Scandinavian Journal of Statistics Theory and Applications

Scandinavian Journal of Statistics, Vol. 42: 609–626, 2015
doi: 10.1111/sjos.12125

Non-parametric Bayesian Hazard Regression for Chronic Disease Risk Assessment

OLLI SAARELA
Dalla Lana School of Public Health, University of Toronto

ELJA ARJAS
Department of Mathematics and Statistics, University of Helsinki and National Institute for Health and Welfare

ABSTRACT. Assessing the absolute risk for a future disease event in presently healthy individuals has an important role in the primary prevention of cardiovascular diseases (CVD) and other chronic conditions. In this paper, we study the use of non-parametric Bayesian hazard regression techniques and posterior predictive inferences in the risk assessment task. We generalize our previously published Bayesian multivariate monotonic regression procedure to a survival analysis setting, combined with a computationally efficient estimation procedure utilizing case–base sampling. To achieve parsimony in the model fit, we allow for multidimensional relationships within specified subsets of risk factors, determined either on *a priori* basis or as a part of the estimation procedure. We apply the proposed methods for 10-year CVD risk assessment in a Finnish population.

Key words: case–base sampling, disease prediction, monotonic regression, non-parametric Bayesian regression, risk assessment

1. Introduction

In this paper, we propose a new statistical methodology for chronic disease risk assessment and apply it in the context of cardiovascular disease (CVD). Our approach is Bayesian, and it applies well-known ideas from posterior predictive inference. Despite the extensive literature on developing and validating prognostic models, the prediction problem is only rarely formulated or addressed from this natural perspective. Here, we first review the theoretical framework needed for the task of prediction, by adopting the concepts and the notation of counting processes. After this, we discuss some new practical approaches for fitting and validating non-parametric Bayesian risk models, emphasizing a symmetry aspect in the modelling of timescales and other predictors.

Most conventional risk models are based on the assumption of proportional hazards, that is, invariance of multiplicative covariate effects over time, usually expressed in terms of exponential relative risk functions. An advantage of this structure is the ability to obtain easily interpretable risk scores, formed as linear combinations of the individual risk factors (e.g. D'Agostino Sr *et al.*, 2008). Non-parametric approaches present a natural alternative when the main interest lies in the predictive accuracy of the method, rather than in individually interpretable parameter estimates. When considering the prediction problem from a Bayesian perspective, it becomes natural to directly employ the concept of posterior predictive distributions. In the context of non-parametric modelling, these predictive probabilities are obtained as expectations over realizations of the non-parametrically specified random hazard functions.

The main computational challenge in fitting non-parametric models to survival data is due to the presence of the continuous-time parameter. Numerical evaluation of the likelihood

expression requires the integration over time along the realizations of the non-parametrically specified hazard function. In order to simplify the task of model fitting and to speed up the computation, we utilize the fact that any hazard model can be approximately represented in terms of a logistic regression model for a collection of indicator variables, which are directly linked to the outcome of interest (e.g. Arjas & Haara, 1987). This approach does not in itself reduce the computational cost, owing to the large proportion of these indicator variables being zero. Computational gains can, however, be achieved when the approach is used in conjunction with the method of case–base sampling (Hanley & Miettinen, 2009). We present this approach as a particular case of conditional likelihood inference, which then implies its applicability also for Bayesian inference.

In principle, any non-parametric model construction for binary responses can be coupled with the case–base sampling approach for estimation. Herein, we concentrate on adapting the multivariate monotonic regression procedure proposed by Saarela & Arjas (2011a) to the considered survival analysis setting. Because fully non-parametric estimation of all dependencies in a high-dimensional multivariate hazard regression model would be impossible in practice, we combine the consequent monotonic construction with the idea of 'packaging' suitably chosen subsets of covariates, as suggested by Arjas & Liu (1996). The covariates in each such package contribute then to the hazard regression via a non-parametrically defined multiplicative factor. The choice of such subsets can be determined either on an *a priori* basis or as a part of the estimation procedure itself. We illustrate these methods in the context of assessing the 10-year risk of CVD, using a few basic classic risk factors of CVD that are also included in the well-known Framingham score (D'Agostino Sr *et al.*, 2008). The assumption of monotonicity can be safely made in this context, as we are focusing on well-established risk factors of CVD.

We can define 'risk' as the probability of an adverse health-related event occurring within a specified time frame, given the individual-level prognostic profile (e.g. Miettinen, 2011, p. 25). This quantity is inherently unobservable, being the limiting relative frequency of the adverse events in an infinite sequence of exchangeable instances with the same prognostic profile (cf. Bernardo & Smith, 1994, p. 173). Because we do not actually have an infinite sequence of such observables, but only a finite one, the prediction problem becomes a posterior predictive one, involving a probability statement about a future observable given the past ones (cf. Bernardo & Smith, 1994, p. 175). We contend that addressing the prediction problem separately from its Bayesian formulation has contributed to the proliferation of the misnomer 'risk prediction', which has become deeply rooted in the epidemiological literature (what we are predicting is the occurrence of the disease event itself, not the risk).

The plan of the paper is as follows. In the following section, we first review the posterior predictive formulation of the problem. In Section 3, we discuss fitting non-proportional hazards models using logistic regression and case–base sampling. In Section 4, we discuss hazard model constructions based on packaging of covariates, while Section 5 contains some notes on numerical evaluation and validation of the risk estimates. We apply the methods to data from Finnish population-based cardiovascular cohorts in Section 6 and conclude with a discussion in Section 7.

2. Posterior predictive distributions

Let $\mathcal{C} \equiv \{1, \ldots, n\}$ denote a cohort of healthy individuals who are followed up for time-to-event outcomes, represented by counting processes $N_{ij}(t) \in \{0, 1\}, i \in \mathcal{C}, j = 0, 1, 2$, with $j = 0$ indicating type I censoring due to end of the follow-up period at time τ, $j = 1$ indicating an incident event of interest (fatal or non-fatal), and $j = 2$ indicating a death due

to a cause other than an event of interest. Type I censoring is here assumed for notational simplicity; however, everything below can be adapted to the general random censoring setting. Because we are only considering the first event of any type, by definition, other counting processes remain at zero after a jump in one of them, so that a jump of $N_i(t) \equiv \sum_{j=0}^{2} N_{ij}(t)$ terminates the follow-up. In addition to the outcomes, we observe realizations of vectors of predictors $X_i = (X_{i1}, \ldots, X_{ip})$, with the values recorded at the outset of the follow-up period. The observed information on individual i up to time t is denoted as $\mathcal{F}_{it} \equiv \sigma\{X_i, N_{ij}(u) : j = 0, 1, 2; 0 \leq u \leq t\}$, with all observed information given by $\mathcal{F} \equiv \bigvee_{i \in \mathcal{C}} \mathcal{F}_{i\tau}$.

By convention, we assume infinite exchangeability of the random vectors $(N_{i0}(t), N_{i1}(t), N_{i2}(t))$ over the individual indices i (cf. Section 1), conditional on the histories $\mathcal{F}_{it-}$, at any t. The objective is to obtain the posterior predictive probabilities $\pi_s(x_i) \equiv P(N_{i1}(s) = 1 \mid \mathcal{F}_{i0} \vee \mathcal{F})$ of an incident event occurring in the time interval $(0, s], s \leq \tau$, for a further exchangeable healthy individual $i \notin \mathcal{C}$. If age is used as the main timescale, the probability of interest is $\pi(b_i, x_i) \equiv P(N_{i1}(b_i + s) = 1 \mid \mathcal{F}_{ib_i} \vee \mathcal{F}), i \neq \mathcal{C}$, where b_i is the age at baseline for individual i. Assuming absolute continuity over time and using the exchangeability postulate, de Finetti's representation theorem (e.g. Bernardo & Smith, 1994, p. 180) gives for any $i \notin \mathcal{C}$

$$\begin{aligned} \pi_s(x_i) &= \int_{\phi} \pi_s(x_i, \phi) P(\mathrm{d}\phi \mid \mathcal{F}) \\ &= \int_{\phi} \int_{t \in (0,s]} \lambda_{i1}(t, \phi)\, \mathrm{d}t \prod_{u \in [0,t)} \left[1 - \sum_{j=1}^{2} \lambda_{ij}(u, \phi)\, \mathrm{d}u \right] P(\mathrm{d}\phi \mid \mathcal{F}), \end{aligned} \tag{1}$$

where

$$\pi_s(x_i, \phi) \equiv P(N_{i1}(s) = 1 \mid \mathcal{F}_{i0}, \phi) \text{ and } \lambda_{ij}(t, \phi)\, \mathrm{d}t \equiv P(\mathrm{d}N_{ij}(t) = 1 \mid \mathcal{F}_{it-}, \phi),$$

