Biostatistics (2014), **15**, 2, *pp.* 222–233
doi:10.1093/biostatistics/kxt050
Advance Access publication on November 29, 2013

Predicting the restricted mean event time with the subject's baseline covariates in survival analysis

LU TIAN

Department of Health Research and Policy, Stanford University, Stanford, CA 94305, USA

LIHUI ZHAO

Department of Preventive Medicine, Northwestern University, Chicago, IL 60611, USA

L. J. WEI*

Department of Biostatistics, Harvard University, Boston, MA 02115, USA

wei@hsph.harvard.edu

Summary

For designing, monitoring, and analyzing a longitudinal study with an event time as the outcome variable, the restricted mean event time (RMET) is an easily interpretable, clinically meaningful summary of the survival function in the presence of censoring. The RMET is the average of all potential event times measured up to a time point τ and can be estimated consistently by the area under the Kaplan–Meier curve over $[0, \tau]$. In this paper, we study a class of regression models, which directly relates the RMET to its "baseline" covariates for predicting the future subjects' RMETs. Since the standard Cox and the accelerated failure time models can also be used for estimating such RMETs, we utilize a cross-validation procedure to select the "best" among all the working models considered in the model building and evaluation process. Lastly, we draw inferences for the predicted RMETs to assess the performance of the final selected model using an independent data set or a "hold-out" sample from the original data set. All the proposals are illustrated with the data from the an HIV clinical trial conducted by the AIDS Clinical Trials Group and the primary biliary cirrhosis study conducted by the Mayo Clinic.

Keywords: Accelerated failure time model; Cox model; Cross-validation; Hold-out sample; Personalized medicine; Perturbation-resampling method.

1. Introduction

For a longitudinal study with time T to a specific event as the primary outcome variable, commonly used summary measures for the distribution of T are the mean, median, or t-year event rate. Due to potential censoring for T, the mean may not be estimable. If the censoring is heavy, the median cannot be empirically identified either. The t-year survival rate can be estimated at a specific time t, but this estimate may not be suitable for summarizing the global profile of T over the duration of the study. On the other hand,

*To whom correspondence should be addressed.

based on the design of the study and clinical considerations, one may pre-specify a time point τ and utilize the expected value μ of $Y = \min(T, \tau)$, the so-called restricted mean event time (RMET), as a summary parameter. This parameter is the mean of T for all potential study patients followed up to time τ, which has heuristic and clinically meaningful interpretation (Chen and Tsiatis, 2001; Andersen *and others*, 2004; Royston and Parmar, 2011; Zhao *and others*, 2012). Moreover, this model-free parameter can be estimated consistently via the standard Kaplan–Meier (KM) curve, that is, the area under the curve up to τ. Inference procedures for the RMET with censored event-time observations were extensively studied by Zhao and Tsiatis (1997, 1999) under a more general setting.

In order to compare two groups, say, A and B, with censored event-time observations, practitioners routinely use the hazard ratio to quantify the group difference. Note that for this between-group contrast measure, there is no "background" rate one can utilize to evaluate whether such a hazard ratio estimate represents a clinically meaningful difference, information which is necessary for the purposes of risk–benefit decision making. Furthermore, when the proportional hazards assumption is not valid, the standard maximum partial likelihood estimator of the hazard ratio approximates a parameter which is difficult, if not impossible, to interpret as the treatment contrast (Lin and Wei, 1989; Rudser *and others*, 2012). Moreover, this parameter depends, oddly, on the nuisance, study-specific censoring distributions (Lin and Wei, 1989). It follows that the hazard ratio estimators at the interim and final analyses from the same study or estimators from independent studies with an identical study populations would estimate different, uninterpretable parameters due to differential follow-up patterns. Therefore, it is highly desirable to consider an estimable, model-free, and censoring-independent parameter to quantify the treatment difference for coherent and consistent assessments between interim and final analyses within a study, as well as across independent studies.

Model-free parameters for the treatment difference can be constructed via two RMETs, say, μ_A and μ_B. As an example, to evaluate the added value of a potent protease inhibitor, indinavir, for HIV patients, a pivotal study ACTG 320 was conducted by the AIDS Clinical Trials Group (ACTG). This randomized, double-blind study (Hammer *and others*, 1997) compared a three-drug combination, indinavir, zidovudine, and lamivudine, with the standard two-drug combination, zidovudine and lamivudine. There were 1156 enrolled for the study. One of the endpoints was the time to AIDS or death with a follow-up time of about 1 year for each patient. Figure 1 presents the KM curves for these two treatment groups. The hazard ratio estimate is 0.50 and the corresponding 0.95 confidence interval is (0.33, 0.76) with a p-value of 0.001. With $\tau = 300$ days, the estimated RMET was 277 days for the control and was 288 days for the three-drug combination. The estimated difference with respect to the RMET is 11 days with the corresponding 0.95 confidence interval of (3.2, 17.3) and a p-value of 0.005. Although the treatment efficacy for the three-drug combination is highly statistically significant, its clinical benefit is debatable using this metric of the absolute difference with respect to the RMET. If we mimic the concept of the hazard ratio or relative risk as a summary measure for the treatment contrast, one may consider a model-free ratio R of $(\tau - \mu_B)$ and $(\tau - \mu_A)$, where μ_A and μ_B are the RMETs for arms A and B, respectively. With the above HIV data, if B is the treatment group receiving the three-drug combination, the estimated R is 0.55 with a p-value of 9.3×10^{-6}, also an impressive statistically significant result. For a single arm, $(\tau - \mu)$ is the average of the years lost from the healthy state up to τ, a meaningful alternative to μ as a summary parameter for the distribution of T. Note that either using the absolute difference or the ratio to quantify the group-contrast via the RMET, there is a background value from the control arm for assessing the added value of indinavir for treating HIV patients from clinical benefit–risk–cost perspectives.

