Received: 9 May 2017 | Accepted: 10 May 2017

DOI: 10.1111/jep.12779

WILEY Journal of Evaluation in Clinical Practice
International Journal of Public Health Policy and Health Services Research

ORIGINAL ARTICLE

Modeling time-to-event (survival) data using classification tree analysis

Ariel Linden DrPH[1,2] | Paul R. Yarnold PhD[3]

[1] President, Linden Consulting Group, LLC, Ann Arbor, MI, USA

[2] Research Scientist, Division of General Medicine, Medical School, University of Michigan, Ann Arbor, MI, USA

[3] President, Optimal Data Analysis, LLC, Chicago, IL, USA

Correspondence
Ariel Linden, Linden Consulting Group, LLC, 1301 North Bay Drive, Ann Arbor, MI 48103, USA.
Email: alinden@lindenconsulting.org

Abstract

Rationale, aims, and objectives: Time to the occurrence of an event is often studied in health research. Survival analysis differs from other designs in that follow-up times for individuals who do not experience the event by the end of the study (called censored) are accounted for in the analysis. Cox regression is the standard method for analysing censored data, but the assumptions required of these models are easily violated. In this paper, we introduce classification tree analysis (CTA) as a flexible alternative for modelling censored data. Classification tree analysis is a "decision-tree"–like classification model that provides parsimonious, transparent (ie, easy to visually display and interpret) decision rules that maximize predictive accuracy, derives exact *P* values via permutation tests, and evaluates model cross-generalizability.

Method: Using empirical data, we identify all statistically valid, reproducible, longitudinally consistent, and cross-generalizable CTA survival models and then compare their predictive accuracy to estimates derived via Cox regression and an unadjusted naïve model. Model performance is assessed using integrated Brier scores and a comparison between estimated survival curves.

Results: The Cox regression model best predicts average incidence of the outcome over time, whereas CTA survival models best predict either relatively high, or low, incidence of the outcome over time.

Conclusions: Classification tree analysis survival models offer many advantages over Cox regression, such as explicit maximization of predictive accuracy, parsimony, statistical robustness, and transparency. Therefore, researchers interested in accurate prognoses and clear decision rules should consider developing models using the CTA-survival framework.

KEYWORDS

censoring, classification tree analysis, machine learning, survival

1 | INTRODUCTION

Time to the occurrence of an event is often studied in health-related research. Typically, the event is survival or, conversely, mortality, over a given period of observation. However, other events may be used as the endpoint, such as hospitalization, development of disease, or reaching a threshold for a physiologic marker.[1-3]

In survival analysis, data from individuals who do not experience the event by the end of the study are used in model estimation. Such individuals' survival times are called *censored*, indicating that that the study terminated before the event occurred or that the individual may have been lost to follow-up at some point during the study. In either case, censored survival times are used—along with the survival times of individuals who experienced the event during the course of the study—to construct the survival analysis model.[4]

Several regression-based models are specifically designed to assess the influence of covariates on survival in the presence of censoring. As with all regression analyses, these methods involve making assumptions about the data, including that variables are independent and that data can be modelled using linear combinations of these variables. Moreover, while these models generally show improved fit when additional variables are included, there is no good indicator of when the model is over-fit.[5] Given that health data are rarely strictly linear, but often exhibit interactions and conditional dependencies, assumptions underlying the validity of these survival models are easily violated.

To avoid these limitations, various machine learning algorithms have been proposed as alternatives for modelling survival data (see, for example[6-10]). Machine learning algorithms find the best-fitting model through automated processes that search through the data to detect patterns that may include interactions between variables, as well as interactions within subsets of variables. This is in contrast to conventional statistics, where a model is chosen and estimated on the basis of an a priori hypothesis about the underlying relationship between the variables, and then statistical tests are performed to evaluate whether the data satisfy crucial assumptions underlying the validity of the findings.[11] In short, machine learning allows the data to dictate the form of the model, whereas conventional statistics attempts to fit the data to an investigator-specified model.

In this paper, we introduce classification tree analysis (CTA) as a machine learning alternative to conventional regression-based models for analysing survival data. Classification tree analysis is a "decision-tree"–like classification model that provides accurate, parsimonious decision rules that are easy to visually display and interpret, while reporting *P* values derived via permutation tests performed at each node. This approach is attractive to clinicians using model-derived prognostic tools in daily practice and to investigators evaluating the effectiveness of interventions in which the outcome is the time to the occurrence of an event.

The paper is organized as follows. Section 2 briefly introduces CTA and describes the data source and analytic framework used in the current study. Section 3 reports and compares the findings of the CTA framework and the Cox semiparametric proportional hazard model,[12] the most widely-used method for analysing survival data. Section 4 describes the specific advantages of the CTA framework for developing risk prediction models compared with regression-based survival models and discusses extending this approach to the evaluation of treatment effects in health care interventions.

2 | METHODS

2.1 | A brief introduction to CTA

Classification tree analysis is an optimal discriminant analysis (ODA) model.[13] Optimal discriminant analysis is a machine learning algorithm used to identify the cutpoint on an ordered attribute (variable), or assignment rule for a categorical attribute, that optimally discriminates between two or more classes (eg, outcome categories).[14] The optimal cutpoint is determined by iterating through every value on the attribute and computing the effect strength for sensitivity (ESS), which is the mean sensitivity across classes standardized using a 0 to 100% scale on which 0 represents the discriminatory accuracy that is expected by chance and 100% represents perfect discrimination. By definition, the maximally accurate predictive model uses the optimal cutpoint that yields the highest ESS versus all other cutpoints. This optimal model is subjected to a nonparametric permutation test to assess the statistical significance of the cutpoint. Finally, the reproducibility and generalizability of the model are assessed using cross-validation methods, such as jackknife, bootstrap, or hold-out analysis, to determine how well it predicts the outcome in new subjects that may differ in their characteristics compared to subjects in the original sample.[15-19]

CTA models use one or more attributes to classify a sample of subjects into two or more subgroups represented as model endpoints (called "terminal nodes" by alternative decision-tree methods). Subgroups are known as "sample strata" because the CTA model stratifies the sample into subgroups that—when considered with respect to model attributes—are homogeneous within and heterogeneous between strata.[19] The hierarchically optimal CTA (HO-CTA) algorithm involves chained ODA models in which the initial ("root") node represents the attribute achieving the highest ESS value for the entire sample, and additional nodes yielding greatest ESS are iteratively added at every step on all model branches.[20,21] In contrast, the enumerated optimal CTA (EO-CTA) algorithm evaluates all possible combinations of the first 3 nodes, which dominate the solution.[16,22] The most robust globally optimal CTA (GO-CTA) algorithm explicitly evaluates all possible solutions (called the descendant family) and identifies the GO-CTA model reflecting the best combination of ESS and parsimony—yielding the highest ESS using the smallest number of strata.[19] The software that implements ODA and CTA models provides an array of options to control the modelling and validation process (see Yarnold and Soltysik[19] for a comprehensive discussion).

