Predictions for parametric and penalised multi-state Markov models

Mark Clements
Karolinska Institutet

Abstract

We describe an efficient algorithm and implementation for predictions for parametic and penalized Markov multi-state models. Predictions include state occupancy probabilities, transition probabilities, prevalence, length of stay, relative survival, screening sensitivity, and costs. The algorithm uses a system of ordinary differential equations to calculate the predictions and their gradients, with standard errors calculated using the delta method. These methods have applications to a range of disciplines, including descriptive epidemiology, causal inference and economic evaluations.

Keywords: survival, multi-state models, Markov models.

1. Introduction

Multi-state models form a very broad class of models that includes standard survival models with an initial and final state, competing risks with multiple final states, and illness-death models, with an initial healthy state, an illness state and a death state. This model class is useful for representing movement through a discrete set of states. As a motivating example, we are interested in the clinical pathways for cancer screening, diagnosis and treatment, taking account of utilities and costs. Using observational data, screening or treatment assignment may be confounded with patient characteristics, so we are also interested in using standardisation or the parametric g-formula to adjust for differences between groups.

Predictions from multi-state models could include: transition hazards between states; transition probabilities for being in a particular state conditional on an initial state; state occupation probabilities for being in a particular state given an observed distribution of initial states; length of stay in a given state; prevalence for the live health states; contrasts of these predictions, including standardisation; and ratios of these predictions (**??**).

Much of the literature on multi-state models has focused on counting processes (e.g. **???**). For implementations, **?** use the Aalen-Johansen estimator to estimate transition probabilities, and **?** predict transition probabilities and their standard errors for transitions modelled using Cox regression. These implementations can also estimate length of stay and contrasts, however variance estimates would then require the non-parametric bootstrap. **?** use stochastic differential equations to transform non-parametric cumulative hazard estimates into a variety of predictions. This approach is shown to work for Aalen's additive hazard model, but it is unclear whether the approach extend to Cox regression models. Their methods are a non-parametric analogue to our development.

Parametric and penalised survival models have potential advantages for multi-state models, including the ready incorporation of time-varying effects and for predictions outside of observed data (**??**). **?** simulate from Poisson regression models on different time scales to predict for multi-state models. **?** provide a simulation framework for multi-state models with random

times based on piece-wise constant hazards. **?** combine parametric time-to-event models to predict transition probabilities and length of stay. The authors use simulations, with variance estimation using the parametric bootstrap. None of these approaches scales well to allow for standardisation across moderate sized datasets.

Predictions for Markov models can use ordinary differential equation solvers to solve Kolmogorov's forward differential equation. Recently, **?** predicted length of stay using differential equations. **?** also demonstrated that the gradients for transition probabilities (or sensitivity equations) can be estimated by augmenting the system of differential equations. Variance estimates for predictions can be calculated using the delta method and gradients for the predictors (**?**). This provides an opportunity to efficiently calculate a range of predictions from multi-state models.

We have two objectives: first, to develop efficient methods for predictions for *smooth, non-homogeneous multi-state Markov models* with variance estimation using the *multivariate delta method*; and, second, to demonstrate that these methods support the use of multi-state models across descriptive statistics, causal inference, and economics. Importantly, our restriction to Markov models is not a heavy constraint: **?** have shown that the state occupation probabilities are consistently estimated under moderate conditions even when the time scale is mis-specified. Length of stay, prevalence, utilities and costs, when they are integrated functions of the state occupation probabilities, are also expected to be consistent under similar conditions.

In outline, we provide a theoretical development of a set of ordinary differential equations, simulate to assess the small sample properties, provide some examples and conclude with a brief discussion.

2. Methods

Assumptions

For the predictions, the main inputs are (i) a multi-state model specification, (ii) the models for the transition intensities, and (iii) and the initial values for the states. We assume that the transition intensities are estimated using maximum (penalised) likelihood estimation, with stacked estimated parameters $\hat{\boldsymbol{\beta}}$ and stacked estimated variance-covariance matrix $\hat{\boldsymbol{\Sigma}}$. The estimated parameters could be from either model fit for all transitions, one model for each transition or a combination of models for different transitions.

For a formal development, the asymptotic properties for parametric survival models were developed by **?** and **?**. Under Cramér-like conditions, the authors show that that the maximum likelihood parameters $\hat{\boldsymbol{\beta}}$ are asymptotically normal. Sufficient conditions include uniform convergence, that the hazard is thrice differentiable with respect to the parameters, boundedness of the hazard function, and that the hazard is bounded away from zero (**?**). We can then estimate the asymptotic variance for the predictions using the multivariate delta method, which will be asymptotically normal if the parameters are asymptotically normal and the gradients of the predictions exist (see Theorem 3.4.6 in **?**). **?** developed asymptotic properties for non-parametric stochastic differential equations under Hadamard continuity. We will assess these asymptotic properties using simulations.

Notation

For vectors $\boldsymbol{v}, \boldsymbol{v}_1$ and $\boldsymbol{v}_2$ and matrices $\boldsymbol{M}, \boldsymbol{M}_1$ and $\boldsymbol{M}_2$, let $\boldsymbol{v}_1 \circ \boldsymbol{v}_2$ and $\boldsymbol{M}_1 \circ \boldsymbol{M}_2$ be the Hadamard element-wise products, and let $\boldsymbol{v}_1 \oslash \boldsymbol{v}_2$ and $\boldsymbol{M}_1 \oslash \boldsymbol{M}_2$ be Hadamard element-wise division. We also define $\boldsymbol{M} \circ \boldsymbol{v} = \boldsymbol{M} \circ (\boldsymbol{v}\mathbf{1}^T)$ and $\boldsymbol{M} \oslash \boldsymbol{v} = \boldsymbol{M} \oslash (\boldsymbol{v}\mathbf{1}^T)$. We assume that $\circ$

and $\oslash$ have higher operator precedence than addition and subtraction. Let $\text{diag}(\boldsymbol{v})$ represent a square matrix with zeros in the off-diagonal and $\boldsymbol{v}$ on the diagonal. We represent a vector of ones using $\mathbf{1}$ and an identity matrix by $\boldsymbol{I}$.

We define the gradient for a prediction $\boldsymbol{\phi}(t_0, t)$ with respect to an estimated model parameter $\hat{\beta}_m$ by $\boldsymbol{\phi}'_m(t_0, t) \equiv \left.\frac{\partial \boldsymbol{\phi}(t_0,t)}{\partial \beta_m}\right|_{\beta_m=\hat{\beta}_m}$.

General approach applied to transition and state occupancy probabilities

Let the set of states be indexed by i and j. We define the matrix of *transition probabilities* $\boldsymbol{P}(t_0, t) = (P_{ij}(t_0, t))$ as the probabilities of being in state j at time t given being in an initial state i at entry time t_0. For smooth hazards, the Markov property is expressed through Kolmogorov's forward differential equation, such that

$$\frac{d\boldsymbol{P}(t_0,t)}{dt} = \boldsymbol{P}(t_0,t)\boldsymbol{Q}(t) \tag{1}$$

$$\boldsymbol{P}(t_0,t_0) = \boldsymbol{I} \tag{2}$$

where $\boldsymbol{Q}(t) = (Q_{ij}(t))$ is a matrix of transition intensities at time t[1] from state i to state j when $i \neq j$ (conceptualised as the rates from state i to state j), and where $Q_{ii}(t) = -\sum_{j\neq i} Q_{ij}(t)$ (conceptualised as the rates out of state i).

