Scientific Research Publishing

Open Journal of Applied Sciences, 2021, 11, 126-148
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

Parametric and Non-Parametric Survival Analysis of Patients with Acute Myeloid Leukemia (AML)

Aditya Chakraborty*, Chris P. Tsokos

Department of Mathematics & Statistics, University of South Florida, Tampa, FL, USA
Email: *adityachakra@usf.edu, ctsokos@usf.edu

How to cite this paper: Chakraborty, A. and Tsokos, C.P. (2021) Parametric and Non-Parametric Survival Analysis of Patients with Acute Myeloid Leukemia (AML). *Open Journal of Applied Sciences*, **11**, 126-148.
https://doi.org/10.4236/ojapps.2021.111009

Received: December 12, 2020
Accepted: January 26, 2021
Published: January 29, 2021

Abstract

Background: Acute Myeloid leukemia (AML) is the most prominent acute leukemia in adults. In the United States, we experience over 20,000 cases per year. Over the past decade, improvements in the diagnosis of subtypes of AML and advances in therapeutic approaches have improved the outlook for patients with AML. However, despite these advancements, the survival rate among patients who are less than 65 years of age is only 40 percent. **Purpose:** The purpose of the paper is to study if there exists any significant difference in the survival probabilities of male and female AML patients. Also, we want to investigate if there is any parametric probability distribution that best fits the male and female patient survival and compare the survival probabilities with the non-parametric Kaplan-Meier (KM) method. **Methods:** We used both parametric and non-parametric statistical methods to perform the survival analysis to assess the survival probabilities of 2015 patients diagnosed with AML. **Results:** We found evidence of a statistically significant difference between the mean survival time of male and female patients diagnosed with AML. We performed parametric survival analysis and found a Generalized Extreme Value (GEV) distribution best fitting the data of the survival time for male and female patients. We then estimated the survival probabilities and compared them with the frequently used non-parametric Kaplan-Meier (KM) survival method. **Conclusion:** The comparison between the survival probability estimates of the two methods revealed a better survival probability estimate by the parametric method than the Kaplan-Meier. We also compared the median survival time of male and female patients individually with descriptive, parametric, and non-parametric methods of analysis. The parametric survival analysis is more robust and efficient because it is based on a well-defined parametric probabilistic distribution, hence preferred over the non-parametric Kaplan-Meier estimate. This study offers therapeutic signi-

ficance for further enhancement to treat patients with Acute Myeloid Leukemia.

Keywords

Acute Myeloid Leukemia (AML), Generalized Extreme Value (GEV) Distribution, Probability Weighted Moment (PWM) Method, Kaplan-Meier (KM) Estimate

1. Introduction

Leukemias are certain types of cancers that start in the cells that naturally develop into different types of blood cells. Most commonly, leukemia starts in early forms of white blood cells, but there are some leukemias that grow in other blood cells. There are different types of leukemia that are divided primarily based on whether the leukemia is acute (rapidly growing) or chronic (slower growing), and whether it starts in myeloid cells or lymphoid cells. Acute myeloid leukemia (AML) develops in the bone marrow (the soft inner part of certain bones, where new blood cells are formed), but most often, it rapidly moves into the blood, as well. It can sometimes spread to other organs that include the lymph nodes, liver, spleen, central nervous system (brain and spinal cord), and testicles. Acute myeloid leukemia (AML) [1] is the most common acute leukemia in adults, accounting for almost 80 percent of the cases in this group. Within the United States, the incidence of AML ranges from three to five cases per 100,000 population. In 2015 alone, an estimated 20,830 new cases were diagnosed, and over 10,000 patients died from this disease. To realize how and why leukemia affects a patient, it is essential to understand how blood cells are made. The body manufactures blood cells in the bone marrow (the soft inner part of bones). The blood cells are produced in a controlled way, as the body needs them. Every blood cell starts as the same type of cell called a stem cell. This stem cell then develops into the following.

- Myeloid stem cells, which become white blood cells called monocytes and neutrophils (granulocyte), red blood cells, and platelets. And,
- Lymphoid stem cells, which become white blood cells called lymphocytes.

Figure 1 describes the process of forming blood cells from stem cells.

In the case of acute myeloid leukemia, the bone marrow produces a plethora of monocytes or granulocytes. These cells are often not fully developed and are not able to function regularly. Figure 2 illustrates a possible path of development of AML from a stem cell.

Figure 1 & Figure 2 have been obtained from [2].

In the United States, AML increases progressively with age, to a peak of 12.6 per 100,000 adults 65 years of age or older [3]. Until the 1970s, the diagnosis was based solely on the pathological and cytologic examination of bone marrow and blood.

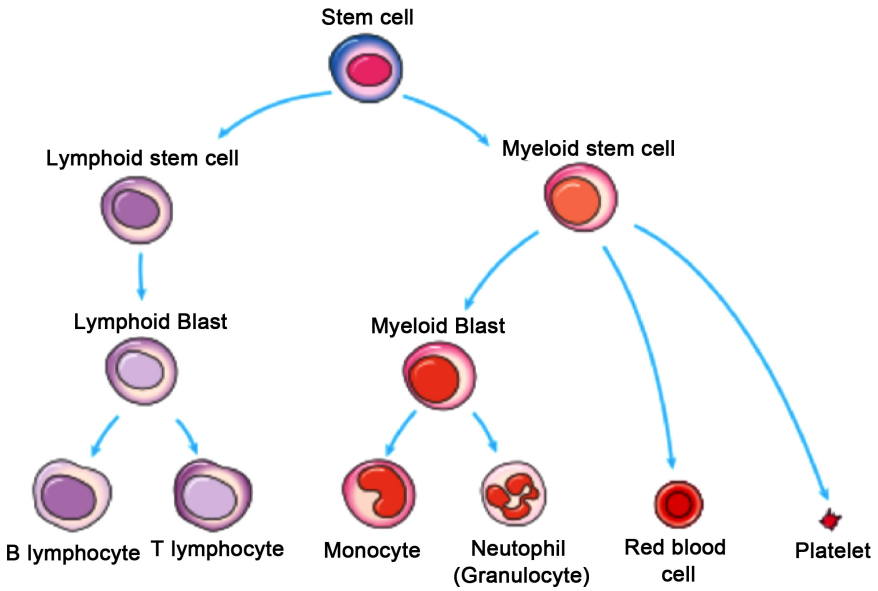

Figure 1. Steps of forming blood cells from stem cells.

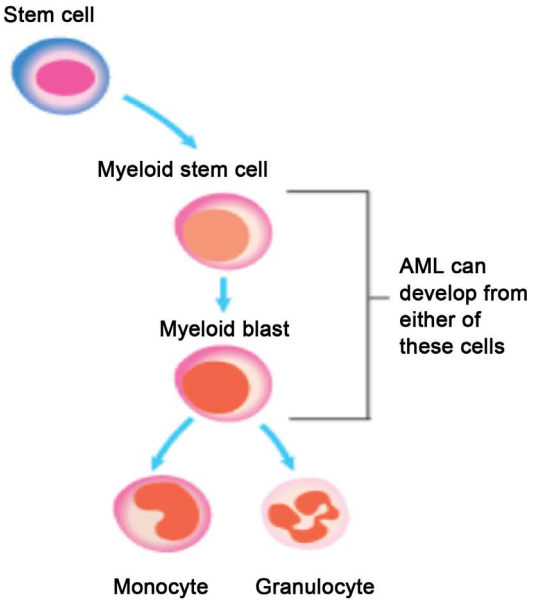

Figure 2. Development of AML from a stem cell.

