Statistical Modelling 2010; **10**(**3**): 291–314

Flexible hazard ratio curves for continuous predictors in multi-state models: an application to breast cancer data

Carmen Cadarso-Suárez[1,5], Luís Meira-Machado[2], Thomas Kneib[3] and Francisco Gude[4,5]
[1]Department of Statistics and Operations Research, University of Santiago de Compostela, Spain
[2]Department of Mathematics and Applications, University of Minho, Portugal
[3]Department of Mathematics, Carl von Ossietzky University, Oldenburg, Germany
[4]Clinical Epidemiology Unit, Hospital Clínico Universitario de Santiago, Santiago de Compostela, Spain
[5]Instituto de Investigación Sanitaria de Santiago (IDIS), Santiago de Compostela, Spain

Abstract: Multi-state models (MSMs) are very useful for describing complicated event history data. These models may be considered as a generalization of survival analysis where survival is the ultimate outcome of interest but where intermediate (transient) states are identified. One major goal in clinical applications of MSMs is to study the relationship between the different covariates and disease evolution. Usually, MSMs are assumed to be parametric, and the effects of continuous predictors on log-hazards are modelled linearly. In practice, however, the effect of a given continuous predictor can be unknown, and its form may be different in all transitions. In this paper, we propose a P-spline approach that allows for non-linear relationships between continuous predictors and survival in the multi-state framework. To better understand the effects at each transition, results are expressed in terms of hazard ratio curves, taking a specific covariate value as reference. Confidence bands for these curves are also derived. The proposed methodology was applied to a database on breast cancer, using a progressive three-state model. This application revealed hitherto unreported effects: whereas DNA index is only an important non-linear predictor of recurrence, the percentage of cells in phase S is a significant predictor of both recurrence and mortality.

Key words: Cox model; hazard ratio; multi-state model; penalized splines

Received December 2007; **revised** July 2008; **accepted** October 2008

1 Introduction

An important aim in survival analysis is to assess the possible effect of a set of prognostic factors on the course of a disease. Typically, statistical methods for the analysis of survival data consider the transition from the initial state ('alive'), to

Address for correspondence: Carmen Cadarso-Suárez, Unit of Biostatistics-School of Medicine, University of Santiago de Compostela, Spain, C/San Francisco s/n, 15782-Santiago de Compostela-Spain. E-mail: carmen.cadarso@usc.es

 10.1177/1471082X0801000303

a single ultimate state or endpoint ('dead'), and the effect of covariates on disease progression is generally modelled using (extensions of) the Cox proportional hazards model (Cox, 1972). In many instances, however, more than one state and more than one transition can be defined. For instance, in a breast cancer study, some states, such as 'disease free', 'local recurrence' or 'distant metastasis', play a role in predicting the prognosis of patients. Analyses of such studies, in which individuals may experience several events, can be successfully performed by multi-state models (MSMs) (Andersen *et al*., 1993; Hougaard, 1999; Andersen and Keiding, 2002; Putter *et al*., 2007). In the MSM framework, the so-called transition intensities characterize the hazard for movement from one state to another. Introduction of covariates can also explain differences in illness progression. In this way, multi-state regression models provide more detailed information on the disease progress, revealing how the different covariates may affect the various permitted transitions.

A common strategy that greatly simplifies inference for MSMs is to decouple the whole process into various survival models, by fitting separate intensities to all permitted transitions with semiparametric Cox proportional hazard regression models while making appropriate adjustments of the risk set. The effect of the continuous covariates on the log-hazards is often assumed to have a linear functional form in all intensities. Nevertheless, if the true effect is highly non-linear, this erroneous assumption of linearity may have serious consequences. Fitting the incorrect functional effect of a covariate is a form of misspecification and leads to bias and decreased power of tests for statistical significance (Struthers and Kalbfleisch, 1986; Anderson and Fleming, 1995). Also, incorrect functional form for a covariate is a model failure that can lead to a diagnosis of non-proportional hazards. Diagnostics for non-proportional hazards, such as Schoenfeld residual plots, will suggest non-proportionality when in fact the problem is the incorrect functional form for covariates (Therneau and Grambsch, 2000, 148–49). However, correcting for non-proportional hazards with log-time interactions when the model failure is due to incorrect functional form will not rectify the misspecification.

To deal with this problem, two approaches have traditionally been used in practice: (*i*) categorizing predictors, creating dummy variables and calculating effects vis-à-vis an appropriate reference category or (*ii*) including these predictors in a polynomial model. However, neither approach is fully satisfactory. Although easy to interpret, the categorical approach is likely to result in loss of statistical power, furnishes calculated effects that are averages for each category and raises the problem of how many categories should be used and where their cutpoints should be located (Altman *et al*., 1994). In many instances, polynomial models can provide better power than categorical analyses, but higher order terms (beyond the quadratic term) tend to produce artificial turns in the fitted model (Greenland, 1995). Even when fractional or inverse terms are included (Royston and Altman, 1994), these can complicate interpretation.

The relative lack of flexibility of parametric survival models has, in recent years, led to the development of a variety of non-parametric regression methods based on various statistical models: the additive hazards approach of Aalen (Martinussen and

Scheike, 2006), Cox regression models with additive predictor (Hastie and Tibshirani, 1990a, 1990b) and extensions of regression models with structured additive predictor (Brezger and Lang, 2006) to the analysis of survival data (Hennerfeind *et al.*, 2006; Kneib and Fahrmeir, 2007) and multi-state data (Kneib and Hennerfeind, 2008). These regression techniques have the advantage of not assuming a parametric relationship between the continuous covariates and the response and eliminate the need for the researcher to impose functional assumptions. To introduce flexibility into the Cox model, several smoothing methods may be applied, but B-splines (de Boor, 2001), smoothing splines (Hastie and Tibshirani, 1990a; Wahba, 1990), P-splines (Eilers and Marx, 1996) or Bayesian versions of P-splines (Lang and Brezger, 2004) are being most frequently considered in this context. The first goal of this paper is to investigate the use of P-splines to model non-linear relationships between continuous predictors and possible outcomes in the multi-state framework.

When reporting survival studies, biomedical researchers try to offer interpretable results in a simple and summarized manner. In general, interpretation of results drawn from smooth multi-state regression models is not immediate. For such models to be directly interpreted, employment of an effect measure, such as the hazard ratio (HR) that relies on comparing estimated effects to effects at a pre-specified reference point proves extremely useful. Despite the potential advantages of using P-spline smoothing methods in MSMs, there is currently no proposed analytical method whereby HR curves can be calculated for continuous covariates in this context. The second goal of this work, therefore, is to propose a flexible method for constructing smooth HR curves with confidence bands, which facilitates expression of the results in a manner that is standard in epidemiologic and clinical survival studies.

The layout of this paper is organized as follows: MSMs are presented in Section 2 and HR curves are defined in this context; Section 3 then describes P-spline estimation of HRs and construction of asymptotic confidence intervals for these curves. An example of their application to a breast cancer dataset is presented in Section 4. Finally, we conclude with a discussion section.

2 Statistical methodology

2.1 Progressive three-state model

In this paper, attention will be focused on the progressive three-state model (the MSM used in our dataset analysis; see Figure 1), though the methods considered here can also be used in other MSMs. In this context, the Cox (semi-) Markov model

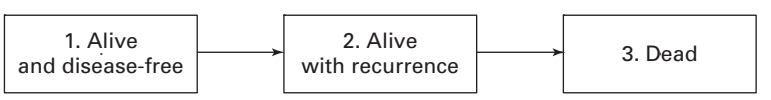

Figure 1 Progressive three-state model for breast cancer data

(Andersen *et al.*, 2000) is usually assumed whenever the interest is to study how covariates affect survival. As already commented earlier, the inference problem can be decoupled into various Cox survival models (two for the progressive three-state model), by fitting separate intensities to all permitted transitions. Depending upon the assumption made (Markovianity, semi-Markovianity), one of the following two MSMs may be used.

2.1.1 Cox Markov model

The Markov assumption states that the future depends on the individual's past solely via his/her current state. In other words, given the individual's present state and event history, the next state to be undergone and the transition time depend exclusively on the present state. Under the Markovian assumption, in a progressive three-state model the corresponding intensities, $\alpha_{hj}(t; \boldsymbol{Z})$ for transitions $h \rightarrow j$, may be modelled semi-parametrically, using Cox-like models of the form

$$\alpha_{hj}(t; \boldsymbol{Z}) = \alpha_{hj0}(t) \exp\left(\boldsymbol{\beta}_{hj}^{\mathrm{T}} \boldsymbol{Z}\right) \tag{2.1}$$

where $\alpha_{hj0}(t)$ is an unspecified non-negative baseline hazard function, $\boldsymbol{Z} = \left(Z_1, \ldots, Z_q\right)$ a vector of q covariates (continuous and/or categorical) and $\boldsymbol{\beta}_{hj}$ the associated vector of unknown regression model parameters.