Following **?**, Kolmogorov's forward differential equation can be augmented to calculate the gradient for $\boldsymbol{P}(t_0, t)$ with respect to the model coefficients. Titman showed that

$$\frac{d\boldsymbol{P}'_m(t_0,t)}{dt} = \boldsymbol{P}'_m(t_0,t)\boldsymbol{Q}(t) + \boldsymbol{P}(t_0,t)\boldsymbol{Q}'_m(t) \tag{3}$$

$$\boldsymbol{P}'_m(t_0,t_0) = \mathbf{0} \tag{4}$$

Note that this requires the evaluation of the gradients for the transition intensities with respect to $\hat{\beta}_m$. *Algorithm 1* is defined as:

input : $\boldsymbol{P}(t_0,t_0), \boldsymbol{Q}(t), \{\boldsymbol{Q}'_m(t)\}\,\forall m, t_0, t_{\max}$
output: $\boldsymbol{P}(t_0,t), \{\boldsymbol{P}'_m(t_0,t)\}\,\forall m$ for $t \in [t_0, t_{\max})$
begin
  define the ODE based on Equations (**??**)–(**??**);
  run an ODE solver from time t_0 to time $t_{\max}$;
end

This algorithm could be done separately for each covariate pattern or the ODEs can be extended to multiple covariates by stacking the equations. Using the multivariate delta method, the estimated covariance matrix for $\boldsymbol{P}(t_0, t)$ is

$$\text{var}(\boldsymbol{P}(t_0,t)) = \boldsymbol{P}'_m(t_0,t)^T\,\hat{\boldsymbol{\Sigma}}\,\boldsymbol{P}'_m(t_0,t)$$

Let the vector $\boldsymbol{\pi}(t_0, t_0) = (\pi_j(t_0, t_0))$ represent the initial *state occupation probabilities* of being in state j at time t_0. The state occupation probabilities of being in state j at time $t > t_0$ can be calculated by $\boldsymbol{\pi}(t_0,t)^T = (\pi_j(t_0,t)) = \boldsymbol{\pi}(t_0,t_0)^T\boldsymbol{P}(t_0,t)$. Differential equations can be readily calculated for the state occupation probabilities.

This general approach can be applied to a variety of predictions. The other predictions may require that a set of differential equations be defined for several related predictions, including their gradients, in combination with the delta method for variance estimation.

[1] Technically, this is $t-$.

Example 1: Length of stay and restricted mean survival time

The ordinary differential equations in Equations (**??**)–(**??**) can be further augmented to calculate the integral $L_{ij}(t_0, t) = \int_{t_0}^{t} P_{ij}(t_0, v)dv$. The additional differential equations are then

$$\frac{d\boldsymbol{L}(t_0,t)}{dt} = \boldsymbol{P}(t_0,t) \tag{5}$$

$$\frac{d\boldsymbol{L}'_m(t_0,t)}{dt} = \boldsymbol{P}'_m(t_0,t) \tag{6}$$

$$\boldsymbol{L}(t_0,t_0) = \boldsymbol{L}'_m(t_0,t_0) = \boldsymbol{0} \tag{7}$$

where the matrix $\boldsymbol{L}(t_0, t)$ is the the *length of stay* or sojourn time for state j given an initial state i at time t_0. Algorithm 1 would be augmented to include Equations (**??**)–(**??**) to calculate transition and state occupation probabilities and the lengths of stay.

For restricted mean survival times and life expectancy, we can weight the length of stay by a vector $\boldsymbol{w} = (w_j)$, where $w_j = 1$ for a live state j, and $w_j = 0$ for a death state j. We then have that the restricted mean survival is $\boldsymbol{L}(t_0, t)\boldsymbol{w}$. Moreover, as $t \to \infty$, $\boldsymbol{L}(t_0, t)\boldsymbol{w}$ will measure life expectancy.

Example 2: Prevalence

Let the prevalence for the live states be defined by

$$\tilde{\boldsymbol{P}}(t_0,t) = (\boldsymbol{P}(t_0,t)\text{diag}(\boldsymbol{w})) \oslash (\boldsymbol{P}(t_0,t)\boldsymbol{w})$$

where the weight vector $\boldsymbol{w}$ again has elements that are 1 for a live state and 0 for a death state. Generalising a result for illness-death models by **?** to multi-state models, we have that

$$\frac{d\tilde{\boldsymbol{P}}(t_0,t)}{dt} = \left(\left(\frac{d\boldsymbol{P}(t_0,t)}{dt}\text{diag}(\boldsymbol{w})\right) \circ (\boldsymbol{P}(t_0,t)\boldsymbol{w}) - (\boldsymbol{P}(t_0,t)\text{diag}(\boldsymbol{w})) \circ \left(\frac{d\boldsymbol{P}(t_0,t)}{dt}\boldsymbol{w}\right)\right) \oslash (\boldsymbol{P}(t_0,t)\boldsymbol{w}) \oslash (\boldsymbol{P}(t_0,t)\boldsymbol{w})$$

with an initial value that $\tilde{\boldsymbol{P}}(t_0, t_0) = \text{diag}(\boldsymbol{w})$. Alternatively, we can calculate $\text{logit}(\tilde{\boldsymbol{P}}(t_0, t)) = \log(\tilde{\boldsymbol{P}}(t_0, t)) - \log(\boldsymbol{1}\boldsymbol{1}^T - \tilde{\boldsymbol{P}}(t_0, t))$, which has a gradient of

$$\frac{\partial}{\partial\beta_m}\text{logit}(\tilde{\boldsymbol{P}}(t_0,t))\Big|_{\beta_m=\hat{\beta}_m} = \left((\boldsymbol{P}'_m(t_0,t)\text{diag}(\boldsymbol{w})) \circ (\boldsymbol{P}(t_0,t)\boldsymbol{w}) - (\boldsymbol{P}(t_0,t)\text{diag}(\boldsymbol{w})) \circ (\boldsymbol{P}'_m(t_0,t)\boldsymbol{w})\right) \oslash \tilde{\boldsymbol{P}}(t_0,t) \oslash \left(\boldsymbol{1}\boldsymbol{1}^T - \tilde{\boldsymbol{P}}(t_0,t)\right) \oslash (\boldsymbol{P}(t_0,t)\boldsymbol{w}) \oslash (\boldsymbol{P}(t_0,t)\boldsymbol{w})$$

Similarly, we could calculate the proportion of person-time in the live health states, which is calculated by $\tilde{\boldsymbol{L}}(t_0, t) = (\boldsymbol{L}(t_0, t)\text{diag}(\boldsymbol{w})) \oslash (\boldsymbol{L}(t_0, t)\boldsymbol{w})$. The development for these predictions follows closely that for prevalence.

Example 3: Linear combinations, differences and standardisation

Linearity combinations of these estimators are straightforward to calculate, as the gradients are then also linear. Let $\boldsymbol{\phi}(t_0, t|\boldsymbol{x}_k)$ represent an estimator, such as transition probabilities, state occupation probabilities or length of stay, conditional on covariate vector $\boldsymbol{x}_k$. Given weights w_k for $k = 1, \ldots, K$, we can calculate the weighted sums

$$\bar{\boldsymbol{\phi}}(t_0,t) = \sum_{k=1}^{K} w_k \boldsymbol{\phi}(t_0,t|\boldsymbol{x}_k)$$

$$\bar{\boldsymbol{\phi}}'_m(t_0,t) = \sum_{k=1}^{K} w_k \boldsymbol{\phi}'_m(t_0,t|\boldsymbol{x}_k)$$

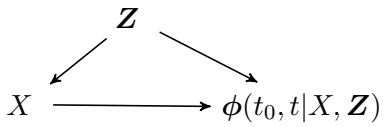

Figure 1: Directed acyclic graph showing predictions $\boldsymbol{\phi}(t_0, t|X, \boldsymbol{Z})$ as a function of the target exposure X and confounders $\boldsymbol{Z}$.

For differences, we can define $K = 2$, $w_1 = 1$ and $w_2 = -1$.