A five-year survival rate during this period was less than 15 percent. Over the past decade, improvements in the diagnosis of subtypes of AML and advances in therapeutic approaches have improved the outlook for patients with AML. However, despite these advancements, the survival rate among patients who are less than 65 years of age is only 40 percent. Although in most of the cases, AML cancer disease remains irremediable, most researches into AML concentrated on how to improve the survival time of patients diagnosed with AML. The Kaplan-Meier (KM) method has been widely used for analyzing cancer survivorship data in recent time due to the simplicity of its usage. It is often used to compare

the survival difference of several groups of patients based on the log-rank test of the null hypothesis that there is no significant difference among the groups. Our study presents a parametric and non-parametric survival analysis of the survival time of patients diagnosed with AML. We believe that finding the unique probability distribution that characterizes the probabilistic behavior of the survival time is essential so that we can proceed to obtain the survival function that is driven by the given data. Such an analysis is more powerful than the non-parametric approach. Feigl and Zelen, [4] have pointed out that assuming exponential distribution works well for studying the survival of cancer related cases [5] [6]. Assuming such a probability distribution without justification will lead to misleading results. Thus, it is important to identify the probability distribution of the survival time among any number of groups (for male/female or age < 65, age > 65). Hence, the probability distribution for a given set of survival time without justification will lead to an incorrect decision. In the present study, we identify the probability distribution that fits the survival time the best and proceeds to obtain the survival function. We also compare our results with the commonly used Kaplan-Meier (KM) method. The structure of the paper will be as follows: In Section 2, we provide the data discussion and perform the log-rank test [7] [8]. In Section 3, we discuss in detail the parametric survival analysis of male and female AML patients. Section 4 talks about the KM estimate and compares the median survival time of male and female patients using the descriptive, parametric, and non-parametric methods. In Section 5, we compare the survival probability estimates of male and female patients using parametric GEV distribution and non-parametric KM estimate. Section 6 and Section 7 provide results & discussion, and conclusion, respectively.

2. Method

Data Description

The data for our study has been extracted from the Surveillance, Epidemiology and End Results (SEER) database. The data contains information on patients diagnosed with AML from 2004 to 2015. We are concerned with the survival time (in months) and cause-specific death (deaths due to AML cancer) for each patient. The survival time of patients is one of the most crucial factors used in all cancer research. It is necessary to evaluate the severity of cancer, which helps to decide the prognosis and help identify the correct treatment methods. We considered a random sample of 2015 patients diagnosed with Acute Myeloid Leukemia (AML) which accounts for almost 80% of the Acute Leukemia cases, [9]. A schematic diagram of the data used in this study with additional details is shown in **Figure 3**. The data for our study has been extracted from the Surveillance, Epidemiology and End Results (SEER) database. The data contains information on patients diagnosed with AML from 2004 to 2015. We are concerned with the survival time (in months) and cause-specific death (deaths due to AML cancer) for each patient. The survival time of patients is one of the most

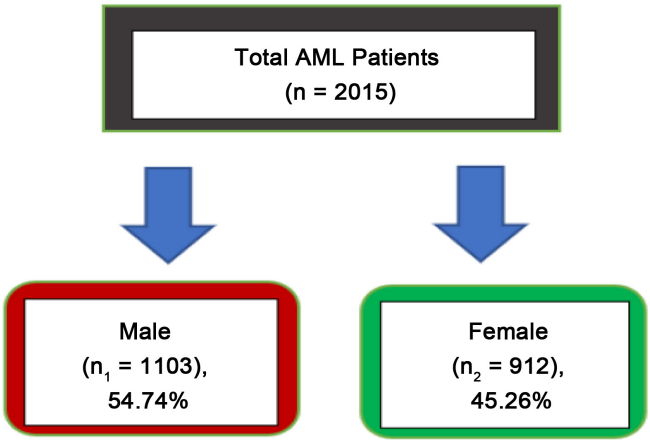

Figure 3. AML patient data sorted by gender.

crucial factors used in all cancer research. It is necessary to evaluate the severity of cancer, which helps to decide the prognosis and help identify the correct treatment methods. As the following schematic diagram illustrates, in our data-set, we have information on survival time regarding 1103 male and 912 female patients diagnosed with AML.

Before we proceed with performing the parametric analysis of the survival time of patients with AML, we need to investigate whether there is a difference in the survival time of gender, *i.e.*, male and female patients. For this purpose, We use the Log Rank test using the following two hypotheses.

H_0: There is no significant difference between the mean survival time of male (μ_M) and mean survival time of female (μ_F) patients. That is, $\mu_M = \mu_F$

Vs.

H_1: Difference exists between male and female survival time. That is, $\mu_M \neq \mu_F$.

The log-rank test produced a p-value of 0.011 (<0.05), implying that there is sufficient sample evidence to reject H_0, which means the distribution of survival time between the Male and Female patients diagnosed with AML is significantly different. Figure 4 illustrates the behavior of survival curves of male and female patients. The male and female survival curves are highlighted in blue and yellow, respectively.

As Figure 4 illustrates, the survival curve of males (blue) lies below the survival curve of females (yellow), which means males have lower survival compared to females diagnosed with AML. In the following section, we describe the parametric analysis of survival time for both genders.

3. Parametric Analysis of the Survival Time

3.1. Descriptive Statistics of the Survival Time of AML Patients

We plotted the histogram and probability density function (pdf) to investigate the distribution of the survival time of male and female patients, as shown in Figure 5 and Figure 6. We can see that the probability distribution of the survival time of AML for both males and females is right skewed.

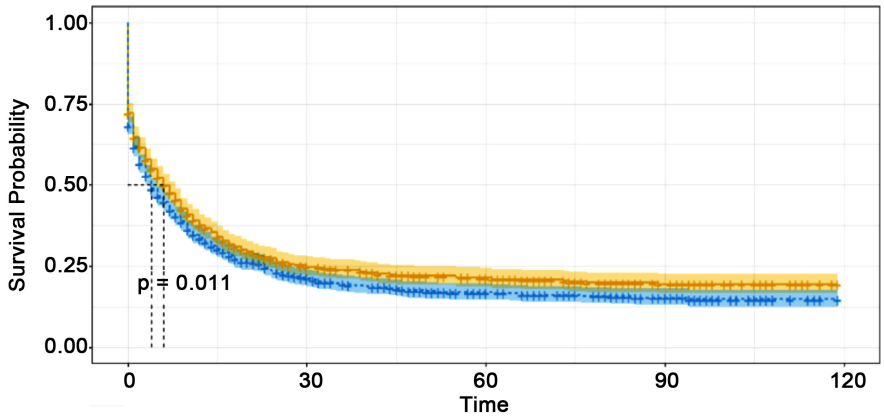

Figure 4. Log-rank test for difference in survival time of gender.

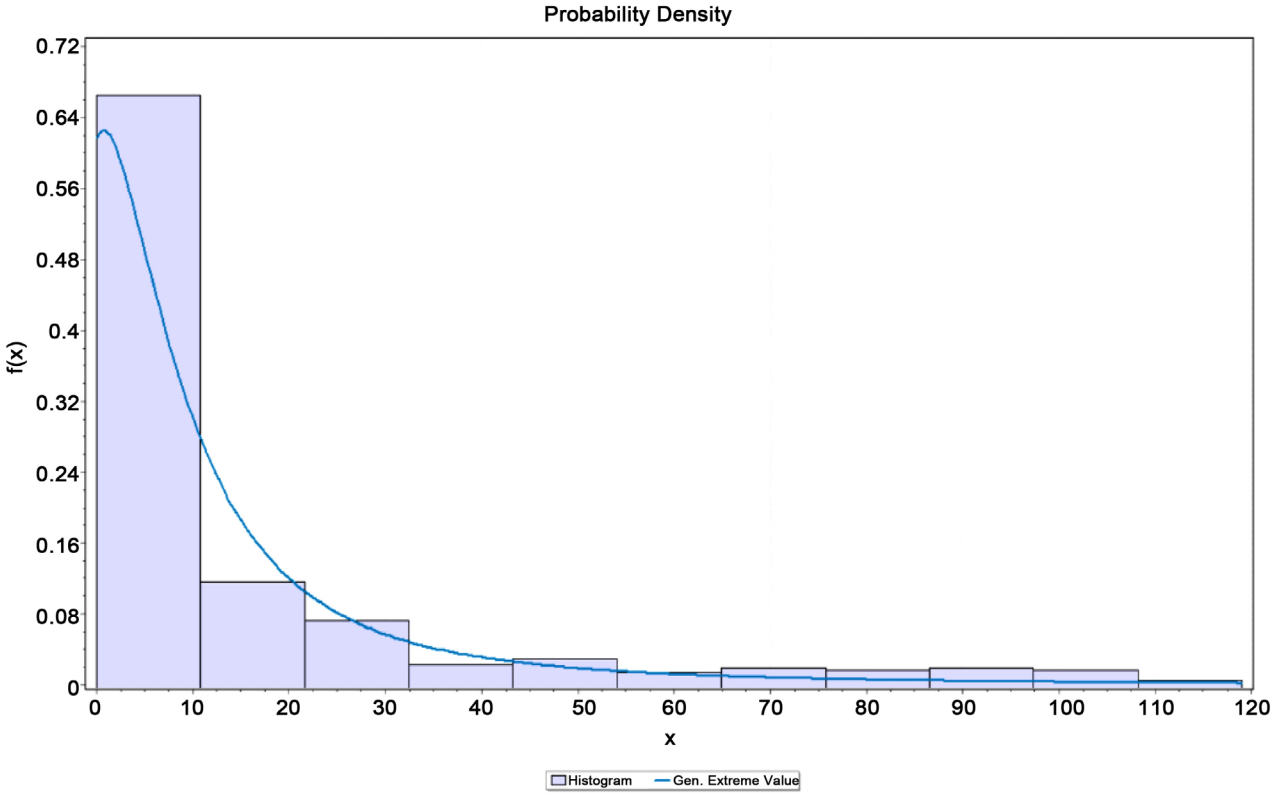

Figure 5. Histogram and probability density of male survival time of AML patients.

