CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 11 (1) (2026), Pages 1318-1333
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Two-Parameter Exponential Estimation via EM Algorithm: Uterine Leiomyosarcoma Survival Risk Analysis

Ardi Kurniawan*, Deby Victoria, and Christabel Lee Angie Sugianto

Department of Mathematics, Faculty of Science and Technology, Airlangga University, Surabaya, Indonesia

Abstract

This study implements the Expectation-Maximization (EM) algorithm for parameter estimation of the two-parameter exponential distribution under right-censored survival data, applied to 122 uterine leiomyosarcoma (uLMS) patients. The model assumes a constant hazard rate, which is an inherent limitation. Using a fixed censoring threshold of 40 months, the EM algorithm converged after 26 iterations, yielding $\hat{\alpha} = 2$ months and $\hat{\beta} = 68.70$ months, corresponding to a monthly hazard rate of approximately 1.46%. A Monte Carlo simulation (1,000 replications) produced a mean simulated scale parameter of 67.96, a bias of -0.74, and a mean squared error of 93.19; the empirical coverage probability of the 95% Wald interval was 93.0%. To quantify extreme survival outcomes, Value-at-Risk (VaR) and Tail Value-at-Risk (TVaR) were derived, giving a 95% VaR of 207.82 months and a TVaR of 276.52 months. Sensitivity analysis across different censoring proportions showed estimates ranging from 74.52 to 78.51 months, indicating reasonable stability. All results are contingent on the constant hazard assumption of the exponential model.

Keywords: EM Algorithm; Right-Censored Survival Data; Two-Parameter Exponential Distribution; Uterine Leiomyosarcoma; Value-at-Risk and Tail Value-at-Risk (VaR and TVaR).

1. Introduction

Survival analysis models time-to-event data, often involving incomplete observations known as censoring, which occurs when patients are lost to follow-up or when studies end prematurely [1]. Excluding or misclassifying censored data can bias parameter estimates and reduce efficiency in survival models [2]. Therefore, methods that explicitly accommodate censored data are essential in applied survival and reliability analysis.

The exponential distribution is widely used in survival analysis for its mathematical tractability, providing a baseline model for small samples [3]. Its two-parameter extension offers a distinct advantage by introducing a location parameter, α, which represents the minimum observable failure time. This capability allows the model to flexibly accommodate delayed event onset, a feature highly plausible in medical prognosis [4]. While recent literature has explored alternative estimation techniques for this model under censored conditions, such as Bayesian estimation with linear exponential loss functions for Type I censoring [5], the model retains the constant hazard assumption. This limitation may not fully capture the time-varying survival dynamics common in oncological data, necessitating cautious interpretation of the results.

*Corresponding author. E-mail: ardi-k@fst.unair.ac.id

Submitted: April 11, 2026 Reviewed: April 24, 2026 Accepted: May 23, 2026
DOI: https://doi.org/10.18860/cauchy.v11i1.41855

Estimating the two-parameter exponential distribution under right-censoring via conventional maximum likelihood estimation (MLE) is challenging due to boundary constraints and non-negligible finite-sample bias. The Expectation-Maximization (EM) algorithm effectively overcomes these limitations by treating unobserved failure times as latent variables. It iteratively alternates between computing the expected complete-data log-likelihood (E-step) and maximizing this expectation (M-step) until convergence, ensuring a monotonic increase in the likelihood without requiring complex second-derivative evaluations [6]. This robustness in handling incomplete data extends to other complex models; for instance, the EM algorithm exhibited superior stability and estimation performance compared to classical methods like Newton-Raphson when applied to the heavy-tailed Burr Type XII distribution under Type II censoring [7].

While existing EM-based estimations for censored data have been developed in controlled or one-parameter settings [8, 9], to the best of our knowledge, the application of the EM algorithm specifically for estimating the two-parameter exponential distribution under right-censored conditions has not yet been established in the literature. Furthermore, standard parametric analyses often restrict inferential outputs to average outcomes such as point estimates and confidence intervals. These summaries may not adequately capture extreme survival risks in the distribution tails, which are highly relevant for long-term clinical prognosis and medical resource planning.

This study addresses these gaps twofold. First, it introduces the EM algorithm for the two-parameter exponential distribution using uterine leiomyosarcoma (uLMS) clinical data, alongside Monte Carlo simulations assessing estimator performance under varying sample sizes and censoring proportions. Second, it integrates Value-at-Risk (VaR) and Tail Value-at-Risk (TVaR) to quantify extreme survival thresholds. While VaR at confidence level p identifies the time by which a proportion p of patients experience the event, TVaR estimates conditional survival beyond this threshold. Integrating EM-based estimation with VaR and TVaR provides a comprehensive, tail-focused framework for clinical risk stratification. Ultimately, by advancing analytical tools for precise cancer prognosis and long-term care planning, this research inherently supports the United Nations Sustainable Development Goal (SDG) 3, which aims to ensure healthy lives and promote well-being through improved management and mitigation of non-communicable diseases.

2. Methods

This section describes the methodological framework used to estimate the parameters of the two-parameter exponential distribution under right-censored survival data. The discussion begins with the data source and variables, followed by the model specification, EM estimation procedure, simulation design, uncertainty quantification, risk forecasting metrics, and model validation strategy.

2.1. Data Source and Variables

The dataset comprises 122 uterine leiomyosarcoma (uLMS) samples obtained from cBioPortal [10]. The primary variable is Disease-Specific Survival (months), accompanied by a right-censoring indicator. The dataset initially follows the original clinical censoring mechanism; however, due to the assumption of a constant hazard rate in the two-parameter exponential model, an additional artificial censoring threshold at 40 months is imposed. This combined censoring scheme restricts the analysis to a temporal window in which the constant hazard assumption is considered appropriate. All censoring mechanisms are assumed to be independent and non-informative. Summary statistics are presented in Table 1.

