

# A spline-based framework for the flexible modelling of continuously observed multistate survival processes

Alessia Eletti<sup>1</sup>, Giampiero Marra<sup>1</sup> and Rosalba Radice<sup>2</sup>

<sup>1</sup>Department of Statistical Science, University College London, UK

<sup>2</sup>Faculty of Actuarial Science and Insurance, Bayes Business School, City, University of London, UK

**Abstract:** Multistate modelling is becoming increasingly popular due to the availability of richer longitudinal health data. When the times at which the events characterising disease progression are known, the modelling of the multistate process is greatly simplified as it can be broken down in a number of traditional survival models. We propose to flexibly model them through the existing general link-based additive framework implemented in the R package GJRM. The associated transition probabilities can then be obtained through a simulation-based approach implemented in the R package *mstate*, which is appealing due to its generality. The integration between the two is seamless and efficient since we model a transformation of the survival function, rather than the hazard function, as is commonly found. This is achieved through the use of shape constrained P-splines which elegantly embed the monotonicity required for the survival functions within the construction of the survival functions themselves. The proposed framework allows for the inclusion of virtually any type of covariate effects, including time-dependent ones, while imposing no restriction on the multistate process assumed. We exemplify the usage of this framework through a case study on breast cancer patients.

**Key words:** additive predictor, multistate process, shape constrained P-splines, survival analysis, transition probabilities

**Received** September 2022; **revised** April 2023; **accepted** April 2023

## 1 Introduction

When considering multistate processes for the modelling of life-history data, a particularly advantageous setting is that in which transition times are known exactly, that is, the process is continuously observed. In this case, in fact, the overall model likelihood can be decomposed into the product of likelihoods referring to each specific transition only. Estimation then becomes equivalent to fitting one standard survival model for each transition, considering only the subset of the data relevant to that transition and including left-truncation times if the transition at hand can only happen once another has occurred. This is referred to as *separate estimation* (Putter et al., 2007; Putter, 2011;

---

Address for correspondence: Alessia Eletti, Department of Statistical Science, University College London, Gower Street, WC1E 6BT London, UK.  
E-mail: [alessia.eletti.19@ucl.ac.uk](mailto:alessia.eletti.19@ucl.ac.uk)



Crowther and Lambert, 2017). An important practical implication of this is that existing tools can be used to fit the transition-specific models. In particular, we propose to model each transition intensity through the general link-based additive modelling framework by Eletti et al. (2022), implemented in the R package GJRM (Marra and Radice, 2023). This modelling framework allows for the inclusion of virtually any type of covariate effects (including time-dependent effects) using any type of smoother (e.g., thin plate and cubic splines, and tensor products). Importantly, the use of shape constrained P-splines (SCOPs) to model time effects permits to approach the multiple univariate survival models directly on the survival scale, rather than on the hazards scale (which would require expensive numerical integration), while retaining a high degree of modelling flexibility. Specifically, SCOPs, developed by Pya and Wood (2015), extending the penalised B-splines discussed in the seminal work of Eilers and Marx (1996), elegantly embed the monotonicity required for the survival functions within the construction of the survival functions themselves, thus enabling very efficient parameter estimation. The exploration of different forms of dependence on past history also becomes considerably easier when the exact transition times are known. Indeed, assuming a semi-Markov process, the most common relaxation considered in the literature, rather than a Markov process, the most commonly made assumption, implies no further methodological difficulty.

When dealing with life-history data, one is often interested in assessing the effects of specific risk-factors on the probability of transitioning between states. When the process is assumed to be time-dependent and/or not-Markov, the computation of the transition probabilities is a nontrivial task. Two main approaches can be identified in the literature to address this problem and are detailed in Supplementary Material A. We adopt a simulation-based approach which allows one to compute the transition probabilities by simulating a number of paths through the assumed multistate process and counting the number of individuals experiencing each transition (Iacobelli and Carstensen, 2013; Touraine et al., 2016). This is appealing due its aptness at supporting any type of multistate process and was proposed in Fiocco et al. (2008) and implemented, amongst others, in the R package `mstate` (Putter et al., 2020), whose tools can be seamlessly integrated with the estimation approach implemented in the R package GJRM.

The remainder of the article is organised as follows. In Section 2, the mathematical setting of multistate survival processes is described, while Section 3 introduces the modelling framework. Sections 4, 5 and 6 discuss model estimation, the extraction of the transition probabilities and inference respectively. In Section 7, the *Rotterdam Breast Cancer Study* is introduced to exemplify the proposed framework. Finally, Section 8 provides some concluding remarks alongside directions of future work.

## 2 Mathematical setting of multistate survival processes

A continuous-time discrete-state stochastic process is a family of random variables  $\{Z(t), t \in \mathcal{T}\}$  with some indexing set given by  $\mathcal{T} = [0, \infty)$  in the survival setting. The set of all values that the process takes  $\mathcal{S} := \{z : Z(t) = z, t \in \mathcal{T}\} \subseteq \{0, 1, 2, \dots\}$  is called the state space, where  $Z(t)$  denotes the state occupied at time  $t$ . A  $p \times 1$  vector of left-continuous, time-dependent covariates is represented by  $X(t)$ . The history of the process, including the evolution of the covariates vector, is denoted by  $\mathcal{F}_t = \{Z(u), X(u), 0 \leq u \leq t\}$ . The transition intensities and the transition probabilities are then the two key quantities associated with the process. The former represent the rates of transition to a state

$s$  for an individual who is currently in another state  $r$ , formally

$$q^{(rs)}(t | \mathcal{F}_{t^-}) = \lim_{\Delta t \downarrow 0} \frac{P(Z(t + \Delta t^-) = s | Z(t^-) = r, \mathcal{F}_{t^-})}{\Delta t}, \quad r \neq s,$$

with  $q^{(rs)}(t | \mathcal{F}_{t^-}) = 0$  if  $r$  is an absorbing state and  $q^{(rr)}(t | \mathcal{F}_{t^-}) = -\sum_{s \neq r} q^{(rs)}(t | \mathcal{F}_{t^-})$ . The matrix

with  $(r, s)$  element given by  $q^{(rs)}(t | \mathcal{F}_{t^-})$  for every  $r, s \in \mathcal{S}$  is called transition intensity matrix or generator matrix and we will denote it by  $\mathbf{Q}(t | \mathcal{F}_{t^-})$ . Similarly, we define the transition probability matrix associated with the time interval  $[u, t]$  as the matrix with  $(r, s)$  element given by  $P(Z(t) = s | Z(u) = r, \mathcal{F}_{u^-})$  and denote this by  $\mathbf{P}(u, t | \mathcal{F}_{u^-})$ . It is common to simplify the dependence on past history and time by assuming either a Markov or a semi-Markov process. The former implies that the probability of being in a given state at a given future time only depends on the current state occupied (Ross et al., 1996). The latter assumes that the future state only depends on the history of the process through the current state and through time since entry to the current state (Pyke, 1961; Yang and Nair, 2011). Exact knowledge of the transition times, as in our setting, allows for both assumptions to be modelled in an equally straightforward manner. The time for intermediate transitions will just need to be re-defined to be the time from entry to the current state.

