Received: 8 March 2023 | Revised: 14 October 2023 | Accepted: 23 October 2023

DOI: 10.1002/sim.9954

RESEARCH ARTICLE

Statistics in Medicine WILEY

A flexible model based on piecewise linear approximation for the analysis of left truncated right censored data with covariates, and applications to Worcester Heart Attack Study data and Channing House data

Ayon Ganguly[1] | **Debanjan Mitra**[2] | **Narayanaswamy Balakrishnan**[3] | **Debasis Kundu**[4]

[1]Department of Mathematics, Indian Institute of Technology Guwahati, Guwahati, Assam, India

[2]Quantitative Methods Division, Indian Institute of Management Udaipur, Udaipur, Rajasthan, India

[3]Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada

[4]Department of Mathematics and Statistics, Indian Institute of Technology Kanpur, Kanpur, Uttar Pradesh, India

Correspondence
Debanjan Mitra, Quantitative Methods Division, Indian Institute of Management Udaipur, Udaipur, Rajasthan, India.
Email: debanjan.mitra@iimu.ac.in

Funding information
Natural Sciences and Engineering Research Council of Canada; Science and Engineering Research Board, Grant/Award Numbers: MTR/2017/000700, MTR/2021/000533

Left truncated right censored (LTRC) data arise quite commonly from survival studies. In this article, a model based on piecewise linear approximation is proposed for the analysis of LTRC data with covariates. Specifically, the model involves a piecewise linear approximation for the cumulative baseline hazard function of the proportional hazards model. The principal advantage of the proposed model is that it does not depend on restrictive parametric assumptions while being flexible and data-driven. Likelihood inference for the model is developed. Through detailed simulation studies, the robustness property of the model is studied by fitting it to LTRC data generated from different processes covering a wide range of lifetime distributions. A sensitivity analysis is also carried out by fitting the model to LTRC data generated from a process with a piecewise constant baseline hazard. It is observed that the performance of the model is quite satisfactory in all those cases. Analyses of two real LTRC datasets by using the model are provided as illustrative examples. Applications of the model in some practical prediction issues are discussed. In summary, the proposed model provides a comprehensive and flexible approach to model a general structure for LTRC lifetime data.

KEYWORDS

cumulative hazard function, left truncation, maximum likelihood estimators, piecewise linear approximation, prediction, proportional hazards model, right censoring

1 | INTRODUCTION

Left truncated right censored (LTRC) data arise quite commonly from survival experiments. Lifetime data, when collected under practical time constraints, often lead to LTRC structure. For example, LTRC data may be obtained from a survival experiment with an observation window (ie, a window for data collection) between two fixed time points, where the failure event of interest (eg, death of a subject) occurring outside the observation window cannot be observed. Left truncation in the data occurs when subjects are included in the study only after they survive beyond a threshold time, often obtained as the difference between the beginning of the study (or the enrollment of the subjects) and the initiation of the lifetimes of the subjects (eg, time of birth, or the starting point of the incubation time of a disease). Naturally, in left truncated data, subjects with shorter survival times are excluded. Right censoring is incorporated into the data when some subjects

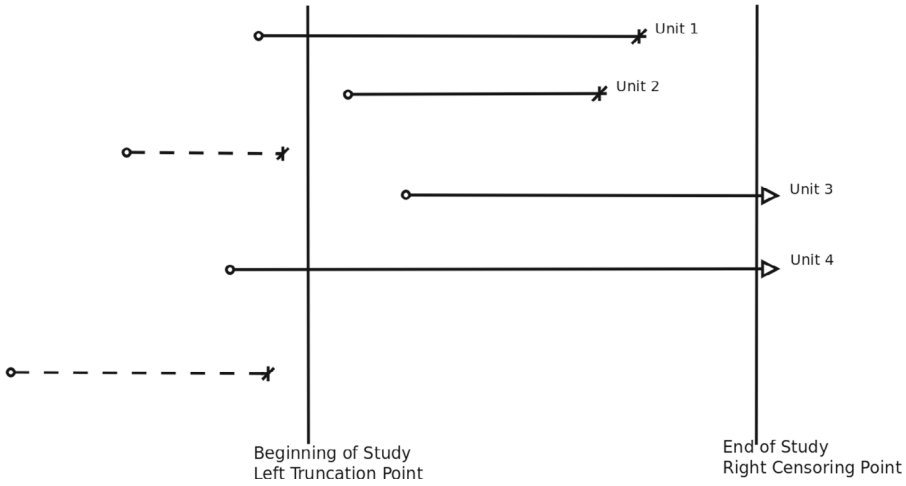

FIGURE 1 All types of units in LTRC data, with different combinations of left truncation and right censoring; the units depicted by dashed lines are not observed at all.

do not experience the failure event at the termination point of the study, implying that they would experience the event sometime after data collection is completed.

A pictorial description of LTRC data is given in Figure 1. The beginning of the study essentially serves as the left truncation point, because occurrence of any failure event before this point would not be known. The right censoring point corresponds to the termination point the study; failures of the right censored subjects are also not observable, but at least partial information is known that their actual lifetimes are greater than their lifetimes observed up to the right censoring point. In Figure 1, to enter the observation window, Subjects 1 and 4 had to survive beyond a threshold time which we refer to as their left truncation times. These subjects are said to be left truncated units. On the other hand, there is no such threshold times for Subjects 2 and 3 to enter the observation window; they are referred to as untruncated units. Subjects 3 and 4 are right censored as they have not experienced the failure event by the end of the study. This structure of LTRC data is quite general in accommodating subjects with different truncation and censoring status; in this structure, a subject can be left truncated and right censored, left truncated and uncensored, untruncated and right censored, and untruncated and uncensored. That is, this setting covers all combinations of left truncation and right censoring with respect to an observation window.

Statistical inference based on LTRC data must be carried out by properly accounting for the incompleteness in the data to avoid any systematic bias in the results. Klein and Moeschberger[1] gave details of some of the methods that are used to analyze LTRC data. Among the more recent works, Hong et al[2] presented a detailed study of parametric analyses of LTRC data by using the log-location-scale family of distributions. Using the data structure of Hong et al, Balakrishnan and Mitra[3] developed the steps of the EM algorithm for different members of the generalized gamma family of distributions; see also Mitra et al[4] in this regard wherein the stochastic EM algorithm was used to model LTRC data from the Lehmann family of distributions. Emura and Michimae[5] conducted a comprehensive review of existing parametric models and methods for LTRC lifetime data. Specifically in health research, among others, a recent study by Pak et al[6] analysed infectious disease data with LTRC structure and interval-censored infection onset times. Among the other recent works on LTRC data, we refer to the following: Chen and Yi[7] for the PH model with covariates containing measurement error; Chen and Yi[8] for model selection with covariates containing measurement error; Chen[9] for a study of the additive hazards model; Huang and Qin[10] for semiparametric inference for the additive hazards model; Huang et al[11] for semiparametric inference; and Su and Wang[12] for the PH model with longitudinal covariates.

In many practical situations, lifetimes of units depend on some explanatory variables or covariates that are observed simultaneously. The proportional hazards (PH) model of Cox[13,14] provides a convenient way to incorporate covariates into the underlying distribution of the lifetime variable. In this approach, the hazard function of the underlying lifetime variable in the presence of a covariate vector $\boldsymbol{x}$ is modeled as

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{\beta}$ is the vector of regression parameters and $\lambda_0(\cdot)$ is the baseline hazard function independent of covariates. The regression parameters of the PH model can be estimated by maximizing a partial likelihood in $\boldsymbol{\beta}$,[13,14] while the baseline hazard $\lambda_0(\cdot)$ can be estimated nonparametrically. This approach results in a semiparametric estimation for the PH model. The PH model has been used for LTRC data; see van Houwelingen and Stijnen.[15] Further, Zhang et al[16] analysed LTRC competing risks data with a PH model for the subdistribution. In this context of regression models for LTRC survival data, the work of Cortese et al[17] that addressed regression for the restricted residual mean life for LTRC data should also be mentioned.

In the present work, we propose an approach to model LTRC data by using a PH model with a functional approximation for the baseline hazard. We specifically approximate the cumulative baseline hazard of the PH model by a piecewise linear function. The proposed model is simple, yet powerful. It is simple in the sense that it does not depend on restrictive parametric assumptions and is also computationally convenient. Its power lies in the fact that it is data-driven and is quite flexible in modeling lifetime data with baseline hazard of different forms. In particular, while most parametric lifetime models can accommodate only a monotone baseline hazard function, the model proposed here can handle a non-monotone baseline hazard. In this way, the proposed model will be suitable for a wide variety of lifetime data involving a general structure of left truncation and right censoring. It may be mentioned here that although a piecewise constant model for hazard rate was discussed by Friedman,[18] that model did not focus on the choice of the cut-points for constructing intervals over which the piecewise constant function was defined, thus assuming the cut-points to be fixed at certain positions over the observed data range. For example, for implementing the `R` function `phreg` of the `R`-package `eha`,[19] the user needs to supply the cut-points for defining the intervals for the piecewise constant hazard rate. However, for applications in real-life, it would be much more practical if the cut-points for the intervals can be chosen, with the help of some guiding criteria, based on the observed data. Appropriate choice of the cut-points would ensure a truly flexible model and a data-driven approach for the analysis. Along with modeling LTRC data by a PH model involving a piecewise linear function, in this article, we also focus on choosing the cut-points suitably to make the model flexible and data-driven. Thus, a practical guidance for appropriately positioning the linear pieces over different ranges of the observed ranges of the LTRC data is also provided as it would facilitate wide application of the proposed model.

It is of interest to mention here that a piecewise linear approximation (PLA) for the hazard function has been used recently by Balakrishnan et al[20] in the context of cure-rate models. We observe that within the framework of the PH model by using a PLA for the cumulative baseline hazard, it is possible to obtain an estimate of the underlying survival function that is quite close to the estimate obtained by using the well-known Breslow estimator.[21] Moreover, the PLA-based approach has the advantage over Breslow estimator for making predictions of future failures. Being a fully nonparametric procedure, Breslow estimator is constructed over the observed range of data, and so cannot be used for prediction of failure events that occur outside the observed data range. However, the proposed PLA model, being effectively a parametric approach, can be extended beyond the observed data range for making predictions. Thus, the proposed approach turns out to be data-driven and flexible with a wider scope, and is also computationally convenient.

Therefore, in summary, the main contributions of the proposed PLA-based model are as follows:

- The model does not depend on strong parametric restrictions and assumptions. It is developed under minimal assumptions, for example, that of continuity of the cumulative hazard at the cut-points, which is quite general;
- The model is fairly data-driven as the cut-points can be chosen suitably based on the observed data. It is possible to provide simple practical guidance on choosing the cut-points;
- The model can accommodate both monotone and non-monotone hazard rates of the underlying lifetime. Additionally, the model has less number of parameters compared to fully parametric models that can accommodate such wide range of hazards rates;
- The model can be used for prediction purposes. There are nonparametric approaches, like Breslow estimator, that do not depend on parametric assumptions for estimating baseline cumulative hazard. But they cannot be used for prediction purposes as they cannot be extrapolated beyond the observed data range due to their nonparametric nature.

