# Analysis of Left Truncated Right Censored Data with Covariates via a Flexible Model based on Piecewise Linear Approximation

Ayon Ganguly [*], Debanjan Mitra [†], Narayanaswamy Balakrishnan [‡] and Debasis Kundu [§]

**Abstract**

Left truncated right censored (LTRC) data quite commonly arise from survival and reliability studies. In this article, a model based on piecewise linear approximation is proposed for the analysis of left truncated right censored data. Specifically, the model involves a piecewise linear approximation of 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 a detailed simulation study, the robustness property of the model is studied by fitting it to left truncated right censored data generated from different processes covering a wide range of lifetime distributions; it is observed that the performance of the model is quite satisfactory in all those cases. Analyses of two LTRC datasets by using the model are provided as illustrative examples. Applications of the model in some practical prediction issues are discussed, including the case of predicting failures at a future interval. In summary, the proposed model provides a comprehensive and flexible approach to model a general structure of LTRC lifetime data.

Key Words and Phrases: Proportional hazards model; Cumulative hazard function; Piecewise linear approximation; Maximum likelihood estimators; Left truncation; Right censoring; Prediction.

## 1 Introduction

Left truncated right censored (LTRC) data quite commonly arise from survival and reliability experiments. Lifetime data, when collected under practical time constraints, may lead to

[*]Department of Mathematics, Indian Institute of Technology Guwahati, Assam 781039, India

[†]Quantitative Methods Division, Indian Institute of Management Udaipur, Rajasthan 313001, India. Corresponding Author. Email: debanjan.mitra@iimu.ac.in; Telephone: +91-294-2477130 (Office)

[‡]Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario L8S 4K1, Canada

[§]Department of Mathematics and Statistics, Indian Institute of Technology Kanpur, Uttar Pradesh 208016, India.

LTRC structure. For example, LTRC data may be obtained from a lifetime experiment with an observation window (i.e., a window for data collection) between two fixed time points, where failures occurring outside the observation window cannot be observed. 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 before this point would not be known. The right censoring point corresponds to the end of the study; failures of the right censored units 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, the starting points of Units 1 and 4 are before the left truncation point; therefore, to enter the study's observation window, both these units have to survive through a certain amount of time which we refer to as their left truncation times. Such units are said to be left truncated units. On the other hand, Units 2 and 3 start after the beginning of the study; that is, there is no threshold that the lifetimes of these units need to cross in order to enter the observation window of the study. Such units are referred to as untruncated. Units 3 and 4 are right censored as they do not fail by the end of the study. This structure of LTRC data that we consider here is quite general wherein a unit can be left truncated and right censored, left truncated and uncensored, untruncated and right censored, and untruncated and uncensored. In this manner, 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 [15] gave details of some of the methods that are used to analyze LTRC data. Among more recent works, Hong et al. [13] presented a detailed study of parametric analyses of LTRC data pertaining to lifetimes of power transformers by using the log-location-scale family of distributions. Using the data structure of Hong et al., Balakrishnan and Mitra [4] developed the steps of the EM algorithm for different members of the generalized gamma family of distributions; see also Mitra et al. [17] in this regard wherein the stochastic EM algorithm was