Table 1 displays the descriptive statistics of the survival time of AML for Male and females. We see that the mean (average) survival time for male and female patients diagnosed with AML is 15.16 months and 17.61 months, respectively. It means that a randomly chosen patient diagnosed with AML, a Male, is expected to survive for 15.16 on an average. Similarly, a randomly chosen female patient diagnosed with AML is expected to survive for 17.61 months on average. Also, the median survival time for male and female patients are four months and five months, respectively, which implies that the probability/chance of survival of a male or female AML patient beyond 4 and 5 months, respectively, is approximately

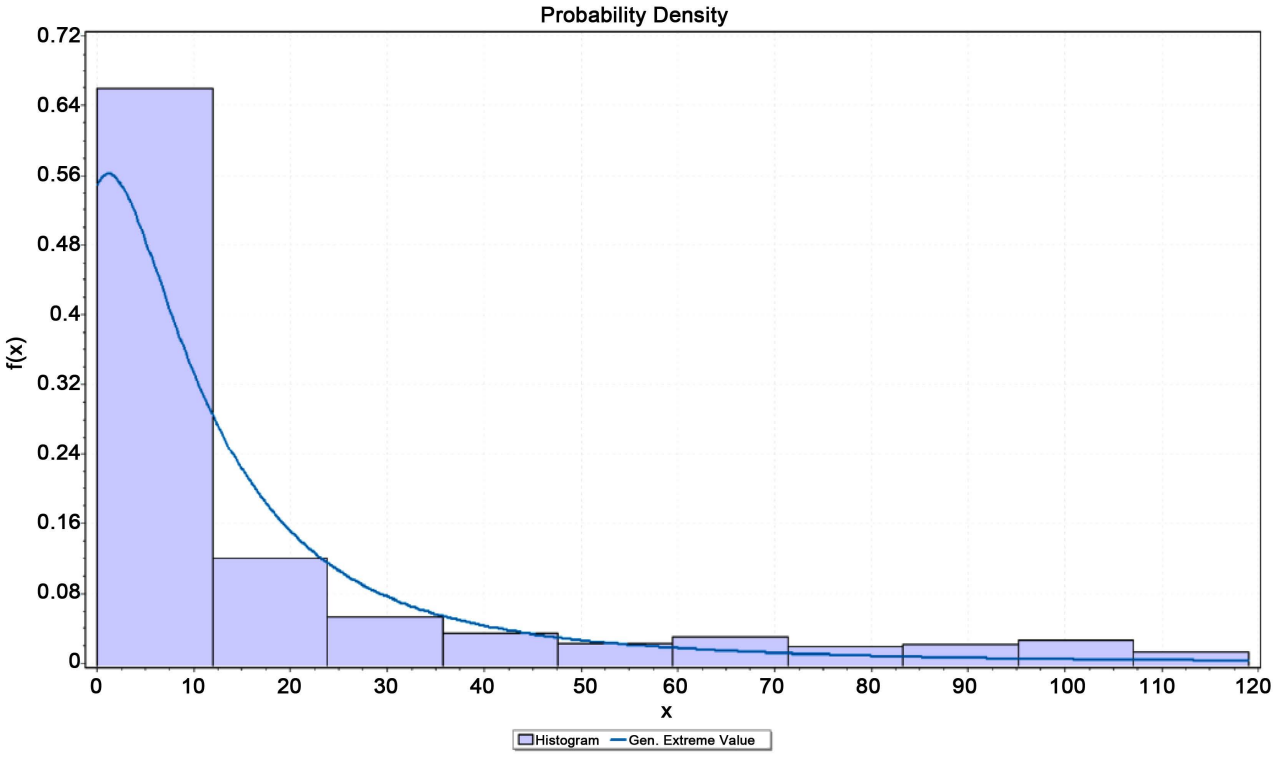

Figure 6. Histogram and probability density of female survival time of AML patients.

Table 1. Descriptive statistics of survival time (in month) of AML patients classified by gender.

SEX	Mean	Median	Std. Dev.	Skewness	Kurtosis	Std. Error
Male	15.16	4	25.22	2.2	4.21	0.76
Female	17.61	5	27.89	1.98	2.99	0.92

50%. A negative (less than zero) skewed value implies that data distribution is left or negatively skewed, and a positive skewed value suggests that data is right or positive skewed. Thus, the positive skewed value of 2.2 and 1.98, as shown in Table 1, for male and female patients, respectively, is further evidence to support the right-skewed behavior of the data, as shown in Figure 5 and Figure 6. Kurtosis supports the assessment of the extreme values of the data, and its positive value illustrates a leptokurtic behavior of the distribution. In contrast, a negative value shows a platykurtic behavior of the data distribution. Thus, the kurtosis value of 4.21 and 2.99 for males and females, respectively, in Table 1 attests to the AML survival time data's leptokurtic behavior.

3.2. Generalized Extreme Value (GEV) Probability Estimation of the Survival Time of Patients with AML

We perform a parametric analysis of the survival time of patients diagnosed with AML to identify the underlying probability distribution, which characterizes the probabilistic behavior of the survival time of AML patients (both genders). In the attempt to obtain the best-fitted probability distribution, a number of clas-

sical distributions were tested to fit the subject data. The three commonly used goodness-of-fit tests, Kolmogorov-Smirnov test, Anderson-Darling test, and Chi-Square fitness test, were used to identify the best probability distribution function that characterizes the probabilistic behavior of the survival time of male and female patients. Also, we estimate the expected survival time and median survival time under each identified probability distribution function. The best fitted probability distribution that characterizes the probabilistic behavior of the survival time of the male and female patients accurately is the Generalized Extreme Value (*GEV*) distribution. We choose the Kolmogorov-Smirnov test, Anderson-Darling test, and Chi-Square fitness test to identify the best probability distribution function as they are very widely used and popular non-parametric goodness of fit (GOF) [10] [11] tests. **Table 2** shows the goodness of fit (GOF) results of the *GEV* distribution.

The above results show that we fail to reject the null hypothesis that the subject data (survival time for males and females) follow a GEV distribution. In this section, we define the probability density function (pdf) of the Generalized Extreme Value (GEV) distribution and the statistical approach to obtain approximate estimates of its parameters. In the domain of probability theory and statistics, the Generalized Extreme Value (GEV) distribution is a family of continuous probability distributions developed based on the extreme value theory, [12]. The distribution combines three probability distribution families, namely, Gumbel, Fréchet, and Weibull. They are also known as type I, II, and III extreme value distributions. GEV distribution was first introduced by Jenkinson, [13] however, in some fields of application, the generalized extreme value distribution is known as the Fisher-Tippett distribution [14], named after Ronald Fisher and L. H. C. Tippett, who recognized the three different forms of the distribution. Let T be a random variable following GEV distribution with location parameter ξ, scale parameter $\alpha > 0$, and shape parameter k. That is,

$X \sim \mathrm{GEV}(\xi, \alpha, k)$ with domain $\left\{t : 1 - k\left(\frac{t-\xi}{\alpha}\right) > 0\right\}$ when $k \neq 0$, and $-\infty < t < \infty$, when $k = 0$. Then, the probability density function (pdf) is given as follows:

$$f_{\mathrm{GEV}}(t;\xi,\alpha,k) = \begin{cases} \frac{1}{\alpha}\exp\left[-\left\{1-k\left(\frac{t-\xi}{\alpha}\right)\right\}^{\frac{1}{k}}\right]\left\{1-k\left(\frac{t-\xi}{\alpha}\right)\right\}^{-1+\frac{1}{k}}, & k \neq 0 \\ \frac{1}{\alpha}\exp\left[-\left\{\left(\frac{t-\xi}{\alpha}\right)-\exp\left(\frac{t-\xi}{\alpha}\right)\right\}\right], & k = 0 \end{cases} \quad (1)$$

Table 2. Goodness-of-fit test of the *GEV* distribution of the survival time of male and female.