The rest of the article is organized as follows. In Section 2, the structure of LTRC data is described. Two real LTRC datasets are introduced as motivating examples in Section 3. The proposed PH model involving a PLA for the cumulative baseline hazard, along with a practical guidance to choose the cut-points for the PLA-based model, are presented in Section 4. The associated likelihood inference for the model based on LTRC data are also discussed in this section. Results of a detailed simulation study that evaluates the performance of the proposed model are presented in Section 5.

In Section 6, two illustrative examples based on the real LTRC datasets are presented. Some important prediction issues based on the proposed model are discussed in Section 7. Finally, some concluding comments are made in Section 8.

2 | STRUCTURE OF LTRC DATA

Consider a survival experiment with ψ_L and ψ_R, $\psi_L < \psi_R$, as the starting and termination points of data collection, respectively. Then, the interval (ψ_L, ψ_R) constitutes the observation window; failure of a subject in the study is observable only if it occurs within (ψ_L, ψ_R). Subjects that have to survive beyond a threshold time to enter the observation window are the left truncated units; otherwise, they are the untruncated units. Let ψ_B denote the starting point of a subject, and $\kappa = \psi_L - \psi_B$. The left truncation time τ_L is defined as

$$\tau_L = \begin{cases} \kappa, & \text{if} \quad \kappa > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Of course, different subjects may start at different time points, and hence will have non-homogenous values of τ_L. Define a truncation indicator ν that takes the value 1 if a unit is truncated, and 0 otherwise.

The study subjects that experience the failure event before ψ_R are the observed failures, otherwise they are the right censored units. The right censoring time of a subject is defined by $\tau_R = \psi_R - \psi_B$. If the underlying lifetime variable is denoted by T, then in the presence of left truncation and right censoring, the observed lifetime for a subject will be

$$Y = Min(T, \tau_R), \quad T > \tau_L.$$

We introduce a censoring indicator, δ, that takes the value 1 if a subject experiences the failure event and 0 if the subject is right censored. Summarizing the above discussion, the available LTRC data is of the form $(Y_i, \tau_{Li}, \delta_i, \nu_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, where $\boldsymbol{x}_i$ denotes the vector of covariates associated with the ith subject in the survival experiment.

3 | MOTIVATING DATA SETS

3.1 | Worcester Heart Attack Study data

The Worcester Heart Attack Study (WHAS) dataset, containing information on 500 patients and called WHAS500 in short, was reported in Hosmer et al.[22] The primary goal of this study was to investigate factors and time patterns related to long-term survival of subjects who were admitted to various hospitals in Worcester, Massachusetts Standard Metropolitan Statistical Area, following acute myocardial infarction (MI). The study began in 1975 and collected data approximately every other year till 2001 for residents of Worcester experiencing MI. While the original data obtained from the study was for more than 11 000 individuals, the WHAS500 dataset was constructed with random samples of subjects from the cohort years 1997, 1999, and 2001. For more details about the dataset, see Hosmer et al.[22]

The dataset WHAS500 has patient-level information on many covariates including age, gender, systolic and diastolic blood pressure, heart rate (HR), body mass index (BMI), history of cardiovascular disease, type of MI, and so on. Medical investigations based on this data can involve study of the effects of these covariates over survival rate of subjects. Three different dates, in the MM/DD/YYYY format, are recorded in the dataset for each patient: date of admission to hospital, date of discharge from hospital, and date of last follow-up. At the date of the last follow-up, a subject may be found to be dead (indicator = 1), or alive (indicator = 0). The lifetime variables that can be constructed using these information are as follows: (a) Total length of follow-up, which is the number of days between the date of last follow-up and the date of admission to hospital, and (b) Length of hospital stay, which is the number of days between the date of discharge from hospital and the date of admission to hospital.

Very recently, Chen and Yi[7] considered the WHAS500 dataset. Out of the 500 subjects in the dataset, they chose only those subjects who were discharged alive from the hospitals. That is, for each of these patients, the total length of follow-up was greater than the corresponding length of hospital stay. This incorporated left truncation in the resulting dataset, as subjects who died before being discharged from hospitals never featured in the sample. Also, there was right censoring in the sample: several subjects were alive at the time of the last follow-up. For this dataset, a question of interest could be

as to how the covariates affect the long-term survival rates of subjects who were discharged alive from the hospitals. For more details on the dataset, we refer to Hosmer et al.[22]

3.2 | Channing House data

The Channing House dataset[23] contains lifetime information pertaining to the residents of a retirement center, called the Channing House, in Palo Alto, California. The data collection started from the time of establishment of the center in 1964, and ended at its closing in 1975. During this time, a total of 462 individuals, of which 365 were women and 97 were men, were admitted to the center.

For each individual, the age at the time of entering the center, and the age of death or leaving the center were recorded. During 1964 to 1975, 130 women and 46 men died at the center. A minimum age of 60 was a condition for admittance to the center, and individuals entered the center at different ages after 60. As the data collection started in 1964, no information on lifetimes of individuals were available before this point. Also, information on lifetimes were available only for those individuals who entered the center at some point after 1964. This incorporated left truncation in the Channing House data. In 1975, when the data collection ended, a total of 286 residents were still alive. Thus, these subjects were not observed failures (ie, deaths), but were right censored. For this study, the years 1964 and 1975 serve as the left truncation and the right censoring points, respectively.

Apart from the lifetimes of individuals in the center, the dataset also contains information on gender of the individuals. This information can be used as a covariate for the lifetimes in this dataset, and one may be interested in assessing the survival rates of male and female residents admitted to Channing House. We refer to Klein and Moeschberger[1] for more details on this data.

4 | PLA-BASED PH MODEL AND INFERENCE

4.1 | PH model involving a PLA

Using the cumulative hazard function $\Lambda(t|\boldsymbol{x})$, defined as $\Lambda(t|\boldsymbol{x}) = \int_0^t \lambda(u|\boldsymbol{x})du$, the PH model can be written as

$$\Lambda(t|\boldsymbol{x}) = \Lambda_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{x}), \tag{2}$$

where $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ is the cumulative baseline hazard function. Then, the model proposed is as follows.

Suppose the lifetimes, including observed failures and right censored subjects, are $y_1, \ldots, y_n$. Let $\xi = \{t_0^*, \ldots, t_N^*\}$ denote a set of $N+1$ cut-points on the time scale such that $t_0^* < t_1^* < \cdots < t_N^*$, with $t_0^* < \text{Min}(y_{\min}, \tau_{L\min})$ and $t_N^* \le y_{\max}$, where $y_{\max} = \text{Max}(y_1, \ldots, y_n)$, $y_{\min} = \text{Min}(y_1, \ldots, y_n)$ and $\tau_{L\min} = \text{Min}(\tau_{L1}, \ldots, \tau_{Ln})$. We consider ξ to be fixed and known at this stage. Later on, we will discuss how we could suitably choose ξ. The PLA-based model for the cumulative baseline hazard involves approximation of $\Lambda_0(t)$ by $H_0(t)$, which is a linear function of time between any two consecutive cut-points of the set ξ, with additional general conditions that $H_0(t)$ is monotone increasing and continuous. That is,

$$H_0(t) = \sum_{k=1}^{N}(a_k + b_k t)I(t_{k-1}^* \le t < t_k^*), \tag{3}$$

where

$$I(t_{k-1}^* \le t < t_k^*) = \begin{cases} 1, & \text{if} \quad t_{k-1}^* \le t < t_k^* \\ 0, & \text{otherwise,} \end{cases}$$

for $k = 1, \ldots, N$, with $b_k > 0$, for all values of k. We also assume that

$$H_0(t_0^*) = 0, \tag{4}$$

and

$$\lim_{\epsilon \to 0+} H_0(t_k^* - \epsilon) = H_0(t_k^*), \quad k = 1, \ldots, N. \tag{5}$$

Note that (5) implies continuity of $H_0(t)$ over the entire range. Clearly, continuity and $b_k > 0$ for all values of k imply that $H_0(t)$ is monotone increasing. Finally, the PLA $H_0(t)$ in (3) is used in the PH model in (2) to obtain the proposed model

$$H(t|\boldsymbol{x}) = H_0(t) \exp(\boldsymbol{\beta}'\boldsymbol{x}). \tag{6}$$

Hence, the main idea here is to approximate the cumulative baseline hazard of the PH model by N linear pieces. This approach, as we shall see in the subsequent sections, is data-driven and is easy to implement.

It may be noted here that the above PLA of the cumulative baseline hazard is equivalent to estimating the baseline hazard rate by constants over ranges specified by the cut-points. That is, the above model can equivalently be expressed as

$$h_0(t) = \sum_{k=1}^{N} b_k I(t_{k-1}^* \le t < t_k^*), \tag{7}$$

where $h_0(\cdot)$ is the hazard rate corresponding to the cumulative hazard $H_0(\cdot)$. Therefore, the PLA model tantamounts to approximating the underlying lifetime distribution by several exponential models (with different rate parameters) over the ranges specified by the cut-points.

Further, it should be noted that it is sufficient to estimate the slopes b_i's for the linear pieces, and then the PLA model can be written in the form of (7). To express the model in the form of (3), the intercepts a_i's can be obtained by using the condition of continuity given in (5) as follows:

$$a_k = \sum_{i=1}^{k-1} b_i (t_i^* - t_{i-1}^*) - b_k t_{k-1}^*, \quad k = 2, \ldots, N, \tag{8}$$

and $a_1 = -b_1 t_0^*$.

4.1.1 | Choosing the cut-points

The cut-points may be suitably chosen for a given data. Choosing a large number of cut-points amounts to approximating the cumulative baseline hazard by linear pieces within a large number of small intervals; clearly, this will provide a close local approximation, and is naturally expected to significantly reduce bias in the resulting estimates of parameters and inference. However, in addition to being computationally intensive due to a large number of cut-points involved, it is also expected to increase variability of the estimates due to the classical bias-variance trade-off. So, the number of cut-points should be chosen carefully. A convenient way to choose the number of cut-points and their positions is by looking at the plot of Breslow estimator of the cumulative baseline hazard. Obviously, areas of the lifetime data wherein the Breslow estimator changes significantly are natural choices for the cut-points. In case this is difficult to implement in practice, especially while evaluating the efficiency of the proposed method through a simulation study, a simpler way would be to use some suitable quantiles of the observed data as the cut-points.

For choosing the number of cut-points and their positions placed at different quantiles of the observed LTRC data, the following is an effective method. Apart from t_0^* which is always taken as the first cut-point, the quantiles $q_{0.2}$, $q_{0.4}$, $q_{0.6}$, and $q_{0.8}$ are taken as candidate positions for placing the cut-points, where q_α represents the αth quantile of the data. Then, for an observed LTRC data, PLA-based models for all possible combinations of various number and position of the cut-points, varying them over the four quantiles $q_{0.2}$, $q_{0.4}$, $q_{0.6}$, and $q_{0.8}$, are fitted. This amounts to fitting a total of 15 PLA-based models to a LTRC data, as $\binom{4}{1} + \binom{4}{2} + \binom{4}{3} + \binom{4}{4} = 15$. Among these 15 PLA-based models, the one with the lowest value of Akaike's Information Criterion (AIC)[24] is chosen as the best model for the LTRC data.