$\prod$ represents a product integral (Gill & Johansen, 1990) over time, and ϕ is a parameter vector specifying the cause-specific hazard functions. Herein, we concentrate on modelling approaches where the dimension of ϕ is not specified on an *a priori* basis. In the present problem, this is a natural approach, as we are not primarily interested in model parameters but in posterior predictive distributions. The required integration can then be carried out using Markov chain Monte Carlo (MCMC) methods, by drawing random samples from the posterior distribution

$$P(\mathrm{d}\phi \mid \mathcal{F}) \overset{\phi}{\propto} \prod_{i=1}^{n} \prod_{t \in (0,\tau]} \left(\prod_{j=1}^{2} \lambda_{ij}(t, \phi)^{\mathrm{d}N_{ij}(t)} \left[1 - \sum_{j=1}^{2} \lambda_{ij}(t, \phi)\, \mathrm{d}t \right]^{1 - N_i(t)} \right) P(\mathrm{d}\phi).$$

While the additivity of the preceding hazards always applies without further assumptions if the event definitions are mutually exclusive, additional conditional independence assumptions $\mathrm{d}N_{i1}(t) \perp\!\!\!\perp \phi_2 \mid (\mathcal{F}_{it-}, \phi_1)$ and $\mathrm{d}N_{i2}(t) \perp\!\!\!\perp \phi_1 \mid (\mathcal{F}_{it-}, \phi_2)$ are usually made for the purposes of estimation. These correspond to the non-informative, or non-innovative, censoring discussed by Arjas (1989) and, more generally, may be understood as local independence conditions (e.g. Schweder, 1970; Arjas & Haara, 1984). These assumptions enable partitioning of the parameter vector $\phi = (\phi_1, \phi_2)$ into components describing the events of interest and the competing causes of death, respectively. Thus, while the posterior predictive probability (1) is a function of both, under non-informative censoring, estimation of the two hazard functions can be carried out separately. It depends on the covariates x_i whether these assumptions are reasonable;

it may well be, for example, that the conditional independence applies only after conditioning on a latent individual frailty term. The predictive probability (1) (the cause-specific cumulative incidence function) takes into account competing causes of death and thus is directly relevant to real-life events. In contrast, as is well known (e.g. Kalbfleisch & Prentice, 2002, p. 252), a 'risk' based on considering only a single cause-specific hazard function in isolation would not have a direct probability interpretation, with the exception of the special case where the outcome of interest is all-cause mortality.

3. Fitting hazard models using logistic/multinomial regression

3.1. Formulation as conditional likelihood inference

Motivated by the arguments provided in Section 1, rather than expressing the likelihood in terms of the cumulative hazard (which of course remains an alternative and which we have discussed in Saarela & Arjas, 2011b), we specify our statistical model in terms of time-specific log-odds (Arjas & Haara, 1987, p. 3)

$$\begin{aligned} &\log\left(\frac{P(\Delta N_{ij}(t)=1 \mid \mathcal{F}_{it-})}{P(\Delta N_{ij}(t)=0 \mid \mathcal{F}_{it-})}\right) = \omega_j(t,x_i) \\ &\quad\Leftrightarrow\quad P(\Delta N_{ij}(t)=1 \mid \mathcal{F}_{it-}) = \frac{\exp\{\omega_j(t,x_i)\}}{1+\exp\{\omega_j(t,x_i)\}}, \end{aligned} \tag{2}$$

where $\Delta N_{ij}(t) \equiv N_{ij}(t) - N_{ij}(t-\Delta t)$, $\mathcal{F}_{it-} \equiv \sigma\{X_i, N_{ij}(u) : j = 0,1,2; 0 \le u \le t-\Delta t\}$ and $\omega_j(t,x_i)$ is a non-parametric regression function of time and the predictors on the log-odds scale. By splitting the time axis into bins of the form $(0,\Delta t], (\Delta t, 2\Delta t], \ldots$, realizations of the regression functions $\omega_j(t,x_i)$ could be drawn from the posterior distribution

$$\begin{aligned} P(\mathrm{d}\omega_j \mid \mathcal{F}) \overset{\omega_j}{\propto} \prod_{i=1}^{n}\prod_{k=1}^{\infty}\Bigg[&\left(\frac{\exp\{\omega_j(k\Delta t,x_i)\}}{1+\exp\{\omega_j(k\Delta t,x_i)\}}\right)^{\Delta N_{ij}(k\Delta t)} \\ &\times\left(\frac{1}{1+\exp\{\omega_j(k\Delta t,x_i)\}}\right)^{1-N_i(k\Delta t)}\Bigg] P(\mathrm{d}\omega_j). \end{aligned} \tag{3}$$

However, such splitting results in a very large number of contributions with $\Delta N_{ij}(t) = 0$, infinitely many in the limit when approaching the continuous-time setting, and this makes the likelihood (3) computationally impractical in approximating hazard models. This motivates the use of the case–base sampling approach of Hanley & Miettinen (2009). It operates in continuous time and is based on considering a finite sample of 'person-moments', a term that we can take to mean the observed history $\mathcal{F}_{it}$ at the 'person' coordinate i and the time, or 'moment', coordinate t. (Person-moments cannot be characterized as marked point processes because they are not locally finite in number; however, the sampled person-moments do constitute a marked point process.)

We introduce a counting process $M_i(t) \in \{0,1,2,\ldots\}$ for sampled person-moments contributed by individual i, with $\Delta M_i(t) = 1$ indicating that a person-moment was recorded for individual i in the interval $(t-\Delta t, t]$. To ascertain the two case series, each person-moment corresponding to a CVD event or death from another cause is selected, that is,

$$\begin{aligned} &P(\Delta M_i(t)=1 \mid \Delta N_{i1}(t)=1, \Delta N_{i2}(t)=0, \mathcal{F}_{it-}) \\ &\quad = P(\Delta M_i(t)=1 \mid \Delta N_{i1}(t)=0, \Delta N_{i2}(t)=1, \mathcal{F}_{it-}) = 1. \end{aligned}$$

The corresponding base series is ascertained by sampling m person-moments randomly (and independently of the eventual event history of the individual) from the study base, which here consists of the cohort $\mathcal{C}$ and its follow-up up to time τ (cf. Miettinen & Karp, 2012, p. 56). The sampling probabilities $P(\Delta M_i(t) = 1 \mid \Delta N_i(t) = 0, \mathcal{F}_{it-})$ can be allowed to depend on time and on the predictors. A formal justification to the case–base sampling approach to hazard modelling can be given by considering time-specific conditional likelihood contributions of the form

$$P(\Delta N_{i1}(t) = 1, \Delta N_{i2}(t) = 0 \mid \Delta M_i(t) = 1, \mathcal{F}_{it-}) = \frac{P(\Delta N_{i1}(t) = 1 \mid \mathcal{F}_{it-})}{P(\Delta M_i(t) = 1 \mid \mathcal{F}_{it-})} \tag{4}$$

$$P(\Delta N_{i1}(t) = 0, \Delta N_{i2}(t) = 1 \mid \Delta M_i(t) = 1, \mathcal{F}_{it-}) = \frac{P(\Delta N_{i2}(t) = 1 \mid \mathcal{F}_{it-})}{P(\Delta M_i(t) = 1 \mid \mathcal{F}_{it-})} \tag{5}$$

and

$$\begin{aligned} &P(\Delta N_{i1}(t) = 0, \Delta N_{i2}(t) = 0 \mid \Delta M_i(t) = 1, \mathcal{F}_{it-}) \\ &= \frac{P(\Delta M_i(t) = 1 \mid \Delta N_i(t) = 0, \mathcal{F}_{it-}) \left[1 - \sum_{j=1}^{2} P\left(\Delta N_{ij}(t) = 1 \mid \mathcal{F}_{it-}\right)\right]}{P(\Delta M_i(t) = 1 \mid \mathcal{F}_{it-})}, \end{aligned} \tag{6}$$

where we utilized the fact that $N_{i1}(t)$ and $N_{i2}(t)$ cannot both jump and where

$$\begin{aligned} P(\Delta M_i(t) = 1 \mid \mathcal{F}_{it-}) = &\sum_{j=1}^{2} P(\Delta N_{ij}(t) = 1 \mid \mathcal{F}_{it-}) \\ &+ P(\Delta M_i(t) = 1 \mid \Delta N_i(t) = 0, \mathcal{F}_{it-}) \left[1 - \sum_{j=1}^{2} P(\Delta N_{ij}(t) = 1 \mid \mathcal{F}_{it-})\right]. \end{aligned} \tag{7}$$

