Statistics
in Medicine

Research Article

Received 26 December 2010, Accepted 12 October 2012 Published online 9 November 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5673

A K-nearest neighbors survival probability prediction method

D. J. Lowsky,[a*†] Y. Ding,[b] D. K. K. Lee,[c] C. E. McCulloch,[d] L. F. Ross,[e] J. R. Thistlethwaite[e] and S. A. Zenios[f]

We introduce a nonparametric survival prediction method for right-censored data. The method generates a survival curve prediction by constructing a (weighted) Kaplan–Meier estimator using the outcomes of the K most similar training observations. Each observation has an associated set of covariates, and a metric on the covariate space is used to measure similarity between observations. We apply our method to a kidney transplantation data set to generate patient-specific distributions of graft survival and to a simulated data set in which the proportional hazards assumption is explicitly violated. We compare the performance of our method with the standard Cox model and the random survival forests method. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: nonparametric survival analysis; survivor function; right-censored data; Kaplan–Meier; K-nearest neighbors; Mahalanobis distance; Cox regression; organ transplantation

1. Introduction

The semiparametric Cox proportional hazards model [1] is the gold standard for modeling the effect of covariates on survival outcomes. However, its underlying proportional hazards assumption is not suitable in all settings. According to [2], some strategies available for dealing with violations of the assumption include (1) incorporating nonproportional covariates as stratification factors rather than regressors, (2) partitioning the time axis into intervals so that the proportional hazards assumption holds within each interval, (3) using time-dependent coefficients, or (4) using an accelerated failure time model or Aalen's additive hazards model [3]. Although each of these approaches has advantages and settings in which it is most suitable, all of them require considerable time, effort, and modeling expertise in order to analyze the data set, verify that the associated structural assumptions hold, and make appropriate modeling choices. For example, incorporating stratification factors requires selecting upon which variables to stratify and the strata per variable, whereas using time-dependent coefficients requires selecting a form for the time dependency.

Another important distinction is that the main purpose of the Cox model is not to predict survival but to identify the relative hazard associated with different covariates. In applications where the main goal is to predict survival probabilities, it is preferable to make as few (possibly unwarranted) parametric assumptions as possible to improve prediction accuracy.

In this paper, we introduce a simple nonparametric method for predicting survival probability under the setting of right-censored data without competing risks. Our method constructs Kaplan–Meier survival curves on the basis of the observed survival of the K-nearest neighbors in a training set. Distance between data points is measured with a metric on the covariates associated with the observations.

[a]*RAND Corporation, Santa Monica, CA 90407, U.S.A.*
[b]*Sauder School of Business, University of British Columbia, Vancouver, BC V6T 1Z2, Canada*
[c]*School of Management, Yale University, New Haven, CT 06520, U.S.A.*
[d]*Division of Biostatistics, UCSF, San Francisco, CA 94107, U.S.A.*
[e]*University of Chicago Medical Center, Chicago, IL 60637, U.S.A.*
[f]*Graduate School of Business, Stanford University, Stanford, CA 94305, U.S.A.*
**Correspondence to: D. J. Lowsky, RAND Corporation, Santa Monica, CA 90407, U.S.A.*
[†]*E-mail: dlowsky@rand.org*

Copyright © 2012 John Wiley & Sons, Ltd.

Our initial choice for the metric is the Mahalanobis distance, a broadly applicable metric with the advantage of imposing no structural assumptions on the data.

Of the nonparametric survival prediction methods in existing literature, those based on classification and regression trees [4] are perhaps the most well known [5]. A recent example is bagged survival trees [6], where multiple survival trees are built from bootstrapped samples of the data. Then a survival prediction for a new observation is created by constructing a Kaplan–Meier estimator using similar training observations, which are the (possibly duplicated) observations belonging to the same terminal node as the target in any of the bootstrapped survival trees. Along another direction, the random survival forests (RSF) method [7] extends the random forest algorithm [8] to the analysis of right-censored survival data. The method generates a prediction for the cumulative hazard (Nelson–Aalen estimator) from which a predicted survival curve can be computed. Although such ensemble methods can generate superior predictions, they can require substantial computational resources and the determination of a number of tuning parameters. In contrast, the method introduced in this paper appears to be substantially less computationally intensive and only requires choosing how many nearest neighbors to use.