For the PLA-based modeling approach, choosing cut-points near the two ends of the observed data should be avoided in general, as there may not be enough datapoints near the ends to provide a good fit for the PLA. Therefore, for choosing cut-points by the above method, choice of quantiles near the two ends is not recommended. One can increase the number

of candidate quantiles, which effectively increase the number of candidate PLA-based models for a given LTRC data. However, in a sensitivity analysis study presented in a later section, we show that considering four candidate quantiles ($q_{0.2}$, $q_{0.4}$, $q_{0.6}$, and $q_{0.8}$) for the cut-points results in a satisfactory final model, and increasing the number of candidate quantiles may not be beneficial.

4.2 | Likelihood inference

Let $h_0(t)$ denote the baseline hazard rate corresponding to the PLA $H_0(t)$ of the cumulative baseline hazard $\Lambda_0(t)$. Further, let $S(t|\boldsymbol{x})$ denote the survival function of the underlying lifetime variable; corresponding to the PLA $H_0(t)$, then the survival function can be written as

$$S_{\text{PLA}}(t|\boldsymbol{x}) = \exp(-H_0(t)e^{\boldsymbol{\beta}'\boldsymbol{x}}).$$

Let $f_{\text{PLA}}(t|\boldsymbol{x})$ denote the PDF corresponding to the survival function $S_{\text{PLA}}(t|\boldsymbol{x})$.

The contribution to the likelihood by different subjects according to their left truncation and right censoring status are as follows:

(i) $\delta = 1, \nu = 0$:

$$f_{\text{PLA}}(t|\boldsymbol{x}) = h_0(t)e^{\boldsymbol{\beta}'\boldsymbol{x}} \exp(-H_0(t)e^{\boldsymbol{\beta}'\boldsymbol{x}});$$

(ii) $\delta = 0, \nu = 0$:

$$S_{\text{PLA}}(t|\boldsymbol{x}) = \exp(-H_0(t)e^{\boldsymbol{\beta}'\boldsymbol{x}});$$

(iii) $\delta = 1, \nu = 1$:

$$\frac{f_{\text{PLA}}(t|\boldsymbol{x})}{S_{\text{PLA}}(\tau_L|\boldsymbol{x})} = \frac{h_0(t)e^{\boldsymbol{\beta}'\boldsymbol{x}} \exp(-H_0(t)e^{\boldsymbol{\beta}'\boldsymbol{x}})}{\exp(-H_0(\tau_L)e^{\boldsymbol{\beta}'\boldsymbol{x}})};$$

(iv) $\delta = 0, \nu = 1$:

$$\frac{S_{\text{PLA}}(t|\boldsymbol{x})}{S_{\text{PLA}}(\tau_L|\boldsymbol{x})} = \frac{\exp(-H_0(t)e^{\boldsymbol{\beta}'\boldsymbol{x}})}{\exp(-H_0(\tau_L)e^{\boldsymbol{\beta}'\boldsymbol{x}})}.$$

Combining all these cases, the general likelihood for LTRC survival data based on the proposed model is obtained as

$$\begin{aligned} L(\boldsymbol{\theta}|DATA) = \prod_{i=1}^{n} &\left\{ h_0(t_i)e^{\boldsymbol{\beta}'\boldsymbol{x}_i} \exp(-H_0(t_i)e^{\boldsymbol{\beta}'\boldsymbol{x}_i}) \right\}^{(\delta_i(1-\nu_i))} \left\{ \exp(-H_0(t_i)e^{\boldsymbol{\beta}'\boldsymbol{x}_i}) \right\}^{(1-\delta_i)(1-\nu_i)} \\ &\times \left\{ \frac{h_0(t_i)e^{\boldsymbol{\beta}'\boldsymbol{x}_i} \exp(-H_0(t_i)e^{\boldsymbol{\beta}'\boldsymbol{x}_i})}{\exp(-H_0(\tau_{Li})e^{\boldsymbol{\beta}'\boldsymbol{x}_i})} \right\}^{\delta_i\nu_i} \left\{ \frac{\exp(-H_0(t_i)e^{\boldsymbol{\beta}'\boldsymbol{x}_i})}{\exp(-H_0(\tau_{Li})e^{\boldsymbol{\beta}'\boldsymbol{x}_i})} \right\}^{(1-\delta_i)\nu_i}. \end{aligned} \tag{9}$$

Here, $\boldsymbol{\theta}$ is the vector of all relevant parameters, including the slopes of the linear pieces that constitute the PLA $H_0(t)$ and the vector of regression parameters $\boldsymbol{\beta}$. Upon simplification, the corresponding log-likelihood function can be given as

$$\log L(\boldsymbol{\theta}|DATA) = \sum_{i=1}^{n} \left[\delta_i(1-\nu_i) \log h_0(t_i) + \delta_i \boldsymbol{\beta}'\boldsymbol{x}_i - H_0(t_i)e^{\boldsymbol{\beta}'\boldsymbol{x}_i} + \nu_i H_0(\tau_{Li})e^{\boldsymbol{\beta}'\boldsymbol{x}_i} \right]. \tag{10}$$

The log-likelihood in (10) needs to be maximized to obtain the maximum likelihood estimates (MLEs) of the model parameters. It may be noted here that the dimension of $\boldsymbol{\theta}$, and hence the complexity of the optimization problem, depends on the number of linear pieces used in the PLA $H_0(t)$.

4.2.1 | Confidence intervals

Let $I(\boldsymbol{\theta}) = E(J(\boldsymbol{\theta}))$ denote the expected Fisher information matrix, where $J(\boldsymbol{\theta})$, given by

$$\boldsymbol{J}(\boldsymbol{\theta}) = -\nabla^2(\log L(\boldsymbol{\theta})),$$

is the negative of the second derivatives of the log-likelihood function in (10) with respect to $\boldsymbol{\theta}$. Inverse of $I(\boldsymbol{\theta})$, computed at the MLE $\widehat{\boldsymbol{\theta}}$, gives an estimate of the asymptotic variance-covariance matrix of the estimates. Using asymptotic normality of the MLEs, pointwise 95% confidence intervals for each of the model parameters can be constructed by using the estimated asymptotic variances. In fact, instead of the expected Fisher information matrix $I(\boldsymbol{\theta})$, the observed Fisher information matrix $J(\boldsymbol{\theta})$ may also be used for this purpose. Using asymptotic normality of the MLEs, we have

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \rightarrow \boldsymbol{N}_d(\boldsymbol{0}, \boldsymbol{J}^{-1}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}),$$

where the dimension d of the multivariate normal distribution depends on the dimension of $\boldsymbol{\theta}$. Asymptotic 95% confidence intervals (CIs) for the model parameters can then be constructed easily using these results; for example, for the regression parameter β_1, the CI is given by

$$\widehat{\beta}_1 \pm 1.96\sqrt{\widehat{Var}(\widehat{\beta}_1)},$$

where $\widehat{Var}(\widehat{\beta}_1)$ is the estimated variance of $\widehat{\beta}_1$.

5 | SIMULATION STUDY

The simulation study presented here is intended to examine the performance of the PLA-based model under different scenarios. We consider three scenarios for generating LTRC data in the simulation study; LTRC data of a given size are generated from (a) Weibull distributions, (b) a PH model with the cumulative baseline hazard as a quadratic function of time, and (c) a PH model with a mixture of Weibull distributions as the cumulative baseline hazard. These scenarios cover a wide range of data observed in practice; for example, LTRC data with a monotone hazard rate (chosen by specifying the shape parameter of the Weibull distribution in Case (a)), or with a non-monotone hazard rate (for data generated in Case (c)). By fitting the proposed PLA model to the data generated from these processes, we examine the robustness property of the PLA-based model in providing a fit.

For the Weibull distribution, the cumulative hazard is given by

$$\Lambda(t|x) = (\lambda t)^{\gamma} e^{-\beta x},$$

where λ, γ, and β are scale, shape, and regression parameters, respectively. The PH model with quadratic baseline hazard used for data generation is given by the cumulative hazard

$$\Lambda(t|x) = (\lambda t + \gamma t^2)e^{-\beta x}.$$

It is of interest to mention here that statistical distributions with hazard function as a lower order polynomial have been used earlier in life-testing studies; the linear-exponential distribution is an example of this kind. One may refer to the works of Bain[25] and Balakrishnan and Malik,[26] and the references therein. The cumulative hazard of a PH model with a mixture of Weibull distributions as the cumulative baseline hazard is given by

$$\Lambda(t|x) = (\lambda_1 t^{\alpha_1} + \lambda_2 t^{\alpha_2})e^{-\beta x}.$$

Different values of the parameters for these processes are chosen for generating data in this empirical study.

For assessing the performance of the PLA-based model, we define an absolute integrated error (AIE) measure. This measure is calculated based on the survival function and the cumulative hazard as follows:

$$\text{AIE}_{\text{SF}} = \frac{1}{R}\sum_{k=1}^{R}\frac{1}{\zeta_k - \xi_k}\int_{\xi_k}^{\zeta_k}|\widehat{S}_{\text{PLA}}(t) - S_{\text{TGP}}(t)|dt,$$

$$\text{AIE}_{\text{CHF}} = \frac{1}{R}\sum_{k=1}^{R}\frac{1}{\zeta_k - \xi_k}\int_{\xi_k}^{\zeta_k}|\widehat{H}_{\text{PLA}}(t) - \Lambda(t|x)|dt,$$

where ξ_k and ζ_k are the minimum and maximum of the observed lifetime data for the kth Monte Carlo sample, $\widehat{S}_{\text{PLA}}(\cdot)$ and $\widehat{H}_{\text{PLA}}(\cdot)$ are the estimated survival and cumulative hazard functions by using the PLA-based model, respectively, and $S_{\text{TGP}}(\cdot)$ is the survival function of the true generating process. Here, R is the number of Monte Carlo repetitions. These measures provide an overall assessment of the PLA-based model over the entire observed data range. While the measure AIE is not scaled, it is evident that lower values of AIE imply a good approximation of the true data generation process.

For generating LTRC survival data, we use the following simulation settings: sample size $n = 300$ and 500, and left truncation percentage $p = 10\%$ and 30%. As the right censoring is random, we choose the relevant simulation parameters in such a way that corresponding to each of the truncation percentages, we obtain two approximate percentages of right censoring: 20% and 40%. The other relevant values of simulation parameters are indicated in the tables.

To the generated LTRC data, we fit different versions of the PLA-based model with four cut-points (ie, three linear pieces). The first cut-point is always set at a point that is less the minimum of the observed data values while the other three cut-points are selected by varying their positions across different quantiles of the data. In particular, the following sets of cut-points were used: $(t_0^*, q_{0.2}, q_{0.5}, q_{0.7})$, $(t_0^*, q_{0.2}, q_{0.5}, q_{0.8})$, $(t_0^*, q_{0.3}, q_{0.5}, q_{0.7})$, and $(t_0^*, q_{0.3}, q_{0.5}, q_{0.8})$, where $t_0^* < \text{Min}(y_{\min}, \tau_{L\min})$ (as mentioned in Section 4). The simulations were carried out in `R` software, using parallel computing facilities provided by the `doMC` and `doRNG` packages.

