City Research Online

City, University of London Institutional Repository

Citation: Eletti, A., Marra, G. & Radice, R. (2023). A Spline-Based Framework for the Flexible Modelling of Continuously Observed Multistate Survival Processes. Statistical Modelling: An International Journal,

This is the accepted version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/30316/

Link to published version:

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

A Spline-Based Framework for the Flexible Modelling of Continuously Observed Multistate Survival Processes

Alessia Eletti [1], Giampiero Marra [1] and Rosalba Radice [2]

[1] Department of Statistical Science, University College London, Gower Street, WC1E 6BT London, UK.

[2] Faculty of Actuarial Science and Insurance, Bayes Business School, City, University of London, 106 Bunhill Row, EC1Y 8TZ London, UK.

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`.

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.

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, i.e. 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; 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, 2022). 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 paper 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 $\mathcal{S} := \{z : Z(t) = z, t \in \mathcal{T}\} \subseteq \{0, 1, 2, ...\}$ 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

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

with $q^{(rs)}(t \mid \mathcal{F}_{t^-}) = 0$ if r is an absorbing state and $q^{(rr)}(t \mid \mathcal{F}_{t^-}) = -\sum_{s \neq r} q^{(rs)}(t \mid \mathcal{F}_{t^-})$. The matrix with (r, s) element given by $q^{(rs)}(t \mid \mathcal{F}_{t^-})$ for every $r, s \in \mathcal{S}$ is called transition intensity matrix or generator matrix and we will denote it by $\mathbf{Q}(t \mid \mathcal{F}_{t^-})$. Similarly, we define the transition probability matrix associated with the time interval $[u, t]$ as the matrix with (r, s) element given by $P(Z(t) = s \mid Z(u) = r, \mathcal{F}_{u^-})$ and denote this by $\mathbf{P}(u, t \mid \mathcal{F}_{u^-})$. 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 viewed as a standalone time-to-event and can thus be modelled through traditional survival analysis. It is well know 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 $\mathcal{A} = \{(r,s) \mid r \neq s \in \mathcal{S}, q^{(rs)}(t_i) \neq 0\}$ be the set of transitions and N represent the sample size. For individual $i = 1, \ldots, N$ 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 $H^{(rs)}(t_i \mid \mathbf{x}_i; \boldsymbol{\beta}^{(rs)}) = \int_0^{t_i} q^{(rs)}(u \mid \mathbf{x}_i; \boldsymbol{\beta}^{(rs)}) du$. Then we will have a conditional survival function denoted by $S^{(rs)}(t_i \mid \mathbf{x}_i; \boldsymbol{\beta}^{(rs)}) = \exp\left\{-H^{(rs)}(t_i \mid \mathbf{x}_i; \boldsymbol{\beta}^{(rs)})\right\} \in (0,1)$, where $\mathbf{x}_i$ represents a generic vector of patient characteristics that has an associated regression coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^w$, where w is the length of $\boldsymbol{\beta}^{(rs)}$. A link-based additive transition-specific survival model can then be written as

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

where $g : (0,1) \to \mathbb{R}$ is a monotone and twice continuously differentiable link function with bounded derivatives, hence invertible, which determines the scale of the analysis, $\eta_i^{(rs)}(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}^{(rs)})) \in \mathbb{R}$ is an additive predictor which includes a baseline function of time and several types of covariate effects and $\mathbf{f}(\boldsymbol{\beta}^{(rs)})$ is a vector function of $\boldsymbol{\beta}^{(rs)}$ through which the monotonicity required for the survival functions is imposed (see Section 3.2). Rearranging (3.1) yields $S^{(rs)}(t_i \mid \mathbf{x}_i; \boldsymbol{\beta}^{(rs)}) = G\left\{\eta_i^{(rs)}(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}^{(rs)}))\right\}$, 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 $H^{(rs)}(t_i \mid \mathbf{x}_i; \boldsymbol{\beta}^{(rs)}) = -\log\left[G\left\{\eta_i^{(rs)}(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}^{(rs)}))\right\}\right]$ and the transition intensity function is defined as

$$q^{(rs)}(t_i \mid \mathbf{x}_i; \boldsymbol{\beta}^{(rs)}) = -\frac{G'\left\{\eta_i^{(rs)}(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}^{(rs)}))\right\}}{G\left\{\eta_i^{(rs)}(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}^{(rs)}))\right\}} \frac{\partial \eta_i^{(rs)}(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}^{(rs)}))}{\partial t_i}, \tag{3.2}$$

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

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

Table 1: Functions implemented in `GJRM`. Φ and ϕ 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; e.g. `margin = "PH"`.

3.1 Additive predictor

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

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

where $\beta_0^{(rs)} \in \mathbb{R}$ is an overall intercept, $\mathbf{z}_{ki}$ denotes the k^{th} sub-vector of the complete vector $\mathbf{z}_i$ and the $K^{(rs)}$ functions $s_k^{(rs)}(\mathbf{z}_{ki})$ 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 $\mathbf{b}_k(\mathbf{z}_{ki}) = (b_{k1}^{(rs)}(\mathbf{z}_{ki}), \ldots, b_{kJ_k}^{(rs)}(\mathbf{z}_{ki}))^\intercal$ and regression coefficients $\mathbf{f}_k^{(rs)}(\boldsymbol{\beta}_k^{(rs)}) = (f_{k1}^{(rs)}(\beta_{k1}^{(rs)}), \ldots, f_{kJ_k}^{(rs)}(\beta_{kJ_k}^{(rs)}))^\intercal \in \mathbb{R}^{J_k}$, that is $s_k^{(rs)}(\mathbf{z}_{ki}) =$

