

Calibrating parametric subject-specific risk estimation

By T. CAI

Department of Biostatistics, Harvard University, Boston, Massachusetts 02115, U.S.A.
tcai@hsph.harvard.edu

L. TIAN

*Department of Health Research and Policy, Stanford University School of Medicine,
Stanford, California 94305, U.S.A.*
lutian@stanford.edu

HAJIME UNO

*Department of Biostatistics and Computational Biology, Dana Farber Cancer Institute,
Boston, Massachusetts 02115, U.S.A.*
huno@jimmy.harvard.edu

SCOTT D. SOLOMON

Cardiovascular Division, Brigham & Women's Hospital, Boston, Massachusetts 02115, U.S.A.
ssolomon@partners.org

AND L. J. WEI

Department of Biostatistics, Harvard University, Boston, Massachusetts 02115, U.S.A.
wei@hsph.harvard.edu

SUMMARY

For modern evidence-based medicine, decisions on disease prevention or management strategies are often guided by a risk index system. For each individual, the system uses his/her baseline information to estimate the risk of experiencing a future disease-related clinical event. Such a risk scoring scheme is usually derived from an overly simplified parametric model. To validate a model-based procedure, one may perform a standard global evaluation via, for instance, a receiver operating characteristic analysis. In this article, we propose a method to calibrate the risk index system at a subject level. Specifically, we developed point and interval estimation procedures for t -year mortality rates conditional on the estimated parametric risk score. The proposals are illustrated with a dataset from a large clinical trial with post-myocardial infarction patients.

Some key words: Cardiovascular diseases; Cox model; Nonparametric functional estimation; Risk index; ROC analysis; Survival analysis.

1. INTRODUCTION

The choice of disease prevention or management strategy for an individual from a population of interest is often made based on his/her risk of experiencing a certain clinical event during a specific follow-up time period. Subjects classified as high risk may be recommended for intensive

prevention or therapy. The risk is usually estimated via a parametric or semiparametric regression model using baseline risk factors. For example, [Eagle et al. \(2004\)](#) created a simple bedside risk index system to predict six-month post-discharge mortality for patients hospitalized due to acute coronary syndrome. A proportional hazards model ([Cox, 1972](#)) was utilized to fit the data on post-discharge follow-up mortality and all potential predictors taken at admission as well as during hospitalization. The six-month post-discharge mortality rate is then estimated with the fitted model for each future individual patient.

An empirical model is merely an approximation to a true model, and it is crucial to assess its performance for appropriate clinical decision making. In addition to checking the fitted model with routine goodness-of-fit techniques, the model-based scoring system is often evaluated or validated globally based on specific summary measures, such as the receiver operating characteristic curve ([Pepe et al., 2004](#); [Wang et al., 2006](#); [Ware, 2006](#); [Cook, 2007](#)), expected Brier score ([Gerds & Schumacher, 2006](#)) or the net reclassification index ([Pencina et al., 2008](#)). For example, [Uno et al. \(2007\)](#) developed procedures for evaluating the accuracy of empirical parametric models in predicting t -year survival with respect to various global accuracy measures. For the aforementioned post-discharge mortality example, the investigators used an independent cohort to validate their system with an overall c -statistic. However, even if the risk index system is acceptable using a global average measure, for the group of patients with the same model-based risk score, the corresponding parametric risk estimator may not approximate its true mortality rate well. Consequently, clinicians may be misguided in choosing therapies for treating these individuals.

In this paper, we propose a procedure for calibrating the parametric risk estimates. Specifically, under a general survival analysis setting, a consistent estimation procedure is proposed for t -year mortality rates of patients grouped by a parametric risk score. Furthermore, we provide pointwise and simultaneous confidence intervals for such risks over a range of estimated risk scores. These interval estimates quantify the precision of our consistent point estimator. Moreover, the upper or lower bounds of the intervals provide valuable information for complex cost-benefit decision making. Pointwise confidence interval estimates of the true average risk may be of interest to individual patients with a specific risk score. Simultaneous interval estimates can be used to identify subgroups with differential risks and to assist in selecting target populations for appropriate interventions. Our proposal is also useful for calibrating a risk index system when applied to populations different from the study population. For example, the Framingham Risk Score ([Anderson et al., 1991](#)) for predicting the risk of coronary heart disease was developed based on a U.S. population. Its applicability to other populations has been extensively investigated in recent years ([D'Agostino et al., 2001](#); [Brindle et al., 2003](#); [Zhang et al., 2005](#)). For a new population, our procedure can be used to assess the true risk of coronary heart disease for subjects with any given Framingham Risk Score.

2. A CONSISTENT ESTIMATOR FOR THE MEAN RISK OF SUBJECTS WITH THE SAME MODEL-BASED RISK SCORE

Consider a subject randomly drawn from the study population. Let \tilde{T} be the time to the occurrence of a specific event since baseline and U be the corresponding set of risk factors/markers ascertained at baseline; some marker values may be collected repeatedly over time. In this paper, we assume that markers are measured at a well-defined time zero. Discussions on the setting where time zero is not well defined are given in §5. Now, assume that \tilde{T} has a continuous distribution given U . Also, let $\mathcal{P}(u) = \text{pr}(\tilde{T} < t_0 \mid U = u)$. The event time \tilde{T} may be censored by a random variable C , which is assumed to be independent of \tilde{T} and U . For \tilde{T} , one can only observe $T = \min(\tilde{T}, C)$ and $\delta = I(\tilde{T} \leq C)$, where $I(\cdot)$ is the indicator function. Our data $\{(T_i, \delta_i, U_i), i = 1, \dots, n\}$ consist of n independent copies of (T, δ, U) .

One may estimate $\mathcal{P}(u)$ fully nonparametrically. Li & Doss (1995) and Nielsen (1998) considered local linear estimators of the conditional hazard function by smoothing over the q -dimensional U . However, such nonparametric estimates may not behave well when $q > 1$ and n is not large. A standard, feasible way to reduce the dimension of u is to approximate $\mathcal{P}(u)$ with a working parametric or semiparametric model such as the proportional hazards model:

$$\mathcal{P}(u) = g\{\log \Lambda(t_0) + \beta^T x\}, \tag{1}$$

where $g(s) = 1 - \exp(-e^s)$, $\Lambda(\cdot)$ is the working baseline cumulative hazard function for \tilde{T} ; x , a $p \times 1$ vector, is a function of u ; and β is an unknown vector of regression parameters.

To obtain an estimate for $\mathcal{P}(u)$ via (1), one may employ the maximum partial likelihood estimator $\hat{\beta}$ for β with all mortality information from the data collected up to time t_0 . That is, $\hat{\beta}$ is the maximizer of the log partial likelihood function,

$$\sum_{i=1}^n \int_0^{t_0} \left[\beta^T X_i - \log \left\{ \sum_{j=1}^n Y_j(t) \exp(\beta^T X_j) \right\} \right] dN_i(t),$$

where $N_i(t) = I(T_i \leq t)\delta_i$ and $Y_i(t) = I(T_i \geq t)$, for $i = 1, \dots, n$. When the model (1) is correctly specified, $\hat{\beta}$ consistently estimates the true value of β . On the other hand, if there is no vector ζ such that $\text{pr}(T_1 > T_2 \mid \zeta^T X_1 > \zeta^T X_2, T_2 \leq t_0) = 1$, then $\hat{\beta}$ converges to a finite constant β_0 , as $n \rightarrow \infty$, even when the model (1) is misspecified (Hjort, 1992). Alternatively, one may use an estimate of β by fitting a global Cox model without truncating at t_0 , denoted by $\tilde{\beta}$. When the Cox model fits the data well, $\tilde{\beta}$ may be more efficient than $\hat{\beta}$.

Next, we estimate $\Lambda(t)$ in (1) with Breslow’s estimator (Kalbfleisch & Prentice, 2002),

$$\hat{\Lambda}(t) = \sum_{i=1}^n \int_0^t \frac{dN_i(s)}{\sum_{j=1}^n Y_j(s) \exp(\hat{\beta}^T X_j)},$$

which is a step function that only jumps at observed failure times. It follows that a model-based estimate of $\mathcal{P}(u)$ is $\hat{\mathcal{P}}(u) = g(\hat{\gamma}^T x_+)$, where $\hat{\gamma} = [\log\{\hat{\Lambda}(t_0)\}, \hat{\beta}^T]^T$ and $x_+ = (1, x^T)^T$. Following from the convergence of $\hat{\beta}$ and arguments given in Hjort (1992), $\hat{\mathcal{P}}(u)$ converges to a deterministic function $\bar{\mathcal{P}}(u)$ in probability, as $n \rightarrow \infty$. When (1) is correctly specified, $\bar{\mathcal{P}}(u) = \mathcal{P}(u)$.

Although most likely (1) is not the true model, the parametric risk scores $\{\hat{\mathcal{P}}(u)\}$ or $\{\hat{\gamma}^T x_+\}$ can be used as a risk index system for future populations similar to the study population. Consider a future subject with $(\tilde{T}, U, X) = (\tilde{T}^0, U^0, X^0)$ and $\hat{\mathcal{P}}(U^0) = g(\hat{\gamma}^T X_+^0) = v$. When (1) is misspecified, the risk of $\tilde{T}^0 < t_0$ may be quite different from v . To calibrate the subject-level model-based risk estimate, one needs a consistent estimate for

$$\tau(v; t_0) = \text{pr}\{\tilde{T}^0 < t_0 \mid \hat{\mathcal{P}}(U^0) = v\},$$

the mortality rate among subjects with risk score $\hat{\mathcal{P}}(U^0) = v$. Here, the probability is with respect to \tilde{T}^0, U^0 and $\{(T_i, \delta_i, U_i), i = 1, \dots, n\}$.