In this paper, we are interested in building prediction models for the RMET with the subject's "baseline" covariates and make inferences of such predictions. Existing regression models, such as the Cox model, can be candidates to create such a "personalized" prediction scheme. However, it seems more natural to model the RMET with the covariates directly, not via the hazard function (Andersen *and others*, 2004). In this article, we consider a class of models which takes this approach and study the properties of the corresponding

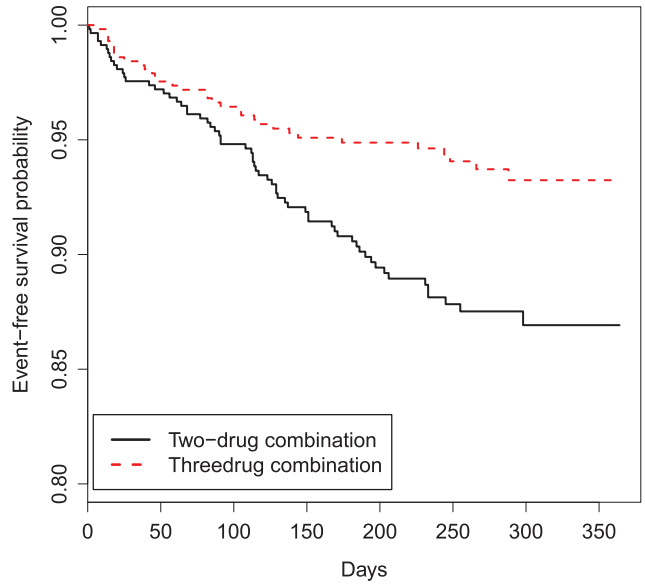

Fig. 1. KM estimates of the survival functions of the two randomized groups based on the ACTG 320 data.

inference procedures. Since it is unlikely that any model will be precisely correct, our ultimate goal is to choose the best "fitted" model among a set of candidate working models to stratify the future patients. To avoid overly optimistic results, we randomly split the data set into two pieces. Using the first piece, called the training-evaluation set, we utilize a cross-validation (CV) procedure to build and select the final model. We then use the second data set, called the hold-out set, to make inferences about the RMETs over a range of scores created from the final selected model. This final step is crucial to have a valid assessment of the performance of the proposed prediction procedure. We use a data set from a well-known clinical study conducted at Mayo Clinic (Therneau and Grambsch, 2000) for treating a liver disease to illustrate the proposals.

2. Regression models for RMET

For a typical subject with event time T, let Z be the corresponding q-dimensional baseline covariate vector. Suppose that T is subject to right censoring by a random variable C, which is assumed to be independent of T and Z. The observable quantities are (U, Δ, Z), where $U = \min(T, C)$, $\Delta = I(T \leqslant C)$, and $I(\cdot)$ is the indicator function. The data, $\{(U_i, \Delta_i, Z_i);\ i = 1, \ldots, n\}$, consist of n independent copies of (U, Δ, Z). Suppose that for a time point τ, $\mathrm{pr}(U \geqslant \tau) > 0$. The restricted survival time $Y = \min(T, \tau)$ may also be censored, but its expected value μ is estimable. Let Y_i be the corresponding Y for the ith subject, $i = 1, \ldots, n$. A natural estimator for μ is $\hat{\mu} = \int_0^\tau \hat{S}(u)\,\mathrm{d}u$, where $\hat{S}(u)$ is the KM estimator for the survival function of T based on $\{(U_i, \Delta_i),\ i = 1, \ldots, n\}$. The inference procedures for μ have been extensively studies, for example, by Zhao and Tsiatis (1997, 1999) and Zhao *and others* (2012).

Now, let $\mu(z) = E(Y|Z = z)$. It follows from Andersen *and others* (2004), one may model this directly with Z:

$$\eta\{\mu(z)\} = \beta' X, \tag{2.1}$$

where $\eta(\cdot)$ is a given smooth and strictly increasing link function, β is a $(q + 1)$-dimension unknown vector and $X' = (1, Z')$. The link function can be the identity function. On the other hand, since the support of

the restricted event time Y is finite, it may be appropriate to consider $\eta(\cdot)$ being an increasing function mapping $[0, \tau]$ to the real line. A special link function is $\eta(a) = \log\{a/(\tau - a)\}$, which mimics the logistic regression. Note that with this specific link, for the two-sample problem, the regression coefficient of the treatment indicator is

$$\log\left\{\frac{\mu_B(\tau - \mu_A)}{\mu_A(\tau - \mu_B)}\right\},$$

an odds-ratio like summary for the group contrast.