2.1.2 Cox semi-Markov model

By ignoring disease history behaviour, Markov models such as (2.1) earlier may have severe limitations, thus rendering them inappropriate (e.g., the future health of recently diseased individuals may be different from that of individuals who have been ill for a long time). Apart from the fact that a patient has experienced recurrence (state 2, Figure 1), in cancer studies the amount of time spent in the healthy state (sojourn time) is often of interest. An alternative approach to Cox Markov model (CMM) is to use a semi-Markov assumption, in which the future of a process depends not on the current time, but rather on its duration in the current state. Again, assuming a progressive three-state model, the only difference between CMM and Cox semi-Markov model (CSMM) (Andersen *et al.*, 2000) resides in transition $2 \rightarrow 3$, in which intensity α_{23} is modelled in a different way. Specifically, the corresponding intensities in the CSMM are

$$\begin{aligned} \alpha_{12}(t; \boldsymbol{Z}) &= \alpha_{120}(t) \exp\left(\boldsymbol{\beta}_{12}^{\mathrm{T}} \boldsymbol{Z}\right), \\ \alpha_{23}(t - T_{12}; \boldsymbol{Z}) &= \alpha_{230}(t - T_{12}) \exp\left(\boldsymbol{\beta}_{23}^{\mathrm{T}} \boldsymbol{Z}\right), \end{aligned} \tag{2.2}$$

where T_{12} is the entry time into state 2. CSMMs like (2.2) are also called 'clock reset' models, because each time the patient enters a new state, time is reset to 0.

In both models, CMM (2.1) and CSMM (2.2), the inference for $\boldsymbol{\beta}_{hj}$ on transition $h \rightarrow j$ is classically based on the partial likelihood function, $PL\left(\boldsymbol{\beta}_{hj}\right)$ (Cox, 1972), assuming that the observations are independent. Baseline hazards α_{hj0} can

be obtained in a second step, see Therneau and Grambsch (2000) for details. Given a sample of n individuals, models (2.1)–(2.2) partial likelihood-based estimates can be obtained with most statistical packages (R, SAS, etc.), as long as a counting process notation is used, with each patient being represented by several observations (Therneau and Grambsch, 2000; Meira-Machado *et al.*, 2009). As a supplement to the two-step partial likelihood approach, we will consider a joint estimation procedure for the regression coefficients and the baseline intensity based on the full likelihood that relies on representing the baseline intensity as a P-spline (see the following subsection).

2.2 Additive MSMs

One possible approach allowing for non-linear effects in CMM model (2.1) is to express the hazard on transition $h \rightarrow j$ as an additive Cox model of the form

$$\alpha_{hj}(t; \boldsymbol{Z}) = \alpha_{hj0}(t) \exp\left(\sum_{i=1}^{q} f_{i,hj}(Z_i)\right), \tag{2.3}$$

where $f_{i,hj}(\cdot)$, $i = 1, \ldots, q$ are smooth covariate functions.

Alternatively, CSMM model (2.2) can be generalized, by considering the intensities

$$\begin{aligned} \alpha_{12}(t; \boldsymbol{Z}) &= \alpha_{120}(t) \exp\left(\sum_{i=1}^{q} f_{i,12}(Z_i)\right), \\ \alpha_{23}(t - T_{12}; \boldsymbol{Z}) &= \alpha_{230}(t - T_{12}) \exp\left(\sum_{i=1}^{q} f_{i,23}(Z_i)\right). \end{aligned} \tag{2.4}$$

Models (2.3)–(2.4) are more flexible than models (2.1)–(2.2) because, rather than assuming a parametric form for the effects of the continuous covariates, the researcher merely assumes that these effects may be represented by arbitrary unknown smooth functions.

It is worth noting that models (2.3)–(2.4) avoid the curse of dimensionality, by restricting a multidimensional non-parametric regression problem to an additive model. Hence, additive models are easy to interpret because the additive components simply describe the influence of each covariate separately. Likewise, non-parametric models, such as (2.3)–(2.4), can be used to check (visually) the presence of non-linear effects of the covariates, as well as to identify their correct functional form (overcoming the problems mentioned earlier).

Additive Cox regression models have been studied by several authors in the framework of the mortality model for survival data (i.e., the simplest form of an MSM with states 'alive' and 'dead', and only one possible transition) using various smoothing techniques (see, e.g., Hastie and Tibshirani, 1990b; Gray, 1992; Huang *et al.*, 2000; Huang and Liu, 2006). In this paper, P-splines have been chosen as smoothers, since

in the context of Cox-type hazard regression models, they have been shown to result in satisfactory reproduction of effect curves (Hennerfeind *et al.*, 2006; Govindarajulu *et al.*, 2007; Kneib and Fahrmeir, 2007). Moreover, the existence of standard software (e.g., in R and S-plus) makes it easy for this type of smoother to be implemented in an MSM framework.

2.3 Smooth HR curves in MSMs

As mentioned in Section 1, the primary goal in survival studies is to understand the effect that each covariate exerts on the outcome. One of the most commonly used measures of this effect is the HR function. Recently, a flexible P-spline approach to this measure was proposed by Strasak *et al.* (2009) for the additive Cox model. Here, we propose a generalization of the HR curves to additive MSMs.

For the sake of simplicity, a Markovian progressive three-state model (2.3) will henceforth be assumed. Extensions to semi-Markovian models like (2.4) are straightforward.

Assuming (2.3), the adjusted HR for a subject with covariate value Z_i compared to a subject with covariate value $z_{i,ref}$, on transition $h \rightarrow j$ is given by

$$HR_{i,hj}\left(Z_i, z_{i,ref}\right) = \frac{\exp\left(f_{i,hj}\left(Z_i\right)\right)}{\exp\left(f_{i,hj}\left(z_{i,ref}\right)\right)} = \exp\left\{f_{i,hj}\left(Z_i\right) - f_{i,hj}\left(z_{i,ref}\right)\right\}. \qquad (2.5)$$

The HR curve so derived provides an estimate of the relative risk for the event of interest (represented by state j). Hence, $HR_{i,hj}\left(Z_i, z_{i,ref}\right)$ denotes the HR curve that indicates the expected change in the risk of the event with the change in covariate values.

Note that when $f_{i,hj}\left(Z_i\right) = \beta_{i,hj} Z_i$ (i.e., the linear case), the logarithm of HR curve (2.5) is reduced to a straight line

$$\log HR_{i,hj}\left(Z_i, z_{i,ref}\right) = \log \frac{\exp\left(\beta_{i,hj} Z_i\right)}{\exp\left(\beta_{i,hj} z_{i,ref}\right)} = \beta_{i,hj}\left(Z_i - z_{i,ref}\right), \qquad (2.6)$$

thus indicating that the change in risk for a $\left(Z_i - z_{i,ref}\right)$ change in Z_i is a constant value. Both (2.5) and (2.6) are independent of time t, relating to the well-known proportional hazards assumption. A good explanation of the HR and its interpretability can be found in the papers by Kay (2004) and Spruance *et al.* (2004).

3 Penalized spline smoothing

This section introduces the use of P-splines within the additive Cox proportional hazards model to reflect the nature of continuous covariates. The inclusion of

non-linear covariate effects in the multi-state framework (CMM, CSMM) is then straightforward.

3.1 Additive models

An additive Cox model (i.e., the simplest additive MSM for describing mortality) is given by

$$\alpha(t;\mathbf{Z}) = \alpha_0(t)\exp\left(\sum_{i=1}^{q} f_i(Z_i)\right),$$

where each of the functions f_i represents a flexible, non-linear function of a continuous covariate Z_i and all functions are combined in an additive fashion.

The key assumption underlying P-splines is that the unknown smooth functions, $f_i, i = 1, \ldots, q$, can be approximated by polynomial splines of degree l, defined on a set of knots within the domain of Z_i. Different parameterizations (or bases) of polynomial splines exist, but we will employ the representation

$$f_i(Z_i) = \theta_{i0}Z_i + \sum_{k=1}^{K+2}\theta_{ik}B_{ik}(Z_i)$$

(introduced by Gray, 1992, in the context of survival modelling) in the following, due to its intuitive decomposition into a linear part and the non-linear deviation and since it forms the basis for P-splines as available in the R-package *survival*. In this representation, K corresponds to the number of interior knots for each of the splines, B_{ik} are the basis functions for the non-linear part of the effects and $\theta_{ik}, k = 0, 1, \ldots, K+2$, are the unknown regression coefficients associated with these basis functions. Note that different parameterizations have been considered in additive models; in particular, Kneib and Fahrmeir (2007) employ a parameterization in terms of B-splines.