Considering these expressions now in continuous time, we assume the existence of the limits

$$\lim_{\Delta t \to 0} \frac{P\left(\Delta N_{ij}(t) = 1 \mid \mathcal{F}_{it-}\right)}{\Delta t} = \lambda_{ij}(t, \phi_j) \equiv \exp\{\phi_j(t, x_i)\}$$

and

$$\lim_{\Delta t \to 0} \frac{P(\Delta M_i(t) = 1 \mid \Delta N_i(t) = 0, \mathcal{F}_{it-})}{\Delta t} \equiv \rho(t, x_i),$$

where the intensity function $\rho(t, x_i)$ specifies the base series sampling mechanism (Section 3.2). We then obtain from (7) that

$$\lim_{\Delta t \to 0} \frac{P(\Delta M_i(t) = 1 \mid \mathcal{F}_{it-})}{\Delta t} = \rho(t, x_i) + \sum_{j=1}^{2} \lambda_{ij}(t, \phi_j).$$

Dividing both the numerator and denominator of (4)–(6) by Δt and taking the limit as $\Delta t \to 0$, we obtain the expression

$$\lim_{\Delta t \to 0} P(\Delta N_{i1}(t), \Delta N_{i2}(t) \mid \Delta M_i(t) = 1, \mathcal{F}_{it-}) \overset{\phi}{\propto} \frac{\prod_{j=1}^{2} \lambda_{ij}(t, \phi_j)^{\mathrm{d}N_{ij}(t)}}{\rho(t, x_i) + \sum_{j=1}^{2} \lambda_{ij}(t, \phi_j)}.$$

Thus, in the continuous-time setting, the conditional likelihood contribution of individual i can be written as

$$L_i(\phi) \equiv \prod_{t \in (0,\tau]} \left(\frac{\prod_{j=1}^{2} \lambda_{ij}(t, \phi_j)^{\mathrm{d}N_{ij}(t)}}{\rho(t, x_i) + \sum_{j=1}^{2} \lambda_{ij}(t, \phi_j)} \right)^{\mathrm{d}M_i(t)}. \tag{8}$$

This expression has a form that is familiar from the context of multinomial regression, now with an offset term $\log(1/\rho(t, x_i))$. Further, by holding ϕ_2 fixed, we have

$$L_i(\phi) \overset{\phi_1}{\propto} \prod_{t \in (0,\tau]} \left(\frac{\exp\left\{ \mathrm{d}N_{i1}(t) \left[\phi_1(t, x_i) + \log\left(\frac{1}{\rho(t,x_i) + \lambda_{i2}(t,\phi_2)} \right) \right] \right\}}{1 + \exp\left\{ \phi_1(t, x_i) + \log\left(\frac{1}{\rho(t,x_i) + \lambda_{i2}(t,\phi_2)} \right) \right\}} \right)^{\mathrm{d}M_i(t)}.$$

A symmetric expression exists for ϕ_2 by holding ϕ_1 fixed. Thus, we can maximize the likelihood expression $\prod_{i=1}^{n} L_i(\phi)$ by iterating two logistic regressions with offset terms of the form $\log(1/[\rho(t, x_i) + \lambda_{ij}(t, \phi_j)])$ until convergence. Instead of this, however, our aim is to use (8) as the likelihood function in Bayes' formula, in conjunction with MCMC sampling.

As a side note, in non-absolute risk applications, we would rather avoid the estimation of the nuisance parameters ϕ_2 characterizing non-CVD mortality. For this purpose, we can define another 'sampler' counting process $M_i^*(t)$ such that $P\left(\Delta M_i^*(t) = 1 \mid \Delta N_{i1}(t) = 0, \Delta N_{i2}(t) = 1, \mathcal{F}_{it-}\right) = 0$, in which case we do not have the likelihood contributions (5), and the denominator in (8) becomes

$$\lim_{\Delta t \to 0} \frac{P(\Delta M_i^*(t) = 1 \mid \mathcal{F}_{it-})}{\Delta t} = \rho(t, x_i) + \lambda_{i1}(t, \phi_1).$$

The consequent conditional likelihood expression then becomes

$$L_i^*(\phi_1) \equiv \prod_{t \in (0,\tau]} \left(\frac{\exp\left\{ \mathrm{d}N_{i1}(t) \left[\phi_1(t, x_i) + \log\left(\frac{1}{\rho(t,x_i)} \right) \right] \right\}}{1 + \exp\left\{ \phi_1(t, x_i) + \log\left(\frac{1}{\rho(t,x_i)} \right) \right\}} \right)^{\mathrm{d}M_i^*(t)}, \tag{9}$$

which corresponds to the form suggested by Hanley & Miettinen (2009). Note that, for evaluating the absolute risk (1), we need to estimate both ϕ_1 and ϕ_2. For this, we need the conditional likelihood expression (8), which utilizes information from both case series.

Finally, because $P(\Delta N_{i1}(t) = 0, \Delta N_{i2}(t) = 0 \mid \Delta M_i(t) = 0, \mathcal{F}_{it-}) = 1$, the full likelihood in the continuous-time setting is given by

$$\prod_{i=1}^{n} \prod_{t\in(0,\tau]} [P(\mathrm{d}N_{i1}(t), \mathrm{d}N_{i2}(t) \mid \mathrm{d}M_i(t), \mathcal{F}_{it-}, \phi) P(\mathrm{d}M_i(t) \mid \mathcal{F}_{it-}, \phi)] \\ = \prod_{i=1}^{n} L_i(\phi) \times \prod_{i=1}^{n} \prod_{t\in(0,\tau]} P(\mathrm{d}M_i(t) \mid \mathcal{F}_{it-}, \phi), \tag{10}$$

and thus, the first product in (10) is indeed a conditional likelihood in the sense of Cox (1975, p. 269). Such conditional likelihoods share the properties of a likelihood (zero mean score function and information equality, e.g. Appendix C of Saarela *et al.*, 2012), so using them in conjunction with Bayesian inferences is unproblematic, cf. Theorem 3 of Chernozhukov & Hong (2003, p. 308).

3.2. Efficient sampling of the base series

Hanley & Miettinen (2009, p. 8) sampled the base series completely at random from the study base, corresponding to sampling probabilities of the form

$$\rho(t, x_i)\,\mathrm{d}t = 1 - \left(1 - \frac{t_i}{\sum_{i=1}^{n} t_i} \frac{\mathrm{d}t}{t_i}\right)^m \approx m \frac{\mathrm{d}t}{\sum_{i=1}^{n} t_i}, \tag{11}$$

where t_i is the follow-up time until the first event, including censoring, for individual i, m is the number of person-moments to be sampled and $\sum_{i=1}^{n} t_i \equiv y$ are the total person-years of follow-up in the study base. The term t_i/y corresponds to the multinomial sampling probability of individual i (the 'person') in a single trial (of the total of m trials), while the 'moment' is then sampled from the uniform $U(0, t_i)$-distribution. The preceding sampling scheme corresponds to the density of $\rho = m/y$ and offset term $\log(y/m)$.

We generalize this procedure by noting that, for a given m, the efficiency of the sampling can potentially be improved by allowing the sampling probabilities to depend on the determinants of the outcome, most importantly, on age. This is comparable with the efficient case–cohort design suggested by Kim & De Gruttola (1999, p. 155–156). While it might seem tempting to obtain the sampling probabilities through an initial fit of some simple parametric model to the data, in principle, these have to be fixed on an *a priori* basis in order to avoid using the data twice. Let $\lambda^*_{ij}(t)$ denote the 'prior' hazard function for individual i experiencing an incident event of type j and $\Lambda^*_{ij}(t)$ the corresponding cumulative hazard. We want to roughly match the risk factor distribution of the sampled person-moments to that of the events, so that $\rho(t, x_i) \propto \lambda^*_{ij}(t)$. Generally, the selection probabilities may now be chosen as

$$\rho(t, x_i)\,\mathrm{d}t = 1 - \left(1 - \frac{\Lambda^*_{ij}(t_i)}{\sum_{i=1}^{n} \Lambda^*_{ij}(t_i)} \frac{\lambda^*_{ij}(t)\,\mathrm{d}t}{\Lambda^*_{ij}(t_i)}\right)^m \approx m \frac{\lambda^*_{ij}(t)\,\mathrm{d}t}{\sum_{i=1}^{n} \Lambda^*_{ij}(t_i)}. \tag{12}$$

It should be noted that (12) reduces to (11) when the sampling rate $\lambda^*_{ij}(t)$ does not depend on i or t. As usual in case–base/case–cohort type sampling, the same base series can be used for estimating several cause-specific hazard rates (Langholz & Thomas, 1990). Thus, while the base series is sampled to roughly match the resulting distribution of (t, x_i) to that of the outcome events of interest ($j = 1$), the same base series may also be used to estimate the mortality rate due to other causes ($j = 2$).