Standardisation can be defined in terms of the mean prediction under a counterfactual exposure. For example, consider a binary exposure X with confounders $\boldsymbol{Z}_i$ across a sample or population indexed by i. The standardised estimator for everyone been unexposed or exposed to X is then

$$\bar{\boldsymbol{\phi}}_0(t_0, t) = E_{\boldsymbol{Z}}(\boldsymbol{\phi}(t_0, t|X = 0, \boldsymbol{Z})) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\phi}(t_0, t|X = 0, \boldsymbol{Z}_i)$$
$$\bar{\boldsymbol{\phi}}_1(t_0, t) = E_{\boldsymbol{Z}}(\boldsymbol{\phi}(t_0, t|X = 1, \boldsymbol{Z}))$$

Under no unmeasured confounding and positivity, we could give a causal interpretation to contrasts such as $\bar{\boldsymbol{\phi}}_1(t_0, t) - \bar{\boldsymbol{\phi}}_0(t_0, t)$, $\bar{\boldsymbol{\phi}}_1(t_0, t) \oslash \bar{\boldsymbol{\phi}}_1(t_0, t)$ or $\mathbf{1}\mathbf{1}^T - \bar{\boldsymbol{\phi}}_1(t_0, t) \oslash E_{\boldsymbol{Z},X}(\boldsymbol{\phi}(t_0, t|X\boldsymbol{Z}))$. These marginal estimators can be interpreted as applications of the parametric g-formula .

Example 4: Ratios

We can augment the system of differential equations to calculate the ratios for specific estimators (**?**). Let a matrix of ratios of state occupation probabilities be $\mathbf{R}(t_0, t|\boldsymbol{x}_1, \boldsymbol{x}_2) = \boldsymbol{\pi}(t_0, t|\boldsymbol{x}_1) \oslash \boldsymbol{\pi}(t_0, t|\boldsymbol{x}_2)$. Then

$$\begin{aligned}\frac{d\mathbf{R}(t_0, t|\boldsymbol{x}_1, \boldsymbol{x}_2)^T}{dt} &= \Big(\Big(\boldsymbol{\pi}(t_0, t|\boldsymbol{x}_1)^T \boldsymbol{P}(t_0, t|\boldsymbol{x}_1)\boldsymbol{Q}(t|\boldsymbol{x}_1)\Big) - \Big(\boldsymbol{\pi}(t_0, t|\boldsymbol{x}_2)^T \boldsymbol{P}(t_0, t|\boldsymbol{x}_2)\boldsymbol{Q}(t|\boldsymbol{x}_2)\Big) \circ \mathbf{R}(t_0, t|\boldsymbol{x}_1, \boldsymbol{x}_2)\Big) \\ &\quad \oslash \boldsymbol{P}(t_0, t|\boldsymbol{x}_2) \\ \mathbf{R}(t_0, t_0) &= \mathbf{1}\end{aligned}$$

where $\boldsymbol{P}(t_0, t|\boldsymbol{x}_2) > 0$. Alternatively, we can calculate the gradient for $\log(\boldsymbol{\phi}(t_0, t|\boldsymbol{x}_1) \oslash \boldsymbol{\phi}(t_0, t|\boldsymbol{x}_2))$ (that is, the log ratio), which is

$$\frac{\partial}{\partial \beta_m} \log\left(\boldsymbol{\phi}(t_0, t|\boldsymbol{x}_1) \oslash \boldsymbol{\phi}(t_0, t|\boldsymbol{x}_2)\right)\Big|_{\beta_m = \hat{\beta}_m} = \boldsymbol{\phi}'_m(t_0, t|\boldsymbol{x}_1) \oslash \boldsymbol{\phi}(t_0, t|\boldsymbol{x}_1) - \boldsymbol{\phi}'_m(t_0, t|\boldsymbol{x}_2) \oslash \boldsymbol{\phi}(t_0, t|\boldsymbol{x}_2)$$

This evaluation depends on $\boldsymbol{\phi}(t_0, t|\boldsymbol{x}_2)$ being non-zero.

Example 5: Utilities and costs

The approach readily incorporates utilities and costs. For utilities, we have the cumulative discounted utilities $U_i(t_0, t) = \sum_j \int_{t_0}^{t} P_{ij}(t_0, v)u_j(v)e^{-\lambda v}dv$, where $u_j(v)$ is the utility for state j at time v and $\lambda = \log(1 + \delta)$ is the rate of decline for a discount rate δ. The augmented differential equations are then

$$\frac{d\boldsymbol{U}(t_0, t)}{dt} = \boldsymbol{P}(t_0, t)\boldsymbol{u}(t)e^{-\lambda t} \tag{8}$$
$$\frac{d\boldsymbol{U}'_m(t_0, t)}{dt} = \left(\boldsymbol{P}'_m(t_0, t)\boldsymbol{u}(t) + \boldsymbol{P}(t_0, t)\boldsymbol{u}'_m(t)\right)e^{-\lambda t} \tag{9}$$
$$\boldsymbol{U}(t_0, t_0) = \boldsymbol{U}'_m(t_0, t_0) = \mathbf{0} \tag{10}$$

In health economics, the discounted, quality-adjusted life-years are calculated by the product of the initial state probabilities and the utilities, such that $\text{QALY}(t_0, t) = \boldsymbol{\pi}(t_0, t_0)^T \boldsymbol{U}(t_0, t)$, with gradient $\text{QALY}'_m(t_0, t) = \boldsymbol{\pi}(t_0, t_0)^T \boldsymbol{U}'_m(t_0, t)$.

Costs may be represented as accumulated costs or costs at a point in time. Let costs per unit time for being in state i at time t be represented by the vector function $\boldsymbol{c}(t) = (c_i(t))$ and model for costs for transitions from state i to state j at time t, represented by the matrix $\boldsymbol{\mathcal{C}}(t) = (\mathcal{C}_{ij}(t))$, with $\mathcal{C}_{ii}(t) = 0$. Then the cumulative discounted costs $\boldsymbol{C}(t_0, t)$ can be represented by the equations

$$\frac{d\boldsymbol{C}(t_0,t)}{dt} = \boldsymbol{P}(t_0,t)\left(\boldsymbol{c}(t) + (\boldsymbol{Q}(t) \circ \boldsymbol{\mathcal{C}}(t))\,\mathbf{1}\right) e^{-\lambda t} \tag{11}$$

$$\begin{aligned}\frac{d\boldsymbol{C}'_m(t_0,t)}{dt} = \Big(&\boldsymbol{P}'_m(t_0,t)\left(\boldsymbol{c}(t) + (\boldsymbol{Q}(t) \circ \boldsymbol{\mathcal{C}}(t))\,\mathbf{1}\right) + \\ &\boldsymbol{P}(t_0,t)\left(\boldsymbol{c}'_m(t) + (\boldsymbol{Q}'_m(t) \circ \boldsymbol{\mathcal{C}}(t) + \boldsymbol{Q}(t) \circ \boldsymbol{\mathcal{C}}'_m(t))\,\mathbf{1}\right)\Big) e^{-\lambda t}\end{aligned} \tag{12}$$

$$\boldsymbol{C}(t_0, t_0) = \boldsymbol{C}'_m(t_0, t_0) = \mathbf{0} \tag{13}$$

Note that these require the evaluation of the gradients for the utility and cost functions. The total costs are calculated by weighting by the initial state probabilities, such that $\text{Costs}(t_0, t) = \boldsymbol{\pi}(t_0, t_0)^T \boldsymbol{C}(t_0, t)$, with gradient $\text{Costs}'_m(t_0, t) = \boldsymbol{\pi}(t_0, t_0)^T \boldsymbol{C}'_m(t_0, t)$. We can also consider incremental cost-effectiveness ratios, estimated by

$$\text{ICER}(t_0, t|\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{\text{Costs}(t_0, t|\boldsymbol{x}_1) - \text{Costs}(t_0, t|\boldsymbol{x}_2)}{\text{QALY}(t_0, t|\boldsymbol{x}_1) - \text{QALY}(t_0, t|\boldsymbol{x}_2)}$$

