The Log-normal-Log-logistic Mixture Model Analysis on the Survival Data of Diffuse Large B-Cell Lymphoma

Sulasri Suddin, Gunardi, Mardiah Suci Hardianti, and Fajar Adi-Kusumo

***Abstract*—Diffuse large B-cell lymphoma (DLBCL) is a prevalent and heterogeneous type of blood cancer classified under the non-Hodgkin lymphoma subtype. Early diagnosis significantly improves patient survival, and a deeper understanding of DLBCL's statistical characteristics can assist physicians in improving treatment outcomes. Several studies have shown that mixture survival modeling effectively captures survival data heterogeneity and identifies diverse survival patterns. Therefore, this study aims to propose a new mixture model combining log-logistic and log-normal distributions to analyze DLBCL survival data. The model was optimized using the expectation-maximization algorithm to estimate parameters via maximum likelihood. Model accuracy was then evaluated using mean squared error (MSE) comparisons and Kolmogorov-Smirnov (K-S) test. The results showed that the proposed mixture model, combining log-normal and log-logistic distributions, effectively represented survival data for DLBCL patients in Indonesia. In addition, it provided the best fit based on MSE and statistical significance. This application demonstrated the model's suitability for analyzing heterogeneous datasets in DLBCL cases, providing a foundation for further studies on its generalization to broader patient survival data in Indonesia.**

***Index Terms*—expectation-maximization algorithm, mixture models, survival, diffuse large b-cell lymphoma.**

I. Introduction

NON-HODGKIN lymphoma is a cancer that develops in the human lymphatic system, accounting for approximately $553,000$ new cases and $250,475$ deaths in 2022 [1]. According to the American Cancer Society, approximately $80,550$ new cases and $20,180$ related deaths were reported in the United States in 2022 [2]. Diffuse large B-cell lymphoma (DLBCL) has been reported to be the most prevalent form, representing approximately 25% to 30% of all cases worldwide [3]. DLBCL refers to an aggressive form of lymphoma marked by substantial variability in clinical presentation and prognosis. Despite achieving a 5-year survival rate of 60% to 70% with the first-line standard treatment involving rituximab combined with cyclophosphamide, doxorubicin, vincristine, and prednisone (R-CHOP), 40% to 50% of patients experience relapse or develop resistance following therapy [4]. This indicates that, although improvements in treatment have been made, the survival rate of patients with DLBCL remains relatively low, showing the ongoing need for more effective therapies.

According to previous studies, overall survival (OS) is a clinical indicator for measuring the duration of survival among individuals diagnosed with cancer. From a statistical perspective, OS is considered an important indicator for evaluating the effectiveness of treatment provided. A valuable prognostic instrument in medical practice, the International Prognostic Index (IPI) was developed during the CHOP era to identify 4 distinct risk groups. However, the inclusion of rituximab in the treatment has limited its effectiveness in distinguishing between these groups. Improved scoring models, such as R-IPI [5] and NCCN-IPI [6], have been created to assist in predicting the patient's prognosis. Although these 2 scoring systems provide greater prognostic, both are still not fully effective in identifying very high-risk subgroups of patients [7]. Consequently, developing an assessment tool that can address all variations and levels of risk in patients with DLBCL is still a challenge.

Survival analysis is a statistical method used to analyze data where the outcome variable represents the time until a specific event takes place [8]. The evaluation of survival periods or failure intervals is a significant focus in numerous fields, especially in health sciences. Historically, survival analysis was usually represented using nonparametric methods or classical parametric models, such as Gamma, Exponential, and Weibull distributions [9], [10], [11], [12], [13]. When data deviates from a predetermined distribution, applying the classical model becomes inappropriate. Data that has more than 1 distribution is called heterogeneous survival data.

Classical parametric models with a single distribution are often inadequate to model heterogeneous survival data since the assumption of population homogeneity is unrealistic. Therefore, the current study proposes the use of a mixture model of 2 parametric distributions (identical or different) that allows the grouping of subpopulations with varying survival characteristics. This provides a more accurate description of the data. Mixture models can accommodate survival distributions that have complex hazard rate patterns, such as non-monotonicity, which often cannot be described by a single parametric distribution. The combination of 2 distributions allows a more flexible approach to modeling various survival patterns. In the context of medical survival data, a subpopulation can represent groups of patients with good and poor prognosis. Mixture models are more flexible than standard parametric models and can fit heterogeneous

Manuscript received January 27, 2025; revised June 2, 2025.

S. Suddin is a doctoral candidate of the Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia (email: sulasri.suddin@gmail.com)

Gunardi is a professor of Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia (corresponding author to provide e-mail: gunardi@ugm.ac.id)

M. S. Hardianti is an associate professor of Department of Internal Medicine, Faculty of Medicine, Public Health and Nursing, Universitas Gadjah Mada, Yogyakarta, 55281, Indonesia (e-mail: mardiah.suci@ugm.ac.id)

F. Adi-Kusumo is a professor of Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia (e-mail: f_adikusumo@ugm.ac.id)

survival data. The models produce Kaplan-Meier curves that remain high over time, which is common in immunotherapy trials. This potentially leads to more accurate estimates of median survival time compared to standard models [14].

Several studies have implemented mixture modeling approaches to analyze survival data. De Angelis et al. [15] proposed a parametric mixture model for the relative survival rates of colon cancer patients using data from a Finnish population-based cancer registry. In a Bayesian context, Marín et al. [16] introduced a mixture Weibull with an unspecified number of components to analyze heterogeneous survival data, including cases with right censoring, using the birth-death Markov chain Monte Carlo method. Erişoğlu and Erol [17] applied a mixture model combining the exponential and geometric distributions, enhanced with the Expectation-Maximization algorithm, to analyze heterogeneous survival data. In addition, the method can provide a more accurate estimate of overall survival compared to the standard model in patients with relapsed/refractory DLBCL [18].

Studies have shown that therapeutic vaccines have different effects on lung cancer patient populations with short-term and long-term survival and can increase the proportion of patients with longer survival [19]. In addition, the mixture parametric model is more suitable for detecting the effectiveness of immunotherapy compared to the standard model [20]. This suggests that modeling with 2 mixture distributions can serve as an alternative approach to defining 2 groups of patients with different intrinsic mortality rates. Meanwhile, some studies have shown slightly different results [21], [22], [23]. For instance, a study on survival in patients diagnosed with Hodgkin Lymphoma who received autologous hematopoietic stem cell transplantation (HSCT) showed that the non-mixture model gave better results than the mixture model when using distributions from the GMW (Generalized Modified Weibull) family [21].

Several studies have also shown that the mixture model can be considered in analyzing heterogeneous survival data. The current study explores the use of a 2-distribution mixture model in developing a novel survival time framework to estimate overall survival in DLBCL patients. In addition, mixture and standard parametric models were compared to determine the best for predicting the OS of patients based on their respective mean squares error (MSE) values and statistical significance. The remaining part of this study is structured as follows. Section 2 outlines the dataset and the statistical methods used, while Section 3 shows the mixture model and estimation in survival analysis. Section 4 represents real data application in DLBCL along with the discussion, and Section 5 provides the conclusion of the study.