Figure 1 presents an illustration of case–base sampling in the study base defined by the first 10 years of follow-up of 12,609 men aged 25–64 years from the cohorts described in more detail

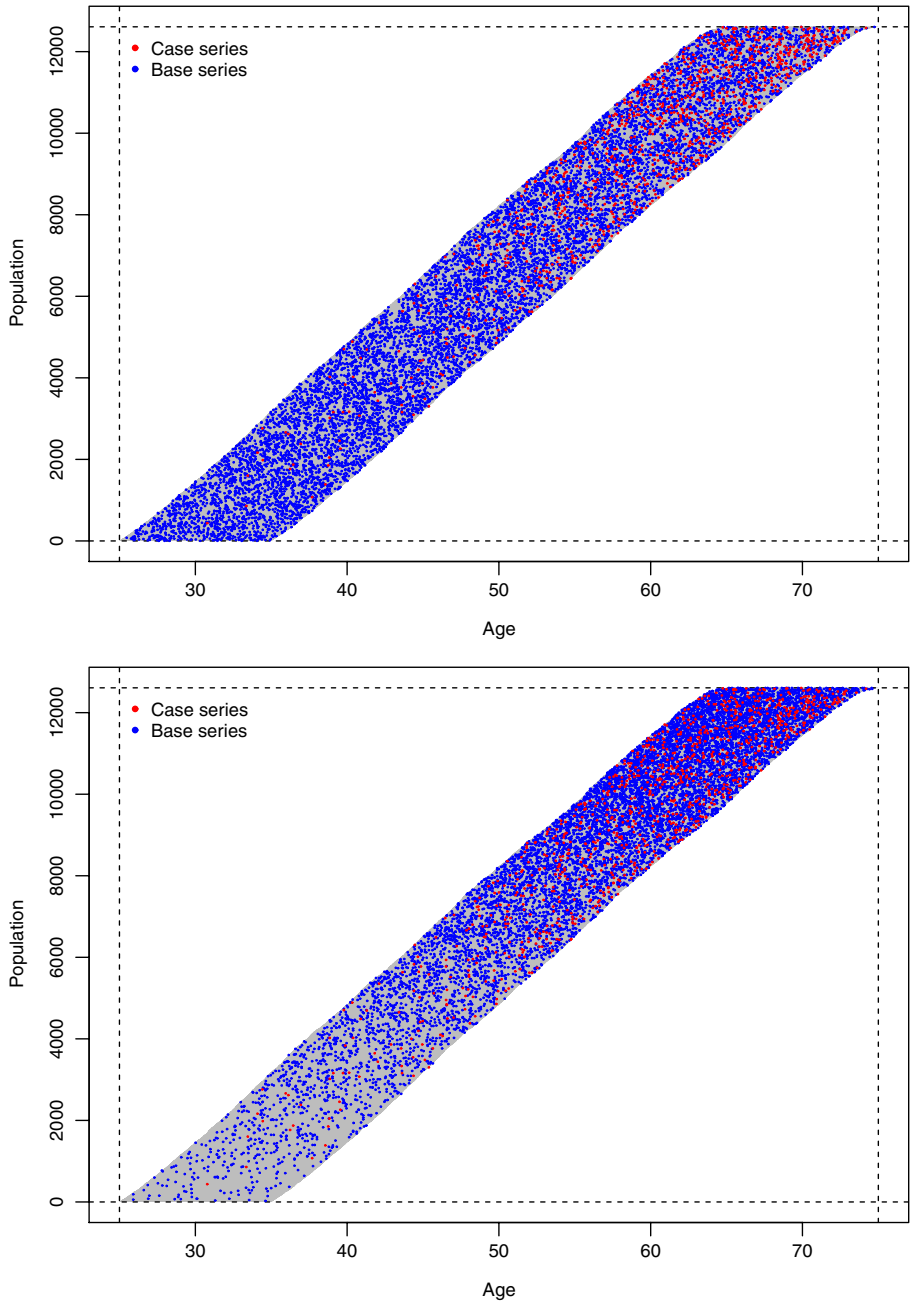

Fig. 1. Top panel: case–base sampling with uniform probabilities. Bottom panel: case–base sampling with probabilities proportional to the cardiovascular disease event rate given by the Framingham score. The number of cases is 1014, and the size of the base series $m = 10{,}140$.

in Section 6. The top panel illustrates sampling of a base series of $m = 10{,}140$ person-moments, 10 times the number of incident CVD events in the study base, with uniform probabilities as in Hanley & Miettinen (2009, p. 8). The bottom panel illustrates sampling of the base series with

probabilities (12), where the (constant over time) *a priori* rate λ_{i1}^* is given by the 10-year risk of CVD based on the Framingham score (D'Agostino Sr *et al.*, 2008, p. 751–752):

$$\begin{aligned}\lambda_{i1}^* = -\exp\{&3.06117 \times \log(\text{age at baseline}_i)\\ &+ 1.12370 \times \log(\text{total cholesterol}_i)\\ &- 0.93263 \times \log(\text{high-density lipoprotein cholesterol}_i)\\ &+ 1.99881 \times \log(\text{treated systolic blood pressure}_i)\\ &+ 1.93303 \times \log(\text{untreated systolic blood pressure}_i)\\ &+ 0.65451 \times \text{smoker}_i\\ &+ 0.57367 \times \text{prevalent diabetes}_i\} \times \log(0.88936)/10.\end{aligned}$$

The sampling of person-moments proceeds by first selecting the persons with replacement by a single draw from the

$$\text{Multinomial}\left(m, \frac{t_1\lambda_{11}^*}{\sum_{i=1}^n t_i\lambda_{i1}^*}, \cdots, \frac{t_n\lambda_{n1}^*}{\sum_{i=1}^n t_i\lambda_{i1}^*}\right)$$

distribution. Because the sampling rate was taken to be constant in time on study, the moments (with their number being given by the multinomial counts) for each of the selected persons are then drawn independently from $U(0, t_i)$. The resulting offset terms are given by $\log\{(\sum_{i=1}^n t_i\lambda_{i1}^*)/(m\lambda_{i1}^*)\}$. The realized case–base sample in Figure 1 illustrates how the low-risk individuals contribute fewer person-moments, which should result in a gain in efficiency. To have an idea of the magnitude of the improvement, we fitted a Cox model similar to the Framingham model to the full dataset and compared the estimated variances of the regression coefficients with those based on fitting a corresponding logistic regression to the case–base samples. We repeated the base series sampling 1000 times using both sampling mechanisms with $m = 10{,}140$, fitted the logistic models to each sample and compared the average variances of the logistic regression estimates to the Cox model variances. When using the uniform sampling mechanism, the variances inflated on average by 38%, while using sampling proportional to the Framingham CVD rate increased the variances by only 12%. With $m = 101{,}400$, the corresponding numbers were 4.7% and 1.2%.

4. Non-parametric hazard regression and packaging of covariates

In this section, we consider flexible non-parametric specifications for the hazard functions λ_j that would enable us to better approximate the integrals of type (1). In addition, non-parametric modelling allows us to learn about the functional forms of the associations between the outcomes and risk factors. To emphasize the symmetric role of timescales and covariates in such a modelling task (Berzuini & Clayton, 1994; Arjas & Liu, 1996), in the following, we consider cause-specific hazards as functions of a $(p + 1)$-dimensional argument $z_i = (z_{i1}, z_{i2}, \ldots, z_{ip}, z_{i(p+1)}) \equiv (t, x_{i1}, \ldots, x_{ip})$. This is in contrast to conventional semi-parametric hazard regression modelling, where the hazard over time is handled non-parametrically, while parametric functions are used for the covariate effects of interest. Under the present notation, consideration of multiple timescales would also be straightforward, although we do not pursue this aspect here further.

