Biometrika (2025), **112**, 2, asaf008 https://doi.org/10.1093/biomet/asaf008
Advance Access publication 31 January 2025

Super-efficient estimation of future conditional hazards based on time-homogeneous high-quality marker information

By D. BAGKAVOS

Department of Mathematics, University of Ioannina, Ioannina 45100, Greece
d.bagkavos@uoi.gr

A. ISAKSON

Bayes Business School, City St George's, University of London,
106 Bunhill Row, London EC1 Y 8TZ, U.K.
alex.isakson.2@city.ac.uk

E. MAMMEN

Institute of Applied Mathematics, Heidelberg University,
Im Neuenheimer Feld 205, Heidelberg 69120, Germany
mammen@math.uni-heidelberg.de

J. P. NIELSEN

Bayes Business School, City St George's, University of London,
106 Bunhill Row, London EC1 Y 8TZ, U.K.
jens.nielsen.1@city.ac.uk

and C. PROUST–LIMA

Université de Bordeaux, INSERM, Bordeaux Population Health Research Center,
U1219, F-33000 Bordeaux, France
cecile.proust-lima@u-bordeaux.fr

Summary

We introduce a new concept for forecasting future events based on marker information. The model is developed in the nonparametric counting process setting under the assumptions that the marker is of so-called high quality and with a time-homogeneous conditional distribution. Despite the model having nonparametric parts, it is established herein that it attains a parametric rate of uniform consistency and uniform asymptotic normality. In usual nonparametric scenarios, reaching such a fast convergence rate is not possible, so one can say that the proposed approach is super-efficient. These theoretical results are employed in the construction of simultaneous confidence bands directly for the hazard rate. Extensive simulation studies validate and compare the proposed methodology with the joint modelling approach and illustrate its robustness for mild violations of the assumptions. Its use

in practice is illustrated in the computation of individual dynamic predictions in the context of primary biliary cirrhosis of the liver.

Some key words: Counting process; Dynamic prediction; Kernel hazard estimation; Nonparametric smoothing; Survival analysis.

1. Introduction

This paper investigates a novel approach to understanding future survival when the hazard depends on a developing marker process. Given some natural assumptions on the marker process and the marker-dependent hazard, we establish herein that the proposed technique achieves parametric rates of convergence. That is so even though the specifications of the marker process and the marker-dependent hazard are fully nonparametric. Nonparametric estimators with this property have been called super-efficient; see, e.g., Nielsen (1999). This should not be mixed up with the notion of super-efficiency used in asymptotic parametric statistics when discussing optimality of estimators. The estimator proposed here allows one to analyse and visualize survival forecasting using methodology related to square-root-n-consistent mathematical statistics. In particular, we derive uniform confidence bands based on approximations by Gaussian processes, proceeding similarly as in statistical inference based on cumulative distribution functions, Kaplan–Meier estimators or Nelson–Aalen estimators. While most of the mathematics of this paper is new, interpretation of its results requires only conventional intuition. A key asymptotic requirement of the proposed technique is that the nonparametric components involved are undersmoothed with vanishing asymptotic bias. In this sense our approach is related to semiparametric statistics. Implementation of the estimator in practice is facilitated via a fully automatic smoothing methodology, also developed herein, based on an adapted version of cross-validation. Our final approach is therefore fully data driven. Additionally, the present research develops theory for uniform confidence bands for future development of conditional hazards for individuals with a certain present marker level, thus allowing its first practical implementation.

Development of the new methodology contained in this article requires two crucial assumptions. The first one is that conditional hazards depend on time only through the value of the marker process at this time point. We regard such markers as being of high quality, a notion that goes back to Nielsen (1999); see also Nielsen (2000). Secondly, for the marker process, we make a Markov-type assumption that allows us to predict the further development of the marker. In our model we take a purely nonparametric view. The main intuition behind our new approach to modelling the full survival system was already given by Nielsen (2000) with some improved technical indications by Mammen & Nielsen (2007). However, the practicalities and full technical consequences of this new approach were never investigated and, as a result, the approach has never entered mainstream statistics nor has it yet been implemented on real data.

The proposed methodology applies in many fields such as data mining or asset-liability management; see Nielsen (2000). Below we discuss a health research example that requires the forecast of the hazard rate function as a disease progresses. Specifically, in analysing the primary biliary cirrhosis dataset of the Mayo Clinic (Therneau & Grambsch, 2000), the objective is the prediction of clinical progression of patients diagnosed with primary biliary cirrhosis, based on repeated measures of different biological markers and time to

clinical progression. There exist multiple examples in the biostatistical literature with similar modelling objectives, including, e.g., prostate-specific antigen and prostate cancer recurrence (Proust-Lima & Taylor, 2009) or CD4 counts and HIV infection (Fusaro et al., 1993; Cui et al., 2023).

For the analysis of such data, individual dynamic prediction techniques have been proposed, e.g., landmarking (Anderson et al., 1983; Van Houwelingen, 2007) or joint modelling (Henderson et al., 2000; Rizopoulos, 2012). The landmarking approach uses individuals alive at t and information on the biomarker up to t to fit a proportional hazard model and estimate the probability of surviving up to $t + s$ (Ferrer et al., 2019); see Van Houwelingen (2007) for an implementation where smoothing with respect to t has been included. The joint modelling methodology estimates the distribution of the biomarker and the time to event, and derives the conditional future event probability using past biomarker data (Proust-Lima & Taylor, 2009; Rizopoulos, 2012). In this model, the marker process is usually described through a linear mixed model with a stochastic component W_i for individual i, and the time to event is usually modelled via a mixed proportional hazard model that includes the same stochastic component W_i. The two submodels for the marker and survival are linked by sharing the same stochastic component W_i, and the marker process is thus conditionally independent of the survival times given W_i. Both approaches have been shown as being powerful tools in statistical inference, especially for individual dynamic prediction (Ferrer et al., 2019). However, both methodologies do not fully capture the stochastic structure of the data. This is so because in landmarking one has a separate model for each value of the landmark time t with biomarker data only considered up to t so that the two processes are not temporally and mutually linked for all times. This induces that predictions at landmark t are not consistent (in the sense of Jewell & Nielsen, 1993) with predictions at other landmark times (Suresh et al., 2017; Ferrer et al., 2019). In contrast, the joint model approach relies on the joint distribution of the two processes and is thus likely to provide consistent predictions. However, in practice, joint models suffer from simplifying parametric assumptions that highly reduce the flexibility and stochasticity of the model and association between the two processes (Ferrer et al., 2019). For instance, the shared stochastic component W_i is almost always limited to a vector of time-independent random effects.

This paper contains a full asymptotic study for the case of one-dimensional marker processes. We show that, under our two assumptions, the full survival system can be estimated with parametric rates. This has important technical consequences. For example, our resulting forecast of the future hazard can be estimated uniformly consistently for any individual starting the forecasting at any value of the marker process. Extended to this methodology, a wild bootstrap approach, see Mammen (1992), is developed and provides both pointwise and uniform confidence bands of the entire range of the future hazard. This is very powerful and yields a real alternative to the commonly used parametric approach. Because of the mathematical challenges of our approach, we have developed the full mathematical insight described above for the one-dimensional marker case only. Possible extensions to higher-dimensional markers are briefly discussed at the end of the paper.