Instead of simply treating the coefficients θ_{ik} as usual regression effects, their estimation is restricted by adding a penalty to the likelihood in the context of P-splines. This avoids the necessity for sophisticated knot placement strategies and allows to use a moderately large number of equidistant knots. The most commonly used penalty in smoothing approaches is the integral of the squared second derivative

$$\frac{1}{2}\lambda_i\int_0^{\infty}\left[f_i''(z_i)\right]^2 \boldsymbol{d}z_i, \tag{3.1}$$

since a small value of this penalty relates to visually smooth function estimates. The impact of the penalty is governed by smoothing parameters λ_i, and determining appropriate values of these is the most important part of an estimation scheme for additive hazard rate models.

Since (3.1) is a quadratic function of $\boldsymbol{\theta}_i^{\mathrm{T}} = (\theta_{i1}, \ldots, \theta_{iK})$, it can be rewritten as $\frac{1}{2}\lambda_i \boldsymbol{\theta}_i^{\mathrm{T}} \boldsymbol{P}_i \boldsymbol{\theta}_i$, where $\boldsymbol{P}_i$ is a positive semidefinite matrix. With this parameterization, the full penalized log-partial likelihood (PPL) is $PPL(\boldsymbol{\theta}) = l_p(\boldsymbol{\theta}) - \frac{1}{2}\sum_{i=1}^{q} \lambda_i \boldsymbol{\theta}_i^{\mathrm{T}} \boldsymbol{P}_i \boldsymbol{\theta}_i$, where $l_p(\boldsymbol{\theta}) = \ln[PL(\boldsymbol{\theta})]$ denotes the logarithm of the partial likelihood. The parameter estimates for the regression coefficients are then obtained by maximizing *PPL*.

When the penalized splines are represented in terms of B-splines, a convenient approximation to the squared second derivative penalty can be derived based on squared differences of parameters associated with adjacent basis functions (Eilers and Marx, 1996). Still the penalty can be expressed as a quadratic form, enabling the application of penalized Fisher scoring algorithms for estimation. While the integral-based penalty forms the basis of the classical penalized likelihood approach implemented in the *survival* package in R, an empirical Bayes approach has been derived from the difference approximation in Kneib and Fahrmeir (2007). The latter is implemented in the stand-alone software package *BayesX* (Brezger *et al*., 2005) available from http://www.stat.uni-muenchen.de/∼bayesx. Note that the empirical Bayes approach allows for the simultaneous estimation of the baseline hazard rate (also represented as a penalized spline) and the flexible covariate effects based on the full instead of the partial likelihood.

3.2 Controlling the amount of smoothing

One particular concern in fitting P-splines is the selection of reasonable values for smoothing parameters, λ_i (used in the penalized partial likelihood fit (3.1)). For the sake of simplicity, we will discuss a univariate setting first, where only one non-parametric function is estimated. Within the R-function *pspline* (available in the package *survival*), the standard approach is to choose the smoothing parameter according to Akaike's information criterion (AIC; Akaike, 1974) $AIC = -2 \times PPL + 2 \times df$, where *df* represents the equivalent degrees of freedom of the penalized spline fit. The *df* provides an intuitive measure for the complexity of a penalized spline fit that appropriately accounts for the effective dimensionality reduction induced by the penalty (see Gray, 1992, for a definition). Hurvich *et al*. (1998) show that in non-parametric regression, the AIC can under-penalize, leading to models with excessively large degrees of freedom, especially when data are dispersed. They suggest a corrected AIC (cAIC) which uses $n \times (df + 1)/(n - (df + 2))$ as the correction term instead of *df*. In the case of a Cox model, *n* is replaced by the total number of events (Therneau *et al*., 2000). In R/S-plus, the corrected AIC option for the *pspline* is also available.

Minimization of the AIC can simply be achieved by a line search in the univariate setting but becomes increasingly complex in the multivariate setting. The usual work-around is to consider separate one-parameter minimizers, i.e., to determine the amount of smoothness for each of the parameters separately. Note that this strategy crucially depends on the effects being approximately orthogonal and is likely to fail

Table 1 Optimal degrees of freedom for P-spline estimation using both univariate and multivariate survival models. Degrees of freedom df_{AIC} were obtained using the corrected AIC criterion and df_{REML} using REML

(a) Additive Cox model

	Univariate		Multivariate
	df_{AIC}	df_{REML}	df_{REML}
Age	2.25	2.28	1.22
Size	3.32	3.32	1.75
LNI	1.96	1.96	2.03
SPF	2.70	2.97	2.28
DI	11.10	10.36	4.49

(b) Additive multi-state model (recurrence transition)

	Univariate		Multivariate
	df_{AIC}	df_{REML}	df_{REML}
Age	1.73	1.74	1.08
Size	3.38	3.36	2.14
LNI	2.32	2.17	2.06
SPF	2.61	2.91	2.03
DI	12.60	11.39	5.89

(c) Additive multi-state model (mortality transition)

	Univariate		Multivariate
	df_{AIC}	df_{REML}	df_{REML}
Age	1.52	1.57	1.23
Size	1.00	1.18	1.14
LNI	2.90	2.09	1.12
SPF	1.92	1.85	2.00
DI	1.00	1.20	1.08

in circumstances where this assumption is not fulfilled (as will be demonstrated in the application section, see Table 1).

An approach that allows to determine smoothing parameters in both the univariate and the multivariate setting has been proposed by Kneib and Fahrmeir (2007) based on a mixed model representation of penalized splines. The basic idea is to interpret the penalty term as a random effects distribution assigned to the vector of regression effects, which effectively turns the smoothing parameter into a variance component. Concepts from mixed model methodology such as restricted maximum likelihood (REML) estimation can then be adapted to the additive hazard model setting. We will employ this approach as a benchmark reference to the simpler AIC standard approach in our application in Section 4. A fully Bayesian analogue to the empirical Bayes approach by Kneib and Fahrmeir (2007) is proposed in Hennerfeind *et al.* (2006).

3.3 Smooth HR curves

Regardless of the specific parameterization and estimation technique employed to determine the additive model fit, a natural estimate of the adjusted HR curve $HR_{i,hj}(Z_i, z_{i,ref})$ in (2.5), can be constructed as $\widehat{HR}_{i,hj}(Z_i, z_{i,ref}) = \exp\{\hat{f}_{i,hj}(Z_i) - \hat{f}_{i,hj}(z_{i,ref})\}$ by replacing $f_{i,hj}(\cdot)$ by the corresponding P-spline estimate, $\widehat{f}_{i,hj}(\cdot)$. The variance of the log HR estimate, $Ln\widehat{HR}_{i,hj}(Z_i, z_{i,ref})$, can then be expressed as

$$Var\left(Ln\widehat{HR}_{i,hj}\left(Z_i, z_{i,ref}\right)\right) = Var\left(\hat{f}_{i,hj}\left(Z_i\right)\right) + Var\left(\hat{f}_{i,hj}\left(z_{i,ref}\right)\right) - 2Cov\left(\hat{f}_{i,hj}\left(Z_i\right), \hat{f}_{i,hj}\left(z_{i,ref}\right)\right),$$

where the asymptotic covariance matrix $\mathrm{Cov}\left(\hat{f}_{i,hj}\left(Z_i\right), \hat{f}_{i,hj}\left(z_{i,ref}\right)\right)$ takes the form of $H^{-1}IH^{-1}$, with I being the usual observed information and $H = I + P$, where P is the second derivative matrix of the penalty function. The latter is a block-diagonal matrix with blocks $\lambda_i P_i$ in the terms corresponding to the θ_i parameters and zeros elsewhere (Gray, 1992; Kneib and Fahrmeir, 2007). Finally, assuming normality, $(1-\alpha)\,100\%$ pointwise confidence bands can be constructed around the $HR_{i,hj}\left(Z_i, z_{i,ref}\right)$ curve

$$\exp\left\{Ln\widehat{HR}_{i,hj}\left(Z_i, z_{i,ref}\right) \pm z_{1-\alpha/2} SE\left(Ln\widehat{HR}_{i,hj}\left(Z_i, z_{i,ref}\right)\right)\right\},$$

where $SE(Ln\widehat{HR}_{i,hj}(Z_i, z_{i,ref})) = \sqrt{Var(Ln\widehat{HR}_{i,hj}(Z_i, z_{i,ref}))}$ is the standard error of $Ln\widehat{HR}_{i,hj}\left(Z_i, z_{i,ref}\right)$ and $z_{1-\alpha/2}$ is the upper $1-\alpha/2$ quantile of the standard normal distribution.