Let now $\mathcal{S} \subseteq 2^{\{1,2,\ldots,p,p+1\}}$ represent a subset of the collection of all non-empty subsets of the covariates (including timescales). In a general form, the cause-specific hazard functions could be specified as

$$\lambda_j(t, x_i, \phi_j) \equiv \lambda_j(z_i, \phi_j) = \exp\left\{\phi_{j0} + \sum_{k=1}^{|\mathcal{S}|} \phi_{jk}(z_{iS_k})\right\},$$

where $S_k \in \mathcal{S}$ and $z_{iS_k} \equiv (z_{il} : l \in S_k)$. Here, $\phi_{jk} : \mathbb{R}^{|S_k|} \to \mathbb{R}$ are realizations of random (to be interpreted here as synonymous to 'unknown') regression functions to be estimated from the data. We do not yet concern ourselves with the identifiability of such functions, but note that in situations where the interest lies only in the posterior predictive distributions for observable quantities, identifiability may not be necessary. It is now possible to represent several familiar special cases by using this notation. First, taking $\mathcal{S} = \{\{1\}, \{2\}, \ldots, \{p\}, \{p+1\}\}$, that is, $S_k = \{k\}$ for all $k \in \{1, 2, \ldots, p, p+1\}$, and fixing $\phi_{jk}(z_{iS_k}) = \beta_{jk} z_{ik}$ for all $k \in \{2, 3, \ldots, p, p+1\}$ give a semi-parametric formulation

$$\lambda_j(z_i, \phi_j) = \exp\left\{\phi_{j0} + \phi_{j1}(t) + \sum_{k=2}^{p+1} \beta_{jk} x_{ik}\right\}$$

with a non-parametrically specified baseline log-hazard rate function. Relaxing the assumption of exponential covariate effects, but retaining the log-additive structure of independent covariate effects, gives a generalized additive model-(GAM) type model

$$\lambda_j(z_i, \phi_j) = \exp\left\{\phi_{j0} + \sum_{k=1}^{p+1} \phi_{jk}(z_{jk})\right\}.$$

To allow for interactions between specific covariates, we might want to generalize this so that $\mathcal{S}$ represents a partition of the set of covariates, that is, $\bigcup_{k=1}^{|\mathcal{S}|} S_k = \{1, 2, \ldots, p, p+1\}$ and $S_k \cap S_l = \emptyset$ for all $k \neq l$. This was the approach taken by Arjas & Liu (1996), with the partition specified on an *a priori* basis. Finally, the general formulation allows us to specify even $\mathcal{S} = \{\{1, 2, \ldots, p, p+1\}, \ldots, \{1, 2, \ldots, p, p+1\}\}$, with $|\mathcal{S}|$, the number of full-rank regression functions allowed, fixed. Such a specification is useful if we want allow for possibilities for model selection that would happen by controlling the specification of the random ϕ_{jk} rather than the specification of $\mathcal{S}$.

The general modelling framework outlined earlier does not yet indicate how the function realizations ϕ_{jk} should be specified. Even though these functions are considered unknown, in practice, some structure is needed to enable their estimation. Here, we adopt the non-parametric monotonic regression procedure proposed by Saarela & Arjas (2011a); other possible approaches for non-parametric Bayesian regression are briefly reviewed therein, but comparisons between different methods, although potentially of interest, are outside the scope of the present paper. For our present purposes, the procedure of Saarela & Arjas (2011a) has the advantage of being able to accommodate any number of covariates and a built-in model selection functionality. Further computational details are given in the Appendix.

5. Model validation in case–base samples

In evaluating the risk (1), integration over the regression function realizations $\phi_j(t, x_i)$ is computationally less critical compared with evaluating the likelihood function. This is because the

risk estimates need not be evaluated every time a modification is proposed to the regression function. Thus, we approximate the risk (1) by

$$\pi_s(x_i) \equiv E_{\phi|\mathcal{F}}[\pi_s(x_i,\phi)] \approx \frac{1}{l}\sum_{k=1}^{l}\int_{t\in(0,s]}\exp\left\{\phi_1^{(k)}(t,x_i)\right\}\exp\left(-\sum_{j=1}^{2}\int_{u\in(0,t]}\exp\left\{\phi_j^{(k)}(t,x_i)\right\}\,\mathrm{d}u\right)\mathrm{d}t,$$

where $\phi^{(k)}, k = 1,\ldots,l$, is an MCMC sample from the posterior distribution $P(\mathrm{d}\phi \mid \mathcal{F})$ and the integrals over time are evaluated by quadrature integration if needed; when using the model described in the Appendix, these are available analytically, as the regression function realizations are piecewise constant. While the posterior variances $V_{\phi|\mathcal{F}}[\pi_s(x_i,\phi)]$ are not of direct interest, the variability of the parameter-conditional risks, or alternatively, predicted time-to-event outcomes simulated from the hazard model given the realizations $\phi^{(k)}$, may be used for obtaining posterior credible intervals for various model validation statistics.

As illustrated in the following section, in population cohorts with a wide age range, age dominates the absolute risk of chronic diseases and consequently the discrimination statistics. Thus, one might be interested in discrimination statistics where the (multiplicative) effect of (baseline) age has been statistically removed. This can be achieved by constructing a validation set where the case and base series have been matched for their age distribution. Although a similar effect could be achieved using age-matched risk set sampling, following the approach of Section 3, we apply case–base sampling. Informally, the construction of such a validation sample can be motivated by noting that the cohort-level hazard rate and the 'fitted' log-odds in the case–base sample, denoted as $\tilde{\omega}_1(t,x_i)$, are connected through

$$\begin{aligned}
\exp\{\tilde{\omega}_1(t,x_i)\} &= \frac{P\left(\mathrm{d}N_{i1}(t)=1 \mid \mathrm{d}M_i^*(t)=1,\mathcal{F}_{it-}\right)}{P\left(\mathrm{d}N_{i1}(t)=0 \mid \mathrm{d}M_i^*(t)=1,\mathcal{F}_{it-}\right)}\\
&= \frac{P\left(\mathrm{d}N_{i2}(t)=0 \mid \mathrm{d}N_{i1}(t)=1,\mathrm{d}M_i^*(t)=1,\mathcal{F}_{it-}\right)}{P\left(\mathrm{d}N_{i2}(t)=0 \mid \mathrm{d}N_{i1}(t)=0,\mathrm{d}M_i^*(t)=1,\mathcal{F}_{it-}\right)}\\
&\quad\times\frac{P\left(\mathrm{d}N_{i1}(t)=1 \mid \mathrm{d}M_i^*(t)=1,\mathcal{F}_{it-}\right)}{P\left(\mathrm{d}N_{i1}(t)=0 \mid \mathrm{d}M_i^*(t)=1,\mathcal{F}_{it-}\right)}\\
&= \frac{P\left(\mathrm{d}M_i^*(t)=1 \mid \mathrm{d}N_{i1}(t)=1,\mathrm{d}N_{i2}(t)=0,\mathcal{F}_{it-}\right)}{P\left(\mathrm{d}M_i^*(t)=1 \mid \mathrm{d}N_{i1}(t)=0,\mathrm{d}N_{i2}(t)=0,\mathcal{F}_{it-}\right)/\mathrm{d}t}\\
&\quad\times\frac{P(\mathrm{d}N_{i1}(t)=1 \mid \mathcal{F}_{it-})/\mathrm{d}t}{P(\mathrm{d}N_{i1}(t)=0,\mathrm{d}N_{i2}(t)=0 \mid \mathcal{F}_{it-})}\\
&= \frac{\sum_{i=1}^{n}\Lambda_{i1}^*(t_i)}{m\lambda_{i1}^*(t)}\lambda_{i1}(t).
\end{aligned} \tag{13}$$

Here, $\lambda_{i1}^*(t)$ may be obtained for instance from a simple parametric survival model fitted to the validation cohort, using age at baseline as a covariate. For instance, in the case of an exponential survival model, because the fitted log-odds do not depend on age, $E[\lambda_{i1}(t)] = mE[\lambda_{i1}^*]/(\sum_{i=1}^{n} t_i\lambda_{i1}^*)\exp\{\tilde{\omega}_1(t,x_i)\}$, where the expectation is with respect to the age distribution of the sampled person-moments. The corresponding offset terms for estimating such age-averaged rates are $\log((\sum_{i=1}^{n} t_i\lambda_{i1}^*)/(m\bar{\lambda}_1^*))$, where $\bar{\lambda}_1^*$ is the mean sampling rate over the m sampled person-moments.