### 3 Flexible transition-specific modelling

When a multistate process is continuously observed, each transition time can be viewed as a standalone time-to-event and can thus be modelled through traditional survival analysis. It is well known that survival analysis can be undertaken on different scales. One such option is to model transformations of the survival function using generalised survival models, a class that was first introduced by Younes and Lachin (1997). Subsequent works further developed this approach (e.g., Royston and Parmar, 2002; Liu et al., 2018), each allowing for more modelling flexibility and ensuring the monotonicity of the survival function in different ways. More recently Marra and Radice (2020) proposed a generalised survival modelling framework which elegantly embeds the monotonicity of the survival function within the model design matrix by exploiting the properties of P-splines (see Section 3.2). We adopt this approach and thus describe it in the following in the context of transition-specific modelling.

Let  $\mathcal{A} = \{(r, s) | r \neq s \in \mathcal{S}, q^{(rs)}(t_i) \neq 0\}$  be the set of transitions and  $N$  represent the sample size. For observation  $i = 1, \dots, n$  and for  $(r, s) \in \mathcal{A}$ , let  $H^{(rs)}(\cdot)$  be the cumulative transition-specific hazard defined in terms of the transition intensity  $q^{(rs)}(\cdot)$  as  $H^{(rs)}(t_i | \mathbf{x}_i; \boldsymbol{\beta}^{(rs)}) = \int_0^{t_i} q^{(rs)}(u | \mathbf{x}_i; \boldsymbol{\beta}^{(rs)}) du$ . Then we will have a conditional survival function denoted by  $S^{(rs)}(t_i | \mathbf{x}_i; \boldsymbol{\beta}^{(rs)}) = \exp\{-H^{(rs)}(t_i | \mathbf{x}_i; \boldsymbol{\beta}^{(rs)})\} \in (0, 1)$ , where  $\mathbf{x}_i$  represents a generic vector of patient characteristics that has an associated regression coefficient vector  $\boldsymbol{\beta}^{(rs)} \in \mathbb{R}^w$ , where  $w$  is the length of  $\boldsymbol{\beta}^{(rs)}$ . A link-based additive transition-specific survival model can then be written as

$$g\left\{S^{(rs)}(t_i | \mathbf{x}_i; \boldsymbol{\beta}^{(rs)})\right\} = \eta_i^{(rs)}(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}^{(rs)})), \quad (3.1)$$

**Table 1** Functions implemented in GJRM.  $\Phi$  and  $\phi$  are the cumulative distribution and density functions of a univariate standard normal distribution. Note: the desired link-function can be specified by setting the argument `margin` of the function `gamlss()` in GJRM to the values within brackets; for example, `margin = 'PH'`.

Model	Link $g(\mathcal{S})$	Inverse Link $g^{-1}(\eta) = G(\eta)$	$G'(\eta)$
Prop. hazards or log-log ('PH')	$\log\{-\log(\mathcal{S})\}$	$\exp\{-\exp(\eta)\}$	$-G(\eta)\exp(\eta)$
Prop. odds or logit ('PO')	$-\log\left(\frac{\mathcal{S}}{1-\mathcal{S}}\right)$	$\frac{\exp(-\eta)}{1+\exp(-\eta)}$	$-G^2(\eta)\exp(-\eta)$
Probit ('probit')	$-\Phi^{-1}(\mathcal{S})$	$\Phi(-\eta)$	$-\phi(-\eta)$

where  $g : (0, 1) \rightarrow \mathbb{R}$  is a monotone and twice continuously differentiable link function with bounded derivatives, hence invertible, which determines the scale of the analysis,  $\eta_i^{(r,s)}(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}^{(r,s)})) \in \mathbb{R}$  is an additive predictor which includes a baseline function of time and several types of covariate effects and  $\mathbf{f}(\boldsymbol{\beta}^{(r,s)})$  is a vector function of  $\boldsymbol{\beta}^{(r,s)}$  through which the monotonicity required for the survival functions is imposed (see Section 3.2). Rearranging (3.1) yields  $S^{(r,s)}(t_i | \mathbf{x}_i; \boldsymbol{\beta}^{(r,s)}) = G\left\{\eta_i^{(r,s)}(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}^{(r,s)}))\right\}$ , where  $G$  is an inverse link function. Note that modelling directly on the survival scale implies a considerable advantage in this context (see Section 5). The cumulative transition-specific hazard is then  $H^{(r,s)}(t_i | \mathbf{x}_i; \boldsymbol{\beta}^{(r,s)}) = -\log\left[G\left\{\eta_i^{(r,s)}(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}^{(r,s)}))\right\}\right]$  and the transition intensity function is defined as

$$q^{(r,s)}(t_i | \mathbf{x}_i; \boldsymbol{\beta}^{(r,s)}) = -\frac{G'\left\{\eta_i^{(r,s)}(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}^{(r,s)}))\right\}}{G\left\{\eta_i^{(r,s)}(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}^{(r,s)}))\right\}} \frac{\partial \eta_i^{(r,s)}(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}^{(r,s)}))}{\partial t_i}, \quad (3.2)$$

where  $G'\left\{\eta_i^{(r,s)}(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}^{(r,s)}))\right\} = \partial G\left\{\eta_i^{(r,s)}(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}^{(r,s)}))\right\} / \partial \eta_i^{(r,s)}(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}^{(r,s)}))$ . Table 1 displays the functions  $g$ ,  $G$  and  $G'$  available in the R package GJRM.

### 3.1 Additive predictor

Dropping the dependence on covariates and on parameters for the sake of simplicity, the additive predictor is defined as

$$\eta_i^{(r,s)} = \beta_0^{(r,s)} + \sum_{k=1}^{K^{(r,s)}} s_k^{(r,s)}(\mathbf{z}_{ki}), \quad i = 1, \dots, n, \quad (3.3)$$

where  $\beta_0^{(r,s)} \in \mathbb{R}$  is an overall intercept,  $\mathbf{z}_{ki}$  denotes the  $k^{\text{th}}$  sub-vector of the complete vector  $\mathbf{z}_i$  and the  $K^{(r,s)}$  functions  $s_k^{(r,s)}(\mathbf{z}_{ki})$  denote effects which are chosen according to the type of covariate(s) considered. These functions can be expressed as a linear combination of basis functions  $\mathbf{b}_k(\mathbf{z}_{ki}) = (b_{k1}^{(r,s)}(\mathbf{z}_{ki}), \dots, b_{kJ_k}^{(r,s)}(\mathbf{z}_{ki}))^\top$  and regression coefficients  $\mathbf{f}_k^{(r,s)}(\boldsymbol{\beta}_k^{(r,s)}) = (f_{k1}^{(r,s)}(\beta_{k1}^{(r,s)}), \dots, f_{kJ_k}^{(r,s)}(\beta_{kJ_k}^{(r,s)}))^\top \in \mathbb{R}^{J_k}$ , that is  $s_k^{(r,s)}(\mathbf{z}_{ki}) = \mathbf{b}_k(\mathbf{z}_{ki})^\top \mathbf{f}_k^{(r,s)}(\boldsymbol{\beta}_k^{(r,s)})$  (e.g., Wood, 2017). We can then write