2.2 | Data

To demonstrate the use of CTA for survival analysis and to compare this approach to the standard Cox-regression model, we use a subset of data from the Framingham Heart Study, which has been collecting longitudinal data on residents of Framingham, Massachusetts since 1948, to gain insight into the epidemiology of cardiovascular disease (CVD) and its risk factors (see Mahmood et al[23] for an excellent historical perspective). We use data that comprise 4699 individuals free of CVD at their baseline exam and followed for up to 11 688 days (32 years). The variables include systolic and diastolic blood pressure (mmHg), age (years), serum cholesterol (mg/100 mL), body mass index (kg/m^2), gender, follow-up time (days), and an indicator of whether the individual developed CVD or was otherwise censored. The dataset was accessed as a supplement to the book "*Statistical Modeling for Biomedical Researchers*"[24] (http://biostat.mc.vanderbilt.edu/dupontwd/wddtext/index.html#datasets).

2.3 | Analytic approach

Split-half cross-validation methodology is used throughout the analytic process to evaluate model reproducibility and generalizability. This entails randomly drawing subjects from the full sample and assigning them into 1 of 2 groups: split-half 1 (SH1) or split-half 2 (SH2). Next, models are generated using SH1 as the training sample, and these models are used to make out-of-sample predictions for subjects in SH2 (the test or "hold-out" sample). This process is then repeated after switching the roles of SH1 (test sample) and SH2 (training sample). By definition, perfectly reproducible (generalizable) models are identical, and parallel models are identical except for the values of numerical cutpoints used on model branches.[18,19] Unless otherwise noted, all models make use of all available follow-up data.

For comparative purposes, estimates were derived by implementing the widely used Cox semiparametric proportional hazards model,[12] which models the effects of covariates on survival time. The quantity of interest in Cox regression is the *hazard* function, which may be described as the risk that the event will occur for a subject within an observation (ie, follow-up) period, given that the subject did not already have the event. A high hazard function indicates a high event rate (low survival probability), and conversely, a low hazard function indicates a low event rate (high survival probability). While Cox regression requires no assumptions about the distribution of failure (eg, development of CVD) times, it is assumed that the hazards between any 2 subjects are proportional over time (hence, the name *proportional hazards regression*), with the proportion being a function of the covariates.[4] For the current example, we estimate the time to the development of CVD at the end of follow-up, incorporating all of the covariates in the model as main effects. We estimate a model for the full sample and separately for both split-half samples. Standard errors are computed using a bootstrap procedure with 2000 repetitions.[25] Following the modelling procedure, we test if the proportional hazards assumption was violated.[26]

In nonweighted CTA, every subject has a weight of one and the model identifies attributes that classify subjects with maximum accuracy. In contrast, for a weighted CTA survival model, the weight of every subject is their follow-up time (ie, the number of days of follow-up), and the model identifies attributes that classify subject-days with maximum accuracy. For example, a subject without CVD and lost to follow-up after 1000 days is coded as class = 0 (no CVD), weight = 1000; a subject without CVD after maximum follow-up (eg, 10 585 d) is coded as class = 0, weight = 10 585; and a subject experiencing an event after 7919 days is coded as class = 1 (CVD), weight = 7919. The optimal cutpoint is identified by iterating through every value of the attribute and computing the weighted ESS (WESS), which is the mean weighted sensitivity (ie, percent of correctly predicted subject-days for each class) across the classes, standardized to a 0 to 100% scale on which 0 represents the weighted discriminatory accuracy expected by chance and 100% represents perfect discrimination. By definition, the maximally accurate predictive model uses the optimal cutpoint that achieves the highest WESS, versus all other cutpoints. The optimal model is subjected to a nonparametric permutation test to assess the statistical validity of the cutpoint. Model reproducibility and cross-generalizability are assessed using a hold-out (split-half) method, which is one of several possible cross-validation techniques typically implemented as part of the machine learning process.[16,19]

The present study demonstrates CTA-based survival analysis—which is implemented in 5 sequential steps. The first step uses weighted CTA for each attribute considered individually, separately for SH1 and SH2, to provide a "benchmark" for evaluating comparative predictive performance of multivariable models using two or more attributes.[27] The second step obtains the descendant family of all possible weighted CTA survival models using all available attributes, separately for SH1 and SH2. Identical and parallel models identified in SH1 and SH2 are considered reproducible and are hypothesized to cross-generalize to new independent random samples of subjects. The third step evaluates intermodel agreement of outcome predictions made by corresponding SH1 and SH2 survival models.[19] The fourth step involves a sensitivity analysis[28] to assess the consistency of the predictive accuracy yielded by the SH1 model used to classify the SH2 sample—and by the SH2 model used to classify the SH1 sample—over increasing annual follow-up lengths: The first analysis omits subjects with <1 year of follow-up; the second analysis omits subjects with <2 years of follow-up, and so on, until either all follow-up periods have been evaluated or until the point at which a follow-up period yields samples that provide inadequate statistical power.[19] The fifth and final step evaluates whether computing new CTA survival models for the SH1 and/or the SH2 samples improves WESS at strategic follow-up times identified in the sensitivity analysis. This is determined by whether or not the WESS decreases beyond some empirically defined level—indicating poor model fit. In such a circumstance, the modelling process is repeated beginning with step 2.