For the general link function $\eta(\cdot)$, following the least squares principle, an inverse probability censoring weighted estimating function of β is

$$S_n(\beta) = n^{-1}\sum_{i=1}^{n}\frac{\tilde{\Delta}_i}{\hat{G}(Y_i)}X_i\{Y_i - \eta^{-1}(\beta' X_i)\},$$

where $\tilde{\Delta}_i = I(Y_i \leqslant C_i)$ and $\hat{G}(\cdot)$ is the KM estimator of the censoring time C based on $\{(U_i, 1 - \Delta_i),\ i = 1, \ldots, n\}$. Let $\hat{\beta}$ be the unique root of $S_n(\beta) = 0$. In Appendix A of supplementary material available at *Biostatistics* online, we show that under mild regularity conditions, $\hat{\beta}$ converges to a constant $\bar{\beta}$ in probability, even when Model (1) is misspecified, an important property for building a prediction model. In addition, we show that as $n \to \infty$, $n^{1/2}(\hat{\beta} - \bar{\beta})$ converges weakly to a mean zero Gaussian distribution. We also provide inference procedures for $\bar{\beta}$ in Appendix A of supplementary material available at *Biostatistics* online. Note that Andersen *and others* (2004) studied Model (1) via a specific log-link $\eta(\cdot)$ using a pseudo-observation technique to make inferences about the regression coefficient. However, there is no systematic procedure in the literature for evaluating the adequacy of such a working model from the prediction point of view.

Using the above model, one may estimate $\mu(z)$ by $\hat{\mu}(z) = \eta^{-1}(\hat{\beta}' x)$, for any fixed $Z = z$, where $x' = (1, z')$. The distribution of $\{\hat{\mu}(z) - \eta^{-1}(\bar{\beta}' x)\}$ can be approximated by the delta method. Note that $\mu(z)$ can also be estimated via, for example, a Cox model (Cox, 1972). Specifically, let the hazard function for given z be $\lambda(t|Z = z) = \lambda_0(t)\,\mathrm{e}^{\gamma' z}$, where γ is a q-dimensional unknown vector and $\lambda_0(\cdot)$ is the nuisance baseline hazard function. It follows that $\mu(z)$ can be estimated by

$$\hat{\mu}(z) = \int_0^{\tau}\exp\{-\hat{\Lambda}_0(s)\,\mathrm{e}^{\hat{\gamma}' z}\}\,\mathrm{d}s,$$

where $\hat{\gamma}$ and $\hat{\Lambda}_0(s)$ are the maximum partial likelihood estimator for γ and the Breslow estimator for $\Lambda_0(s) = \int_0^s \lambda_0(v)\,\mathrm{d}v$, respectively.

Alternatively, one may use the accelerated failure time (AFT) model (Kalbfleisch and Prentice, 2002), $\log(T) = \gamma' Z + \epsilon$, to make inference about $\mu(z)$, where γ is a q-dimensional unknown vector and ϵ is the error term whose distribution is entirely unspecified. Here, γ can be estimated via a rank-based estimating function (Tsiatis, 1990; Jin *and others*, 2003). Let $\hat{\gamma}$ be the corresponding estimator for γ. One may estimate the survival function of e^{ϵ} by KM estimator based on the data $\{(U_i\,\mathrm{e}^{-\hat{\gamma}' Z_i}, \Delta_i),\ i = 1, \ldots, n\}$. Let the resulting estimator be denoted by $\hat{S}_0(\cdot)$. Then one can estimate $\mu(z)$ by $\hat{\mu}(z) = \int_0^{\tau}\hat{S}_0(\mathrm{e}^{-\hat{\gamma}' z}s)\,\mathrm{d}s$. Note that when $\mathrm{pr}(C > \mathrm{e}^{\gamma' Z}\sup \mathrm{e}^{\epsilon}) > 0$, $\hat{\mu}(z)$ is estimable for any given covariate z. In practice, we can always set the censoring indicator at one for the observation with the largest $U_i\,\mathrm{e}^{-\hat{\gamma}' Z_i}$ in estimating the survival function of e^{ϵ}. Although these estimators for $\mu(z)$ may be biased, and under general model misspecification, will depend on the censoring distribution, they may still produce reasonable predictions for the RMET. Note that the parameters in either Cox or AFT working model may be estimated with the entire observed data, not restricted by the follow-up information up to time point τ.