II. Materials and Methods

A. Data

A total of 387 patients diagnosed with DLBCL between 2012 and 2020 were first identified in medical record data, which contained detailed clinical information about the characteristics of DLBCL at primary diagnosis. These included Ann Arbor staging, performance status, lactate dehydrogenase serum, extranodal site, and type of therapy (CHOP/R-CHOP). Patients were divided into risk groups according to their IPI, with 1 point given for each factor. In this study, 225 patients were included. The reason for exclusion in the detailed data review was that the medical records considered in this study were incomplete or could not be identified ($n = 164$, 42.6%). This study used DLBCL medical record data which had ethical approval from the Medical and Health Study Ethics Committee (MHREC) Faculty of Medicine, Public Health, and Nursing, Universitas Gadjah Mada, Ref. No.: KE/FK/1356/EC/2023.

TABLE I: The features of the theoretical distributions used in the study.

Model	Parameters	Probability Density Function	Survival Function
Exponential	λ	$f(t) = \lambda e^{-\lambda t}$	$S(t) = e^{-\lambda t}$
Gamma	α, β	$f(t) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} t^{\alpha-1} e^{-\frac{t}{\beta}}$, $\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} dt$	No closed form
Weibull	λ, γ	$f(t) = \lambda\gamma t^{\gamma-1} e^{-\lambda t^{\gamma}}$	$S(t) = e^{-\lambda t^{\gamma}}$
Log-normal	μ, σ	$f(t) = \frac{1}{\sqrt{2\pi}\sigma t} e^{-\frac{1}{2\sigma^2}[\ln t - \mu]^2}$	$S(t) = 1 - \Phi\left(\frac{\ln t - \mu}{\sigma}\right)$
Log-logistic	α, β	$f(t) = \frac{\frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1}}{\left(1+\left(\frac{t}{\alpha}\right)^{\beta}\right)^2}$	$S(t) = \frac{1}{1+\left(\frac{t}{\alpha}\right)^{\beta}}$

TABLE II: Goodness-of-fit measures for the five distribution models.

	Exponential	Gamma	Weibull	Log-normal	Log-logistic
A-D	4.5285	1.5724	2.0224	1.0567	0.6876

B. Statistical Analysis

This section explained the fundamental concepts of survival analysis. Survival time data captured the duration until a specific event occurred, such as failure, death, response, or disease progression. Let T represent the survival time, and the survival function, $S(t)$, was defined as the probability of an individual surviving beyond time t, namely., $S(t) = P(T > t)$. A summary of the probability density and survival functions used in this study was presented in TABLE I.

III. Mixture Model Estimation in Survival Analysis

A. Model Description

In this section, a mixture model was introduced, which consisted of 2 identical distributions in survival analysis, along with a mixture of 2 distinct distributions applied to the survival analysis of DLBCL patients. To apply a theoretical probability distribution function to the survival time data, the R statistical software was used. The quality of the fitted distributions was evaluated using Goodness-of-Fit test statistics. In addition, the null hypothesis assumed that the data follow the proposed distribution. A distribution was deemed appropriate for the data when the discrepancy between the data and the fitted distribution was below a predefined threshold (critical value). The distribution with the lowest statistical value was considered the best fit.

Based on TABLE II, the distribution with the lower Anderson-Darling value was log-logistic with a statistical value of 0.6876. Therefore, the log-logistic distribution was the closest to the actual data based on the given Anderson-Darling (A-D) goodness-of-fit statistics. This was followed by the log-normal and gamma distributions. Consequently, it was important to consider 2 or 3 of these distributions to describe the characteristics of observed data. In this study, only 2 mixture distributions were considered, namely log-logistic and log-normal.

Based on the A-D statistic test, the survival time probabilities were best modeled by the log-logistic and log-normal distributions, with the probability density and survival function, were as follows.

1) Log-logistic: $f(t;\alpha,\beta) = \frac{\frac{\alpha}{\beta}\left(\frac{t}{\beta}\right)^{\alpha-1}}{\left(1+\left(\frac{t}{\beta}\right)^{\alpha}\right)^2}$, $S(t;\alpha,\beta) = \frac{1}{1+\left(\frac{t}{\beta}\right)^{\alpha}}$, $h(t;\alpha,\beta) = \frac{\frac{\alpha}{\beta}\left(\frac{t}{\beta}\right)^{\alpha-1}}{1+\left(\frac{t}{\beta}\right)^{\alpha}}$, where $t>0$, $\alpha>0$, $\beta>0$.
2) Log-normal: $f(t;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma t}e^{-\frac{1}{2\sigma^2}[\ln t-\mu]^2}$, $S(t;\mu,\sigma) = 1-\Phi\left(\frac{\ln t-\mu}{\sigma}\right)$, $h(t;\mu,\sigma) = \frac{\phi\left(\frac{\ln t-\mu}{\sigma}\right)}{\sigma t\left[1-\Phi\left(\frac{\ln t-\mu}{\sigma}\right)\right]}$, where $t>0$.

To represent a heterogeneous survival data set, a combination of 2 distinct distributions was used, namely Log-logistic and Log-normal. For identical distribution pairs, the equations could be written as

$$\begin{aligned} f_{Llogis-Llogis}(t) &= \pi_1 f_{Llogis}(t)+\pi_2 f_{Llogis}(t) \\ f_{Lnorm-Lnorm}(t) &= \pi_1 f_{Lnorm}(t)+\pi_2 f_{Lnorm}(t). \end{aligned}$$

Non-identical distribution pairs, such as Log-logistic and Log-normal, were as follows.

$$f_{Llogis-Lnorm}(t) = \pi_1 f_{Llogis}(t)+\pi_2 f_{Lnorm}(t),$$

where π_1, π_2 were the mixing weights and $\pi_1+\pi_2=1$, $0<\pi_1,\pi_2<1$ for each mixture distribution model.

To determine when a specific distribution was more suitable, the goodness-of-fit test was applied using the MSE. The MSE value was defined as

$$MSE = \frac{\sum_{i=1}^{n}\left[F(t_i)-F_e(t_i)\right]^2}{n},$$

where $F(t)$ was the cumulative distribution function proposed to model heterogeneous survival data sets and $F_e(t)$ was the empirical distribution, the most suitable distribution was the one with the smallest MSE value.

The goodness-of-fit was also evaluated using the Kolmogorov-Smirnov (K-S) test to assess how well the proposed model fitted the data, as conducted in previous studies that compared several mixture models using the K-S test statistic [23], [24]. Meanwhile, the best model was the one with the smallest K-S value, and the K-S test was defined as

$$K-S = max\left|F(t)-F_e(t)\right|.$$

B. Parametric Mixture Model of Two Distribution in Survival Analysis

In the context of a finite mixture distribution model, the EM algorithm was a commonly applied method that delivered an iterative process for determining the maximum likelihood estimator of the unknown parameter [25]. In this study, a model was constructed from a mixture of 2 distributions, namely $f_1(x;\theta_1)$ and $f_2(x;\theta_1)$. The 2-distribution mixture model of the density function T had the form

$$f(t_i;\psi) = \sum_{k=1}^{2}\pi_k f_k(t_i;\theta_k),$$

where the vector $\psi=(\pi,\theta_k)$ included all the unknown parameters, such as π and $\theta_k=(\theta_1,\theta_2)$, which represented the parameters of the model. The function $f_k(t_i;\theta_k)$ denoted the density function for the component $k=1,2$. In mixture models, π were the weights given to 2 distributions, $\pi_k\in(0,1)$. The sum of the weights must be 1 which could be formulated as $\sum_{k=1}^{2}\pi_k=1$.