To estimate $\tau(v; t_0)$, let $\Lambda_v(t) = -\log\{1 - \tau(v; t)\}$ be the corresponding cumulative hazard function and we propose a nonparametric kernel Nelson–Aalen estimator for $\Lambda_v(t)$ based on $\{T_i, \delta_i, \hat{\mathcal{P}}(U_i)\}$, for $i = 1, \dots, n$. As for the standard one-sample estimation of the cumulative hazard function, we focus on the class of potential estimators which are step functions over t and only jump at the observed failure times with jump sizes $\Delta\Lambda_v(t)$. First, we consider a local constant estimator by assuming that for v' in a small neighbourhood of v , $\Delta\Lambda_{v'}(t) \approx \Delta\Lambda_v(t)$. Specifically, for any given t and v , we obtain $\Delta\tilde{\Lambda}_v(t)$ as the following minimizer:

$$\arg \min_a \sum_{i=1}^n K_h(\hat{D}_{vi}) \{\Delta N_i(t) - Y_i(t)a\}^2 = \frac{\sum_{i=1}^n K_h(\hat{D}_{vi}) \Delta N_i(t)}{\sum_{i=1}^n K_h(\hat{D}_{vi}) Y_i(t)}, \tag{2}$$

where $\Delta N_i(t) = N_i(t) - N_i(t-) = \lim_{w \rightarrow 0} \int_{t-w}^t dN_i(s)$; $\hat{D}_{vi} = \psi\{\hat{\mathcal{P}}(U_i)\} - \psi(v)$; $K(\cdot)$ is a smooth symmetric density function; $K_h(x) = K(x/h)/h$, $h = O(n^{-\nu})$ is a bandwidth with $1/2 > \nu > 0$; and $\psi(\cdot) : (0, 1) \rightarrow (-\infty, \infty)$ is a known increasing smooth function (Wand et al., 1991; Park et al., 1997). Based on (2), we obtain a local constant estimator for $\Lambda_v(t_0)$ as

$$\tilde{\Lambda}_v(t_0) = \int_0^{t_0} \frac{\sum_{i=1}^n K_h(\hat{D}_{vi}) dN_i(t)}{\sum_{i=1}^n K_h(\hat{D}_{vi}) Y_i(t)}.$$

Here, we adopt the notation $\int \Delta N_i(t) W(t) dt \equiv \int W(t) dN_i(t)$ as in Nielsen & Tanggaard (2001). When $\hat{\mathcal{P}}(\cdot)$ is deterministic, the limiting distribution of such a nonparametric estimator with a fixed v has been derived by Nielsen & Linton (1995) and Du & Akritas (2002). Here, we investigate the asymptotic properties of $\tilde{\Lambda}_v(t_0)$ over a set of v at a given time t_0 in the presence of the additional variability in $\hat{\mathcal{P}}(\cdot)$.

The above estimator can potentially be improved by considering a local linear approximation to $\Lambda_v(t)$ (Fan & Gijbels, 1996; Li & Doss, 1995). That is, for v' in a small neighbourhood of v , we assume $\Delta \Lambda_{v'}(t) \approx \Delta \Lambda_v(t) + b_v(t)\{\psi(v') - \psi(v)\}$. Thus, for any given t and v , we replace a in (2) by a local linear function $a + b\hat{D}_{vi}$ with unknown intercept and slope parameters a and b . The resulting estimator $\Delta \hat{\Lambda}_v(t)$ is the intercept of the vector that minimizes

$$\sum_{i=1}^n K_h(\hat{D}_{vi}) \{\Delta N_i(t) - Y_i(t)(a + b\hat{D}_{vi})\}^2$$

with respect to (a, b) . The corresponding estimator $\hat{\Lambda}_v(t)$ for $\Lambda_v(t)$ is the sum of $\Delta \hat{\Lambda}_v(\cdot)$ over all distinct observed death times by t . The resulting estimator for $\tau(v; t_0)$ is $\hat{\tau}(v; t_0) = 1 - \exp\{-\hat{\Lambda}_v(t_0)\}$. In Appendix A, we show that under mild regularity conditions, if $h = O(n^{-\nu})$ with $1/5 < \nu < 1/2$, $\hat{\tau}(v; t_0)$ is consistent for $\tau(v; t_0)$, uniformly in $v \in \mathcal{J} = [\psi^{-1}(\rho_l + h), \psi^{-1}(\rho_r - h)]$, where (ρ_l, ρ_r) is a subset contained in the support of $\psi\{\bar{\mathcal{P}}(U)\}$.

3. POINTWISE AND SIMULTANEOUS INTERVAL ESTIMATION PROCEDURES FOR $\tau(v; t_0)$ OVER RISK SCORE V

For any fixed $v \in \mathcal{J}$, we show in Appendix A that the distribution of

$$(nh)^{1/2} \{\hat{\Lambda}_v(t_0) - \Lambda_v(t_0)\} \tag{3}$$

can be approximated well by the conditional distribution of a zero-mean normal distribution

$$n^{-1/2} h^{1/2} \sum_{i=1}^n K_h(\hat{D}_{vi}) \hat{V}_i(v; t_0) \xi_i + (nh)^{1/2} \{\hat{\Lambda}_v(t_0, \hat{\gamma}^*) - \hat{\Lambda}_v(t_0)\}, \tag{4}$$

given the data for large n and $h = O(n^{-\nu})$ with $1/5 < \nu < 1/2$, where $\Xi = (\xi_1, \dots, \xi_n)$ are standard normal random variables independent of the data, $\hat{V}_i(v; t_0) = \int_0^{t_0} \{dN_i(s) - Y_i(s) d\hat{\Lambda}_v(s)\} / \{n^{-1} \sum_{j=1}^n K_h(\hat{D}_{vj}) Y_j(s)\}$, $\hat{\Lambda}_v(t_0, \gamma)$ is given in (A3) and $\hat{\gamma}^*$ is given in (A5). A perturbed version of the observed $\hat{\Lambda}_v(t_0)$ is $\hat{\Lambda}_v(t_0, \hat{\gamma}^*)$ with the same perturbation variables Ξ , which accounts for the extra variability of $\hat{\gamma}$. Although asymptotically one only needs the first term of (4) since $\hat{\gamma} = O(n^{-1/2})$, the inclusion of the second term is likely to improve the approximation to the distribution of (3) in finite sample. This perturbation method is similar to the so-called wild bootstrap (Wu, 1986; Härdle, 1990; Mammen, 1992) and has been used successfully for a number of interesting problems in survival analysis (Jin et al., 2001; Park & Wei, 2003; Cai et al., 2005). The distribution of (4) can be easily approximated by generating a large number M of realizations of Ξ .

With the above approximation, for any $v \in \mathcal{J}$, one may obtain a variance estimator of (3), $\hat{\sigma}_v^2(t_0)$, based on the empirical variance of M realizations from (4). For any given $\alpha \in (0, 1)$, a $100(1 - \alpha)\%$ confidence interval for $\Lambda_v(t_0)$ can be obtained as $\hat{\Lambda}_v(t_0) \pm (nh)^{-1/2}c\hat{\sigma}_v(t_0)$, where c is the $100(1 - \alpha/2)$ th percentile of the standard normal. The corresponding pointwise $100(1 - \alpha)\%$ confidence interval for $\tau(v; t_0)$, the mortality rate with score v , is

$$1 - \exp \left\{ - \hat{\Lambda}_v(t_0) \pm (nh)^{-1/2}c\hat{\sigma}_v(t_0) \right\}.$$

To make inference about the mortality rate over a range of v , one may construct simultaneous confidence intervals for $\{\tau(v; t_0), v \in \mathcal{J}\}$ by considering a sup-type statistic

$$W = \sup_{v \in \mathcal{J}} \left| (nh)^{1/2} \frac{\hat{\Lambda}_v(t_0) - \Lambda_v(t_0)}{\hat{\sigma}_v(t_0)} \right|. \tag{5}$$

However, the distribution of $(nh)^{1/2}\{\hat{\Lambda}_v(t_0) - \Lambda_v(t_0)\}$ does not converge as a process in v , as $n \rightarrow \infty$. Therefore, we cannot use the standard large sample theory for stochastic processes to approximate the distribution of W . On the other hand, by the strong approximation arguments and extreme value limit theorem (Bickel & Rosenblatt, 1973), we show in Appendix B that a standardized version of W converges in distribution to a proper random variable. In practice, for large n , one can approximate the distribution of W by W^* , the supremum of the absolute value of (4) divided by $\hat{\sigma}_v(t_0)$, with (4) perturbed by the same set of Ξ for all $v \in \mathcal{J}$. It follows that the $100(1 - \alpha)\%$ simultaneous confidence interval for $\tau(v; t_0)$ is

$$1 - \exp \left\{ - \hat{\Lambda}_v(t_0) \pm (nh)^{-1/2}d \hat{\sigma}_v(t_0) \right\},$$

where the cut-off point d is chosen such that $\text{pr}(W^* < d) \geq 1 - \alpha$.