$\mathbf{b}_k(\mathbf{z}_{ki})^\mathsf{T}\mathbf{f}_k^{(rs)}(\boldsymbol{\beta}_k^{(rs)})$ (e.g., Wood, 2017). We can then write (3.3) compactly as $\eta_i^{(rs)} = \mathbf{Z}_i^{(rs)\mathsf{T}}\mathbf{f}^{(rs)}(\boldsymbol{\beta}^{(rs)})$, where $\mathbf{Z}_i^{(rs)} = (1, \mathbf{b}_1(\mathbf{z}_{1i})^\mathsf{T}, \ldots, \mathbf{b}_{K^{(rs)}}(\mathbf{z}_{K^{(rs)}i})^\mathsf{T})^\mathsf{T}$ and $\mathbf{f}^{(rs)}(\boldsymbol{\beta}^{(rs)}) = (\beta_0^{(rs)}, \mathbf{f}_1^{(rs)}(\boldsymbol{\beta}_1^{(rs)})^\mathsf{T}, \ldots, \mathbf{f}_{K^{(rs)}}^{(rs)}(\boldsymbol{\beta}_{K^{(rs)}}^{(rs)})^\mathsf{T})^\mathsf{T}$. Observe that $\partial\eta_i^{(rs)}(t_i, \mathbf{x}_i; \mathbf{f}^{(rs)}(\boldsymbol{\beta}^{(rs)}))/\partial t_i$ is required in (3.2). This can be expressed as $\mathbf{Z}_i^{(rs)}(t_i, \mathbf{x}_i)'^\mathsf{T}\mathbf{f}^{(rs)}(\boldsymbol{\beta}^{(rs)})$ where, depending on the type of spline basis employed, $\mathbf{Z}_i(t_i, \mathbf{x}_i)' = \lim_{\varepsilon\to 0}\frac{\mathbf{Z}_i^{(rs)}(t_i+\varepsilon,\mathbf{x}_i)-\mathbf{Z}_i^{(rs)}(t_i-\varepsilon,\mathbf{x}_i)}{2\varepsilon}$ can be calculated either by a finite-difference method or analytically. Each $\boldsymbol{\beta}_k^{(rs)}$ has an associated quadratic penalty $\lambda_k^{(rs)}\boldsymbol{\beta}_k^{(rs)\mathsf{T}}\mathbf{D}_k^{(rs)}\boldsymbol{\beta}_k^{(rs)}$, used in fitting, whose role is to enforce specific properties on the k^{th} function, such as smoothness, with matrix $\mathbf{D}_k^{(rs)}$ depending only on the choice of the basis functions. The smoothing parameter $\lambda_k^{(rs)} \in [0, \infty)$ 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 $\boldsymbol{\beta}^{(rs)\mathsf{T}}\mathbf{S}_{\boldsymbol{\lambda}^{(rs)}}^{(rs)}\boldsymbol{\beta}^{(rs)}$, where $\mathbf{S}_{\boldsymbol{\lambda}^{(rs)}}^{(rs)} = \text{diag}(0, \lambda_1^{(rs)}\mathbf{D}_1^{(rs)}, \ldots, \lambda_{K^{(rs)}}^{(rs)}\mathbf{D}_{K^{(rs)}}^{(rs)})$ is a block diagonal matrix where each block is given by the k^{th} penalty, and where $\boldsymbol{\lambda}^{(rs)} = (\lambda_1^{(rs)}, \ldots, \lambda_{K^{(rs)}}^{(rs)})^\mathsf{T}$ 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, e.g. 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 $s^{(rs)}(t_i) = \sum_{j=1}^{J^{(rs)}} f_j^{(rs)}(\beta_j^{(rs)}) b_j^{(rs)}(t_i)$, where the $b_j^{(rs)}(\cdot)$ are B-spline basis functions of at least second order built over the interval $[a, b]$, based on equally spaced knots, and the $f_j^{(rs)}(\beta_j)^{(rs)}$ are spline coefficients. Given the link functions listed in Table 1, we need $s^{(rs)'}(t_i) \geq 0$. 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 $s^{(rs)'}(t_i) \geq 0$ over $[a, b]$ is that $f_j^{(rs)}(\beta_j^{(rs)}) \geq f_{j-1}^{(rs)}(\beta_{j-1}^{(rs)}), \forall j$. 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 $\partial\eta(x)/\partial x \leq 0$ since $B_j(x, q-1) \geq 0$ (Leitenstorfer and Tutz, 2007). Such condition can be imposed

by defining the vector function $\mathbf{f}^{(rs)}(\boldsymbol{\beta}^{(rs)}) = \boldsymbol{\Sigma}\left\{\beta_1^{(rs)}, \exp(\beta_2^{(rs)}), \ldots, \exp(\beta_{J^{(rs)}}^{(rs)})\right\}^{\mathsf{T}}$, 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 ι_1 and ι_2 denoting the row and column entries of $\boldsymbol{\Sigma}$, and $\boldsymbol{\beta}^{(rs)\mathsf{T}} = (\beta_1^{(rs)}, \beta_2^{(rs)}, \ldots, \beta_{J^{(rs)}}^{(rs)})$ 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 "wiggliness". 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 $\beta_j^{(rs)}$, starting from $\beta_2^{(rs)}$, using $\mathbf{D}^{(rs)} = \mathbf{D}^{(rs)*\mathsf{T}}\mathbf{D}^{*}$ where $\mathbf{D}^{(rs)*}$ is a $(J^{(rs)}-2)\times J^{(rs)}$ matrix made up of zeros except that $\mathbf{D}^{(rs)*}[\iota, \iota+1] = -\mathbf{D}^{(rs)*}[\iota, \iota+2] = 1$ for $\iota = 1, \ldots, J^{(rs)} - 2$. The penalty is zeroes when all the $\beta_j^{(rs)}$ after $\beta_1^{(rs)}$ are equal so that the $f_j^{(rs)}(\beta_j^{(rs)})$ form a uniformly increasing sequence and $s^{(rs)}(t_i)$ 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, i.e. $\ell(\boldsymbol{\theta}) = \sum_{(r,s)\in\mathcal{A}} \ell^{(rs)}(\boldsymbol{\beta}^{(rs)})$, where $\boldsymbol{\theta} = \{\boldsymbol{\beta}^{(rs)} \mid (r,s) \in \mathcal{A}\}$ is an overall model parameter vector. Re-writing 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 $\ell_p^{(rs)}(\boldsymbol{\beta}^{(rs)}) = \ell^{(rs)}(\boldsymbol{\beta}^{(rs)}) - \frac{1}{2}\boldsymbol{\beta}^{(rs)\mathsf{T}}\mathbf{S}^{(rs)}_{\boldsymbol{\lambda}^{(rs)}}\boldsymbol{\beta}^{(rs)}$ where $\mathbf{S}^{(rs)}_{\boldsymbol{\lambda}^{(rs)}}$ 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 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 $H^{(rs)}(t)$ be the cumulative transition-specific hazard function for transition $r \to s$ and $H^{(r\cdot)}(t) = \sum_{s \in \mathcal{B}} H^{(rs)}(t)$ 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 $dH^{(rs)}(t^*)/dH^{(r\cdot)}(t^*)$, which provides a weight for the specific transition $r \to s$ 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 $\boldsymbol{\beta}^{(rs)}$, i.e. $f_{\boldsymbol{\beta}^{(rs)}} \propto \exp\left\{-\boldsymbol{\beta}^{(rs)\mathsf{T}} \mathbf{S}^{(rs)}_{\boldsymbol{\lambda}^{(rs)}} \boldsymbol{\beta}^{(rs)}/2\right\}$. This leads to the Bayesian

large sample approximation $\boldsymbol{\beta}^{(rs)} \dot{\sim} \mathcal{N}(\widehat{\boldsymbol{\beta}}^{(rs)}, \mathbf{V}_{\boldsymbol{\beta}^{(rs)}})$, where $\mathbf{V}_{\boldsymbol{\beta}^{(rs)}} = -\mathbf{H}_p(\widehat{\boldsymbol{\beta}}^{(rs)})^{-1}$; using $\mathbf{V}_{\boldsymbol{\beta}^{(rs)}}$ 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 $\mathbf{f}^{(rs)}(\boldsymbol{\beta}^{(rs)})$ around $\mathbf{f}^{(rs)}(\tilde{\boldsymbol{\beta}}^{(rs)})$. This gives $\mathbf{f}^{(rs)}(\boldsymbol{\beta}^{(rs)}) - \mathbf{f}^{(rs)}(\tilde{\boldsymbol{\beta}}^{(rs)}) \approx \text{diag}(\mathbf{E}^{(rs)})\big(\boldsymbol{\beta}^{(rs)} - \tilde{\boldsymbol{\beta}}^{(rs)}\big)$, where $\mathbf{E}^{(rs)}[k_{j_k}] = 1$ if $f^{(rs)}_{k_{j_k}}(\beta^{(rs)}_{k_{j_k}}) = \beta_{k_{j_k}}$ and $\exp(\beta^{(rs)}_{k_{j_k}})$ otherwise, showing that $\mathbf{f}^{(rs)}(\boldsymbol{\beta}^{(rs)}) - \mathbf{f}^{(rs)}(\tilde{\boldsymbol{\beta}}^{(rs)})$ is approximately a linear function of $\boldsymbol{\beta}^{(rs)}$. Combining this with the result above we have that $\mathbf{f}^{(rs)}(\boldsymbol{\beta}^{(rs)}) \dot{\sim} \mathcal{N}(\mathbf{f}^{(rs)}(\tilde{\boldsymbol{\beta}}^{(rs)}), \mathbf{V}_{\mathbf{f}^{(rs)}(\boldsymbol{\beta}^{(rs)})})$ where $\mathbf{V}_{\mathbf{f}^{(rs)}(\boldsymbol{\beta}^{(rs)})} = \text{diag}(\mathbf{E}^{(rs)})\mathbf{V}_{\boldsymbol{\beta}^{(rs)}}\,\text{diag}(\mathbf{E}^{(rs)})$, 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, e.g. 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, e.g., Sauerbrei et al. (2007) for further details on the *Rotterdam Breast Cancer Study* from which the data origi-

