or log logistic) are the only PH (or PO) models that may be written in AFT form.

Natural cubic splines are defined as cubic splines constrained to be linear beyond boundary knots $k_{\min}$, $k_{\max}$. Such knots are usually, but not necessarily, placed at the extreme observed x-values. In addition, m distinct internal knots $k_1 < \cdots < k_m$ with $k_1 > k_{\min}$ and $k_m < k_{\max}$ are specified. A natural cubic spline may be written as

$$s(x;\boldsymbol{\gamma}) = \gamma_0 + \gamma_1 x + \gamma_2 v_1(x) + \cdots + \gamma_{m+1} v_m(x)$$

where the jth basis function is defined for $j = 1,\ldots,m$ as

$$v_j(x) = (x - k_j)_+^3 - \lambda_j(x - k_{\min})_+^3 - (1 - \lambda_j)(x - k_{\max})_+^3$$

Table I. Internal knot placement for spline models with different degrees of freedom (d.f.). 'Centiles' are centiles of the distribution of the uncensored log survival times (for example, 50 denotes the median).

d.f.	Centiles		
2	50		
3	33	67	
4	25	50	75

and

$$\lambda_j = \frac{k_{\max} - k_j}{k_{\max} - k_{\min}}$$

$$(x-a)_+ = \max(0, x-a)$$

See Appendix B for algebraic details. The curve complexity is governed by the number of degrees of freedom (d.f.) which ignoring γ_0 equals $m+1$. By convention, $m=0$ (or d.f. $=1$) is taken to mean that no internal and no boundary knots are specified, that is, $s(x;\boldsymbol{\gamma}) = \gamma_0 + \gamma_1 x$. The baseline distribution is then Weibull or log logistic.

The placement of the internal knots is an issue (see for example the discussion by Durrleman and Simon [20]). Fortunately, however, 'optimal' (optimized) knot selection does not appear critical for a good fit and may even be undesirable in that the fitted curve may follow small-scale features of the data too closely (see also Section 3.6). Furthermore, uncertainty in parameter estimates due to knot optimization should be reflected in enlarged standard errors and confidence intervals, but this is not straightforward to do. We chose to place boundary knots at the extreme uncensored log survival times. For internal knots, we selected the centile-based positions given in Table I. The positions are essentially those recommended by Durrleman and Simon [20]. The rationale for choosing knots not too far from the median uncensored log survival time is to allow the data to be most closely modelled in the region of greatest density and hence usually of lowest variance. Models with d.f. >4 were not entertained since we expected the resulting curves to be potentially unstable. If desired, however, Table I could be extended to higher d.f.

3.4. Estimation

Estimation is by full maximum likelihood. Suppose the sample comprises n independent observations $\{t_i, \delta_i, \mathbf{z}_i\}_{i=1,\ldots,n}$, where δ_i is 0 for a right-censored observation and 1 for a complete one. Using an abbreviated notation, let the likelihood function for a typical observation be l, so that the likelihood for the whole sample is $\Pi_{\text{sample}}\, l$. Let $\eta = s(x;\boldsymbol{\gamma}) + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}$. For models based on smoothing the log cumulative odds function, we have

$$l = \begin{cases} \dfrac{1}{t}\dfrac{\mathrm{d}s(x;\boldsymbol{\gamma})}{\mathrm{d}x}\exp(\eta)(1+\exp\eta)^{-2} & \text{for an uncensored observation} \\ (1+\exp\eta)^{-1} & \text{otherwise} \end{cases}$$

For models based on smoothing the log cumulative hazard function, we have

$$l = \begin{cases} \dfrac{1}{t}\dfrac{\mathrm{d}s(x;\boldsymbol{\gamma})}{\mathrm{d}x}\exp(\eta - \exp\eta) & \text{for an uncensored observation} \\ \exp(-\exp\eta) & \text{otherwise} \end{cases}$$

Also

$$\frac{\mathrm{d}s(x;\boldsymbol{\gamma})}{\mathrm{d}x} = \gamma_1 + \sum_{j=2}^{m}\gamma_j\frac{\mathrm{d}v_j(x)}{\mathrm{d}x}$$

$$= \gamma_1 + \sum_{j=2}^{m}\gamma_j\left[3(x-k_j)_+^2 - 3\lambda_j(x-k_{\min})_+^2 - 3(1-\lambda_j)(x-k_{\max})_+^2\right]$$

To obtain maximum likelihood (ML) estimates, suitable starting values for $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are needed. Initial guesses at $O(t;\mathbf{z})$ and $H(t;\mathbf{z})$ were found by first fitting a standard (semi-parametric) Cox model with covariates $\mathbf{z}$. The survival function $\hat{S}(t_i;\mathbf{z}_i)$ at the ith observation was estimated as the baseline survival function at t_i raised to the power of the estimated relative hazard, $\exp(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{z}_i)$. For the PH spline model, the initial guess at $H(t_i;\mathbf{z}_i)$ was $-\ln\hat{S}(t_i;\mathbf{z}_i)$. For the PO spline model, the initial guess at $O(t_i;\mathbf{z}_i)$ was $[1-\hat{S}(t_i;\mathbf{z}_i)]/\hat{S}(t_i;\mathbf{z}_i)$. Starting values of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ were determined by simple least-squares regression of the logarithms of these initial cumulative hazards or cumulative odds values on the x_i, the $\mathbf{z}_i$ and the spline basis functions with the desired number of knots. Only the uncensored failure times were used in these initial computations. Full ML estimation was then performed by using the `ml` routine in the package Stata (StataCorp [21]). Depending of course on model complexity, convergence was usually achieved in about 3–10 iterations.

3.5. *Interval estimation*

Asymptotic standard errors of the fitted values $\hat{\eta}_i$ were obtained in standard fashion from the information matrix for $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\beta}}$ following ML estimation. Pointwise 95 per cent confidence intervals for the log cumulative odds function or the log cumulative hazard function were determined as $\hat{\eta}_i \pm 1.96\,\widehat{\mathrm{var}}^{1/2}(\hat{\eta}_i)$. These were transformed to confidence intervals for the survival and cumulative hazard functions. Approximate confidence intervals for hazard functions and hazard ratios were obtained using bootstrapping (250 replicates), since analytic methods to determine these intervals were not apparently available. This was done by computing bootstrap estimates of the log hazard function and its standard error, calculating pointwise 95 per cent confidence intervals as (estimate $\pm 1.96 \times$ standard error) and exponentiating the estimate and the confidence limits.