The survival function was also used to represent the model, that is,

$$S(t_i;\psi) = \sum_{k=1}^{2}\pi_k S_k(t_i;\theta_k),$$

where $S_k(t_i;\theta_k)$ denotes the kth component survival function.

Due to the time survival DLBCL, which comprised 2 distinct distributions, namely log-logistic and log-normal, the mixture of the densities of log-logistic and log-normal was represented as

$$f(t_i;\psi) = \pi_1\frac{\frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1}}{\left(1+\left(\frac{t}{\alpha}\right)^{\beta}\right)^2}+\pi_2\frac{1}{\sqrt{2\pi}\sigma t}e^{-\frac{1}{2\sigma^2}[\ln t-\mu]^2} \quad (1)$$

where π_1 and π_2 represented the proportion of the mixture distribution, $\theta=(\alpha,\beta,\mu,\sigma)$ was a set of parameters. These were denoted by α, β, a set of parameters for the log-logistic distribution, and μ, σ, a set of parameters for the log-normal distribution.

C. Estimation and Expectation-Maximization Algorithm in the Survival Mixture Model

The purpose of this analysis was to estimate the π weights and parameters α, β, μ, σ given t_i from (1). New distribution parameters could be estimated using the Expectation-Maximization (EM) Algorithm. The Expectation-Maximization algorithm was shown to be the most appropriate method for estimating the mixture parameters [26]. This was an iterative procedure used to determine the maximum likelihood (ML) estimate. The Expectation-Maximization algorithm consisted of 2 steps, namely Expectation (E) and Maximization (M). To obtain the estimated value, the first step was to formulate the likelihood function.

Right censored survival data could be represented as a pair of survival observation values with their censored status, namely (t_i,δ_i), $i=1,2,\cdots,n$ where $\delta_i=0$ if i was censored and $\delta_i=1$ if i obtained an event. From the (T_i,δ_i) pairs, which were independent from one another, the likelihood function for right-censored data was written as follows [27],

$$L = \prod_{i=1}^{n}\left(f(t_i;\theta)\right)^{\delta_i}\left(S(t_i;\theta)\right)^{(1-\delta_i)}.$$

In this case, the sample was right censored, and the likelihood function for a model for a mixture of the distributions based on type-I censored samples was written as

$$L = \prod_{i=1}^{n} \prod_{k=1}^{2} \left(f_k\left(t_i;\theta_k\right)\right)^{\delta_i} \left(S_k\left(t_i;\theta_k\right)\right)^{(1-\delta_i)}, \quad (2)$$

where $t = (t_1, t_2, \cdots, t_n)$ and δ_i was an indicator function.

The natural logarithm of (2) was given by

$$\begin{aligned} \ln L &= \sum_{i=1}^{n} \sum_{k=1}^{2} [\delta_i \ln f_k\left(t_i;\theta_k\right) + (1-\delta_i) \\ &\quad \ln S_k\left(t_i;\theta_k\right)]. \end{aligned} \quad (3)$$

This was well-established that the maximum likelihood estimate of the parameter vector corresponded to the values of the parameters that maximized the likelihood function (2) or the logarithm of the likelihood function (3). The Expectation-Maximization algorithm, proposed by [25], used the concept of missing data. In this context, the missing data referred to the unknown group membership of each observation in the sample. These missing values could be represented by the random vector $z = (z_1, \cdots, z_n)$ where $z_i = (z_{1i}, \cdots, z_{ki})$ and

$$z_{ki} = \begin{cases} 1, \, if \; T_i \; belongs \; to \; group \; k \\ 0, \, otherwise \end{cases}.$$

In addition, it was assumed that z_{ki} was a multinomial i.i.d with probability π_k and its density function was given by

$$f\left(z_i\right) = \prod_{k=1}^{2} \pi_k^{z_{ki}}. \quad (4)$$

The probability density function for t_i when z_i was known, could be represented as follows

$$f\left(t_i|z_i\right) = \prod_{k=1}^{2} \left(f_k\left(t_i;\theta_k\right)\right)^{z_{ki}}, \quad (5)$$

where vectors $z_{ki} = (z_{1i}, z_{2i})^T$, $\theta = (\theta_1, \theta_2)^T$, and z_{ki} considered as unobserved data.

As a result, the joint distribution of T and Z was

$$f\left(t_i, z_i\right) = \prod_{k=1}^{2} \left(f_k\left(t_i;\theta_k\right)\right)^{z_{ki}} \pi_k^{z_{ki}}.$$

Therefore, for an i.i.d sample of size n that consisted of observed pairs of (t_i, δ_i), $i = 1, 2, \cdots, n$, the likelihood function (2) was expressed as

$$L_C = \prod_{i=1}^{n} \prod_{k=1}^{2} \left(\pi_k f_k\left(t_i;\theta_k\right)\right)^{z_{ki}\delta_i} \left(\pi_k S_k\left(t_i;\theta_k\right)\right)^{z_{ki}(1-\delta_i)}$$

or this expression could be restated as

$$L_C = \prod_{i=1}^{n} \prod_{k=1}^{2} \left(h_k\left(t_i\right)\right)^{z_{ki}\delta_i} \left(\pi_k S_k\left(t_i\right)\right)^{z_{ki}}. \quad (6)$$

The estimation results were obtained by maximizing the function L_C, taking into account π, α, β, μ, σ. Subsequently, the value of ln-likelihood was obtained based on (6) as follows

$$\ln L_C = \sum_{i=1}^{n} \sum_{k=1}^{2} \ln \left(\left(h_k\left(t_i\right)\right)^{z_{ki}\delta_i} \left(\pi_k S_k\left(t_i\right)\right)^{z_{ki}} \right)$$

or

$$\begin{aligned} \ln L_C &= \sum_{i=1}^{n} \sum_{k=1}^{2} [z_{ki}\delta_i \ln h_k\left(t_i\right) + z_{ki} \ln S_k\left(t_i\right) \\ &\quad + z_{ki} \ln \pi_k]. \end{aligned} \quad (7)$$

The complete ln-likelihood in (7) was maximized using the Expectation-Maximization iteration algorithm. In addition, the Expectation-Maximization algorithm required an iteration between what was called E (expectation) and was called Step M (Maximization). The following were the steps to get the parameter estimation formula. Step E of $\ln L_C$, $E\left(\ln L_C\right)$, was calculated considering the conditional distribution of the unobserved data z_{ki}.