Table 1: Summary statistics of survival data

Description	Value
Sample size	122
Number of events	54
Number of censored	68
Censored (%)	55.74
Mean survival time (months)	74.84
Median survival time (months)	50

2.2. Model and Estimation Procedure

This study focuses on parameter estimation of the two-parameter exponential distribution in the presence of right-censored data using the Expectation-Maximization (EM) algorithm. The methodological framework formally defines the distributional assumptions, the censoring mechanism, and the iterative estimation procedure.

2.2.1. Two-Parameter Exponential Distribution

The two-parameter exponential distribution is an extension of the standard exponential distribution obtained by incorporating a location parameter α. This distribution allows modeling situations in which events (failures) cannot occur before a specified threshold time α [11]. The probability density function (PDF) is defined as follows [12]:

$$f(t;\,\alpha,\beta) = \frac{1}{\beta}\exp\!\left(-\frac{t-\alpha}{\beta}\right), \quad t \geq \alpha,\ \beta > 0$$

Here, α represents the location parameter indicating the minimum survival time, while β is the scale parameter. While this model assumes a constant hazard rate given by $h(t) = \frac{1}{\beta}$ for $t \geq \alpha$, it provides a fundamental mathematical baseline for evaluating delayed event onset in clinical survival data before extending to more complex parametric models.

2.2.2. Censored Data Framework

Although clinical data inherently involve random individual-specific censoring [1], this study adopts a fixed Type II right-censoring framework to ensure a tractable EM-based likelihood formulation [13, 14]. For a sample of size n, ordered failure times $y_1 \leq \cdots \leq y_r$ are observed, while the remaining $n-r$ observations are treated uniformly as latent variables $z_i > y_r$. To handle these unobserved values, the EM algorithm iteratively computes their conditional expectation in the E-step as $E(z_i \mid z_i > y_r) = y_r + \beta$. It must be explicitly acknowledged that enforcing this uniform cutoff (y_r) artificially modifies the original data structure by overriding individual censoring times. While this simplification guarantees analytical consistency for the EM derivation, it causes inevitable information loss in the distribution's tail, a limitation that requires careful consideration during interpretation.

2.2.3. EM Algorithm for Parameter Estimation

The EM algorithm iteratively obtains maximum likelihood estimates for incomplete data, such as unobserved exact failure times in right-censored datasets [6]. The procedure relies on complete-data log-likelihood estimation and alternates between two main steps:

1. **Expectation step (E-step):** For the $n-r$ right-censored observations, the exact latent survival times z_i $(i = r+1, \ldots, n)$ are known only to exceed the fixed threshold y_r. Given the current estimate of the scale parameter $\beta^{(k)}$ at the k-th iteration, the expected complete survival time for the censored data is evaluated via conditional expectation:

$$E(z_i \mid z_i > y_r) = y_r + \beta^{(k)}$$

Thus, the unobserved failure times are imputed uniformly as the fixed censoring time (y_r) plus the expected residual life ($\beta^{(k)}$).

2. **Maximization step (M-step):** The expected complete-data log-likelihood function is maximized to update the parameters. However, because the partial derivative with respect to the location parameter α is strictly positive, its maximum likelihood estimate is boundary-constrained and fixed at the minimum observed failure time, $\hat{\alpha} = y_1$. Consequently, the EM iterations effectively update only the scale parameter β. The new estimate $\beta^{(k+1)}$ is obtained by maximizing the expected log-likelihood, incorporating both the r fully observed survival times and the $n - r$ imputed expected times derived from the E-step.

These E and M steps are iteratively repeated until the parameter estimates stabilize within a prescribed convergence tolerance. The detailed mathematical derivations of these estimators are explicitly presented in the Results section.

2.3. Monte Carlo Simulation Study

A Monte Carlo simulation study systematically evaluates the finite-sample performance of the EM-based parameter estimation under right-censored conditions. Simulation is a standard approach in survival analysis to assess estimator behavior—particularly Bias and Mean Squared Error (MSE), when analytical evaluation is complex [8, 15].

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables from a two-parameter exponential distribution with true parameter values (α, β). The total number of Monte Carlo replications is set to $B = 1000$, which is commonly used to obtain stable estimates of coverage probability and bias [16]. To closely mirror the empirical application, the simulation considers varying sample sizes of n (e.g., $n = 30, 50, 100$). For each replication $k = 1, 2, \ldots, B$, the following steps are executed:

1. **Step 1: Data Generation.** A random sample is generated from the two-parameter exponential distribution using the inverse transform method: $X_i = \alpha - \beta \ln(1 - U_i)$, where $U_i \sim U(0, 1)$.
2. **Step 2: Censoring Mechanism.** To align with the empirical framework applied to the clinical dataset, a fixed Type II right-censoring threshold of $y_r = 40$ months is imposed. Observations exceeding y_r are truncated and recorded as right-censored latent variables (z_i), while those below the threshold are fully observed (y_i).
3. **Step 3: Parameter Estimation.** The iterative EM algorithm detailed in Section 2.2.3 is applied to the right-censored simulated data. The expected complete-data log-likelihood is iteratively maximized to obtain the simulated scale parameter estimate $\hat{\beta}^{(k)}$ for the k-th replication.
4. **Step 4: Performance Evaluation.** The estimation accuracy for the scale parameter is formally assessed by calculating the empirical Bias and MSE across all B replications [17]:

$$\text{Bias}(\hat{\beta}) = \frac{1}{B} \sum_{k=1}^{B} (\hat{\beta}^{(k)} - \beta)$$

$$\text{MSE}(\hat{\beta}) = \frac{1}{B} \sum_{k=1}^{B} (\hat{\beta}^{(k)} - \beta)^2$$

2.4. Uncertainty Quantification

To evaluate the precision and reliability of the EM-based estimates, multiple approaches are employed for uncertainty quantification [18], allowing for a comprehensive assessment of interval estimation performance.