3.6. *Model selection*

Experience so far suggests that a worthwhile improvement in fit over a straight-line model is often obtained by using a spline model with a single internal knot, but often little is gained by adding further knots. For that reason we suggest a 2 d.f. spline model as a reasonable initial choice for a given scale (PH or PO). However, it is obviously sensible to check the fit

of other models. For guidance, we suggest looking informally at the AIC (Akaike information criterion) of models with between 1 and 4 d.f. The AIC is defined to be minus twice the log-likelihood plus twice the number of model parameters. Formally, the preferred model is the one with minimum AIC, but to encourage parsimony (Occam's razor) this criterion should not be applied mechanically. Because the smaller sets of knots given in Table I are not generally subsets of the larger sets, not all of the models within a given class (PH or PO) are nested within those with higher d.f. Thus formal significance testing between the models may be inappropriate.

An important issue is how to build possibly complex multivariable models from many potential predictors within the PH and PO spline framework. Construction of prognostic models is a typical example. As presently implemented, the PH and PO spline models are somewhat computer-intensive to fit, so searching a large model space could take considerable CPU time. However, results such as those presented in Tables II and III suggest that estimated regression coefficients for covariates are robust to the misspecification of the spline part of the model. Therefore, we anticipate that $\hat{\boldsymbol{\beta}}$ will generally be similar between Weibull, Cox and PH spline models, and similar (but taking a different value) between log-logistic, semi-parametric PO and PO spline models. If so, then the choice of covariates, and of the functional form for continuous covariates, within PH spline models may be explored within the Weibull or Cox model. Analogously, covariates and functional forms for PO spline models may be investigated within the log-logistic model or even the log-normal model, since the latter two densities are fairly similar. This is mainly a pragmatic consideration, but it is of practical importance if such models are to be useful in routine data analysis. In the package Stata, for example, the Cox and log-normal models appear to be particularly efficiently implemented. Fitting is very fast, and so Stata is well-suited to exploring fairly large spaces of such models. Given the covariates and functional forms chosen within the Weibull/Cox or log-logistic/log-normal frameworks, final models would be fitted within the PH or PO spline class. For example, the fractional polynomials approach to selection of variables and transformation of continuous covariates described by Sauerbrei and Royston [18] could easily be applied within the above paradigm.

4. MODEL EXTENSIONS AND ASSESSMENT OF SCALE

4.1. Time-varying hazard ratios and odds ratios

One well-known extension of the Cox model attempts to deal with non-proportionality of hazards (non-PH) by allowing regression coefficient(s) to depend on some predefined function $f(t)$ of time, such as t or $\ln t$ (Cox [22]). Consider the simplest case of a single covariate z_1 and a predefined function $f(t)$. In the extended model, the log hazard ratio (log HR) for z_1 versus $z_1 = 0$ is specified as $\beta_1 z_1[1 + \beta_1^* f(t)]$. Time-dependency is induced through the multiplicative interaction between z_1 and $f(t)$. The time-varying regression coefficient is $\beta_1(t) = \beta_1[1 + \beta_1^* f(t)]$. The parameter β_1^* is estimated in a time-dependent Cox model with fixed covariate z_1 and time-dependent covariate $f(t)z_1$. A special expanded data set must be constructed. The details are not important here, except to note that the computational burden may be quite severe due to the potentially very large sample size of the expanded data set.

The specific choice of $f(t)$ may influence the power of a test of $\beta_1^* = 0$ (that is, of a test of non-PH for z_1), and more importantly, will have a strong impact on the estimate of $\beta_1(t)$, the

time-varying pattern of the HR. The behaviour of the HR over time is of substantive interest in survival studies with long follow-up, such as prognostic factors studies. For this reason, more flexible formulations than a simple fixed $f(t)$ have been used, such as step functions and spline functions (for example, Gore *et al.* [23], Gray [24], Hastie and Tibshirani [7] and Hess [8]). Hess [8] includes a brief but useful review of the methods.

A closely related approach may be used with the models developed here. Consider first a PH spline model with a single covariate z_1 and a single knot, written as

$$\ln H(t;z_1) = \gamma_0 + \gamma_{10}x + \gamma_{20}v_1(x) + \beta_1 z_1 \tag{2}$$

The log cumulative hazard ratio (CHR) and (since proportional hazards obtains) the log HR for z_1 versus $z_1 = 0$ is $\beta_1 z_1$, independent of t. If model (2) is extended analogously to the predefined function approach discussed above, we obtain

$$\ln H(t;z_1) = \gamma_0 + \gamma_{10}x + \gamma_{20}v_1(x) + \beta_1 z_1 + z_1\beta_1^* f(t) \tag{3}$$

Note that this extended model is defined on the log CHR, not the log HR. A more general extension of (2) to allow a time-dependent log CHR for z_1 is

$$\ln H(t;z_1) = \gamma_0 + (\gamma_{10} + \gamma_{11}z_1)x + (\gamma_{20} + \gamma_{21}z_1)v_1(x) + \beta_1 z_1 \tag{4}$$

We obtain the special case (3) from (4) by setting $x = f(t) = \ln t$, $\gamma_{11} = \beta_1^*$, $\gamma_{21} = 0$. As noted above, $f(t) = \ln t$ may be too restrictive for practical use. The more flexible class (4), in which γ_{21} is not constrained to zero, is preferable. With (4), the log CHR for z_1 versus $z_1 = 0$ is

$$\ln H(t;z_1) - \ln H(t;0) = \beta_1 z_1 + z_1[\gamma_{11}x + \gamma_{21}v_1(x)] = z_1[\beta_1 + \gamma_{11}x + \gamma_{21}v_1(x)]$$

that is, z_1 multiplied by a natural cubic spline in x. This flexible model is related to the time-varying Cox model based on $\beta_1^* f(t) = \gamma_{11}\ln t + \gamma_{21}v_1\ (\ln t)$, essentially as in Hess [8], except that we model the log CHR not the log HR, and our spline functions have argument $\ln t$ rather than t.

Class (4) may be formalized for general m as follows. Without loss of generality, suppose that the log CHR (compared with zero) for each of $z_1,\ldots,z_k$, the first k covariates in $\mathbf{z}$, is to be modelled as time-dependent, the remainder assumed to satisfy PH. By convention, let $k = 0$ represent the PH spline model. When $k = \dim(\mathbf{z})$ every covariate has a non-proportional effect on the hazard function. The generalized class of PH spline models may be written

$$\ln H(t;\mathbf{z}) = \gamma_0 + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{v}(x) + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{z} \tag{5}$$

where $\mathbf{v}(x) = (x, v_1(x),\ldots,v_m(x))^{\mathrm{T}}$, $\dim(\boldsymbol{\gamma}) = m+1$ and the jth component of $\boldsymbol{\gamma}$ $(1 \leqslant j \leqslant m+1)$ is

$$\gamma_j = \begin{cases} \gamma_{j0}, & k = 0 \text{ (PH model)} \\ \gamma_{j0} + \sum_{l=1}^{k}\gamma_{jl}z_l, & k \geqslant 1 \text{ (non-PH for } z_1,\ldots,z_k) \end{cases}$$

For example, the time-varying log CHR for z_1 at fixed values $z_2,\ldots$ of the other covariates is

$$\ln H(t;z_1,z_2,\ldots) - \ln H(t;0,z_2,\ldots) = z_1[\beta_1 + \gamma_{11}x + \gamma_{21}v_1(x) + \cdots + \gamma_{m+1,1}v_m(x)]$$

Of course, an identical extension to the PO model may be made to accommodate non-proportional odds for a subset of the covariates.