3. MODEL SELECTION AND EVALUATION

All the models for estimating $\mu(z)$ discussed in the previous section are approximations to the true model. To compare these models, one may compare the restricted event time Y with the covariate vector z and its predicted $\hat{\mu}(z)$. A reasonable predicted error measure is $E|Y - \hat{\mu}(Z)|$, where the expected value is with respect to the data and the future subject's (Y, Z). If there is no censoring, the empirical apparent prediction error is $n^{-1}\sum_{i=1}^{n}|Y_i - \hat{\mu}(Z_i)|$, which is obtained by first using the entire data to compute $\hat{\mu}(\cdot)$ and then using the same data to estimate the predicted error. To avoid bias, we utilize a CV procedure to estimate such a predicted error (Tian *and others*, 2007). Specifically, consider a class of models for $\mu(Z)$. For each model, we randomly split the data set into K disjoint subsets of approximately equal sizes, denoted by $\{\mathcal{I}_k,\ k = 1, \ldots, K\}$. For each k, we use all observations which are not in $\mathcal{I}_k$ to obtain a model-based prediction rule $\hat{\mu}_{(-k)}(Z)$ for Y, and then estimate the total absolute prediction error for observations in $\mathcal{I}_k$ by

$$\hat{D}_k = \sum_{j \in \mathcal{I}_k} \frac{\tilde{\Delta}_j}{\hat{G}(Y_j)} |Y_j - \hat{\mu}_{(-k)}(Z_j)|.$$

Then we use the average $n^{-1}\sum_{k=1}^{K}\hat{D}_k$ as a K-fold CV estimate for the absolute prediction error. We may repeat the aforementioned procedure a large number of, say J, times with different random partitions. Then the average of the resulting J cross-validated estimates is the final random K-fold CV estimate for the absolute prediction error of the fitted regression model. Generally, the model which yields the smallest cross-validated absolute prediction error estimate among all candidate models is chosen as the final model. On the other hand, a parsimonious model may be preferable if its empirical predicted error is comparable with a more complex "optimal" model. We then refit the entire training-evaluation data set with this selected model for making predictions based on $\hat{\mu}(\cdot)$.

Note that in the training stage of this CV process, a candidate model may be obtained via a complex variable selection process. For example, a Cox model may be built with a stepwise regression or lasso procedure. In this case, the final choice for creating the score would be refitting the entire training-evaluation data set with the selected model building algorithm.

4. INFERENCE ABOUT SUBJECT-SPECIFIC RMET

Ideally, one would use a model-free estimate of $\mu(z) = E(Y|Z = z)$ for subject-specific prediction of RMET. However, a fully non-parametric estimate of $\mu(z)$ is not feasible due to the curse of dimensionality. A practical way to create a prediction scheme using the baseline covariates is to utilize the "best" candidate among all the working models considered in the previous section to create a scoring system for the future subject's RMET. Then we use this univariate score to stratify subjects and make inferences about the stratum-specific RMET with a data set from an independent study or the hold-out sample from the same study.

To this end, let the estimated $\hat{\mu}(\cdot)$ from the final selected model be denoted by $\hat{\mu}_f(\cdot)$ and for a future subject with (Y, Z), let its prediction score be denoted by $V = \hat{\mu}_f(Z)$. That is, for each future subject, the covariate vector Z is reduced to a univariate V which is a function of Z. If the selected model is close to the true one, we expect that $E(Y|V) \approx \mu(Z) \approx V$. In general, however, the group mean $\xi(v) = E(Y|V = v)$ by clustering all subjects with Z, whose $\hat{\mu}_f(Z) = v$, may be different from v. Therefore, the conventional parametric inferences about predicting $\xi(v)$ via the selected model may not be valid. On the other hand, since we reduce the covariate information to a univariate score V, one may utilize a non-parametric estimation procedure to draw valid inferences about $\xi(\cdot)$.

To make non-parametric inference about $\xi(v)$ simultaneously across a range of the score v, we use a fresh independent data set or "hold-out" set from the original data set. With slight abuse of notation, let

such a fresh data set be denoted by $\{(U_i, \Delta_i, V_i),\ i = 1, \ldots, n\}$. We propose to use local linear smoothing method to estimate $\xi(v)$ non-parametrically. To this end, for a score v inside the support of V, let $\hat{a}$ and $\hat{b}$ be the solution of the estimating equation

$$S_n(a, b; v) = \sum_{i=1}^{n} \frac{K_h(V_i - v)\tilde{\Delta}_i}{\hat{G}(Y_i|v)} \begin{pmatrix} 1 \\ V_i - v \end{pmatrix} [Y_i - \tilde{\eta}^{-1}\{a + b(V_i - v)\}] = 0,$$

where $K(\cdot)$ is a smooth symmetric kernel function with a finite support, $K_h(s) = K(s/h)/h$, $h = o_p(1)$ is the smoothing bandwidth,

$$\hat{G}(t|v) = \exp\left\{ -\sum_{i=1}^{n} \int_0^t \frac{\mathrm{d}N_i^C(u) K_h(V_i - v)}{\sum_{j=1}^{n} (U_j \geqslant u) K_h(V_j - v)} \right\}$$