2. Model formulation

Consider n individuals observed in the time interval $[0, T]$ with survival times $T_1, \ldots, T_n$. The methodology proposed herein is developed under the same model formulation as in Nielsen (2000). For $i = 1, \ldots, n$, let N_i be the counting process, which indicates the observed

event for the ith individual, and let Z_i be the exposure process taking values in $\{0, 1\}$, with 1 indicating that the ith individual is at risk to encounter the event of interest. Furthermore, additional information for every individual is available in the form of a one-dimensional predictable càdlàg marker process $X_i(t)$, $t \in [0, T]$. We assume that $N^{(n)} = (N_1, \ldots, N_n)$ is an n-dimensional counting process with respect to the increasing, right-continuous filtration $\mathcal{F}_t = \sigma\{N^{(n)}(s), X^{(n)}(s), Z^{(n)}(s); s \leqslant t\}$, $t \in [0, T]$, where $Z^{(n)} = (Z_1, \ldots, Z_n)$ and $X^{(n)} = (X_1, \ldots, X_n)$. The observed data are denoted by $\mathbb{X}_n = \{X^{(n)}(t), Z^{(n)}(t), N^{(n)}(t); t \in [0, T]\}$. When, for each individual i, $X_i(t) = x_i \in \mathbb{R}$ for all $t \in [0, T]$, the present formulation defaults to the time-invariant covariate setting.

The following assumptions are used throughout the article. The main assumptions are the first two.

Assumption 1 (*High-quality marker information*). The stochastic processes (N_1, X_1, Z_1), $\ldots$, (N_n, X_n, Z_n) are independent and identically distributed with predictable intensity

$$\lambda_i(t)\,\mathrm{d}t = E\{\mathrm{d}N_i(t) \mid \mathcal{F}_{t-}\} = \alpha\{X_i(t)\}Z_i(t)\,\mathrm{d}t,$$

where the conditional hazard $\alpha(\cdot)$ depends on time t only via marker $X_i(t)$.

Assumption 1 expresses the fact that the conditional hazard $\alpha(\cdot)$ depends only on the marker information and, in particular, not on time. Our second main assumption is a Markov-type condition and concerns the dynamics of the marker processes $X_i(t)$. In what follows, s can be seen as landmark time, i.e., the time at which the prediction is to be computed.

Assumption 2 (*Markov-type assumption on the marker process*). The conditional distribution of $X_i(s + t)$, given $\mathcal{F}_s$, $T_i \geqslant s + t$, $Z_i(s + t) = Z_i(s) = 1$, depends only on $X_i(s)$ and t and, in particular, not on s.

Assumptions 1 and 2 define a setting where it is possible to use previous data for future prediction based only on marker information. A situation, as stated in Assumption 1, where $\alpha(\cdot)$ depends only on the marker is described by Yong et al. (1997), who argued that time since infection of AIDS has little implication on its hazard. In this particular example, Assumption 1 is covering the fact that the count of CD4 cells in the blood is way more important than just the time since infection. Time is still part of the model through the time-dependent covariates X_i, N_i and Z_i. Time invariance of the marker process transition probabilities stated in Assumption 2 allows using already observed developments of patients in the past for future predictions. Methodology for assessing the validity of Assumptions 1 and 2 before applying the proposed approach in practice is provided in § B.2 of the Supplementary Material. The same section also contains numerical evidence on the sensitivity of the approach when either assumption is violated; the results indicate that the method is robust to mild violations of both assumptions.

For x_* in the interior of the support of $X_i(t)$, we make the following additional assumptions that are discussed in the Supplementary Material.

Assumption 3. The support A of $X_i(t)$ is a finite closed interval that does not depend on t. The conditional density $f(s, s+t, x, z)$ of $\{X(s), X(s+t)\}$, given $\{Z(s), Z(s+t)\} = (1, 1)$ exists, and is twice continuously differentiable with respect to x and z for $\delta_T \leqslant s \leqslant s + t \leqslant T - \delta_T$, $x, z \in A$. The conditional density $f(s, z)$ of $X(s)$, given $Z(s) = 1$ and the functions

$$\int_A f(s, s+t, x, z)\,\mathrm{d}z, \qquad \int_A \alpha(z) f(s, s+t, x, z)\,\mathrm{d}z$$

exist and are twice continuously differentiable with respect to x for $0 \leqslant s \leqslant s+t \leqslant T$, $x \in A$. The conditional hazard $\alpha(x)$ is twice continuously differentiable for $x \in A$ and bounded away from 0. For some constant $C > 0$, we have $|f(s, s+t_1, x, z) - f(s, s+t_2, x, z)| \leqslant C|t_1 - t_2|$ for $s, t_1, t_2 \geqslant 0$, $s + t_1, s + t_2 \leqslant T$, $\delta_T \leqslant t_1, t_2 \leqslant T - \delta_T$, $x, z \in A$.

Assumption 4. The expectations $\gamma(s) = E\{Z_i(s)\}$ and $\gamma(t, s) = E\{Z_i(t+s)Z_i(s)\}$ exist and are continuous. For some constant $C > 0$, we have $|\gamma(t_1, s) - \gamma(t_2, s)| \leqslant C|t_1 - t_2|$ for $s, t_1, t_2 \geqslant 0$, $s + t_1, s + t_2 \leqslant T$, $\delta_T \leqslant t_1, t_2 \leqslant T - \delta_T$. Furthermore, for $\delta_T \leqslant t \leqslant T - \delta_t$, the term $\Gamma(t, x_*)$ is bounded from below and, for $x \in A$, the function $E(x)$ is bounded from below, where

$$\Gamma(t, x_*) = \int_0^{T-t} \gamma(t, s) f(s, t+s, x_*, z)\,\mathrm{d}s\,\mathrm{d}z, \qquad E(x) = \int_0^T \gamma(s) f(s, x)\,\mathrm{d}s.$$

Assumption 5. Kernel K has bounded support, $[-1, 1]$ say, and is continuously differentiable on $[-1, 1]$. The bandwidths b_1 and b_2 are equal to $c_{b,1}n^{-\rho_1}$ or $c_{b,2}n^{-\rho_2}$ for some $c_{b,1}, c_{b,2} > 0$, $1/4 < \rho_1, \rho_2 < 1/3$.

Assumption 6. It holds that

$$\mathrm{pr}\{|X_i(s+t) - X_i(s)| \leqslant \delta,\ Z_i(s+t) = 1 \mid X_i(s) = z,\ Z_i(s) = 1\} \leqslant \delta\kappa(t)$$

for all $z \in A$, $0 \leqslant s < s+t \leqslant T$ and small enough $\delta > 0$, where κ is a positive function with $\int_0^T \kappa(t)\,\mathrm{d}t \leqslant C_\kappa$ for a constant $C_\kappa > 0$.

Assumption 7. For some constant $C > 0$, we have $E\{|X_i(s+t_1) - X_i(s+t_2)|^4\} \leqslant C|t_1 - t_2|^2$ for $s, t_1, t_2 \geqslant 0$, $s + t_1, s + t_2 \leqslant T$, $\delta_T \leqslant t_1, t_2 \leqslant T - \delta_T$.