We apply our method to data on kidney transplantation outcomes, where we attempt to make predictions for the graft survival probability. We also apply our method to a simulated data set in which the proportional hazards assumption is explicitly violated. We compare our method's performance with the standard Cox model and the RSF method, where performance is measured using the integrated prediction error curve (IPEC) [9]. We find that our method outperforms the Cox model in nonproportional hazards settings and is outperformed by the more sophisticated and computationally intensive RSF method.

Our paper is organized as follows. Section 2 describes the method. Section 3 describes the application of our method to the kidney transplantation setting and the simulated data set. Performance results are also presented in this section. Section 4 concludes by discussing the strengths and limitations of the method and directions for future research.

2. Method description

Given a new observation j, our method generates a prediction for j's survival curve by creating a Kaplan–Meier curve from (possibly censored) survival times of 'similar' observations from an existing data set $\mathcal{S}$.

Each observation $i \in \mathcal{S}$ is associated with a set of covariates $x_i \in \mathbb{R}^P$, an event time $t_i \geqslant 0$, and a censoring indicator δ_i, where 0/1 indicates a censored/uncensored observation. The set of covariates x_j is associated with the new observation j. Let $d(x_i, x_j)$ be a metric on x that measures how similar observations i and j are. The K-nearest neighbors of j in $\mathcal{S}$ are chosen according to $d(x_i, x_j)$ to form the set $\mathcal{S}_j^K \subseteq \mathcal{S}$. A weighted analogue of the Kaplan–Meier survival curve [10] generated from the observations in $\mathcal{S}_j^K$,

$$\hat{S}^K(t|x_j; w) = \prod_{i \in \mathcal{S}_j^K : t_i < t} \left(1 - \frac{d_i^w}{n_i^w}\right), \tag{1}$$

serves as a prediction for observation j's survival curve, where $n_i^w = \sum_{k \in \mathcal{S}_j^K} w(d(x_j, x_k)) \cdot I(t_k \geqslant t_i)$ is the weighted number of observations in $\mathcal{S}_j^K$ at risk just before t_i and $d_i^w = \sum_{k \in \mathcal{S}_j^K} w(d(x_j, x_k)) \cdot \delta_k \cdot I(t_k = t_i)$ is the weighted number of deaths at time t_i. The weighting function $w(\cdot)$ for the observations in $\mathcal{S}_j^K$ is non-increasing in the distance $d(\cdot, \cdot)$ of the observation to x_j, thereby placing greater emphasis on training observations that are more similar to the new observation. Because $\hat{S}^K(t|x_j; w)$ is invariant to scalings of w, without loss of generality, we require that $w(0) = 1$. Note that setting $w(\cdot) \equiv 1$ gives equal weighting to each of the K-nearest neighbors and hence the original Kaplan–Meier estimator is recovered.

Choice of K: K can be chosen by separating $\mathcal{S}$ into training and validation sets and performing validation testing. The K-nearest neighbors used to generate the predicted survival curve are drawn from the training set. The validation set is a holdout data set used to test the performance of a range of K values and select the best performing K.

Copyright © 2012 John Wiley & Sons, Ltd.

Choice of metric $d(x_i, x_j)$: There are many possible options for the metric, and the choice should be informed by the problem context. One versatile option is the Mahalanobis distance [11], which is used in this paper:

$$d(x_i, x_j) = \sqrt{(x_i - x_j)' \Sigma_{\mathcal{S}}^{-1} (x_i - x_j)}, \tag{2}$$

where $\Sigma_{\mathcal{S}}$ is the covariance matrix of x for observations belonging to $\mathcal{S}$. This inverse weighting in the Mahalanobis distance renders the metric scale-invariant, and therefore it is independent of the units in which the covariates are expressed. Note that (2) can also be applied to categorical variables by first transforming them into a set of indicator variables.