is the local non-parametric estimator for the survival function of C (Dabrowska, 1987, 1989) and $N_i^C(u) = I(Y_i \leqslant u)(1 - \tilde{\Delta}_i)$. Here, $\tilde{\eta}(\cdot)$ is a strictly increasing function from $[0, \tau]$ to the entire real line given *a priori*. The resulting local linear estimator for $\xi(v)$ is $\hat{\xi}(v) = \tilde{\eta}^{-1}(\hat{a})$. As $n \to \infty$ and $nh^5 = o_p(1)$, $(nh)^{1/2}\{\hat{\xi}(v) - \xi(v)\}$ converges weakly to a mean zero Gaussian. The details are given in Appendix B of supplementary material available at *Biostatistics* online. Since the censoring time C is assumed to be independent of V, generally the non-parametric KM estimator based on the entire sample is used in the inverse probability weighting method for $S_n(a, b; v)$. Here, we use the local estimator $\hat{G}(t|v)$ in above estimating equation. In Appendix B of supplementary material available at *Biostatistics* online, we show that this estimation procedure results in a more accurate estimator for $\xi(v)$ than that using $\hat{G}(\cdot)$. Note that when the empirical distribution of $\{V_i,\ i = 1, \ldots, n\}$ is quite non-uniform, transforming the score via an appropriate function before smoothing could potentially improve the performance of the kernel estimation (Park *and others*, 1997; Cai *and others*, 2010).

The perturbation-resampling method proposed by Gilbert *and others* (2002) and Tian *and others* (2005) can be used to construct pointwise and simultaneous confidence intervals for $\xi(v)$ over v. To this end, let $\hat{a}^*$ and $\hat{b}^*$ be the solution of the perturbed estimating equation

$$S_n^*(a, b; v) = \sum_{i=1}^{n} Q_i \frac{K_h(V_i - v)\tilde{\Delta}_i}{G^*(Y_i|v)} \begin{pmatrix} 1 \\ V_i - v \end{pmatrix} [Y_i - \tilde{\eta}^{-1}\{a + b(V_i - v)\}] = 0,$$

where $\{Q_1, \ldots, Q_n\}$ are positive random variables with unit mean and variance and independent of the observed data, and

$$G^*(t|v) = \exp\left\{ -\sum_{i=1}^{n} \int_0^t \frac{Q_i\, \mathrm{d}N_i^C(u) K_h(V_i - v)}{\sum_{j=1}^{n} Q_j (Y_j \geqslant u) K_h(V_j - v)} \right\}.$$

Then a perturbed estimator for $\xi(v)$ is $\hat{\xi}^*(v) = \tilde{\eta}^{-1}(\hat{a}^*)$. Conditional on the observed data, the limiting distribution of $(nh)^{1/2}\{\hat{\xi}^*(v) - \hat{\xi}(v)\}$ approximates the unconditional counterpart of $(nh)^{1/2}\{\hat{\xi}(v) - \xi(v)\}$. It follows that one can estimate the variance of $\hat{\xi}(v)$ by $\hat{\sigma}^2(v)$, the empirical variance of M realized $\hat{\xi}^*(v)'$s using M independent sets of $\{Q_1, \ldots, Q_n\}$. Based on generated $\hat{\xi}^*(v)$, one may construct $(1 - 2\alpha)$ confidence interval of $\xi(v)$ as $[\hat{\xi}(v) - c_\alpha \hat{\sigma}(v), \hat{\xi}(v) + c_\alpha \hat{\sigma}(v)]$, where c_α is the upper 100α percentage point of the standard normal. For an interval $[v_1, v_2]$, a subset of the support of V, the $(1 - 2\alpha)$ simultaneous confidence band of $\xi(v)$, $v \in [v_1, v_2]$ can be constructed similarly as $[\hat{\xi}(v) - d_\alpha \hat{\sigma}(v), \hat{\xi}(v) + d_\alpha \hat{\sigma}(v)]$,

where

$$\text{pr}\left\{\sup_{v\in[v_1,v_2]}\frac{|\hat{\xi}^*(v)-\hat{\xi}(v)|}{\hat{\sigma}(v)}<d_\alpha\,\middle|\,(U_i,\Delta_i,V_i),\ i=1,\ldots,n\right\}=1-2\alpha.$$

As with any non-parametric function estimation problem, it is crucial to choose an appropriate bandwidth h in order to make proper inference about $\xi(v)$. In Appendix C of supplementary material available at *Biostatistics* online, we propose a CV procedure to choose an optimal h value which minimizes a weighted cross-validated absolute prediction error.

5. Example for subject-specific prediction

In this section, we use a well-known data set from a liver study to illustrate how to build and select a model, and make inferences simultaneously about the RMETs over a range of scores created by the final model. This liver disease study in primary biliary cirrhosis (PBC) was conducted between 1974 and 1984 to evaluate the drug D-penicillamine, which was found to be futile with respect to the patient's mortality. The investigators for the study then used this rich data set to build a prediction model with respect to mortality (Fleming and Harrington, 1991). There were a total of 418 patients involved in the study, including 112 patients who did not participate in the clinical trials, but had available baseline and mortality information. For illustration, any missing baseline value was imputed by the corresponding sample mean calculated from its observed counterparts in the study. We randomly split the data set with equal sizes as the training and hold-out sets.