In the sequel it is shown that, under these conditions, it is possible to estimate the marker conditional future hazard with starting point s,

$$h_{x,s}(t) = \mathrm{pr}\{T_i \in (s+t, s+t+\mathrm{d}t) \mid X_i(s) = x,\ Z_i(s+t) = Z_i(s) = 1\}/\mathrm{d}t, \tag{1}$$

where T_i is the survival time of individual i. Assumption 2 allows us to write $h_{x,s}(t) = h_x(t)$. We will use $h_{x,s}(t)$ if we want to underline when the prediction is made, but keep the notation $h_x(t)$ most of the time. Our main results contain limiting distributions for estimators of process $h_x(t)$ for a fixed value of x and for the consistency of uniform confidence bands based on wild bootstrap. Trivially, we have

$$h_{x,s}(t) = h_x(t) = \mathrm{pr}\{T_i \in (s+t, s+t+\mathrm{d}t) \mid X_i(s) = x,\ T_i > s+t\}/\mathrm{d}t,$$

under the additional assumption of noninformative censoring, i.e.,

$$\begin{aligned}&\mathrm{pr}\{T_i \in (s+t, s+t+\mathrm{d}t) \mid X_i(s) = x,\ T_i > s+t\}\\ &\quad = \mathrm{pr}\{T_i \in (s+t, s+t+\mathrm{d}t) \mid X_i(s) = x,\ Z_i(s+t) = Z_i(s) = 1\}.\end{aligned}$$

Under this assumption, our approach yields an estimator of the conditional survival function, defined by simply integrating the conditional hazard estimator appropriately.

Consequently, the prediction of the future conditional hazard becomes an estimation problem. Thus, our approach can be considered as an in-sample forecasting method. In-sample forecasting has been introduced using structural models by Martínez-Miranda et al. (2013) and has been used in reserving. See also Hiabu et al. (2016), Bischofberger et al. (2019) and Mammen et al. (2021). Our framework allows us to write the marker conditional future hazard as

$$h_x(t) = E[\alpha\{X_i(s+t)\} \mid X_i(s) = x,\, Z_i(s) = Z_i(s+t) = 1].$$

Let $K_b(\cdot) = b^{-1}K(\cdot/b)$ be a kernel with bandwidth b. By Assumption 5, K is a continuously differentiable function with bounded support. Estimators for $\alpha(z)$ and $h_x(t)$ have been proposed by Nielsen (2000) and we use them here as well. For bandwidths b_1 and b_2, define

$$\hat{\alpha}_{i,b_1}(z) = \frac{\sum_{k\neq i}\int_0^T K_{b_1}\{z - X_k(s)\}\,\mathrm{d}N_k(s)}{\sum_{k\neq i}\int_0^T K_{b_1}\{z - X_k(s)\}Z_k(s)\,\mathrm{d}s}$$

and

$$\hat{h}_{x,b_1,b_2}(t) = \frac{\sum_{i=1}^n \int_0^T \hat{\alpha}_{i,b_1}\{X_i(t+s)\}Z_i(t+s)Z_i(s)K_{b_2}\{x - X_i(s)\}\,\mathrm{d}s}{\sum_{i=1}^n \int_0^T Z_i(t+s)Z_i(s)K_{b_2}\{x - X_i(s)\}\,\mathrm{d}s}. \tag{2}$$

To simplify the notation, we also write $\hat{\alpha}_i$ and $\hat{h}_x$ for $\hat{\alpha}_{i,b_1}$ and $\hat{h}_{x,b_1,b_2}$ if dependence on bandwidths does not need to be highlighted. Estimator $\hat{\alpha}_i$ is a usual leave-one-out estimator for the hazard. Its use as an approximation of $\alpha(\cdot)$ is intuitive since $\hat{\alpha}_i$ can be thought of as the natural extension to the present setting of the local likelihood principle in modelling a constant hazard rate function; see also Nielsen & Linton (1995). Also, $\hat{h}_x(t)$ arises from using a kernel estimator of the conditional density of $X_i(s)$ and combining it with $\hat{\alpha}_i$. For achieving parametric rates, it is important that the bias terms arising in the smoothing are of order $o(n^{-1/2})$. This can be achieved by choosing the bandwidths b_1 and b_2 such that $b_1^2, b_2^2 = o(n^{-1/2})$; again, see Assumption 5. Our estimators also work when some data are right censored; this is controlled via the individual time-dependent exposure measures $Z_i(\cdot)$.

3. Asymptotic theory

This section formulates the main result of the present research that states that the proposed hazard estimator yields a parametric rate of uniform convergence and is asymptotically normal. On first sight this fact seems very surprising since we estimate the hazard rate nonparametrically. For the pointwise case, this has been already observed by Nielsen (2000); see also related models studied by Castellana & Leadbetter (1986), Kutoyants (2004), Bosq (2012) and Aeckerle-Willems & Strauch (2022), where parametric rates show up in nonparametric settings. Thus, these models differ from other nonparametric models where estimators only allow a slower convergence rate compared to parametric estimation problems. The main difference lies in the fact that typically in nonparametric problems only local

information can be used, whereas in our model all individuals with markers $X_i(\cdot)$ having visited the neighbourhood of x at some time point $s \leqslant T - t$ add information for the estimation of $h_x(t)$.

THEOREM 1. *Suppose that Assumptions 1–7 apply for an $x = x_*$ in the interior of the support of $X_i(t)$. Then, for $\delta_T > 0$, it holds that*

$$n^{-1/2}(\hat{h}_{x_*} - h_{x_*}) \to \mathbb{G}_{x_*}$$

in distribution, weakly in $\ell^\infty([\delta_T, T-\delta_T])$, where $\mathbb{G}_{x_*}$ is a tight Gaussian process taking values in $\ell^\infty([\delta_T, T-\delta_T])$ with mean 0 and covariance $\Sigma(t_1, t_2)$ stated in the Supplementary Material.

For the proof of Theorem 1, see the Supplementary Material. There, we also argue that the result of the theorem also holds for the boundary points x_* of A if in the smoothing step (2) the convolution kernel $K_{b_2}(x-u)$ is replaced by a boundary-corrected kernel or if we use local linear estimation instead of local constant smoothing. We conjecture that the result of Theorem 1 also holds if the interval $[\delta_T, T-\delta_T]$ is replaced by $[0, T-\delta_T]$. But this would require additional detailed assumptions on the speed of convergence of the joint density of $\{X_i(t_1), X_i(t_2)\}$ to infinity for $t_1 - t_2 \to 0$. Theorem 1 is used in the next section as the basis for constructing pointwise and uniform confidence bands. Implementation of the proposed estimator in practice requires the development of a data-driven, consistent bandwidth selection rule that is discussed in detail in the Supplementary Material.