Based on Bayes' theorem from (4) and (5), the conditional distribution of z_i over t_i, the following equation was given by

$$f\left(z_{ki}|t_i\right) = \frac{\left(f_k\left(t_i;\theta_k\right)\right)^{z_{ki}} \pi_k^{z_{ki}}}{\sum_{k=1}^{2} \left(f_k\left(t_i;\theta_k\right)\right)^{z_{ki}} \pi_k^{z_{ki}}}. \quad (8)$$

To calculate the conditional expectation of z_{ki}, Equation (8) was used, which resulted in the following expectation value,

$$E\left(z_{ki}|t_i\right) = \sum_{z_{ki}=0}^{1} z_{ki} f\left(z_{ki}|t_i\right) = \frac{\pi_k f_k\left(t_i;\theta_k\right)}{\sum_{k=1}^{2} \pi_k f_k\left(t_i;\theta_k\right)}.$$

The z_k variables were considered as missing data in the E-step and the hidden variable vector z_k was estimated by evaluation of the expectation $E(z_{ki}|t_i)$, resulting in

$$\hat{z}_{ki} = E\left(z_{ki}|t_i\right) = \frac{\pi_k f_k\left(t_i;\theta_k\right)}{\sum_{k=1}^{2} \pi_k f_k\left(t_i;\theta_k\right)}, \quad (9)$$

so that

$$\begin{aligned} \hat{z}_{1i} &= E\left(z_{1i}|t_i\right) = \frac{\pi_1 f_1\left(t_i;\theta_1\right)}{\pi_1 f_1\left(t_i;\theta_1\right) + \pi_2 f_2\left(t_i;\theta_2\right)}; \\ \hat{z}_{2i} &= E\left(z_{2i}|t_i\right) = \frac{\pi_2 f_2\left(t_i;\theta_1\right)}{\pi_1 f_1\left(t_i;\theta_1\right) + \pi_2 f_2\left(t_i;\theta_2\right)}. \end{aligned} \quad (10)$$

At the E-step of the Expectation-Maximization algorithm, the term of z_{ki} in (7) could be replaced by the expected value $\hat{z}_{ki}$ calculated in (10)

$$\begin{aligned} \ln L_C &= \sum_{i=1}^{n} \sum_{k=1}^{2} [\hat{z}_{ki}\delta_i \ln h_k\left(t_i\right) + \hat{z}_{ki} \ln S_k\left(t_i\right) \\ &\quad + \hat{z}_{ki} \ln \pi_k]. \end{aligned} \quad (11)$$

In the M-step of the algorithm, Equation (11) must be maximized, taking into account. π_k and θ_k assuming $\hat{z}_{ki}$ remained at this step. Maximizing (11) by considering π_k, subject to the constraints in (11) could be obtained by the Lagrange multiplier method. The maximized function was written as follows $f\left(t_i, z_{ki}\right) = \ln L_C$.

The extreme value (optimization) of the function f was observed with certain constraints that needed to be satisfied, such as., $\sum_{k=1}^{2} \pi_k = 1$, then the Lagrange function was formed: $F\left(\lambda, t_i, z_{ki}\right) = f\left(t_i, z_{ki}\right) + \lambda \; g\left(t_i, z_{ki}\right)$. The constraint function $g\left(t_i, z_{ki}\right)$ must be equal to zero to ensure that $g\left(t_i, z_{ki}\right) = \sum_{k=1}^{2} \pi_k - 1$ was obtained. As a result, from

the function to be maximized and the constraint function, the Lagrange function was formed as follows.

$$F(t_i, z_{ki}) = \sum_{i=1}^{n}\sum_{k=1}^{2}[\hat{z}_{ki}\delta_i \ln h_k(t_i) + \hat{z}_{ki}\ln S_k(t_i) + \hat{z}_{ki}\ln \pi_k] + \lambda\left(\sum_{k=1}^{2}\pi_k - 1\right), \quad (12)$$

where λ was the Lagrange Multiplier. Maximizing (12) concerning θ_k was equivalent to maximizing the independence of each expression k.

The determination of the mixing probabilities π_k and parameter vector θ was carried out through the application of the Lagrange method. In addition, the mixing probabilities will be obtained by

$$\hat{\pi}_k = \frac{\sum_{i=1}^{n}\hat{z}_{ki}}{n}. \quad (13)$$

In this study, the first part on the left side l and the second part on the right side h were assigned. When (11) was derived against θ_k, then the value of h equals zero. To maximize (11) taking into account θ_k was equivalent to maximizing the independence of each expression k below,

$$l_k = \sum_{i=1}^{n}\sum_{k=1}^{2}[\hat{z}_{ki}\delta_i \ln h_k(t_i) + \hat{z}_{ki}\ln S_k(t_i)] \quad (14)$$

or

$$l = \sum_{i=1}^{n}([\hat{z}_{1i}\delta_i \ln h(t_i;\alpha,\beta) + \hat{z}_{1i}\ln S(t_i;\alpha,\beta)] + [\hat{z}_{2i}\delta_i \ln h(t_i;\mu,\sigma) + \hat{z}_{2i}\ln S(t;\mu,\sigma)]).$$

The maximum likelihood estimator of the parameter $\theta_1 = \left(\hat{\alpha}, \hat{\beta}\right)$ of the log-logistic distribution in the proposed model were determined by assuming $\hat{z}_{1i}\delta_i \ln\left[\frac{\frac{\alpha}{\beta}\left(\frac{t_i}{\beta}\right)^{\alpha-1}}{1+\left(\frac{t_i}{\beta}\right)^{\alpha}}\right] + \hat{z}_{1i}\ln\left[\frac{1}{1+\left(\frac{t_i}{\beta}\right)^{\alpha}}\right] = l_1$, then

$$\frac{\partial l_1}{\partial \alpha} = \hat{z}_{1i}\delta_i\left(\frac{1}{\alpha} + \ln\frac{t_i}{\beta}\right) - \hat{z}_{1i}(\delta_i + 1)\frac{\left(\frac{t_i}{\beta}\right)^{\alpha}}{1+\left(\frac{t_i}{\beta}\right)^{\alpha}}\ln\frac{t_i}{\beta} \quad (15)$$

and

$$\frac{\partial l_1}{\partial \beta} = -\hat{z}_{1i}\delta_i\frac{\alpha}{\beta} + \hat{z}_{1i}(\delta_i+1)\frac{\alpha}{\beta}\frac{\left(\frac{t_i}{\beta}\right)^{\alpha}}{1+\left(\frac{t_i}{\beta}\right)^{\alpha}}. \quad (16)$$

The maximum likelihood estimator of the parameter $\theta_2 = (\hat{\mu}, \hat{\sigma})$ of the log-normal distribution for the proposed model could be found by considering $\hat{z}_{2i}\delta_i \ln\left[\frac{\phi\left[\frac{\ln t_i-\mu}{\sigma}\right]}{\sigma t_i\left(1-\Phi\left[\frac{\ln t_i-\mu}{\sigma}\right]\right)}\right] + \hat{z}_{2i}\ln\left[1-\Phi\left[\frac{\ln t_i-\mu}{\sigma}\right]\right] = l_2$ first, followed by

$$\frac{\partial l_2}{\partial \mu} = \hat{z}_{2i}\delta_i\frac{1}{\sigma}\left(\frac{\ln t_i - \mu}{\sigma}\right) + \hat{z}_{2i}(1-\delta_i)\frac{1}{\sigma}\frac{\phi\left[\frac{\ln t_i-\mu}{\sigma}\right]}{1-\Phi\left[\frac{\ln t_i-\mu}{\sigma}\right]} \quad (17)$$

and

$$\frac{\partial l_2}{\partial \sigma^2} = \hat{z}_{2i}\delta_i\frac{1}{2\sigma^2}\left[-1+\left(\frac{\ln t_i-\mu}{\sigma}\right)^2\right] + \hat{z}_{2i}(1-\delta_i)\frac{1}{2\sigma^2}\frac{\left(\frac{\ln t_i-\mu}{\sigma}\right)\phi\left[\frac{\ln t_i-\mu}{\sigma}\right]}{1-\Phi\left[\frac{\ln t_i-\mu}{\sigma}\right]}. \quad (18)$$

Given the results of the derivative of the log-likelihood concerning the parameters α, β, μ, and σ^2 which were equal to zero in closed form could not be fulfilled (see (15)-(18)). As a result, a numerical method was needed, in this case, the Newton-Raphson method. Estimation of parameter values using Newton Raphson required a matrix of the first and second derivative of the maximized ln-likelihood equation. The value $\hat{\theta}$ was an estimator, when $\hat{\theta}$ provided the maximum value of $l(\theta)$. In addition, the value of $\hat{\theta}$ was maximum when the second partial derivative matrix of $l(\theta)$, namely the Hessian matrix, was a negative definite matrix, and in this case, it was fulfilled.