Choice of weighting function $w(\cdot)$: Multiple options exist for the weighting function, and there may even exist an optimal weighting scheme (the investigation of which is beyond the scope of this paper). Given that we are introducing this method for the first time, for simplicity we selected the constant weight function $w(\cdot) \equiv 1$.

For the remainder of this paper, we shall refer to this method as the Mahalanobis K-nearest neighbor (MKNN) method.

3. Application

We evaluate the performance of our method using two approaches. First, we apply it to a data set of kidney transplant recipients to test prediction of graft survival probability. Second, using the original data set, we created a simulated data set that explicitly violates the proportional hazards assumption and tested the performance of our method in such a setting. In both settings, we compare our method's performance with that of the Cox model and the RSF method.

3.1. Original data description

Our data set is provided by the United States Renal Data System (USRDS) [12] and contains all kidney transplant procedures performed from 1996 to 1999 that were covered by Medicare. The event of interest is graft failure or death, whichever occurs first. Hence, graft survival time is defined as the time elapsed from transplant date until graft failure or death due to any cause, censored at 31 December 2004. The covariates consist of sex (male, female), race (Native American, Asian, Black, White, unknown, other), age (0–9, 10–19, 20–39, 40–49, 50–59, 60–69, 70+ years), pre-event dialysis time (none, time (counted every half year before 5 years, counted every year for 5–9 years)), blood type (O, A, B, AB), peak panel-reactive antibody (0%, 1–20%, 21–40%, 41–60%, 61–80%, 81–100%), body mass index quintile (1st, 2nd, 3rd, 4th, 5th, missing), disease cause of end-stage renal disease (diabetes, hypertension, glomerulonephritis, cystic kidney, other urologic, other cause, unknown cause), pre-transplant blood transfusion (no, yes, unknown), previous transplant (no, yes), whether the kidney came from a living donor (no, yes), whether the kidney came from an expanded criteria donor (no, yes), and (year of transplant–1995). The data set included 51,088 total observations.

To assess to what extent the graft survival times satisfy the proportional hazards assumption, we ran a Schoenfeld residuals test [13] on the covariates and found that four covariates (age, pre-event dialysis time, living donor, expanded criteria donor) violated the assumption at the 5% significance level.

3.2. Simulated data description

We created a new simulated data set in which the proportional hazards assumption is explicitly violated, using a model fitted to the USRDS graft survival data as follows. We first fit an exponential hazard model $\lambda(t|x) = \exp(\beta_0 + \beta' x)$ to the data set, yielding estimates for β_0 and β. Then we simulated the survival times of all patients in our full data set using three hazard models, producing three new data series. The hazard models are of the form

$$\lambda(t|x) = \begin{cases} \exp(\beta_0 + \beta' x) & t \leqslant 2.5 \text{ years} \\ \exp(\beta_0 + \beta' x + \theta \cdot I_{(\text{age} \geqslant 50)}) & t > 2.5 \text{ years} \end{cases}$$

for $\theta = 0, -2, -4$. We randomly generated all censoring times. Figure 1 shows the survival curves for a sample patient under these distributions. A more negative value of θ results in a greater violation of the proportional hazards assumption. For $\theta = 0$, the hazard rate is time independent; hence, its survival

Copyright © 2012 John Wiley & Sons, Ltd.