4. CONFIDENCE BANDS

We now introduce a bootstrap procedure for the construction of pointwise and uniform confidence sets. For a critical discussion of confidence sets based on Gaussian approximations, see Bie et al. (1987). Our approach is a slight modification of wild bootstrap procedures recently proposed in counting process models (Beyersmann et al., 2013; Bluhmki et al., 2019). Because in our case the estimation error is not a martingale, we need to adapt this approach to our setting. We use the multiplier bootstrap, as presented by Chernozhukov et al. (2013). Fix x_* in the interior of the support of $X_i(t)$. Define

$$W_i(t) = \int_0^T \hat{\alpha}_{i,b_1}\{X_i(t+s)\} Z_i(t+s) Z_i(s) K_{b_2}\{x_*, X_i(s)\}\, \mathrm{d}s.$$

Then

$$n^{-1} \sum_{i=1}^n \hat{\Gamma}(t, x_*)^{-1} W_i(t) = \hat{h}_{x_*}(t),$$

$$\text{where} \quad \hat{\Gamma}(t,x) = n^{-1} \sum_{i=1}^n \int_0^{T-t} Z_i(t+s) Z_i(s) K_{b_2}\{x, X_i(s)\}\, \mathrm{d}s.$$

In the proof of Theorem 1 it is shown that $\hat{h}_{x_*}(t) = A(t) + B(t) + o_P(n^{-1/2})$, where

$$A(t) = n^{-1} \sum_{i=1}^{n} \int_0^T g_{t,x_*}\{X_i(s)\}\, \mathrm{d}M_i(s),$$

$$B(t) = n^{-1} \sum_{i=1}^{n} \int_0^T [\alpha\{X_i(t+s)\} - h_{X_i(s)}(t)] Z_i(t+s) Z_i(s) K_{b_2}\{x_*, X_i(s)\}\, \mathrm{d}s$$

and

$$g_{t,x}(z) = \frac{\int_0^{T-t} \gamma(t,s) f(s, t+s, x, z)\, \mathrm{d}s}{\int_0^T \gamma(s) f(s,z)\, \mathrm{d}s}.$$

Our bootstrap estimator of the distribution of the process $\hat{h}_{x_*}(\cdot) - h_{x_*}(\cdot)$ is the conditional distribution of $\bar{h}_{x,B}(\cdot) = A_B(\cdot) + B_B(\cdot)$, given all observations $\mathbb{X}_n = \{N_i(t), Z_i(t), X_i(t) \colon 1 \leqslant i \leqslant n,\ 0 \leqslant t \leqslant T \text{ with } Z_i(t) = 1\}$, where

$$A_B(t) = n^{-1/2} \sum_{i=1}^{n} \int_0^T \hat{g}_{i,t,x_*}\{X_i(s)\} V_i [\mathrm{d}N_i(s) - \hat{\alpha}_i\{X_i(s)\} Z_i(s)\, \mathrm{d}s],$$

$$B_B(t) = n^{-1/2} \sum_{i=1}^{n} V_i \{\hat{\Gamma}(t, x_*)^{-1} W_i(t, x_*) - \hat{h}_{x_*}(t)\}.$$

Here the V_i are independent and identically distributed normal random variables independent of (N_i, X_i, Z_i) with expectation 0 and variance 1. Furthermore, we have used the following leave-one-out estimator of $g_{t,x}(z)$:

$$\hat{g}_{i,t,x}(z) = n^{-1} \sum_{j=1, j\neq i}^{n} \int_0^{T-t} \hat{E}_j\{X_j(t+s)\}^{-1} K_{b_2}\{z, X_j(t+s)\} Z_j(t+s) \times Z_j(s) K_{b_2}\{x, X_j(s)\}\, \mathrm{d}s.$$

We apply the bootstrap to get an approximation for the quantiles of the distribution of the random variables $\sigma^{-1}_{\mathbb{G}_{x_*}}(t)\{\hat{h}_{x_*}(t) - h_{x_*}(t)\}$ for fixed t and for the $\sup_{t\in[\delta_T, T-\delta_T]} \sigma^{-1}_{\mathbb{G}_{x_*}}(t)|\hat{h}_{x_*}(t) - h_{x_*}(t)|$. Denote the asymptotic distributions of the random variable by $\mathrm{pr}^{\mathbb{G}^*_x(t)}$ and of the supremum by $\mathrm{pr}^{\mathbb{G}^*_{x,M}}$. Here, $\sigma^2_{\mathbb{G}_{x_*}}(t)$ is the variance of $\hat{h}_{x_*}(t) - h_{x_*}(t)$. The bootstrap estimators of these quantities are given by the conditional distribution $\mathrm{pr}^{h^*_{x_*,B}(t)}$ of

$$h^*_{x_*,B}(t) = \hat{\sigma}_{\mathbb{G}_{x_*}}(t)^{-1}\{A_B(t) + B_B(t)\},$$

and of the conditional distribution $\mathrm{pr}^{h^*_{x_*,B,M}}$ of $\sup_{\delta_T \leqslant t \leqslant T-\delta_T} |h^*_{x_*,B}(t)|$. Here $\hat{\sigma}^2_{\mathbb{G}_{x_*}}(t)$ is an estimator of the variance of $\hat{h}_{x_*}(t)$ and is equal to $E\{\bar{h}_{x_*,B}(\cdot)^2 \mid \mathbb{X}_n\}$. In the proof of Theorem 2 below, which is provided in the Supplementary Material, we argue that $\hat{\sigma}^2_{\mathbb{G}_{x_*}}(t)$ converges to the variance $\sigma^2_{\mathbb{G}_{x_*}}(t)$ of the limiting Gaussian process. The consistency of the bootstrap approach is established in the next theorem.

THEOREM 2. *Under the assumptions of Theorem 1, $h^*_{x_*,B}(t)$ approximates the distribution of $\mathbb{G}^*_{x_*}(t)$ pointwise and uniformly, i.e., in the pointwise case it holds that*

$$d_{\mathrm{K}}(\mathrm{pr}^{h^*_{x_*,B}(t)}, \mathrm{pr}^{\mathbb{G}^*_{x_*}(t)}) \to 0$$

in probability for all $t \in (0, T)$ and in the uniform case it holds that

$$d_{\mathrm{K}}(\mathrm{pr}^{h^*_{x_*,B,M}}, \mathrm{pr}^{\mathbb{G}^*_{x_*,M}}) \to 0$$

in probability, where $d_{\mathrm{K}}(\cdot,\cdot)$ denotes the Kolmogorov distance.

Implementation details for the $1-\alpha$ pointwise bootstrap confidence intervals and uniform bootstrap confidence bands are provided in the Supplementary Material.