nated. 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, i.e. the transition $2 \to 3$ in this case, one must account for the fact that the patient is at risk only after entering the new starting state, i.e. 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 $2 \to 3$. 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, i.e. 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, i.e. 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 $2 \to 3$, 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 $1 \to 3$ 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, i.e. 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: ≤ 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

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

where $s_0^{(rs)}(\log(\mathrm{t}_i))$ 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; $s_1^{(rs)}(\text{age}_i)$, $s_2^{(rs)}(\text{nodes}_i)$ and $s_3^{(rs)}(\text{pr}_i)$ are thin-plate splines, while $s_4^{(rs)}(\log(\text{t}_i), \text{pr}_i)$ is a pure smooth interaction between time and the progesterone level, i.e. a time-dependent effect. In regard to the penalty associated with a nonlinear term, e.g., $s_1(\text{age}_i)$, 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, i.e. $\int \mathbf{d}_k(z_k)\mathbf{d}_k(z_k)^\mathsf{T} dz_k$ with the j_k^{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 $s^{(rs)}(\text{hormon}_i) = \beta_3^{(rs)}\text{hormon}_i$. 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 $health \rightarrow relapse$. 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

		Estimate	Std. Error	Pr($>\|z\|$)
Transition $1 \to 2$	(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 $1 \to 3$	(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 $2 \to 3$	(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

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

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* $\to$ *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

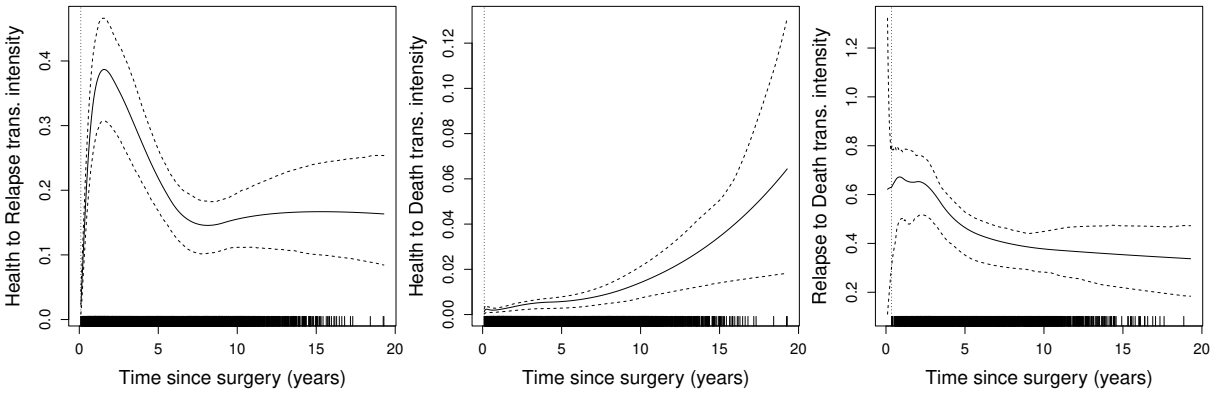

Figure 1: Fitted transition intensities and 95% confidence intervals (CIs) for a 54 year old patient under hormonal therapy with tumour size $\geq$ 50 mm, 10 nodes and progesterone level of 3, over 20 years. The vertical dashed line marks the smallest observed time: the transition intensities estimated 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.

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, i.e. a 54 year old patient under hormonal therapy, progesterone level of 3, 20 positive nodes and tumour size $\leq$ 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

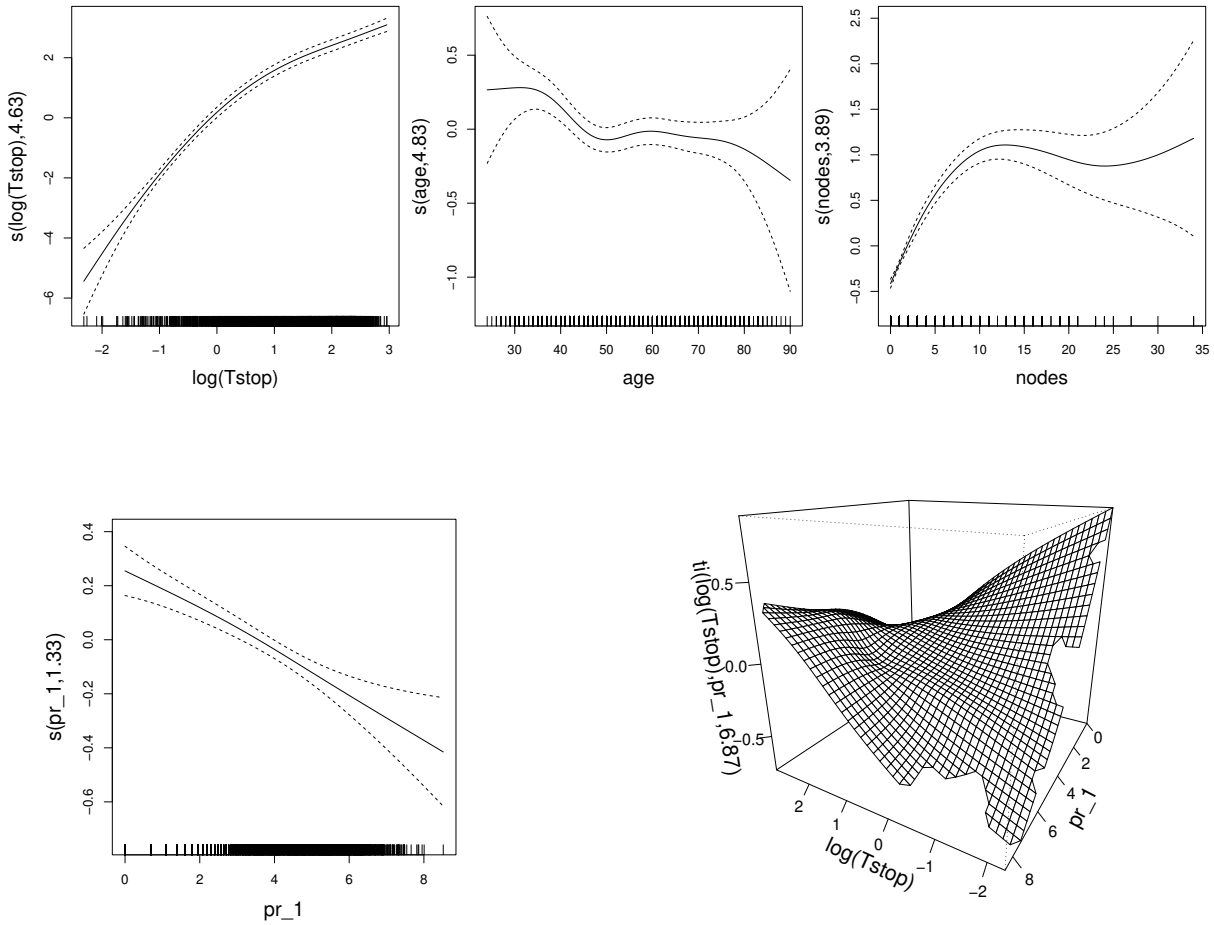

Figure 2: Smooth of log-time (top left), smooth of age (top middle), smooth of the number of positive nodes (top right), smooth of the progesterone level (bottom left) and smooth interaction between log-time and progesterone level (bottom right) for the transition $health \rightarrow relapse$.