As an example of validation statistics, we consider the calculation of receiver operating characteristic (ROC) curves. While numerous direct and indirect non-parametric regression techniques for estimating (potentially covariate-dependent) ROC curves have been presented in the literature (e.g. Cai & Pepe, 2002; Alonzo & Pepe, 2002; Zheng & Heagerty, 2004; Rodriguez & Martinez, 2014), herein we concentrate mainly on the problem on estimating ROC curves from censored time-to-event data. Our general approach follows that of Heagerty *et al.* (2000), but replacing Kaplan–Meier survival probabilities with ones obtained from fitting a hazard model to a case–base sample of person-moments, using the techniques introduced in Sections 3 and 4. For this purpose, we note that for any given risk cut-off π^*, the true positive probability (sensitivity) can be calculated as

$$P(\pi_s(x_i)>\pi^* \mid N_{i1}(s)=1)=\frac{\int_{t\in(0,s]}\int_{r\in(\pi^*,1]}P(\mathrm{d}N_{i1}(t)=1 \mid \pi_s(x_i)=r)P(\pi_s(x_i)\in \mathrm{d}r)}{\int_{t\in(0,s]}\int_{r\in(0,1]}P(\mathrm{d}N_{i1}(t)=1 \mid \pi_s(x_i)=r)P(\pi_s(x_i)\in \mathrm{d}r)}$$
$$\approx \frac{\sum_{i=1}^n \mathbf{1}_{\{\pi_s(x_i)>\pi^*\}}\int_{t\in(0,s]}P(\mathrm{d}N_{i1}(t)=1 \mid \pi_s(x_i)=r)}{\sum_{i=1}^n \int_{t\in(0,s]}P(\mathrm{d}N_{i1}(t)=1 \mid \pi_s(x_i)=r)}. \tag{14}$$

The false positive probability (1 − specificity) is obtained analogously as

$$P(\pi_s(x_i)>\pi^* \mid N_{i1}(s)=0)\approx \frac{\sum_{i=1}^n \mathbf{1}_{\{\pi_s(x_i)>\pi^*\}}\left[1-\int_{t\in(0,s]}P(\mathrm{d}N_{i1}(t)=1 \mid \pi_s(x_i)=r)\right]}{\sum_{i=1}^n \left[1-\int_{t\in(0,s]}P(\mathrm{d}N_{i1}(t)=1 \mid \pi_s(x_i)=r)\right]}. \tag{15}$$

In (14) and (15), the conditional risk $\int_{t\in(0,s]}P(\mathrm{d}N_{i1}(t)=1 \mid \pi_s(x_i)=r)$ is given by formula (1) and can in turn be obtained from a hazard model of the form $\lambda_j(t,\pi_s(x_i),\phi_j)=\exp\{\phi_{j0}+\phi_{j1}(t,\pi_s(x_i))\}$, where ϕ_{j1} is a non-parametrically specified regression function; we specify this using the monotonic construction discussed in the Appendix. Given the risks $\pi_s(x_i)$ from the predictive model fitted in the training set, the hazard model may be fitted to the validation case–base sample as outlined in Sections 3 and 4; optionally, the validation set can be age matched, as outlined earlier.

6. Application

The data we use in the illustration come from the National FINRISK Study (Salomaa *et al.*, 1996; Vartiainen *et al.*, 2000). The FINRISK cohorts are risk factor surveys carried out in Finland every five years, recruited as cross-sectional random samples from geographical populations and followed up thereafter for various outcomes of interest, including CVD and total mortality. We included 12,609 men aged 25–64 years at baseline from the FINRISK 82, 87, 92 and 97 cohorts. Although all of the cohorts have been followed up until recent years, we included only the first 10 years of follow-up for the purpose of assessing the 10-year risk of CVD. During this period, there occurred a total of 1014 incident (fatal or non-fatal) major CVD events when using a composite definition comprising myocardial infarction, unstable angina, cardiac revascularization and ischaemic stroke and 539 deaths due to causes other than CVD.

In the illustration, we consider only a few basic risk factors of CVD, also contained in the Framingham score (D'Agostino Sr *et al.*, 2008), namely non-high-density lipoprotein (HDL) and HDL cholesterol (the former obtained from total cholesterol by subtracting the latter,

resulting in two nearly uncorrelated variables), systolic blood pressure (separately for those with and without antihypertensive medication), smoking and prevalent diabetes ($z_{i\{2,\ldots,7\}}$) and age (z_{i1}, used as the timescale, except for model 2 where follow-up time was used instead, with age at baseline included in the risk score). To observe the effect of relaxing parametric modelling assumptions, we consider a small number of alternative model specifications, all capable of reducing to a generalized additive formulation when multidimensional relationships are not supported by data. These are specified as follows (in the order of perceived flexibility):

(1) $\lambda_1(z_i, \phi_1) = \exp\{\phi_{10} + \phi_{11}(z_{i1})\}$ (age-only model for comparison).
(2) $\lambda_1(z_i, \phi_1) = \exp\{\phi_{10} + \phi_{11}(t) + f(z_{i\{1,\ldots,7\}})\}$, where f is the CVD Framingham score, with the baseline hazard estimated from the data on hand.
(3) $\lambda_1(z_i, \phi_1) = \exp\{\phi_{10} + \phi_{11}(z_{i1}) + \sum_{k=2}^{7} \beta_{1k} z_{ik}\}$ (semi-parametric).
(4) $\lambda_1(z_i, \phi_1) = \exp\{\phi_{10} + \sum_{k=1}^{7} \phi_{1k}(z_{ik})\}$ (GAM type).
(5) $\lambda_1(z_i, \phi_1) = \exp\{\phi_{10} + \sum_{k=1}^{7} \phi_{1k}(z_{ik}) + \sum_{k=8}^{9} \phi_{1k}(z_{i\{1,2,\ldots,7\}})\}$, allowing for up to two-dimensional relationships within each of the two blocks ϕ_{18} and ϕ_{19}.
(6) $\lambda_1(z_i, \phi_1) = \exp\{\phi_{10} + \sum_{k=1}^{7} \phi_{1k}(z_{ik}) + \sum_{k=8}^{11} \phi_{1k}(z_{i\{1,2,\ldots,7\}})\}$, allowing for up to two-dimensional relationships within each of the four blocks $\phi_{18}, \ldots, \phi_{1(11)}$.
(7) $\lambda_1(z_i, \phi_1) = \exp\{\phi_{10} + \sum_{k=1}^{7} \phi_{1k}(z_{ik}) + \phi_{18}(z_{i\{1,2,\ldots,7\}})\}$ (allowing for all dependencies within the block ϕ_{18}).

The model for the competing causes of death was in each case taken to be of the semi-parametric form $\lambda_2(z_i, \phi_2) = \exp\{\phi_{20} + \phi_{21}(z_{i1}) + \sum_{k=2}^{7} \beta_{2k} z_{ik}\}$. The models were fitted to the case–base sample depicted in the lower panel of Figure 1 for checking the model fit and calibration, while the actual predictive ability of the models was compared using fivefold cross-validation, with the data split randomly into five equal-sized groups and the models fitted to the remaining set when each of these is removed in turn (in conjunction with case–base sampling, this works by removing all person-moments contributed by the removed individuals). We refer to the Appendix for computational details and prior specifications.

Figure 2 shows the model fit in terms of log-likelihood and the complexity in terms of the total number of support points used in the fit. These results indicate, as expected, that the

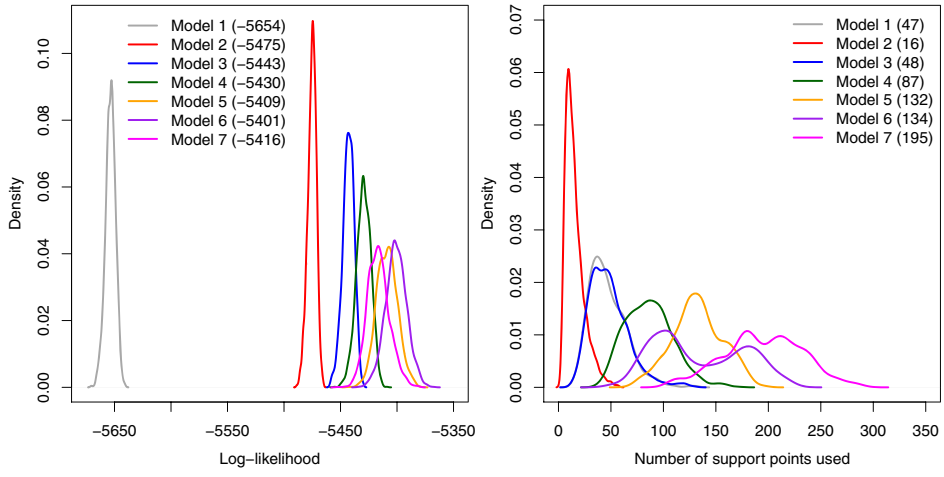

Fig. 2. Model fit and complexity. The left panel shows posterior log-likelihood distributions. The right panel shows the numbers of support points used in the model fit. The numbers in brackets are posterior means.

more flexible models allow better fit to the data and generally use more support points in doing so. The downside of more flexible model specifications is the added noise in the predictions. Figure 3 shows the discriminative ability of the models in terms of ROC curves, when estimated without cross-validation. The left panel demonstrates the dominating effect of age in the absolute risk; the age-only model gives an area under the curve (AUC) of 76%, with the other risk factor predictors adding only few percentage points on top of this. The right panel depicts the ROC curves in a validation case–base sample where the baseline age distribution of the base series has been matched to that of the case series, as described in Section 5. The absolute level