Software for computing the HR either for the empirical Bayes approach or for the approach based on AIC minimization has been developed and is freely available from the authors. The online supplement to this paper demonstrates their application based on a random sample from the data presented in the following section.

4 Application to breast cancer

The American Society of Clinical Oncology (ASCO) publishes evidence-based clinical practice guidelines for the use of tumour markers in breast cancer since 1996. For the 2007 update (Harris *et al*., 2007), the panel expanded the scope of the guideline to include a broader range of markers in breast cancer, like flow cytometry-based markers. The literature on this is abundant and often not in agreement. Many studies have found DNA measurements of prognostic worth, while others have found DNA measurements of questionable use. Following ASCO, published data are insufficient to recommend use of flow cytometry-based markers to assign patients to prognostic groupings. The disagreements within the literature may be due to the multiple

methods of statistical modelling as well as the variability in the data and the lack of standardization for procedures (Lynn-Eudey, 1996). We believe that the introduction of flexible MSMs, like those proposed in this paper, might support that DNA flow cytometry has an independent prognostic value for predicting the clinical outcome of breast carcinoma patients.

To this aim, all the methods proposed in Section 3 were applied to a Galician breast cancer dataset, using a progressive three-state model with states (*i*) 'alive and disease free', (*ii*) 'alive with recurrence' and (*iii*) 'dead' (see Figure 1). The main aims of this study were: (*i*) to use flexible P-spline methods to study the relationship between different flow cytometry-based proliferation markers (DNA index, DI, and S-phase fraction, SPF) and the evolution of breast cancer disease in a multi-state framework; (*ii*) to furnish effect measures to compare the risk in continuous prognostic factors using a reference value; (*iii*) to compare the Cox model with multi-state modelling with respect to covariate effects and (*iv*) to prove that this methodology could be useful in the identification of prognostic factors associated with different clinical features after surgery for breast cancer.

4.1 Data description

To illustrate our approach, we re-analyzed survival data on 584 incident cases of breast cancer, diagnosed at the Santiago University Teaching Hospital (Hospital Clínico Universitario de Santiago, Santiago de Compostela, Spain) from 1991 through 2000. This study sought to assess the prognostic value of flow cytometry-based proliferation markers, DI and SPF, in breast cancer. The primary medical use of flow cytometry is the indirect measurement of intracellular DNA content. Measurement of the amount of DNA content in tumour cells gives an indication of cell proliferation, as well as cells with an abnormal amount of DNA, and thus may be of prognostic value in cancer studies.

Patients' vital status and date of relapse or death were obtained from their physicians, until end of follow-up on 31 December, 2004. Median follow-up time was 102 months (range, 48–156). During this time, 402 women (69%) were alive and disease free, 167 (29%) experienced a recurrence (local regional or metastases), 117 patients (20%) died due to cancer and 11 due to other causes. These 11 patients are included in the proportional hazards model but in the MSM they are treated as censored on the recurrence transition and they are not considered on the mortality transition from the 'alive with recurrence' state. In this prospective study, the patients—mean age 59 years (range, 23–90)—were not participating in a clinical trial, and treatments were based on physicians' choice.

SPF is defined as the percentage of cells in phase S, in which the cell duplicates its DNA. DI is defined as the ratio of the G0/G1 channel number of tumour cells to the G0/G1 channel number of diploid cells.

Although the study focused mainly on flow cytometry parameters (DI and SPF), factors which are well recognized as being associated with a worse prognosis, namely,

age, tumour size, histological grade (SBR, stages I to III), lymph node involvement (LNI) and hormone receptor status (ER), are also included in the analyses. Detailed descriptions of participants and procedures have been published by Chavez-Uribe *et al*. (2007).

4.2 Univariate analysis

To investigate differences between AIC-based determination of smoothing parameters and the empirical Bayes approach, we first conducted univariate analyses, where only one covariate at a time is entered to the model. In this case, the AIC minimizer can be derived from a line search and we do not have to bother about the impact of the approximate solution for the AIC-based approach in the multivariate setting. Table 1 and Figure 2 illustrate our findings: The results are in all cases quite close to each other, both for the smoothing parameters themselves but especially for the function estimates. While some differences are observed for larger degrees of freedom, these typically do not relate to significant changes in the function estimates. Note, however, that the results from the univariate analyses differ substantially from a multivariate model where all smoothing parameters are determined simultaneously based on REML. In most cases, we observe a reduced complexity in the multivariate approach since now several functions add together to provide the model fit.

4.3 Multivariate analysis

As mentioned before in Section 3.2, the degrees of freedom for the multivariate case cannot be obtained automatically using AIC (in the R *survival* library). To use *survival*, we fix the degrees of freedom as those obtained automatically in REML (using *BayesX*). It should be noticed here that, using the same *df*s in the multivariate analyses, the corresponding estimates for the effect of covariates are virtually the same. As commented earlier, however, an advantage of using *BayesX* is that it allows for the simultaneous P-spline estimation of the baseline hazard rate and several non-linear covariate effects.

4.3.1 Proportional hazards model

Marginal distribution of survival times was analyzed using multivariate Cox regression, with flexible covariate effects being introduced through P-splines for all the continuous predictors. For each continuous covariate, separate results were obtained for linear and non-linear terms for the P-spline fit in the Cox model. Due to the symmetry of the P-spline basis functions, the chi-square test for linearity is a test for zero slope of the regression of the spline coefficients on the centers of the basis functions (Therneau and Grambsch, 2000, 124–26). In addition, a chi-square test for non-linearity was used to test the significance of the deviation from linearity. These