IV. Real Data Application and Discussion

In this study, data from 225 DLBCL patients treated at Dr. Sardjito Hospital between 2012 and 2020 were analyzed and described using a 2-component model combining Log-logistic and Log-normal distributions ($k = 2$), as shown in TABLE II. The histogram and 3 density functions, illustrating the probability of a better fit for the survival times of patients compared to others, were shown in Fig. 1. Meanwhile, the survival function of each model was shown in Fig. 2.

Based on the graphical comparison in Fig. 1, it could be observed that the empirical density function closely matched the density functions of the non-identical mixture model. Moreover, it appeared that the survival function of the mixture model, and the classical log-logistic/log-normal model also fitted the empirical survival function, as shown in Fig. 2. Since the survival graphs of the identical and non-identical mixture models closely approximated the real data, model selection criteria, and significance tests were applied to determine the optimal model (see TABLE III).

The criteria for selecting the best model used the MSE and K-S value of each model, while the model with the smallest MSE and K-S value was good as presented in TABLE III.

Based on the MSE and K-S values in TABLE III, the non-identical mixture model emerged as one of the suitable models for DLBCL data. The second smallest MSE corresponded to the log-logistic model, followed by its identical mixture model. However, the estimated value was not sufficiently small. The model that was reasonably well-fitted and could also be considered based on MSE and its statistical significance was the standard log-normal model and its identical mixture. Among the fitted models, the most suitable one was the non-identical mixture model, which emerged as an alternative for modeling and analyzing the survival time of DLBCL patients in Indonesia.

Although the difference in MSE between the mixture and standard models was small, the non-identical mixture models still showed superiority in long-term prediction. Guidez et al. [22], noted that the mixture models were able to extrapolate relapse estimates, even though the observational data were limited to a shorter period. Other studies have also shown

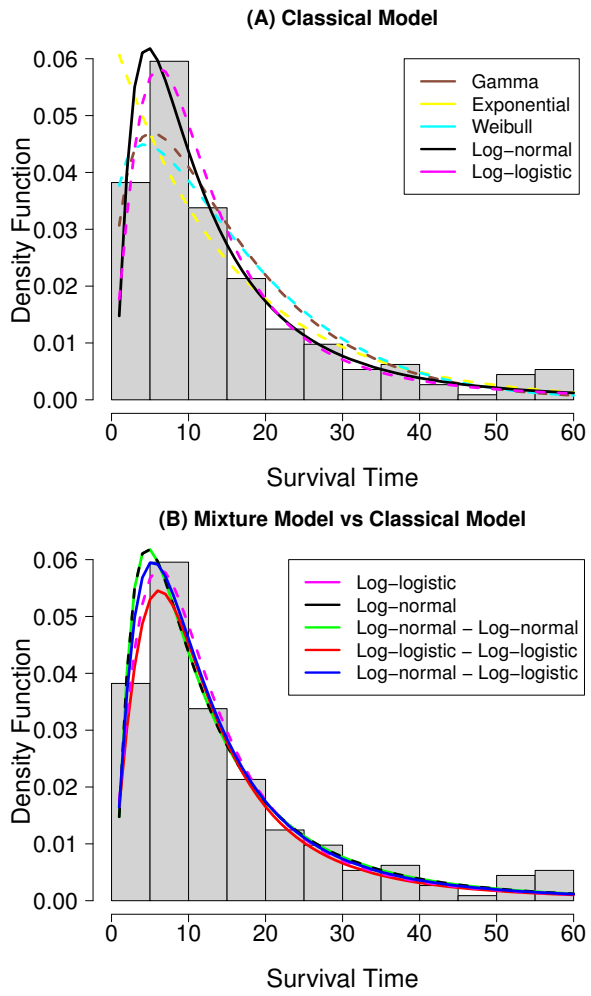

Fig. 1: The probability density functions of the fitted distributions: Classical model (A) and Mixture Model versus Classical Model (B) for survival times of 225 DLBCL patients.

that a mixture of 2 non-identical distributions was as useful as a mixture of identical distributions [23].

Fig. 2 showed that the survival function graphs of both the mixture models and the single log-logistic/log-normal models fit relatively well with the Kaplan-Meier curve of the data. The survival of all models revealed the highest probability of survival (best prognosis) within the first year, but the lowest probability of survival (poor prognosis) thereafter.

Fig. 3 presented the cumulative hazard (A-C) and hazard function (D) derived from a model using a combination of log-normal and log-logistic distributions applied to the survival time data of DLBCL patients. The cumulative hazard between the mixture model and the empirical data was compared in this study. In addition, the results generally indicated that the survival time data fitted the mixture model, particularly the Log-normal-Log-logistic model. The cumulative hazard value in DLBCL cases increased over time, with an initial peak at 10 months, followed by a continuous rise, indicating that the risk or danger continued to increase as time progressed. This represented a situation where the longer a patient suffered from DLBCL, the greater the probability of death.

In this study, it was observed that the cumulative hazard

TABLE III: Estimated parameters, K-S, and MSE values for the survival times of 225 patients with DLBCL.

Distributions	Parameter Estimation	K-S	MSE
Exponential	$\lambda = 0.064710958$	0.0009221	0.001183402
Gamma	$\alpha = 1.506565549$, $\beta = 10.257483$	0.07875	0.0008052499
Weibull	$\lambda = 1.228679$, $\gamma = 16.605594$	0.05087	0.0009668863
Log-normal	$\mu = 2.3706153$, $\sigma = 0.9102271$	0.0784	0.0002095301
Log-logistic	$\alpha = 11.009$, $\beta = 1.954$	0.4661	0.0001927844
Lnorm-Lnorm	$\pi_1 = 0.7004305$, $\pi_2 = 0.2995695$, $\mu_1 = 2.390732$, $\sigma_1 = 0.9086296$, $\mu_2 = 2.32358$, $\sigma_2 = 0.9122216$	0.07852	0.0002079922
Llogis-Llogis	$\pi_1 = 0.5714271$, $\pi_2 = 0.4285729$, $\alpha_1 = 1.954213$, $\beta_1 = 11.0094$, $\alpha_2 = 1.954213$, $\beta_2 = 1.954213$	0.4669	0.0001931253
Llogis-Lnorm	$\pi_1 = 0.5058073$, $\pi_2 = 0.4941927$, $\alpha = 1.952846$, $\beta = 10.97164$, $\mu = 2.367018$, $\sigma = 0.9109686$	$< 2.2e{-}16$	0.0001833573

value was less than 1 in the 1 year following diagnosis (the first 10 months). Subsequently, in the second year after diagnosis, the cumulative hazard value increased by an estimated 2. In this case, it was said that there was a significant change in the second year. This suggested that the risk of death or other events for these patients increased rapidly in the second year, which revealed the development of more serious DLBCL or a response to less effective treatment. This increase could indicate changes in behavior or other factors that elevated the risk of death in the second year.