(3.3) compactly as  $\eta_i^{(r,s)} = \mathbf{Z}_i^{(r,s)\top} \mathbf{f}^{(r,s)}(\boldsymbol{\beta}^{(r,s)})$ , where  $\mathbf{Z}_i^{(r,s)} = (1, \mathbf{b}_1(\mathbf{z}_{1i})^\top, \dots, \mathbf{b}_{K^{(r,s)}}(\mathbf{z}_{K^{(r,s)}i})^\top)^\top$  and  $\mathbf{f}^{(r,s)}(\boldsymbol{\beta}^{(r,s)}) = (\beta_0^{(r,s)}, \mathbf{f}_1^{(r,s)}(\boldsymbol{\beta}_1^{(r,s)})^\top, \dots, \mathbf{f}_{K^{(r,s)}}^{(r,s)}(\boldsymbol{\beta}_{K^{(r,s)}}^{(r,s)})^\top)^\top$ . Observe that  $\partial \eta_i^{(r,s)}(t_i, \mathbf{x}_i; \mathbf{f}^{(r,s)}(\boldsymbol{\beta}^{(r,s)})) / \partial t_i$  is required in (3.2). This can be expressed as  $\mathbf{Z}_i^{(r,s)}(t_i, \mathbf{x}_i)^\top \mathbf{f}^{(r,s)}(\boldsymbol{\beta}^{(r,s)})$  where, depending on the type of spline basis employed,  $\mathbf{Z}_i^{(r,s)}(t_i, \mathbf{x}_i)' = \lim_{\varepsilon \rightarrow 0} \frac{\mathbf{Z}_i^{(r,s)}(t_i + \varepsilon, \mathbf{x}_i) - \mathbf{Z}_i^{(r,s)}(t_i - \varepsilon, \mathbf{x}_i)}{2\varepsilon}$  can be calculated either by a finite-difference method or analytically. Each  $\boldsymbol{\beta}_k^{(r,s)}$  has an associated quadratic penalty  $\lambda_k^{(r,s)} \boldsymbol{\beta}_k^{(r,s)\top} \mathbf{D}_k^{(r,s)} \boldsymbol{\beta}_k^{(r,s)}$ , used in fitting, whose role is to enforce specific properties on the  $k^{\text{th}}$  function, such as smoothness, with matrix  $\mathbf{D}_k^{(r,s)}$  depending only on the choice of the basis functions. The smoothing parameter  $\lambda_k^{(r,s)} \in [0, \infty)$  controls the trade-off between fit and smoothness, and hence determines the shape of the estimated smooth function. The overall penalty can be defined as  $\boldsymbol{\beta}^{(r,s)\top} \mathbf{S}_{\boldsymbol{\lambda}^{(r,s)}}^{(r,s)} \boldsymbol{\beta}^{(r,s)}$ , where  $\mathbf{S}_{\boldsymbol{\lambda}^{(r,s)}}^{(r,s)} = \text{diag}(0, \lambda_1^{(r,s)} \mathbf{D}_1^{(r,s)}, \dots, \lambda_{K^{(r,s)}}^{(r,s)} \mathbf{D}_{K^{(r,s)}}^{(r,s)})$  is a block diagonal matrix where each block is given by the  $k^{\text{th}}$  penalty, and where  $\boldsymbol{\lambda}^{(r,s)} = (\lambda_1^{(r,s)}, \dots, \lambda_{K^{(r,s)}}^{(r,s)})^\top$  is the transition-specific overall smoothing parameter vector. Depending on the types of covariate effects one wishes to model, several definitions of basis functions are possible, for example, thin plate, cubic and P- regression splines, tensor products, Markov random fields, random effects, Gaussian process smooths. These are handled automatically within the software proposed. We refer the reader to Section 7 for practical examples of the effects mentioned above and to Wood (2017) for the other available options.

### 3.2 Imposing monotonicity by means of SCOPs

When modelling life-history data through multistate processes, one is often interested in making statements in terms of the probabilities of transitioning from one state to another for specific combinations of risk-factors. In Section 5, it will be shown that we compute these by first extracting the transition-specific cumulative hazards at various time points. Direct modelling of the survival functions thus allows us to obtain the transition probabilities more cheaply, as we drop the intermediate step of having to first integrate the transition intensities. The only caveat is that one needs to ensure the survival functions are monotonically decreasing. Liu et al. (2018) propose to do this by means of a penalty applied to the hazard function such that the associated coefficient is iteratively doubled until the estimated hazard functions of all individuals are not negative. We employ a more theoretically founded approach. Indeed, in the proposed framework the properties of P-splines are exploited to elegantly embed the monotonicity within the construction of the survival functions themselves, while allowing for the flexible modelling of the time effect.

Let  $s^{(r,s)}(t_i) = \sum_{j=1}^{J^{(r,s)}} f_j^{(r,s)}(\boldsymbol{\beta}_j^{(r,s)}) b_j^{(r,s)}(t_i)$ , where the  $b_j^{(r,s)}(\cdot)$  are B-spline basis functions of at least second order built over the interval  $[a, b]$ , based on equally spaced knots, and the  $f_j^{(r,s)}(\boldsymbol{\beta}_j^{(r,s)})$  are spline coefficients. Given the link functions listed in Table 1, we need  $s^{(r,s)}(t_i) \geq 0$ . Eilers and Marx (1996) combined B-spline basis functions with discrete penalties in the basis coefficients to produce the popular P-spline smoothers. Then Pya and Wood (2015) proposed shape constrained P-splines through a mildly nonlinear extension of these P-splines, with corresponding novel discrete penalties, thus allowing the development of efficient and stable model estimation