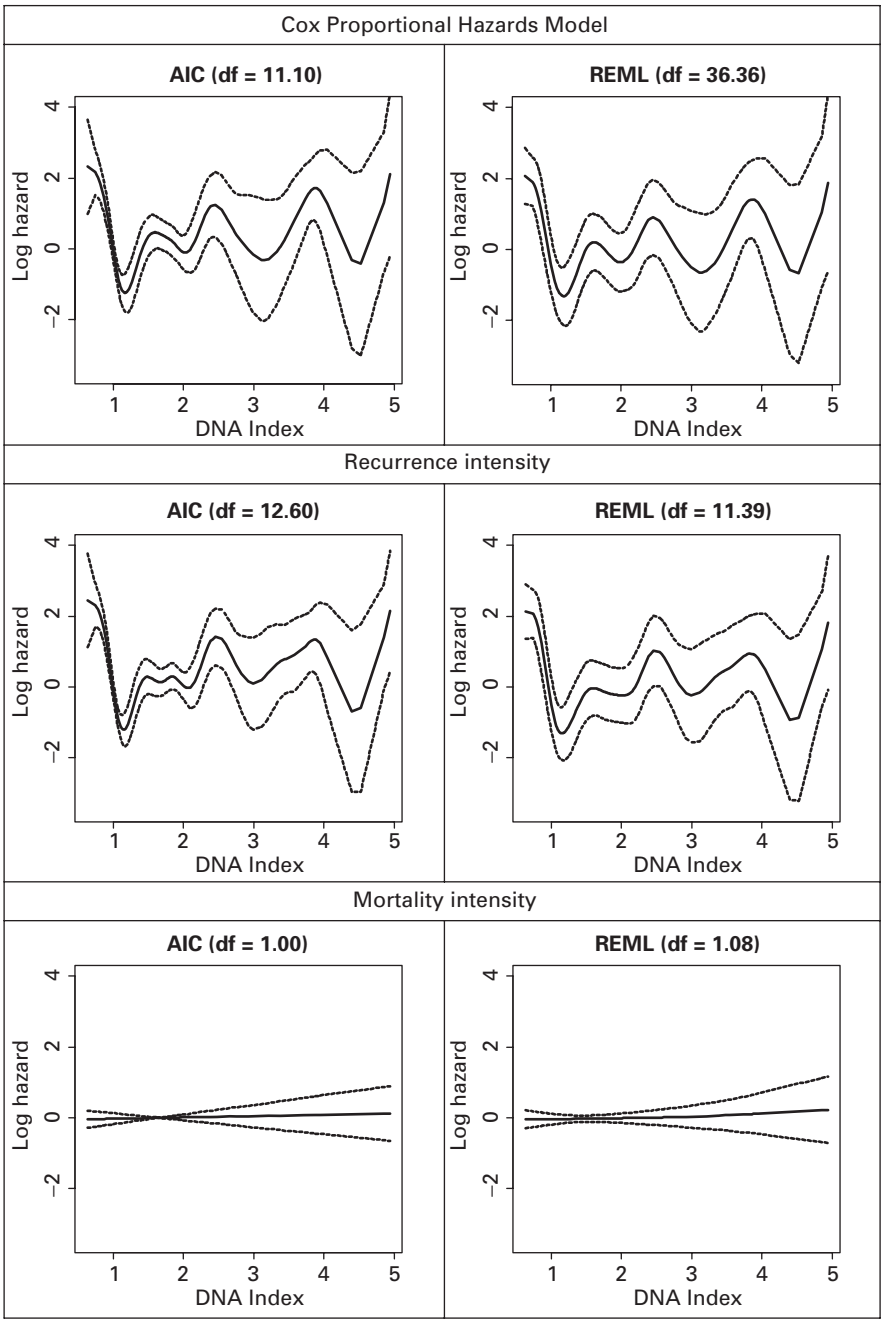

Figure 2 Smooth log hazard with 95% pointwise confidence bands for *DI*. Crude analysis using *survival* (left-hand side) and *BayesX* (right-hand side). Degrees of freedom chosen automatically with cAIC criterion in *survival* and REML in *BayesX*

Table 2 Estimated effects in the additive Cox model

Covariate	$\hat{\beta}$	SE	HR	95% CI	*p*-value
Age (years)					
Age	0.018	0.008	1.019	1.002–1.036	0.028
ps (age, df = 1.22) non-linear	—	—	—	—	ns
Size (mm)					
Size	0.039	0.061	1.039	0.921–1.172	ns
ps (size, df = 1.75) non-linear	—	–	—	—	0.025
LNI (%)					
LNI	0.020	0.003	1.020	1.013–1.027	<0.001
ps (LNI, df = 2.03) non-linear	—	—	—	—	0.037
SBR					
I	–	—	1.000	—	
II	0.143	0.332	1.154	0.601–2.210	ns
III	0.353	0.385	1.423	0.669–3.030	ns
ER					
No	—		1.000	—	
Yes	−0.964	0.279	0.382	0.221–0.659	<0.001
SPF (%)					
SPF	0.094	0.020	1.098	1.056–1.142	<0.001
ps (SPF, df = 2.28) non-linear	—	—	—	—	0.028
DI					
DI	−0.016	0.138	0.984	0.751–1.291	ns
ps (DI, df = 4.49) non-linear	—	—	–	—	0.005

Notes: SE = standard error; HR = hazard ratio; 95% CI = 95% confidence interval; ns = not significant; ps = pspline() function; Size = tumour size; LNI = lymph node involvement; SBR = histologic grading system; ER = estrogen receptors; SPF = S-phase fraction; DI = DNA Index

tests are implemented in R/S-plus within the package *survival*. However, it should be noted that they can only be applied with pre-specified degrees of freedom but not with estimated smoothing parameters. Estimating the degrees of freedom leads to overly sensitive tests that 'detect' even very small deviations from linearity (see Wood, 2006: 194–96, for an explanation).

Results summarized in Table 2 show that the prognostic impact of DI and *Size* proved to be significant only when P-splines were used to estimate the relationship between risk and predictors. Furthermore, percentage of LNI and SPF revealed a significant non-linear effect on survival. *Age*, on the other hand, showed itself to be an important predictor, indicating that the risk of death increases linearly with age. Finally, in contrast to SBR, which had no significant effect on survival, the presence of estrogen receptors (ER) displayed a strong effect. The presence of non-linear effects can be visually inspected and their correct functional form is identified in the graphs shown in Figure 3. In this plot, a P-spline estimate of the (log)-baseline intensity is also added.

Numerical results shown in Table 2 can easily be obtained via the majority of statistical packages. In particular, R and S-plus use the *coxph* function available in *survival* package (Therneau and Grambsch, 2000). Users may also use the *tdc.msm*

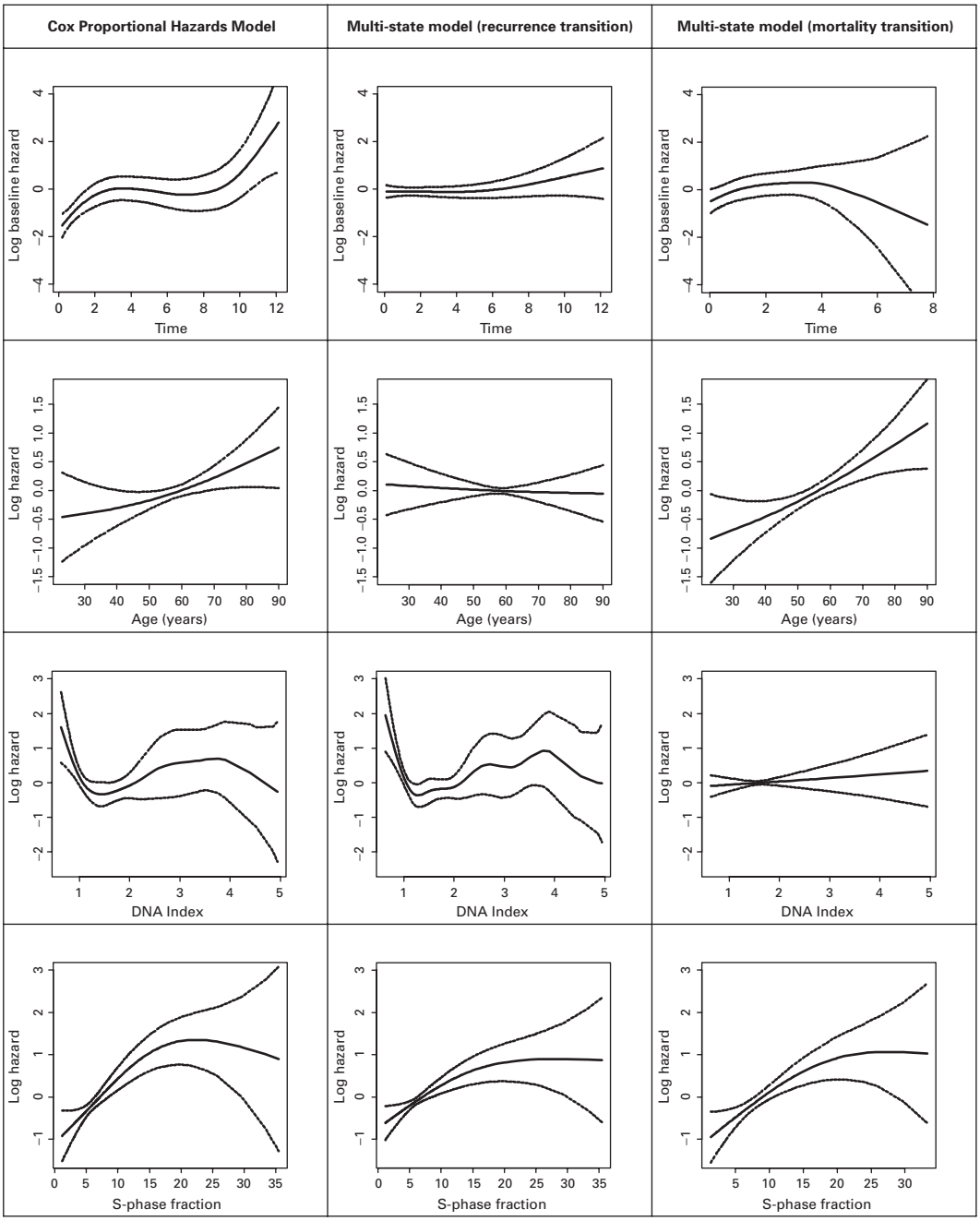

Figure 3 Adjusted smooth log baseline hazard and effects of *age, SPF* and *DI* with corresponding 95% pointwise confidence bands

package (Meira-Machado *et al*., 2007). This software is a user-friendly R package, which can be used not only for the Cox model but also for fitting some MSMs, such as the CSMM. For a detailed description of *tdc.msm* including the input data structure, see Meira-Machado *et al*. (2007).