The gradient of the log of the ICER is

$$\begin{aligned}\frac{\partial}{\partial \beta_m} \log(\text{ICER}(t_0, t|\boldsymbol{x}_1, \boldsymbol{x}_2))\Big|_{\beta_m = \hat{\beta}_m} = &\frac{\text{Costs}'_m(t_0, t|\boldsymbol{x}_1) - \text{Costs}'_m(t_0, t|\boldsymbol{x}_2)}{\text{Costs}(t_0, t|\boldsymbol{x}_2) - \text{Costs}(t_0, t|\boldsymbol{x}_1)} - \\ &\frac{\text{QALY}'_m(t_0, t|\boldsymbol{x}_1) - \text{QALY}'_m(t_0, t|\boldsymbol{x}_2)}{\text{QALY}(t_0, t|\boldsymbol{x}_1) - \text{QALY}(t_0, t|\boldsymbol{x}_2)}\end{aligned}$$

Finally, we can expand the equations to record the utilities or costs for different states. We can extend Equations (**??**) and (**??**) to

$$\begin{aligned}\frac{d\tilde{\boldsymbol{U}}(t_0,t)}{dt} &= \left(\boldsymbol{P}(t_0,t) \circ \left(\mathbf{1}\,\boldsymbol{u}(t)^T\right)\right) e^{-\lambda t} \\ \frac{d\tilde{\boldsymbol{C}}(t_0,t)}{dt} &= \left(\boldsymbol{P}(t_0,t) \circ \left(\mathbf{1}\,\boldsymbol{c}(t)^T + \mathbf{1}\left((\boldsymbol{Q}(t) \circ \boldsymbol{\mathcal{C}}(t))\,\mathbf{1}\right)^T\right)\right) e^{-\lambda t}\end{aligned}$$

for matrices $\tilde{\boldsymbol{U}}$ and $\tilde{\boldsymbol{C}}$, where $\tilde{U}_{ij}$ is the cumulative utility for state j given an initial state i, and $\tilde{C}_{ij}$ is the cumulative cost for being in state j or a transition from state j given an initial state i.

Example 6: Transformations

For causal inference and health economic evaluations, it is often useful to combine evidence from different data sources to transform the baseline transitions. We introduce five such transformations. First, we can turn off a transition, where the transformed hazard $h^*(t|\boldsymbol{x}) = 0 \times h(t|\boldsymbol{x})$. One advantage of this zero-model formulation is that we can keep the same variance-covariance matrix for the baseline transitions, which supports a straightforward comparison between the interventions. Second, we may have a hazard ratio $\exp(\phi)$ with 95% confidence interval $(\exp(\phi_l), \exp(\phi_u))$, such as from a meta-analysis. The variance for the log hazard ratio is

$((\phi_u - \phi_l)/2/1.96)^2$. The transformed hazard is $h^*(t|\boldsymbol{x}, \phi) = \exp(\phi)h(t|\boldsymbol{x})$, with the gradient of the hazard h^* with respect to the parameters of $h(t|\boldsymbol{x})$ being $\exp(\phi)\nabla h(t|\boldsymbol{x})$, and the gradient with respect to ϕ being $\exp(\phi)h(t|\boldsymbol{x})$. Third, we may have an acceleration factor $\exp(\phi)$. Then the transformed hazard is $h^*(t|\boldsymbol{x}, \phi) = \exp(\phi)h(\exp(\phi)t|\boldsymbol{x})$, with the gradient of the hazard h^* with respect to the parameters of $h(t\exp(\phi)|\boldsymbol{x})$ being $\exp(\phi)\nabla h(t\exp(\phi)|\boldsymbol{x})$, and the gradient with respect to ϕ being $t\exp(2\phi)h'(t\exp(\phi)|\boldsymbol{x}) + \exp(\phi)h(t\exp(\phi)|\boldsymbol{x})$. Fourth, we can add transition hazards together, such that the transformed hazard is $h^*(t|\boldsymbol{x}) = h_1(t|\boldsymbol{x}) + h_2(t|\boldsymbol{x})$. The gradient of the sum is equal to the sum of the gradients. Fifth, we can include smooth mathematical functions with no uncertainty. A useful example of this would be to use a spline interpolation function for the log of background mortality rates from vital statistics for use in relative survival (or excess hazards) modelling. The spline interpolation has the advantage of being smooth for the ordinary differential equation solver.

Variance and interval estimation

For an estimator $\boldsymbol{\phi}(t_0, t)$ with support on the real line, a confidence interval can be calculated from the variance-covariance matrix $\boldsymbol{V} = \boldsymbol{\phi}'_m(t_0, t)\,\hat{\boldsymbol{\Sigma}}\,\boldsymbol{\phi}'_m(t_0, t)$ and normal quantiles z at the α level with bounds $\boldsymbol{\phi}(t_0, t) \pm z_{(1-\frac{\alpha}{2})}\sqrt{\operatorname{diag}(\boldsymbol{V})}$. For estimators that are on the open interval $(0, 1)$, the gradient can be calculated using an identity or logit transformations. Similarly, for estimators that are on the open interval $(0, \infty)$, the gradient can be calculated using an identity of log transformations.

Integration of hazards that are functions of log(time)

Many parametric survival models are implemented in terms of log(time), including flexible parametric survival models and accelerated failure time models. Integration of the ordinary differential equation solvers from the origin for such models can lead to numerical issues. We offer two approaches to address these issues. First, we can truncate small values for time, such that $t^* = \max(t, \epsilon)$, e.g. using ϵ =1e-5. Defining cumulative hazards as $H_{ij}(t_0, t) = \int_{t_0}^{t} Q_{ij}(u)du$, the value for $H_{ij}(0, \epsilon)$ may be poorly estimated by $\epsilon Q_{ij}(\epsilon)$. Moreover, the hazards will generally not be smooth at $t = \epsilon$.

Second, we could directly calculate initial values at $t = \epsilon$ using cumulative hazard estimates. Let the matrix of cumulative hazard intensities from 0 to ϵ be $\boldsymbol{H}(t_0, t) = (H_{ij}(t_0, t))$ and let $\boldsymbol{H}'_m(t_0, t) = \int_{t_0}^{t} \boldsymbol{Q}'_m(u)du$. The transition probabilities at ϵ can be calculated approximately by

$$
\begin{aligned}
\boldsymbol{P}(0, \epsilon) &\approx \operatorname{mexp}(\boldsymbol{H}(0, \epsilon)) \\
\boldsymbol{P}'_m(0, \epsilon) &\approx \boldsymbol{H}'_m(0, \epsilon) \\
\boldsymbol{L}(0, \epsilon) &\approx \epsilon(\boldsymbol{I} + \boldsymbol{P}(0, \epsilon))/2 \\
\boldsymbol{L}'_m(0, \epsilon) &\approx \epsilon\boldsymbol{P}'_m(0, \epsilon)/2 \\
\boldsymbol{U}(0, \epsilon) &\approx \epsilon\boldsymbol{P}(0, \epsilon)\boldsymbol{u}(\epsilon)e^{-\lambda\epsilon}/2 \\
\boldsymbol{U}'_m(0, \epsilon) &\approx \epsilon(\boldsymbol{P}'_m(0, \epsilon)\boldsymbol{u}(\epsilon) + \boldsymbol{P}(0, \epsilon)\boldsymbol{u}'_m(\epsilon))e^{-\lambda\epsilon}/2
\end{aligned}
$$

where $\operatorname{mexp}(\boldsymbol{M})$ is the matrix exponential for a matrix $\boldsymbol{M}$.