As for any nonparametric functional estimation problem, the choice of h for $\hat{\tau}(v; t_0)$ is crucial for making inferences about $\tau(v; t_0)$. To incorporate censoring, we propose obtaining an optimal h by minimizing mean integrated squared martingale residuals over time interval $(0, t_0)$ through K -fold crossvalidation. Such a procedure has been successfully used for bandwidth selection in Tian et al. (2005). Specifically, we randomly split the data into K disjoint subsets of about equal sizes denoted by $\{\mathcal{I}_k, k = 1, \dots, K\}$. For each k , we use all observations not in \mathcal{I}_k to obtain $\Lambda_v(t)$ with a given h . Let the resulting estimators be denoted by $\hat{\Lambda}_{v(k)}(t)$. We then use the observations from \mathcal{I}_k to calculate the sum of integrated squared martingale residuals

$$\int_0^{t_0} \sum_{j \in \mathcal{I}_k} \left\{ N_j(t) - \int_0^t Y_j(s) d\hat{\Lambda}_{\hat{P}(u_j)(k)}(s) \right\}^2 d \left\{ \sum_{j^i \in \mathcal{I}_k} N_{j^i}(t) \right\}. \tag{6}$$

Lastly, we sum (6) over $k = 1, \dots, K$, and then choose h_{opt} as the minimizer of the sum of K martingale residuals. Since the order of h_{opt} is expected to be $n^{-1/5}$ (Fan & Gijbels, 1995), the bandwidth we use for estimation is $h = h_{\text{opt}} \times n^{-d_0}$ with $0 < d_0 < 3/10$ such that $h = n^{-\nu}$ with $1/5 < \nu < 1/2$. This ensures that the resulting functional estimator $\hat{\Lambda}_v(t_0)$ with the data-dependent smooth parameter has the above desirable large sample properties.

4. NUMERICAL STUDIES

4.1. Example

We illustrate the new proposal with a dataset from the Valsartan in Acute Myocardial Infarction study (Pfeffer et al., 2003). This large clinical study, often referred to as the VALIANT study, was conducted to evaluate the effect of angiotensin II receptor blocker and angiotensin converting

Table 1. Estimates of the regression coefficients, the corresponding standard errors and p -values, multiplied by 100, derived from fitting the Cox model to the VALIANT dataset based on survival time information up to t_0 and based on the entire dataset

		X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
$t_0 = 6$	Est	2.8	28.9	-1.2	20.6	24.2	-43.3	25.2	24.1	16.1	22.0	16.4
	SE	0.3	3.2	0.2	6.6	7.2	10.7	7.2	6.1	8.8	10.8	6.2
	p	0.0	0.0	0.0	0.2	0.1	0.0	0.1	0.0	7.0	4.2	0.8
$t_0 = 24$	Est	3.0	23.7	-1.0	29.5	32.7	-46.3	29.6	26.8	24.3	23.1	23.5
	SE	0.2	2.3	0.1	4.7	5.1	7.7	5.2	4.4	6.2	8.0	4.5
	p	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0
Global	Est	2.7	27.3	-1.1	24.1	28.3	-50.6	26.3	26.4	19.6	22.7	18.2
	SE	0.3	3.0	0.2	5.6	6.8	9.3	6.2	5.2	7.6	9.5	5.8
	p	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.7	0.2

Est, estimates; SE, standard error; p, p -value; Global, entire dataset.

enzyme inhibitors on overall mortality among patients with myocardial infarction complicated by left ventricular systolic dysfunction and/or heart failure. The trial was designed to compare three treatments, angiotensin II receptor blocker valsartan, angiotensin converting enzyme inhibitor captopril and a combination of these two drugs, for treating high-risk patients after myocardial infarction with respect to mortality and morbidity. The study was conducted from 1999 to 2003 with a total of 14 703 patients assigned randomly and equally to the three groups. The median follow-up time was 24.7 months. There are no significant differences among the three groups with respect to the overall mortality and thus we combine all three groups to develop a risk scoring system for mortality.

For simplicity, we used 11 baseline covariates from each study subject considered by Solomon et al. (2005) in our analysis. These 11 predictors are potentially the most significant ones among the risk factors identified by Anavekar et al. (2004) for the overall mortality based on this study. We fitted the survival data with an additive Cox model using these 11 covariates to obtain a parametric risk score for each individual. Specifically, the vector $U = X$ consists of age, X_1 ; Killip class, X_2 ; estimated glomerular filtration rate, X_3 ; history of myocardial infarction, X_4 ; history of congestive heart failure, X_5 ; percutaneous coronary intervention after index myocardial infarction, X_6 ; atrial fibrillation after index myocardial infarction, X_7 ; history of diabetes, X_8 ; history of chronic obstructive pulmonary disease, X_9 ; new left bundle-branch block, X_{10} ; and history of angina, X_{11} . Our analysis includes $n = 14\,088$ patients who had complete information on these 11 covariates.

First, suppose that we are interested in predicting the six-month mortality rates of future patients. To this end, we let $t_0 = 6$ months and fitted the survival observations truncated slightly after six months with a Cox model (1) and $x = u$. The estimated regression coefficient vector $\hat{\beta}$ is given in Table 1. These estimates, coupled with the estimated intercept $\hat{\gamma} = -6.02$, create a risk score $\hat{P}(u)$ for six-month mortality of future patients. To obtain $\hat{\tau}(v, 6)$ for a given $\hat{P}(u) = v$, we let $K(\cdot)$ be the Epanechnikov kernel, and $\psi(v) = \log\{-\log(1 - v)\}$, which leads to $\psi\{\hat{P}(U)\} = \hat{\gamma}^T X_+$. The smoothing parameter $h = 0.18$ was obtained by multiplying h_{opt} by a factor of $n^{-0.07}$ as defined in § 3, with ten-fold crossvalidation. We chose the second and 98th percentiles of the empirical distribution of $\{\hat{\gamma}^T X_{+i}\}_{i=1, \dots, n}$ as the boundary points ρ_l and ρ_r for interval \mathcal{J} . To approximate the distribution of (3) and W in (5), we used the perturbation-resampling method (4) with $M = 500$ independent realized standard normal samples Ξ .

The smoothed density estimate for $\hat{P}(U)$, shown in Fig. 1(a), provides useful information regarding the relative size of the subgroup of individuals with $\hat{P}(U) = v$, where

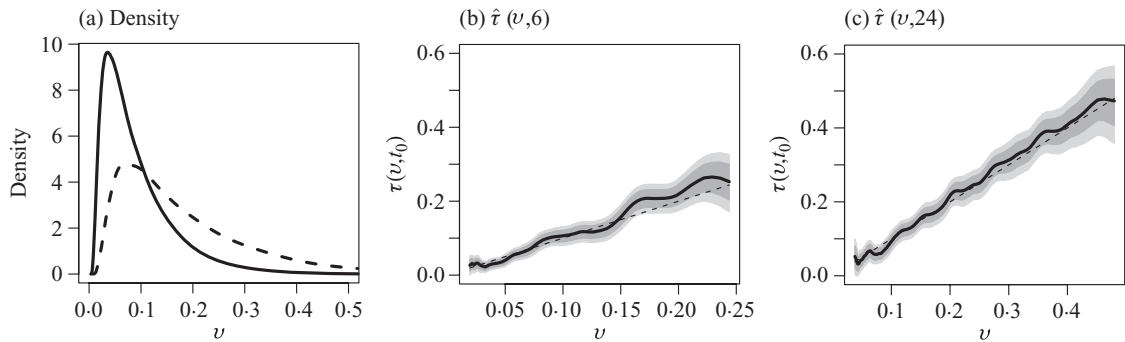


Fig. 1. Prediction of six-month and 24-month mortality risks based on the 11 risk factors. Panel (a) shows the estimated density function of the model-based risk estimate for $t_0 = 6$ (solid line) and 24 (dashed line) months. Shown also are the point (thick solid curve) estimate of the true risk function $\tau(v, t_0)$ along with their pointwise (dark shaded region) and simultaneous (light shaded region) confidence intervals for (b) $t_0 = 6$ months; and (c) $t_0 = 24$ months. The dashed line is the 45° reference line.

$v \in \mathcal{J} = [0.02, 0.24]$. In Fig. 1(b), we present the point and interval estimates for $\tau(v, 6)$. The estimated risk scores for most patients are less than 0.15. For groups of subjects with low risk scores, their interval estimates tend to be tight. For example, among subjects with $\hat{P}(u) = 0.045$, the six-month mortality rate is likely to be between 0.03 and 0.05, based on the 95% pointwise confidence interval. The corresponding simultaneous confidence interval is [0.02, 0.06]. For this subgroup, the Cox model appears to be slightly overestimating the risk with a p -value of 0.08 when testing the difference $\tau(v, 6) - v = 0$. For patients whose risk scores are greater than 0.15, the interval estimates are relatively wide, as expected. For example, the mortality rate among patients with $\hat{P}(u) = 0.17$ is estimated as 0.21 with a 95% pointwise confidence interval of [0.18, 0.23] and a simultaneous interval of [0.17, 0.25]. The Cox model underestimates the risk of this subgroup substantially with a pointwise p -value of 0.001 and a familywise adjusted p -value of 0.04 after controlling for the overall Type I error. For comparison, we also obtained the regression coefficient estimates based on the global Cox model without truncation. As shown in Table 1, the estimates are slightly different from the ones obtained based on survival information up to six months only. Under the global model, when the Cox model predicted risk is 0.045, the calibrated risk estimate is 0.034 with 95% pointwise and simultaneous confidence intervals being [0.03, 0.04] and [0.02, 0.05]. Overall, we find that the global model and the truncated version lead to very similar interval estimates for the present case.