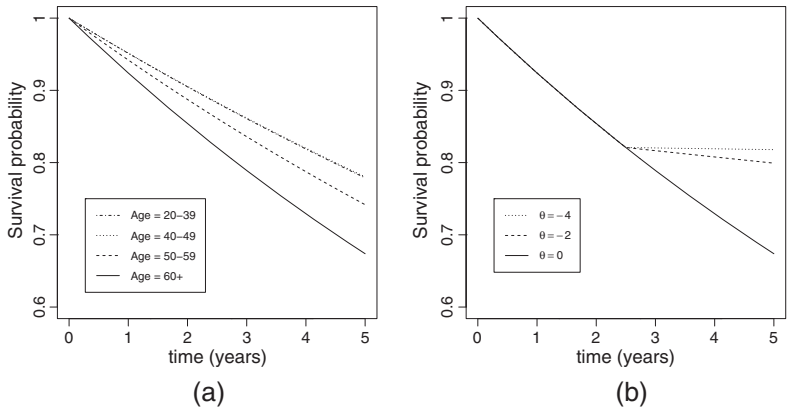

Figure 1. Illustration of model used to generate simulated data. Patient covariates in this sample are sex = male, race = White, no pre-transplant dialysis, blood type = O, peak panel-reactive antibody = 0%, body mass index = 3rd quintile, disease = diabetes, prior blood transfusion = no, prior transplant = no, and transplant year = 1996. (a) Exponential distribution fits for different values of age (20–39, 40–49, 50–59, 60+ years) and (b) survival distributions for simulated data, for age = 60+ years, and θ=0 (exponential fit), $-2, -4$.

curve is smooth as seen in Figure 1(a). For the other two values of θ, we observe a kink at $t = 2.5$ years, the sharpness of which increases as θ becomes more negative, as seen in Figure 1(b).

3.3. Performance metric

We used the IPEC [9] as our metric of predictive accuracy, where smaller IPEC values indicate greater predictive accuracy. For a single test observation j, $IPEC_j$ is the integrated squared difference between the actual and predicted survival, weighted in a way to account for censoring bias:

$$IPEC_j = \frac{1}{T}\int_0^T \hat{W}_j(t)\left\{I(t_j > t) - \hat{S}^K(t|x_j; w)\right\}^2 \mathrm{d}t. \tag{3}$$

$[0, T]$ represents the time interval of interest and $\hat{W}_j(t)$ are the inverse probability of censoring weights, defined as follows:

$$\hat{W}_j(t) = \frac{\delta_j I(t_j \leq t)}{\hat{G}(t_j - |x_j)} + \frac{I(t_j > t)}{\hat{G}(t|x_j)}, \tag{4}$$

where $\hat{G}(t|x)$ is a consistent estimator for the survivor function of the censoring distribution. Our aggregate performance is the average $IPEC_j$ over M test observations:

$$IPEC = \frac{1}{M}\sum_j IPEC_j. \tag{5}$$

We numerically integrated (3) for $T = 5$ years using a step size of 0.25 years (using other step sizes minimally changed the results). For $\hat{G}$, we used the Kaplan–Meier estimator, thereby implicitly assuming that the censoring distribution is independent of t_j and x_j.

3.4. Testing procedure

To evaluate the performance of our method against the Cox model and the RSF method, we consider 16 different settings: combinations of four data sets (original distribution, simulated data with $\theta = \{0, -2, -4\}$) and four training set sizes $\{500, 1000, 3000, 7500\}$. The maximum training set size was limited to 7500 observations because of the computation time required to perform repeated tests on larger training set sizes.

To generate a distribution for the prediction error in each setting, we bootstrapped independent samples from each of the four data distributions, as follows:

Copyright © 2012 John Wiley & Sons, Ltd.

Original data set: We randomly divided the original data set of 51,088 observations into master training, validation, and test sets. We randomly split 80/20 all transplants from years 1996 to 1998, generating the master training and validation sets of size 30,051 and 7512, respectively. All transplants performed in 1999 comprised the master test set of size 13,525. For each of the four training set sizes, we drew 20 independent samples from the master training set and proportionately sized samples from the master validation set. We used the full master test set in testing across all training set sizes and replicates.

Simulated data set: For each combination of θ and training set size (12 settings in total), we generated 20 independent training set samples and corresponding validation set samples (using the same 80/20 size ratio). For each θ, we generated and used a single test set of size 13,525 across all training set sizes.

In each of the 16 settings, neighbors were drawn from the training set, and the validation set was used to determine the optimal K. The values of K tested were 100–500 at increments of 100 and from 1000 to 7000 at increments of 1000.

We used the R method **rsf** in package **randomSurvivalForest** [14] to compute predicted survival probabilities for the RSF method. We used the default parameters for all calls to **rsf**. We performed all other computations in MATLAB. We performed all computations on a Linux server with an eight-core 2.7 GHz AMD Opteron processor with 32 GB of RAM. We also tracked the average running time for each training set size and method.