For our analysis, we consider 16 baseline covariates: gender, histological stage of the disease (1, 2, 3, and 4), presence of ascites, edema, hepatomegaly or enlarged liver, blood vessel malformations in the skin, log-transformed age, serum albumin, alkaline phosphatase, aspartate aminotransferase, serum bilirubin, serum cholesterol, urine copper, platelet count, standardized blood clotting time, and triglycerides. Three models discussed in Section 2 with these covariates included additively were considered in the model selection. They are the Cox model, the AFT model, and the new RMET model. Moreover, since a more parsimonious Cox model using five of these covariates (edema, log-transformed age, bilirubin, albumin, and standardized blood clotting time) has been established as a prediction model in the literature (Fleming and Harrington, 1991), we also considered the aforementioned three types of models with these five covariates additively in our analysis. There are, therefore, six different models were considered. Note that there was no variable selection procedure involved in the model building stage for this illustration.

Figure 2 shows the KM curve for the patients' survival, estimated using the entire data set. None of the patients' follow-up times exceed 13 years. Since the tail part of the KM estimate is not stable. We let $\tau = 10$ years for illustration. The overall 10-year survival rate is about 44%. Table 1 presents the L_1 prediction error estimates for the RMET up to 10 years for the three model building procedures based on 100 random 5-fold CVs. With CV, the L_1 prediction error is minimized at 1.94 in years when the proposed regression Model (1) with the logistic link function $\eta(\mu) = \log\{\mu/(\tau-\mu)\}$ based on five baseline covariates is utilized.

The final model is obtained by fitting the RMET model with five covariates:

$$\begin{aligned}\eta\{\mu(z)\} = 6.36 &- 0.14\times\text{edema} - 2.95\times\log(\text{age}) + 0.99\times\log(\text{bilirubin})\\ &- 0.35\times\log(\text{albumin}) + 0.34\times\log(\text{clotting time}).\end{aligned}$$

We then use the score created by this model to make prediction and stratification for subjects in the hold-out set.

For predicting future restricted event time, we use the procedures proposed in Section 4 to estimate the subject-specific RMET $\xi(v)$ over a range of score v's, and the perturbation-resampling method with $M = 500$ independent sets of $\{Q_1,\ldots,Q_n\}$ from the unit exponential to construct its 0.95 pointwise and

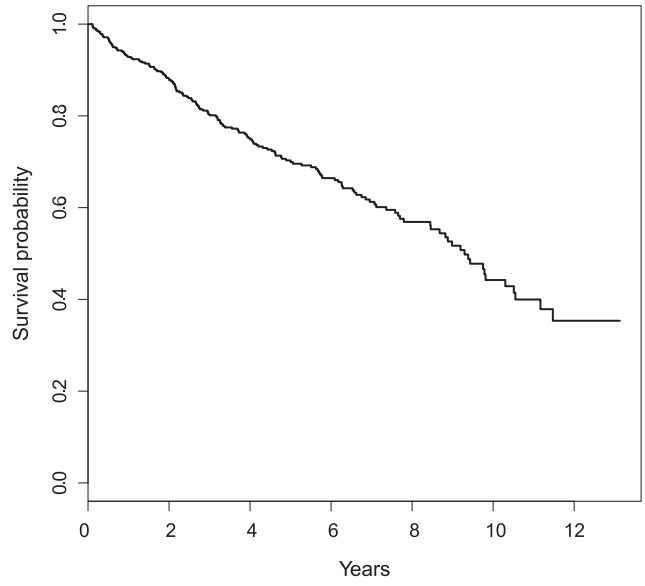

Fig. 2. KM estimate of the overall patient survival function based on the PBC data.

Table 1. *L_1 prediction error estimates for the RMET up to* 10 *years of the three model building procedures based on* 100 *random* 5*-fold CVs*

	L_1 prediction error with CV		
	Cox model	AFT model	New model
5 covariates	2.34	2.00	1.94
16 covariates	2.34	2.11	2.10

simultaneous confidence intervals over the interval [0.07, 9.52] in years, where 0.07 and 9.52 are the 2nd and 98th percentiles of observed scores in the hold-out set. Here, we let $K(\cdot)$ be the Epanechnikov kernel and the bandwidth be 2.1, as selected via CV discussed in Appendix C of supplementary material available at *Biostatistics* online. The results are presented in Figure 3(a). For comparison, we also present the corresponding results in Figure 3(b) with the survival function of the censoring time C being estimated based on the entire sample rather than locally as proposed in Section 4. As expected, the resulting estimator for $\xi(\upsilon)$ is less accurate, e.g. the 95% confidence interval for $\xi(5)$ is 24.8% wider when the survival function of C is estimated based on the entire sample.