2.4.1. Asymptotic Variance and Fisher Information

The standard error of the scale parameter β is derived from the observed Fisher information [9]. For the exponential distribution under right-censoring, the asymptotic variance of the estimator $\hat{\beta}$ is given by:

$$\mathrm{Var}(\hat{\beta}) \approx I^{-1}(\hat{\beta}) = \frac{\hat{\beta}^2}{r}$$

where r denotes the number of uncensored observations.

Accordingly, the $100(1-\nu)\%$ Wald confidence interval is constructed as:

$$\mathrm{CI}_{\mathrm{Fisher}} = \hat{\beta} \pm z_{1-\nu/2} \frac{\hat{\beta}}{\sqrt{r}}$$

2.4.2. Non-Parametric Bootstrap

To obtain a robust estimate without relying on asymptotic normality, a non-parametric bootstrap procedure is implemented. Let $\mathbf{X}^* = \{x_1^*, x_2^*, \ldots, x_n^*\}$ be a resample of size n drawn with replacement from the original dataset. The bootstrap confidence interval is constructed from the empirical distribution of $B = 1000$ bootstrap replicates $\{\hat{\beta}_1^*, \ldots, \hat{\beta}_B^*\}$ using the percentile method:

$$\mathrm{CI}_{\mathrm{Boot}} = \left[\hat{\beta}^*_{(\nu/2)},\ \hat{\beta}^*_{(1-\nu/2)}\right]$$

2.4.3. Profile Likelihood Confidence Interval

To account for potential asymmetry in the likelihood function, particularly in small samples with substantial censoring, the profile likelihood confidence interval is also considered. Given that the location parameter is fixed at its boundary estimate $\hat{\alpha} = \min(y_i)$, the profile log-likelihood for β, denoted as $pl(\beta)$, is constructed.

The $100(1-\nu)\%$ profile likelihood confidence interval consists of all values of β satisfying:

$$2\left[\ln L(\hat{\alpha}, \hat{\beta}) - \ln L(\hat{\alpha}, \beta)\right] \leq \chi^2_{1,\,1-\nu}$$

where $\chi^2_{1,\,1-\nu}$ denotes the $(1-\nu)$-th quantile of the chi-square distribution with one degree of freedom.

2.5. Risk Forecasting Metrics in Clinical Prognosis

This study extends the traditional survival analysis framework into risk management by adapting two actuarial metrics derived from the survival function $S(t) = \exp\left(-\frac{t-\alpha}{\beta}\right)$. Although these metrics are conventionally used in financial risk assessment, they can provide intuitive and interpretable measures for clinical prognosis.

2.5.1. Value at Risk (VaR)

In a clinical context, the Value at Risk (VaR) at confidence level p (e.g., $p = 0.95$) represents a lower bound for survival time under a pessimistic scenario. Specifically, it defines the minimum time t such that the probability of experiencing the event by time t is p, or equivalently, the probability of surviving beyond this threshold is $1-p$. Formally, VaR corresponds to the p-th quantile of the survival distribution and is given by:

$$\mathrm{VaR}_p = \alpha - \beta \ln(1-p)$$

While commonly used in financial risk analysis [19], VaR is adapted in this study as an alternative metric for interpreting survival risk.

2.5.2. Tail Value at Risk (TVaR)

The Tail Value at Risk (TVaR) provides a complementary measure by capturing the expected survival time conditional on exceeding the VaR threshold. Utilizing the memoryless property of the exponential distribution, TVaR is expressed as [20]:

$$\mathrm{TVaR}_p = \mathbb{E}[X \mid X > \mathrm{VaR}_p] = \frac{\int_{\mathrm{VaR}_p}^{\infty} t f(t)\, dt}{S(\mathrm{VaR}_p)} = \mathrm{VaR}_p + \beta$$

Together, these metrics provide a bridge between statistical estimation and risk-based interpretation, offering additional perspectives for survival assessment beyond conventional summary measures.

2.6. Model Validation and Sensitivity Analysis

2.6.1. Goodness-of-Fit Assessment and Model Comparison

Goodness-of-fit is evaluated using censoring-adjusted Q-Q plots. To address bias due to right-censoring, empirical quantiles are estimated via the Kaplan–Meier method and compared with the theoretical quantiles of the fitted model $(\hat{\alpha}, \hat{\beta})$. An adequate model fit is indicated when the plotted points lie approximately along the 45-degree reference line [21].

For model comparison, the Akaike Information Criterion (AIC) based on the right-censored log-likelihood is employed. Suppose there are n observations with survival times t_i and event indicators δ_i (where $\delta_i = 1$ if the event is observed and $\delta_i = 0$ if censored). The log-likelihood function is defined as

$$\ell(\theta) = \sum_{i=1}^{n} \left[\delta_i \ln f(t_i; \theta) + (1 - \delta_i) \ln S(t_i; \theta)\right],$$

where $f(\cdot)$ and $S(\cdot)$ denote the probability density function and survival function, respectively.

The AIC is computed as

$$\mathrm{AIC} = -2\ell(\hat{\theta}) + 2k.$$

The proposed two-parameter exponential model ($k = 2$) is then compared with the standard Weibull model ($k = 2$).

2.6.2. Sensitivity Analysis

To assess the stability and robustness of the EM-based estimator under varying levels of data truncation, a sensitivity analysis is conducted by altering the proportion of fully observed events. A set of observation proportions is defined as $\mathcal{P} = \{0.70, 0.80, 0.90\}$.

For each $p \in \mathcal{P}$, the corresponding number of observed events is determined as $R = \lfloor p \cdot n \rfloor$, where n is the total sample size. The scale parameter estimate under each scenario, denoted as $\hat{\beta}_p$, is obtained by applying the EM algorithm to the truncated dataset:

$$\hat{\beta}_p = \Phi(y_1, y_2, \ldots, y_R)$$

where $\Phi(\cdot)$ represents the EM estimation mapping.