Next, suppose that we are interested in predicting long-term survival for future patients similar to the VALIANT study population. Here, we let $t_0 = 24$ months and fitted the survival observations slightly truncated at 24 months with a Cox model and $x = u$. The resulting regression coefficient estimates are also given in Table 1. For this case, the bandwidth $h = 0.15$ and the range of the estimated risk score v is from 0.04 to 0.48. The smoothed density function estimation of the parametric score is also given in Fig. 1(a). Relatively few patients have scores beyond 0.3. Our point and interval estimates for the true mortality rate are given in Fig. 1(c). For example, for subjects with a risk score of 0.1, their 24-month mortality rate is estimated as 0.09 with 95% pointwise and simultaneous confidence intervals being [0.08, 0.11] and [0.07, 0.12], respectively. For patients whose risk scores are high, as expected, their interval estimates can be quite wide. For example, for subjects with a risk score of 0.3, the true mortality rate is likely to be between 0.26 and 0.37 based on the 95% simultaneous confidence interval. Under the global model, the calibrated risk estimates at risk scores of 0.1 and 0.3 are similar to the above results based on the truncated fitting.

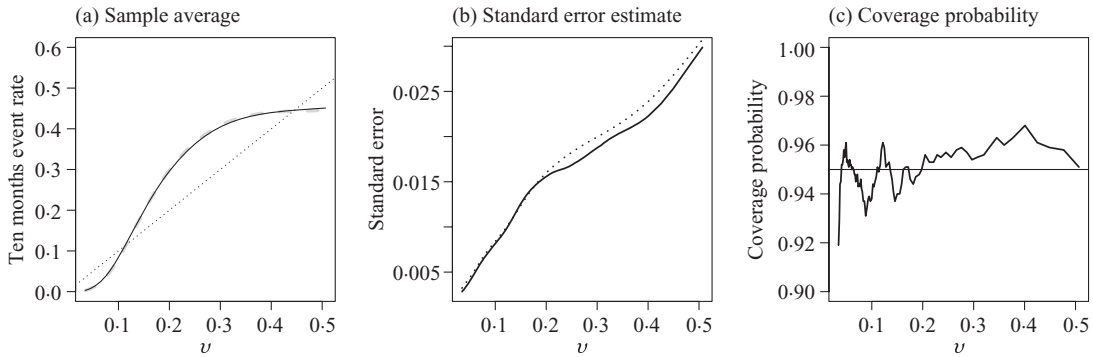


Fig. 2. Performance of the new procedure under a misspecified model with sample size 5000 for $t_0 = 10$: (a) the sample average of $\hat{\tau}(v, t_0)$ (solid) compared with truth (dashed); (b) the empirical standard error estimate (solid) and the average of the estimated standard errors (dotted); and (c) the empirical coverage of the 95% confidence intervals based on the proposed resampling procedures. The 45° reference line in (a) corresponds to the Cox-model-based risk estimates.

4.2. Simulation studies

We conducted simulation studies to examine the validity of the proposed inference procedures; and compared the performance of the calibrated prediction procedures with that of the prediction based on the Cox model. We first simulated data from a normal mixture model under which the Cox model is misspecified. In particular, we generated all discrete covariates, $X_D = (X_2, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11})^T$, based on their empirical distribution from the observed data. Then, for each given value of X_D , we generated the corresponding value of $X_C = (X_1, X_3)$, from $N(\mu_{X_D}, \Sigma)$, where μ_{X_D} is the empirical mean of X_C given X_D and Σ is the empirical covariance matrix of X_C in the VALIANT dataset. Then, we generated T from

$$\log \tilde{T} = \alpha_{X_4 X_8} + \beta_{X_4 X_8}^T X_{-(48)} + \sigma_{X_4 X_8} \varepsilon, \quad \varepsilon \sim N(0, 1),$$

where for $(X_4, X_8) = (0, 0), (0, 1), (1, 0)$ and $(1, 1)$, we let $\alpha_{X_4 X_8} = 11.5, 8.5, 5.5, 8.5$ and $\sigma_{X_4 X_8} = 0.9, 1.5, 1.3, 1.0$; $X_{-(48)}$ denotes the covariate vector with (X_4, X_8) excluded and $\beta_{X_4 X_8}$ is obtained from fitting a lognormal model to the subset of VALIANT data with a given value of (X_4, X_8) . The censoring was generated from a Weibull (5, 30) that was obtained by fitting the Weibull model to the VALIANT data. This results in about 77% censoring and overall event rates of 17% by month 10 and 22% by month 24. We considered a moderate sample size of 5000 and a relatively larger sample size of 10 000. For all the simulation studies, we obtain $\hat{\beta}$ based on the truncated partial likelihood function at t_0 . For ease of computation, the bandwidth for constructing the nonparametric estimate was fixed at $h = 0.37, 0.38, 0.29$ and 0.27 for (i) $n = 5000, t_0 = 10$; (ii) $n = 5000, t_0 = 24$; (iii) $n = 10\,000, t_0 = 10$; and (iv) $n = 5000, t_0 = 24$, respectively. Here, h was chosen as the average of the bandwidths selected based on (6) with $d_0 = 0.1$ from 10 simulated datasets.

Across all four settings, the nonparametric estimator has negligible bias, the estimated standard errors are close to their empirical counterparts, and the empirical coverage levels of the 95% confidence intervals are close to the nominal level. The coverage levels of the simultaneous confidence band were 93.7%, 94.5%, 93.1% and 94.1%, for settings (i), (ii), (iii) and (iv), respectively. In Fig. 2, we summarize the performance of the point and interval estimates of the proposed nonparametric procedure for $n = 5000$ and $t_0 = 10$. Under the correct model specification, we would expect that the true risk against the predicted risk from the Cox model is a 45° diagonal line. For the present case, the predicted risk based on the Cox model is severely

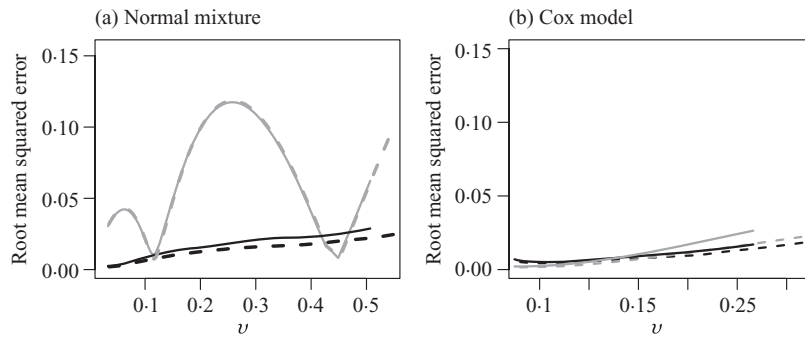


Fig. 3. The root mean squared error of the risk estimates for $t_0 = 10$ based on the Cox model and the nonparametric calibrated procedure when the data were generated from (a) a normal mixture; and (b) a Cox model, with $n = 5000$ (solid) and $n = 10\,000$ (dashed) on both panels.

biased. For example, the event rate at $t_0 = 10$ is 0.31 among the subjects whose predicted risk is 0.20 based on the Cox model. It is interesting that the standard test has very low power in detecting the lack of fit for the misspecified model. For example, the test based on the Schoenfeld residuals has about 15% power when $n = 5000$ and 25% power when $n = 10\,000$.

We also conducted simulation studies to evaluate the performance of the proposed procedure under the correct model specification. To this end, we generated the covariates and the censoring time from the same distribution as described above, but generated \tilde{T} from an exponential model with coefficients obtained by fitting the model to the VALIANT data. The proposed estimators also perform well with little bias for the point estimation and proper coverage level for intervals estimators. For example, when $n = 5000$ and $t_0 = 10$, the empirical coverage level ranges from 0.94 to 0.98 for the pointwise confidence intervals and is about 0.96 for the simultaneous confidence interval. Similar patterns were observed for $t_0 = 24$.

Lastly, we examined the performance of the nonparametric calibrated risk estimates with respect to their true predicted risks conditional on the observed data under correctly, and also incorrectly, specified models. Specifically, for each simulated dataset, we obtained $\hat{\gamma}_{\text{obs}}$ based on the above Cox models to calculate $\hat{P}(U^0) = g(\hat{\gamma}_{\text{obs}}^T X_+^0)$ for any given covariate U^0 , and then assign future subjects to a risk group $\Omega_v = \{U^0 : \hat{P}(U^0) = v\}$. A standard Cox model would estimate the event rate for the subgroup Ω_v as v . On the other hand, the nonparametric procedure would produce a calibrated estimate $\hat{\tau}(v; t_0)$ for the event rate among Ω_v , while the true event rate conditional on the observed data is $\tau(v; t_0, \hat{\gamma}_{\text{obs}}) = \text{pr}\{T^0 \leq t_0 \mid g(\hat{\gamma}_{\text{obs}}^T X_+^0) = v, \hat{\gamma}_{\text{obs}}\}$. In the previous simulation studies, $\tau(v; t_0) = E\{\tau(v; t_0, \hat{\gamma}_{\text{obs}})\}$ with expectation taken over the observed data. For each dataset, we summarized the performance of the proposed and the parametric risk prediction procedures based on their squared distance from the true event rate. The root mean squared errors of the Cox and calibrated estimates for $t_0 = 10$, denoted rmSE_{Cox} and $\text{rmSE}_{\text{Calib}}$, respectively, are shown in Fig. 3. Under the above misspecified normal mixture model, the parametric procedure is severely biased. For instance, when $n = 5000$, rmSE_{Cox} can be as high as 0.12 while $\text{rmSE}_{\text{Calib}}$ is below 0.03. Under model misspecification, over 90% of the root mean squared error is due to bias for most of the risk estimates. On the other hand, under the correct specification of the Cox model, the risk estimators from the nonparametric procedure are only slightly inefficient compared to the Cox procedure. For example, at $t_0 = 10$ with a sample size of 5000, $(\text{rmSE}_{\text{Cox}}, \text{rmSE}_{\text{Calib}})$ is (0.003, 0.006) when $v = 0.05$, is (0.005, 0.008) when $v = 0.10$ and is (0.017, 0.021) when $v = 0.20$.