2.4 | Performance metrics

Several methods proposed to assess the accuracy of predictions derived from survival models—including the concordance index,[29,30] omnibus goodness-of-fit tests,[31,32] and measures of explained variation versus explained randomness[33]—have been criticized on methodological grounds. Accordingly, we use the widely used Integrated Brier Score (IBS)[34] and introduce comparisons of WESS and of generated survival curves.

First, we use the IBS to compare performance of the naïve nonadjusted model, Cox regression with covariates, and weighted CTA approaches, because its computation relies on 2 quantities available in every survival model—event status and predicted survival probability. For a given follow-up time, the Brier score is calculated by taking the squared difference between each individual's true survival status and their predicted survival probability, weighted by their probability of censoring—and then averaged across all subjects. The resulting score ranges between 0 and 1, with lower values indicating lower prediction error. The IBS is an omnibus measure of the weighted mean-squared error (MSE) in predictive accuracy of the model across all follow-up times. For the present data, the IBS was estimated using the *riskRegression* package in R[35] truncating the maximum follow-up time at 10 590 days (29 y), because of the sparsity of events occurring beyond that time.

Second, we perform pairwise comparisons between models on estimated survival curves to determine whether they are discriminable at any point along the follow-up continuum. The Kaplan-Meier (KM) product-limit estimator[36] was used to estimate a survivor function for the naïve (nonadjusted) model and each of the three-strata EO-CTA survival model endpoints and both two-strata GO-CTA survival model endpoints. A postestimation survival function was computed following estimation of the Cox regression model with covariates. Stata 14.1 (StataCorp, College Station, Texas) was used to estimate the Cox regression model and generate all survival curves. To be consistent with the 10-year (3650 d) event horizon used for predicting CVD in the Framingham project,[2] we also estimated survival curves for a 10-year follow-up period. Nonweighted GO-CTA was used to perform pairwise comparisons between the naïve model and each of the 6 adjusted survival curves; and between the Cox regression model with covariates and each of the 5 CTA survival model-based curves.

3 | RESULTS

3.1 | Baseline characteristics, follow-up length, split-half samples, and Cox regression results

Table 1 presents baseline characteristics and length of follow-up of study participants, by CVD outcome status and sample (ie, split-half or full), as well as the results of the Cox regression analysis. By the end of the follow-up period (11 688 d), approximately 46% of the study population developed CVD; on average, these were older males with higher baseline blood pressure, serum cholesterol level, and BMI. The SH1 and SH2 samples were comparable on all baseline characteristics, outcomes, and follow-up times, as assessed by CTA: No statistically significant model emerged (all $P > .05$), indicating that SH1 and SH2 could not be discriminated on the basis of these variables. In the Cox regression analysis, all covariates were statistically significant ($P < .001$) except diastolic blood pressure. All models (full, SH1, and SH2) produced similar estimates, further demonstrating comparability of the samples. Postestimation tests revealed that the proportional-hazards assumption was violated for gender. While beyond the scope of the present study, in general, if the proportional-hazards assumption fails, then alternative modelling choices should be considered.[37]

TABLE 1 Baseline characteristics and length of follow-up of study participants, by CVD outcome status and sample (full and by split-half), and Cox regression results

		No CVD			CVD			Cox regression results	
Variable	Sample	N	Mean	SD	N	Mean	SD	Coefficient	SE
SBP, mmHg	SH1	1602	130.2	21.2	725	137.7	23.6	0.011*	0.003
	SH2	1624	129.8	21.4	748	139.9	25.6	0.009*	0.002
	Full	3226	130.0	21.3	1473	138.8	24.7	0.010*	0.002
DBP, mmHg	SH1	1602	81.2	12.2	725	85.4	12.7	0.001	0.005
	SH2	1624	80.9	12.3	748	86.3	13.7	0.008	0.005
	Full	3226	81.0	12.2	1,473	85.8	13.2	0.005	0.003
SCL, mg/100 mL	SH1	1586	222.9	43.0	720	238.8	46.1	0.004*	0.001
	SH2	1614	223.1	41.8	746	240.8	47.8	0.006*	0.001
	Full	3200	223.0	42.4	1466	239.8	46.9	0.005*	0.001
Age, years	SH1	1602	45.4	8.5	725	47.9	8.4	0.040*	0.005
	SH2	1624	45.0	8.3	748	47.9	8.3	0.043*	0.005
	Full	3226	45.2	8.4	1473	47.9	8.4	0.041*	0.003
BMI, kg/m^2	SH1	1601	25.2	4.0	724	26.6	4.2	0.037*	0.009
	SH2	1617	25.2	4.0	748	26.5	4.0	0.030*	0.009
	Full	3218	25.2	4.0	1,472	26.6	4.1	0.033*	0.007
Male, %	SH1	624	39.0		417	57.5		0.779*	0.079
	SH2	602	37.1		406	54.3		0.764*	0.079
	Full	1226	38.0		823	55.9		0.771*	0.055
Follow-up, days	SH1	1602	9012.0	3395	725	5957	3178		
	SH2	1624	9041.0	3362	748	5937	3066		
	Full	3226	9027.0	3378	1473	5947	3121		

Notes: Cox regression values represent coefficients, with bootstrapped standard errors in parentheses. SBP, systolic blood pressure; DBP, diastolic blood pressure; SCL, serum cholesterol level; BMI, body mass index.

*$P < .001$.