of the three transitions, over 100 distinct time points, and $M = 10000$ 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.

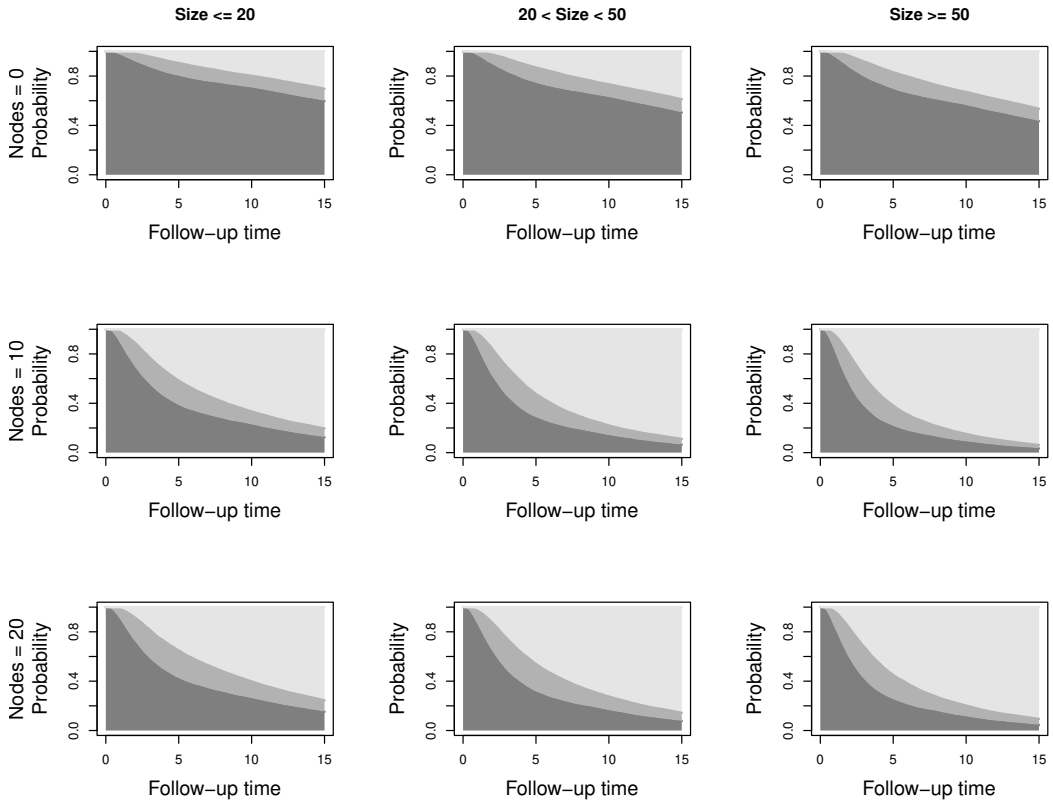

Figure 3: Stacked representation of estimated transition probabilities (dark grey: post-surgery; grey: relapse; light grey: death) for each combination of nodes (0, 10 and 20) and tumour sizes (≤ 20, $(20, 50)$ and ≥ 50) considered in a 54 year old patient under hormonal therapy with progesterone level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

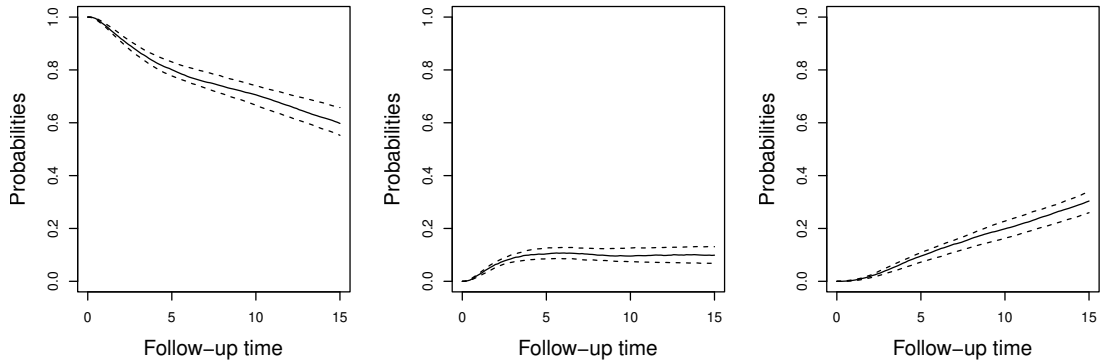

Figure 4: Estimated transition probabilities (left: post-surgery; middle: relapse; right: death) for the top-left pane in Figure 3.

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.

AE was supported by the UCL Departmental Teaching Assistantship Scholarship. GM and RR were supported by the EPSRC grant EP/T033061/1.

References

Clements, M., Liu, X.-R., 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, M. J. 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, M. J. and Lambert, P. C. (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**(29), 4719–4742.

De Wreede, L. C., 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**(3), 261–274.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. Statistical science, **11**(2), 89–121.

Eletti, A., Marra, G., Quaresma, M., Radice, R., and Rubio, F. J. (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, H. C. (2008). Reduced-rank proportional hazards regression and simulation-based prediction for multi-state models. Statistics in Medicine, **27**(21), 4340–4358.

Iacobelli, S. and Carstensen, B. (2013). Multiple time scales in multi-state models. Statistics in medicine, **32**(30), 5315–5327.

Jackson, C. (2021). flexsurv: Flexible Parametric Survival and Multi-State Modelsflexsurv: 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**(3), 654–673.

Liu, X.-R., Pawitan, Y., and Clements, M. (2018). Parametric and penalized generalized survival models. Statistical Methods in Medical Research, **27**(5), 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**(530), 886–895.

Marra, G. and Radice, R. (2022). GJRM: Generalised Joint Regression Modelling. URL https://CRAN.R-project.org/package=GJRM. R package version 0.2-6.

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, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. Statistics in medicine, **26**(11), 2389–2430.

Putter, H., de Wreede, L. C., 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**(3), 543–559.

Pyke, R. (1961). Markov renewal processes: definitions and preliminary properties. The Annals of Mathematical Statistics, pages 1231–1242.

Ross, S. M., Kelly, J. J., Sullivan, R. J., Perry, W. J., Mercer, D., Davis, R. M., Washburn, T. D., Sager, E. V., Boyce, J. B., and Bristow, V. L. (1996). Stochastic processes, volume 2. Wiley New York.