GOF Tests	P-Value Male	P-Value Female
Kolmogorov-Smrinov	0.5156	0.3972
Anderson-Darling	0.2214	0.1956
Chi Squared	0.3856	0.2946

The corresponding cumulative distribution function (cdf) is given as follows:

$$F_{\mathrm{GEV}}\left(t;\xi,\alpha,k\right)=\begin{cases}\exp\left[-\left\{1-k\left(\dfrac{t-\xi}{\alpha}\right)\right\}^{\frac{1}{k}}\right], & k\neq 0\\ \exp\left[-\left\{-\exp\left(\dfrac{t-\xi}{\alpha}\right)\right\}\right], & k=0\end{cases} \tag{2}$$

There are several methods to estimate the parameters ξ, α, and k of the GEV distribution. Some of these methods include Jenkinson's (1969) method of sextiles and the method of maximum likelihood (Jenkinson 1969; Prescott and Walden 1980, 1983). Neither of these methods is completely accurate [15] [16]. We use the Probability-Weighted Moments (PWM) method [15], introduced by Greenwood *et al.* (1979), which is a generalization of the method of moments of a probability distribution to estimate the set of parameters.

3.3. Parameter Estimation of GEV Distribution Using the Method of Probability Weighted Moments (PWM)

In general, the probability-weighted moments of a random variable X with cumulative distribution function $F\left(x\right)=P\left(X\le x\right)$ is given by,

$$M_{p,r,s}=E\left[X^{p}\left\{F\left(x\right)\right\}^{r}\left\{1-F\left(x\right)\right\}^{s}\right] \tag{3}$$

where p, r, and s are real numbers. Probability-weighted moments [16] [17] are most useful when it is written as a function of the inverse distribution function $F^{-1}\left(x\right)=x\left(F\right)$ in closed form in the following way.

$$M_{p,r,s}=\int_{0}^{1}\left\{x\left(F\right)\right\}^{p}F^{r}\left\{1-F\right\}^{s} \tag{4}$$

The two special cases of $M_{p,r,s}$ which are commonly used are

$$\begin{aligned}\alpha_{r}&=M_{1,0,s}=E\left[X\left\{1-F\left(x\right)\right\}^{s}\right],\ s=0,1,2,\cdots\\ \beta_{r}&=M_{1,r,0}=E\left[X\left\{F\left(x\right)\right\}^{r}\right],\ r=0,1,2,\cdots\end{aligned} \tag{5}$$

where X inside the $E\left[\cdot\right]$ is the inverse distribution of X, denoted by $x\left(F\right)$. To estimate the parameters of GEV distribution, we use β_{r} from (5) according to the approach used by Hosking *et al.* [15].

Given a random sample of size n from the cdf, F, the estimate of β_{r}, is based on the ordered sample $x_{1}\le x_{2}\le\cdots\le x_{n}$. The unbiased estimate of the statistic β_{r} (Landwehr *et al.* 1979) is b_{r} which is given by:

$$b_{r}=\frac{1}{n}\sum_{i=1}^{n}\frac{\left(j-1\right)\left(j-2\right)\cdots\left(j-r\right)}{\left(n-1\right)\left(n-2\right)\cdots\left(n-r\right)}x_{j}, \tag{6}$$

b_{r} will be used to estimate β_{r} which will lead us to achieve our goal successfully.

Instead of b_{r}, one might use the estimate

$$\widehat{\beta}_r\left[p_{j,n}\right]=\frac{1}{n}\sum_{j=1}^{n}p_{j,n}^{r}x_j, \tag{7}$$

where $p_{j,n}$ is a plotting position, that is, a distribution-free estimate of $F\left(x_j\right)$. From (2) we can solve for X to obtain the inverse cdf, $x\left(F\right)$. The inverse distribution function is given by,

$$x\left(F\right)=\begin{cases}\xi+\dfrac{\alpha}{k}\left\{1-\left(-\log\left(F\right)^{k}\right)\right\}, & \text{if } k\neq 0\\ \xi-\alpha\log\left(-\log\left(F\right)\right), & \text{if } k=0\end{cases} \tag{8}$$

Now we proceed to derive the analytical form of β_r for the GEV distribution using expressions (4) and (7). From (5), we have

$$\begin{aligned}\beta_r=M_{1,r,0}&=\int_0^1\xi+\frac{\alpha}{k}\left\{1-\left(-\log\left(F\right)^{k}\right)\right\}F^{r}\mathrm{d}F\\&=\int_0^\infty\xi+\frac{\alpha}{k}\left(1-u^{k}\right)\exp\left\{-\left(r+1\right)u\right\}\mathrm{d}u\\&\text{Substituting } u=-\log\left(F\right),\\&=\left(\xi+\frac{\alpha}{k}\right)\int_0^\infty\exp\left\{-\left(r+1\right)u\right\}\mathrm{d}u-\frac{\alpha}{k}\int_0^\infty u^{k}\exp\left\{-\left(r+1\right)u\mathrm{d}u\right\}\\&=\left(\xi+\frac{\alpha}{k}\right)\left(r+1\right)^{-1}-\frac{\alpha}{k}\left(r+1\right)^{-1-k}\Gamma\left(1+k\right),\ \text{provided } k>-1\\&=\left(r+1\right)^{-1}\left\{\xi+\alpha\left(1-\left(r+1\right)^{-k}\right)\frac{\Gamma\left(1+k\right)}{k}\right\}\end{aligned} \tag{9}$$

Thus, for $k\neq 0$; the probability-weighted moments of the GEV distribution is given by (9). When $k\leq -1$, it can be shown that, (the mean of the distribution, β_0) and the rest of the $\beta_r, r=0,1,2,\cdots$ do not exist. In Equation (9), substituting $r=0$; $r=1$; and $r=2$ we can obtain explicit expressions of β_0, β_1 and β_2 in terms of ξ, α, and k. That is,

$$\beta_0=\xi+\frac{\alpha}{k}\left[1-\Gamma\left(1+k\right)\right] \tag{10}$$

$$\left(2\beta_1-\beta_0\right)=\frac{\alpha}{k}\Gamma\left(1+k\right)\left(1-2^{-k}\right) \tag{11}$$

and

$$\frac{2\beta_2-\beta_0}{2\beta_1-\beta_0}=\frac{1-3^{-k}}{1-2^{-k}} \tag{12}$$

The PWM estimates of the parameters $\left(\hat{\xi},\hat{\alpha},\hat{k}\right)$ can be obtained by solving the Equations (10), (11) and (12) for ξ, α, and k by replacing β_r by their estimators b_r or $\widehat{\beta}_r\left[p_{j,n}\right]$ from (6) and (7). To estimate the shape parameter k, we have to solve,

$$\frac{2\beta_2-\beta_0}{2\beta_1-\beta_0}=\frac{1-3^{-\hat{k}}}{1-2^{-\hat{k}}} \tag{13}$$

The exact solution requires some iterative methods. Hosking *et al.* (1985) used

a low order polynomial approximation for $\hat{k}$ which is given by,

$$\hat{k} = 7.859c + 2.9554c^2, \text{ where } c = \frac{2b_1 - b_0}{3b_2 - b_0} - \frac{\log 2}{\log 3}. \tag{14}$$

Once, we have obtained $\hat{k}$, the estimates of scale and location parameters, $\hat{\xi}$ and $\hat{\alpha}$ we can be estimated successively from Equations (11) and (10), that is,

$$\begin{aligned} \hat{\alpha} &= \frac{\left(2b_1 - b_0\right)\hat{k}}{\Gamma\left(1+\hat{k}\right)\left(1-2^{-\hat{k}}\right)}, \\ \hat{\xi} &= b_0 + \frac{\hat{\alpha}}{\hat{k}}\left\{\Gamma\left(1+\hat{k}\right)-1\right\}, \end{aligned} \tag{15}$$

Table 3 shows the approximate parameter estimates of GEV distribution for male and female survival time.