TABLE 2 Weighted CTA models discriminating subjects with versus without CVD, by variable (attribute) and split-half

Variable	Split-half	Cutpoint predicting disease status	Weighted sensitivity (CVD)	Weighted specificity (No disease)	WESS (Exact P<)
SBP, mmHg	SH1	≤145	26.6	84.3	10.8 (.001)
	SH2	≤149	24.2	87.2	11.4 (.001)
DBP, mmHg	SH1	≤99	12.4	93.4	5.8 (.001)
	SH2	≤99	15.0	94.6	9.6 (.001)
SCL, mg/100 mL	SH1	≤292	11.4	94.1	5.5 (.001)
	SH2	≤276	16.6	90.9	7.5 (.001)
Age, years	SH1	≤54	21.5	87.3	8.8 (.001)
	SH2	≤58	9.2	95.0	4.2 (.004)
BMI, kg/m^2	SH1	≤30.8	13.2	93.4	6.6 (.001)
	SH2	≤34.1	4.4	98.0	2.4 (.023)
Gender	SH1	Male	54.8	62.1	16.9 (.001)
	SH2	Male	51.6	63.9	15.5 (.001)

Notes: All estimates are weighted by follow-up. For WESS, 0 = weighted ESS expected by chance, 100 = perfect prediction. For every attribute the first row of results is for analysis involving SH1, and the second row of results is for analysis involving SH2. For all models subjects having values less than or equal to the tabled threshold value (computed by the ODA algorithm) are predicted to be from the no disease group (coded as 0), and subjects having values greater than the tabled threshold are predicted to be from the disease group (coded as 1). Exact *P* values are given for all WESS values. SBP, systolic blood pressure; DBP, diastolic blood pressure; SCL, serum cholesterol level; BMI, body mass index.

3.2 | Weighted CTA for attributes evaluated individually (step 1)

Table 2 presents the results of applying weighted CTA to individual attributes to predict subject CVD status, separately by split-half sample. All models yielded relatively weak WESS and were statistically significant (age and body mass index effects for SH2 were only statistically significant if evaluated using the "per-comparison" $P < .05$ criterion[18]).

3.3 | Obtaining all possible weighted CTA models separately for SH1 and SH2 and identifying the identical and parallel models (step 2)

Table 3 presents all of the statistically valid CTA survival models obtained for predicting CVD status, separately for SH1 and SH2. Two models were retained. First, an identical two-strata weighted GO-CTA model using gender as the only attribute (if gender = male then predict disease; if gender = female then predict no disease) was identified as model 8 for SH1 and SH2. This model estimated probability of disease as 0.1787 or 0.1863 for females, and 0.3020 or 0.3019 for males, for SH1 and SH2, respectively. Second, parallel weighted EO-CTA models were obtained as models 6 and 7—in which SH1 and SH2 models were identical except for the systolic blood pressure (SBP) cutpoint value. Model 6 used gender as the root attribute and then SBP, and model 7 used SBP as the root attribute and then gender. These models had identical WESS within complementary split-half samples, but model 7 was selected as the three-strata model because it had the largest minimum strata N, thus providing greatest statistical power.[19] Figure 1 illustrates model 7 for SH1 and SH2 and summarizes hold-out validity classification results obtained by applying the SH1 model to the SH2 sample, and vice versa. Consistent with findings for model 8, predicted outcomes for model 7 are highly consistent between SH1 and SH2.

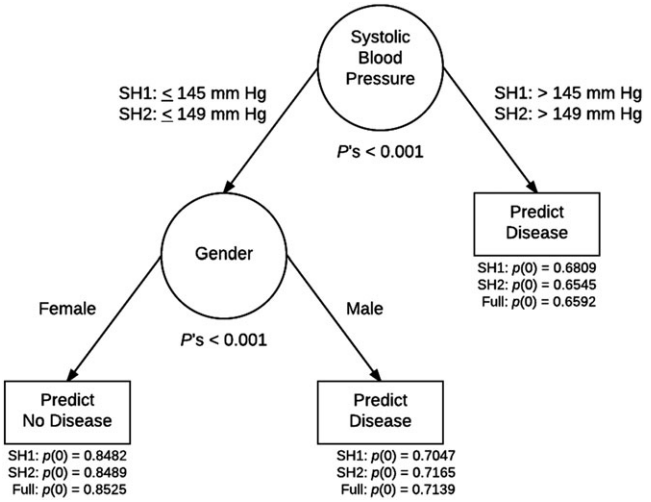

FIGURE 1 The three-strata weighted CTA survival model 7 obtained using SH1 (with probability of disease estimated for the SH2 sample), for SH2 (with probability of disease estimated for the SH1 sample), and for the full sample (with probability of disease estimated for the full sample). CTA, classification tree analysis

3.4 | Evaluating intermodel agreement of outcome predictions made by corresponding SH1 and SH2 survival models (step 3)

Agreement of outcome predictions made by models 7 and 8 was assessed for the full sample. Cross-classifying predicted disease status of subjects using model 8 (which was identical for SH1 and SH2) yielded 2650 subjects similarly classified as having no disease, and 2049 subjects similarly classified as having disease—revealing perfect congruence (ESS = 100). In contrast, model 7 differed for SH1 and SH2 (ie, the SBP threshold value was 145.5 versus 149.5 mmHG, respectively). Cross-classifying predicted disease status of subjects obtained using SH1- and SH2-based model 7 yielded 1962 of 2055 subjects similarly classified as having no disease, and 2644 subjects similarly classified as having disease—indicating near-perfect agreement: ESS = 95.5 for SH1 model and ESS = 96.6 for SH2 model.

3.5 | Sensitivity analysis assessing model validity for increasingly longer follow-ups (step 4)

Sensitivity analysis was conducted to assess stationarity of the predictive accuracy (WESS) for models 7 and 8 for SH1 and SH2, over increasing follow-up periods. In each analysis, the WESS for SH1 was

TABLE 3 All weighted CTA survival models predicting CVD identified for each split-half training sample

SH1						SH2					
Model	Strata	WESS	Efficiency	D	Smallest strata N	Model	Strata	WESS	Efficiency	D	Smallest strata N
1	6	25.38	4.23	17.64	42	1	8	27.03	3.38	21.60	50
2	5	24.28	4.86	15.59	80	2	7	25.59	3.66	20.35	52
3	5	22.95	4.59	16.79	114	3	6	25.13	4.12	17.88	79
4	4	22.73	5.68	13.60	217	4	4	24.74	6.18	12.17	225
5	4	22.65	5.66	13.66	222	5	4	22.95	5.74	13.43	260
6	3	22.57	7.52	10.29	339	6	3	22.77	7.59	10.18	299
7	3	22.57	7.52	10.29	538	7	3	22.77	7.59	10.18	464
8	2	16.91	8.46	9.83	1,041	8	2	15.53	7.76	10.88	1008

Notes: Strata is the number of model endpoints (terminal nodes); WESS measures normed weighted predictive accuracy (0 = weighted accuracy expected by chance; 100 = perfect accuracy); efficiency is WESS divided by strata—a measure of the magnitude of normed accuracy yielded per model endpoint; the distance statistic D indicates the number of additional effects with equivalent WESS that are needed to obtain a theoretically ideal model yielding perfect accuracy and maximum possible parsimony for the application; and smallest strata N is the number of subjects in the endpoint representing the fewest subjects among all endpoints in the model.[19,46]

assessed by applying the SH1 model to classify the SH2 ("hold-out") sample, and vice versa. Model WESS was computed using data from subsamples representing increasing annual follow-up periods—ranging from more than 1 year to more than 28 years (models for longer periods failed statistical power criteria). Table A.1 presents the findings of the sensitivity analysis, which indicates that models 7 and 8 each have stable, comparable hold-out validity WESS between split-half samples across time.