The values of the measures AIE_{SF} and AIE_{CHF} are reported in Tables 1–4 for different simulation settings. For LTRC data generated from Weibull distributions (Tables 1 and 2), with respect to AIE_{SF}, observe that it increases with censoring percentage for a given level of truncation. However, it remains more or less at the same level irrespective of the truncation percentage. This shows that estimation based on this model is more affected by censoring than by truncation. Also, as expected, with an increase in the sample size (from 300 to 500), the measure improves, implying that the model gets better for larger sample sizes. The values of the measure AIE_{CHF} although show a close approximation of the underlying data generation process by the PLA-based model, they do behave somewhat differently. It may be observed that AIE_{CHF}, for a given level of truncation, decreases with an increase in censoring percentage. This apparent anomalous behavior of AIE_{CHF} can be attributed to the cumulative hazard function being unbounded above. Also, it is more heavily affected by truncation, increasing with truncation percentage.

For the LTRC data generated from a PH model with quadratic cumulative baseline hazard, we see almost similar trends when the PLA-based model is fitted (Table 3). In this case as well, the overall model fit is quite good as reflected by both AIE_{SF} and AIE_{CHF}. However, when the parent process is a PH model with a mixture of Weibull as the cumulative baseline hazard (Table 4), the performance of the PLA-based model in fitting the generated LTRC data is somewhat inferior compared to the other two cases discussed above. This relatively poor performance in this case may be attributed to the non-monotone nature of the hazard function of the parent process.

Choice of cut-points for the PLA model do not seem to play a key role in any of the above scenarios considered. However, choice of cut-points near the two ends, for example, at $q_{0.05}$ and $q_{0.95}$ of the observed data range should be avoided. If the cut-points are near the two ends, sufficient number of data points may not be available to fit the linear pieces at the two terminal intervals, resulting in a poor fit. As long as the cut-points are at interior positions of the data (apart from the first cut-point at t_0^*) and there are enough datapoints available at each segment of the PLA, the approximation of the underlying cumulative hazard by the PLA-based model seems to work quite well.

Overall, our conclusion from the above elaborate simulation study is that the proposed PLA-based model works successfully in fitting the LTRC data under different scenarios. The quality of fit, as assessed by the measures AIE_{SF} and AIE_{CHF}, shows that the proposed approach is quite robust, and can accommodate various forms of the underlying cumulative baseline hazard function.

TABLE 1 Absolute integrated error (AIE) for the proposed PLA model for data generated from Weibull distribution with $\lambda = 0.3$, $\gamma = 0.5$, $\beta = -0.3$.

			AIE_{SF}		AIE_{CHF}	
n	Truncation	Censoring	$x = 0$	$x = 1$	$x = 0$	$x = 1$
Cut-points: $(t_0^*, q_{0.2}, q_{0.5}, q_{0.7})$						
300	10%	20%	0.038	0.024	0.185	0.074
		40%	0.048	0.029	0.133	0.058
	30%	20%	0.038	0.025	0.180	0.074
		40%	0.049	0.030	0.130	0.060
500	10%	20%	0.030	0.019	0.144	0.058
		40%	0.039	0.024	0.105	0.047
	30%	20%	0.031	0.020	0.141	0.058
		40%	0.040	0.025	0.104	0.048
Cut-points: $(t_0^*, q_{0.2}, q_{0.5}, q_{0.8})$						
300	10%	20%	0.039	0.025	0.190	0.079
		40%	0.048	0.029	0.132	0.059
	30%	20%	0.039	0.025	0.183	0.077
		40%	0.049	0.030	0.130	0.060
500	10%	20%	0.031	0.020	0.148	0.063
		40%	0.039	0.024	0.105	0.047
	30%	20%	0.031	0.020	0.143	0.061
		40%	0.040	0.025	0.104	0.048
Cut-points: $(t_0^*, q_{0.3}, q_{0.5}, q_{0.7})$						
300	10%	20%	0.040	0.025	0.188	0.075
		40%	0.051	0.030	0.138	0.059
	30%	20%	0.041	0.027	0.182	0.077
		40%	0.054	0.035	0.136	0.065
500	10%	20%	0.032	0.020	0.147	0.059
		40%	0.042	0.025	0.110	0.048
	30%	20%	0.033	0.022	0.143	0.060
		40%	0.045	0.029	0.109	0.052
Cut-points: $(t_0^*, q_{0.3}, q_{0.5}, q_{0.8})$						
300	10%	20%	0.040	0.025	0.193	0.079
		40%	0.051	0.030	0.137	0.060
	30%	20%	0.041	0.028	0.185	0.081
		40%	0.054	0.035	0.136	0.065
500	10%	20%	0.032	0.021	0.151	0.063
		40%	0.042	0.025	0.110	0.048
	30%	20%	0.034	0.023	0.146	0.065
		40%	0.045	0.029	0.109	0.052

TABLE 2 Absolute integrated error (AIE) for the proposed PLA model for data generated from Weibull distribution with $\lambda = 0.3$, $\gamma = 3.0$, $\beta = -0.3$.

			AIE_{SF}		AIE_{CHF}	
n	**Truncation**	**Censoring**	$x = 0$	$x = 1$	$x = 0$	$x = 1$
Cut-points: $(t_0^*, q_{0.2}, q_{0.5}, q_{0.7})$						
300	10%	20%	0.046	0.033	0.164	0.088
		40%	0.052	0.038	0.142	0.080
	30%	20%	0.047	0.034	0.181	0.096
		40%	0.053	0.038	0.148	0.080
500	10%	20%	0.041	0.031	0.147	0.084
		40%	0.047	0.034	0.126	0.073
	30%	20%	0.042	0.032	0.158	0.089
		40%	0.047	0.035	0.125	0.071
Cut-points: $(t_0^*, q_{0.2}, q_{0.5}, q_{0.8})$						
300	10%	20%	0.046	0.033	0.162	0.087
		40%	0.052	0.038	0.142	0.080
	30%	20%	0.047	0.034	0.178	0.093
		40%	0.053	0.038	0.148	0.079
500	10%	20%	0.041	0.030	0.143	0.080
		40%	0.046	0.034	0.123	0.070
	30%	20%	0.042	0.031	0.152	0.083
		40%	0.047	0.035	0.123	0.068
Cut-points: $(t_0^*, q_{0.3}, q_{0.5}, q_{0.7})$						
300	10%	20%	0.061	0.045	0.181	0.103
		40%	0.069	0.052	0.162	0.097
	30%	20%	0.061	0.046	0.197	0.109
		40%	0.069	0.052	0.167	0.096
500	10%	20%	0.056	0.043	0.167	0.098
		40%	0.063	0.048	0.148	0.090
	30%	20%	0.056	0.043	0.176	0.102
		40%	0.064	0.049	0.146	0.087
Cut-points: $(t_0^*, q_{0.3}, q_{0.5}, q_{0.8})$						
300	10%	20%	0.061	0.045	0.180	0.101
		40%	0.068	0.051	0.161	0.096
	30%	20%	0.061	0.046	0.195	0.106
		40%	0.069	0.051	0.166	0.095
500	10%	20%	0.055	0.042	0.162	0.094
		40%	0.063	0.048	0.144	0.086
	30%	20%	0.056	0.043	0.169	0.097
		40%	0.064	0.048	0.143	0.084

TABLE 3 Absolute integrated error (AIE) for the proposed PLA model for data generated from PH model with a quadratic function as cumulative baseline hazard, $\lambda = 0.3$, $\gamma = 0.5$, $\beta = 0.3$.

			AIE_{SF}		AIE_{CHF}	
n	**Truncation**	**Censoring**	$x = 0$	$x = 1$	$x = 0$	$x = 1$
Cut-points: $(t_0^*, q_{0.2}, q_{0.5}, q_{0.7})$						
300	10%	20%	0.029	0.020	0.230	0.115
		40%	0.036	0.025	0.213	0.110
	30%	20%	0.028	0.020	0.242	0.118
		40%	0.035	0.024	0.200	0.095
500	10%	20%	0.022	0.017	0.205	0.113
		40%	0.028	0.021	0.183	0.102
	30%	20%	0.022	0.017	0.203	0.110
		40%	0.027	0.020	0.160	0.082
Cut-points: $(t_0^*, q_{0.2}, q_{0.5}, q_{0.8})$						
300	10%	20%	0.029	0.020	0.229	0.112
		40%	0.035	0.025	0.211	0.107
	30%	20%	0.028	0.020	0.240	0.112
		40%	0.034	0.023	0.202	0.093
500	10%	20%	0.022	0.016	0.198	0.105
		40%	0.027	0.020	0.175	0.093
	30%	20%	0.022	0.017	0.194	0.100
		40%	0.027	0.019	0.156	0.076
Cut-points: $(t_0^*, q_{0.3}, q_{0.5}, q_{0.7})$						
300	10%	20%	0.031	0.023	0.233	0.118
		40%	0.038	0.028	0.215	0.113
	30%	20%	0.032	0.024	0.246	0.122
		40%	0.036	0.025	0.202	0.097
500	10%	20%	0.025	0.019	0.209	0.116
		40%	0.031	0.023	0.187	0.105
	30%	20%	0.026	0.021	0.208	0.114
		40%	0.029	0.021	0.162	0.084
Cut-points: $(t_0^*, q_{0.3}, q_{0.5}, q_{0.8})$						
300	10%	20%	0.031	0.023	0.231	0.115
		40%	0.038	0.027	0.213	0.109
	30%	20%	0.032	0.023	0.244	0.116
		40%	0.036	0.025	0.203	0.094
500	10%	20%	0.025	0.019	0.202	0.107
		40%	0.030	0.022	0.178	0.096
	30%	20%	0.026	0.020	0.198	0.103
		40%	0.028	0.020	0.158	0.078

TABLE 4 Absolute integrated error (AIE) for the proposed PLA model for data generated from PH model with a mixture of Weibull cumulative hazards as cumulative baseline hazard, $\alpha_1 = 0.5$, $\lambda_1 = 0.3$, $\alpha_2 = 2.0$, $\lambda_2 = 0.3$, $\beta = -0.3$.