3. Implementation

We have implemented the algorithm in R as the `rstpm2::markov_msm` function. The current implementation allows for: independent models for each of the transitions; a rich set of models

for each transitions (see Appendix B); predictions for state occupation and transition intensities, length of stay, utilities and costs; and useful post-processing facilities, including weighted standardisation, differences and ratios.

4. Simulations

To assess the small sample properties compared with the asymptotics properties, we undertook several simulations. We fitted a Markov illness-death model with Weibull transition rates with a shape parameter of 1.5 and a scale parameter of 10 for the transitions Healthy → Illness, Healthy → Death, and Illness → Death. We simulated for 1000 individuals with censoring min(20,Uniform(0,30)). We fitted the transitions using Weibull regression using the `aftreg` model from the **eha** package. The expected state occupation probabilities were predicted using the ODE solver with the true parameters. For each simulation, we predicted the state occupation probabilities, bias, confidence intervals using five transformations (plain confidence intervals, and confidence intervals based on log-log, log, logit and arcsin transformations), and coverage. We then calculated the mean bias, mean squared error and coverage across 1000 simulations.

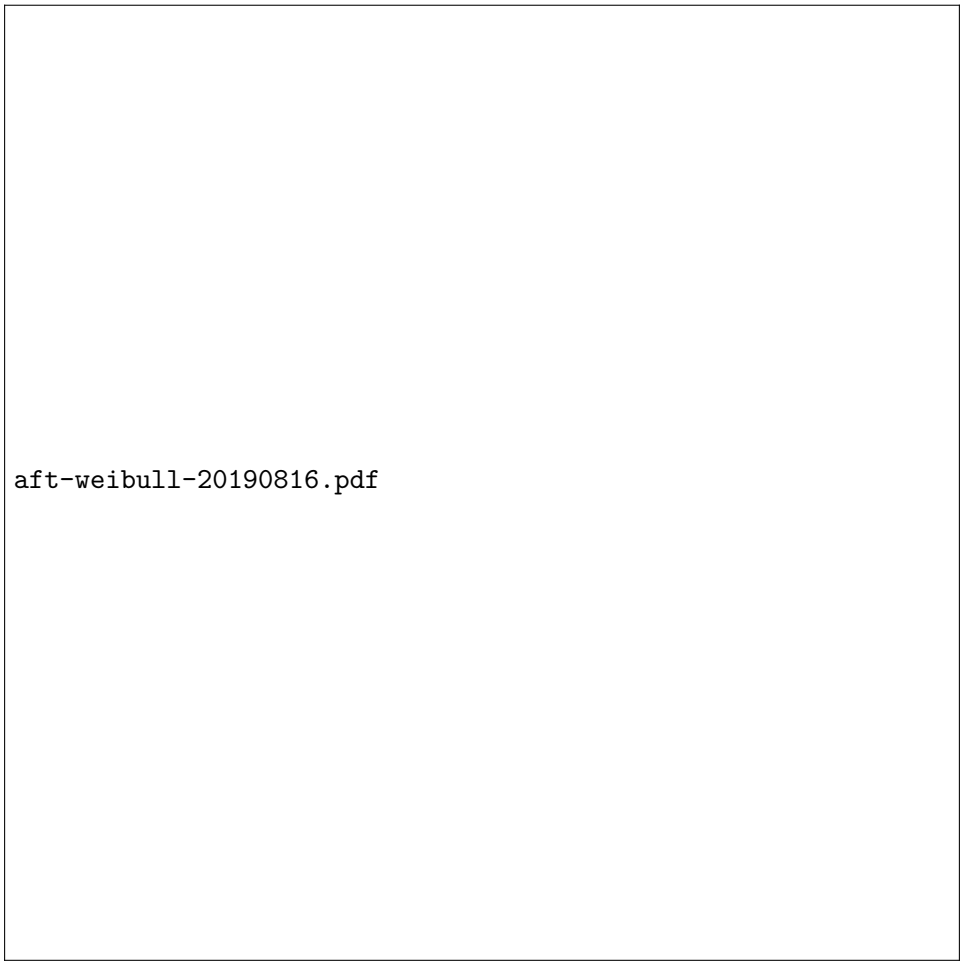

Figure 2: State occupation coverage for a Markov illness-death model with Weibull-distributed times, fitted using Weibull regression models. The panels represent confidence intervals based on different transformation methods.

Across the 1000 simulations, the mean bias for the state occupation probabilities for any of the three states across time varied between -0.0006 and 0.0008, while the mean squared error

varied between 0 and 0.0002. Coverage was generally close to 0.95 for each of the transformation methods (Figure **??**), particularly given that the binomial variability at 950/1000 has a 95% confidence interval of (0.945,0.963). The "plain" untransformed approach performs poorly close to start for the initial state.

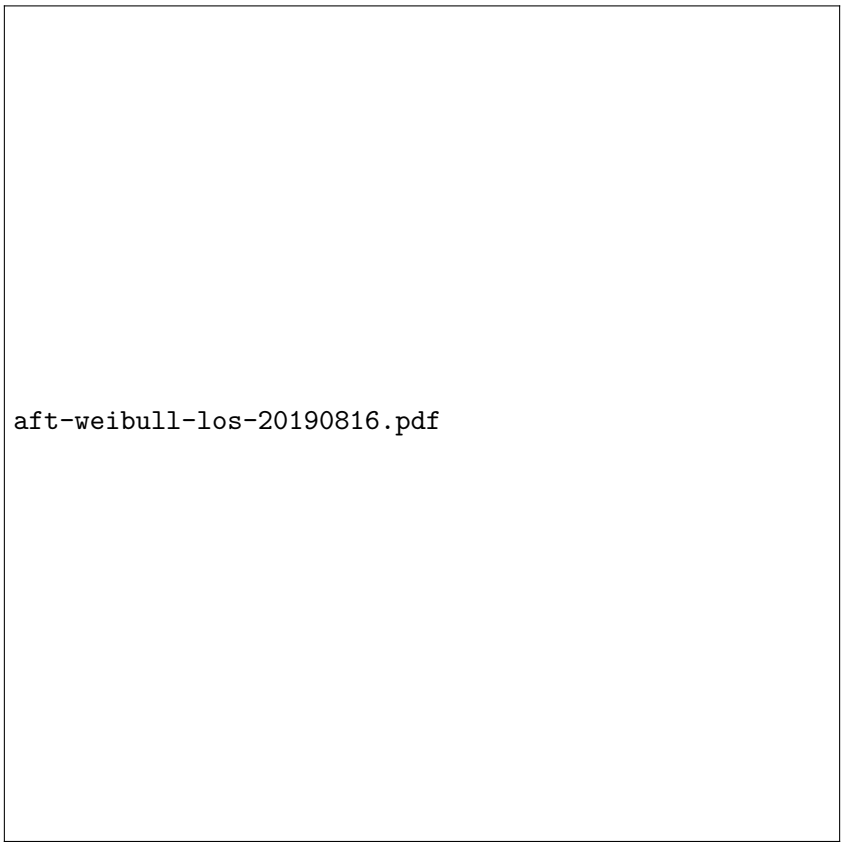

Figure 3: Length of stay coverage for a Markov illness-death model with Weibull-distributed times, fitted using Weibull regression models. The panels represent confidence intervals based on different transformation methods.

Across the 1000 simulations, the mean bias for the length of stay for any of the three states across time varied between -0.006 and 0.008 , while the mean squared error varied between 0 and 0.004. Coverage was generally close to 0.95 for both of the transformation methods (Figure **??**).