4.3.2 Multi-state modelling approach

Cancer patients who have experienced a recurrence are known to be at a substantially higher risk of death, thus rendering it essential to understand what characteristics of the patient or the tumour predispose to recurrence. Furthermore, recurrence may possibly alter disease progression, and consequently the role of some prognostic factors may change after such a recurrence. In the analysis of the breast cancer dataset, recurrence (which is a time-dependent covariate, coded as: 0 = no, 1 = yes) may be regarded as an associated state of risk, and the progressive three-state model depicted in Figure 1 can thus be used. This model enables the prognostic factors affecting both recurrence and death to be studied.

Aside from the fact that a patient has experienced recurrence (state 2, Figure 1), in cancer studies the amount of time spent in the healthy state (sojourn time) is often of interest. By including covariates depending on his/her history (Kay, 1986), we verified the assumption that the transition rate from state 2 to state 3 is affected by the time spent in the previous state (p-value < 0.05). This allowed us to conclude that the Markov model of the form (2.3) was unsatisfactory for the Galician breast cancer dataset. In our application, a strong (negative) effect of time since the recurrence on the mortality transition is verified. In such cases, there is a duration effect of $\alpha_{23}(\cdot)$, and so semi-Markov models are often considered (see Meira-Machado *et al*., 2009, for a detailed discussion on this topic). Hence, a Cox semi-Markov model like (2.4) was used to study the way in which prognostic factors affect both recurrence and survival. The results obtained from fitting this model are reported in Table 3.

While the *survival* package requires some extra effort insofar as data preparation is concerned using the counting process notation (Therneau and Grambsch, 2000; Meira-Machado *et al*., 2009), an advantage of using the *tdc.msm* package is the use of the same dataset previously used for Cox analyses. In addition, this program also performs a test for checking the Markov assumption.

The results shown in Table 3 indicate that, save for covariates *Age* and SBR, all the remaining predictors were considered important for recurrence transition. Interestingly, *Age* and SPF displayed a strong linear effect on mortality transition, whereas tumour *size*, LNI, and DI failed to show any significant effects on this transition.

Comparison between the results of fitting a semi-Markov model and those obtained from the traditional Cox model yielded some important biologic insights. For instance, while *Age* registers a statistically significant effect on survival when using a Cox model, it only displayed a significant effect on mortality intensity under the CSMM. Moreover, the non-linear effect of *Size* and DI, which was found to be significant in the Cox model, only revealed a statistically significant effect on

Table 3 Estimated effects in the additive multi-state Cox (semi-Markov) model. MSM for recurrence intensity, $\alpha_{12}(t)$, and MSM for mortality, $\alpha_{23}(t - T_{12})$, after recurrence

MSM for recurrence transition (1 → 2)					
Covariate	$\hat{\beta}_{12}$	SE	HR_{12}	95% CI	*p*-value
Age (years)					
Age	−0.002	0.007	0.998	0.984–1.011	ns
ps(age, df = 1.08) non-linear	—	—	—	—	ns
Size (mm)					
Size	0.019	0.051	1.019	0.922–1.127	ns
ps (size, df = 2.14) non-linear	—	—	—	—	0.018
LNI (%)					
LNI	0.017	0.003	1.017	1.012–1.023	<0.001
ps (LNI, df = 2.06) non-linear	—	—	—	—	0.019
SBR					
I	—	—	1.000	—	
II	0.136	0.258	1.146	0.692–1.900	ns
III	0.046	0.309	1.047	0.572–1.920	NS
ER					
No	—	—	1.000	—	
Yes	−0.667	0.232	0.513	0.326–0.809	0.004
SPF (%)					
SPF	0.060	0.016	1.062	1.030–1.096	<0.001
ps (SPF, df = 2.03) non-linear	—	—	—	—	ns
DI					
DI	0.004	0.114	1.004	0.802–1.256	ns
ps (DI, df = 5.89) non-linear	—	—	—	—	<0.001
MSM for mortality transition (2 → 3)					
Covariate	$\hat{\beta}_{23}$	SE	HR_{23}	95% CI	*p*-value
Age (years)					
Age	0.031	0.009	1.031	1.014–1.049	<0.001
ps (age, df = 1.23) non-linear	—	—	—	—	ns
Size (mm)					
size	0.066	0.064	1.069	0.942–1.212	ns
ps (size, df = 1.14) non-linear	—	—	—	—	ns
LNI (%)					
LNI	0.004	0.003	1.004	0.998–1.011	ns
ps (LNI, df = 1.12) non-linear	—	—	—	—	ns
SBR					
I	—	—	1.000	—	
II	0.253	0.341	1.287	0.659–2.510	ns
III	0.863	0.392	2.371	1.100–5.110	0.027
ER					
No	—	—	1.000	—	
Yes	−0.520	0.264	0.595	0.354–0.998	0.049
SPF (%)					
SPF	0.084	0.021	1.087	1.043–1.133	<0.001
ps (SPF, df = 2) non-linear	—	—	—	—	ns
DI					
DI	0.099	0.142	1.104	0.835–1.459	ns
ps (DI, df = 1.08) non-linear	—	—	—	—	ns

Notes: SE = standard error; HR = hazard ratio; 95% CI = 95% confidence interval; ns = not significant; ps = pspline() function; Size = tumour size; LNI = lymph node involvement; SBR = histologic grading system; ER = estrogen receptors; SPF = S-phase fraction; DI = DNA Index

recurrence. The graphs depicted in Figure 3 illustrate the advantages of using MSMs for analyzing both quantitative predictors and baseline hazards, when compared against the classical Cox model.

Numerical results in Tables 2 and 3 showed the presence of significant non-linear effects for DI (checked through a formal test) when using both Cox and recurrence intensity models. As can be seen in Figure 3, the estimated (non-linear) effects of DI indicated that risk decreased sharply until about 1.28, increased gradually until a value of 3 and then remained roughly constant. A more technical analysis of this quantitative predictor will be given later. The functional form for SPF (in the time-dependent Cox model) indicates that the risk increases rapidly until about value 20% and then remained roughly constant.

For any given covariate, the corresponding parametric HR estimate in transition $h \rightarrow j$, $\exp\left(\hat{\beta}_{hj}\right)$, provides important (and interpretable) information about how such a covariate affects survival (the estimated coefficient $\hat{\beta}_{hj}$ gives the change in log hazard for one-unit change in the covariate). However, the P-spline estimate obtained directly from the output is not yet interpretable. Although the smoothed log hazard curves shown in Figure 3 provide important information about covariate effect on hazard, interpretation is not straightforward since we do not have a reference value. To obtain interpretable results in a simple and summarized manner, we constructed smooth log HR curves with 95% confidence intervals (as proposed in Section 3.3) to describe the relationship between the continuous predictor and risk, when a specific value is taken as reference. Figure 4 shows the corresponding curves for *Age*, SPF and DI. Reference values for these covariates were chosen for biological reasons. For each transition, adjusted log HR curves and corresponding 95% confidence intervals can be obtained using either our new *smooth.HR* software (implemented in R) or the *BayesX* software.

With regard to *Age*, a reference value of 50 years was selected as a possible value for the beginning of menopause. The corresponding plots for this covariate (see Figure 4) showed that in mortality transition, the risk of death was higher for older patients, e.g., a HR of exp(0.642) = 1.9 was obtained when patients aged 70 years were compared with patients aged 50 years (reference value).

The non-linear effect of DI on recurrence is important and worth mentioning. The log HR for DI was obtained taking $DI = 1$ as the reference value (alternatively, one could obtain the corresponding plot using the value with the lower risk as reference). The corresponding log HR curve depicted in Figure 4 can be interpreted as follows: the risk of the event of interest (recurrence) diminishes sharply until a value of 1.28, increases gradually until a value of 3 and then remains roughly constant. The apparent decrease after 3.8 is not significant due to the wide confidence intervals. The main curve represents the log HR for DI, i.e., the (log) expected change in the risk of the event (of interest) when *DI* deviates from 1. For example, an HR of exp(1.813) = 6.129 is obtained when patients with $DI = 0.63$ are compared with someone with a *DI* reference value of 1. This value indicates that a patient having a $DI = 0.63$ at a given point in time has about six times the chance of experiencing