			AIE_{SF}		AIE_{CHF}	
n	**Truncation**	**Censoring**	$x = 0$	$x = 1$	$x = 0$	$x = 1$
Cut-points: $(t_0^*, q_{0.2}, q_{0.5}, q_{0.7})$						
300	10%	20%	0.128	0.165	0.481	0.518
		40%	0.146	0.188	0.431	0.471
	30%	20%	0.128	0.163	0.507	0.534
		40%	0.151	0.189	0.461	0.483
500	10%	20%	0.124	0.161	0.452	0.517
		40%	0.143	0.186	0.411	0.472
	30%	20%	0.127	0.162	0.487	0.533
		40%	0.151	0.189	0.449	0.481
Cut-points: $(t_0^*, q_{0.2}, q_{0.5}, q_{0.8})$						
300	10%	20%	0.129	0.166	0.492	0.528
		40%	0.147	0.189	0.442	0.481
	30%	20%	0.128	0.163	0.518	0.544
		40%	0.151	0.189	0.466	0.488
500	10%	20%	0.124	0.162	0.464	0.530
		40%	0.144	0.187	0.423	0.483
	30%	20%	0.127	0.162	0.498	0.543
		40%	0.151	0.189	0.455	0.486
Cut-points: $(t_0^*, q_{0.3}, q_{0.5}, q_{0.7})$						
300	10%	20%	0.129	0.166	0.490	0.524
		40%	0.146	0.188	0.438	0.476
	30%	20%	0.125	0.160	0.511	0.535
		40%	0.149	0.187	0.463	0.483
500	10%	20%	0.124	0.162	0.460	0.523
		40%	0.143	0.186	0.419	0.477
	30%	20%	0.123	0.159	0.491	0.534
		40%	0.148	0.186	0.451	0.481
Cut-points: $(t_0^*, q_{0.3}, q_{0.5}, q_{0.8})$						
300	10%	20%	0.129	0.167	0.501	0.534
		40%	0.147	0.189	0.449	0.486
	30%	20%	0.125	0.160	0.522	0.545
		40%	0.149	0.187	0.468	0.488
500	10%	20%	0.125	0.163	0.473	0.536
		40%	0.144	0.187	0.431	0.488
	30%	20%	0.124	0.159	0.502	0.544
		40%	0.149	0.187	0.456	0.486

5.1 | Sensitivity analysis

In the sensitivity analysis, we examine the PLA-based model in the following way. First, a LTRC dataset is generated from a true process whose baseline hazard rate is a piecewise constant function. The true baseline hazard rate, spread over four intervals, is chosen to be non-monotone, with slopes 0.5, 0.3, 1.0, and 2.0 for the intervals, respectively. The cut-points defining the intervals for the true baseline hazard are placed at the quantiles $q_{0.3}$, $q_{0.5}$, and $q_{0.7}$ of the LTRC data. Using the slopes, the numerical values of the quantiles can be worked out to be 0.71, 1.83, and 2.34. Therefore, in summary, the true baseline hazard rate of the data generating process is taken as

$$h(t) = \begin{cases} 0.5 & \text{if } 0 < t \leq 0.71 \\ 0.3 & \text{if } 0.71 < t \leq 1.83 \\ 1.0 & \text{if } 1.83 < t \leq 2.34 \\ 2.0 & \text{if } t \geq 2.34. \end{cases}$$

Also, in the data generation process, one covariate with two levels (ie, $x = 1$ and 0) is considered. Without loss of any generality, the regression coefficient is taken as $\beta = -0.3$.

To the LTRC data thus generated, we fit the proposed PLA-based model, by applying the method for choosing the number of cut-points and their positions as described in Section 4.1.1. That is, we fit all 15 candidate PLA-based models to the data, by varying the number of cut-points and their positions over the different quantiles $q_{0.2}$, $q_{0.4}$, $q_{0.6}$, and $q_{0.8}$, and select the PLA-based model with the lowest AIC value as the best model. In addition, we also fit a PLA-based model to the generated LTRC data, by placing the cut-points at $q_{0.3}$, $q_{0.5}$, and $q_{0.7}$ (ie, at the same positions as the true data generating process). Finally, we calculate a metric involving the AIEs, both with respect to the survival function and the cumulative hazard function, as follows.

Suppose $\mathcal{M}^*$ is the best PLA-based model among the 15 candidate PLA-based models, and $\mathcal{M}^T$ is the PLA-based model fitted to the data by placing the cut-points at the same positions as the data generation process. Define

$$R^{(i)}_{\text{SF}} = \frac{\text{AIE}^{(i)}_{\text{SF}}(\mathcal{M}^*)}{\text{AIE}^{(i)}_{\text{SF}}(\mathcal{M}^T)},$$

$$R^{(i)}_{\text{CHF}} = \frac{\text{AIE}^{(i)}_{\text{CHF}}(\mathcal{M}^*)}{\text{AIE}^{(i)}_{\text{CHF}}(\mathcal{M}^T)},$$

where $\text{AIE}^{(i)}_{\text{SF}}(\mathcal{M})$ and $\text{AIE}^{(i)}_{\text{CHF}}(\mathcal{M})$ are the AIE_{SF} and AIE_{CHF}, respectively, for model $\mathcal{M}$, evaluated for covariate level i (ie, $x = i$), $i = 0, 1$. Clearly, lower values of $R^{(i)}_{\text{SF}}$ and $R^{(i)}_{\text{CHF}}$ indicate strength of the proposed PLA-based model and the model selection procedure.

In Table 5, the results of a simulation study on this sensitivity analysis are presented. The simulation parameters vary across different sample sizes, truncation and censoring percentages. Two different percentages of units across the two levels of the covariate are chosen; denoted by $P(x = 1)$, the table reports percentage of units with covariate value $x = 1$.

It may be noted from Table 5 that the values of the ratios are always close to 1, and quite often less than 1, indicating that the proposed PLA-based model, with the model selection procedure described in Section 4.1.1, works very well. Effect of left truncation percentage is relatively heavier on the sensitivity of the model than that of the right censoring percentage. Sample size and covariate information percentage do not seem to have any significant effect.

Therefore, in summary, based on the overall results of the simulation studies presented in this section, it may be concluded that the proposed PLA-based model can be applied quite effectively for LTRC data.

6 | ILLUSTRATIVE DATA ANALYSES

In this section, analyses of two real datasets, introduced earlier in Section 3, are presented for illustrative purposes. The proposed PLA-based model is used for the analyses.

TABLE 5 Ratio of the AIEs of the PLA-based models.

n	Truncation	Censoring (approx)	$P(x=1)$	$R^{(0)}_{SF}$	$R^{(1)}_{SF}$	$R^{(0)}_{CHF}$	$R^{(1)}_{CHF}$
300	10%	20%	0.80	0.849	0.833	0.879	0.877
		40%	0.80	1.032	1.017	0.917	0.901
	30%	20%	0.80	1.186	1.171	1.004	1.041
		40%	0.80	1.248	1.231	1.037	1.017
500	10%	20%	0.80	0.815	0.787	0.821	0.852
		40%	0.80	1.024	0.994	0.876	0.882
	30%	20%	0.80	1.186	1.179	0.951	1.034
		40%	0.80	1.245	1.227	0.983	0.999
300	10%	20%	0.50	0.843	0.833	0.798	0.853
		40%	0.50	1.024	1.010	0.869	0.911
	30%	20%	0.50	1.016	1.045	0.856	0.957
		40%	0.50	1.127	1.140	0.912	0.974
500	10%	20%	0.50	0.821	0.832	0.741	0.849
		40%	0.50	1.016	1.021	0.810	0.895
	30%	20%	0.50	0.971	0.991	0.797	0.949
		40%	0.50	1.101	1.106	0.871	0.961

6.1 | Analysis of Worcester Heart Attack Study data

We analyse the WHAS dataset that was analysed by Chen and Yi.[7] Note that by design, as detailed in Section 3, all the subjects in this sample are left truncated, and some of the subjects are right censored. There are 461 subjects in this LTRC dataset, among whom almost 62% are right censored. Before analysis, without any loss of generality, we transform the time scale of the sample according to the formula $l^* = \frac{l}{1000}$ (ie, by dividing all the lifetimes, left truncation times, and right censoring times by the constant 1000). These transformations will not affect the inferences based on this sample in any way. We fit the proposed PLA-based model for the cumulative hazard to this data. Although there were many covariates, we use the information on BMI (kg/m^2) and initial heart rate (HR) (beats per minute) of subjects as the two covariates in our analyses, following Chen and Yi.[7] The method of model fitting would be similar if other covariates were used.

The cut-points of the PLA-based model for this data are chosen by applying the methods described earlier in Section 4.1.1, that is, in two different ways: by using the method based on quantiles, and by eye-estimating the points where Breslow estimator changes significantly. In both cases, the first cut-point is placed at $t_0^* = 0.0$, as usual. For the quantile-based method, the candidate cut-points are $q_{0.20} = 0.3446$, $q_{0.40} = 0.5361$, $q_{0.60} = 1.1904$, and $q_{0.80} = 1.5774$, the values of which are obtained from the data in transformed time scale. For the other method, apart from t_0^*, the other cut-points are placed at $q_{0.15} = 0.1690$ and $q_{0.85} = 1.9191$ by eye-estimation.

Note that when there are k cut-points, that is, k linear pieces for $H_0(\cdot)$, there are $k+2$ parameters to estimate from the WHAS sample: the k slopes corresponding to the k linear pieces, and the two regression parameters β_1 and β_2 corresponding to the covariates BMI and HR, respectively. The intercepts of the linear pieces can be expressed in terms of the slopes when the condition of continuity given in (5) of the cumulative hazard is used, as given by (8). Table 6 gives the estimates of slopes of the linear pieces and the regression parameters for different models. Here, $\mathcal{M}_1, \ldots, \mathcal{M}_{15}$ represent the 15 PLA-based models obtained by applying the quantile-based cut-point selection method, and $\mathcal{M}_{16}$ represents the model obtained by eye-estimating the positions of cut-points, based on Breslow estimator.

Among $\mathcal{M}_1, \ldots, \mathcal{M}_{15}$, model $\mathcal{M}_1$ with cut-points at t_0^* and $q_{0.20}$ provides the best fit with respect to the AIC value. However, overall, the AIC value is the lowest for model $\mathcal{M}_{16}$, and hence for the WHAS data, $\mathcal{M}_{16}$ turns out to be the best PLA-based model.

It may be of interest to compare the fits provided by models $\mathcal{M}_1$ and $\mathcal{M}_{16}$, and also to compare the fits with that provided by the standard technique like semiparametric Cox regression model in which the cumulative baseline hazard is

TABLE 6 Fitting of different PLA-based models to the Worcester Heart Attack Study data.