This procedure evaluates how the parameter estimates respond to increasingly heavy artificial right-censoring, providing insight into the robustness of the model under different levels of information loss.

3. Results and Discussion

This section presents the analytical derivation, empirical implementation, and validation results of the proposed EM-based estimation framework. The discussion begins with the mathematical derivation of the estimators, followed by the application to clinical survival data, simulation-based evaluation, uncertainty analysis, risk forecasting, goodness-of-fit assessment, sensitivity analysis, and methodological interpretation.

3.1. Mathematical Derivation of EM Estimators

Before analyzing the clinical dataset, it is necessary to derive the estimators for the two-parameter exponential distribution under a fixed right-censoring framework using the Expectation–Maximization (EM) algorithm. The derivation proceeds as follows.

1. Initial Parameter Estimates

Let $x_1, x_2, \ldots, x_n$ be lifetime data from a random sample of size n, assumed to follow a two-parameter exponential distribution with location parameter α and scale parameter β. The probability density function (PDF) is:

$$f(x) = \frac{1}{\beta} \exp\left(-\frac{x-\alpha}{\beta}\right), \quad x \geq \alpha,\ \beta > 0$$

The likelihood function is:

$$L(\alpha, \beta \mid \mathbf{x}) = \frac{1}{\beta^n} \exp\left(-\frac{\sum_{i=1}^{n}(x_i - \alpha)}{\beta}\right)$$

Thus, the log-likelihood function becomes:

$$\ell(\alpha, \beta) = -n \ln \beta - \frac{1}{\beta} \sum_{i=1}^{n} (x_i - \alpha)$$

Simplifying:

$$\ell(\alpha, \beta) = -n \ln \beta + \frac{n\alpha}{\beta} - \frac{1}{\beta} \sum_{i=1}^{n} x_i$$

Differentiating with respect to α:

$$\frac{\partial \ell}{\partial \alpha} = \frac{n}{\beta}$$

Since this derivative is strictly positive, the maximum likelihood estimate of α occurs at the boundary:

$$\hat{\alpha} = \min(x_i)$$

Differentiating with respect to β:

$$\frac{\partial \ell}{\partial \beta} = -\frac{n}{\beta} - \frac{1}{\beta^2} \left(n\alpha - \sum_{i=1}^{n} x_i\right)$$

Setting to zero yields:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i - n\alpha}{n} = \bar{x} - \hat{\alpha}$$

Thus, the initial estimates are:

$$\hat{\alpha}^{(0)} = \min(x_i), \quad \hat{\beta}^{(0)} = \bar{x} - \hat{\alpha}^{(0)}$$

2. Expectation Step (E-step)

Under the fixed right-censoring framework at threshold y_r, the data are partitioned into observed values y_i $(i = 1, \ldots, r)$ and censored values z_i $(i = r+1, \ldots, n)$. The log-likelihood becomes:

$$\ell(\alpha, \beta) = -n \ln \beta + \frac{n\alpha}{\beta} - \frac{1}{\beta} \left(\sum_{i=1}^{r} y_i + \sum_{i=r+1}^{n} z_i\right)$$

Taking expectation with respect to the missing data:

$$\mathbb{E}[\ell(\alpha,\beta)] = -n\ln\beta + \frac{n\alpha}{\beta} - \frac{1}{\beta}\left(\sum_{i=1}^{r} y_i + \sum_{i=r+1}^{n} \mathbb{E}(z_i \mid z_i > y_r)\right)$$

The conditional expectation is:

$$\mathbb{E}(z_i \mid z_i > y_r) = \int_{y_r}^{\infty} z\, f(z \mid z > y_r)\, dz$$

Using the exponential distribution and its memoryless property:

$$\mathbb{E}(z_i \mid z_i > y_r) = y_r + \beta$$

Thus:

$$\mathbb{E}[\ell(\alpha,\beta)] = -n\ln\beta + \frac{n\alpha}{\beta} - \frac{1}{\beta}\left(\sum_{i=1}^{r} y_i + (n-r)(y_r+\beta)\right)$$

3. Maximization Step (M-step)

Define the expected log-likelihood function:

$$Q(\alpha,\beta) = -n\ln\beta + \frac{n\alpha}{\beta} - \frac{1}{\beta}\left(\sum_{i=1}^{r} y_i + (n-r)(y_r+\beta^{(k)})\right)$$

Differentiating with respect to β:

$$\frac{\partial Q}{\partial \beta} = -\frac{n}{\beta} - \frac{n\alpha}{\beta^2} + \frac{1}{\beta^2}\left(\sum_{i=1}^{r} y_i + (n-r)(y_r+\beta^{(k)})\right)$$

Setting to zero gives:

$$\beta^{(k+1)} = \frac{\sum_{i=1}^{r} y_i + (n-r)(y_r+\beta^{(k)}) - n\alpha}{n}$$

The location parameter remains:

$$\hat{\alpha} = \min(x_i) \tag{1}$$

The iterative process continues until convergence:

$$|\beta^{(k+1)} - \beta^{(k)}| < 10^{-6}$$

3.2. Implementation on Clinical Survival Data

The derived EM algorithm framework was applied to a real-world dataset obtained from the cBioPortal platform [10], comprising 122 uterine leiomyosarcoma (uLMS) patients. The primary endpoint, Disease-Specific Survival, yielded an observed survival time range of $x \in [2, 283]$ months.

To evaluate the performance of the algorithm under the proposed right-censoring framework, an administrative censoring threshold was imposed at $y_r = 40$ months. Accordingly, the dataset was partitioned into $r = 54$ fully observed events ($y_i \leq 40$) and $n - r = 68$ right-censored observations ($z_i > 40$).