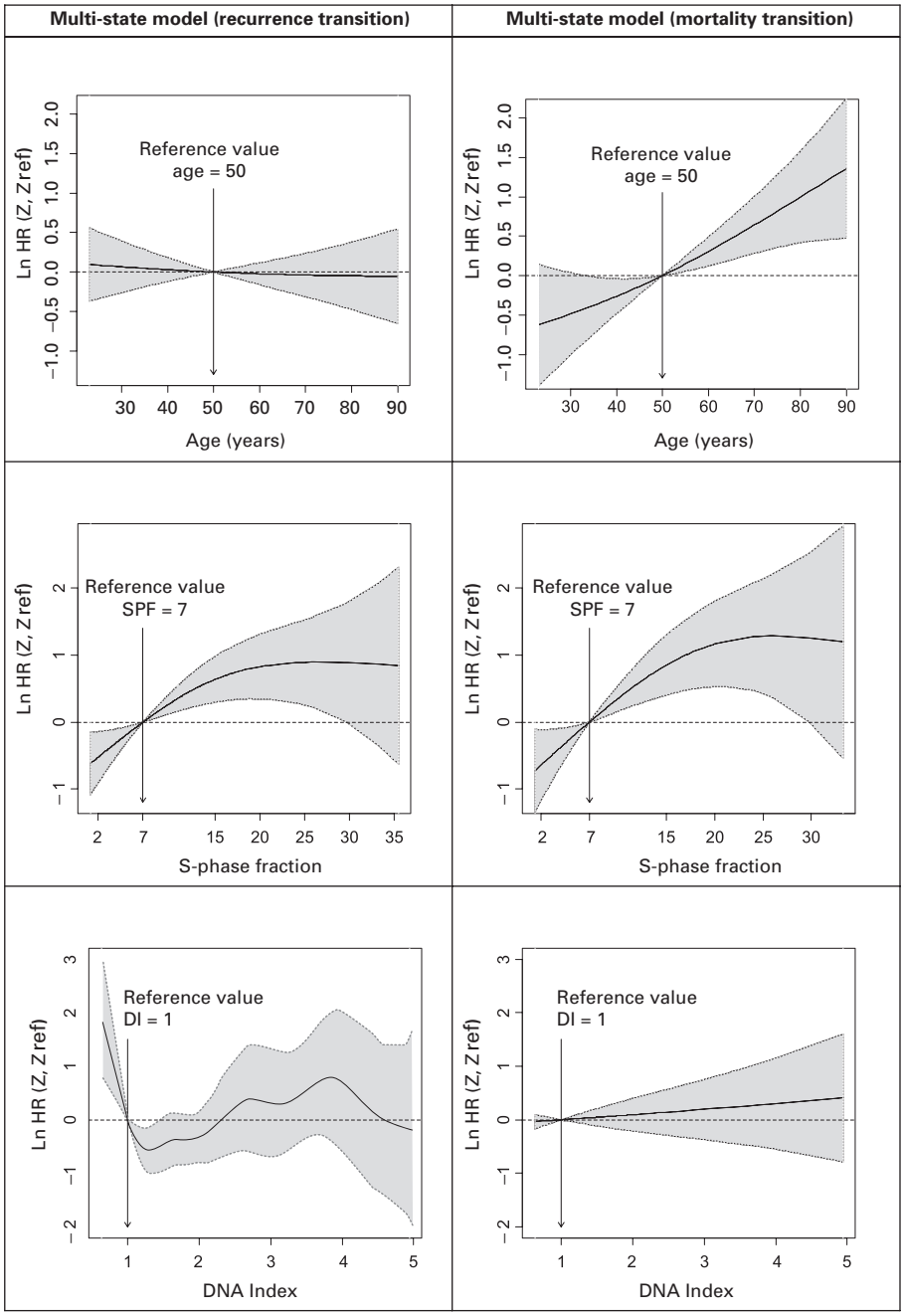

Figure 4 Adjusted smooth log hazard ratio estimates with 95% pointwise confidence bands for age (value 50 as reference), SPF (value 7 as the reference) and DI (value 1 as reference)

Table 4 Recurrence transition (1 → 2). Adjusted log hazard ratio (HR) with 95% confidence intervals for some values of *DI*, taking a value of 1 as reference

	DI	$LnHR_{12}$	95% CI
	0.63	1.813	0.770, 2.856
	0.78	0.994	0.474, 1.514
	0.90	0.404	0.197, 0.611
	0.95	0.190	0.090, 0.291
Reference =	1.00	0.000	—
	1.10	−0.299	−0.484, −0.114
	1.25	−0.505	−0.876, −0.134
	1.50	−0.388	−0.867, 0.090
	2.00	−0.277	−0.741, 0.188
	2.60	0.354	−0.533, 1.242
	4.85	−0.139	−1.590, 1.311

recurrence (being diseased) at the next point in time compared to some patient who is in the same condition in all respects except for having a *DI* of 1. Note that this value would be different if the reference value were chosen differently, e.g., higher values are to be expected for a reference value of $DI = 1.28$. The confidence bands for the log HR are also important, showing that patients with *DI* values lower than 1 are at higher risk and patients with *DI* values in the interval from 1 through 1.4 are at lower risk. Interestingly, there is no evidence of statistical differences in the risk for those patients with a *DI* higher than 1.4 when compared with those having the reference value of $DI = 1$.

In the original analysis of the dataset (Chavez-Uribe *et al.*, 2007), the authors categorized the variable *DI* into three groups, i.e., Group I diploid ($0.96 \leq DI \leq 1.15$), Group II hyperploids ($DI \geq 1.16$) and Group III hypoploids ($DI \leq 0.95$), and they concluded that the patients with hypoploid tumours had a worse prognosis. From the biologic point of view, clinicians expect that tumours with *DI* values close to 1 will have a better prognosis than those distant from 1. This was our reason for choosing the reference value for *DI*. However, Figure 4 shows that patients with or close to $DI = 1.28$ had a significantly smaller risk when compared to those with $DI = 1$. This observation prompted us to carry out a detailed analysis of this seemingly contradictory result. One possible explanation is that a relatively large group of patients may be present with two diploid populations, with $DI = 1$, the same as for normal cells, but one of these populations has abnormal, larger sized cells, thus accounting for a more aggressive biologic behaviour.

Finally, Table 4 lists the adjusted log HRs (with 95% confidence intervals) in recurrence transition for some *DI* values, with 1 being taken as the reference value.

5 Discussion

A flexible approach using additive MSMs for estimating the possible effects of continuous predictors on response is proposed in this paper. This approach not only

allows for visual exploration of the possible effects without prior assumption of a specific functional form but also facilitates expression of the results in a manner that is standard in survival studies.

In this paper, P-splines-based approaches were used as the smoothing technique. Although the P-spline approach was shown to be a good option, other smoothers (e.g., smoothing splines, kernel smoothers) could also be used in this context. It may well be worthwhile for a comparative study to be conducted on the behaviour of the different types of smoothers within the context of the multi-state modelling framework.

When using P-splines, special attention is called for when selecting the optimal amount of smoothing. Usual rules of thumb such as using four degrees of freedom for each of the smooth components are likely to produce artificial results by assuming a fixed, not data-dependent amount of smoothing. Automatic selection based on minimizing the AIC somewhat mitigates this problem in univariate models but is difficult to extend to the multivariate setting typically observed in practice. A possible route for future research might be to consider an extension of the methodology of Eilers *et al*. (2006) to implement a multivariate AIC-based criterion in survival analysis. REML-based estimates obtained in an empirical Bayes approach finally enabled us to determine optimal smoothing parameters also in models with multiple smoothing parameters. Possible extensions that might provide additional insights for variable such as *DI*, where the estimated effects obey considerable variation, would be to consider adaptive function estimation where the single smoothing parameter is replaced by a sequence of smoothing parameters that allow the function estimate to adapt to abrupt changes and local variability. In this context, considering a fully Bayesian MCMC approach might be worthwhile, taking advantage of the hierarchical model formulation of adaptive spline smoothing models.

When using MSMs for analysis of survival data, some care is needed when choosing the appropriate model. Traditionally, statistical methods for analyzing such models depend on Markov's assumption. When using multi-state methodology to analyze the Galician breast cancer dataset, we verified that the amount of time spent in a healthy state was important when modelling mortality (after recurrence) transition. To avoid the possibility of the Markovian assumption proving to be flawed, a (Cox) semi-Markov model was used.

Flexible MSMs can be used as a diagnostic tool for (traditional) semi-parametric MSMs. The Galician breast cancer data nicely illustrate the advantages of using these flexible methods for assessing the possible effect of quantitative predictors not only on survival (through the traditional Cox model) but also on recurrence and mortality after recurrence (through a MSM). The use of MSMs also provides important disease information regarding the effect of flow cytometry parameters considered. Thus, our study revealed that whereas *DI* is only an important predictor of recurrence intensity, *SPF* is a significant predictor of both recurrence and mortality. What must be stressed here, however, is that the important (non-linear) effect of *DI* in the recurrence transition would probably not have been detected by a parametric analysis (in both the linear Cox model and the linear multi-state

framework), in the absence of any prior knowledge of the shape of the corresponding HR curve.