frameworks, such as the one proposed. In particular, a sufficient condition for  $s^{(r,s)}(t_i) \geq 0$  over  $[a, b]$  is that  $f_j^{(r,s)}(\beta_j^{(r,s)}) \geq f_{j-1}^{(r,s)}(\beta_{j-1}^{(r,s)})$ ,  $\forall j$ . Indeed, given a function  $\eta(x) = a_0 + \sum_{j=1}^m a_j B_j(x, q)$ , where  $B_j(x, q)$  are the bases for a  $(q+1)^{th}$  order B-spline,  $m$  is the number of basis functions,  $\partial\eta(x)/\partial x = \frac{1}{h} \sum_{j=1}^{m-1} (a_{j+1} - a_j) B_j(x, q-1)$  with  $h$  the distance between equally spaced knots and so  $a_{j+1} \geq a_j$  implies  $\partial\eta(x)/\partial x \geq 0$  since  $B_j(x, q-1) \geq 0$  (Leitenstorfer and Tutz, 2007). Such condition can be imposed by defining the vector function  $\mathbf{f}^{(r,s)}(\boldsymbol{\beta}^{(r,s)}) = \boldsymbol{\Sigma} \left\{ \beta_1^{(r,s)}, \exp(\beta_2^{(r,s)}), \dots, \exp(\beta_{J^{(r,s)}}^{(r,s)}) \right\}^\top$ , where  $\boldsymbol{\Sigma}[\iota_1, \iota_2] = 0$  if  $\iota_1 < \iota_2$  and  $\boldsymbol{\Sigma}[\iota_1, \iota_2] = 1$  if  $\iota_1 \geq \iota_2$ , with  $\iota_1$  and  $\iota_2$  denoting the row and column entries of  $\boldsymbol{\Sigma}$ , and  $\boldsymbol{\beta}^{(r,s)\top} = (\beta_1^{(r,s)}, \beta_2^{(r,s)}, \dots, \beta_{J^{(r,s)}}^{(r,s)})$  is the parameter vector to estimate. Crucially, in practice  $\boldsymbol{\Sigma}$  is absorbed into the design matrix containing the B-spline basis functions  $\mathbf{Z}$ , hence allowing the constraint to be elegantly embedded within the construction of the model design matrix itself. Finally, in a smoothing context, we are interested in having a penalty on the smooth function to control its ‘wiggleness’. Eilers and Marx (1996) introduced the notion of directly penalising the difference in the basis coefficients of a B-splines basis, which is used with a relatively large number of basis functions to avoid underfitting. The adaptation to the shape-constrained case is straightforward as it implies penalising the squared differences between adjacent  $\beta_j^{(r,s)}$ , starting from  $\beta_2^{(r,s)}$ , using  $\mathbf{D}^{(r,s)} = \mathbf{D}^{(r,s)*\top} \mathbf{D}^{(r,s)*}$  where  $\mathbf{D}^{(r,s)*}$  is a  $(J^{(r,s)} - 2) \times J^{(r,s)}$  matrix made up of zeros except that  $\mathbf{D}^{(r,s)*}[l, \iota + 1] = -\mathbf{D}^{(r,s)*}[l, \iota + 2] = 1$  for  $\iota = 1, \dots, J^{(r,s)} - 2$ . The penalty is zeroes when all the  $\beta_j^{(r,s)}$  after  $\beta_1^{(r,s)}$  are equal so that the  $f_j^{(r,s)}(\beta_j^{(r,s)})$  form a uniformly increasing sequence and  $s^{(r,s)}(t_i)$  is an increasing straight line. As a result, the proposed penalty shares the basic feature of smoothing towards a straight line, but in a manner that is computationally convenient for constrained smoothing.

## 4 Estimation

Since each likelihood contribution refers to a specific transition only and every transition is exactly observed if and only if it occurs, it can be shown (see Supplementary Material B) that the overall model log-likelihood can be broken down into the sum of the log-likelihoods associated with each transition, which are functions only of the parameters relating to that transition, that is,  $\ell(\boldsymbol{\theta}) = \sum_{(r,s) \in \mathcal{A}} \ell^{(r,s)}(\boldsymbol{\beta}^{(r,s)})$ , where  $\boldsymbol{\theta} = \{\boldsymbol{\beta}^{(r,s)} \mid (r, s) \in \mathcal{A}\}$  is an overall model parameter vector. Rewriting the log-likelihood in this way, rather than as a sum of contributions associated with each observation time, is more convenient as it breaks down the estimation task into a number of traditional survival problems, one for each transition. It is precisely to each of these transition-specific models that the framework developed in Eletti et al. (2022) is applied. Briefly, as the model allows for a high degree of flexibility, to prevent over-fitting, the log-likelihood is augmented with a penalty term  $\ell_p^{(r,s)}(\boldsymbol{\beta}^{(r,s)}) = \ell^{(r,s)}(\boldsymbol{\beta}^{(r,s)}) - \frac{1}{2} \boldsymbol{\beta}^{(r,s)\top} \mathbf{S}_{\lambda^{(r,s)}}^{(r,s)} \boldsymbol{\beta}^{(r,s)}$  where  $\mathbf{S}_{\lambda^{(r,s)}}^{(r,s)}$  is an overall penalty term defined in Section 3. The estimation framework then combines a carefully structured trust region algorithm which uses the analytical expressions of the gradient and Hessian of the log-likelihood and properly chosen starting values with a general automatic multiple smoothing parameter selection algorithm based on an approximate AIC measure.

## 5 Prediction on the transition probabilities scale

While estimation can be carried out entirely by-passing the computation of the transition probabilities, one is often interested in making statements in terms of the probability of transitioning from one state to another given a specific combination of risk-factors. We choose the simulation-based approach proposed in Fiocco et al. (2008), which we briefly describe in the following. Let  $r$  be the starting state, entered at time  $t_r = 0$ , and  $t_{\max}$  the maximum follow-up time. Then

- Let  $\mathcal{B}$  be the set of states that can be reached from state  $r$ . If  $\mathcal{B}$  is empty, stop. Otherwise, for  $s \in \mathcal{B}$ , let  $H^{(r,s)}(t)$  be the cumulative transition-specific hazard function for transition  $r \rightarrow s$  and  $H^{(r\cdot)}(t) = \sum_{s \in \mathcal{B}} H^{(r,s)}(t)$  refer to the event of leaving state  $r$ .
- Sample  $t^*$  from  $H^{(r\cdot)}(t) - H^{(r\cdot)}(t_r)$ . This refers to the conditional distribution of leaving state  $r$  given that the process is known to be in state  $r$  until time  $t_r$ , thus ensuring that the sampled time  $t^* > t_r$ .
- If  $t^* > t_{\max}$ , select the next state  $s$  with probability  $dH^{(r,s)}(t^*)/dH^{(r\cdot)}(t^*)$ , which provides a weight for the specific transition  $r \rightarrow s$  out of state  $r$  for each  $s \in \mathcal{B}$  at the given time  $t^*$ , and set the new starting points for the next iteration,  $r = s$  and  $t_r = t^*$ . Otherwise, stop: a full path through the process was obtained.

This is repeated to obtain  $M$  paths through the multistate model and to compute the transition probabilities by counting the number of paths for which each event occurred. This approach is implemented in the function `mssample()` of the R package `mstate` and is straightforward to use given the estimated transition-specific cumulative hazards for both Markov and semi-Markov models.

## 6 Inference

One view of the smoothing process is that the penalty employed during fitting imposes the belief that the true function is more likely to be smooth than wiggly. This belief can be expressed in a Bayesian manner through the form of a prior distribution on  $\boldsymbol{\beta}^{(r,s)}$ , that is,  $f_{\boldsymbol{\beta}^{(r,s)}} \propto \exp\left\{-\boldsymbol{\beta}^{(r,s)\top} \mathbf{S}_{\lambda^{(r,s)}} \boldsymbol{\beta}^{(r,s)} / 2\right\}$ .