5. NUMERICAL EXAMPLES

5.1. *Simulation study*

The simulation study in this section assesses the performance of the proposed estimator and the accuracy level of the associated confidence sets. We generated data according to the high-quality marker (HQM) model described in § 2 with a univariate marker. We considered three different functional shapes for the marker-only hazard, $\alpha_1(x) = \exp(2x-2)/15$, $\alpha_2(x) = 4(x-0.3)^4$ and $\alpha_3(x) = 4[1+\exp\{-4(x-1)\}]^{-1}$, displayed in Fig. 5 in the Supplementary Material. These hazards are dependent only on a marker measurement x and, thus, satisfy Assumption 1. Marker X is simulated in the following way: first, we simulate a Gaussian random walk $G(t_g)$ with jump sizes that vary according to a normal distribution $N(0, 0.07^2)$ on the grid $\{t_d/10\colon t_d = 0,\ldots,100\} \ni t_g$ and with uniform randomly chosen starting points $G(0) = x_0 = 0.1, 0.2, \ldots, 0.9$. The resulting continuous process, marker $X(t)$ for all $t \in [0, 10]$, was derived via linear interpolation between the values $G(t_g)$. To mimic cohort follow-ups, we then considered that this marker was only observed at individual-specific discrete follow-up visits $\tilde{t} = a + D$, where $a = 1,\ldots,9$ and $D \sim N(0, 0.07^2)$. Since the Gaussian random walk is a homogeneous Markov process, Assumption 2 is also satisfied.

Given the hazard α_j, $j = 1, 2, 3$, and marker X, we then simulate n independent one-jump counting processes, each with intensity $\alpha_j\{X(t)\}$ before a jump and 0 after. If a jump did not happen before 10, the survival time T was right censored by 10.

We evaluated six settings based on samples of size $n = 300, 600$ and the three marker-only hazards. In each setting we calculated 1000 realizations of estimator $\hat{h}_x(t)$ defined in (2) together with local and uniform confidence bands; see also (A1) and (A2) in the Supplementary Material. This was done for marker values x equal to $q_{0.1}, q_{0.25}, q_{0.5}, q_{0.75}, q_{0.9}$, where q_z is the empirical z quantile of all observed marker values. We compared $\hat{h}_x(t)$ with the true hazard $h_x(t)$ defined in (1). Since the true hazard $h_x(t)$ does not have a closed form, it was numerically approximated using the relation $h_x(t) = \partial/\partial t\{-\log S_x(t)\}$, where $S_x(t)$ was approximated through simulations. The bandwidths were set to $b_1 = b_2 = b$ with b chosen as a minimizer of the mean integrated squared error

$$mise(h_x, \hat{h}_{x,b,b}) = E\left[\int_0^{10} \{\hat{h}_x(t) - h_x(t)\}^2 \,\mathrm{d}t\right].$$

Different values of b are selected for each combination of x, n and α. We did not simulate the cross-validation bandwidths as proposed in the Supplementary Material because, due to the large number of settings, such a calculation would be computationally too complex.

Under the aforementioned settings, we performed extensive simulations to quantify the relative bias of $\hat{h}_x(t)$ and the coverage rates of the associated pointwise and uniform confidence bands. Because of space restrictions, graphical illustrations of the results are deferred to the Supplementary Material. In summary, the simulations indicate that the methodology generally provides a very small bias, which decreases as the sample size increases. Furthermore, the standard deviation is much larger compared to the bias, which aligns with the theoretical findings. Nevertheless, for marker-only hazards with a very low value, the relative bias is substantial. This is not surprising, as a low marker-only hazard value leads to a smaller number of observed events that do not allow for accurate estimations. Increased bias is also observed at boundary values of time, e.g., at the first or last year that affect the coverage rates of the 95% pointwise and uniform confidence bands, as can indicatively be seen in Fig. 1. The figure suggests that boundary effects in combination with small marker-only hazard values increase the coverage error of the proposed technique. In the Supplementary Material we argue that increasing the sample size slightly improves the coverage error (see Fig. 10 therein); nevertheless, effective treatment of the issue necessitates the use of a local linear estimator, i.e., estimator $\tilde{h}_x(t)$ that is defined in the Supplementary Material.

Additional simulations were performed for a smaller sample size of $n = 50$ patients. The estimator captured the key elements of the marker-only hazard and gave reasonable predictions, but the quality of inference was not perfect (results not shown). In the perspective of proposing prediction tools, we encourage the use of large enough sample sizes, as recommended in the literature (e.g., Riley et al., 2020). Indeed, too small samples may provide inaccurate individual predictions, in particular, in the case of complex relationships with the marker.

5.2. *Comparison with the joint modelling methodology*

In biomedical research, the joint modelling methodology of the longitudinal marker and the time to event have been proposed to compute dynamic individual predictions (Rizopoulos, 2012; Ferrer et al., 2019). Here we compare the HQM estimator and joint modelling estimators when data are generated according to the framework described in § 5.1. The comparison is based on the values of the model's estimated MISE, area under the curve and Brier score metrics. The future conditional survival function (given the information $\mathcal{F}_s$ up to s) is estimated by

$$S_\star^{\text{TRUE}}(s+t \mid \mathcal{F}_s) = \text{pr}\{T_\star > s+t \mid \mathcal{X}_\star(s), T_\star > s\}, \tag{3}$$

where $\mathcal{X}_\star(s)$ is the marker history of an individual $\star$ before landmark time s.

The joint modelling estimator consists of two parametric submodels that are linked by a shared latent structure. We consider here the classical specification with a linear mixed model for the longitudinal marker measurements and a proportional hazard model for the time to event:

$$\begin{aligned} X_i(t) &= m_i(t) + \varepsilon_i(t) = F_i(t)^{\mathrm{T}}\beta + R_i(t)^{\mathrm{T}}B_i + \varepsilon_i(t), \\ \lambda_i(t) &= \lambda_0(t)\exp\{W_i^{\mathrm{T}}\gamma + m_i(t)\eta\}. \end{aligned}$$

Fig. 1. Coverage rate of the pointwise (red) and uniform (blue) confidence bands for the marker-only hazards α_j, $j = 1, 2, 3$, for $n = 300$ individuals and marker values x at the 0.1, 0.25, 0.50, 0.75 and 0.90 quantiles of the empirical marker distribution. The exact rates for the uniform confidence band computed on the total time window (blue) or computed while ignoring the first and last year (green) are also reported. The confidence bands are based on 1000 realizations and a bootstrap with 1000 repetitions.

Here $F_i(t)$ and $R_i(t)$ are covariate vectors, including functions of time, that are associated with the vector of fixed effects $\beta \in \mathbb{R}^p$ and the vector of individual random effects $B_i \sim N(0, D)$ with unknown D, respectively. The measurement errors $\varepsilon_i(t) \sim N(0, \sigma^2)$ and are independent of B_i. The instantaneous risk of event is defined according to the baseline hazard λ_0, and a linear predictor that includes covariates W_i and a function of

the marker trajectory, in this example $m_i(t)$, the underlying true current marker level of individual i. In this specific setting, the baseline hazard λ_0 is approximated by cubic splines with five internal knots, and considered a linear regression with time, i.e., $F_i(t) = t$, $R_i(t) = t$, and no adjustment for covariates, i.e., $W_i = 0$. The joint model is fitted on all the longitudinal and survival information within the maximum joint likelihood framework using the `JM` package (Rizopoulos, 2010) in `R` (R Development Core Team, 2025). The numerical integration involved in the loglikelihood computation is performed by a pseudo-adaptive Gauss–Hermite quadrature with nine knots, and optimization of the loglikelihood was achieved by an expectation-maximization algorithm. See Rizopoulos (2012) and Ferrer et al. (2019) for further details.