Model	Cut-points	$\widehat{b}_1$	$\widehat{b}_2$	$\widehat{b}_3$	$\widehat{b}_4$	$\widehat{b}_5$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	AIC
$\mathcal{M}_1$	t_0^*, $q_{0.2}$	1.577	0.696	–	–	–	−0.093	0.017	574.19
$\mathcal{M}_2$	t_0^*, $q_{0.4}$	1.431	0.666	–	–	–	−0.094	0.017	578.59
$\mathcal{M}_3$	t_0^*, $q_{0.6}$	1.023	0.760	–	–	–	−0.096	0.018	600.61
$\mathcal{M}_4$	t_0^*, $q_{0.8}$	1.003	0.616	–	–	–	−0.097	0.019	600.15
$\mathcal{M}_5$	t_0^*, $q_{0.2}$, $q_{0.4}$	1.603	0.920	0.649	–	–	−0.093	0.017	574.28
$\mathcal{M}_6$	t_0^*, $q_{0.2}$, $q_{0.6}$	1.572	0.663	0.767	–	–	−0.093	0.017	575.81
$\mathcal{M}_7$	t_0^*, $q_{0.2}$, $q_{0.8}$	1.581	0.708	0.627	–	–	−0.093	0.017	576.06
$\mathcal{M}_8$	t_0^*, $q_{0.4}$, $q_{0.6}$	1.429	0.584	0.800	–	–	−0.094	0.017	579.11
$\mathcal{M}_9$	t_0^*, $q_{0.4}$, $q_{0.8}$	1.429	0.668	0.650	–	–	−0.094	0.017	580.56
$\mathcal{M}_{10}$	t_0^*, $q_{0.6}$, $q_{0.8}$	1.023	0.881	0.618	–	–	−0.097	0.018	601.80
$\mathcal{M}_{11}$	t_0^*, $q_{0.2}$, $q_{0.4}$, $q_{0.6}$	1.605	0.921	0.571	0.782	–	−0.092	0.017	574.82
$\mathcal{M}_{12}$	t_0^*, $q_{0.2}$, $q_{0.4}$, $q_{0.8}$	1.609	0.924	0.654	0.638	–	−0.093	0.017	576.27
$\mathcal{M}_{13}$	t_0^*, $q_{0.2}$, $q_{0.6}$, $q_{0.8}$	1.577	0.665	0.890	0.626	–	−0.093	0.017	577.01
$\mathcal{M}_{14}$	t_0^*, $q_{0.4}$, $q_{0.6}$, $q_{0.8}$	1.425	0.582	0.923	0.649	–	−0.094	0.017	580.31
$\mathcal{M}_{15}$	t_0^*, $q_{0.2}$, $q_{0.4}$, $q_{0.6}$, $q_{0.8}$	1.623	0.932	0.578	0.915	0.643	−0.093	0.017	576.02
$\mathcal{M}_{16}$	t_0^*, $q_{0.15}$, $q_{0.85}$	2.184	0.680	1.353	–	–	−0.091	0.017	555.45

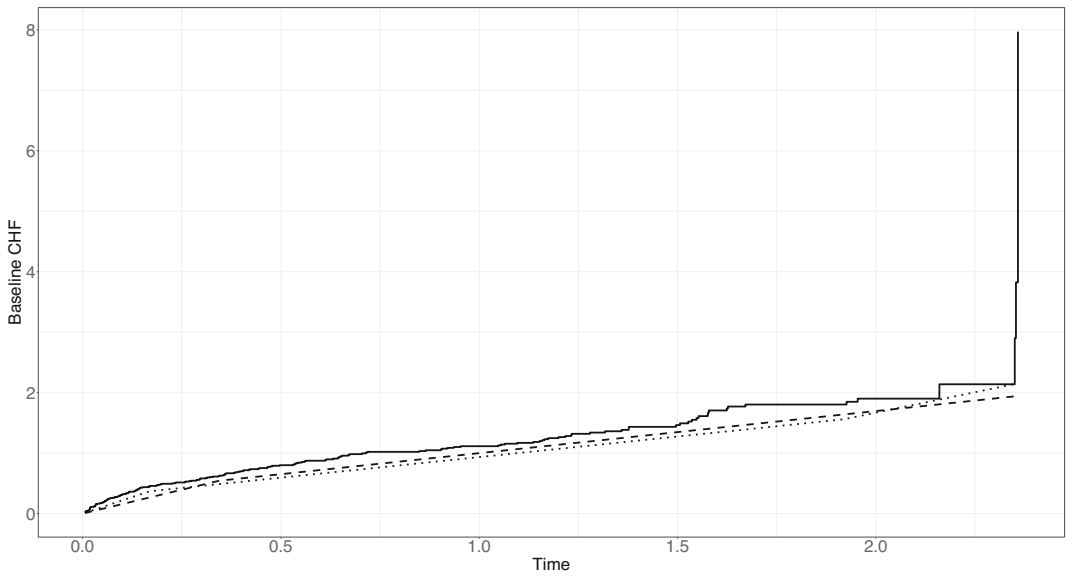

FIGURE 2 Plot of the estimated cumulative baseline hazard for the WHAS sample corresponding to Breslow estimator (solid line), $\mathcal{M}_1$ (dashed line), and $\mathcal{M}_{16}$ (dotted line).

estimated by Breslow estimator. In Figure 2, the fitted cumulative baseline hazards for $\mathcal{M}_1$ and $\mathcal{M}_{16}$ are plotted against Breslow estimator for the purpose of comparison. There are three subjects in the WHAS sample with very high survival times. Due to their presence, Breslow estimator increases very sharply near the right end; see Figure 2. For same reason, the proposed PLA-based estimators underestimate the baseline cumulative hazard as compared to Breslow estimator. Table 7 details the point estimates and the corresponding 95% confidence intervals for the slopes and regression parameters of $\mathcal{M}_1$ and $\mathcal{M}_{16}$. It also includes the estimates of regression parameters along with corresponding confidence intervals when the semiparametric Cox model is fitted to the WHAS data. It may be noticed that although the point estimates of

TABLE 7 Details of the PLA-based models $\mathcal{M}_1$ and $\mathcal{M}_{16}$, and the semiparametric Cox model for the Worcester Heart Attack Study data: Point estimates and corresponding 95% confidence intervals.

Model	CI for $\hat{b}_1$	CI for $\hat{b}_2$	CI for $\hat{b}_3$	CI for $\hat{\beta}_1$	CI for $\hat{\beta}_2$
$\mathcal{M}_1$	1.577	0.696	–	−0.093	0.017
	(0.319, 2.836)	(0.451, 0.940)	–	(−0.0934, −0.0937)	(0.0173, 0.0174)
$\mathcal{M}_{16}$	2.184	0.680	1.353	−0.091	0.017
	(0, 4.644)	(0.447, 0.916)	(0, 2.870)	(−0.092, −0.091)	$(1.731 \times 10^{-2}, 1.734 \times 10^{-2})$
SP Cox	–	–	–	−0.094	0.016
	–	–	–	(−0.126, −0.062)	(0.010, 0.022)

the regression parameters for the three models are quite close, the corresponding 95% confidence intervals for them are significantly wider in the case of semiparametric Cox model.

Note that for $\mathcal{M}_{16}$, which is selected as the best PLA-based model for the WHAS data, the estimated slopes of the first and third linear pieces ($b_1 = 2.193$ and $b_3 = 1.358$, respectively) are larger than that of the second linear piece ($b_2 = 0.683$). It implies that for a subject with a given value of BMI and HR, the hazard of experiencing the event (in this case, death) is higher immediately after discharge from the hospital and at the later phase of the subject's life, compared to that at the middle period after discharge from the hospital. In other words, a subject is at higher risk of death immediately after discharge from the hospital (perhaps because of the immediate after-effects of experiencing MI) and at the later phase of their life (naturally due to old age). After some time elapses since discharge from the hospital, the subject is at a comparatively lower risk of death. This clearly shows that in this scenario, subjects should be under significant level of medical observation immediately after they are discharged from the hospitals.

6.2 | Analysis of Channing House data

The Channing House data is available in the `R` package `boot`. Note that each observation in this dataset is left truncated. For the study subjects, the lifetime variable is the age (recorded in months) of death or of leaving the center at the end of the study. Out of the 462 total subjects, there were 4 subjects who left the retirement center immediately after admittance (lifetime = 0). So, in our analyses, we use information on 458 subjects of the center, removing those 4 subjects. It may be noted here that about 62% of the observations are right censored in this dataset.

Recall that the minimum age to enter the retirement center was 60 years. Without loss of any generality, we transform the lifetimes by changing their location and scale: for each lifetime, we consider the time transformation $l^* = \frac{l-u}{v}$, where l is the original lifetime, $u = 720$, and $v = 100$. The left truncation times are transformed similarly as well. These transformations will not affect the inferences based on this data in anyway.

For fitting the PLA-based model to this transformed Channing House data, the first cut-point t_0^* is at 0.12. For the quantile-based method for choosing cut-points, apart from t_0^*, we have $q_{0.2} = 2.070$, $q_{0.4} = 2.518$, $q_{0.6} = 2.860$, and $q_{0.8} = 3.226$ as the candidate cut-points. For the eye-estimation method, the cut-points are placed at $q_{0.2}$ and $q_{0.5} = 2.705$ along with t_0^*, as the Breslow estimator changes significantly at these points. In this process, we naturally avoid choosing the cut-points near the tails of the data, that is, at $q_{0.1}$ and $q_{0.9}$, in order to avoid the situation of not having enough datapoints in the intervals at the two ends. The only covariate in the Channing House data is the gender of the residents. The residents were either male or female. Thus, we code the covariate information by using the binary variable $x = 1$ for a female resident, and $x = 0$ for a male resident.

Note that for k cut-points, there are $k + 1$ parameters to estimate based on the Channing House data: the k slopes corresponding to the k linear pieces, and the regression parameter corresponding to the covariate (gender). Here, $\mathcal{M}_1$ to $\mathcal{M}_{15}$ represent the 15 PLA-based models obtained by applying the quantile-based cut-point choosing method, and $\mathcal{M}_{16}$ represents the model based on eye-estimation from Breslow estimator. Table 8 gives the details of fits of the various PLA-based models for the Channing House data. It may be noted that among the models $\mathcal{M}_1$ to $\mathcal{M}_{15}$, $\mathcal{M}_8$ provides the best fit in terms of AIC. However, the model $\mathcal{M}_{16}$, with three cut-points placed at t_0^*, $q_{0.2}$ and $q_{0.5}$, turns out to be the most suitable model for the Channing House data as it has the lowest value of AIC among all the PLA-based models considered here. In fact, with the cut-points at t_0^*, $q_{0.4}$ and $q_{0.6}$, the model $\mathcal{M}_8$ produces a very close fit.

TABLE 8 Fitting of different PLA-based models to the Channing House data.