This leads to the Bayesian large sample approximation  $\boldsymbol{\beta}^{(r,s)} \sim \mathcal{N}(\widehat{\boldsymbol{\beta}}^{(r,s)}, \mathbf{V}_{\boldsymbol{\beta}^{(r,s)}})$ , where  $\mathbf{V}_{\boldsymbol{\beta}^{(r,s)}} = -\mathbf{H}_p(\widehat{\boldsymbol{\beta}}^{(r,s)})^{-1}$ ; using  $\mathbf{V}_{\boldsymbol{\beta}^{(r,s)}}$  gives close to across-the-function frequentist coverage probabilities because it accounts for both sampling variability and smoothing bias, a feature that is particularly relevant at finite sample sizes (Wood et al., 2016). Following Pya and Wood (2015), we then consider the Taylor series expansion of  $\mathbf{f}^{(r,s)}(\boldsymbol{\beta}^{(r,s)})$  around  $\mathbf{f}^{(r,s)}(\tilde{\boldsymbol{\beta}}^{(r,s)})$ . This gives  $\mathbf{f}^{(r,s)}(\boldsymbol{\beta}^{(r,s)}) - \mathbf{f}^{(r,s)}(\tilde{\boldsymbol{\beta}}^{(r,s)}) \approx \text{diag}(\mathbf{E}^{(r,s)})(\boldsymbol{\beta}^{(r,s)} - \tilde{\boldsymbol{\beta}}^{(r,s)})$ , where  $\mathbf{E}^{(r,s)}[k_{jk}] = 1$  if  $f_{k_{jk}}^{(r,s)}(\beta_{k_{jk}}^{(r,s)}) = \beta_{k_{jk}}^{(r,s)}$  and  $\exp(\beta_{k_{jk}}^{(r,s)})$  otherwise, showing that  $\mathbf{f}^{(r,s)}(\boldsymbol{\beta}^{(r,s)}) - \mathbf{f}^{(r,s)}(\tilde{\boldsymbol{\beta}}^{(r,s)})$  is approximately a linear function of  $\boldsymbol{\beta}^{(r,s)}$ . Combining this with the result above we have that  $\mathbf{f}^{(r,s)}(\boldsymbol{\beta}^{(r,s)}) \sim \mathcal{N}(\mathbf{f}^{(r,s)}(\tilde{\boldsymbol{\beta}}^{(r,s)}), \mathbf{V}_{\mathbf{f}^{(r,s)}(\boldsymbol{\beta}^{(r,s)})})$  where  $\mathbf{V}_{\mathbf{f}^{(r,s)}(\boldsymbol{\beta}^{(r,s)})} = \text{diag}(\mathbf{E}^{(r,s)})\mathbf{V}_{\boldsymbol{\beta}^{(r,s)}}\text{diag}(\mathbf{E}^{(r,s)})$ , since linear functions of normally distributed random variables follow normal distributions. Confidence intervals for linear functions of the model coefficient

can then be obtained using this result. P-values for the smooth components in the model are derived by adapting the result discussed in Wood (2017) and using  $\mathbf{V}_{\mathbf{f}^{(rs)}(\boldsymbol{\beta}^{(rs)})}$  as covariance matrix. For nonlinear functions of the model coefficients, for example, the transition-specific cumulative hazard functions, instead, the intervals can be conveniently obtained by posterior simulations, hence avoiding computationally expensive parametric bootstrap or frequentist approximations, for instance.

## 7 Primary breast cancer modelling case study

To illustrate what the proposed approach adds compared to the existing literature, we consider the case study described in Crowther and Lambert (2017) which is based on data from 2892 patients with primary breast cancer for which the time to relapse and/or the time to death is known. See, for example, Sauerbrei et al. (2007) for further details on the *Rotterdam Breast Cancer Study* from which the data originated. The code used to produce this analysis can be found in the public repository <https://github.com/AlessiaEletti/ContinObsMultistateProcesses>. All patients begin in the initial post-surgery state, 1518 patients experience relapse, 195 die without relapse and 1075 die after experiencing relapse. A Markov illness-death model (IDM, see Figure 5 in Supplementary Material C) will thus be used to model the data. As an aside, note that an attempt assuming semi-Markovianity was also made but this was not supported by the data according to the AIC values found for the fitted models. As there are three transitions in the assumed IDM, three survival models will be fitted. For transitions which can occur only given that another transition has already taken place, that is, the transition  $2 \rightarrow 3$  in this case, one must account for the fact that the patient is at risk only after entering the new starting state, that is, state 2. As long as this is done, each transition can be treated as a separate survival problem. The time at which the individual entered state 2 thus becomes the left-truncation time for the new transition  $2 \rightarrow 3$ . To clarify how the separate estimations are carried out, recall that longitudinal survival data are characterised by multiple observations through time of at least one quantity of interest for the same individual. Typically the data are formatted in the so-called stacked (or long) form, that is, each row represents a single time point per subject. In particular, each subject will have at least  $v$  rows, where  $v$  is the number of possible transitions exiting the initial state. Here,  $v = 2$  as there are two ways of exiting state 1, that is, going in state 2 or 3. A start and a stop time will then indicate, respectively, the first time after which the patient becomes at risk of the given transition and the time at which the transition itself occurred. The start time for transitions exiting the first state is 0, as is usually the case here. If the patient transitions to an intermediate state,  $u$  rows will be added, where  $u$  is the number of transitions exiting the intermediate transition state reached. Here,  $u = 1$ , as the only possible transition out of state 2 is  $2 \rightarrow 3$ , where 3 is an absorbing state. When estimating  $q^{(12)}(\cdot)$ , all of the rows relating to this transition are included in the estimation. Since every patient will at least have one row for each transition exiting the first state, this implies that the entire population is included. The same is true for  $q^{(13)}(\cdot)$ , for which the rows relating to the  $1 \rightarrow 3$  transition will be used for estimation. The two resulting separate datasets can then be treated as traditional survival data with uncensored and right-censored observations and with the event of interest given by the transition to the new state, that is, state 2 for the former and state 3 for the latter. When estimating  $q^{(23)}(\cdot)$ , only individuals who have transitioned to state 2



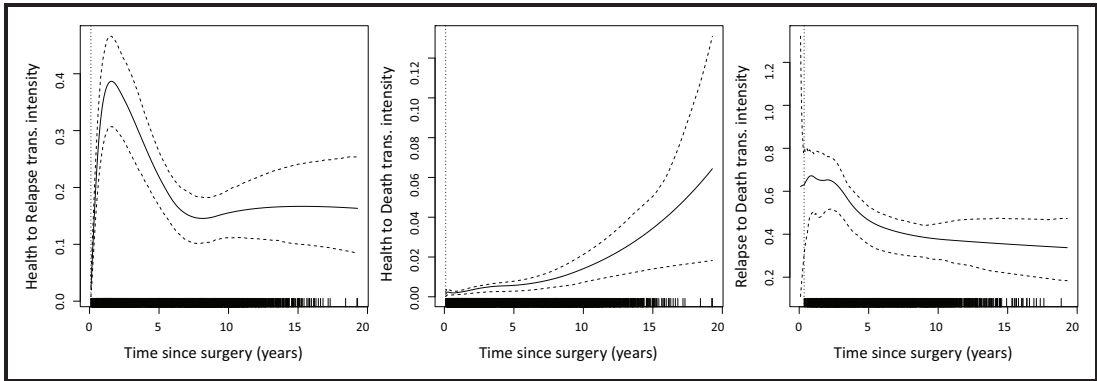
at some point are included in the estimation. The data are then treated as traditional survival data with left-truncated uncensored and left-truncated right-censored observations and where the event of interest is the transition to the absorbing state 3. We refer the reader to Supplementary Material D for further details on the format of the data in this setting.