The hazard rate function in this survival analysis generally followed a modal pattern (Fig. 3 (D)). This pattern illustrated that the hazard rate increased gradually, reaching a peak in a certain period, and then tended to decrease over time. Early observations indicated a low hazard rate (0.03), suggesting a relatively low risk of death for patients suffering from DLBCL. This could be due to several factors, including effective initial treatment (CHOP/R-CHOP), relatively mild patient characteristics, or a disease that had not yet reached a severe stage. The peak increase in hazard rate occurred between the 5th month and the 18th month (approximately $0.075-0.085$) which indicated changes in disease behavior or other factors influencing mortality risk. In this month, there was a significant increase in the risk of death, which revealed a worsening in the condition of the patient, potentially caused by complications or a poor response to more effective treatment. A decrease in the estimated hazard rate after 18 months to reach less than 0.04 at the end time indicated that the risk of death in patients decreased or stabilized after reaching a peak. This could be caused by several factors, such as a positive response to subsequent treatment, the body's

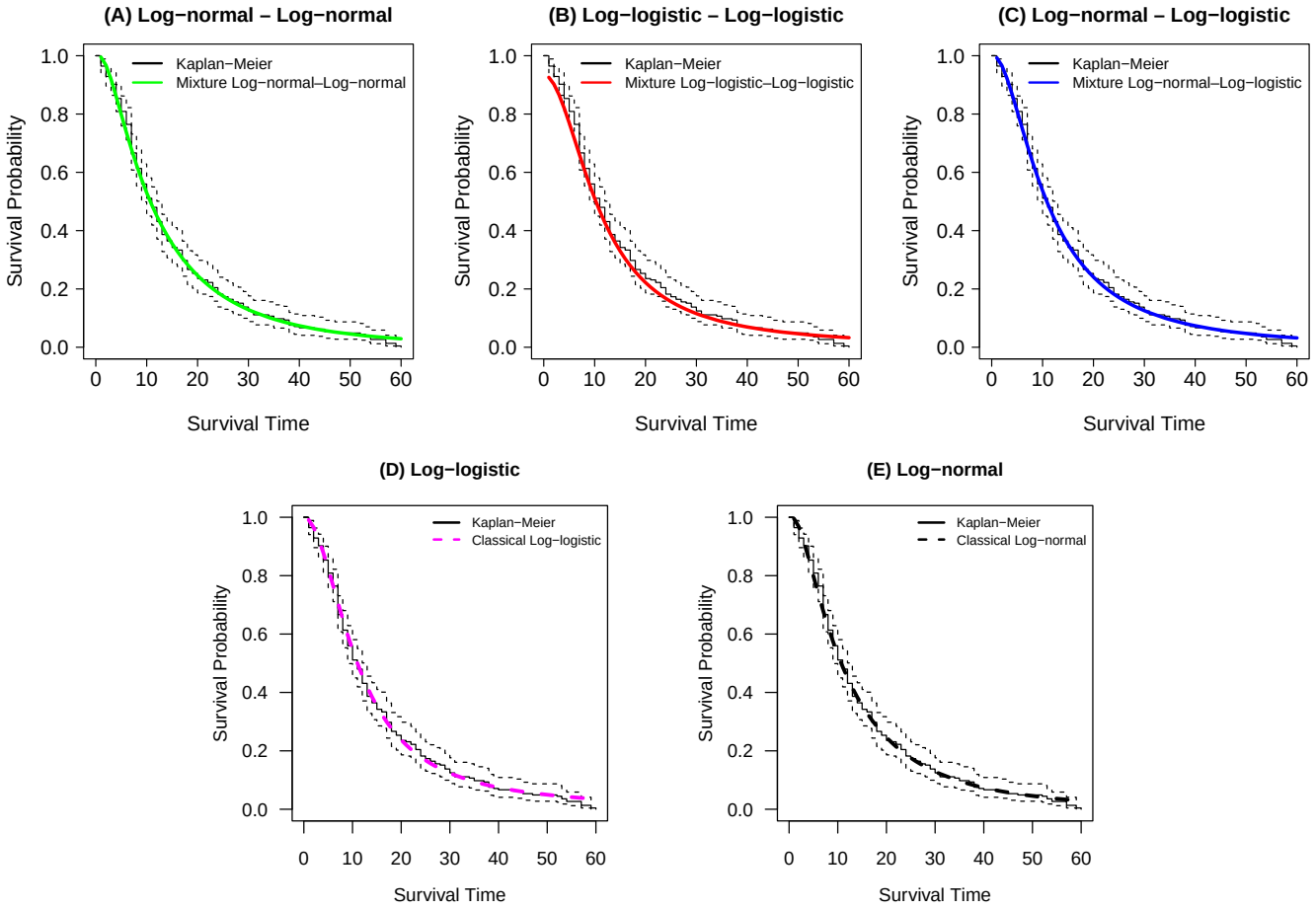

Fig. 2: The Survival of the fitted distributions: Mixture Model (A-C) versus Classical Model (D,E) for survival data of 225 DLBCL patients.

adaptation to the disease, or the effects of palliative care between the 5th month and the 18th month.

Further analysis was conducted by constructing a life table to illustrate the survival patterns in this dataset. The life table presented the probability of survival ($S(t)$) at each time interval (in months), as well as the probability of an event occurring (q_t) in each interval. Using the model estimation results, this table helped identify high-risk periods and provided a more structured understanding of survival trends. The following were the results of the life table based on the estimated non-identical mixture survival model, as seen in TABLE IV.

Based on the results of TABLE IV, the survival probability decreased over time, with a sharper decline in the early intervals. A similar trend was also observed in the previously presented survival function, displayed in Fig. 2. This suggested that the estimation of the non-identical mixture model effectively represented the distribution of survival time based on real data. In the 5th month, approximately 81% of individuals remained alive. However, by the 10th month, the survival probability decreased sharply to 53.68%, indicating that almost half of the individuals had experienced an event. By the 30th month, the percentage reduced to approximately 12.55%, and by the 60th month, only about 3.2% of individuals remained alive. This indicated that the majority of the population experienced an event before reaching 60 months. The highest probability of an event (q_t) occurred within the first 10 months, showing that this period was a critical phase with the greatest risk. After 30 months, q_t remained above 20% but showed a decreasing trend, suggesting the presence of a group of individuals who survived longer than the majority of the population.

TABLE IV: Life table results were based on the non-identical mixture model.