It may be worth pointing out that although the functional form of the HR for a given predictor does not depend on the value used as reference point, the choice of this point does affect HR values and must be taken into account in their interpretation. In this study, the behaviour of the HR curves obtained illustrates the arbitrariness of referring HRs to minima or maxima of an HR curve (very common in practice) when there are several transitions, since the position or presence of such extremes may depend on transition that is being considered (see, e.g., in Figure 4 the different shape of HR for *DI*). In multi-state models, it seems then preferable to proceed as we did in this paper, i.e., to take a point from the background literature as one of the common clinical reference values or to use some value related to clinical normality as the reference point (e.g., value 1 for *DI*).

It is necessary to highlight that the pointwise confidence bands for the HR represent the $100(1-\alpha)/\%$ confidence intervals of the true HR at each value of the covariate, but they do not formally allow us to make global inferences about the true HR curve. However, for the variables that were studied here, the resulting HR curves give us quite a reasonable representation of their relationships with survival.

The identification of prognostic factors associated with different clinical features following surgery for breast cancer remains a major goal in oncology. In this paper, we propose a flexible MSM-based methodology which has proved to be very useful when analyzing breast cancer data. In this multi-state framework, we have described non-linear relationships between continuous predictors and survival and expressed these results as smooth HR curves with confidence intervals where a specific value is taken as reference. These HR curves provide an effective measure that greatly simplifies the information, yielding important biologic insights when applied to breast cancer. The use of this methodology showed which of the prognostic factors were associated: solely with recurrence (*Size*, *LNI* and *DI*); solely with mortality (*Age* and *SBR*) and with both recurrence and mortality (*ER* and *SPF*).

Lastly, it should be noted that, even though the methodology described in this paper has been designed with continuous predictors in mind, it can easily be adapted to structures representing 'factor-by-curve' interactions, in which it might be of interest to calculate HR curves for a continuous predictor that may vary among levels of a categorical covariate. Despite being beyond the scope of the present paper, this is nonetheless an important topic for further research.

Acknowledgements

The authors would gratefully like to acknowledge the financial support received in the form of Spanish Ministry of Education & Science grants MTM2005-00818 and MTM2008-01603 (European FEDER support included) and Galician Regional Authority (*Xunta de Galicia*) grants PGIDIT06PXIC208043PN and INCITE08PXIB208113PR. Thanks are given to María Durbán for helpful

discussions. We are grateful to both the associate editor and the two peer referees for their valuable comments and suggestions, which served to make a substantial improvement to this paper.

References

Akaike H (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–23.

Altman DG, Lausen B, Sauerbrei W and Schumacher M (1994) The dangers of using 'optimal' cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute*, **86**, 829–35.

Andersen PK, Borgan O, Gill RD and Keiding N (1993) *Statistical models based on counting processes*. New York: Springer.

Andersen PK, Esbjerj S and Sorensen TIA (2000) Multi-state models for bleeding episodes and mortality in liver cirrhosis. *Statistics in Medicine*, **19**, 587–99.

Andersen PK and Keiding N (2002) Multi-state models for event history analysis. *Statistical Methods in Medical Research*, **11**, 91–115.

Anderson GL and Fleming TR (1995) Model misspecification in proportional hazards regression. *Biometrika*, **82**, 527–41.

Brezger A, Kneib T and Lang S (2005) BayesX: analyzing bayesian structured additive regression models. *Journal of Statistical Software*, **14**, 1–22 (http://www.jstatsoft.org/).

Brezger A and Lang S (2006) Generalized structured additive regression based on Bayesian P-Splines. *Computational Statistics and Data Analysis*, **50**, 967–91.

Chavez-Uribe E, Cameselle-Teijeiro J, Vinuela JE, *et al.* (2007) Hypoploidy defines patients with poor prognosis in breast cancer. *Oncology Reports*, **17**, 1109–14.

Cox DR (1972) Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 187–220.

de Boor C (2001) *A practical guide to splines* (rev. edn). New York: Springer.

Eilers PHC, Currie ID and Durbán M (2006) Fast and compact smoothing on large multidimensional grids. *Computational Statistics and Data Analysis*, **50**, 61–76.

Eilers PHC and Marx BD (1996) Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89–121.

Govindarajulu US, Spiegelman D, Thurston SW, Ganguli B, *et al.* (2007) Comparing smoothing techniques in Cox models for exposure-response relationships. *Statistics in Medicine*, **26**, 3735–52.

Gray RJ (1992) Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. *Journal of the American Statistical Association*, **87**, 942–51.

Greenland S (1995) Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. *Epidemiology*, **6**, 356–65.

Harris L, Fritsche H, Mennel R, Norton L, *et al.* (2007) American Society of Clinical Oncoloy 2007 update of recommendations for the use of tumor markers in breast cancer. *Journal of Clinical Oncology*, **25**, 5287–312.

Hastie TJ and Tibshirani RJ (1990a) *Generalized additive models*. London: Chapman & Hall.

Hastie TJ and Tibshirani RJ (1990b) Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, **46**, 1005–16.

Hennerfeind A, Brezger A and Fahrmeir L (2006) Geoadditive survival models. *Journal of the American Statistical Association*, **101**, 1065–75.

Hougaard P (1999) Multi-state models: a review. *Lifetime Data Analysis*, **5**, 239–64.

Huang JZ, Kooperberg C, Stone CJ and Truong YK (2000) Functional ANOVA modelling for proportional hazards regression. *Annals of Statistics*, **28**, 960–99.

Huang JZ and Liu L (2006) Polynomial spline estimation and inference of proportional hazards regression models with flexible relative risk form. *Biometrics*, **62**, 793–802.

Hurvich CM, Simonoff JS and Tsai ChL (1998) Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society, Series B*, **60**, 271–93.

Kay R (1986) A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics*, **42**, 855–65.

Kay R (2004) An explanation of the hazard ratio. *Pharmaceutical Statistics*, **3**, 295–97.

Kneib T and Fahrmeir L (2007) A mixed model approach for geoadditive hazard regression. *Scandinavian Journal of Statistics*, **34**, 207–28.

Kneib T and Hennerfeind A (2008) Bayesian semiparametric multi-state models. *Statistical Modelling*, **8**, 169–98.

Lang S and Brezger A (2004) Bayesian P-splines. *Journal of Computational and Graphical Statistics*, **13**, 183–212.

Lynn-Eudey T (1996) Statistical considerations in DNA flow cytometry. *Statistical Science*, **11**, 320–34.

Martinussen T and Scheike T (2006) *Dynamic regression models for survival data*. Berlin: Springer.

Meira-Machado L, Cadarso-Suárez C and de Uña-Álvarez J (2007) tdc.msm: an R library for the analysis of multi-state survival data. *Computer Methods and Programs in Biomedicine*, **86**, 131–40.

Meira-Machado L, de Uña-Álvarez J, Cadarso-Suárez C and Andersen PK (2009) Multi-state models for the analysis of time to event data. *Statistical Methods in Medical Research*, **18**, 195–222.

Putter H, Fiocco M and Geskus RB (2007) Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, **11**, 2389–430.

Royston P and Altman DG (1994) Regression using fractional polynomials of continuous covariates: parsimonious parametric modeling. *Applied Statistics*, **43**, 425–67.

Spruance SL, Reid JE, Grace M and Samore M (2004) Hazard ratio in clinical trials. *Antimicrobial Agents and Chemotherapy*, **48**, 2787–92.

Strasak AM, Lang S, Kneib T, Brant L, *et al.* (2009) Use of penalized splines in extended cox-type additive hazard regression to flexibly estimate the effect of time-varying serum uric acid on risk of cancer incidence: a prospective, population-based study in 78850 men. *Annals of Epidemiology*, **19**, 15–24.

Struthers CA and Kalbfeisch JD (1986) Misspecifed proportional hazards models. *Biometrika*, **73**, 363–69.

Therneau TM and Grambsch PM (2000) *Modelling survival data: extending the Cox model*. New York: Springer.

Therneau TM, Grambsch PM, and Pankratz VS (2000) Penalized survival models and frailty. Technical report 66. Department of Health Science Research, Mayo Clinic, Rochester, Minnesota.

Wahba G (1990) *Spline functions for observational data* (CBMSNSF regional conference series). Philadelphia, PA: Society for Industrial and Applied Mathematics.

Wood SN (2006) *Generalized additive models: an introduction with R*. Boca Raton, FL: Chapman & Hall/CRC press.