Royston, P. and Parmar, M. K. (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**(15), 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**(1), 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**(3), 453–473.

Titman, A. C. (2011). Flexible nonhomogeneous markov models for panel observed data. Biometrics, **67**(3), 780–787.

Touraine, C., Helmer, C., and Joly, P. (2016). Predictions in an illness-death model. Statistical methods in medical research, **25**(4), 1452–1470.

Wood, S. N. (2017). Generalized Additive Models: An Introduction With R. Second Edition, Chapman & Hall/CRC, London.

Wood, S. N., Pya, N., and Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. Journal of the American Statistical Association, **111**(516), 1548–1563.

Yang, Y. and Nair, V. N. (2011). Parametric inference for time-to-failure in multi-state semi-markov models: A comparison of marginal and process approaches. Canadian Journal of Statistics, **39**(3), 537–555.

Younes, N. and Lachin, J. (1997). Link-based models for survival data with interval and continuous time censoring. Biometrics, pages 1199–1211.

Supplementary Material: "A Spline-Based Framework for the Flexible Modelling of Continuously Observed Multistate Survival Processes"

Alessia Eletti [1] Giampiero Marra [2] Rosalba Radice [3]

A Computing the transition probabilities in the continuously observed setting: an overview

When the process is assumed to be time-inhomogeneous, computing the transition probabilities from the estimated transition intensities is a nontrivial problem since closed form expressions of the former as functions of the latter are not available. Two main approaches can be identified in the literature of continuously observed processes to address this problem.

The first approach is to solve, by means of packages such as `deSolve` in `R` (Titman, 2011), the ordinary differential equations that tie the transition probability matrix to the transition intensity matrix, when the process is assumed to be Markov. This method is appealing in that it provides the entire transition probability matrix in one step and is the tech-

[1]Department of Statistical Science, University College London, Gower Street, WC1E 6BT London, UK, `alessia.eletti.19@ucl.ac.uk`

[2]Department of Statistical Science, University College London, Gower Street, WC1E 6BT London, UK, `giampiero.marra@ucl.ac.uk`.

[3]Faculty of Actuarial Science and Insurance, Bayes Business School, City, University of London, 106 Bunhill Row, EC1Y 8TZ London, UK. `rosalba.radice@city.ac.uk`.

nique implemented in the `R` packages `rstpm2` (Clements et al., 2021), through the function `markov_msm()`, and `flexsurv` (Jackson, 2021), through the function `pmatrix.fs()`. In both cases, the main required inputs are the fitted transition intensities. In the former case the transition intensities can be specified in a number of ways using a handful of existing survival modelling `R` packages, with the most flexible options provided by the `stpm2()` function present within the `rstpm2` package and the `survPen()` function from the `R` package `survPen` (Fauvernier et al., 2020). With regard to `flexsurv`, the most common parametric forms found in survival analysis, e.g. Weibull, can be assumed for the transition intensities through the function `flexsurvreg()` as well as the Royston-Parmar model through the function `flexsurvspline()`. Overall, the drawback of this approach is that it is difficult to generalise to the case in which the process is not assumed to be Markov, e.g. when it is semi-Markov, another common type of dependence on past history. Confidence intervals can then be obtained by using the covariance matrix computed from the knowledge of the first derivative of the transition probability matrix, obtained by simultaneously solving these ODEs mentioned above and an augmented version of them obtained by taking the derivative of the left and right hand-side with respect to time.

The second approach, and indeed the one that we adopt, is a simulation-based approach which allows one to estimate the transition probabilities by simulating a number M 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 method benefits of the generality lacking in the previous one, i.e. both Markov and semi-Markov processes are supported, thus tying nicely with the flexibility available for the transition-specific modelling. Indeed, is the only such approach which is general in this respect. It was proposed in Fiocco et al. (2008) and implemented in the `Stata` package `multistate` (Crowther and Lambert, 2016) and in the `R` packages `flexsurv` and `mstate` (Putter et al., 2020). In the following we will focus only on `R` packages. In particular, we will use the

latter as it can be seamlessly integrated with our estimation approach, implemented in the `R` package `GJRM`. Indeed, this package allows the user to obtained simulation-based estimates of the transition probability matrix at any vector of time points through function `mssample()` by providing the estimated transition-specific cumulative hazards computed at the time points of interest. Whether the process is Markov or semi-Markov is then simply accounted for by specifying the argument `clock = 'forward'` in the former case and `clock = 'reset'` in the latter. Note that the estimated cumulative hazards can in turn be straightforwardly obtained through function `hazsurv.plot()` from the `GJRM` package. Confidence intervals can then be obtained by simulation from the asymptotic distribution of the maximum likelihood estimates of the model parameters. This is what is done in `flexsurv` and `mstate`. We too adopt this approach as well by exploiting the fact that `hazsurv.plot()` already has a built-in way of simulating cumulative hazard functions given the asymptotic distribution of the model parameters. These can then be used as one would with a single cumulative hazard curve to obtain many corresponding transition probability matrices and thus compute the quantiles on these, as explained in Section 5. For more details on how to fit the transition intensities and then obtain transition probabilities and the related confidence intervals for a profile of interest, we refer the reader to Supplementary Material C and to the code accessible in the public repository `https://github.com/AlessiaEletti/ContinObsMultistateProcesses`, through which the results reported in the case study from Section 7 can be reproduced.

As an aside, in a nonparametric setting, one may also obtain the estimated transition probabilities through the Aalen-Johansen estimator which provides a way to compute the product integral tying the transition probability matrix to the matrix containing the transition-specific cumulative hazard functions, when the process is assumed to Markov (De Wreede et al., 2010). This is one of the approaches implemented in the `R` package `mstate`. Indeed, the transition specific cumulative hazard functions are computed through the `msfit()`

function either via the Aalen estimator (by specifying `vartype = 'Aalen'`) or the Greenwood estimator (by specifying `vartype = 'Greenwood'`). These estimates are then used in the `probtrans()` function to compute the transition probability matrix via the mentioned Aalen-Johansen estimator.

B Rewriting the model log-likelihood when only exact transitions are observed

For a multistate survival process assumed to be observed continuously and for individual $i = 1, \ldots, N$, where N represents the sample size, let $T_i^{(rs)}$ be the transition-specific true event time. This can be either uncensored, i.e. exactly observed, or right-censored if the transition $r \to s$ did not occur prior to the maximum follow-up time $T_{\max}$, in which case the transition is only known to have occurred after this time. In either case, the time may also be left-truncated if the event it relates to is an intermediate one, i.e one which requires the individual to have transition to the starting state considered prior to the current observation time. Indeed, left-truncation of survival data occurs when only individuals whose event time lies within a window (T_i^{td}, ∞) are observed, otherwise no information on the individuals is available and thus the subjects are not considered for inclusion into the study. This is precisely the case here. Indeed, given an intermediate state r, an individual is at risk of experiencing the transition $r \to s$ at a given time only if they are in state r at that time. In particular, if they are known to have transitioned to state r at time T_i^{td}, then they are at risk of the transition $r \to s$ only after this time, i.e. in the window (T_i^{td}, ∞), which is thus the left-truncation time associated with the transition.