We substituted the parameter estimates of ξ, α, k in (1) to obtain the analytical form of the probability density function (pdf) of male and female survival time. The analytical form of the GEV probability density function (pdf) for male survival time is given by:

$$\begin{aligned} f_{\text{Male}}(t) = \frac{1}{7.15}\exp\left[-\left\{1-0.52\left(\frac{t-3.44}{7.15}\right)\right\}^{\frac{1}{0.52}}\right] \\ \times\left\{1-0.52\left(\frac{t-3.44}{7.15}\right)\right\}^{\frac{1}{0.52}-1}, \quad -\infty < t < \infty \end{aligned} \tag{16}$$

Similarly, the analytical form of the GEV probability distribution function (pdf) for female survival time is given by:

$$\begin{aligned} f_{\text{Female}}(t) = \frac{1}{8.66}\exp\left[-\left\{1-0.5\left(\frac{t-4.34}{8.66}\right)\right\}^{\frac{1}{0.5}}\right] \\ \times\left\{1-0.5\left(\frac{t-4.34}{8.66}\right)\right\}^{\frac{1}{0.5}-1}, \quad -\infty < t < \infty \end{aligned} \tag{17}$$

The above probability density functions characterize the probabilistic behavior of the survival time of male and female patients with AML cancer.

We now proceed to calculate the expected survival time of male and female patients. Using estimates given in Table 3, we can find the expectation and median survival time for both male and female patients that follow $\text{GEV}(3.44, 7.15, 0.52)$ and $\text{GEV}(4.34, 8.66, 0.5)$ distribution, respectively. The expectation of a random

Table 3. Parameter estimates of GEV distribution of survival time of male and female AML patients.

Estimates	Male	Female
Location ($\hat{\xi}$)	3.44	4.34
Scale ($\hat{\alpha}$)	7.15	8.66
Shape ($\hat{k}$)	0.5	0.52

variable T following $\mathrm{GEV}(\xi,\alpha,k)$ is given by The above probability density functions characterize the probabilistic behavior of the survival time of male and female patients with AML cancer.

We now proceed to calculate the expected survival time of male and female patients. Using estimates given in Table 3, we can find the expectation and median survival time for both male and female patients that follow $\mathrm{GEV}(3.44,7.15,0.52)$ and $\mathrm{GEV}(4.34,8.66,0.5)$ distribution, respectively. The expectation of a random variable T following $\mathrm{GEV}(\xi,\alpha,k)$ is given by

$$E(T)=\hat{\xi}+(g_1-1)\frac{\hat{\alpha}}{\hat{k}},\ \hat{k}<1,\ \hat{k}\neq 0 \tag{18}$$

where $g_i=\Gamma(1-ik), i=1,2,3,4$ and $\Gamma(a)=\int_0^\infty t^{a-1}\mathrm{e}^{-t}\mathrm{d}t, a>0.$

Using Equation (18), the expected survival time of male and female AML patients are given by,

$$E_{\text{Male}}(T)=3.44+\{\Gamma(1-0.52)-1\}\frac{7.15}{0.52}=14.99 \text{ months}$$

and

$$E_{\text{Female}}(T)=4.34+\{\Gamma(1-0.5)-1\}\frac{8.66}{0.5}=17.68 \text{ months}$$

The median of the survival time T of $\mathrm{GEV}(\xi,\alpha,k)$ is given by,

$$f_{\text{GEV}}(t;\xi,\alpha,k)=\begin{cases}\xi+\dfrac{\alpha}{k}(\ln 2)^{-k}-1, & \text{if } k\neq 0\\ \xi-\alpha\ln(\ln 2), & \text{if } k=0\end{cases} \tag{19}$$

From Equation (19), the median survival time of male and female AML patients are given by,

$$\text{Med}_{\text{Male}}(T)=3.44+\frac{7.15}{0.52}(\ln 2)^{-0.52}-1=6.3 \text{ month}$$

and

$$\text{Med}_{\text{Female}}(T)=4.34+\frac{8.66}{0.5}(\ln 2)^{-0.52}-1=7.8 \text{ month}$$

Once we have the analytical forms of the pdf for males and females categories of survival time, we can obtain the cumulative distribution functions (cdf). The analytical form of the GEV cdf for Male survival time is given by,

$$F_{\text{Male}}(t)=\exp\left\{-\left(1-0.52\left(\frac{t-3.34}{7.15}\right)\right)^{\frac{1}{0.52}}\right\},\ -\infty<t<\infty \tag{20}$$

Similarly, the analytical form of the GEV cdf for female survival time is given as follows.

$$F_{\text{Female}}(t)=\exp\left\{-\left(1-0.5\left(\frac{t-4.34}{8.66}\right)\right)^{\frac{1}{0.5}}\right\},\ -\infty<t<\infty \tag{21}$$

Figure 7 and Figure 8, illustrate the cdf plots of the male and female survival time.

As the figures illustrate, the CDF plots are very helpful to estimate the probabilities that a certain male or female patient diagnosed with AML will survive up to a particular point of time. For example, from Figure 7, the probability that a male patient will survive up to time $t = 20$ months is approximately 0.8. However, this probability is slightly lower for a randomly selected female patient, which is evident from Figure 8. In the next section, we will present the parametric survival analysis of the survival time of males and females AML patients, which is one of the most important aspects of this study.

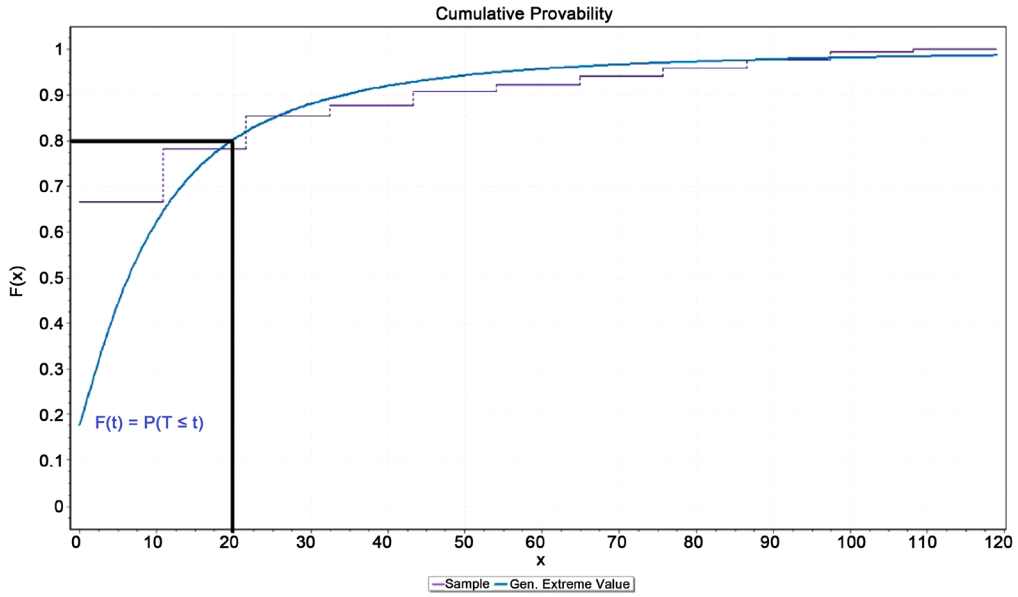

Figure 7. Cdf Plot for the survival time of male AML patients.

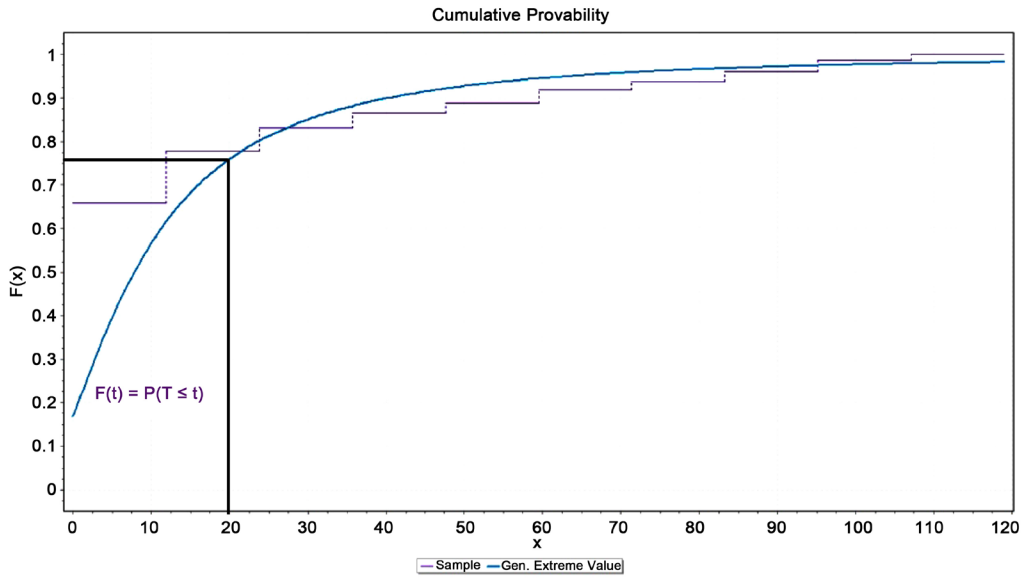

Figure 8. Cdf plot for the survival time of female AML patients.