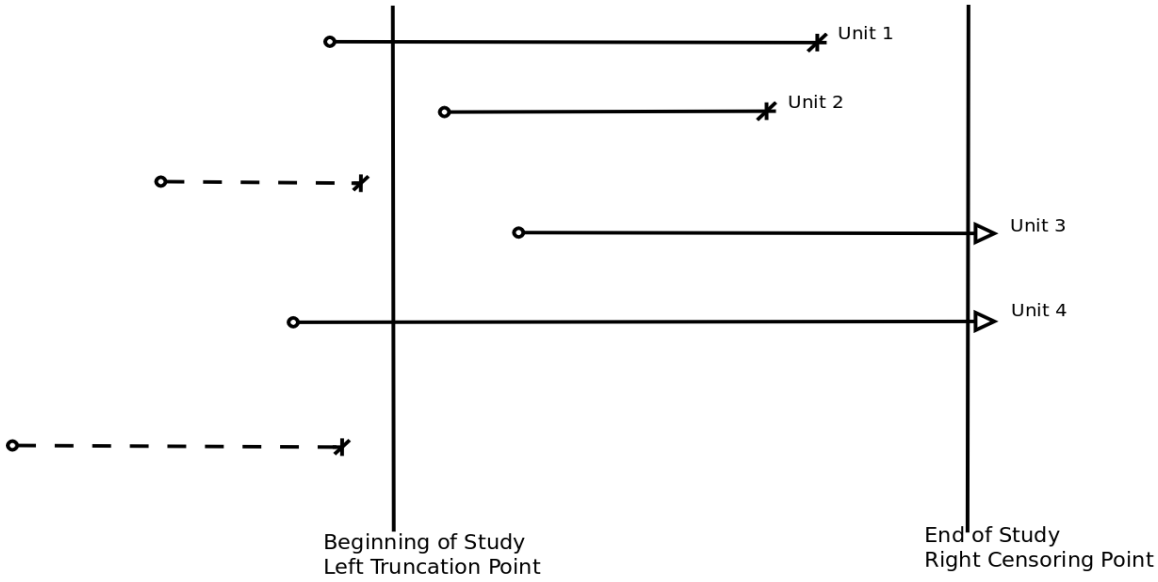


**Figure 1:** All types of units in LTRC data, with different combinations of left truncation and right censoring; the units depicted by dashed lines are not observed at all

used to model LTRC data with the use of the Lehmann family of distributions. Very recently, Emura and Michimae [9] have conducted a comprehensive review of existing parametric models and methods for LTRC lifetime data.

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 [7, 8] 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 modelled 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 $\boldsymbol{\beta}$ [7, 8], 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 [18].

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 modelling 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 is of interest to mention here that a piecewise linear approximation (PLA) for the hazard function has been used recently by Balakrishnan et al. [5] 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 funtion that is quite close to the estimate obtained by using the well-known Breslow estimator [16]. Moreover, the PLA-based approach has the advantage over the Breslow estimator for making predictions of future failures. Being a fully nonparametric procedure, the Breslow estimator is constructed over the observed range of data, and it cannot be used for prediction of failures 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 wider scope, and also computationally convenient.

The rest of the paper is organized as follows. In Section 2, the structure of LTRC data is described. Two motivating LTRC data sets are presented as examples in Section 3. The proposed PH model, involving a PLA for the cumulative baseline hazard, and the associated likelihood inference for the model based on LTRC data are discussed in Section 4. 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 real 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 lifetime experiment with $\psi_L$ and $\psi_R$, $\psi_L < \psi_R$, as the starting and end points of data collection, respectively. Then, the interval $(\psi_L, \psi_R)$ constitutes the observation window; a failure of a unit in the study is observable only if it occurs within $(\psi_L, \psi_R)$. Starting points of individual experimental units may lie before or after $\psi_L$. In case a unit started before $\psi_L$, it has to survive a certain amount of time to enter the observation window of the experiment, otherwise the unit would not be part of the study at all. A unit that started before $\psi_L$ and is part of the experiment, is said to be a left truncated unit. If a unit started after $\psi_L$, it is said to be an untruncated unit. Let $\psi_B$ denote the starting point of a unit, and $\kappa = \psi_L - \psi_B$. The left truncation time $\tau_L$ is defined as

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

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

If a unit in the experiment fails before $\psi_R$, it is an observed failure, otherwise, it is a right censored lifetime. The right censoring time of a unit 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 the unit will be

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

We introduce a censoring indicator, $\delta$, that takes the value 1 if a unit is an observed failure and 0 if the unit is right censored. Summarizing the above discussion, the available LTRC data is of the form $(Y_i, \tau_{Li}, \delta_i, \nu_i, \boldsymbol{x}_i)$, $i = 1, ..., n$, where $\boldsymbol{x}_i$ denotes the vector of covariates associated with the $i$-th unit in the life-testing 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. [12]. 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, Massachusettes 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 11000 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. [12].

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 etc. Medical investigations for 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 date of admission to hospital, and (b) Length of hospital stay, which is the number of days between

date of discharge from hospital and date of admission to hospital.

Very recently, Chen and Yi [6] 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 hospital 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, the questions of interest could be 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. [12], and Chen and Yi [6].

## 3.2 Channing House data

The Channing House dataset [14] contains lifetime information pertaining to the residents of a retirement centre, called the Channing House, in Palo Alto, California. The data collection started from the time of establishment of the centre 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 centre.

For each individual, the age at the time of entering the centre, and the age of death or leaving the centre were recorded. During 1964 to 1975, 130 women and 46 men died at the centre. A minimum age of 60 was a condition for admittance to the centre, and individuals entered the centre at different ages after 60. As the data collection started in 1964, no information on lifetimes of individuals were available before this point. 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 (i.e., 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 centre, the data also contained 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 survival rates of male and female residents admitted to Channing House. We refer to Klein and Moeschberger [15] 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 units, be $y_1, ..., y_n$. Let $\xi = \{t_0^*, ..., t_N^*\}$ denote a set of $N+1$ cut-points on the time scale such that $t_0^* < t_1^* < ... < t_N^*$, with $t_0^* \leq \text{Min}(y_{min}, \tau_{Lmin})$ and $t_N^* \leq y_{max}$, where $y_{max} = \text{Max}(y_1, ..., y_n)$, $y_{min} = \text{Min}(y_1, ..., y_n)$ and $\tau_{Lmin} = \text{Min}(\tau_{L1}, ..., \tau_{Ln})$. We consider $\xi$ to be fixed and known at this stage. Later on, we will discuss how we could suitably choose $\xi$. 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 cut-points of the set $\xi$, with additional general conditions that $H_0(t)$ is monotone increasing and continuous. That is,