3.6 | Evaluating new (recalibrated) CTA survival models for the SH1 and SH2 samples at follow-up times indicated in sensitivity analysis (step 5)

Three attempted model recalibration analyses were indicated by the sensitivity analysis. The first attempted recalibration occurred at >12 years follow-up—the first time model 8 had a WESS value 10% lower than obtained in initial (all data) analysis. For models 7 and 8, for SH1 and SH2, restricting follow-up to >12 years yielded the identical models 7 and 8. Identical results were also obtained for the second recalibration analysis at >13 years follow-up—the first time WESS for model 7 was 10% lower than the initial value. And identical results were obtained for the final recalibration analysis at >21 years follow-up—the first time WESS for models 7 and 8 was 15% lower than the initial value. In summary, sensitivity and model recalibration analyses confirmed findings consistent with the total sample analysis, over the range of follow-up periods studied.

3.7 | Comparing naïve, Cox regression, and CTA survival model survival curves

Figure 2 illustrates the estimated survival curves for all models derived in the present study. These include the CTA two-strata gender model obtained for SH1 and SH2, the CTA three-strata model obtained for SH1 (using a SBP cutpoint of 145.5 mmHg, which yielded the greatest ESS for SH1 and for the full sample), the naïve (unadjusted) Kaplan-Meier estimate, and the covariate-adjusted Cox regression. As shown, the highest survival rate was predicted by the three-strata CTA model with the rule SBP ≤ 145.5 mmHg and female. Conversely, the lowest survival rate was predicted by the three-strata CTA model with the rule SBP > 145.5 mmHg. Females (from the two-strata CTA model) had the second highest predicted survival, while males (also from the two-strata CTA model) had the second lowest predicted survival rates. The adjusted Cox regression model produced survival rates similar to the naïve (Kaplan-Meier) estimate, and both were positioned in the middle of the range of models. Thus, the Cox model best predicts average (omnibus) incidence of the outcome across follow-up whereas the CTA models best predict either relatively high, or low, incidence of the outcome over time.

Two methods were used to compare the prediction error/accuracy between these survival modelling approaches. First, the IBSs for predicted survival estimates for the 7 models are presented in Table A.2. The naïve survival estimate, which serves as a general benchmark, produced a weighted mean-squared prediction error of 0.122, while the adjusted Cox regression had a slightly lower level of 0.104. Different CTA survival model strata produced varying levels of prediction error. For example, the overall best IBS score (0.069) was achieved by the stratum of the three-strata model in which survival was predicted by SBP ≤ 145.5 mmHg and gender = female. Conversely, in the same model, the stratum in which survival was predicted simply as SBP > 145.5 mmHg produced the highest weighted mean-squared prediction error (0.168). In the two-strata model, predicting survival for females elicited less prediction error than that for males (0.096 vs 0.148, respectively). Overall, the greater the predicted incidence of CVD for a model endpoint (all incidence estimates were <50%), the greater the heterogeneity in class status of subjects in the model endpoint—and therefore the greater the IBS score.

In the second approach used to compare survival estimates, the first of 2 sets of analyses compared naïve (nonadjusted) model (ie, actual) 10-year disease-free survival versus 10-year disease-free survival estimated using each adjusted model (data available from authors). Ten-year disease-free survival predicted by the Cox model (relatively weak ESS = 20.9), and by the leftmost endpoints of the two-strata (moderate ESS = 28.0) and the three-strata (relatively strong ESS = 51.9) CTA survival models, was significantly greater than unadjusted 10-year survival. Conversely, 10-year disease-free survival predicted by the rightmost endpoint of the two-strata CTA survival

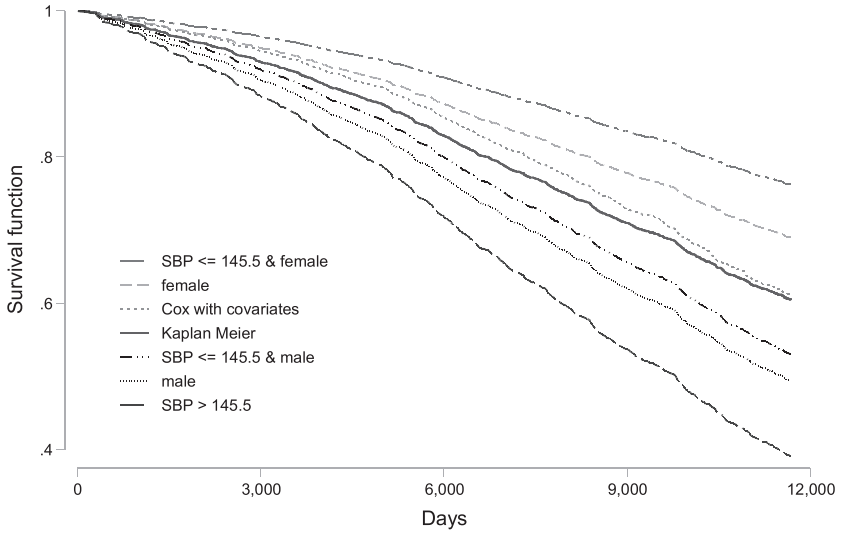

FIGURE 2 Estimated survival curves for all models in the present study. Ordering of models listed in the legend corresponds to ordering of the curves, from top to bottom

model (moderate ESS = 27.4), and by the middle (relatively weak ESS = 14.6) and rightmost (moderate ESS = 38.8) endpoint of the three-strata CTA survival model, was significantly lower than unadjusted 10-year survival.