3.4. Parametric Survival Analysis

Estimation of a parametric survival function is a process to evaluate the survival probabilities of male and female AML patients as a function of the survival time.

We have determined the cdf of the survival time for male and female patients diagnosed with AML in Equations (20) and (21), we can proceed to estimate the survival function of male and female AML patients.

Thus, the parametric survival function of male patients diagnosed with AML is given by,

$$\begin{aligned}\hat{S}_{\text{Male}}(t;\xi,\alpha,k) &= 1-\hat{F}_{\text{Male}}(t;\xi,\alpha,k)\\ &= 1-\exp\left\{-\left(1-0.52\left(\frac{t-3.44}{7.15}\right)\right)^{\frac{1}{0.52}}\right\},\ -\infty<t<\infty\end{aligned} \tag{22}$$

The survival function $S(\cdot,\cdot)$ can be used to estimate the probability that a male patient diagnosed with AML would survive beyond time t, which is denoted by $P(T \geq t)$. For example, we can compute the probability that a male patient diagnosed with AML would survive beyond 20 months. That is, for t = 20 in Equation (22), we estimate the probability as 0.2. Thus, we can infer that a randomly chosen male AML patient has a 20% chance of survival beyond 20 months. Figure 9 describes the parametric survival plot for male AML patients generated using *GEV* distribution.

Figure 9 attests the fact that the survival probability is 0.2 for a male AML patient. As expected, it can be seen that the survival function of the survival time is decreasing with time and approximately zero beyond time t = 100.

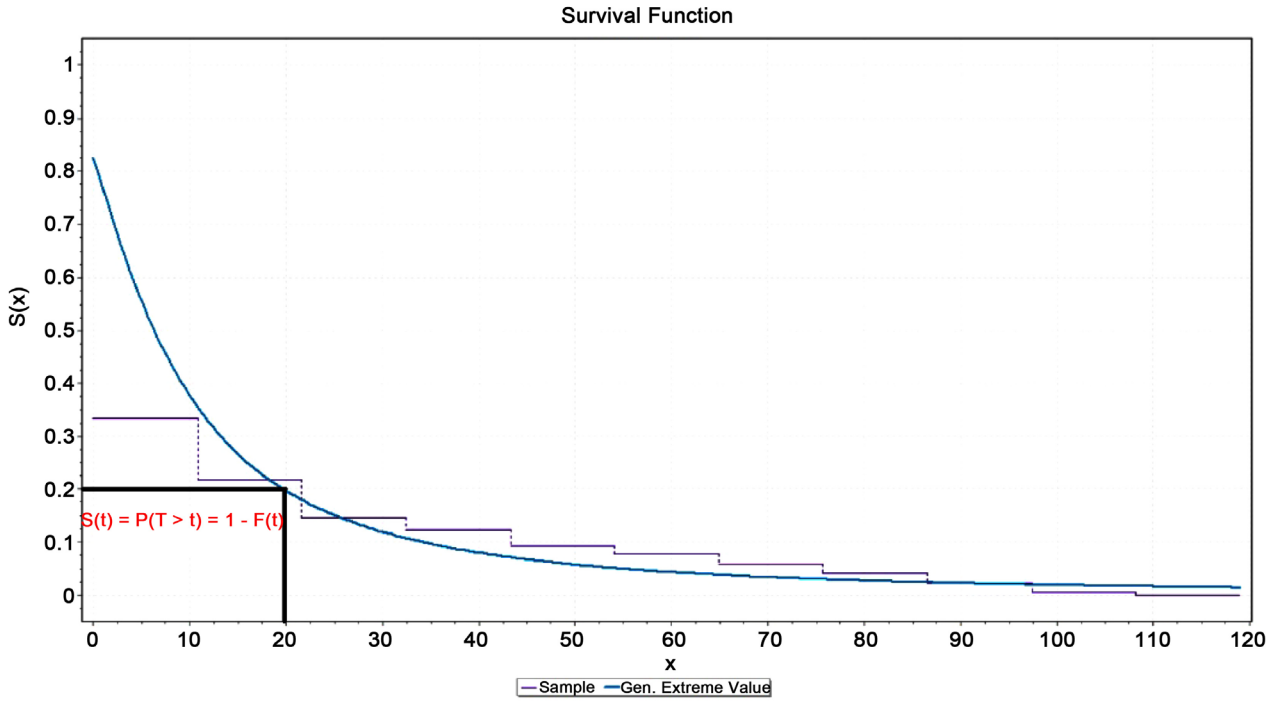

Figure 9. Parametric survival plot of male AML patients.

Similarly, the parametric survival function driven by the GEV distribution for female AML patients is given by,

$$\begin{aligned}\hat{S}_{\text{Female}}(t;\xi,\alpha,k) &= 1-\hat{F}_{\text{Female}}(t;\xi,\alpha,k)\\ &= 1-\exp\left\{-\left(1-0.5\left(\frac{t-4.34}{8.66}\right)\right)^{\frac{1}{0.5}}\right\},\ -\infty<t<\infty\end{aligned} \tag{23}$$

From the above parametric survival function, we can compute the probability that a female patient diagnosed with AML would survive beyond 20 months. By inserting $t = 20$ in Equation (23), we compute the probability of approximately 0.25, which is greater than the survival probability of a male AML patient. Thus, we can infer that a randomly chosen female AML patient has an approximately 25% chance of survival beyond 20 months. Figure 10 describes the parametric survival plot for female AML patients generated using the GEV distribution.

Figure 10 attests to the fact that the survival probability is approximately 0.25 for a female patient diagnosed with AML. Thus, a randomly chosen female AML patient has better survival than a male patient diagnosed with AML. In the next section, we discuss the non-parametric Kaplan-Meier Survival function for AML cancer briefly.

4. Kaplan-Meier Estimation of Survival Probability of the Survival Time of Patients with AML

The most frequently used parametric estimation methods for distributions of

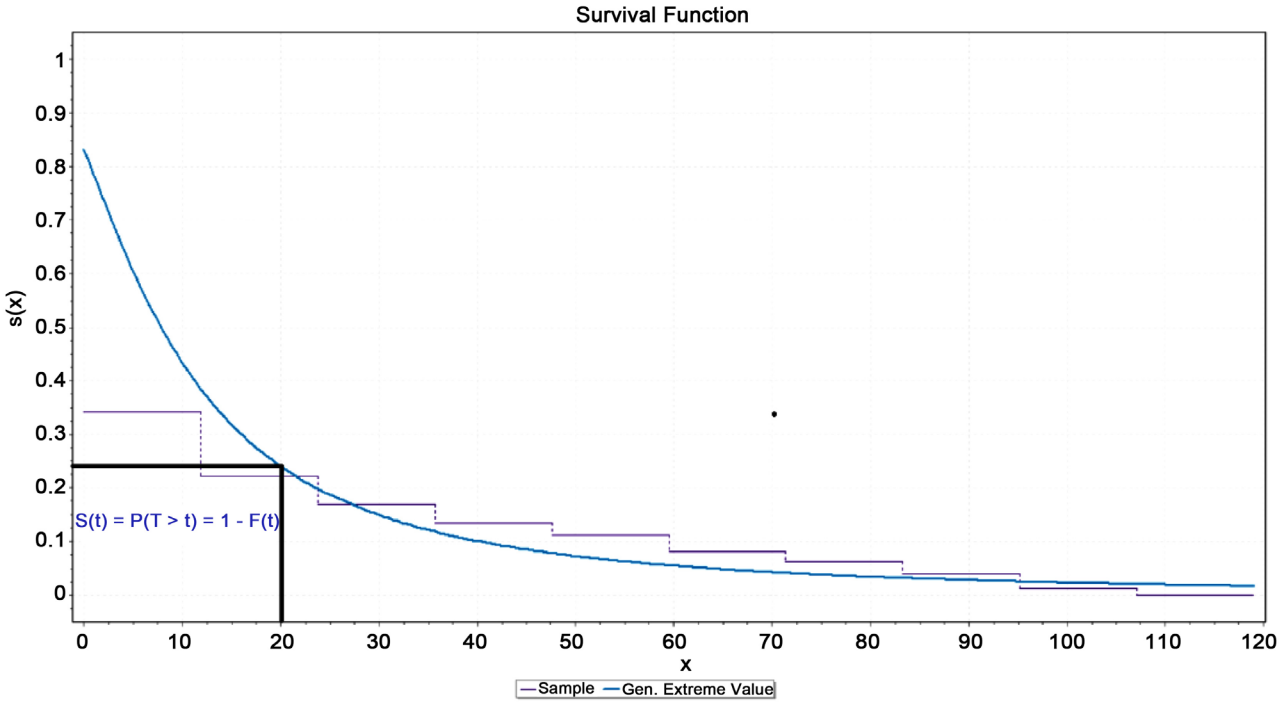

Figure 10. Parametric survival plot of female AML patients.