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

where

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

for $k = 1, ..., N$, with $b_k > 0$, for all values of $k$. To account for the condition of monotonicity and continuity of $H_0(t)$, we have

$$H_0(t_0^*) = 0, \quad \lim_{t \to \infty} H_0(t) \to \infty, \tag{4}$$

and

$$H_0(t_{k-1}^*) = H_0(t_k^*), \quad k = 1, ..., N. \tag{5}$$

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

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

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:

$$\begin{aligned} a_k &= \sum_{i=1}^{k-1}(b_i - b_{i+1})t_i^* \\ &= a_{k-1} + (b_{k-1}b_k)t_{k-1}^*, \quad k = 2, ..., N, \end{aligned} \tag{8}$$

with $a_1 = 0$.

### 4.1.1 Choosing the cut-points

The cut-points need to 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 related 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 estmates 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 and position of the cut-points 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.

## 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

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

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

The contribution to the likelihood by units of different types are as follows:

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

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

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

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

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

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

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

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

Combining all these cases, the general likelihood for LTRC 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

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

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 second derivatives of the log-likelihood function in Eq.(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 be constructed easily using these results; for example, for one of the regression parameters $\beta_1$, 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 study 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 $\lambda$, $\gamma$ and $\beta$ 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 reliability and life-testing studies; the linear-exponential distribution is an example of this kind. One may refer to the works of Bain [2] and Balakrishnan and Malik [3], 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:

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

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

where $\xi_k$ and $\zeta_k$ are the minimum and maximum of the observed lifetime data for the $k$-th Monte Carlo sample, $\widehat{S}_{PLA}(\cdot)$, $\widehat{H}_{PLA}(\cdot)$, and $\widehat{S}_{TGP}(\cdot)$, $\widehat{H}_{TGP}(\cdot)$ are the estimated survival and cumulative hazard functions by using the PLA-based model, and the true generating process, respectively. 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 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 fitted different versions of the PLA-based model with four cut-points (i.e., three linear pieces). The first cut-point is always set at 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 $t_0^* = \text{Min}(y_{min}, \tau_{Lmin})$ (as mentioned in Section 4) and $q_\alpha$ represents the $\alpha$-th quantile of the data. The simulations were carried out in `R` software, using parellel 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 the PLA-based 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 approximation gets better for larger sample sizes. The values of the measure $AIE_{CHF}$, although show a close approximation of the underlying process by the PLA-based model, do behave somewhat differently. It may be observed that $AIE_{CHF}$, for a given level of truncation, decreases with increase in censoring percentage. This apparently anomalous behaviour of $AIE_{CHF}$ can be attributed to the cumulative hazard function being unbounded above.

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, 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. It is also seen that while the behaviour of $AIE_{SF}$ is along the expected lines (like the cases discussed above), $AIE_{CHF}$ behaves differently. For a given level of truncation, it decreases with

**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 |

an increase in censoring percentage. Also, it is more heavily affected by truncation, increasing steeply with truncation percentage.

Choice of cut-points for the PLA does not seem to play a key role in any of the above

**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 |

scenarios considered. As long as the cut-points are at interior positions (apart from the first cut-point at $t_0^*$), implying that there are enough datapoints available at each segment of the PLA, the approximation of the underlying cumulative hazard by the PLA seems to work quite

**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 |

reasonably 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

**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 |

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.

# 6 Illustrative data analysis

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

## 6.1 Analysis of Worcester Heart Attack Study data

We analyse the WHAS sample that was analysed by Chen and Yi [6]. Note that by the 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 sample, with almost 62% censoring. Before analysis, without any loss of generality, we transform the time scale of the sample according to the formula $l^* = \frac{l}{1000}$ (i.e., by dividing all the lifetimes by the constant 1000). These transformations will not affect the inferences based on this sample in any way.

We fit the PLA-based model using three cut-points to this WHAS sample. 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 [6]. The method of model fitting would be similar if other covariates are used. The first cut-point is placed at $t_0^*$, while the other two cut-points are placed at different sample quantiles. The value of $t_0^*$ and the sample quantiles of the transformed data are as follows: $t_0^* = 0.0$, $q_{0.15} = 0.1690$, $q_{0.20} = 0.3446$, $q_{0.80} = 1.5774$, and $q_{0.85} = 1.9191$.