5. REMARKS

Due to the complexity of the disease process and heterogeneity among study subjects, it is unlikely, if not impossible, that a fitted parametric or semiparametric model is the true one. On the other hand, a reasonable approximation to the true model can be quite useful in practice. Generally, the standard goodness-of-fit tests only provide qualitative assessment of the model fitting and may have little power when n is not large. An alternative way to evaluate the overall adequacy of the fitted model is to use a reasonable, physically interpretable distance function between the observed and predicted responses averaged over the entire study population (Cook, 2007; Uno et al., 2007). In this paper, we make an extra effort to calibrate the estimation of the mean risk for subgroups indexed by a parametric score. Through extensive numerical studies, we find that our two-stage procedure significantly outperforms purely model-based counterparts.

While one may obtain $\hat{\beta}$ by fitting a global Cox model, the parametric risk score $\hat{P}(U^0) = g\{\hat{\Lambda}(t_0) + \hat{\beta}^T X^0\}$ established based on the survival time observations up to t_0 is preferable when the primary interest is in estimating the t_0 -year survival rate and the Cox model does not fit the data well. For such settings, the covariate effects on the t_0 -year survival rate may vary markedly across t_0 . Indeed, we often find that certain covariates are essential for predicting short-term survival, but not so for predicting long-term survival. On the other hand, for overall decision making, it may be interesting to construct a unified risk score system, not specific to a time-point t_0 . Such a global score may be obtained by fitting the full dataset with a Cox model and obtaining $\hat{\beta}$ as the maximum partial likelihood estimator. Subsequently, one can group future subjects with $\hat{\beta}^T X^0 = u$ and consistently estimate $\tau_u(t) = \text{pr}(\tilde{T}^0 \leq t \mid \hat{\beta}^T X^0 = u)$ for various t s. For a finite set of time-points, using similar techniques utilized in this article, it is straightforward to obtain the pointwise and simultaneous confidence intervals for such multiple survival probabilities. For example, for patients with a risk score of $\hat{\beta}^T X^0 = 1.40$ in the aforementioned VALIANT study, the simultaneous confidence region for $\{\tau_{1.40}(6), \tau_{1.40}(24)\}$ is $[0.03, 0.05] \times [0.05, 0.09]$, for $t = 6, 24$. For patients with a parametric risk score of $\hat{\beta}^T X^0 = 2.87$, the corresponding simultaneous confidence region is $[0.19, 0.25] \times [0.30, 0.38]$. A challenging problem is how to construct a simultaneous confidence band for the subject-specific survival function over an entire time interval of interest. We expect that the technique developed in Li & Doss (1995) may be extended to the present case with extra care for the additional variability in $\hat{P}(U)$.

For the present paper, we assume that there are so-called baseline covariates available for each study patient. Often the baseline is not clearly defined. That is, it may be difficult to define the time zero for the study patient's follow-up time. Let S denote the study entry time since time zero and let $U(S)$ denote the corresponding covariate vector. For the VALIANT trial, every study participant had a recent nonfatal myocardial infarction with certain complications when entering the trial. For this case, the time zero for the patient's survival can be reasonably defined as the study entry date, which is approximately the time of occurrence of myocardial infarction. Therefore, $S = 0$, \tilde{T} represents the survival time since the myocardial infarction and $U = U(0)$ is the covariate level measured near the time of myocardial infarction. The proposed procedure would be valid for predicting the risk of death within t_0 since experiencing myocardial infarction for a future population whose distribution of (\tilde{T}^0, U^0) is the same as that of (\tilde{T}, U) .

In general, S may not be well defined or observable. For example, in AIDS studies with HIV infected patients, the time zero could be the unknown time of infection and S is the time interval between infection and study entry. For such cases, the observable event time \tilde{T} discussed in this paper, can be viewed as the residual life after S . Even when S is unknown, our new proposal is still valid for future subject-specific risk prediction provided that the joint distribution of $\{T^\dagger, S, U(S)\}$ for the study population is the same as that of $\{T^\dagger, S^0, U^0(S^0)\}$ for the future

population, where T^\dagger is the event time since time zero, and $\tilde{T} = T^\dagger - S$ is the residual life from the entry time. Specifically, one can make inferences about the risk function

$$\tau(v; t_0) = P[T^{\dagger 0} - S^0 \leq t_0 \mid T^{\dagger 0} > S^0, \hat{\mathcal{P}}\{U^0(S^0)\} = v] = P[\tilde{T}^0 \leq t_0 \mid \hat{\mathcal{P}}\{U^0(S^0)\} = v].$$

Here, the probability is taken over $T^{\dagger 0}$, $U^0(S^0)$, S^0 and the observed data. The quantity $\tau(v; t_0)$ is the event rate by t_0 among the subgroup of patients who were event-free at study entry with $\hat{\mathcal{P}}\{U^0(S^0)\} = v$.

When $U(s)$ is measured repeatedly over time, the prediction of residual life with longitudinal markers is an interesting yet challenging problem. It is generally difficult even to identify a coherent prediction model for such settings. While the article focuses on the relatively simple objective of predicting survival time with baseline biomarker value $U(0)$, it is often of interest, though difficult, to predict the residual life repeatedly based on periodically measured time-dependent biomarker $U(s)$. In such a case, it is desirable to make predictions based on a fitted global model describing the dynamic association between the longitudinal markers and survival time. However, it is generally difficult even to specify a coherent prediction model for such a setting. For example, in general, imposing the commonly used proportional hazards or accelerated failure time model on the relationship between the most updated marker value and residual life over a time span of interest leads to an inconsistent model (Jewell & Nielsen, 1993, Theorem 1). See Jewell & Nielsen (1993) and Mammen & Nielsen (2007) for excellent discussions on the prediction of survival times with time-varying markers.

ACKNOWLEDGEMENT

The authors are grateful to the editor and the reviewers for their insightful comments on the article. The work was partially supported by grants from the National Institutes of Health, U.S.A.

APPENDIX

In the Appendix, we use standard notation for the empirical process: \mathbb{P}_n and \mathbb{P} represent the expectation with respect to the empirical distribution generated $\{(T_i, \delta_i, X_i), i = 1, \dots, n\}$ and the distribution generated by (T, δ, X) , respectively. Similarly, $\mathbb{G}_n = n^{1/2}(\mathbb{P}_n - \mathbb{P})$. We assume that the covariate X is bounded, $h = O(n^{-\nu})$ with $1/5 < \nu < 1/2$, $K(x)$ is a symmetric smooth kernel function with a bounded support $[-1, 1]$ and $\int \dot{K}(x)^2 dx < \infty$, where $\dot{K}(x) = dK(x)/dx$. For convenience, we use the following notation: $K_j(x) = K(x)x^j$, $K_{j,h}(x) = K_j(x/h)/h$ and $m_2 = \int_{-1}^1 K^2(x)dx$. Unless noted otherwise, suprema are taken over $[\psi^{-1}(\rho_l + h), \psi^{-1}(\rho_r - h)]$ for v and over $[0, t_0]$ for t .

APPENDIX A

Asymptotic properties of $\hat{\tau}(v; t_0)$

We first derive the asymptotic properties of $\hat{\tau}(v; t_0)$, for $v \in [\psi^{-1}(\rho_l + h), \psi^{-1}(\rho_r - h)]$. Without loss of generality, we choose $\psi(x) = \log\{-\log(1 - x)\}$. Thus, $\psi\{\bar{\mathcal{P}}(U)\} = \gamma_0^\dagger X_+$, $\psi\{\hat{\mathcal{P}}(U_i)\} = \hat{\gamma}^\dagger X_{+i}$, and $\psi\{\hat{\mathcal{P}}(U_i)\} - \psi\{\bar{\mathcal{P}}(U_i)\} = O_p(n^{-1/2})$. Let $\tilde{v} = \psi(v)$, $\lambda_v(t)$ be the hazard function of \tilde{T} conditional on $\gamma_0^\dagger X_+ = \tilde{v}$, $\Lambda_v(t) = \int_0^t \lambda_v(s)ds$; $\zeta(\cdot)$ be the density function of $\gamma_0^\dagger X_+$; and $M_i(t) = N_i(t) - \int_0^t Y_i(s)d\Lambda_{\gamma_0^\dagger X_{+i}}(s)$. In view of $\hat{\tau}(v; t_0) = 1 - \exp\{-\hat{\Lambda}_v(t)\}$ and the delta method, we next establish the asymptotic properties of $\hat{\Lambda}_v(t)$. Furthermore, we will repeatedly use the fact that

$$\sup_{e_0 \in [a, b]} |\mathbb{P}_n[K_h(E - e_0)F] - \mathbb{P}\{K_h(E - e_0)F\}| = O_p[\{(nh)^{-1} \log(n)\}^{1/2}], \tag{A1}$$

where $\{(E_i, F_i)\}_{i=1, \dots, n}$ are independent and identically distributed realizations of (E, F) , satisfying $\sup_{e_0} \mathbb{P}(|F|^s | E = e_0) < \infty$, for all $s > 0$.