The second set of comparative analyses in the second approach, examined the predicted 10-year survival for the Cox model versus for the 5 (one for each endpoint) two- and three-strata CTA survival models. 10-year disease-free survival predicted by the leftmost endpoints of the two-strata (relatively weak ESS = 9.3) and three-strata (moderate ESS = 34.0) CTA survival models was significantly greater than estimated 10-year survival by the Cox model. Conversely, 10-year disease-free survival predicted by the rightmost endpoint of the two-strata CTA survival model (moderate ESS = 39.4), and by the middle (moderate ESS = 32.1) and right-most (relatively strong ESS = 56.2) endpoint of the three-strata CTA survival model, was significantly lower than estimated 10-year survival by the Cox model. In summary, the first set of analyses reveal that all six of the adjusted survival models generated survival curves that were significantly different than the unadjusted (naïve) KM curve, when compared over 10 years of follow-up, and the second set of analyses reveal that all five of the CTA-based survival curves were significantly different (and thus more accurate in predicting either positive or negative CVD status) than the Cox regression survival curve, when compared over 10 years of follow-up.

4 | DISCUSSION

Machine learning techniques are increasingly being used in health care research for applications such as improving diagnostic accuracy, identifying high-risk patients, and extracting concepts in unstructured data.[38] In this paper, we introduce CTA as an appealing machine learning alternative for modelling censored data that offers several important advantages over the commonly used Cox regression.

First, investigators using regression-based models have little guidance in their model-building process. For example, some studies estimate models in which the variable selection process includes only main effects, others estimate completely saturated models (including all possible interactions, and squared and cubed terms), and others use automated forward or backward stepwise procedures to select variables for model inclusion. Such heterogeneous approaches to estimation are likely to produce misspecified or suboptimally fit models. Indeed, postestimation tests following Cox regression in the present study indicated that the proportional-hazards assumption was violated for gender. Moreover, the estimated disease-free survival curve derived from the Cox model followed a similar trajectory to that of the unadjusted Kaplan-Meier estimate, suggesting that the more complex Cox model offered little additional ability to predict disease-free survival probability as compared with a simple model. A unique advantage of GO-CTA survival analysis in this regard is that all statistically valid, reproducible, longitudinally consistent, and generalizable CTA models existing for a given sample are identified by an algorithm-driven process, eliminating concerns of model misspecification.[39]

Second, among CTA's most salient features is the generation of simple decision rules to aid both clinicians and researchers to identify subjects exhibiting specific characteristics that place them at higher or lower risk for realizing the outcome. When supplemented with their respective survival curves, such decision rules become even more compelling. For example, in reviewing Figure 1, we see that the stratum with the lowest predicted probability of disease in the full model (0.6592) has the rule SBP > 145.5 mmHg, which coincides with the lowest estimated disease-free survival function presented in Figure 2. Taken together, a clinician is given a simple, maximally accurate rule for identifying individuals with a modifiable risk, augmented by 2 complementary estimates of disease-free survival. Conversely, regression-based survival models offer no such interpretable formulae or visual displays of the final model.

Third, a measure of weighted MSE in predicted survival, IBS scores increased systematically with decreasing disease-free survival curves (Figure 2). We attribute this relationship to variability in the estimated survival curves, which is maximized when predicted probability of survival is 0.50. The stratum with the highest overall survival estimate (SBP ≤ 145.5 mmHg and female) had the lowest variability and MSE, while the stratum with the lowest predicted survival (SBP > 145.5 mmHg) had the highest variability and MSE. Therefore, an investigator can feel confident that parsimonious CTA-based stratum-specific decision rules predicting highest survival rates produce survival predictions with an associated weighted MSE that is lower than that of more complex Cox models. However, we argue that WESS is a more appropriate measure of a model's predictive accuracy specifically because it is insensitive to the variability in the predicted outcome.

Additionally, we found that Cox regression produced an estimated disease-free survival curve that was statistically different than those of all CTA-based model strata, while converging with the Kaplan Meier estimate over a very long follow-up—possibly indicative of regression to the mean effects.[19,40] In contrast, 2 easily discriminated ($P < .001$) CTA models that predicted lower CVD incidence than the Cox model had comparatively flatter trajectories across follow-up, and 3 easily discriminated (P's < .001) CTA models predicting higher CVD incidence than the Cox model had comparatively accelerated trajectories over follow-up (Figure 2). As is clearly seen, the Cox model best predicts average (omnibus) incidence of the outcome across follow-up, whereas CTA survival models best predict either relatively high, or low, incidence of the outcome across follow-up.

Finally, while this paper has focused on the application of CTA to censored data for developing maximally accurate prognostic models, a logical extension of these methods lies in the evaluation of nonrandomized intervention studies with censored outcomes (eg, targeting poor health behaviours that may cause disease or death). Linden and Yarnold[39] introduced a CTA-based approach to generating propensity score weights. Propensity scoring techniques are in a family of methods that explicitly model treatment assignment to estimate treatment effects in nonrandomized studies.[41,42] To estimate treatment effects with censored data using CTA, propensity score weights would be first generated as described in Linden and Yarnold,[39] and then multiplied by follow-up time. The GO-CTA survival analysis would then be conducted as described herein.

The primary limitation of the CTA framework for developing prognostic models with censored data—as is the case with every analytic approach used for this purpose—is that models are generated using only the available data. No matter how sophisticated the algorithm,

important unobservable factors such as unmeasured motivation to change health behaviours may limit the ability of any model to accurately predict the outcome.[43,44] Another general limitation affecting all prognostic modelling approaches is that the predictive values of the model are highly sensitive to the prevalence rate of the observed outcome in that population evaluated.[45] More specifically, in a population where nearly everyone is disease-free, it would be much easier to predict a person's probability of being disease-free, and much harder to predict who will develop the disease.

5 | CONCLUSION

In summary, this paper introduced a novel machine learning framework for modelling censored data. This framework offers many advantages over broadly used Cox regression, such as explicit maximization of accuracy, parsimony, sensitivity, statistical robustness, and transparency. Therefore, researchers interested in accurate prognoses and clear decision rules should consider developing models using the CTA-survival framework.