The proposed approach to modelling non-proportionality does not require that $z_1,\ldots,z_k$ be binary covariates. They may be ordinal or continuous, but then particular care is needed to ensure that the model fits the data and gives sensible predictions. A fuller exploration of the characteristics and implications of class (5) of models will be the subject of a further report. Here we will restrict ourselves to developing hypothesis tests of scale based on (5) (see Section 4.3).

4.2. Graphical assessment of scale

To make a graphical assessment of possible non-proportional hazards for a categoric covariate in a Cox model, it is recommended classically that one plot the empirical log cumulative hazard function for each level of the covariate against the log survival time. Non-proportional hazards may manifest as non-parallel lines. If the data are sampled from a Weibull distribution, the lines will be approximately parallel and approximately straight.

In practice, reliable interpretation of the plot may be hampered by the strongly dominating trend of the log cumulative hazard function with log survival time. The trend of course contains no information on non-proportionality. More subtly, the eye seems to judge non-parallelism according to the relative lengths of hypothetical line segments crossing the plotted lines at right angles, rather than intersecting them vertically as required. In addition, owing to the use of a log time-scale the area of the plot near the time origin is greatly magnified, but the largest scatter (greatest variance) of the observations is found here, making reliable interpretation difficult.

Attempts were made to improve the standard plot by subtracting the overall log cumulative hazard function of all the observations from the curve for each covariate level, by using an untransformed time-scale and by omitting unreliable points (with a large variance) near $t=0$. These measures did have a positive effect. Nevertheless, we have observed instances in some data sets in which a significance test (see Section 4.3) provided clear evidence of non-proportional hazards which did not manifest itself as convincing non-parallelism in the plot.

For these reasons, and perhaps contrary to expectation, simple graphical methods based on the assessment of parallelism do not appear to be particularly useful for investigating the proportional hazards (or by analogy, proportional odds) assumption. In the PH context it may be that other approaches, such as the Arjas [25] plot of observed and expected cumulative events, are more useful, but we have not pursued the topic further here.

4.3. A test of scale

Suppose we wish to test whether the PH assumption is satisfied jointly by a set of covariates $z_1,\ldots,z_k$. For example, $z_1,\ldots,z_k$ might be binary dummy variables parameterizing a categoric covariate with $c=k+1$ levels, or perhaps continuous predictors each modelled linearly, and so on. In the case of a categoric variable, the null hypothesis is that the log cumulative hazard functions are parallel across the c levels. For a continuous predictor z_1, the null hypothesis is that $\ln H(t;z_{11})$ and $\ln H(t;z_{12})$ are parallel functions of t for any two values z_{11}, z_{12} of z_1.

The approach is the same in each case. It is based on the generalized class of spline model in (5). The null hypothesis is that $\{\gamma_{jl}=0\}_{j=1,\ldots,m+1;l=1,\ldots,k}$ and is assessed by using a likelihood ratio test on $k(m+1)$ degrees of freedom in the usual way. An analogous test may be used to assess appropriateness of PO scaling.

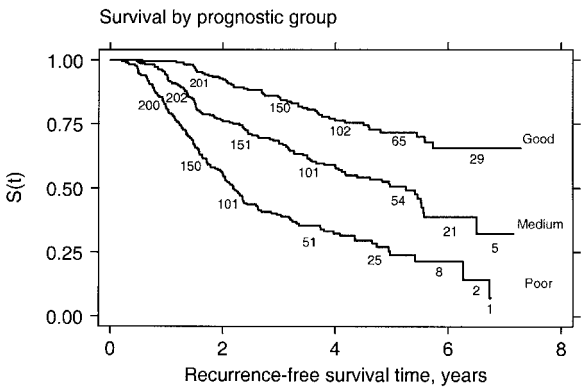

Figure 1. Kaplan–Meier survival curves for prognostic groups in the breast cancer data. Numbers at risk at selected time points are shown.

Table II. AIC values for several spline survival models for the breast cancer data. The regression coefficients are for the dummy variables representing the prognostic groups (see text for details).

d.f.	PH model			PO model		
	AIC	$\hat{\beta}_1$	$\hat{\beta}_2$	AIC	$\hat{\beta}_1$	$\hat{\beta}_2$
1*	1284.9	0.847	1.672	1259.1	1.093	2.260
2	1248.8	0.834	1.612	1241.0	1.052	2.171
3	1248.0	0.834	1.612	1241.5	1.052	2.162
4	1247.0	0.836	1.613	1240.6	1.054	2.159

*The PH model with 1 d.f. is the Weibull model, and the PO model with 1 d.f. is the log-logistic model.

5. EXAMPLES

We will refer to the 1 d.f. PH model interchangeably as Weibull or PH(1), and the 1 d.f. PO model interchangeably as log-logistic or PO(1), depending on context. We will denote by PH(j) or PO(j) the PH or PO spline model with $j \geqslant 2$ d.f.

5.1. Node-positive primary breast cancer

Figure 1 shows Kaplan–Meier survival curves for the three prognostic groups. According to a Cox model, the log hazard ratios (SE) for the medium and poor groups compared with the reference group of good prognosis patients are 0.84 (0.17) and 1.62 (0.16), respectively, highly significantly different from zero and reflecting the clear separation between the survival curves.

Table II gives the AIC values for PH(j) and PO(j) models for $j = 1, \ldots, 4$ and the estimated regression coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ for the dummy variables z_1 and z_2 representing the medium and poor groups in the prognostic model. The results show that the fits of the initial PH(2) and PO(2) models are hardly altered by adding more knots. However, there are reductions in AIC

of >18 between models with 1 and 2 d.f. or each scale, showing that the fit of Weibull and log-logistic models is poorer than that of a spline model. The PO(2) model has an AIC some 8 lower than the PH(2) model, indicating that PO(2) may fit better. However, a formal test is not straightforward since PH(j) and PO(j) models are non-nested for given j. Provided that more than 1 d.f. are used, the estimated regression coefficients are almost identical within a given model class. Even for the evidently misspecified 1d.f. models, the regression coefficients are not very different from those for the more complex models. For comparison, the Cox model gives $\hat{\beta}_1 = 0.840$ and $\hat{\beta}_2 = 1.618$, close to the values for the PH spline models.