The dataset contains information on the age of the patient at primary surgery (in years), tumour size (divided into 3 classes:  $\leq 20$ ,  $20 - 50$  and  $> 50$  mm), number of positive nodes, progesterone levels (in fmol/L) and whether or not the patient was on hormonal therapy. These are all included as covariates. We then include a time-dependent effect for the progesterone level, as this has been found to be relevant in the reference paper, and include age, the progesterone level and the number of positive nodes nonlinearly, as supported by existing literature. Importantly, our chosen framework allows for the exploration of these effects in a more general and flexible manner than previously possible in the literature thanks to the use of splines. In contrast, for instance, Sauerbrei and Royston (1999) modelled the number of positive nodes nonlinearly by using fractional polynomials with the degrees set heuristically. Similarly, in Crowther and Lambert (2017) the time-dependant effect is captured by a single interaction coefficient between time and the progesterone level. In particular, for  $(r, s) \in \{(1, 2), (1, 3), (2, 3)\}$ , we specify the transition-specific models

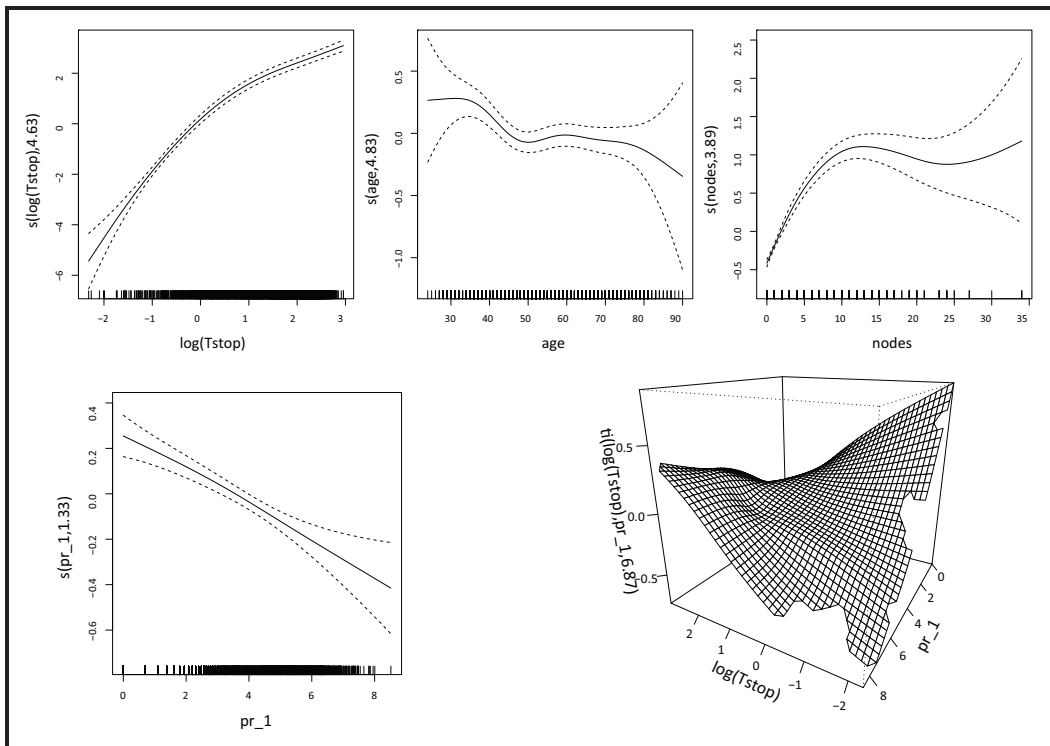
$$\begin{aligned} \eta_i^{(r,s)}(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}^{(r,s)})) &= \beta_0^{(r,s)} + s_0^{(r,s)}(\log(t_i)) + \beta_1^{(r,s)}\mathbf{I}_{\text{size}_i=20-50} + \beta_2^{(r,s)}\mathbf{I}_{\text{size}_i>50} + \beta_3^{(r,s)}\text{hormon}_i \\ &+ s_1^{(r,s)}(\text{age}_i) + s_2^{(r,s)}(\text{nodes}_i) + s_3^{(r,s)}(\text{pr}_i) + s_4^{(r,s)}(\log(t_i), \text{pr}_i), \end{aligned}$$

where  $s_0^{(r,s)}(\log(t_i))$  is a monotonic P-spline of the logarithm of time which ensures the monotonicity of the survival function associated with this transition, as explained in Section 3.2;  $s_1^{(r,s)}(\text{age}_i)$ ,  $s_2^{(r,s)}(\text{nodes}_i)$  and  $s_3^{(r,s)}(\text{pr}_i)$  are thin-plate splines, while  $s_4^{(r,s)}(\log(t_i), \text{pr}_i)$  is a pure smooth interaction between time and the progesterone level, that is, a time-dependent effect. In regard to the penalty associated with a nonlinear term, for example,  $s_1(\text{age}_i)$ , this takes the form of the quadratic penalty defined above with  $\mathbf{D}_k$  given by the integrated square second derivative of the basis functions, that is,  $\int \mathbf{d}_k(z_k)\mathbf{d}_k(z_k)^\top dz_k$  with the  $j_k^{\text{th}}$  element of  $\mathbf{d}_k(z_k)$  defined as  $\partial^2 b_{kj_k}(z_k)/\partial z_k^2$ . The penalty associated with the time-dependent effect is, instead, more complex as it entails combining two penalties (see Wood, 2017, Chapter 5). Finally, note that for parametric effects the spline representation simplifies to  $s^{(r,s)}(\text{hormon}_i) = \beta_3^{(r,s)}\text{hormon}_i$ . No penalty is typically assigned to parametric effects, hence the associated quadratic penalty is  $D = 0$ . Note that in cases such as those in which the categorical variable has many levels with some with few observations, it may be advisable to set the penalty as the identity matrix. In this way, a ridge penalty is imposed and it may help avoid that the parameters associated with the more sparse categories are weakly or nonidentified.

The estimated covariate effects for each transition are reported in Table 2. For the first transition, for instance, they are all significant and in line with our expectations: the larger the size of the tumor the higher the risk of experiencing relapse, while hormonal therapy has a beneficial effect. In Figure 1 we report the estimated transition intensities with their 95% confidence intervals as functions of time for a 54 year old patient with tumour size  $\geq 50$  mm, 10 positive nodes, progesterone level of 3 and under hormonal therapy. We find, for instance, that the risk of experiencing relapse for this profile increases for approximately 2.5 years after surgery, then it decreases and plateaus over time. In Figure 2 we report the plots of the smooths and of the tensor interaction for the transition



**Figure 1** Fitted transition intensities and 95% confidence intervals (CIs) for a 54 year old patient under hormonal therapy with tumour size  $\geq 50$  mm, 10 nodes and progesterone level of 3, over 20 years. The vertical dashed line marks the smallest observed time: the transition intensities estimated at smaller times are extrapolations, thus explaining the wide CIs in the first section of the third plot. The width of the CIs in the final portion of the middle plot can be explained by the scarcity of observations in the final times, as shown by the rug plot. The width of the confidence intervals should also be related to the different range of values in each plot.