ACKNOWLEDGEMENTS

We wish to thank Tanima Banerjee for her analytic assistance in R and Julia Adler-Milstein for her review and feedback on the manuscript.

REFERENCES

1. Linden A, Schweitzer SO. Applying survival analysis to health risk assessment data to predict time to first hospitalization. *AHSRHP Annual Meeting*. 2001;18:26.
2. D'Agostino RB, Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation*. 2008;117:743-753.
3. Biuso TJ, Butterworth S, Linden A. Targeting prediabetes with lifestyle, clinical and behavioral management interventions. *Disease Management*. 2007;10(1):6-15.
4. Linden A, Adams J, Roberts N. Evaluating disease management program effectiveness: An introduction to survival analysis. *Disease Management*. 2004;7:180-190.
5. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*. 1996;15:361-387.
6. Gordon L, Olshen R. Tree-structured survival analysis. *Cancer Treatment Reports*. 1985;69:1065-1068.
7. Brown SF, Branford AJ, Moran W. On the use of artificial neural networks for the analysis of survival data. *IEEE Transactions on Neural Networks*. 1997;8:1071-1077.
8. Kattan MW, Hess KR, Beck JR. Experiments to determine whether recursive partitioning (CART) or an artificial neural network overcomes theoretical limitations of Cox proportional hazards regression. *Computers and Biomedical Research*. 1998;31:363-373.
9. Evers L, Messow CM. Sparse Kernel Methods for High-dimensional Survival Data. *Bioinformatics*. 2008;24:1632-1638.
10. Khan FM, Zubek VB. Support vector regression for censored data (SVRc): a novel tool for survival analysis. *Eighth International Conference on Data Mining*. 2008;863-868.
11. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Statistical Science*. 2001;16:199-231.
12. Cox DR. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society: Series B*. 1972;34:187-220.
13. Yarnold PR, Soltysik RC. Theoretical distributions of optima for univariate discrimination of random data. *Decision Sciences*. 1991;22:739-752.
14. Linden A, Yarnold PR. Combining machine learning and propensity score weighting to estimate causal effects in multivalued treatments. *Journal of Evaluation in Clinical Practice*. 2016a;22:875-885.
15. Linden A, Adams J, Roberts N. The generalizability of disease management program results: getting from here to there. *Managed Care Interface*. 2004;17(7):38-45.
16. Linden A, Yarnold PR. Using data mining techniques to characterize participation in observational studies. *Journal of Evaluation in Clinical Practice*. 2016b;22:839-847.
17. Linden A, Yarnold PR. Using machine learning to assess covariate balance in matching studies. *Journal of Evaluation in Clinical Practice*. 2016c;22:848-854.
18. Yarnold PR, Soltysik RC. *Optimal Data Analysis: Guidebook with Software for Windows*. Washington, D.C.: APA Books; 2005.
19. Yarnold PR, Soltysik RC. *Maximizing Predictive Accuracy*. Chicago, IL: ODA Books, 2016. https://doi.org/10.13140/RG.2.1.1368.3286
20. Yarnold PR. Discriminating geriatric and non-geriatric patients using functional status information: An example of classification tree analysis via UniODA. *Educational and Psychological Measurement*. 1996;56:656-667.
21. Yarnold PR, Soltysik RC, Bennett CL. Predicting in-hospital mortality of patients with AIDS-related *Pneumocystis carinii* pneumonia: An example of hierarchically optimal classification tree analysis. *Statistics in Medicine*. 1997;16:1451-1463.
22. Soltysik RC, Yarnold PR. Automated CTA software: Fundamental concepts and control commands. *Optimal Data Analysis*. 2010;1:144-160.
23. Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *The Lancet*. 2014;383:999-1008.
24. Dupont WD. *Statistical Modeling for Biomedical Researchers*. Cambridge, U.K.: Cambridge University Press; 2009.
25. Linden A, Adams J, Roberts N. Evaluating disease management program effectiveness: An introduction to the bootstrap technique. *Disease Management and Health Outcomes*. 2005;13:159-167.
26. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*. 1994;81:515-526.
27. Yarnold PR, Linden A. Novometric analysis with ordered class variables: The optimal alternative to linear regression analysis. *Optimal Data Analysis*. 2016;22:65-73.
28. Linden A, Adams J, Roberts N. Strengthening the case for disease management effectiveness: unhiding the hidden bias. *Journal of Evaluation in Clinical Practice*. 2006;12:140-147.
29. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *Journal of the American Medical Association*. 1982;247:2543-2546.
30. Gönen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*. 2005;92:965-970.
31. Grønnesby JK, Borgan Ø. A method for checking regression models in survival analysis based on the risk score. *Lifetime Data Analysis*. 1996;2:315-328.
32. May S, Hosmer DW. A simplified method of calculating an overall goodness-of-fit test for the Cox proportional hazards model. *Lifetime Data Analysis*. 1998;4:109-120.
33. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Statistics in Medicine*. 2004;23:723-748.
34. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*. 1999;18:2529-2545.
35. Gerds TA, Scheike TH, Blanche P, Ozenne B. (2017). riskRegression: Risk Regression Models and Prediction Scores for Survival Analysis with Competing Risks. R package version 1.3.7. https://cran.r-project.org/web/packages/riskRegression/index.html [downloaded on April 3, 2017].

36. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of American Statistical Association*. 1958;53:457-481.

37. Cleves M, Gould W, Marchenko Y. An Introduction to Survival Analysis Using Stata (revised 3rd edition). College Station, TX: Stata Press; 2016.

38. Iavindrasana J, Cohen G, Depeursinge A, Müller H, Meyer R, Geissbuhler A. Clinical data mining: a review. In *IMIA Yearbook of Medical Informatics*. (eds A. Geissbuhler, C. Kulikowski). 2009;48(Suppl 1):121–133.

39. Linden A, Yarnold PR. Using classification tree analysis to generate propensity score weights. *Journal of Evaluation in Clinical Practice*. https://doi.org/10.1111/jep.12744

40. Linden A. Estimating the effect of regression to the mean in health management programs. *Disease Management and Health Outcomes*. 2007;15(1):7-12.