Note that when there are $k$ cut-points, i.e., $k$ 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 corresponding to the covariates BMI and HR. 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 5 gives the estimates of slopes of the linear pieces and the regression parameters for different choices of cut-points. In Figure 2, the fitted cumulative baseline hazard is plotted against the corresponding Breslow estimator for the purpose of comparison.

There are three subjects in the WHAS sample with very high survival times. Due to

**Table 5:** Estimates of the parameters for Worcester Heart Attack data-set for different choices of cut points

| Model | Cut Points | $b_1$ | $b_2$ | $b_3$ | $\beta_1$ | $\beta_2$ | AIC |
|---|---|---|---|---|---|---|---|
| $\mathcal{M}_{3,1}$ | $t_0^*$, $q_{0.15}$, $q_{0.85}$ | 2.193 | 0.683 | 1.358 | -0.091 | 0.017 | 555.454 |
| $\mathcal{M}_{3,2}$ | $t_0^*$, $q_{0.15}$, $q_{0.80}$ | 2.169 | 0.707 | 0.600 | -0.091 | 0.017 | 557.396 |
| $\mathcal{M}_{3,3}$ | $t_0^*$, $q_{0.20}$, $q_{0.85}$ | 1.598 | 0.678 | 1.420 | -0.093 | 0.017 | 573.716 |
| $\mathcal{M}_{3,4}$ | $t_0^*$, $q_{0.20}$, $q_{0.80}$ | 1.580 | 0.708 | 0.626 | -0.093 | 0.017 | 576.060 |

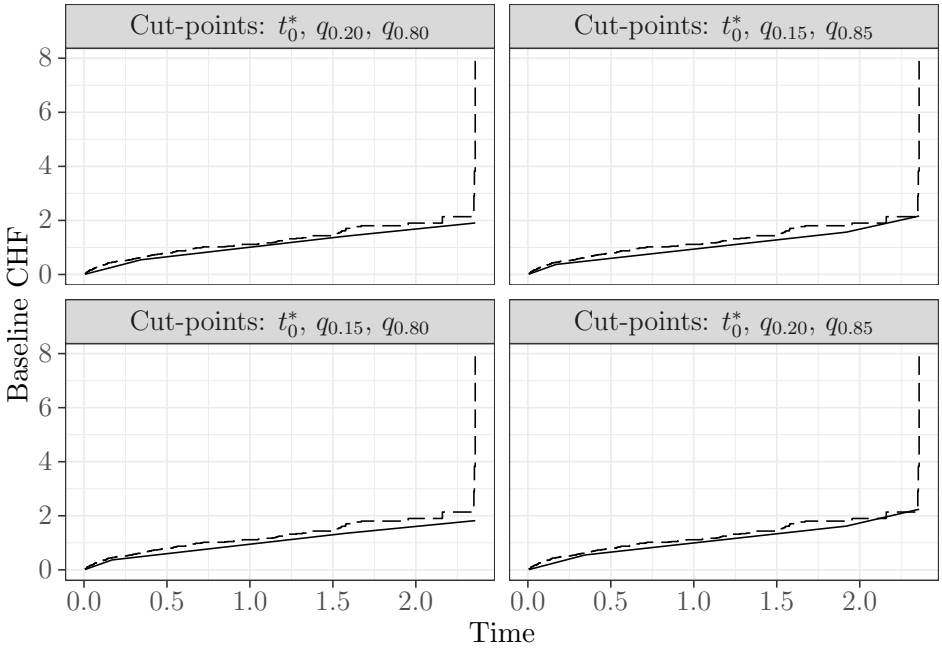


**Figure 2:** Plot of the estimated baseline CHF with three cut points for the WHAS sample; the step function corresponds to the Breslow estimator.

their presence, the Breslow estimator increases very sharply near the right end; see Figure 2. Due to the same reason, the proposed PLA-based estimator of the baseline cumulative hazard underestimates the Breslow estimator.