The equation for the log cumulative odds function within the preferred PO(2) model (with standard errors in parentheses) is as follows:

$$\ln O(t;\mathbf{z}) = s(x;\hat{\boldsymbol{\gamma}}) + 1.052(0.206)z_1 + 2.171(0.209)z_2$$

where

$$s(x;\hat{\boldsymbol{\gamma}}) = -3.451(0.203) + 2.915(0.298)x + 0.191(0.044)v_1(x)$$

$$v_1(x) = \max[0,(x-k_1)^3] - \lambda_1 \max[0,(x-k_{\min})^3] - (1-\lambda_1)\max[0,(x-k_{\max})^3]$$

The interior knot $k_1 = \ln(1.769)$ and boundary knots $k_{\min} = \ln(0.197)$, $k_{\max} = \ln(6.724)$ were obtained, respectively, from the median, minimum and maximum uncensored recurrence-free survival times, as explained in Section 3.3. Thus

$$\lambda_1 = \frac{\ln(6.724) - \ln(1.769)}{\ln(6.724) - \ln(0.197)}$$

$$= 0.378$$

According to the model, the odds ratios for an event in an interval $(0,t]$ are estimated to be $\exp(1.052) = 2.9$ and $\exp(2.171) = 8.8$, respectively, for the medium and poor groups compared with the good group. The baseline survival function, that is, the survival function of the good group, is estimated to be $[1 + \exp s(x;\hat{\boldsymbol{\gamma}})]^{-1}$.

Figure 2 shows the estimated hazard functions for the three prognostic groups from the PO(2) model. It may be seen that the hazard function for all three groups reaches a peak approximately two years after diagnosis. For the poor group, the hazard falls rapidly after this time. By contrast, the hazard for the medium group falls less rapidly, and in the good group remains almost constant. The three functions gradually converge over time. The hazard functions modelled separately for each group (dashed lines), that is, without the constraint of proportional odds, are very similar to those assuming PO scaling.

Figure 3 shows the time-related behaviour of HR for the medium and poor groups according to the PO(2) model, together with 95 per cent pointwise bootstrap confidence intervals. The HRs dwindle monotonically with time. By $t = 6$ years the HRs are about 1.8 and 2.5 for the medium and poor groups, compared with 3.0 and 8.7 near $t = 0$. The (constant) HRs predicted by the Cox model are 2.3 and 5.0 and here are identical to the values from the PH spline model with 2 d.f. Thus for patients with a poorer prognosis, the PO(2) model predicts a relatively higher hazard and hence a more rapid event rate near $t = 0$ than does a Cox or PH(j) model.

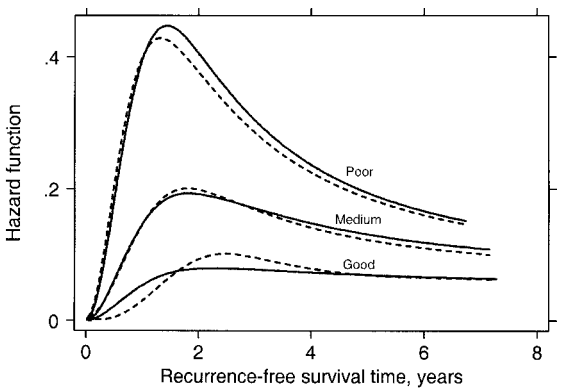

Figure 2. Hazard functions for prognostic groups, breast cancer data, according to PO(2) model. Solid lines, estimated within single model; dashed lines, estimated for each group separately.

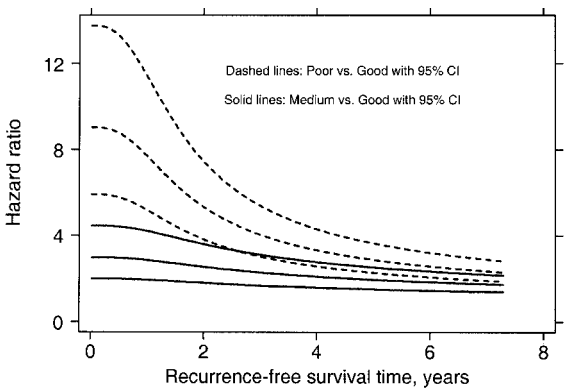

Figure 3. Hazard ratios with 95 per cent pointwise confidence intervals, breast cancer data, derived from PO(2) model. Solid lines, medium group versus good group; dashed lines, poor group versus good group.

Figure 4(a) shows the empirical and fitted cumulative log-odds of an event for recurrence-free survival. Although the model appears to fit poorly for the good group at $t < 1.5$ years, there are in fact only six events during this period compared with 37 for the medium group and 74 for the poor group. The variance of the estimate of $\ln O(t)$ for the good group is high in this region and the fluctuations may reasonably be ascribed to chance. The curvature in the log cumulative odds function is clearly visible. Figure 4(b) shows the observed and fitted Kaplan–Meier survival curves and demonstrates the good fit of the model.

The tests of scale described in Section 4.3 were applied to the PH and PO models with 2 d.f. The χ^2 statistics, each on 4 d.f., were 14.40 for the PH model and 5.74 for the PO model, giving P-values of 0.006 and 0.2, respectively. The results indicate a significantly poor fit of the PH model but no obvious lack of fit of the PO model in this example.

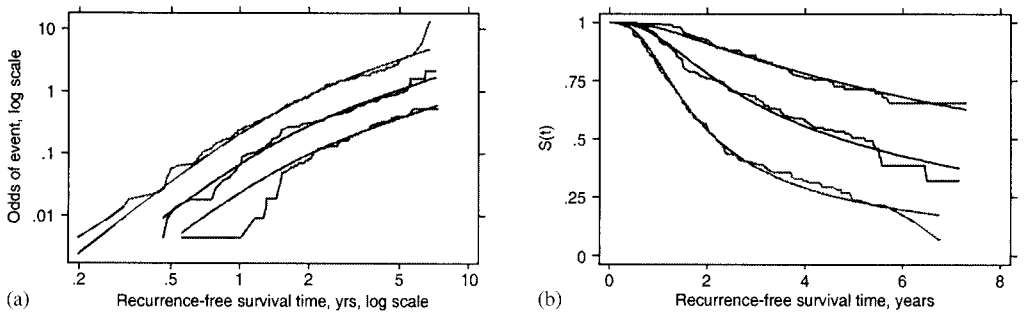

Figure 4. Breast cancer data: (*a*) cumulative odds of death for prognostic groups; (*b*) survival curves. Jagged lines, based on Kaplan–Meier estimates of survival curves; smooth lines, fitted values from PO(2) model.