The estimation procedure was conducted as follows. First, based on the boundary condition derived in Eq. (1), the location parameter was determined as the minimum observed survival time:

$$\hat{\alpha} = 2 \text{ months}$$

This represents the minimum observed time to event occurrence. Next, the initial estimate of the scale parameter was computed as:

$$\hat{\beta}^{(0)} = \bar{x} - \hat{\alpha} = 72.84$$

Subsequently, the EM algorithm was applied iteratively, alternating between the expectation step (imputation of censored observations) and the maximization step (parameter updating), showing stable convergence toward a fixed solution. Convergence was reached when the stopping criterion $\left|\hat{\beta}^{(k+1)} - \hat{\beta}^{(k)}\right| < 10^{-6}$ was satisfied after 26 iterations, yielding the final estimate:

$$\hat{\beta} = 68.70.$$

As noted in the Methods, this estimate is conditional on the imposed fixed censoring threshold of 40 months and the common-threshold assumption in the EM formulation. Using individual censoring times would require a more complex E-step to accommodate varying thresholds, which is not considered in this study.

To illustrate the convergence behavior, the trajectory of $\hat{\beta}$ across iterations is presented in Fig. 1, showing a rapid initial adjustment followed by gradual stabilization.

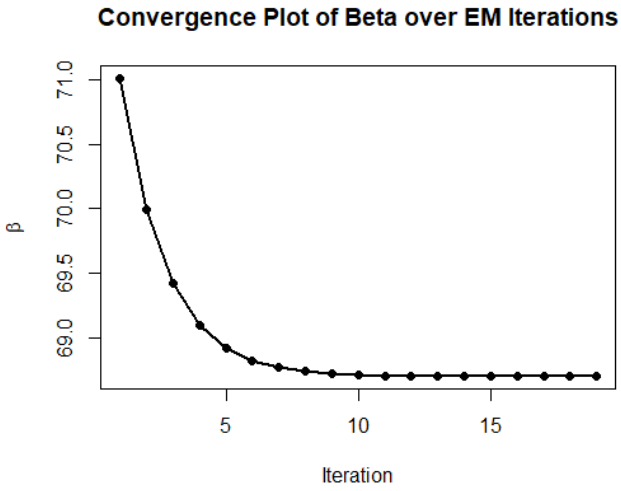

Fig. 1: Convergence trajectory of the estimated scale parameter ($\hat{\beta}$) across EM iterations.

Under the two-parameter exponential model, the hazard function is constant for $t \geq \hat{\alpha}$ and is given by:

$$h(t) = \frac{1}{\hat{\beta}} = \frac{1}{68.70} \approx 0.0146$$

This implies that, conditional on surviving beyond the initial threshold of $\hat{\alpha} = 2$ months, the patients are subject to a constant risk of experiencing the event at approximately 1.46% per month.

It is important to note that this constant hazard rate is an inherent assumption of the exponential model and may not fully capture the dynamic risk patterns typically observed in clinical survival data. Nevertheless, it provides a tractable baseline for subsequent risk-based analysis.

3.3. Monte Carlo Simulation Study

A Monte Carlo simulation study with $B = 1000$ replications was conducted to evaluate the finite-sample performance of the EM algorithm under right-censoring. The data-generating process was calibrated to the uLMS dataset, with sample size $n = 122$ and parameter values

$\hat{\alpha} = 2$ and $\hat{\beta} = 68.70$. In each replication, samples were generated from the two-parameter exponential distribution and subjected to a fixed right-censoring threshold at $y_r = 40$, with censored observations treated as latent variables within the EM framework for parameter estimation. Consistent with the design in Section 2.3, only this scenario is presented as it directly reflects the empirical setting; simulations with varying sample sizes and censoring levels are not included to maintain focus on the clinical application.

The simulation results yielded an average estimated scale parameter of $\hat{\beta}_{\text{sim}} = 67.96$. The corresponding bias was computed as:

$$\text{Bias} = \hat{\beta}_{\text{sim}} - \hat{\beta} = -0.74$$

This negative bias indicates a slight tendency of the EM algorithm to underestimate the true scale parameter under the imposed right-censoring mechanism. Furthermore, the mean squared error (MSE) was calculated as:

$$\text{MSE} = 93.19$$

This level of MSE reflects the sampling variability associated with the relatively high proportion of censored observations, which is approximately 55.7% of the total sample.

3.4. Uncertainty Analysis and Interval Estimation

To evaluate the estimation uncertainty of the EM-derived scale parameter $\hat{\beta} = 68.70$ obtained from the truncated clinical dataset, three inferential approaches were employed to construct 95% confidence intervals (CI).

First, under the assumption of asymptotic normality, the standard error was approximated using the observed Fisher Information. This Wald-type approach produced the following 95% confidence interval:

$$\text{CI}_{\text{Wald}} = [50.38,\ 87.03]$$

To account for potential deviations from normality and the curvature of the likelihood function, a profile likelihood confidence interval was constructed, yielding:

$$\text{CI}_{\text{Profile}} = [53.22,\ 90.82]$$

In parallel, a non-parametric bootstrap procedure with $B = 1000$ resamples was implemented, resulting in the percentile-based confidence interval:

$$\text{CI}_{\text{Bootstrap}} = [50.45,\ 89.30]$$

The right-skewed asymmetry observed in both the profile likelihood and bootstrap intervals, compared to the symmetric Wald interval, suggests that the sampling distribution of the estimator deviates from perfect symmetry. This indicates that computationally intensive methods may provide improved characterization of upper-tail uncertainty in right-censored survival settings.

Finally, the coverage probability (CP) of the Wald interval was evaluated through Monte Carlo simulation. The estimated coverage probability was:

$$\text{CP} = 0.930$$

for a nominal confidence level of 0.95. The observed under-coverage is consistent with the negative bias in the point estimator, indicating that asymptotic intervals may exhibit slight anti-conservative behavior when applied to this truncated survival data framework.