The two methodologies were trained on samples of $n = 300, 600$ individuals and the marker-only hazards $\alpha_1, \alpha_2, \alpha_3$, corresponding to a total of six scenarios. Their predictive performance was then compared on an external sample of 100 new individuals. Once the joint modelling was fitted, the future conditional survival function (3) was computed at different landmark times s and for different horizons t, using the parametric fit described above. For the HQM estimator, we used

$$\hat{S}_\star^{\mathrm{HQM}}(s + t \mid \mathcal{F}_s) = \exp\left\{ - \int_0^t \hat{h}_{x_\star}(u)\, \mathrm{d}u \right\},$$

where $x_\star$ is the last marker measurement $X_\star(s)$ of an individual $\star$ before landmark time s. Note that, since the marker is Markov and the marker-only hazards do not directly depend on time, the future conditional hazard $h_x(t)$ is independent of the landmark time s. For comparing the predictive performances of the two approaches, we considered three different landmark times, 1.5, 3.5 and 5.5, leading to 18 cases. For each case, we computed 100 realizations of the conditional survival estimate of each individual $\star$ based on their marker data up to the landmark time s. Implementation details for all three metrics are provided in the Supplementary Material, along with graphical illustrations of all simulation results.

In summary, in all cases of the MISE and the area-under-the-curve simulations (36 instances in total) the HQM model systematically outperformed the joint modelling approach. This is probably due to the parametric specification of the joint modelling that assumes a log-linear form for the hazard that is not compatible with the α_2 and α_3 cases. For α_1, although both approaches yield close MISES, the HQM model performs slightly better than joint modelling in all instances. This might be due to the misspecification of the parametric model for the marker that was generated as a random walk rather than an individual-specific linear trajectory. The HQM model is better than joint modelling in the Brier score simulations too; however, in this case, their Brier score metric values are very close to each other and, especially for α_2, they can be regarded as equivalent.

6. Application

The methodology developed in the previous sections is applied for predicting the clinical progression of patients diagnosed with primary biliary cirrhosis (PBC) of the liver, based on the publicly available dataset pbc2 of the Mayo Clinic (Therneau & Grambsch, 2000). The dataset contains information on a randomized clinical trial of the D-penicillamine versus placebo for a total of 312 patients who met certain eligibility criteria. The patients were followed up for a maximum of 13 years between 1974 and 1984. Repeated measures of

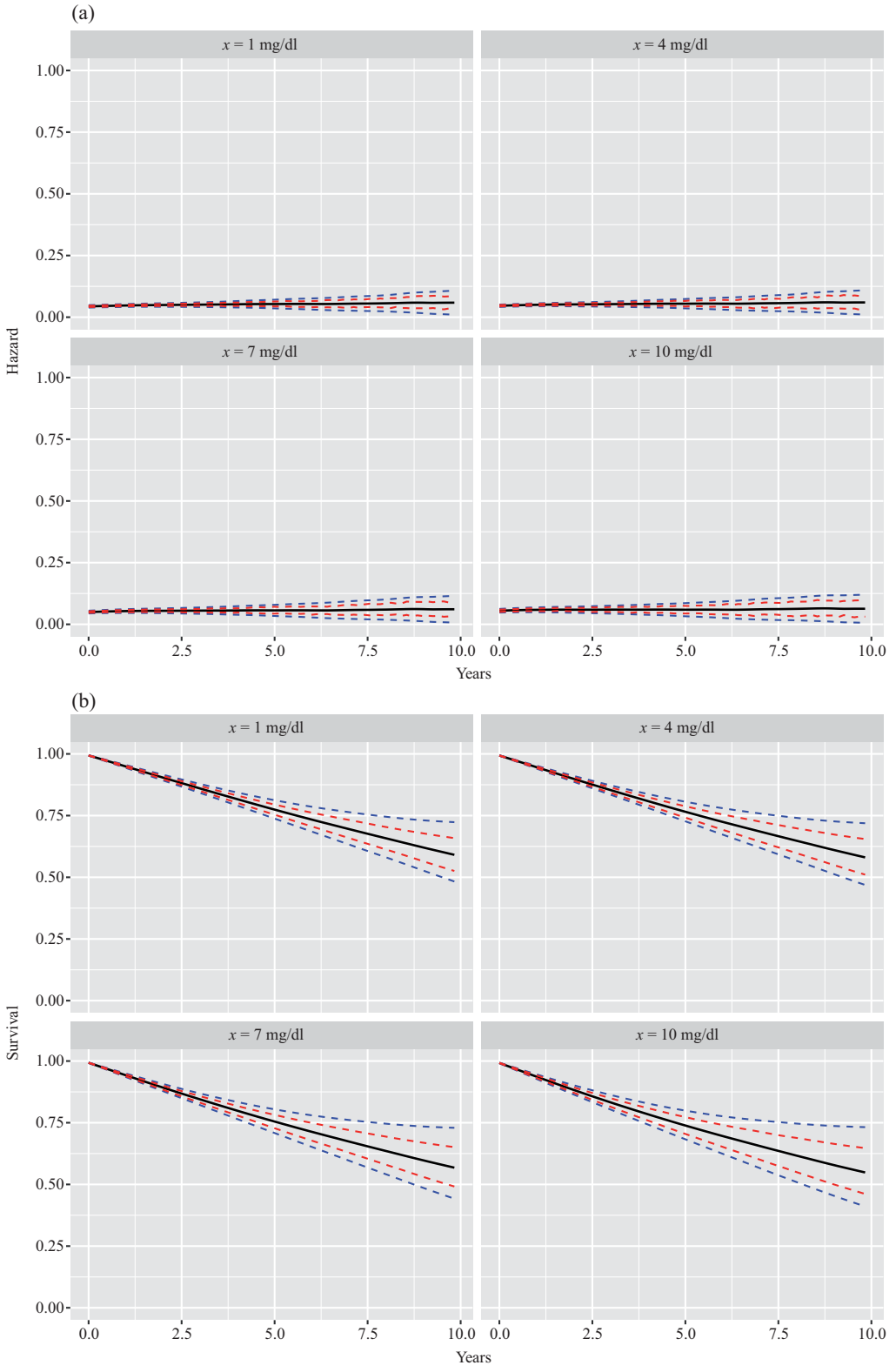

Fig. 2. The HQM estimator (solid) with pointwise (red dashes) and uniform (blue dashes) confidence bands for the future, conditional on bilirubin, (a) hazard and (b) survival function for the next 10 years.