As a conventional practice, we may stratify the subjects in the hold-out set into groups such as low, intermediate, and high risk groups by discretizing the continuous score. For example, we may create four classes based on the quartiles of the scores. Figure 4 presents the KM curves for these four strata. Visually these curves appear quite different. Moreover, their estimated RMETs and the standard error estimates (in parentheses) are 3.59 (0.46), 6.26 (0.53), 8.50 (0.40), and 9.14 (0.31) in years, respectively. These indicate that the scoring system does have reasonable discriminating capability with respect to the patients' RMET. How to construct an "efficient" categorization of the existing scoring system warrants future research.

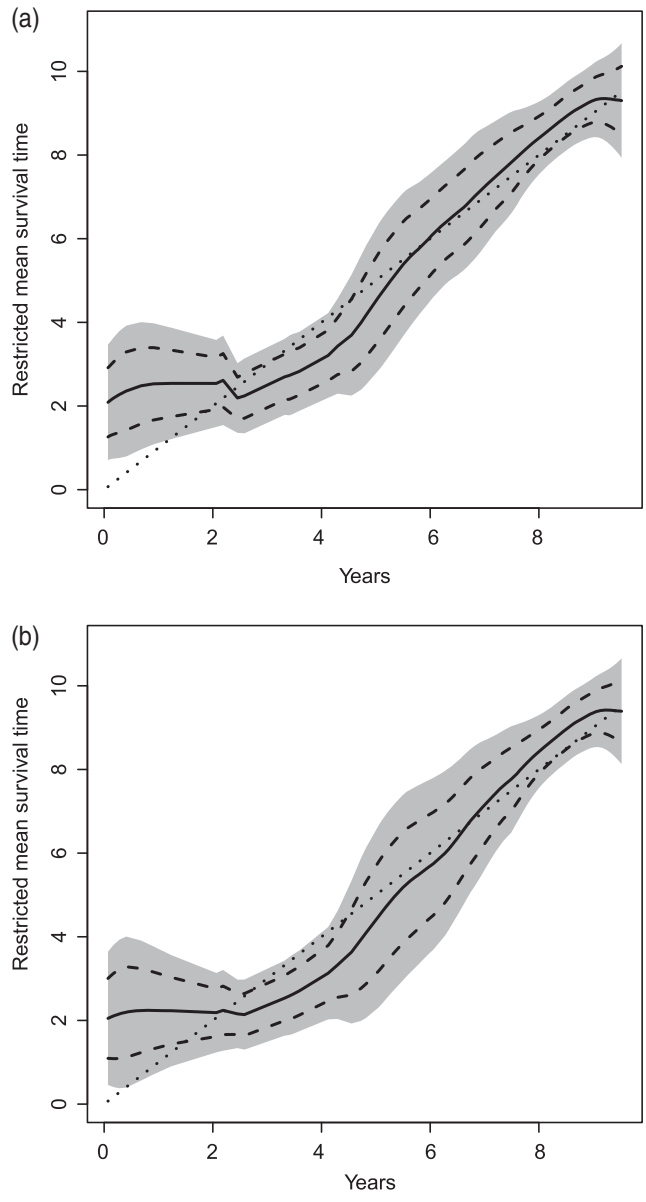

Fig. 3. Estimated subject-specific restricted mean survival time (solid curve) over the score, and its 95% pointwise (dashed curve) and simultaneous confidence intervals (shaded region). The dotted line is the 45° reference line. The survival function of the censoring time C is estimated locally (a) and based on the entire sample (b).

6. REMARKS

In comparing two groups with censored event-time data, the point and interval estimates of the two RMETs and their counterparts for the group contrast provide much more clinically relevant information than, for example, the hazard ratio estimates. The results from the HIV data set from ACTG 320 discussed in Section 1 is a good example, in that the three-drug combination is statistically significantly better than

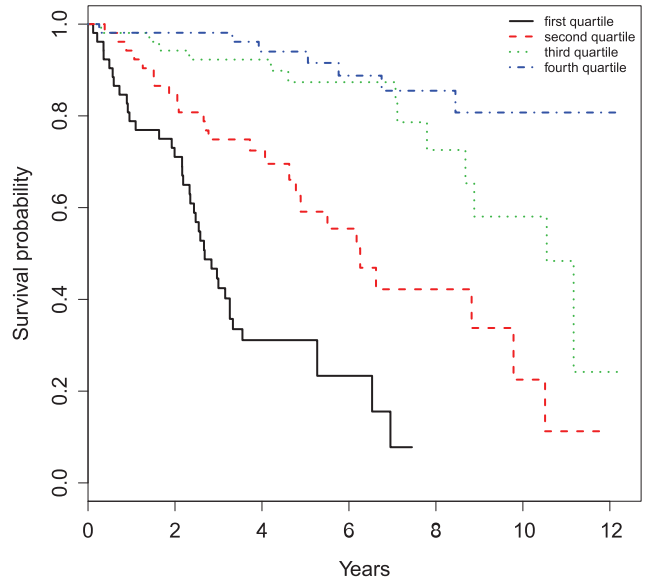

Fig. 4. KM estimates of the survival functions of the four strata divided by quartiles of the scores based on the PBC data.

the conventional therapy, but the gain from the new treatment with respect to RMET was not as impressive from a clinical standpoint, likely due to the relatively short follow-up time. Note that for this case, the median event time cannot be estimated empirically due to heavy censoring. Moreover, we cannot evaluate models using the individual predicted error, such as the L_1 distance function, with the median event time. It follows that the RMET is probably among the most meaningful, model-free, global measures for the distribution of the event time to evaluate the treatment efficacy. The choice of τ to define the RMET is crucial, which may be determined at the study design stage with respect to clinical relevance and feasibility of conducting the study.