Model	Cut-points	$\widehat{b}_1$	$\widehat{b}_2$	$\widehat{b}_3$	$\widehat{b}_4$	$\widehat{b}_5$	$\widehat{\beta}$	AIC
$\mathcal{M}_1$	t_0^*, $q_{0.2}$	0.310	0.828	–	–	–	−0.358	584.09
$\mathcal{M}_2$	t_0^*, $q_{0.4}$	0.333	1.073	–	–	–	−0.323	557.76
$\mathcal{M}_3$	t_0^*, $q_{0.6}$	0.415	1.337	–	–	–	−0.327	560.41
$\mathcal{M}_4$	t_0^*, $q_{0.8}$	0.533	1.471	–	–	–	−0.368	587.94
$\mathcal{M}_5$	t_0^*, $q_{0.2}$, $q_{0.4}$	0.302	0.384	1.072	–	–	−0.322	558.92
$\mathcal{M}_6$	t_0^*, $q_{0.2}$, $q_{0.6}$	0.301	0.523	1.330	–	–	−0.319	555.55
$\mathcal{M}_7$	t_0^*, $q_{0.2}$, $q_{0.8}$	0.307	0.691	1.449	–	–	−0.346	571.41
$\mathcal{M}_8$	t_0^*, $q_{0.4}$, $q_{0.6}$	0.330	0.722	1.324	–	–	−0.312	550.28
$\mathcal{M}_9$	t_0^*, $q_{0.4}$, $q_{0.8}$	0.333	0.935	1.424	–	–	−0.322	555.17
$\mathcal{M}_{10}$	t_0^*, $q_{0.6}$, $q_{0.8}$	0.416	1.251	1.430	–	–	−0.328	562.04
$\mathcal{M}_{11}$	t_0^*, $q_{0.2}$, $q_{0.4}$, $q_{0.6}$	0.299	0.380	0.721	1.322	–	−0.311	551.44
$\mathcal{M}_{12}$	t_0^*, $q_{0.2}$, $q_{0.4}$, $q_{0.8}$	0.302	0.383	0.935	1.423	–	−0.321	556.34
$\mathcal{M}_{13}$	t_0^*, $q_{0.2}$, $q_{0.6}$, $q_{0.8}$	0.301	0.523	1.244	1.422	–	−0.320	557.18
$\mathcal{M}_{14}$	t_0^*, $q_{0.4}$, $q_{0.6}$, $q_{0.8}$	0.330	0.723	1.238	1.415	–	−0.313	551.91
$\mathcal{M}_{15}$	t_0^*, $q_{0.2}$, $q_{0.4}$, $q_{0.6}$, $q_{0.8}$	0.300	0.381	0.722	1.237	1.414	−0.312	553.08
$\mathcal{M}_{16}$	t_0^*, $q_{0.2}$, $q_{0.5}$	0.299	0.427	1.239	–	–	−0.311	549.28

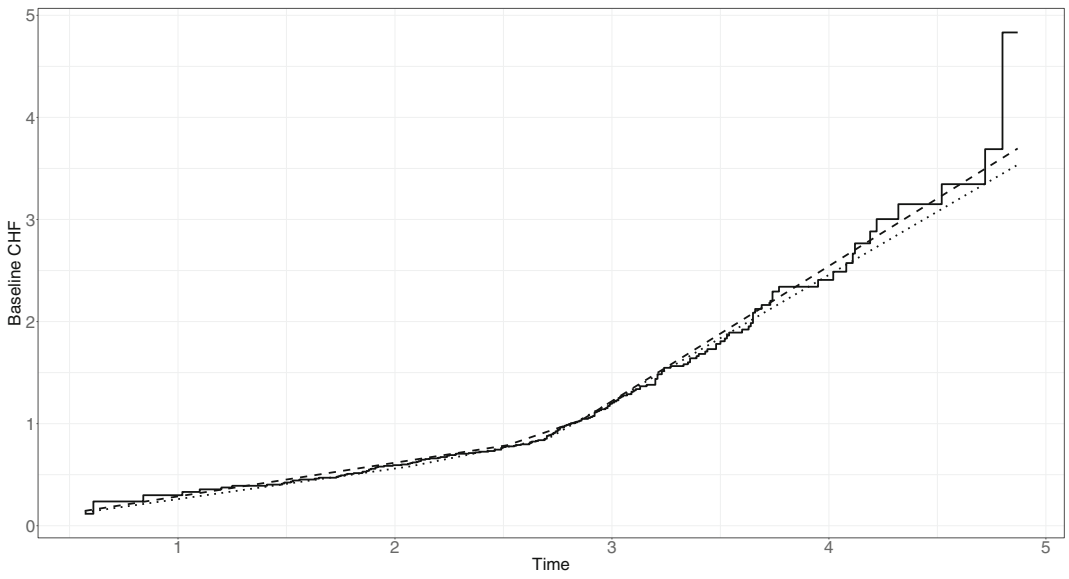

FIGURE 3 Plot of the estimated cumulative baseline hazard for the Channing House data corresponding to Breslow estimator (solid line), $\mathcal{M}_1$ (dashed line), and $\mathcal{M}_{16}$ (dotted line).

A comparison of models between $\mathcal{M}_8$, $\mathcal{M}_{16}$ and the semiparametric Cox regression model is carried out. Figure 3 gives plot of the fitted cumulative baseline hazards for $\mathcal{M}_8$, $\mathcal{M}_{16}$, and Breslow estimator. The estimated cumulative hazard is visibly quite close for all three models. Table 9 gives the point estimates, with corresponding 95% confidence intervals, for the relevant parameters of $\mathcal{M}_8$, $\mathcal{M}_{16}$, and the semiparametric Cox model. While the estimates of the regression parameter is very close for all three models, the 95% confidence intervals for semiparametric Cox model are significantly wider than the other two. For model $\mathcal{M}_{16}$, the estimated regression coefficient β is −0.311. This implies that for a given age, the cumulative hazard of a female resident is 0.733 times that of a male resident.

TABLE 9 Details of the PLA-based models $\mathcal{M}_8$ and $\mathcal{M}_{16}$, and the semiparametric Cox model for the Channing House data: Point estimates and corresponding 95% confidence intervals.

Model	CI for $\widehat{b}_1$	CI for $\widehat{b}_2$	CI for $\widehat{b}_3$	$\widehat{\beta}$
$\mathcal{M}_8$	0.330	0.722	1.324	−0.312
	(0.323, 0.337)	(0.676, 0.769)	(1.230, 1.417)	(−0.370, −0.254)
$\mathcal{M}_{16}$	0.299	0.427	1.239	−0.311
	(0.291, 0.308)	(0.413, 0.442)	(1.164, 1.315)	(−0.369, −0.254)
SP Cox	–	–	–	−0.316
	–	–	–	(−0.655, 0.023)

7 | PREDICTION ISSUES

We would like to reinforce one of the advantages of using the PLA-based model here. The proposed PLA-based model is essentially a parametric modeling approach wherein the unknown parameters are estimated based on LTRC survival data. However, it does not make strict parametric assumptions, such as the case when a single parametric lifetime model is assumed for the whole range of data. Moreover, being parametric in nature, it can be used for prediction purposes, such as predicting conditional survival beyond the observed range of data. Nonparametric procedures, like Breslow estimator for baseline CHF, are valid only for the observed range of data, and cannot be extrapolated outside such ranges. This is a unique advantage of using the PLA-based model proposed here over the fully nonparametric procedures.

However, it is also important to note the limitations of prediction. All predictions are essentially based on extrapolation. For the proposed PLA-based model, although prediction is possible, the quality and accuracy of the prediction depend on the behavior of the hazard rate outside the observed data range. If the behavior of the hazard rate does not change significantly beyond the observed data range, the PLA-based model can be used for prediction by extrapolating the linear piece over the last interval beyond the observed data range to achieve good results. Naturally, the quality of prediction would depend on whether the future time point (for which prediction is being made) is far away from the last observed data point or not. On the other hand, if the behavior of the hazard rate changes drastically outside the observed data range, expectedly, extrapolation of the last linear piece beyond the observed data range would not yield good results in prediction. It may also be noted here that this a general problem of prediction and extrapolation, common to all models that can be used for such purposes.

In survival analysis, some prediction issues are of natural interest. For example, one may want to estimate the probability of conditional survival at a future time point of a subject with right censored lifetime. For the WHAS data, for instance, the probability of conditional survival at a future time point of a patient who has been discharged from the hospital may be of interest. Also, in other survival experiments such as oncology trials, the probability of conditional survival at a future time point is an important quantity of interest.

Consider a subject that is right censored at τ_R with right censoring time t_R, regardless of their truncation status. We are interested in estimating the probability of conditional survival of the subject at a future time point t_f, where $t_f > t_R$. The required probability is given by

$$\begin{aligned}\pi(t_f, t_R) &= P(T > t_f | T > t_R)\\ &= 1 - \frac{F(t_f;\boldsymbol{\theta}) - F(t_R;\boldsymbol{\theta})}{1 - F(t_R;\boldsymbol{\theta})}.\end{aligned}$$

Then, by using the MLE $\widehat{\boldsymbol{\theta}}$ computed from the data, $\pi(t_f, t_R)$ can be estimated as

$$\begin{aligned}\widehat{\pi}(t_f, t_R) &= 1 - \frac{\exp\{-H(t_R;\widehat{\boldsymbol{\theta}})\} - \exp\{-H(t_f;\widehat{\boldsymbol{\theta}})\}}{\exp\{-H(t_R;\widehat{\boldsymbol{\theta}})\}}\\ &= 1 - \frac{\exp\{-H_0(t_R)\exp(\widehat{\boldsymbol{\beta}}'\boldsymbol{x})\} - \exp\{-H_0(t_f)\exp(\widehat{\boldsymbol{\beta}}'\boldsymbol{x})\}}{\exp\{-H_0(t_R)\exp(\widehat{\boldsymbol{\beta}}'\boldsymbol{x})\}}\\ &= \frac{\exp\{-H_0(t_f)\exp(\widehat{\boldsymbol{\beta}}'\boldsymbol{x})\}}{\exp\{-H_0(t_R)\exp(\widehat{\boldsymbol{\beta}}'\boldsymbol{x})\}},\end{aligned} \tag{11}$$

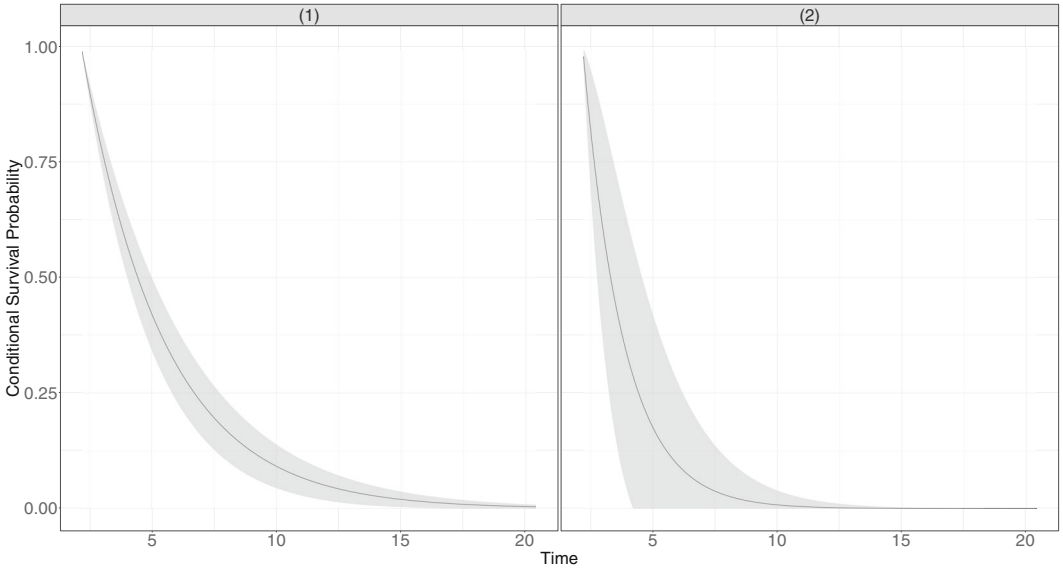

FIGURE 4 Plot of the conditional survival function at future time points for a right censored individuals (ID 1). The shaded region represents the 95% confidence intervals for the conditional survival probability. (1) and (2) corresponds to the models $\mathcal{M}_1$ and $\mathcal{M}_{16}$, respectively.