We will now sketch the steps which show how one can pass from the general overall log-likelihood associated with a Markov multistate process to the re-formulation of it in terms of

ID	t_i^{start}	t_i^{stop}	Transition	Status
i	0	t_{12}	$1 \to 2$	1
i	0	t_{12}	$1 \to 3$	0
i	t_{12}	t^*	$2 \to 3$	0

Table 3: i^{th} individual in the dataset.

a sum of log-likelihoods, each associated with a specific transition, when the exact transition times are known. Showing this for the semi-Markov case is outside of the scope of this paper.

Let us assume that a random *i.i.d.* sample of size N and let $0 = t_{i0} < t_{i1} < \cdots < t_{in_i}$ be the observed transition times for individual i. At these times the process is observed to be in states $z_{i0}, z_{i1}, \ldots, z_{in_i}$. If $\ell_i(\boldsymbol{\theta})$ is the likelihood contribution of individual i, $\mathcal{A} = \{(r,s) \in \mathcal{S} \times \mathcal{S} \mid r \neq s \wedge q^{(rs)}(\cdot) \neq 0\}$ is the set of the pairs of states corresponding to allowed transitions and $\boldsymbol{\theta} = \{\boldsymbol{\beta}^{(rs)} \mid (r,s) \in \mathcal{A}\}$ is an overall model parameter vector, the full log-likelihood is given by

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \ell_{ij}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \log(L_{ij}), \tag{B.1}$$

where

$$L_{ij} = \exp\left[\int_{t_{ij-1}}^{t_{ij}} q_{z_{ij-1},z_{ij-1}}(u;x)du\right] q_{z_{ij-1},z_{ij}}(t_{ij};x).$$

We will now clarify this by specialising it to a simple example which will allow us to write out each term explicitly. In particular, we will do this for a time-homogeneous IDM for simplicity but the same reasoning can be extended to more general contexts as settings. Let us assume we have a dataset with the i^{th} individual characterised by the observed transitions described in Table 3. The i^{th} likelihood contribution associated with the process

at hand will have the following form

$$\begin{aligned} L_i &= \overbrace{q_{12}\exp[t_{12}q_{11}]}^{1^{st}\text{ term}} \cdot \overbrace{p_{22}(t^* - t_{12})}^{2^{nd}\text{ term}} \\ &= q_{12}\exp[t_{12}q_{11}] \cdot \exp[-(t^* - t_{12})q_{23}], \end{aligned} \tag{B.2}$$

i.e. is the product of the contributions associated with the two observation times t_{12} and t^*. In particular, the first term refers to the the exactly observed transition at time t_{12}. Recall that we assume the process stays in state 1 throughout time interval $(0, t_{12})$, hence the term q_{11} in the exponential, and then jumps to state 2 at time t_{12}. The second term, instead, refers to the fact that the process is observed to be in state 2 at time t_{12} and to still be in state 2 at time t^*, the maximum follow-up time. This can be re-written in the following way

$$\begin{aligned} L_i &= q_{12}\exp[t_{12}q_{11}] \cdot \exp[-(t^* - t_{12})q_{23}] \\ &= q_{12}\exp[-t_{12}(q_{12} + q_{13})] \cdot \exp[-(t^* - t_{12})q_{23}] \\ &= q_{12}\exp[-t_{12}q_{12}] \cdot \exp[-t_{12}q_{13}] \cdot \exp[-(t^* - t_{12})q_{23}] \\ &= f_{12}(t_{12}) \cdot S_{13}(t_{12}) \cdot \frac{S_{23}(t^*)}{S_{23}(t_{12})}. \end{aligned} \tag{B.3}$$

In other terms, we broke up the likelihood in the product of terms each corresponding to specific transitions and hence which are functions of nonoverlapping sets of parameters. The usefulness of writing the likelihood contribution as a product of densities and survival functions associated to each transition, rather than as the product of transition probabilities, comes from the fact that one can then group all of the terms relating to the transition $r \to s$ and obtain a transition specific likelihood contribution $L_i^{(rs)}$. In this way we thus have

$$L_i = f_{12}(t_{12}) \cdot S_{13}(t_{12}) \cdot \frac{S_{23}(t^*)}{S_{23}(t_{12})} = L_i^{(12)} \cdot L_i^{(13)} \cdot L_i^{(23)},$$

where each transition specific likelihood can be optimised as a standalone likelihood associated to what then becomes a univariate survival analysis problem. Note, further, that the left-truncation for the $2 \to 3$ transition is apparent in that we have a conditional survival function.

C Further details on the case study and code

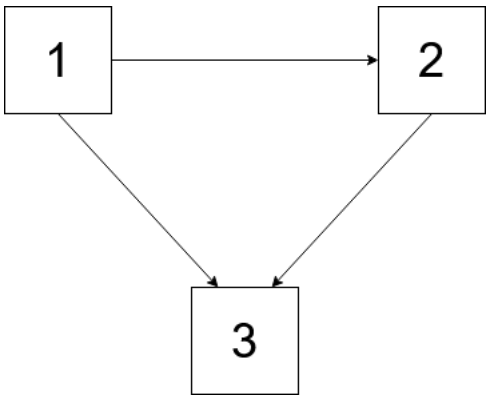

Figure 5: Graphical representation of the IDM assumed to model the data.

The results presented in the case study have been obtained by combining the `R` packages `GJRM` and `mstate`, as mentioned above. Note, however, that a small bug was found in the sampling function `mssample()` of the latter package, which thus had to be modified. To allow for the full reproducibility of the analysis carried out this paper, we therefore not only provide the code used for the case study, but also the modified function. In this way the code provided is entirely self contained. This can be found in the public repository `https://github.com/AlessiaEletti/ContinObsMultistateProcesses`. In the following, instead, we report only a few code snippets to exemplify the usage of the main functions needed for the fitting of the model through our framework, implemented in the `GJRM` package, and the computation of the estimated transition probabilities via the simulation based procedure implemented in the `mstate` package. In particular, the transition-specific models are fitted using the `gamlss()` function from the `GJRM` package, as shown in the following code snippet for the $2 \rightarrow 3$ transition, so as to also show how we account for left-truncation.

```
out.rd = gamlss(list(Tstop ~ s(log(Tstop), bs = 'mpi') +  size2 + size3 + hormon
                           + s(age) + s(nodes) + s(pr_1) + ti(log(Tstop), pr_1)),
                surv = TRUE, margin = 'PH',
```

```
data = mex[mex$trans == 3, ],
truncation.time = 'Tstart',
type = 'mixed',
cens = status.factor[mex$trans == 3])
```

The plots of the smooths included in the three models, and reported in Figure 2 of Section 7 and in Figures 6 and 7 below, can be obtained by using the `plot()` command on the fitted model output.