Note that one of the attractive features of the model which directly relates the RMET to its covariates proposed here is that the score created is free of the censoring distribution even when the model is not correctly specified. On the other hand, those scores built from the Cox or AFT models depend on the study-specific censoring distribution when the model is misspecified.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

We are grateful to the Editor, Associate Editor, two referees, and Dr Brian Claggett for constructive comments on the paper. *Conflict of Interest:* None declared.

Funding

The work is partially supported by the National Institutes of Health grants (R01 AI052817, RC4 CA155940, U01 AI068616, UM1 AI068634, R01 AI024643, U54 LM008748, R01 HL089778) and contracts.

REFERENCES

ANDERSEN, P. K., HANSEN, M. G. AND KLEIN, J. P. (2004). Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Analysis* **10**(4), 335–350.

CAI, T., TIAN, L., UNO, H., SOLOMON, S. D. AND WEI, L. J. (2010). Calibrating parametric subject-specific risk estimation. *Biometrika* **97**(2), 389–404.

CHEN, P. Y. AND TSIATIS, A. A. (2001). Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics* **57**(4), 1030–1038.

COX, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

DABROWSKA, D. M. (1987). Non-parametric regression with censored survival time data. *Scandinavian Journal of Statistics* **14**(3), 181–197.

DABROWSKA, D. M. (1989). Uniform consistency of the kernel conditional Kaplan–Meier estimate. *The Annals of Statistics* **17**(3), 1157–1167.

FLEMING, T. R. AND HARRINGTON, D. P. (1991). *Counting Processes and Survival Analysis*, Volume 8. New York: Wiley Online Library.

GILBERT, P. B., WEI, L. J., KOSOROK, M. R. AND CLEMENS, J. D. (2002). Simultaneous inferences on the contrast of two hazard functions with censored observations. *Biometrics* **58**, 773–780.

HAMMER, S. M., SQUIRES, K. E., HUGHES, M. D., GRIMES, J. M., DEMETER, L. M., CURRIER, J. S., ERON, J. J., FEINBERG, J. E., BALFOUR, H. H., DEYTON, L. R. *and others*. (1997). A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and cd4 cell counts of 200 per cubic millimeter or less. *New England Journal of Medicine-Unbound Volume* **337**(11), 725–733.

JIN, Z., LIN, D. Y., WEI, L. J. AND YING, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika* **90**(2), 341–353.

KALBFLEISCH, J. D. AND PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data*. New York: John Wiley & Sons.

LIN, D. Y. AND WEI, L. J. (1989). The robust inference for the Cox proportional hazards model. *Journal of American Statistical Association* **84**, 1074–1078.

PARK, B. U., KIM, W. C., RUPPERT, D., JONES, M. C., SIGNORINI, D. F. AND KOHN, R. (1997). Simple transformation techniques for improved non-parametric regression. *Scandinavian Journal of Statistics* **24**(2), 145–163.

ROYSTON, P. AND PARMAR, M. K. B. (2011). The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine* **30**(19), 2409–2421.

RUDSER, K. D., LEBLANC, M. L. AND EMERSON, S. S. (2012). Distribution-free inference on contrasts of arbitrary summary measures of survival. *Statistics in Medicine* **31**(16), 1722–1737.

THERNEAU, T. M. AND GRAMBSCH, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Berlin: Springer.

TIAN, L., CAI, T., GOETGHEBEUR, E. AND WEI, L. J. (2007). Model evaluation based on the sampling distribution of estimated absolute prediction error. *Biometrika* **94**(2), 297–311.

TIAN, L., ZUCKER, D. AND WEI, L. J. (2005). On the Cox model with time-varying regression coefficients. *Journal of the American Statistical Association* **100**(469), 172–183.

TSIATIS, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics* **18**(1), 354–372.

ZHAO, H. AND TSIATIS, A. A. (1997). A consistent estimator for the distribution of quality adjusted survival time. *Biometrika* **84**(2), 339–348.

ZHAO, H. AND TSIATIS, A. A. (1999). Efficient estimation of the distribution of quality-adjusted survival time. *Biometrics* **55**(4), 1101–1107.

ZHAO, L., TIAN, L., UNO, H., SOLOMON, S. D., PFEFFER, M. A., SCHINDLER, J. S. AND WEI, L. J. (2012). Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring, and analyzing a comparative clinical study. *Clinical Trials* **9**(5), 570–577.

[*Received June 25, 2013; revised October 7, 2013; accepted for publication October 14, 2013*]