Among the four PLA-based models considered, the model $\mathcal{M}_{3,1}$ with cut-points placed at $t_0^*$, $q_{0.15}$, and $q_{0.85}$ provides the best fit for this data as it has the lowest value of the Akaike's information criterion (AIC) [1]. For the best model $\mathcal{M}_{3,1}$, 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 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 his/her life (naturally due to old age). After some elapsed time since discharge from hospital, the subject is at a comparatively lower risk of death. This clearly shows that the 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 centre at the end of the study. Out of the 462 total subjects, there were 4 subjects who left the retirement centre immediately after admittance (lifetime = 0). In our analyses, we use information on 458 subjects of the centre, removing these 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 centre 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 the transformed lifetime data, we have $q_{0.2} = 2.070$, $q_{0.3} = 2.300$, $q_{0.5} = 2.705$, $q_{0.7} = 3.000$, and $q_{0.8} = 3.226$. To this data, we fit the proposed PLA-based model. The first cut-point is placed at $t_0^* = 0.13$. The other cut-points are placed at various quantiles of the data. In this process, we naturally avoided choosing the cut-points near the tails of the data, i.e., 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). Table 6 gives the estimates of the slopes of the linear pieces and the regression parameter for different choices of cut-points. In Figures 3 - 5, the fitted cumulative baseline hazard is plotted against the corresponding Breslow estimator for the purpose of comparison.

**Table 6:** Estimates of the parameters for different choices of cut points

| Model | Cut Points | $b_1$ | $b_2$ | $b_3$ | $b_4$ | $\beta$ | AIC |
|---|---|---|---|---|---|---|---|
| $\mathcal{M}_{4,1}$ | $t_0^*$, $q_{0.2}$, $q_{0.5}$, $q_{0.7}$ | 0.299 | 0.427 | 1.061 | 1.364 | -0.310 | 549.788 |
| $\mathcal{M}_{4,2}$ | $t_0^*$, $q_{0.2}$, $q_{0.5}$, $q_{0.8}$ | 0.300 | 0.428 | 1.137 | 1.414 | -0.313 | 550.117 |
| $\mathcal{M}_{4,3}$ | $t_0^*$, $q_{0.3}$, $q_{0.5}$, $q_{0.7}$ | 0.329 | 0.428 | 1.062 | 1.365 | -0.311 | 551.017 |
| $\mathcal{M}_{4,4}$ | $t_0^*$, $q_{0.3}$, $q_{0.5}$, $q_{0.8}$ | 0.330 | 0.429 | 1.138 | 1.415 | -0.313 | 551.346 |
| $\mathcal{M}_{3,1}$ | $t_0^*$, $q_{0.2}$, $q_{0.5}$ | 0.299 | 0.427 | 1.239 | – | -0.311 | 549.280 |
| $\mathcal{M}_{3,2}$ | $t_0^*$, $q_{0.3}$, $q_{0.5}$ | 0.329 | 0.429 | 1.240 | – | -0.312 | 550.509 |
| $\mathcal{M}_{3,3}$ | $t_0^*$, $q_{0.5}$, $q_{0.7}$ | 0.358 | 1.063 | 1.367 | – | -0.312 | 550.157 |
| $\mathcal{M}_{3,4}$ | $t_0^*$, $q_{0.5}$, $q_{0.8}$ | 0.359 | 1.139 | 1.417 | – | -0.316 | 550.486 |
| $\mathcal{M}_{2,1}$ | $t_0^*$, $q_{0.5}$ | 0.359 | 1.242 | – | – | -0.314 | 549.649 |

We notice that the model $\mathcal{M}_{3,1}$, with three cut-points placed at $t_0^*$, $q_{0.2}$ and $q_{0.5}$, turn 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, within the class of the four cut-point models, the two models that include $t_0^*$, $q_{0.2}$ and $q_{0.5}$ along with a third percentile, namely, $\mathcal{M}_{4,1}$ and $\mathcal{M}_{4,2}$, produce lower values of AIC compared to the other models of this class. But clearly, from the point of model sufficiency in terms of AIC, a four cut-point model is not necessary for this data. For model $\mathcal{M}_{3,1}$, the estimated regression coefficient $\beta$ 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.