5. Example

We will extend an analysis by **?** of the Rotterdam Breast Cancer Data (**??**). This dataset includes patients who were treated for primary breast cancer in Rotterdam during 1978–1993. Treatment included primary surgery, with either mastectomy or breast conserving therapy, with possible referral for radiation treatment within three months of surgery. Study exclusion criteria included: (i) patient tissue based on biopsy only; (ii) metastatic disease at primary surgery or within one month of surgery; (iii) relapse or residual disease within one month of surgery; (iv) relapse prior to referral to radiation therapy; or (v) a previous primary cancer. For the analysis, patients were also excluded if they had adjuvant treatment but were node-negative, or if they had missing information on the number of positive nodes. After these exclusions, data were available for 2982 patients.

Following the analysis by **?**, we model for three states, including (1) the initial *post-surgery*

state, (2) *relapse* and (3) *death*, with transitions (1)→(2) for death from the post-surgery state with hazard $h_1(t)$, (1)→(3) for relapse with hazard $h_2(t)$, and (2)→(3) for death from the relapse state with hazard $h_3(t)$. All three transitions are modelled using time since surgery as the primary time scale. Crowther and Lambert used flexible parametric survival models on the *log cumulative hazard scale*, where $H_j(t|\boldsymbol{x}) = \int_0^t h_j(u|\boldsymbol{x})du = \exp(s_j(t) + \eta_j(t, \boldsymbol{x}))$ for baseline $s_j(t)$ and linear predictors $\eta_j(t, \boldsymbol{x})$ defined by

$$\begin{aligned}
\eta_1(t, \boldsymbol{x}) &= \beta_1\texttt{age} + \beta_2\texttt{nodes} + \beta_3\texttt{hormon} + I(20 < \texttt{size} \leq 50)s_4(t) + \\
&\quad I(\texttt{size} > 50)s_5(t) + I(\texttt{pr_1})s_6(t) \\
\eta_2(t, \boldsymbol{x}) &= \beta_1\texttt{age} + \beta_2\texttt{nodes} + \beta_3\texttt{hormon} + \\
&\quad \beta_4 I(20 < \texttt{size} \leq 50) + \beta_4 I(\texttt{size} > 50) + \beta_5 I(\texttt{pr_1}) \\
\eta_3(t, \boldsymbol{x}) &= \beta_1\texttt{age} + \beta_2\texttt{nodes} + \beta_3\texttt{hormon} + \\
&\quad \beta_4 I(20 < \texttt{size} \leq 50) + \beta_4 I(\texttt{size} > 50) + I(\texttt{pr_1})s_7(t)
\end{aligned}$$

where `age` is the age in years at cancer treatment, `nodes` is the number of positive lymph nodes, `hormon` is an indicator for whether the patient was on hormonal therapy, `size` is the tumour size (mm), `pr_1` is the log of one plus the progesterone level (fmol/L), $I(p)$ is an indicator function with value 1 when the predicate p is true and value 0 otherwise, $s_0(t) = \sum_{j=0}^{3} B_j(\log(t))\beta_{0j}$ for a natural splines basis function $B_j()$ with knots at the quantiles for log of the events times, and $s_k(t) = \beta_{k0} + \beta_{k1}\log(t)$ for $k > 1$. The parameters are assumed to be distinct between the linear predictors. The hazards are calculated from the cumulative hazard using $h_j(t|\boldsymbol{x}) = H'_j(t|\boldsymbol{x})$. Note that ratios of these hazards may not have a simple interpretation, particularly with continuous or multiple time-varying effects (see Appendix). Moreover, age is expected to be closely related to the mortality rates and modelling using a single parameter may not capture this important variation. For ease of comparison, we do not further investigate changes in the model formulations and focus on predictions from these fitted models.

For predictions, Crowther and Lambert used *conditional* predictions for a patient aged 54 years with a transformed progesterone level of 3 (that is, progesterone = $\exp(3) - 1 = 19.1$ fmol/L), with the number of positive lymph nodes varying between 0, 10 and 20, and with each of the three levels for tumour size. We are able to model and predict for conditional relationships in less than 20 seconds (code included in the documentation for the `rstpm2` package on CRAN). We now focus on *marginal* or *standardised* predictions for patients aged 50–59 years. In particular, we will compare state occupation probabilities and length of stay under counterfactual tumour sizes.

Let the subjects aged 50–59 years be indexed by $k = 1 \ldots K$ and let X represent tumour size and $\boldsymbol{Z}$ represent the other covariates. Then the predictions under the counterfactual $\hat{X} = x$ are modelled by

$$\begin{aligned}
P_{1j}(0, t|\hat{X} = x) &= E_{\boldsymbol{Z}}(P_{1j}(0, t|\boldsymbol{Z}, \hat{X} = x)) \\
&= \frac{1}{K}\sum_{k=1}^{K} P_{1j}(0, t|\boldsymbol{Z} = \boldsymbol{z}_k, \hat{X} = x) \\
L_{1j}(0, t|\hat{X} = x) &= E_{\boldsymbol{Z}}(L_{1j}(0, t|\boldsymbol{Z}, \hat{X} = x))
\end{aligned}$$

The specific counterfactuals are that all patients have a tumour size that is either (i) ≤ 20 mm, (ii) > 20 mm and ≤ 50 mm, or (iii) > 50 mm. The age restriction is due to the strong age dependence for each of the transitions. Differences in state occupation probabilities and length of stay by the counterfactuals are show in Figures **??** and **??**, respectively. There is clear evidence that smaller tumour sizes are associated with fewer early relapses and a lower risk of death for fifteen years after treatment.

Figure 4: Differences in standardised state occupation probabilities by state and by contrasts for counterfactual tumour size, for female breast cancer patients aged 50–59 years and diagnosed 1978–1993, Rotterdam

We also explored other model formulations, including penalised log-hazard models ($h_j(t|\boldsymbol{x}) = \exp(s_j(t) + \eta_j(t, \boldsymbol{x}))$) and accelerated failure time models ($S_j(t|\boldsymbol{x}) = S_{0j}(\int_0^t \exp(\eta_j(t, \boldsymbol{x}))du)$, where $S_{0j}(t) = \exp(-\exp(s_j(t)))$).

6. Discussion

In summary, we describe how to predict from Markov multi-state models with smooth transition intensities using ordinary differential equations. A variety of predictions can be estimated, together with interval estimation based on the multivariate delta method. The method is suitable for a range of models, including Poisson regression, parametric and flexible accelerated failure time models, and parametric and penalised generalised survival models. Applications of these methods could range from descriptive epidemiology, to causal inference, through to health economic evaluations of cost-effectiveness.