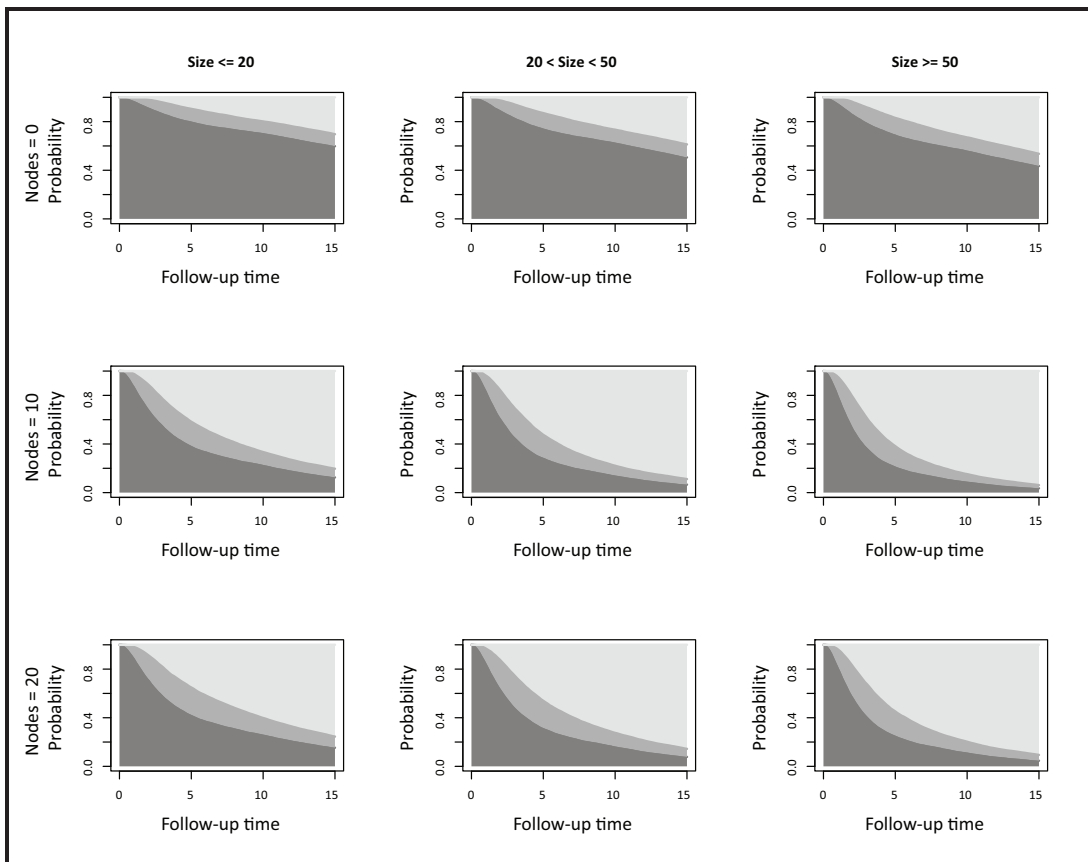
**Figure 2** Smooth of log-time (top left), smooth of age (top middle), smooth of the number of positive nodes (top right), smooth of the progesterone level (bottom left) and smooth interaction between log-time and progesterone level (bottom right) for the transition *health*  $\rightarrow$  *relapse*.

**Table 2** Model estimates, standard errors and p-values for the three transitions.

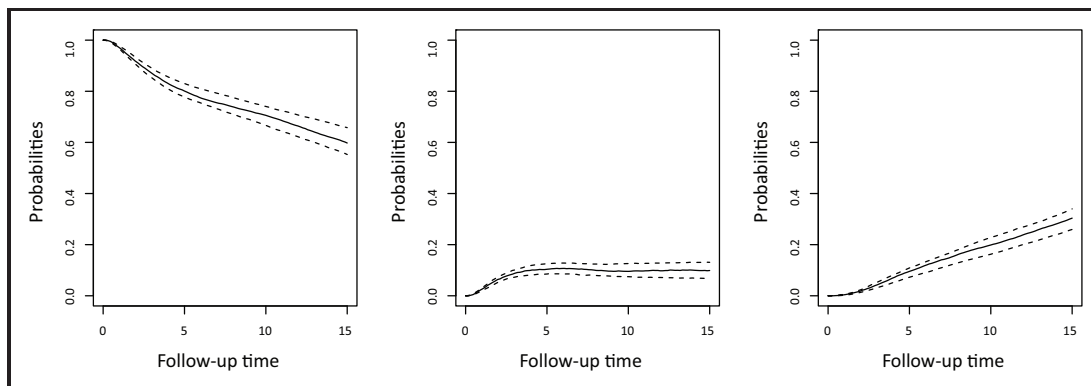
		Estimate	Std. Error	Pr(>  z )
Transition 1 → 2	(Intercept)	−10.630	1.198	< 1e−4
	size20–50	0.284	0.059	< 1e−4
	size>50	0.477	0.089	< 1e−4
	hormon	−0.318	0.085	2e−4
Transition 1 → 3	(Intercept)	−12.543	2.585	< 1e−4
	size20–50	0.153	0.162	0.344
	size>50	0.390	0.236	0.098
	hormon	−0.135	0.236	0.567
Transition 2 → 3	(Intercept)	−2.915	1.023	0.004
	size20–50	0.139	0.072	0.053
	size>50	0.259	0.101	0.010
	hormon	−0.015	0.098	0.881

*health* → *relapse*. These show that the data particularly support nonlinear effects for the age and the number of positive nodes. For instance, the latter exhibits an increasing trend up to about 12 nodes, followed by a plateau. The time-dependence of the progesterone level effect is also clear from the surface representing the smooth interaction, with low levels of progesterone associated with a decreasing risk of experiencing relapse over time and, conversely, high levels of progesterone associated with an increasing trend for the risk of experiencing relapse over time. Any additional complexity not supported by the data is then suppressed automatically through the estimation of the smoothing parameter, rather than requiring the user to make restrictive and potentially arbitrary choices a priori. This can be seen in the plots of the smooths of the remaining two transitions, reported in Figures 6 and 7 of Supplementary Material C. The plot of the smooth of age for the *health* → *death* transition, for instance, shows that the data actually supported a linear effect for this term.

As mentioned above, interest usually lies in making statements in terms of the probabilities of transitioning between states thus, in Figure 3, we report stacked transition probability plots. Representing the probabilities in this stacked manner is a common way of quickly providing an overview of how risk evolves over time, however the uncertainty of the estimates cannot be easily portrayed. For this reason, in Figure 4, we report the predicted probabilities with their 95% confidence intervals for the individual corresponding to the top-left panel, that is, a 54 year old patient under hormonal therapy, progesterone level of 3, 20 positive nodes and tumour size ≤ 20 mm. Note that the computation of the transition probabilities already entails a simulation, thus the process of obtaining confidence intervals for it will result in two nested simulations. The computational burden of this is not prohibitively high, however. Here, they are obtained by using 100 simulated cumulative hazards for each of the three transitions, over 100 distinct time points, and  $M = 10\,000$  simulated paths through the process, which is a larger number of paths than typically needed. This required approximately 37 minutes using a laptop with Windows 10 (2.20 GHz processor, 16 GB RAM, 64-bit). Details on this, on how the model fitting is carried out and how the plots reported in this section were obtained can be found in Supplementary Material C.