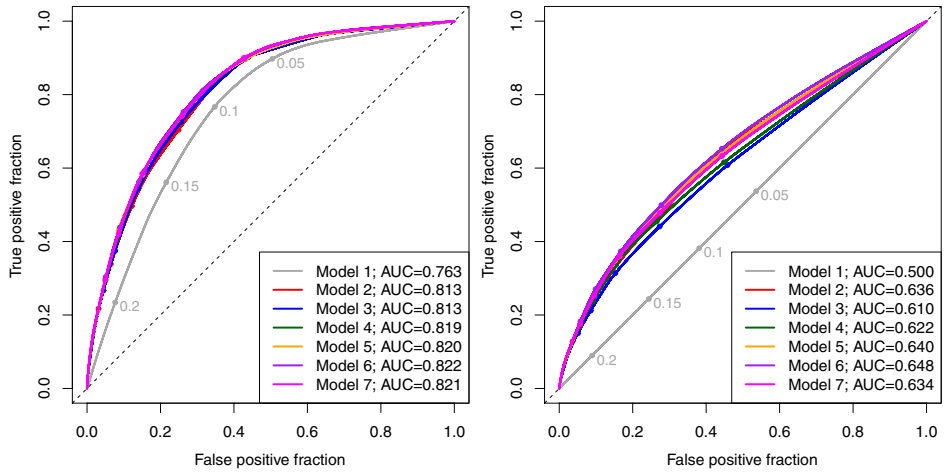

Fig. 3. Model discrimination without cross-validation. The left panel shows receiver operating characteristic (ROC) curves without age adjustment, while the right panel shows ROC curves in a validation set where the effect of age has been removed by matching the age distribution of the case and base series.

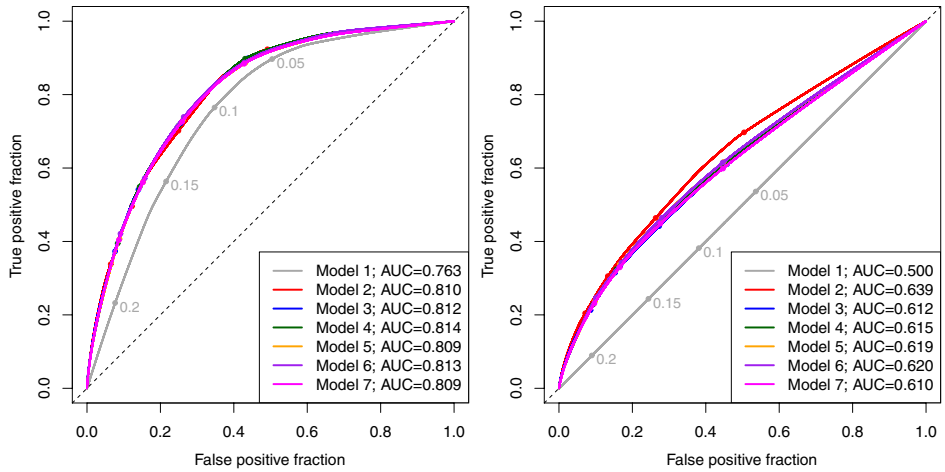

Fig. 4. Model discrimination with cross-validation. The left panel shows receiver operating characteristic (ROC) curves without age adjustment, while the right panel shows ROC curves in a validation set where the effect of age has been removed by matching the age distribution of the case and base series.

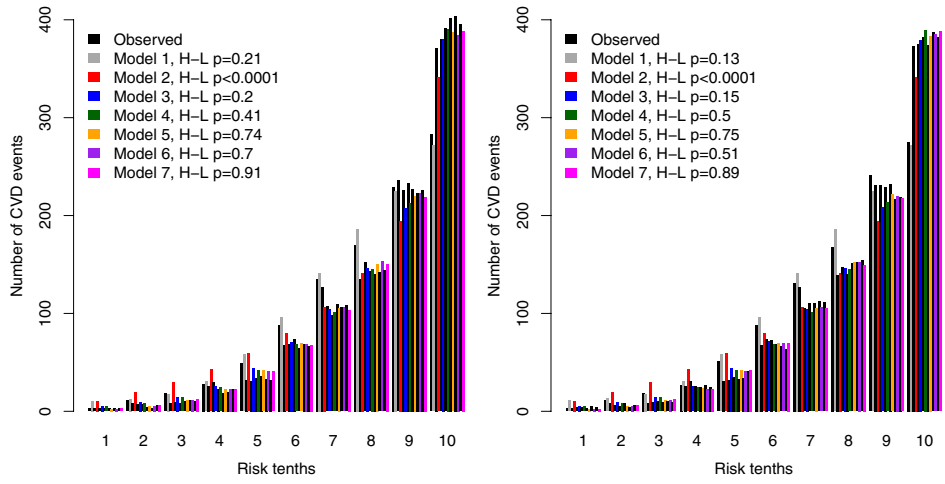

Fig. 5. Model calibration. The left panel shows calibration plots and Hosmer–Lemeshow test p-values without cross-validation. The right panel shows the same statistics with cross-validation.

of AUC is now much lower, but there still are no substantial differences between the discriminative ability of the alternative models resulting from modelling of the baseline risk factors. In the non-cross-validated results of Figure 3, the order of the models in terms of AUC is the same as in terms of likelihood in the left panel of Figure 2. However, in the cross-validated predictions depicted in Figure 4, the more complex models are penalized from having to estimate the features that the proportional hazards models assume known; in fact, in the right panel, the model with the highest AUC is the one using the fixed risk score. However, in terms of calibration (calibration plots and Hosmer–Lemeshow test p-values in Figure 5), the more flexible models 4–7 performed better than the fixed risk score (model 2) or the semi-parametric specification (model 3).

7. Discussion

The combination of case–base sampling and logistic/multinomial form likelihoods provides an easy way to fit non-proportional hazard models utilizing existing software for modelling binary response data. This has natural applications in risk assessment, as the baseline hazard is also obtained from the model fit. This can be contrasted to fitting a Cox partial likelihood, which eliminates the baseline hazard, which is then recovered using the Nelson–Aalen/Breslow estimator. More broadly, the case–base sampling approach provides a whole alternative framework for hazard modelling, alongside risk set sampling and the related semi-parametric inference procedures; thus, we feel that this approach deserves more exposure than it has got so far. In this paper, we formalized this approach as conditional likelihood inference and further outlined how case–base sampling can be used in model validation.

Furthermore, in the approach based on logistic/multinomial form likelihoods, the timescales are handled similarly to baseline covariates, which allows easy incorporation of multiple timescales, as well as computationally convenient modelling of interactions between timescales and baseline covariates, without numerical evaluation of the cumulative hazard or a need to split the time axes. In our case study of 10-year risk of CVD, we used this framework to specify models with additional components allowing for multidimensional relationships, but capable

of reducing into a generalized additive functional form. The results demonstrate that in healthy population cohorts, age dominates the absolute risk, with little gain in discriminative ability to be gained through further modelling. This motivated us to explore ways to remove the age effect from ROC curve comparisons, in order to concentrate on the discrimination due to the other factors in the model.

Finally, some reservations should be made when using Bayesian model averaging techniques (including our proposed monotonic regression procedure) for predictive applications. It is well known that the Bayesian paradigm, when employed in model/variable selection, favours parsimonious models, which generally does not result in optimal predictions (e.g. Burnham & Anderson, 2004). This applies to posterior model probabilities, as well as to Bayes factors and the Bayesian information criterion, all of which use the marginal probability of the data (marginalized over the parameter space of the model) as the selection criterion. Bayesian model selection and model averaging do give optimal predictions in the prequential setting (Dawid, 1984; Kass & Raftery, 1995, p. 777), but this does not correspond to the typical training/validation setting employed in the evaluation of prognostic models, where the models are at best approximations and finite amounts of data are available. Overcoming the slow convergence of Bayesian model selection has been recently studied by van Erven *et al.* (2012), and the lessons from therein should be applied in the prognostic modelling context to obtain the full benefit from the more flexible model formulations.

Acknowledgements

The majority of this work was carried out when the first author was working at the Department of Epidemiology, Biostatistics and Occupational Health of McGill University. The research of the first author was supported by the Finnish Foundation for Technology Promotion, and he also acknowledges support of the Natural Sciences and Engineering Research Council of Canada. The authors would like to thank Prof. James A. Hanley of McGill University for helpful comments and Prof. Veikko Salomaa of the National Institute for Health and Welfare for the permission to use the FINRISK data in our illustration.