The recent article by **?** provides a non-parametric analogue to our development. Those non-parametric methods can readily be extended to many of the estimators described herein, including standardisation, quality-adjusted life-years and costs. The asymptotic theory for the non-parametric approach has been shown to hold for Aalen's additive hazards model; it is unclear whether non-parametric estimators for accelerated failure and proportional hazards

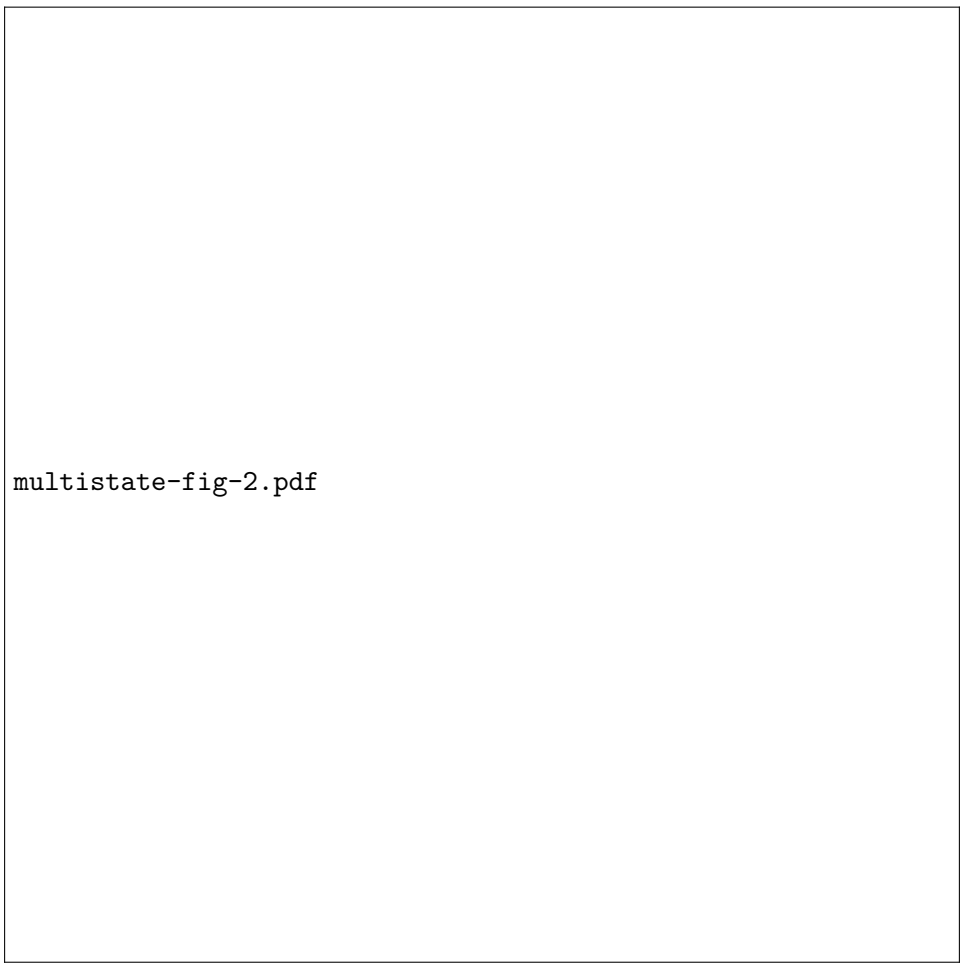

Figure 5: Differences in standardised length of stay by state and by contrasts for counterfactual tumour size, for female breast cancer patients aged 50–59 years and diagnosed 1978–1993, Rotterdam

models will satisfy the assumptions in Theorems 1 and 2 of **?**. Moreover, it is arguable when an additive hazards scale is suitable for covariate adjustment, for example, for modelling all-cause or cause-specific survival with age at cancer diagnosis as a covariate. We suggest that (smooth) accelerated failure time models may provide a useful alternative regression model framework which is also collapsible.

Under the conditions described by **?**, predictions that are functions of the state occupation probability are, under suitable regularity assumptions, expected to be consistently estimated irrespective of whether the Markov assumption holds. However other predictions (e.g. transition probabilities, or the proportion to ever pass through a state) may not be consistently estimated. Moreover, the efficiency of the predictions when the time scale has been mis-specified may be low. Alternative approaches include individual-based simulations, with variance estimation using the bootstrap **?**, which are expected to be computationally expensive.

Appendix A: Time-varying effects for log cumulative hazards models

For modelling on the log cumulative hazard scale with a linear predictor $\eta(t, \boldsymbol{x}_i)$ given time t

and covariates $\boldsymbol{x}_i$, the cumulative hazard $H(t|\boldsymbol{x}_i)$ can be represented by

$$\begin{aligned} H(t|\boldsymbol{x}_i) &= \exp(\eta(t, \boldsymbol{x}_i)) \\ \implies h(t|\boldsymbol{x}_i) &= \exp(\eta(t, \boldsymbol{x}_i)) \frac{\partial}{\partial t} \eta\left(t, \boldsymbol{x}_i\right) \end{aligned}$$

where $h(t|\boldsymbol{x}_i)$ is the hazard. The *hazard ratio* comparing covariates $\boldsymbol{x}_1$ and $\boldsymbol{x}_0$ is then

$$\frac{h(t|\boldsymbol{x}_1)}{h(t|\boldsymbol{x}_0)} = \exp(\eta(t, \boldsymbol{x}_1) - \eta(t, \boldsymbol{x}_0)) \left(1 + \frac{\frac{\partial}{\partial t}\left(\eta\left(t, \boldsymbol{x}_1\right) - \eta\left(t, \boldsymbol{x}_0\right)\right)}{\frac{\partial}{\partial t}\eta\left(t, \boldsymbol{x}_0\right)}\right) \tag{14}$$

We have the following cases for the interpretation of the hazard ratios:

Case 1 *Time-independent effects*: if $\frac{\partial}{\partial t}\left(\eta\left(t, \boldsymbol{x}_1\right) - \eta\left(t, \boldsymbol{x}_0\right)\right) = 0$, then Equation (**??**) gives the ratio $\exp(\eta(t, \boldsymbol{x}_1) - \eta(t, \boldsymbol{x}_0))$. This case will hold when the difference in the linear predictors is independent of t. Usefully, the log hazard ratio is then the difference in the linear predictors, which has a straightforward interpretation.

A sufficient condition is that the covariates that change between $\boldsymbol{x}_1$ and $\boldsymbol{x}_0$ are independent of t. As a specific example, for a linear predictor $\eta(t, \boldsymbol{x}_i) = s_0(t) + \eta_0(\boldsymbol{x}_i)$ for baseline $s_0(t)$ and linear predictor $\eta_0(\boldsymbol{x}_i)$, then the log hazard ratio equals $\eta_0(\boldsymbol{x}_1) - \eta_0(\boldsymbol{x}_0)$, which is a *proportional hazards* model.

Case 2 *Small time-varying effects*: if $\frac{\partial}{\partial t}\left(\eta\left(t, \boldsymbol{x}_1\right) - \eta\left(t, \boldsymbol{x}_0\right)\right) \ll \frac{\partial}{\partial t}\eta\left(t, \boldsymbol{x}_0\right)$, then $\frac{h(t|\boldsymbol{x}_1)}{h(t|\boldsymbol{x}_0)} \approx \exp(\eta(t, \boldsymbol{x}_1) - \eta(t, \boldsymbol{x}_0))$. For this case, the partial derivative with respect to time for the change in the linear predictor is small compared with partial derivative for the linear predictor for $\boldsymbol{x}_0$. This suggests that the interpretation for moderately small time-varying effects will be straightforward (up to an approximation).

Case 3 *Stratified model*: if we have a stratified linear predictor $\eta(t, \boldsymbol{x}_i, j) = s_j(t) + \eta_j(\boldsymbol{x}_i)$ for stratum j with baseline $s_j(t)$ and linear predictor $\eta_j(\boldsymbol{x}_i)$, then

$$\frac{h(t|\boldsymbol{x}_1, j)}{h(t|\boldsymbol{x}_0, k)} = \exp\left(s_j(t) + \eta_j(\boldsymbol{x}_1) + \log(s_j'(t)) - \left(s_k(t) + \eta_k(\boldsymbol{x}_0) + \log(s_k'(t))\right)\right) \tag{15}$$

If $j = k$ then $\frac{h(t|\boldsymbol{x}_1, j)}{h(t|\boldsymbol{x}_0, k)} = \exp(\eta_j(\boldsymbol{x}_1) - \eta_j(\boldsymbol{x}_0))$, which is proportional hazards. If $j \neq k$, then Equation (**??**) can be interpreted as having separable effects (on the log hazard scale) for the time effects and for the other covariates.