**Figure 3** Stacked representation of estimated transition probabilities (dark grey: post-surgery; grey: relapse; light grey: death) for each combination of nodes (0, 10 and 20) and tumour sizes ( $\leq 20$ ,  $(20, 50)$  and  $\geq 50$ ) considered in a 54 year old patient under hormonal therapy with progesterone level of 3.



**Figure 4** Estimated transition probabilities (left: post-surgery; middle: relapse; right: death) for the top-left pane in Figure 3.

## 8 Discussion

In this work we show how one can use existing tools to flexibly model multistate survival processes relating to continuously observed life-history data. In particular, we consider the survival estimation framework described in Eletti et al. (2022) and implemented in the R package GJRM which allows us to model virtually any type of covariate effect, including time-dependent ones. Direct modelling of the survival functions implies a considerable gain in efficiency when it comes to computing the transition probabilities of interest, which in turn are obtained through a simulation-based approach able to support any type of multistate process. Efficient modelling on the survival scale is achieved through shape constrained P-splines, developed by Pya and Wood (2015), building upon the work done in Eilers and Marx (1996). We exemplify our approach on data from the *Rotterdam Breast Cancer Study* and provide the code used for the analysis in the public repository <https://github.com/AlessiaEletti/ContinObsMultistateProcesses>.

With regard to directions of future work, we are interested in integrating the computation of the transition probabilities and the extraction of its confidence intervals directly within the GJRM package, so as to minimise the amount of user-written code needed and thus further simplify the use of these models by the practitioner. Similarly, for the visualisation tools available for the estimated transition probabilities. As the Markov assumption is quite common, we are also interested in implementing the method based on the numerical solution of the differential equations tying the transition probabilities to the intensities as well as to implement our own simulation-based approach within the GJRM package, so that the user has all necessary instruments in the same place and the need for user-written code is reduced to the minimum.

## Acknowledgements

In October 2009, the last two authors attended, as PhD students, the short course ‘Splines, Knots, and Penalties: The Craft of Smoothing’, in Galway delivered by Paul H. C. Eilers and Brian D. Marx. Although at that time they had just started learning about splines, the words used by Brian D. Marx to explain P-splines inspired them and always echoed and tormented their minds until they could appreciate their simplicity and versatility in the context of survival analysis.

## Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

## Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: AE was supported by the UCL Departmental Teaching Assistantship Scholarship. GM and RR were supported by the EPSRC grant EP/T033061/1.

## Supplementary materials

Supplementary materials for this article are available online.

## References

- Clements M, Liu XR and Christoffersen B (2021) rstpm2: Smooth survival models, including generalized survival models. URL <https://cran.r-project.org/package=rstpm2>. R package version 1.5.1.
- Crowther MJ and Lambert P (2016) MULTISTATE: Stata module to perform multistate survival analysis. Statistical Software Components, Boston College Department of Economics. URL <https://ideas.repec.org/c/boc/bocode/s458207.html>.
- Crowther MJ and Lambert PC (2017) Parametric multistate survival models: flexible modelling allowing transition-specific distributions with application to estimating clinically useful measures of effect differences. *Statistics in Medicine*, **36**, 4719–4742.
- DeWreede LC, Fiocco M and Putter H (2010) The mstate package for estimation and prediction in non-and semi-parametric multi-state and competing risks models. *Computer Methods and Programs in Biomedicine*, **99**, 261–274.
- Eilers PH and Marx BD (1996) Flexible smoothing with b-splines and penalties. *Statistical Science*, **11**, 89–121.
- Eletti A, Marra G, Quaresma M, Radice R and Rubio FJ (2022) A unifying framework for flexible excess hazard modelling with applications in cancer epidemiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. URL <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssc.12566>.
- Fauvernier M, Roche L and Remontet L (2020) survPen: Multidimensional penalized splines for survival and net survival models. URL <https://cran.r-project.org/package=survPen>. R package version 1.5.1.
- Fiocco M, Putter H and van Houwelingen HC (2008) Reduced-rank proportional hazards regression and simulation-based prediction for multi-state models. *Statistics in Medicine*, **27**, 4340–4358.
- Iacobelli S and Carstensen B (2013) Multiple time scales in multi-state models. *Statistics in Medicine*, **32**, 5315–5327.
- Jackson C (2021) flexsurv: Flexible parametric survival and multi-state models. URL <https://cran.r-project.org/package=flexsurv>. R package version 2.0.
- Leitenstorfer F and Tutz G (2007) Generalized monotonic regression based on B-splines with an application to air pollution data. *Biostatistics*, **8**, 654–673.
- Liu XR, Pawitan Y and Clements M (2018) Parametric and penalized generalized survival models. *Statistical Methods in Medical Research*, **27**, 1531–1546.
- Marra G and Radice R (2020) Copula link-based additive models for right-censored event time data. *Journal of the American Statistical Association*, **115**, 886–895.
- Marra G and Radice R (2023) GJRM: Generalised joint regression modelling. URL <https://CRAN.R-project.org/package=GJRM>. R package version 0.2-6.4
- Putter H (2011) Tutorial in biostatistics: Competing risks and multi-state models analyses using the mstate package. Companion file for the mstate package.
- Putter H, Fiocco M and Geskus RB (2007) Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, **26**, 2389–2430.
- Putter H, de Wreede LC and Fiocco M (2020) mstate: Data preparation, estimation and prediction in multi-state models. URL <https://cran.r-project.org/package=mstate>. R package version 0.3.1.
- Pya N and Wood S (2015) Shape constrained additive models. *Statistics and Computing*, **25**, 543–559.

- Pyke R (1961) Markov renewal processes: Definitions and preliminary properties. *The Annals of Mathematical Statistics*, **32**, 1231–1242.
- Ross SM, Kelly JJ, Sullivan RJ, Perry WJ, Mercer D, Davis RM, Washburn TD, Sager EV, Boyce JB and Bristow VL (1996) *Stochastic processes*, volume 2. Wiley New York.
- Royston P and Parmar MK (2002) Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, **21**, 2175–2197.
- Sauerbrei W and Royston P (1999) Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **162**, 71–94.
- Sauerbrei W, Royston P and Look M (2007) A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time transformation. *Biometrical Journal*, **49**, 453–473.
- Titman AC (2011) Flexible nonhomogeneous Markov models for panel observed data. *Biometrics*, **67**, 780–787.
- Touraine C, Helmer C and Joly P (2016) Predictions in an illness-death model. *Statistical Methods in Medical Research*, **25**, 1452–1470.
- Wood SN (2017) *Generalized additive models: An introduction with R*, 2nd ed. Chapman & Hall/CRC, London.
- Wood SN, Pya N and Säfken B (2016) Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, **111**, 1548–1563.
- Yang Y and Nair VN (2011) Parametric inference for time-to-failure in multi-state semimarkov models: A comparison of marginal and process approaches. *Canadian Journal of Statistics*, **39**, 537–555.
- Younes N and Lachin J (1997) Link-based models for survival data with interval and continuous time censoring. *Biometrics*, **53**, 1199–1211.