41. Linden A, Adams JL. Using propensity score-based weighting in the evaluation of health management programme effectiveness. *Journal of Evaluation in Clinical Practice*. 2010;16:175-179.

42. Linden A, Adams J. Evaluating disease management program effectiveness: an introduction to instrumental variables. *Journal of Evaluation in Clinical Practice*. 2006;12:148-154.

43. Linden A, Roberts N. Disease management interventions: What's in the black box? *Disease Management*. 2004;7:275-291.

44. Linden A, Butterworth S, Roberts N. Disease management interventions II: What else is in the black box? *Disease Management*. 2006;9:73-85.

45. Altman DG, Bland M. Diagnostic tests 2: predictive values. *British Medical Journal*. 1994;309:102.

46. Yarnold PR, Linden A. Theoretical aspects of the D statistic. *Optimal Data Analysis*. 2016;5:171-174.

How to cite this article: Linden A, Yarnold PR. Modeling time-to-event (survival) data using classification tree analysis. *J Eval Clin Pract*. 2017;23:1299–1308. https://doi.org/10.1111/jep.12779

APPENDIX A

Table A1. Sensitivity analysis applying CTA training models 7 and 8 to independent test (hold-out validity) samples

Training Sample	Model 7, SH1 Model 7, SH2			Model 7, SH2 Model 7, SH1			Model 8, SH1 Model 8, SH2			Model 8, SH2 Model 8, SH1		
Test Sample Minimum Follow-up	Wtd Sensitivity	Wtd Specificity	WESS	Wtd Sensitivity	Wtd Specificity	WESS	Wtd Sensitivity	Wtd Specificity	WESS	Wtd Sensitivity	Wtd Specificity	WESS
Full sample, year	69.1	52.5	21.6	68.4	52.7	21.2	51.6	63.9	15.5	54.8	62.1	16.9
>1	69.1	52.5	21.6	68.4	52.7	21.2	51.6	63.9	15.5	54.8	62.1	16.9
>2	69.1	52.5	21.6	68.4	52.8	21.2	51.6	63.9	15.5	54.8	62.1	16.9
>3	69.0	52.5	21.5	68.3	52.8	21.1	51.5	63.9	15.4	54.7	62.2	16.9
>4	68.9	52.5	21.4	68.2	52.8	21.0	51.5	63.9	15.4	54.6	62.1	16.7
>5	68.9	52.5	21.4	68.0	52.8	20.8	51.4	63.9	15.3	54.6	62.2	16.8
>6	68.7	52.6	21.3	67.8	52.9	20.6	51.3	63.9	15.2	54.5	62.2	16.7
>7	68.6	52.7	21.3	67.6	52.9	20.5	51.2	64.0	15.2	54.4	62.2	16.6
>8	68.4	52.7	21.1	67.4	53.0	20.3	51.1	64.0	15.1	54.1	62.2	16.3
>9	68.2	52.7	20.9	67.1	53.1	20.3	51.0	64.0	15.0	53.8	62.3	16.1
>10	68.1	52.8	20.9	66.8	53.2	19.9	51.0	64.1	15.1	53.7	62.3	15.9
>11	67.8	52.8	20.6	66.2	53.1	19.4	51.3	64.1	15.3	53.3	62.2	15.5
>12	67.6	53.0	20.6	65.8	53.3	19.1	51.1	64.1	15.1	52.8	62.2	15.0
>13	67.5	53.1	20.6	65.6	53.4	19.0	51.1	64.1	15.2	52.6	62.3	14.8
>14	66.9	53.1	20.0	65.2	53.5	18.7	50.4	64.1	14.5	53.0	62.3	15.3
>15	66.7	53.1	19.8	65.4	53.7	19.1	50.6	64.1	14.7	53.2	62.3	15.5
>16	66.3	53.4	19.7	64.9	54.0	19.0	50.9	64.1	15.0	53.4	62.5	15.8
>17	65.6	53.6	19.2	64.8	54.0	18.8	50.6	64.2	14.7	53.2	62.5	15.8
>18	65.0	53.9	18.9	64.3	54.3	18.6	50.1	64.3	14.4	53.3	62.6	15.8
>19	65.4	54.0	19.4	64.4	54.3	18.7	50.6	64.4	14.9	53.6	62.5	16.1
>20	64.7	54.1	18.8	63.8	54.6	18.4	49.2	64.3	13.5	52.6	62.7	15.2
>21	63.2	54.2	17.3	63.4	54.9	18.3	47.3	64.4	11.8	52.6	62.8	15.4
>22	62.6	54.4	17.0	64.5	55.1	19.6	46.2	64.6	10.8	53.0	63.1	16.1
>23	62.3	54.9	17.2	65.0	55.5	20.5	46.1	65.0	11.0	53.4	63.1	16.6
>24	65.0	55.3	20.3	64.9	55.7	20.6	47.3	65.4	12.7	53.3	63.2	16.5
>25	65.6	55.5	20.9	65.6	55.9	21.6	48.9	65.5	14.3	54.1	63.1	17.2
>26	65.1	55.9	21.0	64.1	56.5	20.6	53.3	63.3	16.5	48.7	65.7	14.4
>27	68.7	56.4	25.1	62.6	57.0	19.6	53.0	65.7	18.7	51.8	63.2	15.1
>28	69.6	56.4	25.9	63.4	57.3	20.7	55.0	65.6	20.7	51.8	63.4	15.1

Notes: Tabled values were obtained by applying the indicated training model to the indicated validity sample. Minimum follow-up values >29 y produced samples with an insufficient number of class = 1 (positive CVD outcome) subjects to satisfy the minimum statistical power criterion.[19]

Table A2. Integrated Brier Scores (IBS) for each model in the current study

Model	IBS
Naïve Kaplan Meir (w/o covariates)	0.122
Cox regression with covariates	0.104
Model 7a: SBP ≤ 145.5 and gender = female	0.069
Model 7b: SBP ≤ 145.5 and gender = male	0.135
Model 7c: SBP ≥ 145.5	0.168
Model 8a: Gender = Female	0.096
Model 8b: Gender = Male	0.148