We then obtain the estimated transition probabilities by combining the predicted cumulative hazards obtained using function `hazsurv.plot()` from the `GJRM` package with the (modified) function `mssample()` from the `mstate` package. Indeed, the latter takes the estimated transition-specific cumulative hazards as an input and samples paths through the multistate model outputting either the sampled paths or the estimated transition probabilities depending on the user choice; this is controlled by argument `output`. In particular, for this application, we used $M = 10000$ sampled paths through the multistate model. We show this in the following code snippet for one of the three transitions, as the others are then identical.

```
# 1-3 transition
pred.rd.test =  hazsurv.plot(out.rd, eq = 1, t.vec = times,
                             newdata = newdata, type = 'cumhaz', plot.out = F)

CH13 = pred.rd.test$ch # 1-3 cumulative hazard
# ... (similarly for the others)

Hazprep = data.frame(time = rep(times, 3),
                     Haz = c(CH12, CH13, CH23),
```

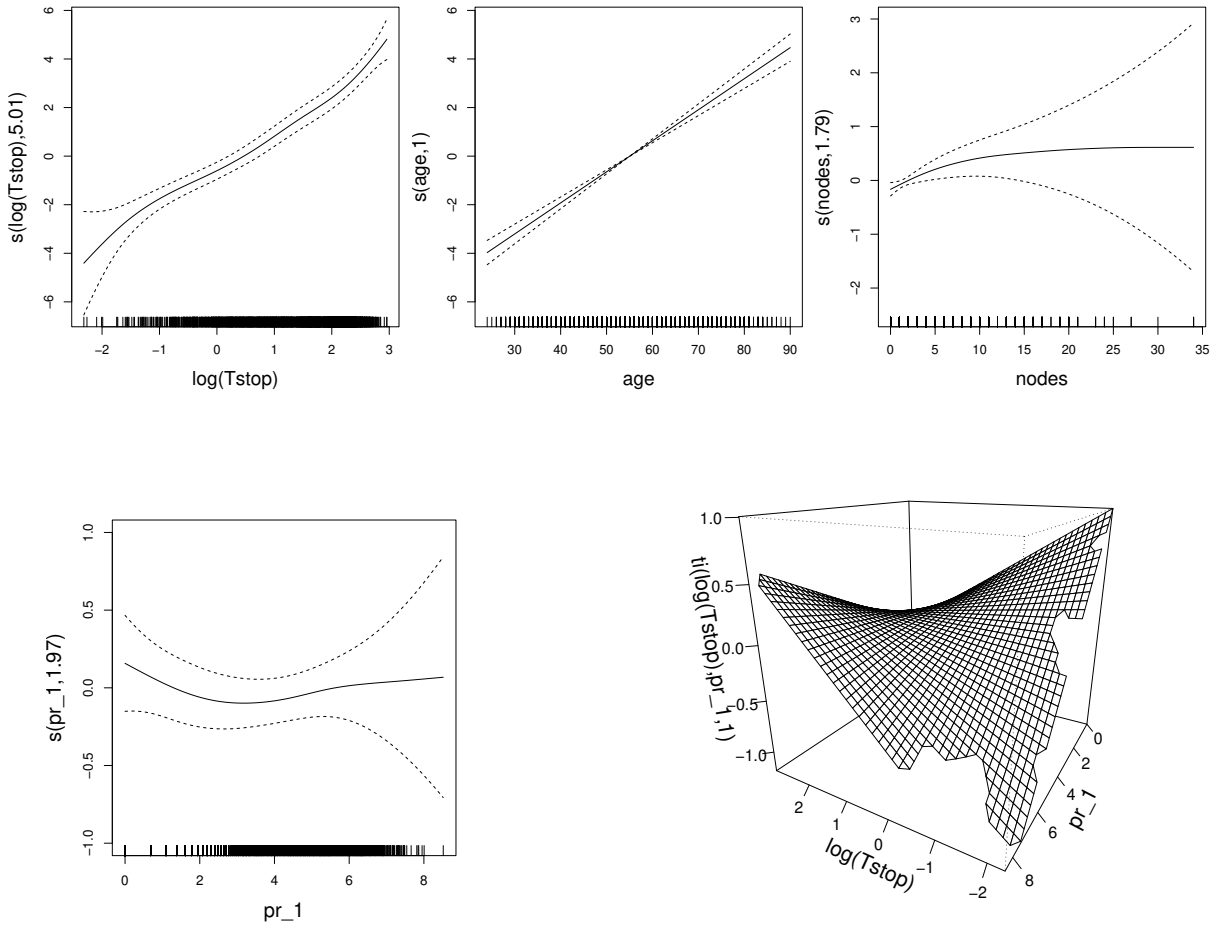

Figure 6: Smooth of log-time (top left), smooth of age (top middle), smooth of the number of positive nodes (top right), smooth of the progesterone level (bottom left) and smooth interaction between log-time and progesterone level (bottom right) for the transition $health \rightarrow death$.

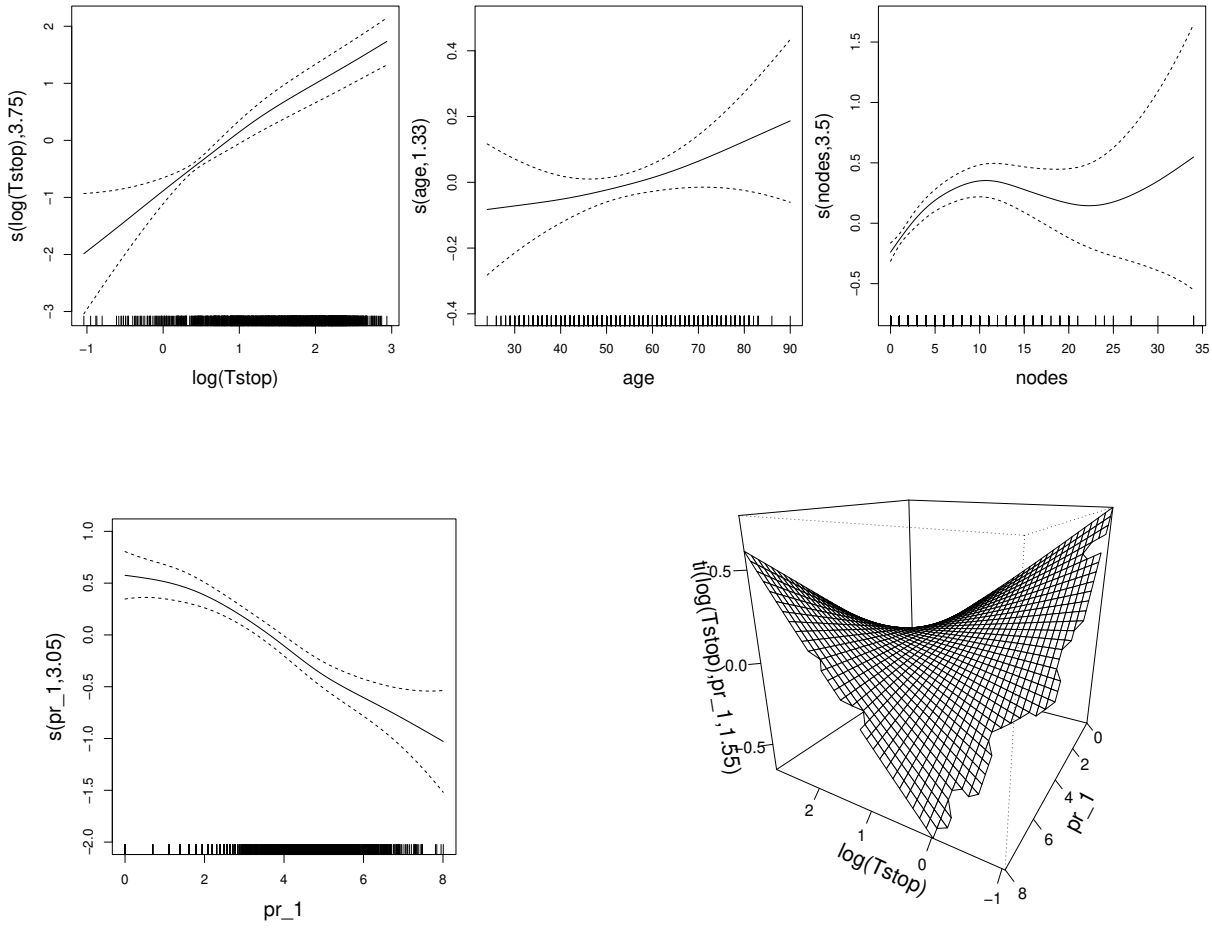

Figure 7: Smooth of log-time (top left), smooth of age (top middle), smooth of the number of positive nodes (top right), smooth of the progesterone level (bottom left) and smooth interaction between log-time and progesterone level (bottom right) for the transition *relapse* $\rightarrow$ *death*.

```
                    trans = c(rep(1, length(times)),
                              rep(2, length(times)),
                              rep(3, length(times))))

probs = mssample(Haz = Hazprep, trans = transmat, tvec = times, M = 10000)
```