3.5. Survival Risk Forecasting: Value-at-Risk (VaR) and Tail Value-at-Risk (TVaR)

In clinical survival analysis, conventional summaries (e.g., median survival or fixed-interval survival rates) mainly describe average outcomes or probabilities at selected time points. To characterize extreme prognostic scenarios, Value-at-Risk (VaR) and Tail Value-at-Risk (TVaR) are used. VaR at confidence level p corresponds to the survival quantile, i.e., the time by which a proportion p of patients experience the event, and serves as an explicit extreme threshold not emphasized by standard summaries. This is complemented by TVaR, which provides the conditional expectation of survival time beyond the VaR threshold, thereby quantifying the expected remaining lifespan of long-term survivors through the tail of the distribution. Together, VaR and TVaR offer a concise and actionable measure for estimating extended care duration and resource needs.

The forecasting results for various confidence levels (p), based on the estimated parameters $(\hat{\alpha}, \hat{\beta})$, are summarized in Table 2.

Table 2: Forecasting of VaR and TVaR for uLMS survival time

Confidence Level (p)	VaR (Months)	TVaR (Months)
0.80	112.57	181.28
0.90	160.20	228.90
0.95	207.82	276.52
0.99	318.39	387.10

As shown in Table 2, at the 95% level, VaR is 207.82 months (5% exceedance probability), while TVaR indicates an expected survival of 276.52 months conditional on exceeding this threshold, capturing the tail behavior relevant for long-term outcomes.

The estimated two-parameter exponential survival curve is overlaid on the Kaplan-Meier estimator in Fig. 2. The parametric curve approximates the empirical step function, and the vertical dashed line marks the 95% VaR threshold, delineating the extreme tail region.

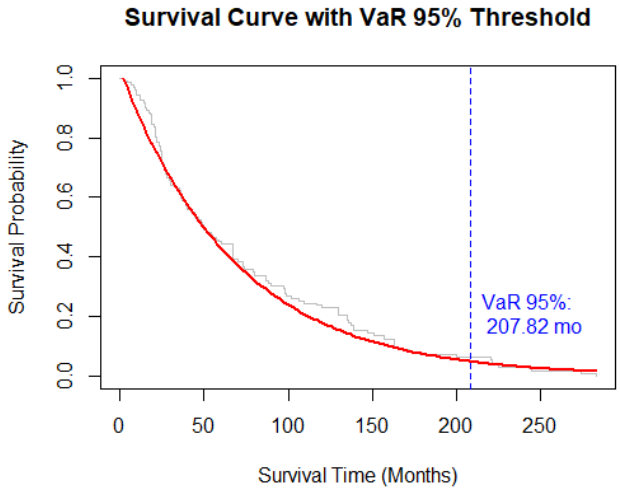

Fig. 2: Survival curve with 95% VaR threshold from Kaplan-Meier and exponential model

To assess analytical stability, a Monte Carlo simulation ($B = 1000$) at the 95% level produced a mean VaR of 206.17 months (SD 18.53), which is close to the theoretical value (207.82). This result indicates consistency between the analytical and simulated estimates under the evaluated censoring condition.

3.6. Goodness of Fit and Model Comparison

To evaluate the adequacy of the estimated two-parameter exponential distribution for the underlying uLMS dataset, analytical, visual, and comparative assessments were performed accounting for the specific right-censored nature of the data.

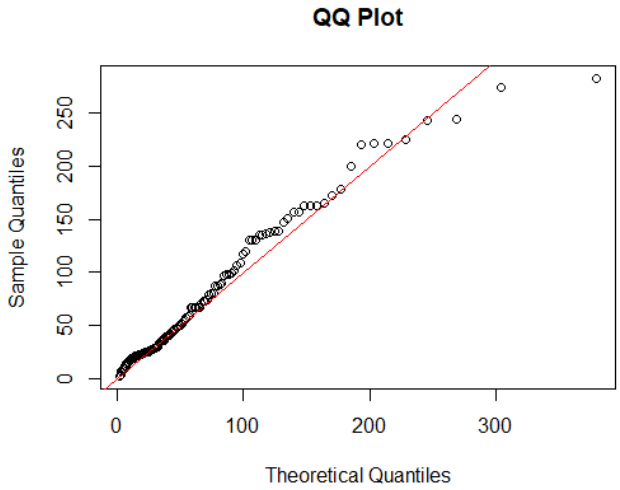

Fig. 3: Censoring-adjusted Q-Q plot comparing empirical (Kaplan-Meier) and theoretical quantiles of the two-parameter exponential model

Visual diagnostics were conducted using a censoring-adjusted Quantile-Quantile (Q-Q) plot, as presented in Fig. 3. By utilizing the Kaplan-Meier estimator to accommodate right-censored observations, the empirical sample quantiles align closely with the theoretical quantiles along the reference line, particularly in the lower and middle regions of the distribution. While minor deviations appear in the extreme upper tail, this behavior is typical for right-skewed, heavily censored clinical data, indicating an acceptable overall fit.

Table 3: Model comparison using Akaike Information Criterion (AIC)

Model	AIC
Two-parameter Exponential Model	568.82
Weibull Model	653.90

Based on Table 3, a formal model comparison using the AIC was conducted. Both models were evaluated using the right-censored log-likelihood formulation with an equal number of estimated parameters ($k = 2$). The results demonstrate that the two-parameter exponential model attains a substantially lower AIC value (568.82) than the Weibull model (653.90), indicating a comparatively better fit to the right-censored data. This finding suggests that the inclusion of the location parameter (α) effectively captures the delayed onset of events observed in the uLMS dataset, providing a more parsimonious and adequate representation compared to the standard Weibull model which lacks a threshold parameter.