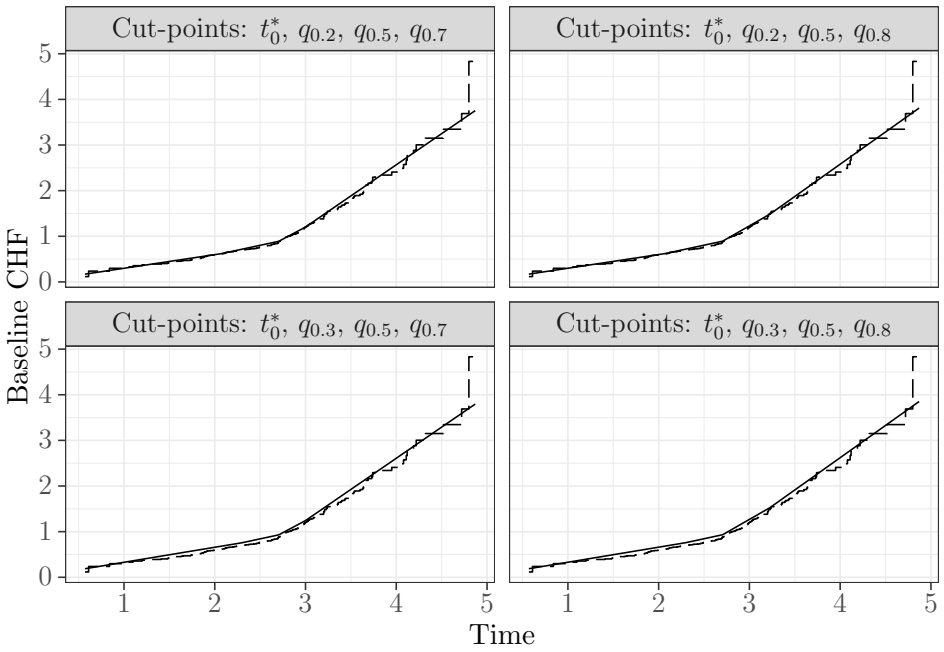


**Figure 3:** Plot of the estimated baseline CHF with four cut points for the Channing House data; the step function corresponds to the Breslow estimator.

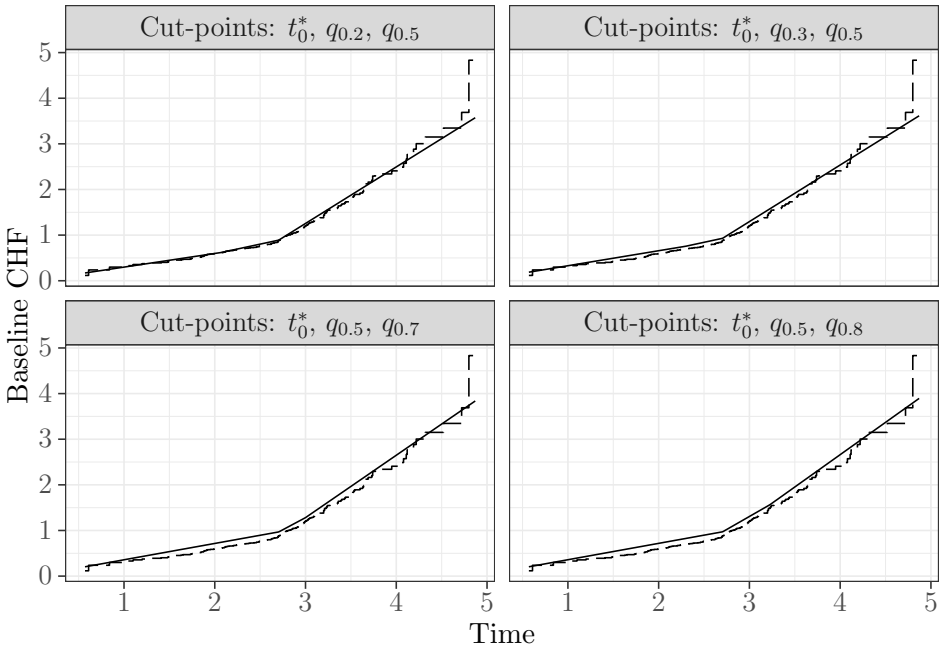


**Figure 4:** Plot of the estimated baseline CHF with three cut points for the Channing House data; the step function corresponds to the Breslow estimator.

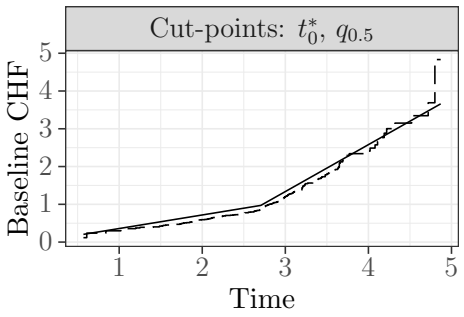


**Figure 5:** Plot of the estimated baseline CHF with two cut points for the Channing House data; the step function corresponds to the Breslow estimator.

# 7 Prediction issues

We would like to reinforce one of the advantages of using the PLA-based model here. The PLA model is essentially a parametric modelling approach wherein the unknown parameters are estimated based on data. However, it does not come with 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 used for predicting failures beyond the observed range of data. Nonparametric procedures like the Breslow estimator for baseline CHF are valid only for the observed range of data, and they cannot be extrapolated outside such ranges. This is a unique advantage of using the PLA-based model over the fully nonparametric procedures.

In lifetime experiments, some prediction issues are of natural interest. For example, one may want to predict the probability of failure of a right censored unit at a future time point, or the number of failures in a future interval. Prediction of these quantities have immense practical implications.