Interval	q_t	$S(t)$	Standard Error
5	0.0000000	0.81008837	0.019402894
10	0.8629979	0.53676927	0.022127018
15	0.5207312	0.35296788	0.022104180
20	0.4665894	0.24067259	0.021015041
25	0.4090981	0.17079903	0.019071616
30	0.3605490	0.12553685	0.016879335
35	0.3211230	0.09502283	0.014774992
40	0.2889971	0.07371842	0.012887782
45	0.2624522	0.05839304	0.011246963
50	0.2401754	0.04708450	0.009840441
55	0.2212077	0.03855568	0.008641675
60	0.2048498	0.03200041	0.007621268

This information was very useful for understanding the survival time pattern of DLBCL patients in Indonesia, particularly in identifying the period of highest risk. The 5 to 10-month interval represented a critical phase that required more attention in clinical intervention planning. By understanding

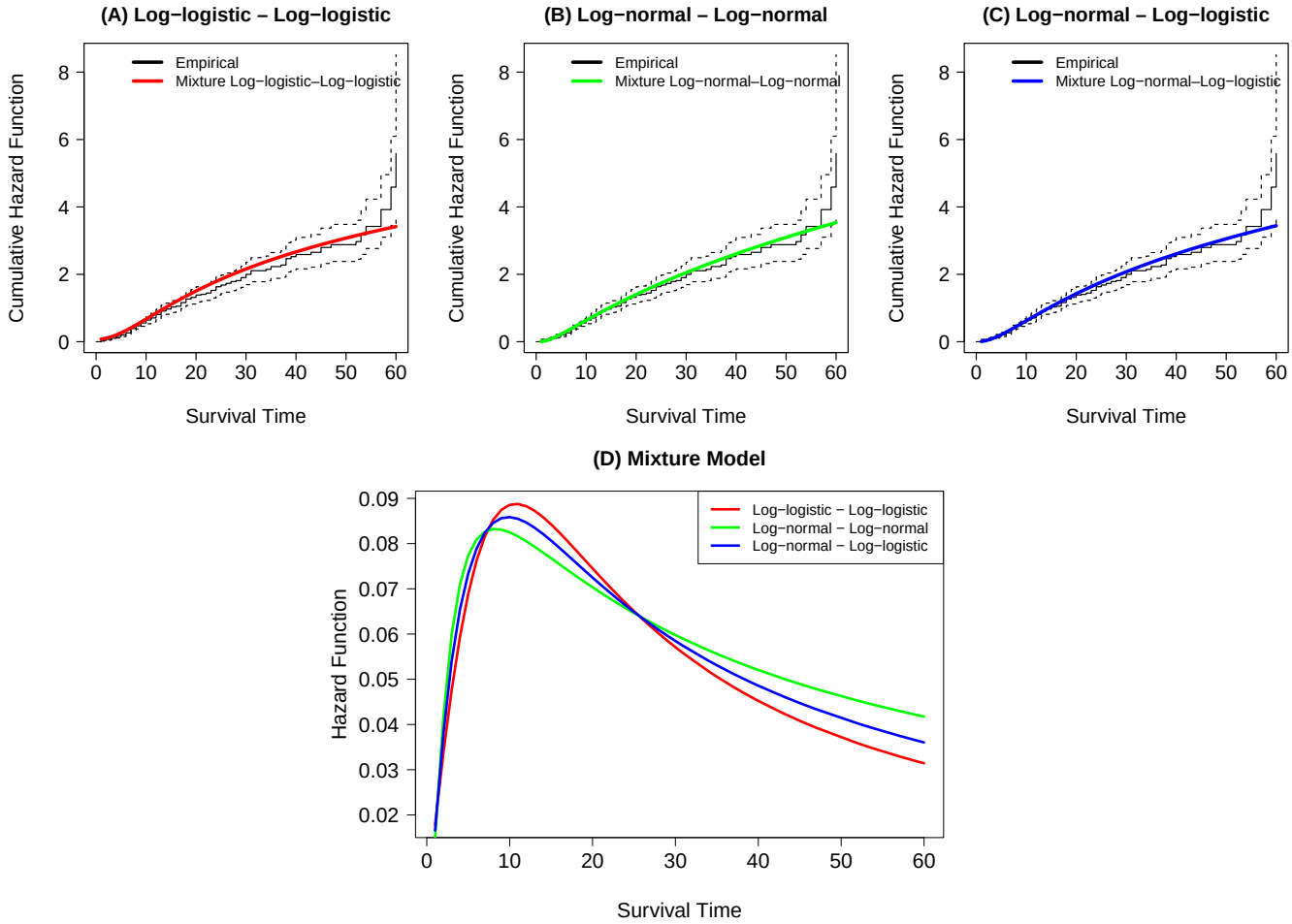

Fig. 3: Cumulative hazard function (A-C) and hazard function (D) for mixture model.

this pattern, clinical decisions were more targeted, including optimizing patient care and monitoring strategies during the highest-risk period, ultimately improving overall patient outcomes.

V. Conclusion

In conclusion, a combination of 2 distinct distributions was applied, namely log-normal and log-logistic, to model the survival time of DLBCL patients in Indonesia. This mixture model represented a novel approach to modeling survival data in this context, as it had not been widely applied to patient survival data in Indonesia. The parameters were estimated using the Expectation-Maximization algorithm. These findings showed that the mixed Log-logistic and Log-normal models were strong candidates for modeling DLBCL survival time data as determined by the MSE and the K-S test. The hazard rate obtained from patient survival data was unimodal, which indicated that there was an increase in the hazard rate at the beginning of time and then slowly decreased at the end of time. One of the effects of CHOP/R-CHOP therapy, which could promote patient healing and ultimately increase the survival rate of patients, could have an impact on this.

Other factors not included in this study also needed to be considered to fully understand the phenomenon that occurred. Overall, the outcome of this study offered deeper insight into the characteristics of survival time data from individuals diagnosed with DLBCL. In addition, knowledge of the parametric mixture model could help patients understand their estimated prognosis, prepare mentally, and make decisions related to treatment with more accurate information. Future work must focus on incorporating covariates as prognostic factors. These were compared with widely applied semi-parametric survival models, such as the Cox model, while also conducting further analysis using larger datasets and appropriate statistical methods to identify factors influencing changes in hazard rates. The comprehensive approach provided valuable insights into DLBCL patient care and disease management at both individual and population levels.

References

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 3, pp. 229–263, 2024.

[2] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," *CA: A Cancer Journal for Clinicians*, vol. 73, no. 1, pp. 17–48, 2023.

[3] S. A. Padala and A. Kallam, *Diffuse Large B-Cell Lymphoma*. Treasure Island, FL: StatPearls Publishing, 2023, https://www.ncbi.nlm.nih.gov/BOOKs/NBK557796/.

[4] M. Crump, S. S. Neelapu, U. Farooq, E. V. D. Neste, J. Kuruvilla, J. Westin, B. K. Link, A. Hay, J. R. Cerhan, L. Zhu, S. Boussetta, L. Feng, M. J. Maurer, L. Navale, J. Wiezorek, W. Y. Go, and C. Gisselbrecht, "Outcomes in refractory diffuse large b-cell lymphoma: Results from the international scholar-1 study," *Blood*, vol. 130, no. 16, pp. 1800–1808, 2017.

[5] L. H. Sehn, B. Berry, M. Chhanabhai, C. Fitzgerald, K. Gill, P. Hoskins, R. Klasa, K. J. Savage, T. Shenkier, J. Sutherland, R. D. Gascoyne, and J. M. Connors, "The revised international prognostic index (r-ipi) is a better predictor of outcome than the standard ipi for patients with diffuse large b-cell lymphoma treated with r-chop," *Blood*, vol. 109, no. 5, pp. 1857–1861, 2007.