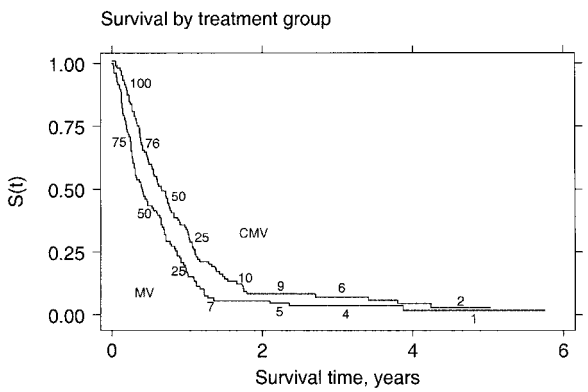

Figure 5. Kaplan–Meier survival curves for treatment groups in the bladder cancer data. Numbers at risk at selected time points are shown.

5.2. Advanced bladder cancer

The setting is a randomized clinical trial where we are concerned only with the treatment effect (CMV versus MV). Two patients in the CMV arm had died by only 2 days after randomization, before they received any treatment. These two very short survival times are influential on the log time-scale and for present purposes we decided to omit them from the analyses.

Figure 5 shows the survival curves by treatment. Almost all the patients died by 4 years after randomization, but there is a significant survival advantage for CMV over MV. According to a Cox model, HR = 0.67 (95 per cent CI 0.51–0.88), $P = 0.004$.

Table III gives the AIC values and regression coefficients for the effect of treatment CMV versus MV. The fit of the initial PH(2) model is little improved by adding knots and is much better than that of the Weibull model. The best PO model and overall best-fitting model is PO(1), the log-logistic. According to the latter model, the odds ratio for death on treatment CMV compared with MV is 0.46 (95 per cent CI 0.29–0.74), $P = 0.001$. The treatment effect is more statistically significant for the log-logistic than for the Cox model or any PH(j) model.

Table III. AIC values for several spline survival models for the advanced bladder cancer data. The regression coefficients are for the effect of treatment CMV versus MV.

d.f.	PH model		PO model	
	AIC	$\hat{\beta}$	AIC	$\hat{\beta}$
1	659.0	−0.393	628.4	−0.772
2	637.2	−0.378	629.7	−0.773
3	636.8	−0.379	631.2	−0.764
4	635.7	−0.380	632.8	−0.762

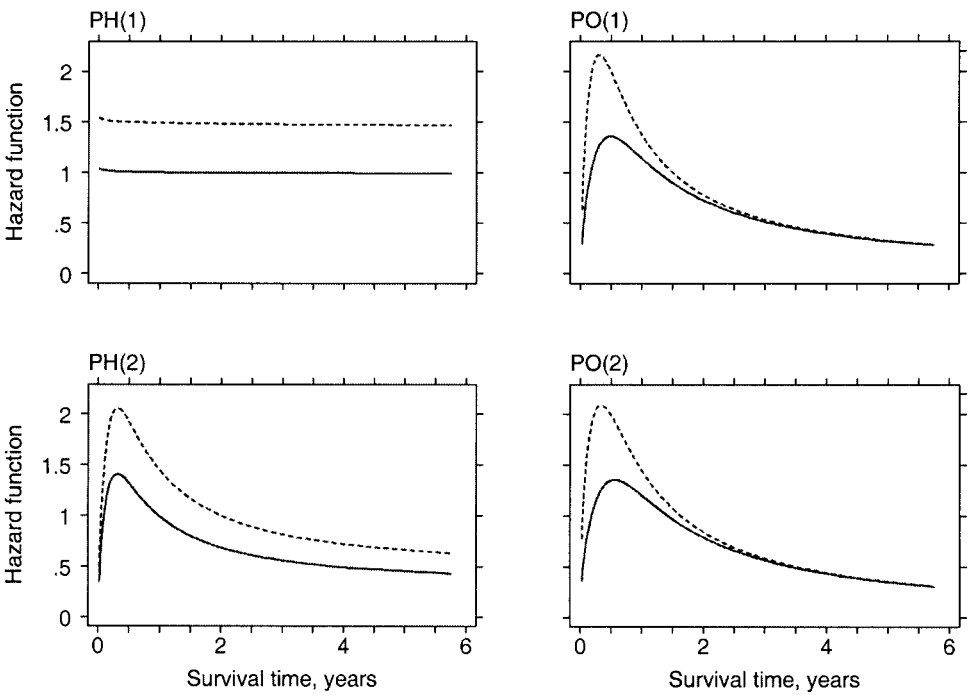

Figure 6. Hazard functions for treatment groups, bladder cancer data, according to low-dimensional PH and PO models. Note that PH(1) denotes a Weibull model and PO(1) a log-logistic model.

For the PH(2) model, for example, the P-value is 0.007. The Cox model gives $\hat{\boldsymbol{\beta}} = -0.405$, close to the values from the PH(j) models.

Figure 6 shows the hazard functions for the two treatment groups, estimated by PH and PO models with 1 and 2 d.f. The PH(1) model suggests that the hazard is nearly constant (exponential distribution), whereas the PO(1) model gives a dramatically different picture. The PH(1) model evidently gives a very poor estimate. By contrast, the hazard functions for the PH(2) model are rather similar to those from the PO models, except that the former are by definition forced to be proportionate between the two treatment groups for all t. The hazard functions from the PO(1) and PO(2) models are almost identical to each other, reaching a

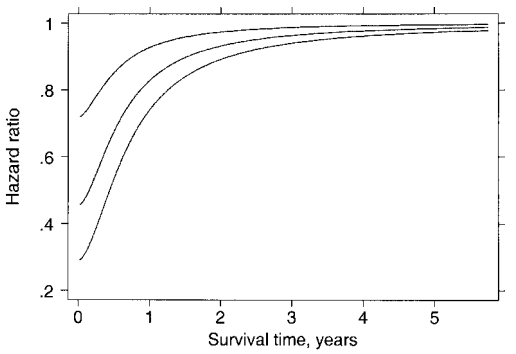

Figure 7. Hazard ratio (treatment CMV versus MV) with 95 per cent pointwise confidence intervals, bladder cancer data, derived from PO(1) model.