We also tested how sensitive our method is to the presence of irrelevant covariates in the following way. We generated a set of N binary covariates and added them to the original data set, where N varied among 5, 10, 20, 30, or 40. The ith irrelevant covariate was drawn from a Bernoulli(p_i) distribution with $p_i \sim U[0, 1]$. Using the same method as described earlier, we compared the performance of MKNN on the original data set with its performance on the augmented data set.

3.5. *Results*

Validation testing: For the MKNN method, validation testing revealed that $K = \{100, 100, 300, 500\}$ was optimal for training set sizes $\{500, 1000, 3000, 7500\}$, respectively.

Performance results: Figure 2 presents results for all 16 combinations of data sets and training set sizes. The figure illustrates the predictive performance of the MKNN and RSF methods relative to the Cox model. The ratio of each method's IPEC score to the Cox model's IPEC score, representing each method's performance relative to the Cox model, is presented for the 20 replicates in box plot format. Lower values indicate better relative performance, with values lower than 1 indicating that the method outperformed the Cox model.

In seven out of the eight simulated data settings where θ is nonzero (i.e., nonproportional hazards), and in 11 out of the 16 total settings evaluated, MKNN outperformed the Cox model more than half the time, that is, the median relative performance was less than 1. Across all the training set sizes, MKNN's performance relative to the Cox model in the simulated data sets generally improved as the proportional hazards violation becomes stronger (θ becomes more negative), reflecting MKNN's advantage in nonproportional hazards settings. MKNN did not show a consistent advantage over the Cox model in the settings of the original data or the exponential hazard model ($\theta = 0$), suggesting that any violation of the proportional hazards assumption in the original data set was mild. RSF outperformed MKNN in all the settings evaluated.

Computation time: Table I presents the running time to train and compute predictions for the test set, averaged over 20 training/validation replicates. The Cox model's running time is always less than 1 min, and MKNN's running time appears to grow sublinearly with training set size. In contrast, RSF's running time roughly triples with every doubling of training set size. As a comparison, for training set size 500, MKNN takes 6 min and RSF takes 1 min, whereas for training set size 7500, MKNN takes 9 min and RSF takes over an hour.

Irrelevant covariates: The performance of MKNN was unchanged when 0, 5, or 10 irrelevant covariates were added to the data set and only worsened by less than 1% when 20, 30, or 40 irrelevant covariates were added. Inclusion of irrelevant covariates had little effect on MKNN performance.

4. Discussion

We present a new method for survival probability prediction that generates a Kaplan–Meier survival curve using the observed outcomes of the K-nearest neighbors in a training set, where proximity is measured using a metric on the covariate space. We tested our method by applying it to four data sets: a data

Copyright © 2012 John Wiley & Sons, Ltd.