To derive asymptotic properties of $\hat{\Lambda}_v(t_0)$, we first show that there is no additional variability due to $\hat{\gamma}$. To this end, we write $\hat{\Lambda}_v(t_0) = \hat{\Lambda}_v(t_0, \hat{\gamma})$ and aim to show that if $\hat{\gamma} = \gamma_0 + O_p(n^{-1/2})$,

$$\sup_v |\hat{\Lambda}_v(t_0, \hat{\gamma}) - \hat{\Lambda}_v(t_0, \gamma_0)| = o_p\{(nh)^{-1/2}\}, \tag{A2}$$

where

$$\hat{\Lambda}_v(t_0, \gamma) = \int_0^{t_0} \{ \hat{R}_v^{(2)}(t, \gamma) d\hat{N}_v^{(0)}(t, \gamma) - \hat{R}_v^{(1)}(t, \gamma) d\hat{N}_v^{(1)}(t, \gamma) \}, \tag{A3}$$

$\hat{R}_v^{(j)} = \hat{\pi}_v^{(j)} / [\hat{\pi}_v^{(2)} \hat{\pi}_v^{(0)} - \{\hat{\pi}_v^{(1)}\}^2]$; $\hat{N}_v^{(j)}(t, \gamma) = \sum_i K_{j,h} \{D_{vi}(\gamma)\} N_i(t) / n$; $D_{vi}(\gamma) = \psi\{g(\gamma^\top X_{+i})\} - \psi(v)$; and $\hat{\pi}_v^{(j)}(t, \gamma) = \sum_{i=1}^n K_{j,h} \{D_{vi}(\gamma)\} Y_i(t) / n$. We next show that

$$\sup_{t,v} |\hat{\pi}_v^{(j)}(t, \hat{\gamma}) - \hat{\pi}_v^{(j)}(t, \gamma_0)| = o_p\{(nh)^{-1/2}\}.$$

Since $\hat{\pi}_v^{(j)}(t, \gamma) = \int K_{j,h}(x - \tilde{v}) d\mathbb{P}_n\{I(\hat{\gamma}^\top X_+ \leq x) Y(t)\}$,

$$\begin{aligned} |\hat{\pi}_v^{(j)}(t, \hat{\gamma}) - \hat{\pi}_v^{(j)}(t, \gamma_0)| &\leq \left| n^{-1/2} \int K_{j,h}(x - \tilde{v}) d\mathbb{G}_n \{I(\hat{\gamma}^\top X_+ \leq x) Y(t) - I(\gamma_0^\top X_+ \leq x) Y(t)\} \right| \\ &\quad + \left| \int K_{j,h}(x - \tilde{v}) \mathbb{P} \{I(\hat{\gamma}^\top X_+ \leq x) Y(t) - I(\gamma_0^\top X_+ \leq x) Y(t)\} \right| \\ &\leq h^{-1} n^{-1/2} \|\mathbb{G}_n\|_{\mathcal{H}_\delta} + O_p(n^{-1/2}), \end{aligned}$$

where $\mathcal{H}_\delta = \{I(\gamma^\top X_+ \leq x) Y(t) - I(\gamma_0^\top X_+ \leq x) Y(t) : x, t, |\gamma - \gamma_0| \leq \delta\}$ is a class of functions indexed by x, t and δ . Furthermore, \mathcal{H}_δ is uniformly bounded by an envelope function of order $\delta^{1/2}$ with respect to the L_2 norm. By the maximum inequality (Van der Vaart & Wellner, 1996, Theorem 2.14.2) and $|\hat{\gamma} - \gamma_0| = O_p(n^{-1/2})$, we have $h^{-1} n^{-1/2} \|\mathbb{G}_n\|_{\mathcal{H}_\delta} \leq O_p\{h^{-1} n^{-1/2} n^{-1/4} \log(n)\}$. It follows that $|\hat{\pi}_v^{(j)}(t, \hat{\gamma}) - \hat{\pi}_v^{(j)}(t, \gamma_0)| = (nh)^{-1/2} (nh^2)^{-1/4} \log(n) + O_p(n^{-1/2}) = o_p\{(nh)^{-1/2}\}$. Consequently, by (A1),

$$\sup_{t,v} |\hat{R}_v^{(j)}(t, \gamma_0) - \hat{R}_v^{(j)}(t, \hat{\gamma})| = o_p\{(nh)^{-1/2}\} \quad (k = 0, 1),$$

$$\sup_{t,v} \{ |\hat{R}_v^{(2)}(t, \gamma_0) - r_v^{(2)}(t)| + |\hat{R}_v^{(1)}(t, \gamma_0)| \} = O_p\{(nh)^{-1/2} \log(n)^{1/2} + h^2\},$$

where $r_v^{(2)}(t) = 1/\{\zeta(\tilde{v})\mathcal{S}_v(t)\}$ and $\mathcal{S}_v(t) = \text{pr}(T \geq t | \gamma_0^\top X_+ = \tilde{v})$. Next, we note that

$$\left| \int_0^{t_0} \hat{R}_v^{(2)}(t, \hat{\gamma}) d\hat{N}_v^{(0)}(t, \hat{\gamma}) - \int_0^{t_0} \hat{R}_v^{(2)}(t, \gamma_0) d\hat{N}_v^{(0)}(t, \gamma_0) \right| \leq \hat{\epsilon}_1 + \hat{\epsilon}_2 + \hat{\epsilon}_3,$$

where

$$\begin{aligned} \hat{\epsilon}_1 &= \mathbb{P}_n \left[\int_0^{t_0} |\hat{R}_v^{(2)}(t, \hat{\gamma}) - \hat{R}_v^{(2)}(t, \gamma_0)| K_h(\hat{\gamma}^\top X_+ - \tilde{v}) dN(t) \right], \\ \hat{\epsilon}_2 &= \mathbb{P}_n \left[\{K_h(\hat{\gamma}^\top X_+ - \tilde{v}) - K_h(\gamma_0^\top X_+ - \tilde{v})\} \int_0^{t_0} |\hat{R}_v^{(2)}(t, \gamma_0) - r_v^{(2)}(t)| dN(t) \right], \\ \hat{\epsilon}_3 &= \left| \mathbb{P}_n \left[\{K_h(\hat{\gamma}^\top X_+ - \tilde{v}) - K_h(\gamma_0^\top X_+ - \tilde{v})\} \int_0^{t_0} r_v^{(2)}(t) dN(t) \right] \right|. \end{aligned}$$

We next bound these three terms. For the first term, $\hat{\epsilon}_1 \leq o_p\{(nh)^{-1/2}\} \mathbb{P}_n\{K_h(\hat{\gamma}^\top X_+ - \tilde{v}) N(t_0)\} = o_p\{(nh)^{-1/2}\}$. For the second term, $\hat{\epsilon}_2 \leq h^{-1} \sup_t |\hat{R}_v^{(2)}(t, \gamma_0) - r_v^{(2)}(t)| \sup_t [\mathbb{P}_n\{I(\hat{\gamma}^\top X_+ \leq x) - I(\gamma_0^\top X_+ \leq x)\} N(t_0)]$, which is bounded by $O_p\{h^{-1} (nh)^{-1/2} \log(n)^{1/2} + h\} O_p(n^{-1/2}) = o_p\{(nh)^{-1/2}\}$. For

the last term,

$$\begin{aligned} \hat{\epsilon}_3 &\leq n^{-1/2}h^{-1} \sup_x \left| \mathbb{G}_n \left[\{I(\hat{\gamma}^\top X_+ \leq x) - I(\gamma_0^\top X_+ \leq x)\} \int_0^{t_0} r_v^{(2)}(t) dN(t) \right] \right| + O_p(n^{-1/2}) \\ &= O_p\{n^{-1/2}h^{-1}n^{-\frac{1}{4}} \log(n)\} = o_p\{(nh)^{-1/2}\}. \end{aligned}$$

Therefore, $\sup_v |\int_0^{t_0} \hat{R}_v^{(2)}(t, \hat{\gamma}) d\hat{N}_v^{(0)}(t, \hat{\gamma}) - \int_0^{t_0} \hat{R}_v^{(2)}(t, \gamma_0) d\hat{N}_v^{(0)}(t, \gamma_0)| = o_p\{(nh)^{-1/2}\}$. Similarly, $\int_0^{t_0} \hat{R}_v^{(1)}(t, \hat{\gamma}) d\hat{N}_v^{(1)}(t, \hat{\gamma}) - \int_0^{t_0} \hat{R}_v^{(1)}(t, \gamma_0) d\hat{N}_v^{(1)}(t, \gamma_0) = o_p\{(nh)^{-1/2}\}$, and thus (A2) holds. On the other hand, following Li & Doss (1995), we can show $\sup_v |\hat{\Lambda}_v(t_0, \gamma_0) - \Lambda_v(t_0)| \rightarrow 0$, in probability. It follows that $\sup_v |\hat{\Lambda}_v(t_0) - \Lambda_v(t_0)| \leq \sup_v |\hat{\Lambda}_v(t_0, \hat{\gamma}) - \hat{\Lambda}_v(t_0, \gamma_0)| + \sup_v |\hat{\Lambda}_v(t_0, \gamma_0) - \Lambda_v(t_0)| \rightarrow 0$, in probability. This establishes the uniform consistency of $\hat{\Lambda}_v(t_0)$.