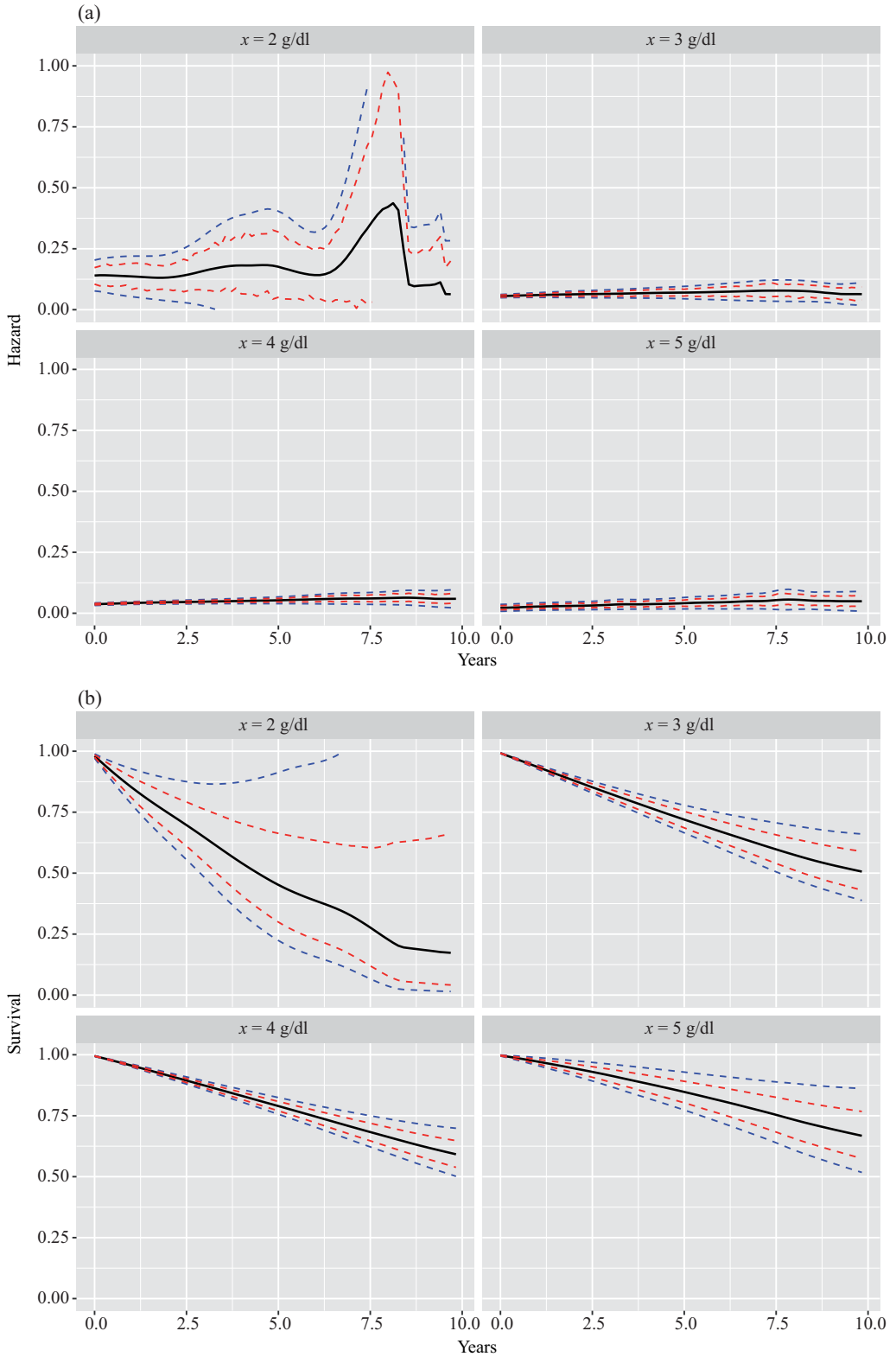

Fig. 3. The HQM estimator (solid) with pointwise (red dashes) and uniform (blue dashes) confidence bands for the future, conditional on albumin, (a) hazard and (b) survival function for the next 10 years.

different biological markers were collected over time, including albumin, bilirubin and alkaline phosphatase along with time to clinical progression defined as the minimum date of death or transplantation.

We consider two high-quality markers for the prediction of clinical progression in PBC. These are bilirubin (normal range 0.2 to 1.2 mg/dl) and albumin (normal range 3.4 to 5.4 g/dl). Both are produced by the liver and according to Lammers et al. (2015) and Hirschfield et al. (2018) are known to correlate with death. A high concentration of bilirubin and a low concentration of albumin indicates a liver dysfunction. Under Assumption 1, the risk of clinical progression is independent of time, given the current marker level. Assumption 2 expresses the fact that the future level of either marker depends on its current level and the time elapsed in between the patient's evaluations. For our methodology to be applicable, we need (a) to assume that the markers are observed without measurement error and (b) to interpolate between the marker measurements and extrapolate linearly at the last measurement. This is necessary as we assume marker X to be continuous. More details are discussed by Fusaro et al. (1993) and Nielsen (1999). Furthermore, we assume that the exposure Z for patient i is $Z_i(s) = I(s < t_i)$, where t_i is the time of clinical progression of patient i and $I(\cdot)$ is the indicator function.

Our model predictions based on each marker are illustrated in Figs. 2 and 3 with four different marker values: 1, 4, 7, 10 mg/dl for bilirubin, and 2, 3, 4, 5 g/dl for albumin. The values are chosen evenly based on the range of around 90% of the respective marker data, which is 0.1–10 mg/dl for bilirubin and 1.5–5 g/dl for albumin. Both the future conditional hazard and the future conditional survival function are predicted for 10 years with the 95% pointwise and uniform confidence bands. In each example, $h_x(t)$ is estimated by $\hat{h}_{x,b_1,b_2}(t)$, given in (2) with bandwidths determined by cross-validation; full details on obtaining the bandwidths for each example are provided in the Supplementary Material. The predicted hazard of clinical progression and the progression-free survival do not seem to change with the bilirubin level. For albumin though, we highlight a higher risk of clinical progression in patients with lower concentrations of albumin. For instance, the predicted five-year progression-free survival decreases from approximately 0.85 for an albumin concentration of 5 g/dL to 0.48 for an albumin concentration of 2 g/dL. Because right censoring is present, the study is not completely run off and, thus, the survival function does not go all the way to zero.

For the albumin marker, the first level considered, $x = 2$ g/dl, corresponds to the first percentile of the marker's distribution and therefore lies on the boundary. Consequently, the confidence bands of the estimator are wider than those of the rest marker levels, due to the underperformance of the local constant estimator on the boundary. This issue has also been demonstrated in the coverage rate simulations in the Supplementary Material where it is also shown that boundary correction remedies this deficiency. Finally, see the Supplementary Material for a simultaneous application of the HQM and joint modelling methodologies using the pbc2 dataset.

7. Conclusion

This paper is a first piece of work integrating the super-efficient approach into mainstream theory of mathematical statistics. The estimation procedure along with confidence bands and a bandwidth selection rule is implemented in the `R` package `HQM` (Bagkavos et al., 2022) together with an estimation example based on time-invariant covariates. We showed

in an illustration and a comparative simulation study that our new method is already of practical use in some important applied problems such as the dynamic individual prediction of health events. However, further extensions are needed so that the method becomes fully operational in applied statistics.

First, an immediate and unproblematic extension would be to allow for categorical variables driving either the marker process or the marker-dependent hazard or both. Second, an extension of the present methodology to multi-dimensional markers may help in some contexts to achieve the high-quality property, i.e., the assumption that the hazard depends only on the marker process. However, this would require a quite different and much more complex mathematical theory that needs very careful modelling of the marker process. An interesting alternative where a large part of the mathematical theory of this paper could be used would be based on one-dimensional marker indices. In such an approach one firstly reduces the markers to a one-dimensional composite indicator defined as a weighted sum of the markers where weights are internally estimated, and secondly applies the super-efficient methodology as described here to the one-dimensional composite indicator. This would thus lead to a semiparametric extension of the super-efficient model for multi-dimensional markers. Another semiparametric extension multiplies the fully nonparametric marker-only hazard model by a parametric dependency on time. Extending our modelling framework to such a semiparametric marker-dependent hazard would still result in a parametric rate of convergence allowing uniform confidence bands based on transformed Gaussian processes.