## 7.1 Predicting the probability of a future failure

Conisder a unit that is right censored at $\tau_R$ with right censoring time $t_R$, regardless of its truncation status. Suppose we are then interested in estimating the probability of failure of the unit by a future time point $t_f$, where $t_f > t_R$.

Using the MLE $\widehat{\boldsymbol{\theta}}$ computed from the data, the required conditional probability may be

estimated as

$$\begin{aligned}
\pi &= P(t_R < T < t_f | T > t_R) \\
&= \frac{F(t_f; \widehat{\boldsymbol{\theta}}) - F(t_R; \widehat{\boldsymbol{\theta}})}{1 - F(t_R; \widehat{\boldsymbol{\theta}})} \\
&= \frac{\exp\{-H(t_R; \widehat{\boldsymbol{\theta}})\} - \exp\{-H(t_f; \widehat{\boldsymbol{\theta}})\}}{\exp\{-H(t_R; \widehat{\boldsymbol{\theta}})\}} \\
&= \frac{\exp\{-H_0(t_R)\exp(\widehat{\boldsymbol{\beta}}'\boldsymbol{x})\} - \exp\{-H_0(t_f)\exp(\widehat{\boldsymbol{\beta}}'\boldsymbol{x})\}}{\exp\{-H_0(t_R)\exp(\widehat{\boldsymbol{\beta}}'\boldsymbol{x})\}},
\end{aligned} \tag{11}$$

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.

## 7.2 Predicting the number of failures in a future interval

Following the procedure of Escobar and Meeker [10], one can give a calibrated prediction interval for the number of failures in a future time interval. Let there be $n_C$ right censored units. We are then interested in predicting the number of failures in an interval of width $\Delta t$ after $\tau_R$, i.e., within the time interval $[\tau_R, \tau_R + \Delta t]$.

Suppose the right censoring time for the $l$-th unit is $t_{rl}$, and $t_{fl} = t_{rl} + \Delta t$, $l = 1, ..., n_C$. If we denote the number of failures in the future interval $[\tau_R, \tau_R + \Delta t]$ by $K$, then

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

where the random variable $K_l$ indicates whether the $l$-th right censored unit fails in the interval $[\tau_R, \tau_R + \Delta t]$ or not. That is,

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

with

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

Clearly, the random variable $K$ is a sum of independent but non-identical Bernoulli random variables $K_l$ with $\rho_l$ as the corresponding success probability, $l = 1, ..., n_C$; $K$ ranges from 0 to $n_C$. The distribution of $K$ is known as the Poisson-binomial distribution; Hong [11] and Zhang et al. [19] described methods for efficient computation of the CDF of the Poisson-binomial distribution.

To provide a prediction interval for the number of failures in the future interval $[\tau_R, \tau_R + \Delta t]$ and to calibrate it, it is necessary to compute the CDF of the Poisson-binomial distribution presented above. A Monte Carlo approach for the same following Escobar and Meeker [10] may also be used, as described below.

Denote the CDF of the Poisson-binomial distribution that is obtained as a sum of $n_C$ independent and non-identical Bernoulli random variables by $F_{PB}(\cdot; \mathbf{1}, \boldsymbol{\rho})$, where $\mathbf{1} = (1, ..., 1)$ and $\boldsymbol{\rho} = (\rho_1, ..., \rho_{n_C})$. Then, $P(K \leq k) = F_{PB}(k; \mathbf{1}, \boldsymbol{\rho})$, $k = 0, ..., n_C$. The $100(1-\alpha)\%$ upper prediction bound for the number of failures in the future interval $[\tau_R, \tau_R + \Delta t]$ is then given by $\widehat{K}_{1-\alpha}$ which is the smallest integer $k$ that satisfies $F_{PB}(k; \mathbf{1}, \widehat{\boldsymbol{\rho}}) \geq 1 - \alpha$.

However, $\widehat{K}_{1-\alpha}$ would be a naive upper prediction bound for the number of failures, as the calculation of it does not take into consideration the uncertainty involved in $\widehat{\boldsymbol{\rho}}$ which is obtained by using the MLE $\widehat{\boldsymbol{\theta}}$; as a result, the confidence coefficient attached to this prediction bound would be actually less than $1 - \alpha$. Hence, the need to calibrate the prediction bound to achieve the desired level of confidence coefficient.