Case 4 *Linear time-varying effect*: if we have a linear predictor $\eta(t, x_i, \boldsymbol{u}_i) = s_0(t) + x_i s_1(t) + \eta_0(\boldsymbol{v}_i)$ for a scalar x_i and a vector of other covariates $\boldsymbol{v}_i$, then

$$\frac{h(t|x_1, \boldsymbol{v}_1)}{h(t|x_0, \boldsymbol{v}_0)} = \exp\left((x_1 - x_0)s_1(t) + \eta_0(\boldsymbol{v}_1) - \eta_0(\boldsymbol{v}_0)\right) \left(1 + \frac{(x_1 - x_0)s_1'(t)}{s_0'(t) + x_0 s_1'(t)}\right)$$

From the ratio $(x_1 - x_0)s_1'(t)/(s_0'(t) + x_0 s_1'(t))$, we see that the time-varying hazard ratio depends on both the baseline value x_0 and the difference $x_1 - x_0$. The interpretation of the effect will be straightforward when x_i is for a binary indicator (e.g. $x_0 = 0$ and $x_1 = 1$).

Case 5 *Multiple time-varying effects*: if we have a linear predictor $\eta(t, x_i, \boldsymbol{u}_i) = s_0(t) + x_i s_1(t) + u_i s_2(t) + \eta_0(\boldsymbol{v}_i)$ for scalars x_i and u_i and a vector of other covariates $\boldsymbol{v}_i$, then

$$\begin{aligned} \frac{h(t|x_1, u_1, \boldsymbol{v}_1)}{h(t|x_0, u_0, \boldsymbol{v}_0)} = \exp\left((x_1 - x_0)s_1(t) + (u_1 - u_0)s_2(t) + \eta_0(\boldsymbol{v}_1) - \eta_0(\boldsymbol{v}_0)\right) \times \\ \left(1 + \frac{(x_1 - x_0)s_1'(t) + (u_1 - u_0)s_2'(t)}{s_0'(t) + x_0 s_1'(t) + u_0 s_2'(t)}\right) \end{aligned}$$

From the ratio in the last line, we see that the hazard ratio depends on both the baseline values x_0 and u_0 and the differences $x_1 - x_0$ and $u_1 - u_0$. The multiple effects will be straightforward to interpret if x_i and u_i are binary indicators for strata (see Case 3).

In summary, time-independent effects and stratified models have a straightforward interpretation, however continuous time-varying effects and multiple time-varying effects on a log cumulative hazard scale are more difficult to interpret in terms of hazard ratios.

Appendix B: Hazard specifications

Class	Model	R function	Specification[a]
Parametric	Poisson regression	`stats::glm`	$\log(h(t\|\boldsymbol{x})) = \eta(t, \boldsymbol{x})$
	Accelerated failure time	`flexsurv::flexsurvreg`	$S(t\|\boldsymbol{x}) = S_0(\exp(\eta(\boldsymbol{x}))t)$
		`eha::aftreg`	$S(t\|\boldsymbol{x}) = S_0(\exp(\eta(\boldsymbol{x}))t)$
		`rstpm2::aft`	$S(t\|\boldsymbol{x}) = S_0(\int_0^t \exp(\eta(u, \boldsymbol{x}))du)$
	Generalized survival	`flexsurv::flexsurvspline`	$\log(H(t\|\boldsymbol{x})) = \eta(t, \boldsymbol{x})$
		`rstpm2::stpm2`	$\log(H(t\|\boldsymbol{x})) = \eta(t, \boldsymbol{x})$
Penalised	Poisson regression	`mgcv::gam`	$\log(h(t\|\boldsymbol{x})) = \eta(t, \boldsymbol{x})$
	Log-hazard	`survPen::survPen`	$\log(h(t\|\boldsymbol{x})) = \eta(t, \boldsymbol{x})$
	Generalised survival	`rstpm2::pstpm2`	$\log(H(t\|\boldsymbol{x})) = \eta(t, \boldsymbol{x})$
Transformation	Zero	`rstpm2::zeroModel`	$h^*(t\|\boldsymbol{x}) = 0$
	Hazard ratio	`rstpm2::hrModel`	$h^*(t\|\boldsymbol{x}, \phi) = h(t\|\boldsymbol{x})\phi$
	Accelerated failure	`rstpm2::aftModel`	$h^*(t\|\boldsymbol{x}, \phi) = h(\phi t\|\boldsymbol{x})\phi$
	Additive models	`rstpm2::addModel`	$h^*(t\|\boldsymbol{x}) = h_1(t\|\boldsymbol{x}) + h_2(t\|\boldsymbol{x})$
	Hazard function	`rstpm2::hazFun`	$h(t\|\boldsymbol{x})$
	Spline interpolation	`rstpm2::splineFun`	$h(t\|\boldsymbol{x}) = \exp(s(t))$

[a]Notes: $\eta(t, \boldsymbol{x})$ and $\eta(\boldsymbol{x})$ are linear predictors for time t and covariates $\boldsymbol{x}$, and $S_0(t)$ is a baseline survival model.

Table 1: Hazard specifications

For Table **??**, the generality of the linear predictor may be constrained by the model class; For example, `flexsurv::flexsurvspline` only allows for spline interactions between time and covariates. The baseline survival models for `flexsurv::flexsurvreg` and `eha::aftreg` are from a parametric family, while the baseline survival for `rstpm2::aft` is where the log cumulative hazards are based on splines. The generalised survival models also allow for other transformations, including proportional odds, probit and additive hazards models. Arbitrary smooth hazards can be defined using `hazFun`; as an example, a spline interpolation for log-hazards is defined using `splineFun`. The transformation functions at the end of the table are meant to take one or more models as inputs and predict rates that are functions of the given models. Formally, Poisson regression models are models for rates rather than for hazards.

Appendix C: Smooth accelerated failure time models

The smooth accelerated failure time models implemented by `rstpm2::aft` and the Stata command `staft` have not been described elsewhere. Survival for these models at time t with covariates $\boldsymbol{x}$ is defined by

$$S(t|\boldsymbol{x}) = S_0\left(\int_0^t \exp(\eta(u, \boldsymbol{x}))\mathrm{d}u\right)$$

where $S_0(t) = \exp\left(-\exp\left(s(\log(t))\right)\right)$ is baseline survival, $s(u)$ is a smooth function with support on the real line (e.g. natural splines), and $\exp(\eta(u, \boldsymbol{x}))$ is a time-varying acceleration factor. We can replace the integration for the acceleration factor by a cumulative function $\exp(\eta_1(\log(t), \boldsymbol{x}))$, such that

$$\begin{aligned}\int_0^t \exp(\eta(u, \boldsymbol{x}))\mathrm{d}u &= \exp(\eta_1(\log(t), \boldsymbol{x})) \\ \implies \eta(t, \boldsymbol{x}) &= \eta_1(\log(t), \boldsymbol{x}) + \log\left(\frac{\partial}{\partial t}\eta_1(\log(t), \boldsymbol{x})\right)\end{aligned}$$

such that we can calculate the time-varying acceleration factor by differentiation. The cumulative linear predictor $\eta_1(\log(t), \boldsymbol{x})$ can be defined using a smooth function of log time (e.g. using natural splines). We can calculate the hazards by

$$\begin{aligned}h(t|\boldsymbol{x}) &= \frac{\mathrm{d}}{\mathrm{d}t}\exp\left(s(\eta_1(\log(t), \boldsymbol{x}))\right) \\ &= \exp\left(s(\eta_1(\log(t), \boldsymbol{x}))\right) s'(\eta_1(\log(t), \boldsymbol{x}))\frac{\partial}{\partial t}\eta_1(\log(t), \boldsymbol{x})\end{aligned}$$

This model has been implemented in R and Stata for left truncated and right censored data.

Affiliation:

Mark Clements
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Email: mark.clements@ki.se