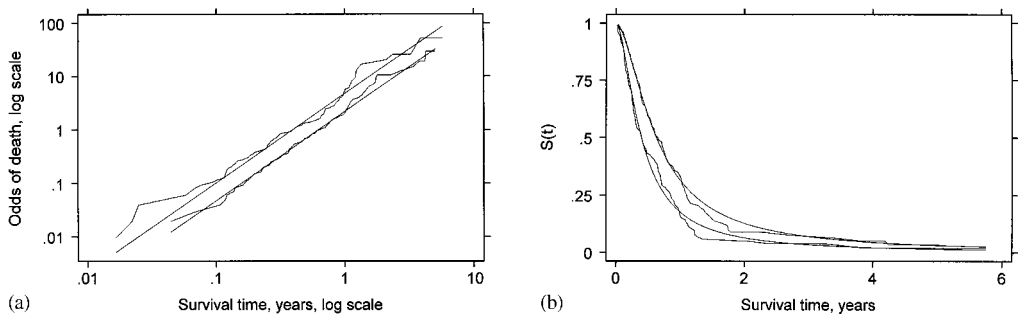

Figure 8. Bladder cancer data: (*a*) cumulative odds of death for treatment groups; (*b*) survival curves. Jagged lines, based on Kaplan–Meier estimates of survival curves; smooth lines, fitted values from PO(1) model.

peak about 6 months after randomization and converging after about 2 years. The pattern found with this model suggests that at the population level, the inclusion of cisplatin (the 'C' in CMV) in the chemotherapy regimen reduces the immediate risk of death, but that the effect does not persist beyond about two years. At the individual level, the effect of cisplatin on the hazard function may be quite different (Aalen [26]).

The *P*-values for scale tests of the PH(2) and PO(1) models are 0.08 and 0.3, respectively, confirming the findings of Table III that proportional odds is a more appropriate scale for the treatment effect than proportional hazards. The test provides weak evidence against the PH assumption.

Figure 7 shows the HR for CMV versus MV according to the PO(1) model, together with 95 per cent pointwise bootstrap confidence intervals. The plot confirms that the hazard reduction conferred by CMV is maximal near the start of treatment and diminishes rapidly. However, the confidence band is very wide and the HR is particularly imprecisely estimated in the period following randomization.

Finally, Figures 8(*a*) and 8(*b*) show, respectively, the empirical and fitted log cumulative odds of death and probabilities of survival according to the PO(1) model. The model fit appears fairly good throughout the follow-up time period. The empirical log cumulative odds

functions appear approximately parallel and linear, a characteristic of log-logistic distributions with the same shape parameter in each treatment group.

6. DISCUSSION

Our (perhaps ambitious) aim was to develop an approach which was flexible, clinically interpretable, robust, transparent and reasonably easily implemented in standard software. Some of this has been achieved, but more remains to be done. Splines certainly appear to offer adequate flexibility for approximating the baseline distribution function, and of course have been widely used to model continuous variables in medicine and epidemiology. Since the PO approach has not been used much in survival analysis, we will restrict this paragraph of the discussion mainly to a comparison between our models (PH and extended) and the unstratified and stratified Cox models, both of which are in common use. Interpretation of the regression parameters in our PH models is exactly as in the Cox model, and the values of the regression estimates and standard errors are likely to be in close agreement. Regarding non-PH of a categoric covariate, a standard approach in the Cox framework is to stratify the hazard function according to the levels of the covariate. No direct estimate of the effect of the covariate on the hazard is then available. The equivalent in our models is to fit different splines for each level of the covariate, formally modelled as a spline by covariate interaction (see Section 4.1). Estimates of the hazard functions at the different levels and of the hazard ratio functions in comparison with a reference level are immediately available from the parameters of the extended spline model, which is a major advantage. An alternative approach to non-PH in the Cox framework is to model the regression coefficients as time-varying. This is fairly closely related to our approach, but may be computationally easier and more convenient with the spline models than with Cox, since the data do not have to be modified to fit the models. However, there are still issues in both approaches with choosing an appropriate time-varying model. Regarding robustness, we are comfortable that our estimated regression coefficients are robust to misspecification of the spline part of the model, as already discussed. Further investigations would be desirable to compare our estimates of the hazard function and the hazard ratio function with the underlying function in simulation studies, and with estimates from other methods. Regarding transparency, we feel our method is more transparent than the Cox model since the complete model is specified in a full likelihood, as opposed to the partial likelihood of the Cox model in which many parameters are conditioned out as nuisance parameters. In a sense, both the strength and the weakness of the Cox model is that it allows the analyst to focus on the regression coefficients without thinking about the underlying distribution, which is a 'black box'. Finally, implementation. We used the maximum likelihood facilities in Stata to fit our models. We understand that implementation in SAS (using the 'interactive matrix language', IML) and S-plus would not be difficult for someone experienced in the relevant language systems, but we ourselves have not attempted it.

The statistical literature on the estimation and visualization of hazard functions is relatively small compared with that on modelling covariate effects and on elaborations of the proportional hazards model to accommodate time-varying effects, competing risks, multivariate failure times, frailty and so on. The medical literature is replete with plots of Kaplan–Meier survival curves but it is rare to find a hazard function in a medical paper. The appearance of Kaplan–Meier curves such as Figure 1 or Figure 5 gives no impression of the shape of the

hazard functions, nor of how the hazard ratio may vary with time. In node-positive primary breast cancer, it appears that patients in the poor prognostic group experience a high hazard of recurrence or death during the first 2 to 3 years after diagnosis (see Figure 2). After that time the hazard diminishes quite quickly and gradually converges towards that of the other two prognostic groups. The implication for breast cancer patients with a poor prognosis is that if no recurrence is experienced during the first 3 years after diagnosis, the outlook will progressively improve. By contrast, the hazard for patients in the good group is of course lower, but it does not appear to change much beyond 2 years. In advanced bladder cancer, CMV reduces the hazard of dying relative to MV during the first 2 years (see Figures 6 and 7), particularly in the very early phase just after treatment, but may have little beneficial effect on long-term survival.

In general, understanding the behaviour of the hazard function may help to elucidate the natural history of the disease, and in clinical trials, to reveal time-related effects of treatment not obvious from survival curves. For example, in cancer many treatments carry significant toxicity and treatment-related deaths are not unknown. Treatment toxicity could manifest as an increase in the hazard function in the early phase. Peaks in the hazard function are seen at approximately 4 months after randomization in the bladder cancer data, but interpretation in terms of toxicity is not straightforward. It is known that the platinum component of CMV increases this regimen's toxicity compared with MV. If anything, we would expect more early deaths on CMV than on MV, but this is not seen. Instead, an early peak appears with both treatments, and the hazard is generally higher with MV than CMV. Similar peaks appear even if the hazard functions for the two groups are estimated independently (data not shown).

The finding that the proportional odds assumption for covariate effects seems more appropriate than the proportional hazards assumption in the two example data sets is interesting. Clearly, much more experience with applying the methods to many data sets would be needed before one could say that PO is generally a more valid assumption than PH. At present, the vast majority of clinical trials and prognostic models in oncology are analysed using a PH model. One possible implication of the reduction in the hazard ratio with follow-up time (see Figures 2 and 6) is that the statistical significance of a treatment effect according to a PH analysis could diminish over time. This may have implications for the monitoring of trials and could be a reason for adopting analyses based on the PO scale, if that has been shown to be more appropriate in earlier trials of the same disease. Similarly, the statistical significance of a treatment effect may be greater if the 'correct' scale is used, and this may increase the power and reduce the sample size calculated for a new trial. How to compute sample sizes using the new models described here is a topic for further research.