We next derive the asymptotic distribution of $\hat{\mathcal{W}}_v(t_0) = (nh)^{1/2}\{\hat{\Lambda}_v(t_0) - \Lambda_v(t_0)\}$. From (A2), we have $\hat{\mathcal{W}}_v(t_0) = \tilde{\mathcal{W}}_v(t_0) + o_p(1)$, where $\tilde{\mathcal{W}}_v(t_0) = (nh)^{1/2}\{\hat{\Lambda}_v(t_0, \gamma_0) - \Lambda_v(t_0)\}$. Noting that $n^{1/2}h^{5/2} = o_p(1)$ and the decomposition that

$$\begin{aligned} \tilde{\mathcal{W}}_v(t_0) &= (nh)^{1/2} \mathbb{P}_n \left[\int_0^{t_0} \{ \hat{R}_v^{(2)}(t, \gamma_0) K_h(\gamma_0^\top X_+ - \tilde{v}) - \hat{R}_v^{(1)}(t, \gamma_0) K_{1,h}(\gamma_0^\top X_+ - \tilde{v}) \} dM(t) \right] \\ &\quad + (nh)^{1/2} \left\{ \frac{h^2 + o_p(h^2)}{2} \int_0^{t_0} \frac{\hat{\pi}_v^{(2)}(t, \gamma_0)^2 - \hat{\pi}_v^{(1)}(t, \gamma_0) \hat{\pi}_v^{(3)}(t, \gamma_0)}{\hat{\pi}_v^{(2)}(t, \gamma_0) \hat{\pi}_v^{(0)}(t, \gamma_0) - \hat{\pi}_v^{(1)}(t, \gamma_0)^2} \frac{\partial^2 \lambda_v(t)}{\partial \tilde{v}^2} ds \right\}, \end{aligned}$$

we have

$$\tilde{\mathcal{W}}_v(t_0) = (nh)^{1/2} \mathbb{P}_n \left[\int_0^{t_0} \{ \hat{R}_v^{(2)}(t, \gamma_0) K_h(\gamma_0^\top X_+ - \tilde{v}) - \hat{R}_v^{(1)}(t, \gamma_0) K_{1,h}(\gamma_0^\top X_+ - \tilde{v}) \} dM(t) \right] + o_p(1).$$

By integration by part and maximum inequality for empirical process, we have

$$\begin{aligned} (nh)^{1/2} \mathbb{P}_n \left\{ K_{1,h}(\gamma_0^\top X_+ - \tilde{v}) \int_0^{t_0} \hat{R}_v^{(1)}(t, \gamma_0) Y(t) dM(t) \right\} &= o_p(1), \\ (nh)^{1/2} \mathbb{P}_n \left[K_h(\gamma_0^\top X_+ - \tilde{v}) \int_0^{t_0} \{ \hat{R}_v^{(2)}(t, \gamma_0) - r_v^{(2)}(t) \} dM(t) \right] &= o_p(1), \end{aligned}$$

uniformly in \tilde{v} . Therefore,

$$\hat{\mathcal{W}}_v(t_0) = (nh)^{1/2} \mathbb{P}_n \left\{ K_h(\gamma_0^\top X_+ - \tilde{v}) \int_0^{t_0} r_v^{(2)}(t) dM(t) \right\} + o_p(1). \tag{A4}$$

For any fixed v or \tilde{v} , by martingale central limit theorem, $(nh)^{1/2} \mathbb{P}_n \{ K_h(\gamma_0^\top X_+ - \tilde{v}) \int_0^{t_0} r_v^{(2)}(t) dM(t) \} \rightarrow N\{0, \sigma_v^2(t_0)\}$, in distribution, as $n \rightarrow \infty$, where $\sigma_v^2(t_0) = m_2 \int_0^{t_0} r_v^{(2)}(t) d\Lambda_v(t)$. It follows that for any fixed v , $\hat{\mathcal{W}}_v(t_0)$ converges in distribution to $N\{0, \sigma_v^2(t_0)\}$.

To demonstrate the validity of the resampling variance estimator, we define

$$\hat{\gamma}^* = [\log\{\hat{\Lambda}_0^*(t_0)\}, \hat{\beta}^{*\top}]^\top, \tag{A5}$$

where

$$\begin{aligned} \hat{\beta}^* &= \hat{\beta} + \hat{\mathbb{A}}^{-1} \sum_{i=1}^n \int \left[\xi_i \left\{ X_i - \frac{\hat{\pi}^{(1)}(t, \hat{\beta})}{\hat{\pi}^{(0)}(t, \hat{\beta})} \right\} - \frac{\sum_{j=1}^n \xi_j Y_j(t) \exp(\hat{\beta}^\top X_j) \{ \hat{\pi}^{(0)}(t, \hat{\beta}) X_j - \hat{\pi}^{(1)}(t, \hat{\beta}) \}}{n \hat{\pi}^{(0)}(t, \hat{\beta})^2} \right] dN_i(t), \\ \log\{\hat{\Lambda}_0^*(t_0)\} &= \log\{\hat{\Lambda}_0(t_0)\} + \frac{n^{-1}}{\hat{\Lambda}_0(t_0)} \sum_{i=1}^n \int_0^t \left[\frac{\xi_i}{\hat{\pi}^{(0)}(s, \hat{\beta})} - \frac{\sum_{j=1}^n \xi_j Y_j(s) \exp(\hat{\beta}^\top X_j) + \hat{\pi}^{(1)}(t, \hat{\beta})^\top (\hat{\beta}^* - \hat{\beta})}{n \hat{\pi}^{(0)}(s, \hat{\beta})^2} \right] dN_i(s), \end{aligned}$$

$\hat{\Pi}^{(j)}(t, \beta) = n^{-1} \sum_{j=1}^n Y_j(t) \exp(\beta^\top X_j) X_j^{\otimes k}$ and for any vector x , $x^{\otimes 0} = 1$, $x^{\otimes 1} = x$ and $x^{\otimes 2} = xx^\top$. It is straightforward to show that $n^{1/2}(\hat{\gamma}^* - \hat{\gamma})$ conditional on the data and $n^{1/2}(\hat{\gamma} - \gamma_0)$ converge to the same

limiting distribution. From $\hat{\gamma}^* = \gamma_0 + O_p(n^{-1/2})$ and (A2), (4) = $n^{-1/2}h^{1/2} \sum_{i=1}^n K_h(\hat{D}_{vi})\hat{V}_i(v; t_0)\xi_i + o_p(1)$. Thus the variance estimator for $(nh)^{1/2}\{\hat{\Lambda}_v(t_0) - \Lambda_v(t_0)\}$ is

$$\begin{aligned} \hat{\sigma}_v^2(t_0) &= \text{var} \left\{ n^{-1/2}h^{1/2} \sum_{i=1}^n K_h(\hat{D}_{vi})\hat{V}_i(v; t_0)\xi_i \mid (T_i, \delta_i, U_i) (i = 1, \dots, n) \right\} + o_p(1) \\ &= \mathbb{P} \left\{ hK_h^2(\gamma_0^\top X_+ - \tilde{v}) \int_0^{t_0} r_v^{(2)}(t)^2 Y(t) d\Lambda_v(t) \right\} + o_p(1) = \sigma_v^2(t_0) + o_p(1), \end{aligned}$$

which converges to $\sigma_v^2(t_0)$ in probability.

APPENDIX B

Justification for the validity of the confidence band for $\hat{\tau}(v; t_0)$

We first justify that with proper standardization, $W = \sup_v |(nh)^{1/2}\hat{\Lambda}_v(t_0) - \Lambda_v(t_0)/\hat{\sigma}_v(t_0)|$ converges weakly. It follows from (A4) and the consistency of $\hat{\sigma}_v(t_0)$ for $\sigma_v(t_0)$ that

$$W = \sup_v \left| (nh)^{1/2} \mathbb{P}_n \left\{ K_h(\gamma_0^\top X_+ - \tilde{v}) \int_0^{t_0} \frac{r_v^{(2)}(t)}{\sigma_v(t_0)} dM_i(t) \right\} \right| + o_p(1).$$

This, together with the continuity of $r_v^{(2)}(t)/\sigma_v(t_0)$, implies that

$$W = \sup_v \left| (nh)^{1/2} \mathbb{P}_n \{ K_h(\gamma_0^\top X_+ - \tilde{v}) \eta \} \right| + o_p(1),$$

where $\eta = \int_0^{t_0} r_{\psi^{-1}(\gamma_0^\top X_+)}^{(2)}(t) dM(t) / \sigma_{\psi^{-1}(\gamma_0^\top X_+)}(t_0)$. It follows from a similar argument in Bickel & Rosenblatt (1973) that $\text{pr}\{a_n(W - d_n) < x\} \rightarrow e^{-2e^{-x}}$, where

$$a_n = \left[2 \log \left\{ \frac{\psi(\rho_r) - \psi(\rho_l)}{h} \right\} \right]^{1/2} \quad \text{and} \quad d_n = a_n + a_n^{-1} \log \left\{ \frac{1}{4m_2\pi} \int \dot{K}(t)^2 dt \right\}.$$