[6] Z. Zhou, L. H. Sehn, A. W. Rademaker, L. I. Gordon, A. S. Lacasce, A. Crosby-Thompson, A. Vanderplas, A. D. Zelenetz, G. A. Abel, M. A. Rodriguez, A. Nademanee, M. S. Kaminski, M. S. Czuczman, M. Millenson, J. Niland, R. D. Gascoyne, J. M. Connors, J. W. Friedberg, and J. N. Winter, "An enhanced international prognostic index (nccn-ipi) for patients with diffuse large b-cell lymphoma treated in the rituximab era," *Blood*, vol. 123, no. 6, pp. 837–842, 2014.

[7] C. Montalbán, A. Díaz-López, I. Dlouhy, J. Rovira, A. Lopez-Guillermo, S. Alonso, A. Martín, J. M. Sancho, O. García, J. M. Sánchez, M. Rodríguez, S. Novelli, A. Salar, A. Gutiérrez, M. J. Rodríguez-Salazar, M. Bastos, J. F. Domínguez, R. Fernández, S. G. de Villambrosia, J. A. Queizan, R. Córdoba, R. de Oña, A. López-Hernandez, J. M. Freue, H. Garrote, L. López, A. M. Martín-Moreno, J. Rodriguez, V. Abraira, J. F. García, and G.-I. P. Investigators, "Validation of the nccn-ipi for diffuse large b-cell lymphoma (dlbcl): the addition of β2-microglobulin yields a more accurate geltamo-ipi," *British journal of haematology*, vol. 176, no. 6, pp. 918–928, 2017.

[8] D. Collett, *Modelling Survival Data in Medical Research*. Chapman & Hall/CRC Texts in Statistical Science.

[9] D. G. Kleinbaum and M. Klein, *Survival Analysis: A Self-Learning Text*, third edition ed. Springer, 2012.

[10] C. Cox, H. Chu, M. F. Schneider, and A. Muñoz, "Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution," *Statistics in Medicine*, vol. 26, no. 23, pp. 4352–4374, 2007.

[11] M. A. Pourhoseingholi, E. Hajizadeh, B. M. Dehkordi, A. Safaee, A. Abadi, and M. R. Zali, "Comparing cox regression and parametric models for survival of patients with gastric carcinoma," *Asian Pacific Journal of Cancer Prevention (APJCP)*, vol. 8, no. 3, pp. 412–416, 2007.

[12] Z. Zhang, "Parametric regression model for survival data: Weibull regression model as an example," *Statistics in Medicine*, vol. 4, no. 24, p. 484, 2016.

[13] N. Taketomi, K. Yamamoto, C. Chesneau, and T. Emura, "Parametric distributions for survival and reliability analyses, a review and historical sketch," *Mathematics*, vol. 10, no. 20, p. 3907, 2022.

[14] P. R. Cislo, B. Emir, J. Cabrera, B. Li, and D. Alemayehu, "Finite mixture models, a flexible alternative to standard modeling techniques for extrapolated mean survival times needed for cost-effectiveness analyses," *Value in Health: the journal of the International Society for Pharmacoeconomics and Outcomes Research*, vol. 24, no. 11, pp. 1643–1650, 2021.

[15] R. D. Angelis, R. Capocaccia, T. Hakulinen, B. Soderman, and A. Verdecchia, "Mixture models for cancer survival analysis: application to population-based data with covariates," *Statistics in Medicine*, vol. 18, no. 4, pp. 441–454, 1999.

[16] J. M. Marín, M. T. Rodríguez-Bernal, and M. P. Wiper, "Using weibull mixture distributions to model heterogeneous survival data," *Communications in Statistics-Simulation and Computation*, vol. 34, no. 3, pp. 673–684, 2005.

[17] Ülkü Erişoğlu and H. Erol, "Modeling heterogeneous survival data using mixture of extended exponential-geometric distributions," *Communications in Statistics - Simulation and Computation*, vol. 39, no. 10, pp. 1939–1952, 2010.

[18] A. Bansal, S. D. Sullivan, V. W. Lin, A. G. Purdum, L. Navale, P. Cheng, and S. D. Ramsey, "Estimating long-term survival for patients with relapsed or refractory large b-cell lymphoma treated with chimeric antigen receptor therapy: A comparison of standard and mixture cure models," *Medical Decision Making: an international journal of the Society for Medical Decision Making*, vol. 39, no. 3, pp. 294–298, 2019.

[19] L. Sanchez, L. Muchene, P. Lorenzo-Luaces, C. Viada, P. C. Rodriguez, S. Alfonso, T. Crombet, E. Neninger, Z. Shkedy, and A. Lage, "Differential effects of two therapeutic cancer vaccines on short- and long-term survival populations among patients with advanced lung cancer," *Seminars in Oncology*, vol. 45, no. 1–2, pp. 52–57, 2018.

[20] L. Sanchez, P. Lorenzo-Luaces, C. Fonte, and A. Lage, "Mixture survival models methodology: an application to cancer immunotherapy assessment in clinical trials," 2019, https://arxiv.org/abs/1911.09765.

[21] D. Kadkhoda, M. Nikoonezhad, A. R. Baghestani, S. Parkhideh, Z. Momeni-Varposhti, and A. A. K. Maboudi, "Prognostic factors for the long-term survival after hematopoietic stem cell transplantation in patients with hodgkin lymphoma," *Asian Pacific Journal of Cancer Prevention (APJCP)*, vol. 24, no. 2, pp. 417–423, 2023.

[22] S. Guidez, S. Glaisner, E. V. D. Neste, E. Gyan, Z. Marjanovic, L.-M. Fornecker, E. Deconinck, M. Fabbro, V. Dorvaux, D. Robu, H. Yokoyama, N. A. Johnson, M. C. Cheung, S. Snauwaert, M. Casanova, Y. Terui, G. Yamamoto, Y. Choudhary, J. R. Mace, D. P. Quick, F. Morschhauser, and Y. Foucher, "Mixture model to predict the cumulative incidence of relapses in follicular lymphoma: Need for longer follow-up or alternative outcomes," *Blood*, 2023.

[23] A. H. Türkan and N. Çalış, "Comparison of two-component mixture distribution models for heterogeneous survival datasets: a review study," *Istatistik Journal of The Turkish Statistical Association*, vol. 7, no. 2, pp. 33–42, 2014.

[24] S. F. Ateya and A. S. Alharthi, "Estimation under a finite mixture of modified weibull distributions based on censored data via em algorithm with application," *Journal of Statistical Theory and Applications*, vol. 13, no. 3, pp. 196–204, 2014.

[25] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–222, 1977.

[26] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the em algorithm," *SIAM Review*, vol. 26, no. 2, pp. 195–239, 1984.

[27] J. F. Lawless, *Statistical Models and Methods for Lifetime Data*, second edition ed. John Wiley & Sons, 2002.

Sulasri Suddin is a doctoral candidate at the Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Yogyakarta, Indonesia. She is also affiliated with Universitas Timor, Kefamenanu, Nusa Tenggara Timur, Indonesia, as a lecturer. Her research interests include survival analysis, mixture models, and biomathematics.