For calibration, the idea is to find the confidence level $1 - \alpha^*$ such that

$$P(K \leq \widehat{K}_{1-\alpha^*}; \widehat{\boldsymbol{\rho}}) = 1 - \alpha.$$

To find $\widehat{K}_{1-\alpha^*}$, use the empirical distribution function of $K$, as follows:

ALGORITHM:

STEP 1: Generate $u_{sl}$ from Bernoulli($\widehat{\rho}_l$), $l = 1, ..., n_C$;

STEP 2: Calculate $u_s = \sum_{l=1}^{n_C} U_{jl}$;

STEP 3: Repeat Steps 1 and 2 $M$ times, to obtain $u_1, ..., u_M$;

STEP 4: From the empirical distribution function of $u_1, ..., u_M$, denoted by $\widehat{F}_{Emp}(\cdot)$, find $u^*$ such that $\widehat{F}_{Emp}(u^*) \geq 1 - \alpha$;

STEP 5: Then, take $\widehat{K}_{1-\alpha^*} = u^*$.

To obtain a calibration curve, the above algorithm can be repeated for different values of $\alpha$.

# 8 CONCLUSIONS

In this work, we have presented a data-driven flexible model for the analysis of LTRC 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 such as Weibull distributions with monotone (decreasing as well as increasing) hazard rates and a PH model with non-monotone hazard rate have been considered for generating the LTRC data. 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 numerical zero thus producing a large number of observed failures with lifetime close to zero, one can fit a PLA-based model to the data with a constraint on $a_1$ (i.e., 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 minor adjustments.

## Acknowledgements

- 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.

## Declaration

Declarations of interest: none.

# References

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

[2] Bain, L. J. (1974). Analysis for the linear failure rate distribution. *Technometrics* **16**, 551–559.

[3] Balakrishnan, N., and Malik, H. J. (1986). Order statistics from the linear-exponential distribution, part I: increasing hazard rate case. *Communications in Statistics - Theory and Methods* **15**, 179–203.

[4] Balakrishnan, N., and Mitra, D. (2014). EM-based likelihood inference for some lifetime distributions based on left truncated and right censored data and associated model discrimination (with discussions). *South African Statistical Journal* **48**, 125–204.

[5] Balakrishnan, N., Koutras, M. V., Milienos, F. S., and Pal, S. (2016). Piecewise linear approximations for cure rate models and associated inferential issues. *Methodology and Computing in Applied Probability*, **18**, 937–966.

[6] Chen, L., and Grace, Y. (2021). Semiparametric methods for left-truncated and right-censored survival data with covariate measurement error. *Annals of the Institute of Statistical Mathematics*, **73**, 481–517.

[7] Cox D. R. (1961). Tests of separate families of hypotheses. In: Proceedings 4th Berkeley Symposium in Mathematical Statistics and Probability, Vol. 1. University of California Press, Berkeley, CA.

[8] Cox D. R. (1962). Further results on test of separate families of hypotheses. Journal of Royal Statistical Society, Series B, **24**, 406–424.

[9] Emura, T., and Michimae, H. (2022). Left-truncated and right-censored field failure data: Review of parametric analysis for reliability. *Quality and Reliability Engineering International*, **38**, 3919–3934.

[10] Escobar, L. A., and Meeker, W. Q. (1999). Statistical prediction based on censored life data. *Technometrics*, **41**, 113–124.

[11] Hong, Y. (2013). On computing the distribution function for the Poisson binomial distribution. *Computational Statistics & Data Analysis*, **59**, 41–51.

[12] Hosmer, D. W., Lemeshaw, S., and May S. (2008). *Applied Survival Analysis: Regression modeling of time-to-event data*, 2nd Ed. Wiley, New York.

[13] Hong, Y., Meeker, W. Q., and McCalley, J. D. (2009). Prediction of remaining life of power transformers based on left truncated and right censored lifetime data. *The Annals of Applied Statistics*, **3**, 857–879.

[14] Hyde, J. (1980). Testing survival with incomplete observations. In: *Biostatistics Casebook*, Eds: R.G. Miller, B. Efron, B.W. Brown and L.E. Moses, pp.31–46. John Wiley, New York.

[15] Klein, J. P., and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*, 2nd Ed. Springer, New York.

[16] Lin, D. Y. (2007). On the Breslow estimator. *Lifetime Data Analysis*, **13**, 471–480.

[17] Mitra, D., Kundu, D., and Balakrishnan, N. (2021). Likelihood analysis and stochastic EM algorithm for left truncated right censored data and associated model selection from the Lehmann family of life distributions. *Japanese Journal of Statistics and Data Science*, **4**, 1019–1048.

[18] van Houwelingen, H. C., and Stijnen, T. (2014). Cox regression model. In: *Handbook of Survival Analysis*, Eds: Klein, J. P., van Houwelingen, H. C., Ibrahim, J. G., and Scheike, T. H. CRC Press, Boca Raton, FL.

[19] Zhang, M., Hong, Y., and Balakrishnan, N. (2018). The generalized Poisson-binomial distribution and the computation of its distribution function. *Journal of Statistical Computation and Simulation*, **88**, 1515–1527.