lifetimes are probably the fitting of a normal probability distribution to the observations or their logarithms by calculating the mean and variance and fitting an exponential distribution by estimating the mean alone. Such assumptions about the form of the distribution are naturally advantageous insofar as they are correct; the estimates are simple and relatively efficient, and a complete distribution is obtained even though the observations may be restricted in range. However, non-parametric estimates have the important functions of suggesting or confirming such assumptions and of supplying the estimate itself in case suitable parametric assumptions are not known. The Kaplan-Meier (KM) estimator [18] [19] also known as the product-limit estimator, is a non-parametric statistic used to estimate the survival function from data related to survival time. In health science, it is generally used to measure the fraction of patients living for a certain amount of time after treatment. It was developed by Edward L. Kaplan and Paul Meier (1958). It is defined as the product over the failure time of the conditional probabilities of surviving to the next failure time. Formally, it is given by,

$$\hat{S}(t)=\prod_{t_i\le t}\left(1-\hat{q}_i\right)=\prod_{t_i\le t}\left(1-\frac{\hat{d}_i}{n_i}\right) \tag{24}$$

where n_i is the number of patients at risk at time t_i, and d_i is the number of individual patients who fail(die) at that time.

Figure 11 demonstrates the survival curves with a risk table for both male and female patients diagnosed with AML.

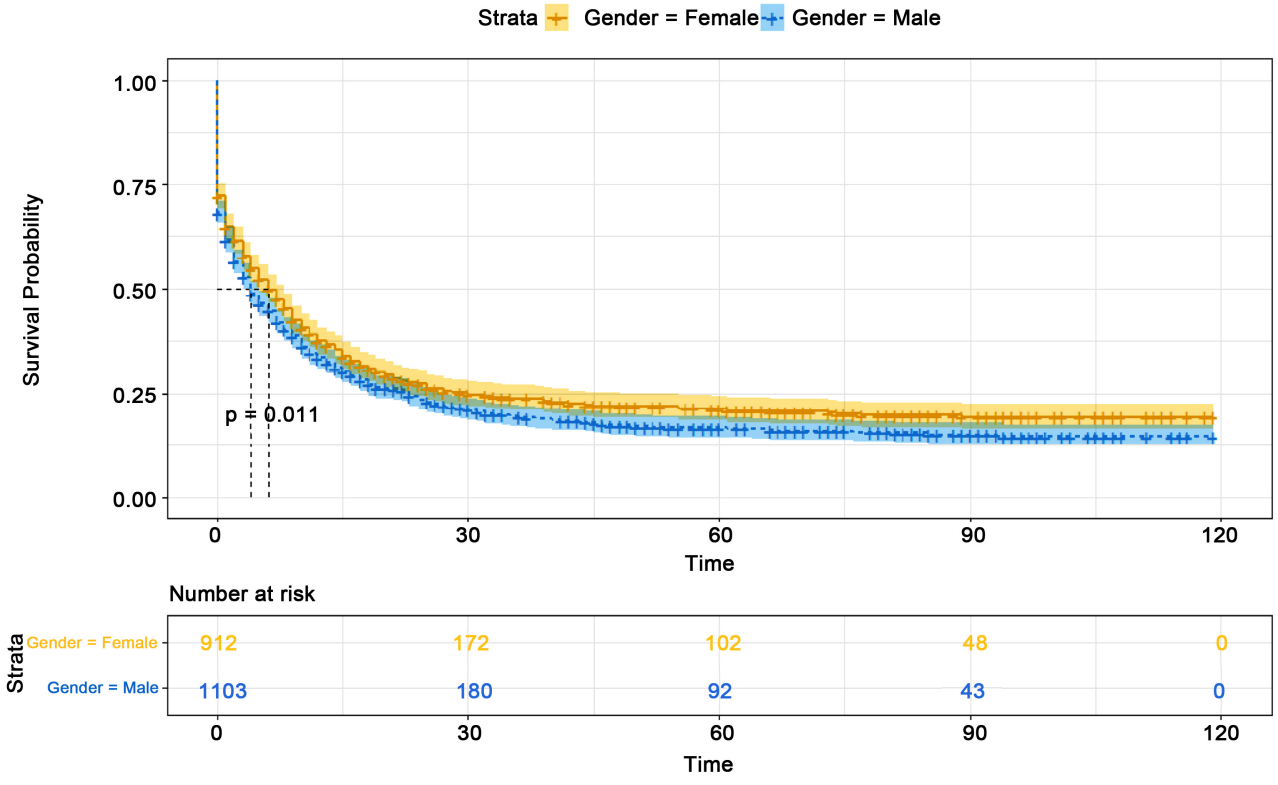

Figure 11. KM survival plot with risk table for both gender diagnosed with AML.

The figure provides information about how many people are at risk at a specific time, t for both male and female patients diagnosed with AML. For example, at time $t = 0$, the number of male and female patients at risk are 1103 and 912, respectively, which is the total number of male and female patients in our data set whom we started our initial analysis with. At the time $t = 60$ (months), the male and female patients that are at risk are respectively 102 and 92. It is important to note that with the passage of time, the number of people at risk gradually decreases for both categories, which is also evident from Figure 11, as the KP survival estimate $\hat{S}(t)$, is a function of the number of patients at risk (n_i).

Median Survival and a Confidence Interval for the Median Using KM Estimate

Median survival time is a statistic that indicates how long a group of patients will survive with an illness in general or after a specific treatment has been applied. It is usually expressed in months or years. Median survival time is when half the patients exposed to a certain disease are anticipated to be alive. It signifies that the probability of surviving beyond that time is 50 percent. It gives an approximate indication of survival and the prognosis of a group of patients with cancer. Median survival is frequently reported in almost every cancer treatment studies. Generally, the median survival time [20] is defined as, $\widehat{t_{med}} = \inf\left\{t : \hat{S}(t) \le 0.5\right\}$. It means that it is the smallest t such that the estimated survival function $\hat{S}(t)$ is less than or equal to 0.5. To compute a $100(1-\alpha)\%$ confidence interval for the median, we consider the following inequality:

$$-z_{\frac{\alpha}{2}} \le \frac{g\left\{\hat{S}(t)\right\} - g(0.5)}{\sqrt{\mathrm{Var}\left[g\left\{\hat{S}(t)\right\}\right]}} \le z_{\frac{\alpha}{2}} \tag{25}$$

where $g(u) = \log\left[-\log(u)\right]$ and $\mathrm{Var}\left[\left\{\hat{S}(t)\right\}\right]$ is given by the following equations:

$$\begin{aligned} &\mathrm{Var}\left\{\hat{S}(t)\right\} \approx \left[\hat{S}(t)\right]^2 \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)} \\ &\mathrm{Var}\left\{\log\left(-\log\left(\hat{S}(t)\right)\right)\right\} \approx \frac{1}{\log\left[\hat{S}(t)\right]^2} \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)} \end{aligned} \tag{26}$$

The confidence interval computed by the first variance formula in (26) might extend below zero or beyond 1. A more realistic approach to compute the variance formula, using the log-log transformation of $\hat{S}(t)$ in the second formula of (26). In order to compute a 95% confidence interval of the non-parametric survival function $\hat{S}(t)$, we look for the smallest value of t, such that the middle portion of the expression (21) is at least −1.96 (the lower limit) and the maximum value of t such that the middle expression does not go beyond 1.96 (the upper limit). The median survival time, computed using non-parametric KM estimator, for male and female patients diagnosed with AML, is given as four

months and six months, respectively. The corresponding 95% confidence interval for the median survival time is given as [4, 5] and [5, 8]. It is very interesting to note that the median survival time we obtained by the descriptive method (Table 1) are very close to what we obtained by non-parametric methods. However, the median survival time we obtained using the parametric method (implementing the GEV distribution) is slightly greater than the descriptive and non-parametric methods. Table 4 compares the median survival time for both male and female patients diagnosed with AML, computed using the three methods.