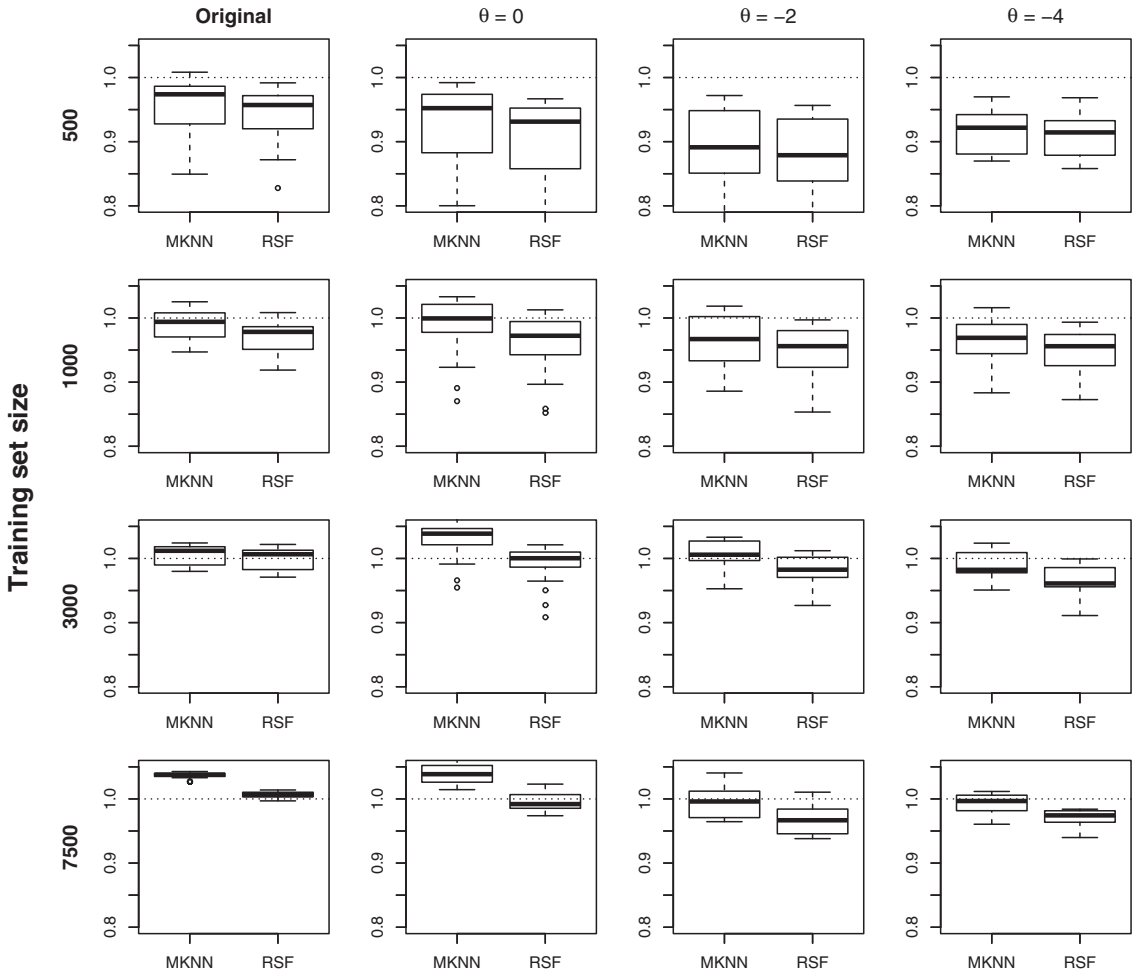

Figure 2. Box plots illustrating performance results of Mahalanobis K-nearest neighbor (MKNN) and random survival forests (RSF) methods relative to the Cox model. Each box represents a different combination of data model and training set size. Columns represent the different data sets, where the first column represents the original data set and the second to fourth columns represent the simulated data for $\theta = 0, -2, -4$. Rows represent different training set sizes $\{500, 1000, 3000, 7500\}$. Each value is the ratio of the method's integrated prediction error curve (IPEC) score to the Cox model's IPEC score, representing performance relative to the Cox model. The box represents the first, second (median), and third quartiles of relative performance across the 20 replicates in each setting. Values below the dotted line indicate cases where the method outperforms the Cox model.

Table I. Running time (minutes) to train and compute predicted survival times for the test set, averaged over 20 training/validation replicates.

	Training set size					
	500	1000	3000	7500	15,000	30,051
MKNN	6	6	6	9	13*	21*
RSF	1	3	14	62	195*	680*
Cox	<1	<1	<1	<1	< 1*	<1*

*Results based on a single training/validation replicate only.

set of kidney transplant recipients to predict graft survival probability and three simulated data sets generated by a family of hazard distributions. Our method outperformed the Cox model in cases where the proportional hazards assumption was violated. Our method was outperformed by the more sophisticated RSF method.

An advantage of our method is that it makes no parametric assumptions, and therefore no particular modeling assumptions need to be imposed. Another advantage is its simplicity, requiring only one tuning

Copyright © 2012 John Wiley & Sons, Ltd.

parameter (K) to be set. It also appears to require substantially lower computational effort than other more sophisticated nonparametric methods such as RSF.