We have not so far made detailed comparisons between our method and the related approach of Younes and Lachin [12]. However, we did apply our method to estimate the models for the well-known Veterans' Administration lung cancer data set (listed in Appendix I of Kalbfleisch and Prentice, reference [2]) equivalent to those in Table 1 of Younes and Lachin (reference [12]). We obtained estimates of $\hat{\boldsymbol{\beta}}$ essentially identical to those from the other methods. However, our standard errors were much closer to those reported for the semi-parametric model of Bennett [9] than those reported by Younes and Lachin [12] for their method, which appear rather implausibly small. Our theoretical concerns about possible instability of Younes and Lachin's [12] method of estimating the hazard function are borne out by the appearance of Figure 4 in their paper. For both the PH and PO models the behaviour of the baseline hazard is quite complex. It reaches a peak of 0.012 about 50 days after diagnosis of lung

cancer, falls and then rises quite sharply after about 300 days. However, the 95 per cent CIs are wide after 300 days. According to our approach, the PO model with one knot fits much better than the PH model with one knot, similar to Younes and Lachin's [12] finding. However, our baseline hazard estimate reaches a peak of 0.012 at about 100 days and then gently falls to about 0.004 at 600 days, compared with 0.012 at this time point according to Younes and Lachin [12]. We suspect the apparent late rise in the hazard is an artefact, probably exacerbated by the small sample size and the extreme sparseness of the observations – there are only 3 deaths beyond 400 days.

An alternative to the log-odds of failure scale is the Normal deviate or probit scale, where the basic (1d.f.) model is the log-normal rather than the log-logistic distribution (see Appendix A). Royston [27], for example, described applications of the log-normal model in breast and ovarian cancer data. The log-normal and log-logistic densities are quite similar to each other, as are the logistic and probit link functions, therefore it may be difficult to distinguish between the two models in real data. This is particularly so when a high proportion of the observation times is censored when most of the long upper tail of the distribution, which contains information about the distributional shape, is unobserved. All three basic models (Weibull, log-logistic and log-normal) have an accelerated failure time interpretation, with covariate effects being interpretable as effects on the mean log survival time. This interpretation is lost with the more general spline models, but covariate effects still have an attractive interpretation in terms of the log hazard and log-odds of failure functions. With the log-normal approach, covariates operate on the perhaps less familiar probit scale, which is one good reason to prefer the log-logistic approach to the log-normal.

The methods described here have been implemented in add-on routines ('ado-files') for the package Stata (StataCorp [21]) and published [28]. Options for late entry and interval censoring, both of which are straightforward to implement in the present parametric framework, have been included.

APPENDIX A: SPLINE MODELS WITH A PROBIT LINK

Spline models with a probit link function $g(.)$ are defined by equation (1) with $g(x)=-\Phi^{-1}(x)$, where $\Phi^{-1}(.)$ is the inverse standard Normal distribution function. Using the notation of Section 4.1, the extended class of spline models with a probit link is defined by

$$-\Phi^{-1}[S(t;\mathbf{z})]=\gamma_0+\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{v}(x)+\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}$$

Letting $\eta=-\Phi^{-1}\ [S(t;\mathbf{z})]$, the survival, density and hazard functions are, respectively

$$S(t;\mathbf{z})=\Phi(-\eta)$$

$$f(t;\mathbf{z})=-\frac{\mathrm{d}\Phi(-\eta)}{\mathrm{d}t}=t^{-1}\,\frac{\mathrm{d}}{\mathrm{d}x}\,[\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{v}(x)]\phi(\eta)$$

$$h(t;\mathbf{z})=f(t;\mathbf{z})/S(t;\mathbf{z})$$

where $\phi(.)$ is the standard Normal density function. The likelihood for a typical observation is $f(t;\mathbf{z})$ for an uncensored observation, $S(t;\mathbf{z})$ for a right-censored one.

APPENDIX B: DERIVATION OF NATURAL CUBIC REGRESSION SPLINE

We derive an expression for a cubic regression spline that is constrained to be linear beyond two 'boundary' knots $k_{\min}$ and $k_{\max}$ (not necessarily placed at the extremes of the argument, x). An unconstrained cubic regression spline with $m+2$ knots at $k_{\min} < k_1 < \cdots < k_m < k_{\max}$ takes the form

$$s(x) = \beta_{00} + \beta_{10}x + \beta_{20}x^2 + \beta_{30}x^3 + \sum_{j=1}^{m} \beta_j (x - k_j)_+^3 + \beta_{k_{\min}}(x - k_{\min})_+^3 + \beta_{k_{\max}}(x - k_{\max})_+^3$$

If $x < k_{\min}$, the linearity constraint implies that all quadratic and cubic terms must vanish, so that $\beta_{20} = \beta_{30} = 0$. If $x > k_{\max}$, the linearity constraint may be handled through the fact that the second and third derivatives of $s(x)$ must be identically zero. For $x > k_{\max}$

$$s(x) = \beta_{00} + \beta_{10}x + \sum_{j=1}^{m} \beta_j (x - k_j)^3 + \beta_{k_{\min}}(x - k_{\min})^3 + \beta_{k_{\max}}(x - k_{\max})^3$$

$$s'(x) = \beta_{10} + 3\sum_{j=1}^{m} \beta_j (x - k_j)^2 + 3\beta_{k_{\min}}(x - k_{\min})^2 + 3\beta_{k_{\max}}(x - k_{\max})^2$$

$$s''(x) = 6\sum_{j=1}^{m} \beta_j (x - k_j) + 6\beta_{k_{\min}}(x - k_{\min}) + 6\beta_{k_{\max}}(x - k_{\max})$$

$$s'''(x) = 6\sum_{j=1}^{m} \beta_j + 6\beta_{k_{\min}} + 6\beta_{\max}$$

$s'''(x) = 0$ implies that $\beta_{k_{\max}} = -\sum_{j=1}^{m} \beta_j - \beta_{k_{\min}}$, whereas $s''(x) = 0$ implies that

$$\begin{aligned} \beta_{k_{\min}}(x - k_{\min}) &= -\sum_{j=1}^{m} \beta_j (x - k_j) - \beta_{k_{\max}}(x - k_{\max}) \\ &= -\sum_{j=1}^{m} \beta_j (x - k_j) + \sum_{j=1}^{m} \beta_j (x - k_{\max}) + \beta_{k_{\min}}(x - k_{\max}) \end{aligned}$$