To justify the validity of the resampling procedure for constructing the confidence band, we consider

$$W^* = \sup_v \left| \frac{(nh)^{1/2}}{\hat{\sigma}_v(t_0)} \sum_{i=1}^n \xi_i \int_0^{t_0} \frac{K_h(\hat{\gamma}^\top X_{+i} - \tilde{v}) \{dN_i(t) - Y_i(t) d\hat{\Lambda}_v(t)\}}{\hat{\pi}_v^{(0)}(t, \hat{\gamma})} + (nh)^{1/2} \{ \hat{\Lambda}_v(t_0, \hat{\gamma}^*) - \hat{\Lambda}_v(t_0, \hat{\gamma}) \} \right|.$$

Again, since $|\hat{\gamma}^* - \hat{\gamma}| = O_p(n^{-1/2})$, from (A2), we have

$$W^* = \sup_v \left| \frac{(nh)^{1/2}}{\hat{\sigma}_v(t_0)} \sum_{i=1}^n \xi_i \int_0^{t_0} \frac{K_h(\hat{\gamma}^\top X_{+i} - \tilde{v}) \{dN_i(t) - Y_i(t) d\hat{\Lambda}_v(t)\}}{\hat{\pi}_v^{(0)}(t, \hat{\gamma})} \right| + o_p(1).$$

It follows from the same argument as given in Tian et al. (2005) that

$$\sup_x \text{pr}\{a_n(W^* - d_n) < x \mid (T_i, \delta_i, U_i), i = 1, \dots, n\} - e^{-2e^{-x}} \rightarrow 0$$

in probability as $n \rightarrow \infty$. Therefore, the conditional distribution of $a_n(W^* - d_n)$ can be used to approximate the distribution of $a_n(W - d_n)$ for large n .

REFERENCES

ANAVEKAR, N., McMURRAY, J., VELAZQUEZ, E., SOLOMON, S., KOBER, L., ROULEAU, J., WHITE, H., NORDLANDER, R., MAGGIONI, A., DICKSTEIN, K., ZELENKOFKSKE, S., LEIMBERGER, J., CALIFF, R. & PFEFFER, M. (2004). Relation between renal dysfunction and cardiovascular outcomes after myocardial infarction. *New Engl. J. Med.* **351**, 1285–95.

- ANDERSON, K., WILSON, P., ODELL, P. & KANNEL, W. (1991). An updated coronary risk profile: a statement for health professionals. *Circulation* **83**, 356–62.
- BICKEL, P. J. & ROSENBLATT, M. (1973). On some global measures of the deviations of density function estimates (Corr: V3 p1370). *Ann. Statist.* **1**, 1071–95.
- BRINDLE, P., EMBERSON, J., LAMPE, F., WALKER, M., WHINCUP, P., FAHEY, T. & EBRAHIM, S. (2003). Predictive accuracy of the Framingham coronary risk score in British men: prospective cohort study. *Br. Med. J.* **327**, 1267.
- CAI, T., TIAN, L. & WEI, L. (2005). Semiparametric Box-Cox power transformation models for censored survival observations. *Biometrika* **92**, 619–32.
- COOK, N. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* **115**, 928–35.
- COX, D. (1972). Regression models and life-tables (with discussion). *J. R. Statist. Soc. B* **34**, 187–220.
- D'AGOSTINO SR, R., GRUNDY, S., SULLIVAN, L. & WILSON, P. (2001). Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *J. Am. Med. Assoc.* **286**, 180–7.
- DU, Y. & AKRITAS, M. (2002). IID representations of the conditional Kaplan–Meier process for arbitrary distributions. *Math. Meth. Statist.* **11**, 152–82.
- EAGLE, K., LIM, M., DABBOUS, O., PIEPER, K., GOLDBERG, R., VAN DE WERF, F., GOODMAN, S., GRANGER, C., STEG, P., GORE, J., BUDAJ A., AVEZUM A., FLATHER M. D., FOX K. A.; GRACE INVESTIGATORS. (2004). A validated prediction model for all forms of acute coronary syndrome: estimating the risk of 6-month postdischarge death in an international registry. *J. Am. Med. Assoc.* **291**, 2727–33.
- FAN, J. & GIJBELS, I. (1995). Data-driven bandwidth selection in local polynomial regression: variable bandwidth selection and spatial adaptation. *J. R. Statist. Soc. B* **57**, 371–94.
- FAN, J. & GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. London: Chapman & Hall.
- GERDS, T. & SCHUMACHER, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biomet. J.* **48**, 1029–40.
- HÄRDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge UK: Cambridge University Press.
- HJORT, N. (1992). On inference in parametric survival data models. *Int. Statist. Rev.* **60**, 355–87.
- JEWELL, N. & NIELSEN, J. (1993). A framework for consistent prediction rules based on markers. *Biometrika* **80**, 153–64.
- JIN, Z., YING, Z. & WEI, L. (2001). A simple resampling method by perturbing the minimand. *Biometrika* **88**, 381–90.
- KALBFLEISCH, J. D. & PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data*. New York: John Wiley & Sons.
- LI, G. & DOSS, H. (1995). An approach to nonparametric regression for life history data using local linear fitting. *Ann. Statist.* **23**, 787–823.
- MAMMEN, E. (1992). Bootstrap, wild bootstrap, and asymptotic normality. *Prob. Theory Rel. Fields* **93**, 439–55.
- MAMMEN, E. & NIELSEN, J. (2007). A general approach to the predictability issue in survival analysis with applications. *Biometrika* **94**, 873–92.
- NIELSEN, J. (1998). Marker dependent kernel hazard estimation from local linear estimation. *Scand. Actuar. J.* **2**, 113–24.
- NIELSEN, J. & LINTON, O. (1995). Kernel estimation in a nonparametric marker dependent hazard model. *Ann. Statist.* **23**, 1735–48.
- NIELSEN, J. & TANGGAARD, C. (2001). Boundary and bias correction in kernel hazard estimation. *Scand. J. Statist.* **28**, 675–98.
- PARK, B., KIM, W., RUPPERT, D., JONES, M., SIGNORINI, D. & KOHN, R. (1997). Simple transformation techniques for improved non-parametric regression. *Scand. J. Statist.* **24**, 145–63.
- PARK, Y. & WEI, L. (2003). Estimating subject-specific survival functions under the accelerated failure time model. *Biometrika* **90**, 717–23.
- PENCINA, M., D'AGOSTINO SR, R., D'AGOSTINO JR, R. & VASAN, R. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statist. Med.* **27**, 157–72.
- PEPE, M., JANES, H., LONGTON, G., LEISENRING, W. & NEWCOMB, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am. J. Epidemiol.* **159**, 882–90.
- PFEFFER, M., McMURRAY, J., VELAZQUEZ, E., ROULEAU, J., KØBER, L., MAGGIONI, A., SOLOMON, S., SWEDBERG, K., VAN DE WERF, F., WHITE, H., LEIMBERGER, J., HENIS, M., EDWARDS, S., ZELENKOFKSKE, S., SELLERS, M., CALIFF, R., FOR THE VALSARTAN IN ACUTE MYOCARDIAL INFARCTION TRIAL INVESTIGATORS. (2003). Valsartan, captopril, or both in myocardial infarction complicated by heart failure, left ventricular dysfunction, or both. *New Engl. J. Med.* **349**, 1893–906.
- SOLOMON, S., SKALI, H., ANAVEKAR, N., BOURGOUN, M., BARVIK, S., GHALI, J. K., WARNICA, J. W., KHRAKOVSKAYA, M., ARNOLD, J. M. O., SCHWARTZ, Y., VELAZQUEZ E. J., CALIFF R. M., McMURRAY J. V. & PFEFFER M. A. (2005). Changes in ventricular size and function in patients treated with valsartan, captopril, or both after myocardial infarction. *Circulation* **111**, 3411–19.

- TIAN, L., ZUCKER, D. & WEI, L. (2005). On the Cox model with time-varying regression coefficients. *J. Am. Statist. Assoc.* **100**, 172–83.
- UNO, H., CAI, T., TIAN, L. & WEI, L. J. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *J. Am. Statist. Assoc.* **102**, 527–37.
- VAN DER VAART, A. & WELLNER, J. (1996). *Weak Convergence and Empirical Processes*. New York: Springer.
- WAND, M., MARRON, J. & RUPPERT, D. (1991). Transformation in density estimation (with comments). *J. Am. Statist. Assoc.* **86**, 343–61.
- WANG, T., GONA, P., LARSON, M., TOFLER, G., LEVY, D., NEWTON-CHEH, C., JACQUES, P., RIFAI, N., SELHUB, J., ROBINS, S., BENJAMIN E. J., D'AGOSTINO R. B. & VASAN R. S. (2006). Multiple biomarkers for the prediction of first major cardiovascular events and death. *New Engl. J. Med.* **355**, 2631–9.
- WARE, J. (2006). The limitations of risk factors as prognostic tools. *New Engl. J. Med.* **355**, 2615–7.
- WU, C. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14**, 1261–95.
- ZHANG, X., ATTIA, J., D'ESTE, C., YU, X. & WU, X. (2005). A risk score predicted coronary heart disease and stroke in a Chinese cohort. *J. Clin. Epidemiol.* **58**, 951–8.

[Received October 2008. Revised November 2009]