3.7. Sensitivity Analysis

To evaluate the sensitivity of the proposed EM estimator to varying degrees of incomplete data, an analysis was conducted by artificially introducing additional right-censoring. Specifically, the dataset was truncated at different order statistics to restrict the proportions of observed events to 70%, 80%, and 90% of the total sample size (n). It is important to explicitly note the inverse relationship in this framework: restricting the observed proportion to 70% corresponds to imposing a high artificial right-censoring level of 30%, thereby deliberately limiting the available

failure time information. The estimation results under these manipulated censoring schemes are summarized in Table 4.

Table 4: Sensitivity of the estimated scale parameter ($\hat{\beta}$) to varying observation proportions

Observed Proportion	Truncated Index (r)	Estimated Scale Parameter ($\hat{\beta}$)
0.70 (70%)	86	76.42
0.80 (80%)	98	78.51
0.90 (90%)	110	74.52

As presented in Table 4, the estimated scale parameter ($\hat{\beta}$) fluctuates between 74.52 and 78.51 months across the different censoring levels. To assess whether this fluctuation implies estimator instability, these variations are evaluated relative to the standard error of the baseline model. The primary unmanipulated estimate ($\hat{\beta} = 68.70$) has a 95% Fisher confidence interval of [50.38, 87.03].

Within this benchmark, when the censoring level is increased to 30% (Observed Proportion = 0.70), the resulting estimate ($\hat{\beta} = 76.42$) remains within the baseline confidence interval. This variation reflects the expected impact of information loss due to higher censoring, while the estimates under different censoring scenarios remain within the range of sampling variability. These results indicate that the EM-based estimation does not exhibit excessive sensitivity to moderate increases in right-censored observations.

3.8. Methodological Justification for the EM Framework

In the context of censored survival data, the use of the Expectation-Maximization (EM) algorithm, as opposed to alternatives such as Newton-Raphson or Bayesian Markov Chain Monte Carlo (MCMC), is related to the structural characteristics of the two-parameter exponential distribution. Gradient-based methods like Newton-Raphson rely on the Hessian matrix, while the presence of the location parameter (α) introduces a boundary constraint ($y_i \geq \alpha$) that may lead to non-regular likelihood behavior and potential convergence issues.

Bayesian MCMC represents another possible approach but requires explicit specification of prior distributions. For the uLMS clinical dataset, the use of subjective priors may influence the resulting survival estimates, particularly under moderate sample sizes [22].

The EM algorithm addresses these considerations by treating unobserved survival times as latent variables, which simplifies the estimation procedure. Although no direct numerical comparison with alternative methods is provided, the EM framework is known to produce a monotonic increase in the likelihood function across iterations without requiring second-derivative calculations. Under these conditions, the EM approach provides a data-driven framework that can be applied to the analysis of this clinical dataset.

4. Conclusion

This study evaluated survival forecasting for right-censored uterine leiomyosarcoma data using the EM algorithm on a two-parameter exponential model. The algorithm converged in 26 iterations ($\hat{\alpha} = 2, \hat{\beta} = 68.70$), demonstrating minimal negative bias (-0.74) via Monte Carlo simulations ($B = 1000$). The proposed model (AIC = 568.82) provided a more parsimonious fit than the standard Weibull distribution (AIC = 653.90). Integrating Value-at-Risk (VaR) and Tail Value-at-Risk (TVaR) quantified extreme survival scenarios, establishing a 95% VaR threshold of 207.82 months and a TVaR (conditional expected survival) of 276.52 months. This framework offers actionable prognostic insights, with inferential reliability supported by consistent Fisher, bootstrap, and profile likelihood intervals.

However, to satisfy the constant hazard assumption and ensure EM tractability, a 40-month Type-II censoring threshold was uniformly imposed. While ensuring computational stability, this administrative cutoff alters the natural data structure and obscures individual-specific

tail information. Future research should explore flexible models (e.g., generalized Gamma or exponentiated Weibull) to avoid restrictive truncation, and incorporate longitudinal covariates via joint modeling to enhance predictive accuracy.

CRediT Authorship Contribution Statement

Ardi Kurniawan: Conceptualization, Methodology, Formal Analysis. **Deby Victoria:** Data Curation, Software, Writing–Original Draft Preparation. **Christabel Lee Angie Sugianto:** Validation, Visualization, Writing–Review & Editing.

Declaration of Generative AI and AI-assisted Technologies

During the preparation of this manuscript, AI tools were used solely for language editing (grammar, clarity, and readability). All research processes—conceptualization, methodology, data analysis, interpretation, and conclusions—were conducted entirely by the authors, who take full responsibility for the content and scientific integrity of this work.

Declaration of Competing Interest

The authors declare no competing interests.

Funding and Acknowledgments

This research received no external funding. The authors would like to express their gratitude to Airlangga University for providing the academic environment and facilities that supported this work. Additionally, acknowledgments are due to cBioPortal for Cancer Genomics for providing the open-access clinical datasets used in this analysis.

Data and Code Availability

The clinical uterine leiomyosarcoma (uLMS) dataset analyzed in this study was originally published by Dermawan et al. [23]. The raw data were accessed via the cBioPortal for Cancer Genomics repository [10], specifically from the "LMS MSK 2024" study cohort (https://www.cbioportal.org/study/clinicalData?id=lms_msk_2024). It is important to clarify that this study performs an independent methodological evaluation using parametric survival modeling and risk forecasting, which is distinct from the genomic analysis conducted in the original research. The analysis was performed using RStudio, and the source code is available from the corresponding author upon reasonable request.

References

[1] Tomasz Burzykowski. “Survival analysis: Methods for analyzing data with censored observations”. In: *Seminars in Orthodontics* 30.1 (2024), pp. 29–36. DOI: 10.1053/j.sodo.2024.01.008.

[2] T. Machimud, L. O. Nashar, D. Fakhriyana, and L. O. Sabran. “Estimasi parameter Cox semiparametric hazards model dengan metode Efron pada data tersensor kanan”. In: *Jurnal Matematika UNAND* 10.3 (2021), pp. 394–405. DOI: 10.25077/jmu.10.3.394-405.2021.