Therefore

$$\begin{aligned} \beta_{k_{\min}} &= -\sum_{j=1}^{m} \beta_j \frac{k_{\max} - k_j}{k_{\max} - k_{\min}} \\ &= -\sum_{j=1}^{m} \beta_j \lambda_j, \text{ say} \end{aligned}$$

where

$$\lambda_j = \frac{k_{\max} - k_j}{k_{\max} - k_{\min}}, \quad 1 - \lambda_j = \frac{k_j - k_{\min}}{k_{\max} - k_{\min}}$$

It follows that for $x > k_{\max}$

$$\begin{aligned}
\beta_{k_{\min}}(x-k_{\min})^3 + \beta_{k_{\max}}(x-k_{\max})^3 &= -\sum_{j=1}^{m}\beta_j\lambda_j(x-k_{\min})^3 - \left(\sum_{j=1}^{m}\beta_j + \beta_{k_{\min}}\right)(x-k_{\max})^3 \\
&= -\sum_{j=1}^{m}\beta_j\lambda_j(x-k_{\min})^3 - \left(\sum_{j=1}^{m}\beta_j - \sum_{j=1}^{m}\beta_j\lambda_j\right)(x-k_{\max})^3 \\
&= -\sum_{j=1}^{m}\beta_j\lambda_j(x-k_{\min})^3 - \left(\sum_{j=1}^{m}\beta_j(1-\lambda_j)\right)(x-k_{\max})^3 \\
&= \sum_{j=1}^{m}\beta_j[-\lambda_j(x-k_{\min})^3 - (1-\lambda_j)(x-k_{\max})^3]
\end{aligned}$$

For any value of x, therefore, the constrained spline may now be written as

$$\begin{aligned}
s(x) &= \beta_{00} + \beta_{10}x + \sum_{j=1}^{m}\beta_j[(x-k_j)_+^3 - \lambda_j(x-k_{\min})_+^3 - (1-\lambda_j)(x-k_{\max})_+^3] \\
&= \gamma_0 + \gamma_1 x + \gamma_2 v_1(x) + \cdots + \gamma_{m+1}v_m(x)
\end{aligned}$$

where $\gamma_0 = \beta_{00}$, $\gamma_1 = \beta_{10}$ and for $j = 1, \ldots, m$

$$\begin{aligned}
\gamma_{j+1} &= \beta_j \\
v_j(x) &= (x-k_j)_+^3 - \lambda_j(x-k_{\min})_+^3 - (1-\lambda_j)(x-k_{\max})_+^3
\end{aligned}$$

ACKNOWLEDGEMENTS

We are grateful to Dr Ben Mead and the collaborators in the MRC Advanced Bladder Cancer Trial (BA07) for allowing us to use the data from this trial, to Sarah Walker, Peter Sasieni, Daniela DeAngelis and two anonymous referees for comments on the manuscript and to Gareth Ambler for assistance with some computational issues.

REFERENCES

1. Hjort NL. On inference in parametric survival data models. *International Statistical Review* 1992; **60**:355–387.
2. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data.* Wiley: New York, 1980.
3. Gelfand AE, Ghosh SK, Christiansen C, Soumerai SB, McLaughlin TJ. Proportional hazards models: a latent competing risk approach. *Applied Statistics* 2000; **49**:385–397.
4. Kooperberg C, Stone CJ, Truong YK. Hazard regression. *Journal of the American Statistical Association* 1995; **90**:78–94.
5. Rosenberg PS. Hazard function estimation using B-splines. *Biometrics* 1995; **51**:874–887.
6. Valsecchi MG, Silvestri D, Sasieni PD. Evaluation of long term survival: use of diagnostics and robust estimators with Cox's proportional hazards models. *Statistics in Medicine* 1996; **15**:2763–2780.
7. Hastie TJ, Tibshirani RJ. Varying-coefficient models (with discussion). *Journal of the Royal Statistical Society, Series B* 1993; **55**:757–796.

8. Hess K. Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline models. *Statistics in Medicine* 1994; **13**:1045–1062.
9. Bennett S. Analysis of survival data by the proportional odds model. *Statistics in Medicine* 1983; **2**:273–277.
10. Yang S, Prentice RL. Semiparametric inference in the proportional odds regression model. *Journal of the American Statistical Association* 1999; **94**:125–136.
11. Rossini AJ, Tsiatis AA. A semiparametric proportional odds model for the analysis of current status data. *Journal of the American Statistical Association* 1996; **91**:713–721.
12. Younes N, Lachin J. Link-based models for survival data with interval and continuous time censoring. *Biometrics* 1997; **53**:1199–1211.
13. Aranda-Ordaz FJ. On two families of transformations to additivity for binary response data. *Biometrika* 1981; **68**:357–363.
14. Shen, XT. Proportional odds regression and sieve maximum likelihood estimation. *Biometrika* 1998; **85**: 165–177.
15. Mackenzie G. Regression models for survival data: the generalized time-dependent logistic family. *Statistician* 1996; **45**:21–34.
16. Mackenzie G. On a non-proportional hazards regression model for repeated medical random counts. *Statistics in Medicine* 1997; **16**:1831–1843.
17. Schmoor C, Olschewski M, Schumacher M. Randomized and non-randomized patients in clinical trials: experiences with comprehensive cohort studies. *Statistics in Medicine* 1996; **15**:263–271.
18. Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors using fractional polynomials. *Journal of the Royal Statistical Society*, *Series A* 1999; **162**: 71–94.
19. Mead GM, Russell M, Harland SJ, Harper PG, Cowan R, Roberts JT, Uscinska BM, Griffiths GO, Parmar MKB. A randomized trial comparing methotrexate and vinblastine (MV) with cisplatin, methotrexate and vinblastine (CMV) in advanced transitional cell carcinoma: a Medical Research Council study. *Lancet* 1998; **78**:1067–1075.
20. Durrleman S, Simon R. Flexible regression-models with cubic-splines. *Statistics in Medicine* 1989; **8**:551–561.
21. StataCorp. *Stata Reference Manual, Release 7.0*. Stata Press: College Station, Texas, 2001.
22. Cox DR. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society*, *Series B* 1972; **34**:187–220.
23. Gore SM, Pocock SJ, Kerr GR. Regression models and non-proportional hazards in the analysis of breast cancer survival. *Applied Statistics* 1984; **33**:176–195.
24. Gray RJ. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association* 1992; **87**:942–951.
25. Arjas EA. A graphical method for assessing goodness of fit in Cox's proportional hazards model. *Journal of the American Statistical Association* 1988; **83**:204–212.
26. Aalen OO. Modelling heterogeneity in survival analysis by the compound poisson distribution. *Annals of Applied Probability* 1992; **2**:951–972.
27. Royston P. The lognormal distribution as a model for survival time in cancer, with an emphasis on prognostic factors. *Statistica Neerlandica* 2000; **53**:1–16.
28. Royston P. Flexible alternatives to the Cox model, and more. *The State Journal* 2001; **1**:1–28.