Finally, we assumed that the marker was measured without error and in continuous time, with the latter assumption being achieved by interpolation. Extending the methodology to handle noisy and intermittently missing marker observations, as done in the parametric joint modelling approach with an underlying latent marker process (Rizopoulos, 2012), is another direction of future research.

Supplementary material

The Supplementary Material includes additional simulation results and illustrations, discussion of the assumptions, auxiliary lemmas, and proofs of Theorems 1 and 2.

References

Aeckerle-Willems, C. & Strauch, C. (2022). Sup-norm adaptive simultaneous drift estimation for ergodic diffusions. *Ann. Statist.* **50**, 3484–509.

Anderson, J. R., Cain, K. C. & Gelber, R. D. (1983). Analysis of survival by tumor response. *J. Clin. Oncol.* **1**, 710–19.

Bagkavos, D., Isakson, A., Proust-Lima, P., Mammen, E. & Nielsen, J. P. (2022). *HQM: superefficient estimation of future conditional hazards based on marker information*. R package version 0.1.1.

Beyersmann, J., Termini, S. D. & Pauly, M. (2013). Weak convergence of the wild bootstrap for the Aalen–Johansen estimator of the cumulative incidence function of a competing risk. *Scand. J. Statist.* **40**, 387–402.

Bie, O., Borgan, Ø. & Liestøl, K. (1987). Confidence intervals and confidence bands for the cumulative hazard rate function and their small sample properties. *Scand. J. Statist.* **14**, 221–33.

Bischofberger, S. M., Hiabu, M., Mammen, E. & Nielsen, J. P. (2019). A comparison of in-sample forecasting methods. *Comp. Statist. Data Anal.* **137**, 133–54.

Bluhmki, T., Dobler, D., Beyersmann, J. & Pauly, M. (2019). The wild bootstrap for multivariate Nelson–Aalen estimators. *Lifetime Data Anal.* **25**, 97–127.

Bosq, D. (2012). *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction* (Lecture Notes Statist. **110**). New York: Springer.

Castellana, J. & Leadbetter, M. (1986). On smoothed probability density estimation for stationary processes. *Stoch. Proces. Appl.* **21**, 179–93.

Chernozhukov, V., Chetverikov, D. & Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41**, 2786–819.
Cui, Y., Michael, H., Tanser, F. & Tchetgen Tchetgen, E. (2023). Instrumental variable estimation of the marginal structural Cox model for time-varying treatments. *Biometrika* **110**, 101–18.
Ferrer, L., Putter, H. & Proust-Lima, C. (2019). Individual dynamic predictions using landmarking and joint modelling: validation of estimators and robustness assessment. *Statist. Meth.: Med. Res.* **28**, 3649–66.
Fusaro, R. E., Nielsen, J. P. & Scheike, T. H. (1993). Marker-dependent hazard estimation: an application to AIDS. *Statist. Med.* **12**, 843–65.
Henderson, R., Diggle, P. & Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1**, 465–80.
Hiabu, M., Mammen, E., Martinez-Miranda, M. D. & Nielsen, J. P. (2016). In-sample forecasting with local linear survival densities. *Biometrika* **103**, 843–59.
Hirschfield, G., Dyson, J., Alexander, G. & Chapman, M. A. (2018). The British Society of Gastroenterology/UK-PBC primary biliary cholangitis treatment and management guidelines. *Gut* **67**, 1568–94.
Jewell, N. P. & Nielsen, J. P. (1993). A framework for consistent prediction rules based on markers. *Biometrika* **80**, 153–64.
Kutoyants, Y. A. (2004). *Statistical Inference for Ergodic Diffusion Processes*. London: Springer.
Lammers, W. J., Hirschfield, G. M., Corpechot, C., Nevens, F., Lindor, K. D., Janssen, H. L., Floreani, A., Ponsioen, C. Y., Mayo, M. J., Invernizzi, P. et al. (2015). Development and validation of a scoring system to predict outcomes of patients with primary biliary cirrhosis receiving ursodeoxycholic acid therapy. *Gastroenterology* **149**, 1804–12.
Mammen, E. (1992). Bootstrap, wild bootstrap, and asymptotic normality. *Prob. Theory Rel. Fields* **93**, 439–55.
Mammen, E., Martínez-Miranda, M. D., Nielsen, J. P. & Vogt, M. (2021). Calendar effect and in-sample forecasting. *Insur.: Math. Econ.* **96**, 31–52.
Mammen, E. & Nielsen, J. P. (2007). A general approach to the predictability issue in survival analysis with applications. *Biometrika* **94**, 873–92.
Martínez-Miranda, M. D., Nielsen, J. P., Sperlich, S. & Verrall, R. J. (2013). Continuous chain ladder: reformulating and generalising a classical insurance problem. *Expert Syst. Appl.* **40**, 5588–603.
Nielsen, J. P. (1999). Super-efficient hazard estimation based on high-quality marker information. *Biometrika* **86**, 227–32.
Nielsen, J. P. (2000). Super-efficient prediction based on high-quality marker information. *ASTIN Bull.* **30**, 295–303.
Nielsen, J. P. & Linton, O. B. (1995). Kernel estimation in a nonparametric marker dependent hazard model. *Ann. Statist.* **23**, 1735–48.
Proust-Lima, C. & Taylor, J. M. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach. *Biostatistics* **10**, 535–49.
R Development Core Team (2025). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.
Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E., Martin, G. P., Reitsma, J. B., Moons, K. G. M., Collins, G. & van Smeden, M. (2020). Calculating the sample size required for developing a clinical prediction model. *Br. Med. J.* **368**, m441.
Rizopoulos, D. (2010). JM: an R package for the joint modelling of longitudinal and time-to-event data. *J. Statist. Software* **35**, 1–33.
Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. Boca Raton, FL: Chapman and Hall/CRC.
Suresh, K., Taylor, J. M., Spratt, D. E., Daignault, S. & Tsodikov, A. (2017). Comparison of joint modeling and landmarking for dynamic prediction under an illness-death model. *Biomet. J.* **59**, 1277–300.
Therneau, T. M. & Grambsch, P. M. (2000). The Cox model. In *Modeling Survival Data: Extending the Cox Model*, pp. 39–77. New York: Springer.
Van Houwelingen, H. C. (2007). Dynamic prediction by landmarking in event history analysis. *Scand. J. Statist.* **34**, 70–85.
Yong, F. H., Taylor, J. M., Bryant, J. L., Chmiel, J. S., Gange, S. J. & Hoover, D. (1997). Dependence of the hazard of AIDS on markers. *AIDS* **11**, 217–28.

[*Received on* 21 *December* 2023. *Editorial decision on* 17 *December* 2024]