Another strength of the method is that it can be used with any metric, whose choice can be tailored to the specific problem context. In this paper, we selected the Mahalanobis distance because it is agnostic to the structure of the problem and therefore can be applied broadly. Its shortcoming is that it does not take the relative explanatory powers of individual covariates into account. When the modeler has a better understanding of when two observations are likely to have similar outcomes, the Mahalanobis distance may be replaced with a more appropriate metric that takes advantage of this knowledge. The treatment of ordinal variables is one example. Using the Mahalanobis distance, an ordinal variable may be included in two ways: (1) it can be converted into categorical variables, thereby losing the ordering information; or (2) it can be treated as an integer-valued ordinal variable where adjacent value pairs are treated as equidistant, even though the natural separation between adjacent value pairs may differ across the range of values. If the modeler has some sense of how the values should be separated, then a metric that incorporates this information can be used instead.

We also note that our method can be extended to handle competing risks. Under competing risks, the primitive of interest is not the survivor function $S(t|x) = \mathbb{P}(T \geqslant t|x)$, but rather the cumulative incidence function $F_j(t|x) = \mathbb{P}(T \leqslant t, \delta = j|x)$ for event type $j = 1, \ldots, J$ (the censoring indicator δ takes on values in $\{1, \ldots, J\}$ if one of the J competing events occurred and is 0 if the observation is censored). An estimate for $F_j(t|x)$ can be obtained using the K-nearest neighbors of x to generate an Aalen–Johansen estimate [15] for the cumulative incidence function. To evaluate the predictive accuracy of $\hat{F}_j(t|x)$, the integrated version of (3) from [16] can be used as the performance metric. This is analogous to $IPEC_j$ in that both scores measure the integrated squared difference between the actual and predicted outcomes, weighted by the inverse probability of censoring. Alternatively, [17] provides a metric based on ROC curves (rather than squared differences) for evaluating the accuracy of $\hat{F}_j(t|x)$. Detailed investigation of extending our method to competing risks is left for future research.

Despite the relatively simple choices made for the metric and weighting function in this introductory use of our method, it outperformed the Cox model when the proportional hazards assumption was violated. With more sophisticated choices for these components, the method's performance should improve further. Additional investigation into these possibilities is also left for future research.

Acknowledgements

The authors thank the review team whose suggestions helped to significantly improve the paper. The authors also wish to thank Trevor Hastie for first suggesting this concept during conversations regarding related research and the United States Renal Data System (USRDS) for providing the data set for this research. The findings and opinions in the manuscript are those of the authors and do not represent USRDS or NIH policy.

References

1. Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society* 1972; **34**(2):187–220.
2. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. Springer, 2000.
3. Aalen OO. A linear regression model for the analysis of life times. *Statistics in Medicine* 1989; **8**:907–925.
4. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Wadsworth: California, 1984.
5. Bou-Hamad I, Larocque D, Ben-Ameur H. A review of survival trees. *Statistics Surveys* 2011; **5**:44–71.
6. Hothorn T, Lausen B, Benner A, Radespiel-Troger M. Bagging survival trees. *Statistics in Medicine* 2004; **23**:77–91.
7. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *The Annals of Applied Statistics* 2008; **2**(3):841–860.
8. Breiman L. Random forests. *Machine Learning* 2001; **45**:5–32.
9. Gerds TA, Schumacher M. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal* 2006; **48**(6):1029–1040.
10. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 1958; **53**:457–481.
11. Mahalanobis PC. On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India* 2008; **2**(1):49–55.
12. U.S. Renal Data System. Standard Analysis Files, 2006.
13. Schoenfeld D. Residuals for the proportional hazards regression model. *Biometrika* 1982; **69**(1):239–241.
14. Ishwaran H, Kogalur U. Random survival forests r-package v.3.6.3, 2010. http://cran.r-project.org/web/packages/randomSurvivalForest/. [Accessed on August 2012].

Copyright © 2012 John Wiley & Sons, Ltd.

15. Aalen OO, Johansen S. An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics* 1978; **5**:141–150.
16. Schoop R, Beyersmann J, Schumacher M, Binder H. Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. *Biometrical Journal* 2011; **53**:88–112.
17. Saha P, Heagerty PJ. Time-dependent predictive accuracy in the presence of competing risks. *Biometrics* 2010; **66**:999–1011.

Copyright © 2012 John Wiley & Sons, Ltd.