[3] A. A. Al-Shomrani. “Various methods of estimating constant-partially accelerated life tests of the Odd Kappa-Exponential distribution based on complete data”. In: *Results in Engineering* 27 (2025). DOI: 10.1016/j.rineng.2025.106907.

[4] Y. Zhuang and S. R. Bapat. "On comparing locations of two-parameter exponential distributions using sequential sampling with applications in cancer research". In: *Communications in Statistics - Simulation and Computation* 51.10 (2022), pp. 6114–6135. DOI: 10.1080/03610918.2020.1794007.

[5] A. Kurniawan, J. T. Victory, and T. Saifudin. "Bayes estimation of a two-parameter exponential distribution and its implementation". In: *TELKOMNIKA (Telecommunication, Computing, Electronics and Control)* 22.6 (2024). DOI: 10.12928/telkomnika.v22i6.26015.

[6] A. J. Turkson, F. Ayiah-Mensah, and V. Nimoh. "Handling Censoring and Censored Data in Survival Analysis: A Standalone Systematic Literature Review". In: *International Journal of Mathematics and Mathematical Sciences* 2021 (2021). DOI: 10.1155/2021/9307475.

[7] A. Kurniawan, A. N. W. Ramadhani, and R. Hikmaya. "Estimation of Burr XII Distribution Parameters on Type II Censored Data Using Expectation-Maximization Algorithm". In: *Pakistan Journal of Statistics* 42.2 (2026), pp. 105–115. https://scholar.unair.ac.id/en/publications/estimation-of-burr-xii-distribution-parameters-on-type-ii-censore/ (visited on 05/24/2026).

[8] S.-F. Wu. "Interval Estimation for the Two-Parameter Exponential Distribution under Progressive Type II Censoring on the Bayesian Approach". In: *Symmetry* 14.4 (2022), p. 808. DOI: 10.3390/sym14040808.

[9] A. Ganguly, S. Mitra, D. Samanta, and D. Kundu. "Exact inference for the two-parameter exponential distribution under Type-II hybrid censoring". In: *Journal of Statistical Planning and Inference* 142.3 (2012), pp. 613–625. DOI: 10.1016/j.jspi.2011.08.001.

[10] cBioPortal for Cancer Genomics. *Clinical Data for LMS MSK 2024 Study*. 2024. https://www.cbioportal.org/study/clinicalData?id=lms_msk_2024 (visited on 02/28/2026).

[11] A. Clifford Cohen and B. Jones Whitten. *Parameter Estimation in Reliability and Life Span Models*. CRC Press, 2020. DOI: 10.1201/9781003066064.

[12] Elsayed A. Elsayed. *Reliability Engineering*. 3rd ed. John Wiley & Sons, 2021. DOI: 10.1002/9781119665946.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data Via the EM Algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–22. DOI: 10.1111/j.2517-6161.1977.tb01600.x.

[14] Kevin P. Murphy. *Probabilistic Machine Learning: An Introduction*. MIT Press, 2022. https://probml.github.io/book1 (visited on 05/24/2026).

[15] D. Luengo, L. Martino, M. Bugallo, V. Elvira, and S. Särkkä. "A survey of Monte Carlo methods for parameter estimation". In: *EURASIP Journal on Advances in Signal Processing* 2020.1 (2020). DOI: 10.1186/s13634-020-00675-6.

[16] P. T. Situma and L. Odongo. "Maximum Likelihood Estimation of Parameters for Poisson-exponential Distribution Under Progressive Type I Interval Censoring". In: *American Journal of Theoretical and Applied Statistics* 9.2 (2020), p. 14. DOI: 10.11648/j.ajtas.20200902.11.

[17] C. Robert and G. Casella. *Introducing Monte Carlo Methods with R*. Springer, 2010. DOI: 10.1007/978-1-4419-1576-4.

[18] Q. Chen and W. Gui. "Statistical Inference of the Generalized Inverted Exponential Distribution under Joint Progressively Type-II Censoring". In: *Entropy* 24.5 (2022), p. 576. DOI: 10.3390/e24050576.

[19] A. M. AboAlkhair, G. G. Hamedani, N. A. Ahmed, M. Ibrahim, M. A. Zayed, and H. M. Yousof. "A New G Family: Properties, Characterizations, Different Estimation Methods and PORT-VaR Analysis for U.K. Insurance Claims and U.S. House Prices Data Sets". In: *Mathematics* 13.19 (2025), p. 3097. DOI: 10.3390/math13193097.

[20] F. Belzunce, A. M. Franco-Pereira, and J. Mulero. "New stochastic comparisons based on tail value at risk measures". In: *Communications in Statistics - Theory and Methods* 51.3 (2022), pp. 767–788. DOI: 10.1080/03610926.2020.1754857.

[21] K. Toffaha, M. Emre Simsekler, A. Shehhi, et al. "Comprehensive survival analysis of breast cancer patients: a bayesian network approach". In: *BMC Medical Informatics and Decision Making* 25 (2025), p. 349. DOI: 10.1186/s12911-025-03197-z.

[22] I. Paolucci, Y. M. Lin, J. Albuquerque Marques Silva, et al. "Bayesian parametric models for survival prediction in medical applications". In: *BMC Medical Research Methodology* 23 (2023), p. 250. DOI: 10.1186/s12874-023-02059-4.

[23] J. K. Dermawan, S. Chiang, S. Singer, B. Jadeja, M. L. Hensley, W. D. Tap, S. Movva, R. G. Maki, and C. R. Antonescu. "Developing Novel Genomic Risk Stratification Models in Soft Tissue and Uterine Leiomyosarcoma". In: *Clinical Cancer Research* 30.10 (2024), pp. 2260–2271. DOI: 10.1158/1078-0432.CCR-24-0148.