References

Alonzo, T. A. & Pepe, M. S. (2002). Distribution-free ROC analysis using binary regression techniques. *Biostatistics* **3**, 421–432.

Arjas, E. (1989). Survival models and martingale dynamics. *Scand. J. Stat.* **16**, 177–225.

Arjas, E. & Haara, P. (1984). A marked point process approach to censored failure data with complicated covariates. *Scand. J. Stat.* **11**, 193–209.

Arjas, E. & Haara, P. (1987). A logistic regression model for hazard: asymptotic results. *Scand. J. Stat.* **14**, 1–18.

Arjas, E. & Liu, L. (1996). Non-parametric Bayesian approach to hazard regression: a case-study with a large number of missing covariate values. *Stat. Med.* **15**, 1757–1770.

Bernardo, J. M. & Smith, A. F. M. (1994). *Bayesian theory*, Wiley, Chichester.

Berzuini, C. & Clayton, D. (1994). Bayesian analysis of survival on multiple time scales. *Stat. Med.* **13**, 823–838.

Burnham, K. P. & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociol. Methods Res.* **33**, 261–304.

Cai, T. & Pepe, M. S. (2002). Semiparametric receiver operating characteristic analysis to evaluate biomarkers for disease. *J. Amer. Statist. Assoc.* **97**, 1099–1107.

Chernozhukov, V. & Hong, H. (2003). An MCMC approach to classical estimation. *J. Econometrics* **115**, 293–346.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

D'Agostino Sr, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M. & Kannel, W. B. (2008). General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* **117**, 743–753.

Dawid, A. P. (1984). Statistical theory: the prequential approach. *J. Roy. Statist. Soc. Ser. A* **147**, 278–292.

Gill, R. D. & Johansen, S. (1990). A survey of product-integration with a view toward application in survival analysis. *Ann. Statist.* **18**, 1501–1555.

Hanley, J. A. & Miettinen, O. S. (2009). Fitting smooth-in-time prognostic risk functions via logistic regression. *Int. J. Biostat.* **5**, (1), Article 3. doi: 10.2202/1557-4679.1125.

Heagerty, P., Lumley, T. & Pepe, M. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56**, 337–344.

Kalbfleisch, J. D. & Prentice, R. L. (2002). *The statistical analysis of failure time data*, (2nd edition)., Wiley, NJ.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90**, 773–795.

Kim, S. & De Gruttola, V. (1999). Strategies for cohort sampling under the Cox proportional hazards model, application to an AIDS clinical trial. *Lifetime Data Anal.* **5**, 149–172.

Langholz, B. & Thomas, D. C. (1990). Nested case–control and case–cohort methods of sampling from a cohort: a critical comparison. *Am. J. Epidemiol.* **131**, 169–176.

Miettinen, O. S. (2011). *Epidemiological research: terms and concepts*, Springer, Dordrecht.

Miettinen, O. S. & Karp, I. (2012). *Epidemiological research: an introduction*, Springer, Dordrecht.

Rodriguez, A. & Martinez, J. C. (2014). Bayesian semiparametric estimation of covariate-dependent ROC curves. *Biostatistics* **15**, 353–369.

Saarela, O. & Arjas, E. (2011a). A method for Bayesian monotonic multiple regression. *Scand. J. Stat.* **38**, 499–513.

Saarela, O. & Arjas, E. (2011b). On non-parametric bayesian regression in cardiovascular disease risk assessment. In *JSM Proceedings, Section on Statistics in Epidemiology* American Statistical Association., Alexandria, VA; 3192–3204.

Saarela, O., Kulathinal, S. & Karvanen, J. (2012). Secondary analysis under cohort sampling designs using conditional likelihood. *J. Probab. Stat.* **2012**, Article 931416. doi: 10.1155/2012/931416.

Salomaa, V., Miettinen, H., Kuulasmaa, K., Niemelä, M., Ketonen, M., Vuorenmaa, T., Lehto, S., Palomäki, P., Mähönen, M., Immonen-Räihä, P., Arstila, M., Kaarsalo, E., Mustaniemi, H., Torppa, J., Tuomilehto, J., Puska, P. & Pyorälä, K. (1996). Decline of coronary heart disease mortality in Finland during 1983 to 1992: roles of incidence, recurrence, and case-fatality. The FINMONICA MI Register Study. *Circulation* **94**, 3130–3137.

Schweder, T. (1970). Composable Markov processes. *J. Appl. Probab.* **7**, 400–410.

van Erven, T., Grünwald, P. D. & de Rooij, S. (2012). Catching up faster by switching sooner: a predictive approach to adaptive estimation with an application to the AIC-BIC dilemma. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **74**, 361–417.

Vartiainen, E., Jousilahti, P., Alfthan, G., Sundvall, J., Pietinen, P. & Puska, P. (2000). Cardiovascular risk factor changes in Finland, 1972–1997. *Int. J. Epidemiol.* **29**, 49–56.

Zheng, Y. & Heagerty, P. J. (2004). Semiparametric estimation of time-dependent ROC curves for longitudinal marker data. *Biostatistics* **5**, 615–632.

Received July 2013, in final form September 2014

Olli Saarela, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada.
E-mail: olli.saarela@utoronto.ca

Appendix: computational details

The function (3) of Saarela & Arjas (2011a), denoted henceforth as $\phi_{jk}(z_{iS_k})$ (with $|S_k|$ specifying the dimension of the construction), can approximate asymptotically, with an increasing number of support points, general monotonic relationships. In principle, we could set directly $\mathcal{S} = \{\{1, 2, \ldots, p, p+1\}\}$ to obtain $\phi_{j1}(z_{iS_1}) = \phi_{j1}(z_i)$ and $\lambda_j(z_i, \phi_j) = \exp\{\phi_{j0} + \phi_{j1}(z_i)\}$. Then, however, the inferences would likely be hampered by the curse of dimensionality even when there are only a moderate number of covariates p, unless vast amounts of data are available. Hence, to achieve parsimony in the model fit, the monotonic

constructions are best utilized as lower-dimensional versions in conjunction with *a priori* packaging of the covariates, with the blocks contributing additively to the log-hazard function. Alternatively, we may allow for several full-dimensional constructions to act additively on the log-hazard and let the model selection capability in the functions $\phi_{jk}(z_{iS_k})$ take care of the desired dimension reduction. Even though in such constructions the individual regression function components are not necessarily identifiable, our main interest here is in posterior predictive distributions instead of function estimation. It is however required that the model components be constrained so that the intercept term ϕ_{j0} specifies the absolute level of log-hazard while the others act additively on this. This can be achieved by setting the constraint $\sum_{i=1}^{n} \phi_{jk}(z_i) = 0$ for all of the additive components, and it will also improve the mixing of the MCMC updating scheme. The computation proceeds as described in the Appendix of Saarela & Arjas (2011a), with an added rescaling on the marks of a point process realization. On the log-hazard scale, it is performed by subtracting from all marks the constant $\sum_{i=1}^{n} \phi^*_{jk}(z_i)/n$, where $\phi^*_{jk}(z_i)$ is the random function after the proposed modification. This is carried out in conjunction with every proposal before accepting or rejecting the modification. Because the marks are assumed *a priori* uniformly distributed (on the log-hazard scale), the rescaling step does not affect the acceptance probabilities. Instead of being *a priori* restricted to a bounded interval as in Saarela & Arjas (2011a, p. 513), for the additive components, it is enough to specify the range of the interval where the log-hazard is allowed to vary.

In the application of Section 6, the sampler described in the Appendix of Saarela & Arjas (2011a) was run for 10,000 iterations after a 5000-iteration burn-in, and saving every fifth state of the chain. When modelled non-parametrically, all the covariates were rescaled to the $[0, 1]$ interval using empirical cumulative distribution function transformation, while the age axis was linearly transformed as $[25, 75) \to [0, 1]$. These transformations were made in order to use the same prior specifications irrespective of the scale of the covariates. The model specifications 1–7 involved specifying 2, 2, 2, 6, 50, 92 and 128 point processes, respectively, including the non-parametrically modelled baseline hazard for the other deaths. These were specified *a priori* as homogeneous Poisson processes, with Gamma$(0.1, 0.1)$-distributed intensity parameters. The range of variation of the non-parametric components was restricted to 4 on the log-relative hazard scale, with the exception of the component ϕ_{18} in model 7, where the maximum range of variation was set to 8.

The computational algorithm is implemented in the R package `monoreg`, which also includes an example for reproducing the ROC curves in Figure 3. The package is available from the homepage of the first author at http://individual.utoronto.ca/osaarela/