As mentioned above, this approach allows us to model both Markov and semi-Markov processes in the exact same way, with the only exception that a different time scale needs to be defined for the latter. In particular, when fitting the model using function `gamlss()` we would reset the time at the moment of entry to each state. When calling function `mssample()` to obtain the transition probabilities one needs to set argument `clock = 'reset'` to specify that the time-scale of the cumulative hazards is the duration in the present state. In this way we are able to fully harness the flexibility allowed when the multistate process is observed continuously through time by combining existing tools. Note that, when the process is observed only intermittently, it becomes considerably more difficult to allow for this degree of flexibility in the assumptions made on the dependence on time and past history. An attempt assuming semi-Markovianity was also made but resulted in inferior AIC values, we thus omit the results here.

We can now obtain confidence intervals for the estimated transition probabilities by simulation. In particular, we already have simulated transition-specific cumulative hazard functions from the previous calls to the `hazsurv.plot()` function. Each of these can thus be used as inputs to obtain the corresponding transition probabilities through the simulation-based procedure, i.e. by iteratively repeating the computation shown in the code snippet above for each simulated transition specific cumulative hazard. Note that the simulated transition-specific cumulative hazard can be extracted through the command `pred.rd.test$s.sim`. The quantiles of the resulting set of transition probabilities extracted

in this way can thus be computed to find the 95% confidence intervals.

Finally, the plots in Figure 1 can be obtained using the `hazsurv.plot()` function, specifying that the curve of interest is the hazard through argument `type = 'hazard'`, as shown below for the $2 \rightarrow 3$ transition intensity.

```
# Transition n. 3 (2-3)
q23 = hazsurv.plot(out.rd, eq = 1,
                   t.vec = seq(min(mex$Tstop), max(mex$Tstop), length.out = 1000),
                    newdata = data.frame(age = 54, size2 = 0, size3 = 1,
                                         nodes = 10, pr_1 = 3, hormon = 1),
                    type = 'hazard',
                    ylab = 'Relapse to Death transition intensity',
                    xlab = 'Time since surgery (years)')
```

In conclusion, note that when fitting the time-only model with our splines-based approach, i.e. when the transition intensities are specified with no covariates, we indeed recover the estimated transition-specific cumulative hazard functions reported in Crowther and Lambert (2017). We report these in Figure 8. These plots can be straightforwardly obtained using function `hazsurv.plot()` from the `R` package `GJRM` by specifying the argument `type = 'cumhaz'`. It can, for instance, be seen that when no covariates are considered, the risk of transitioning to the death state is considerably higher given that relapse occurred compared to the relapse-free setting.

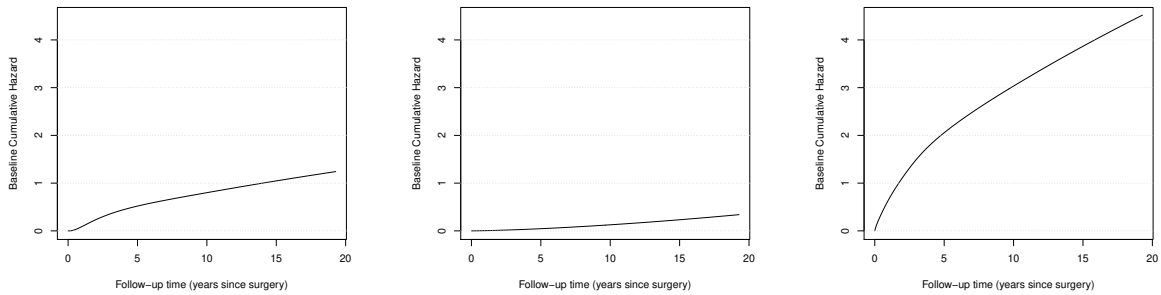

Figure 8: Estimated baseline cumulative hazard functions associated with the health to relapse (left), health to death (middle) and relapse to death (right) transitions with their 95% confidence intervals.

D Further details on the continuously observed setting

Longitudinal data are characterised by multiple observations through time of at least one quantity of interest, for the same individual and generally come in one of two forms, referred to as *stacked* (or long) and *unstacked* (or wide), respectively. In the unstacked (or wide) data format, a subject's repeated responses will be displayed in a single row, i.e. each response is in a separate column. In the stacked (or long) data format, each row represents a single time point per subject. So each subject will have data in multiple consecutive rows. In the continuously observed setting, the data will typically be formatted in the latter form. In particular, assuming an IDM like the one considered in the case study of Section 7, each of the rows corresponding to a given patient will look like either of the four combinations in Tables 4-7. Note that the only type of censoring possible in this setting is right censoring, i.e. the transition has not taken place by the maximum follow-up time. Note also that the start time for transitions exiting the first state will usually be 0. An exception to this is had when the first transition is itself left-truncated, e.g., as a consequence of the nature of the phenomenon of interest. We refer the reader to the tutorial by Putter (2011) as well for further examples on the setup for continuously observed multistate survival processes.

trans	start	stop	event
$1 \to 3$	0	t^{max}	0
$1 \to 2$	0	t^{max}	0

Table 4: The patient does not experience any transition between 0 and t^{max}, the maximum observed follow-up time. The patient is right censored at t^{max} for both transitions.

trans	start	stop	event
$1 \to 3$	0	t_{12}	0
$1 \to 2$	0	t_{12}	1
$2 \to 3$	t_{12}	t^{max}	0

Table 6: The patient experiences a transition to the intermediate state at time t_{12} but does not transition to the following state between t_{12} and t^{max}, i.e. $1 \to 2$ is uncensored. For $2 \to 3$ t_{12} is a left truncation time while t^{max} is a right censoring time. For $1 \to 3$, t_{12} represents a right censoring time.

trans	start	stop	event
$1 \to 3$	0	t_{13}	1
$1 \to 2$	0	t_{13}	0

Table 5: The patient experiences a transition to the absorbing state at time t_{13}. Transition $1 \to 3$ is thus uncensored. For transition $1 \to 2$ this represents a right censoring time.

trans	start	stop	event
$1 \to 3$	0	t_{12}	0
$1 \to 2$	0	t_{12}	1
$2 \to 3$	t_{12}	t_{23}	1

Table 7: The patient experiences a transition to the intermediate state at time t_{12} and then transitions to the absorbing state at time t_{23}. Transitions $1 \to 2$ and $2 \to 3$ are thus uncensored. For transition $2 \to 3$, t_{12} represents a left truncation time. For transition $1 \to 3$, t_{12} represents a right censoring time.