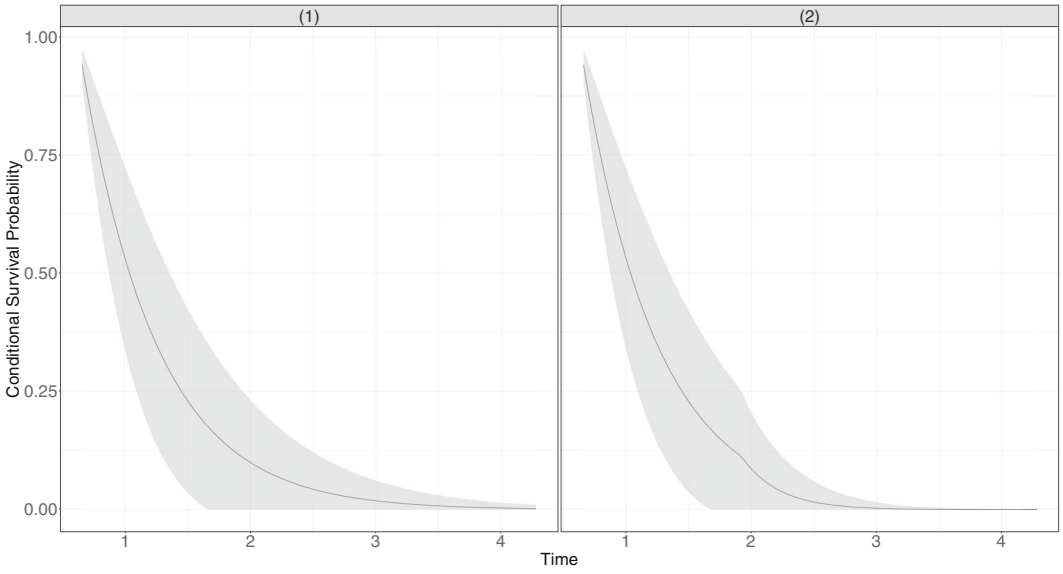

FIGURE 5 Plot of the conditional survival function at future time points for a right censored individuals (ID 432). The shaded region represents the 95% confidence intervals for the conditional survival probability. (1) and (2) corresponds to the models $\mathcal{M}_1$ and $\mathcal{M}_{16}$, respectively.

where $H_0(\cdot)$ is the PLA of the baseline cumulative hazard given in (3). A 95% confidence interval for the probability in (11) can also be readily given by a straightforward application of the delta method.

In Figures 4 and 5, the conditional survival probabilities of two right censored individuals from the WHAS data are shown corresponding to models $\mathcal{M}_1$ and $\mathcal{M}_{16}$ (ie, the two best candidate models obtained by the two methods of cut-points selection). The corresponding 95% point-wise confidence intervals, calculated by using the delta method, are also given.

It is also possible to predict the number of occurrences of the event of interest in a future time interval.[27] Suppose there are n_C right censored subjects, and we are interested in predicting the number of occurrences of the event of interest in a specific interval of width Δt after τ_R, that is, within the time interval $[\tau_R, \tau_R + \Delta t]$.

If t_{rl} is the right censoring time for the lth subject and $t_{fl} = t_{rl} + \Delta t$, $l = 1, \ldots, n_C$, then K, the number of occurrences of the failure event in the future interval $[\tau_R, \tau_R + \Delta t]$ is given by

$$K = \sum_{l=1}^{n_C} K_l,$$

where the random variable K_l indicates whether the lth right censored subject experiences the event in the interval $[\tau_R, \tau_R + \Delta t]$ or not. That is,

$$K_l \sim \text{Bernoulli}(\rho_l),$$

with

$$\rho_l = \frac{P(t_{rl} < T_l < t_{fl})}{P(T_l > t_{rl})} = \frac{F(t_{fl}; \boldsymbol{\theta}) - F(t_{rl}; \boldsymbol{\theta})}{1 - F(t_{rl}; \boldsymbol{\theta})}. \tag{12}$$

Clearly, the random variable K is a sum of independent but non-identical Bernoulli random variables K_l with ρ_l as the corresponding success probability, $l = 1, \ldots, n_C$; K ranges from 0 to n_C. The distribution of K is known as the Poisson-binomial distribution; Hong[28] and Zhang et al[29] described methods for the efficient computation of the CDF of the Poisson-binomial distribution. The MLE $\widehat{\rho}_l$ of ρ_l can be obtained by plugging in the MLE $\widehat{\boldsymbol{\theta}}$ in (12) for subsequent calculations.

8 | CONCLUSIONS

In this work, we have presented a data-driven flexible model for the analysis of LTRC survival data. The model is based on approximating the cumulative baseline hazard of the PH model by a PLA. Likelihood inference for the model is then discussed, and through a detailed simulation study, the quality of the model fit provided by the proposed model has been demonstrated. In the simulation study, different scenarios including Weibull distributions with monotone (decreasing as well as increasing) hazard rates, a PH model with non-monotone hazard rate, and a lifetime model with baseline hazard rate as a piecewise constant function, have been considered for generating the LTRC data, in order to assess the robustness and sensitivity of the proposed PLA-based model. It has been shown that the proposed PLA-based model performs quite well for a wide variety of LTRC data generated from different processes, thus demonstrating the efficacy of the model proposed here. In fact, the PLA-based model can also be easily extended to accommodate data with infant failures. If the underlying lifetime distribution is such that there is a significant mass at zero, thus producing a large number of occurrences of the failure event with lifetime close to zero, one can fit a PLA-based model to the data with a constraint on a_1 (ie, the intercept of the first linear piece) to be greater than zero. Then, the value of a_1 can be estimated from the data. The likelihood estimation and related issues that have been discussed here are applicable for this case as well with some suitable adjustments.

ACKNOWLEDGEMENTS

The authors are grateful to the three anonymous reviewers and the editors for their many constructive comments and suggestion that have helped us in improving the earlier version of the manuscript.

FUNDING INFORMATION

The research of Ayon Ganguly is supported by the Mathematical Research Impact Centric Support (MATRICS; File No. MTR/2017/000700) from the Science and Engineering Research Board (SERB), Government of India. The research of Debanjan Mitra is supported by the Mathematical Research Impact Centric Support (MATRICS; File No.

MTR/2021/000533) from the Science and Engineering Research Board (SERB), Government of India. The research of N. Balakrishnan is supported by the Natural Sciences and Engineering Research Council of Canada through an Individual Discovery Grant.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The real datasets used in this article are publicly available.

ORCID

Debanjan Mitra https://orcid.org/0000-0003-3705-3141

REFERENCES

1. Klein J, Moeschberger M. *Survival Analysis: Techniques for Censored and Truncated Data*. 2nd ed. New York: Springer; 2003.
2. Hong Y, Meeker W, McCalley J. Prediction of remaining life of power transformers based on left truncated and right censored lifetime data. *Ann Appl Stat*. 2009;3:857-879.
3. Balakrishnan N, Mitra D. EM-based likelihood inference for some lifetime distributions based on left truncated and right censored data and associated model discrimination (with discussions). *S Afr Stat J*. 2014;48:125-204.
4. Mitra D, Kundu D, Balakrishnan N. Likelihood analysis and stochastic EM algorithm for left truncated right censored data and associated model selection from the Lehmann family of life distributions. *Jpn J Stat Data Sci*. 2021;4:1019-1048.
5. Emura T, Michimae H. Left-truncated and right-censored field failure data: review of parametric analysis for reliability. *Qual Reliab Eng Int*. 2022;38:3919-3934.
6. Pak D, Liu J, Ning J, Melis G, Shen Y. Analyzing left-truncated and right-censored infectious disease cohort data with interval-censored infection onset. *Stat Med*. 2021;40:287-298.
7. Chen LP, Yi GY. Semiparametric methods for left-truncated and right-censored survival data with covariate measurement error. *Ann Inst Stat Math*. 2021;73:481-517.
8. Chen LP, Yi GY. Model selection and model averaging for analysis of truncated and censored data with measurement error. *Electron J Stat*. 2020;14:4054-4109.
9. Chen LP. Pseudo likelihood estimation for the additive hazards model with data subject to left-truncation and right-censoring. *Stat Interface*. 2019;12:135-148.
10. Huang CY, Qin J. Semiparametric estimation for the additive hazards model with left-truncated and right-censored data. *Biometrika*. 2013;100:877-888.
11. Huang CY, Ning J, Qin J. Semiparametric likelihood inference for left-truncated and right-censored data. *Biostatistics*. 2015;16:785-798.
12. Su YR, Wang JL. Modeling left-truncated and right-censored survival data with longitudinal covariates. *Ann Stat*. 2012;40:1465-1488.
13. Cox DR. Tests of separate families of hypotheses. *Proceedings 4th Berkeley Symposium in Mathematical Statistics and Probability*. Vol 1. Berkeley, CA: University of California Press; 1961.
14. Cox DR. Further results on tests of separate families of hypotheses. *J R Stat Soc Series B Stat Methodol*. 1962;24:406-424.
15. Houwelingen H, Stijnen T. Cox regression model. In: Klein J, Houwelingen H, Ibrahim J, Scheike T, eds. *Handbook of Survival Analysis*. Boca Raton, FL: CRC Press; 2014.
16. Zhang X, Zhang M, Fine J. A proportional hazards regression model for the subdistribution with right-censored and left-truncated competing risks data. *Stat Med*. 2011;30:1933-1951.
17. Cortese G, Holmboe S, Scheike T. Regression models for the restricted residual mean life for right-censored and left-truncated data. *Stat Med*. 2017;36:1803-1822.
18. Friedman M. Piecewise exponential models for survival data with covariates. *Ann Stat*. 1982;10:101-113.
19. Broström G. eha: event history analysis. R package version 2.11.1; 2023.
20. Balakrishnan N, Koutras M, Milienos F, Pal S. Piecewise linear approximations for cure rate models and associated inferential issues. *Methodol Comput Appl Probab*. 2016;18:937-966.
21. Lin D. On the Breslow estimator. *Lifetime Data Anal*. 2007;13:471-480.
22. Hosmer D, Lemeshaw S, May S. *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. 2nd ed. New York: Wiley; 2008.
23. Hyde J. Testing survival with incomplete observations. In: Miller R, Efron B, Brown B, Moses L, eds. *Biostatistics Casebook*. New York: Wiley; 1980:31-46.
24. Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Contr*. 1974;19:716-723.
25. Bain L. Analysis for the linear failure rate distribution. *Dent Tech*. 1974;16:551-559.
26. Balakrishnan N, Malik HJ. Order statistics from the linear-exponential distribution, part I: increasing hazard rate case. *Commun Stat Theory Methods*. 1986;15:179-203.

27. Escobar L, Meeker W. Statistical prediction based on censored life data. *Dent Tech*. 1999;41:113-124.
28. Hong Y. On computing the distribution function for the Poisson binomial distribution. *Comput Stat Data Anal*. 2013;59:41-51.
29. Zhang M, Hong Y, Balakrishnan N. The generalized Poisson-binomial distribution and the computation of its distribution function. *J Stat Comput Simul*. 2018;88:1515-1527.

How to cite this article: Ganguly A, Mitra D, Balakrishnan N, Kundu D. A flexible model based on piecewise linear approximation for the analysis of left truncated right censored data with covariates, and applications to Worcester Heart Attack Study data and Channing House data. *Statistics in Medicine*. 2024;43(2):233-255. doi: 10.1002/sim.9954