5. Comparison of GEV Distribution with the Kaplan-Meier Estimation of the Survival Function

In the parametric analysis (Section 3.3), we found that patients' survival time (both male and female) with acute myeloid leukemia follows a Generalized Extreme Value (GEV) distribution. In Section 3.4, we performed a non-parametric analysis using the Kaplan-Meier to estimate the AML patients' survival probability. We compare the survival probability estimates of the GEV distribution with the Kaplan-Meier survival estimates of the survival time of the AML patients. The importance of the survival function of the two methods is to estimate the survival probability of a patient diagnosed with AML beyond a given time. The survival probabilities corresponding to a specific time (in months) are shown in Table 5 for comparison purposes. We see that the probability estimates computed by the GEV survival function are higher than that of Kaplan-Meier in most cases. However, there are times in which the KM estimates higher survival probabilities than the GEV survival function. Since parametric methods are more powerful, robust, and efficient than non-parametric methods, we must accept the parametric estimates of the probabilities as the most accurate.

In Table 5, $\hat{S}_{PM}(t)$ is the parametric survival probability estimated for male AML patients using GEV distribution. $\hat{S}_{KMM}(t)$ is the non-parametric survival probability estimated for male AML patients using KM estimate. $\hat{S}_{PF}(t)$ is the parametric survival probability estimated for Female AML patients using GEV distribution. $\hat{S}_{KMF}(t)$ is the non-parametric survival probability estimated for Female AML patients using KM estimate.

6. Results and Discussions

Given the risk posed by AML cancer in the past few years, it is imperative to

Table 4. Table of comparison of the median survival time of male and female AML patients.

Methods	Male	Female
Descriptive	4	5
Parametric	6.3	7.8
Non-Parametric	4	6

Table 5. Table of comparison of estimated survival probabilities of male and female AML Patients computed using parametric and non-parametric procedures.

t_i	$\hat{S}_{PM}(t)$	$\hat{S}_{KMM}(t)$	$\hat{S}_{PF}(t)$	$\hat{S}_{KMF}(t)$
0	0.82	0.68	0.83	0.72
1	0.77	0.62	0.78	0.65
2	0.71	0.56	0.74	0.61
3	0.65	0.53	0.7	0.58
4	0.61	0.49	0.65	0.55
5	0.56	0.46	0.6	0.52
6	0.51	0.45	0.56	0.5
7	0.47	0.42	0.53	0.47
8	0.44	0.4	0.5	0.45
9	0.41	0.37	0.46	0.42
10	0.38	0.36	0.43	0.41
11	0.35	0.35	0.41	0.4

investigate the prognosis and enhance the therapeutic/treatment strategy of AML. The primary treatment for most types of AML is chemotherapy, sometimes, along with a targeted therapy drug. A stem cell transplant might follow this. Surgery and radiation therapy do not fall under crucial treatments for AML, but they might be used in exceptional circumstances. Also, the treatment approach for children with AML can be slightly different from that used for adults. Different research approaches and methodologies have been developed to treat AML patients to boost their survival time. In our present study,

- We have shown that there is a significant difference between the survival time of male and female patients diagnosed with AML.
- We identified a well-defined probability distribution that characterizes the survival time of a total of 2015 patients (1103 male and 912 female) diagnosed with AML and used it to estimate the survival function.
- We calculated the survival probabilities utilizing the frequently used non-parametric Kaplan-Meier (KM) cancer survivorship analysis method.
- We compared the median survival time of male and female AML patients using descriptive, parametric, and non-parametric methods.
- We compared the estimated survival probabilities of male and female patients diagnosed with AML by parametric method (driven by GEV probability distribution) and non-parametric method (driven by KM estimate) beyond a given survival time.

At the first stage of our analysis, we tried to investigate if there is any statistically significant difference between the survival time of male and female AML patients using the Log-Rank test. We found that there exists a significant difference between the survival time of both males and females diagnosed with AML. So, we start performing our data analysis using the separate analysis of the males and females AML patients. We found that a GEV distribution best characterizes the survival time's probabilistic behavior for both male and female AML pa-

tients, separately. We believe that finding the most accurate probability distribution that represents the probabilistic behavior of the survival time for a given cancer patient can lead to estimating the survival probability with much more accuracy and efficiency. The fact that we determined a unique probability distribution for our study of the survival time of patients diagnosed with AML contradicts the proposition of the assumption of exponential distribution (Feigl and Zelen ([1965] p. 835) and other authors) or using the non-parametric Kaplan-Meier for the majority of cancer survivorship studies. We found that the GEV distribution most often estimates higher survival probabilities compared to the KM survival function, given by Table 5. We know that KM estimates are very frequently and commonly used tool to analyze the cancer survivorship data, but they are not the best estimates. Statistically, the parametric technique is considered to be more robust and efficient than the non-parametric counterpart.

Therefore, our finding of the parametric GEV probability distribution gives better results in estimating the survival probability of the patients diagnosed with AML than the Kaplan-Meier. The KM technique is most frequently used to compare the difference between the estimated survival probabilities of the survival time of two or more entities or categories, typically based on the log-rank test. However, by obtaining the best parametric probability distribution that characterizes the survival time, we can find the survival function and estimate the survival rate and compare the results of two or more entities with a high degree of accuracy. One of the most useful results that we have obtained from our data analysis is that the survival probabilities for female AML patients are significantly higher than the survival probabilities for male AML patients by both parametric and non-parametric methods, which is evident from Table 5 and also Figure 11.

7. Conclusions

We have determined the survival probability of patients diagnosed with Acute Myeloid Leukemia (AML) using two different statistical methods: the parametric Generalized Extreme Value (GEV) distribution and the non-parametric Kaplan-Meier (KM) estimation. We found the parametric method to give often higher estimates of the survival probabilities than the non-parametric KM method. Despite the fact that there are instances when some of the non-parametric survival probability estimates are the same or higher; all-important arguments favor the parametric approach. The parametric survival analysis's difficulty is the fundamental inherent assumption that the population's survival time under study follows a specific probability distribution.

But if we can overcome such restriction, we can obtain a more robust and efficient result from the parametric analysis, which has greater statistical power. We can also evaluate the hazard function, which determines the rate at which patients die with AML, after finding the right parametric distribution. Depending on the two different methods utilized for estimating the probability of survival of

patients diagnosed with AML, we convey the following important recommendations.

- Given the information regarding male and female cancer patients' survival time, it is customary to investigate first if there exists any statistically significant difference between male and female patients' survival time. If the difference is significant, we must perform a separate analysis for each of the two groups. In the present study, we found that there is a significant difference between the survival time of male and female patients diagnosed with AML.
- If the only information provided about the patient is the survival time, then estimating the survival probability using the parametric technique will yield more accurate, robust, and efficient results than the commonly used non-parametric Kaplan-Meier survival estimate.
- However, if no unique or well-defined parametric probability distribution can be estimated, we still propose using the Kaplan-Meier (KM) technique to estimate the survival probabilities.

Although the use of non-parametric Kaplan-Meier survival analysis may, in certain circumstances, result in a similar or higher probability estimate of the survival rate (such as in our case), the parametric analysis remains more powerful, robust, and efficient. Hence, the parametric analysis must be considered the first stage of data analysis of any given cancer survivorship data. This study provides a more effective and plausible method for estimating the survival probability and analysis of cancer survivorship data to further enhance the therapeutic/treatment process of AML cancer.

Acknowledgements

The authors are thankful to the National Cancer Institute (NIH) for making the Surveillance, Epidemiology and End Results (SEER) database available publicly.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

Funding

Not Applicable.

Availability of Data and Materials

The data for our study has been extracted from SEER cancer registry (https://seer.cancer.gov/) which is available publicly for cancer research and innovation purpose.

References

[1] De Kouchkovsky, I. and Abdul-Hay, M. (2016) Acute Myeloid Leukemia: A Comprehensive Review and 2016 Update. *Blood Cancer Journal*, **6**, e441.

https://doi.org/10.1038/bcj.2016.50

[2] Cancer Research UK, 05-29-2020. https://www.cancerresearchuk.org/about-cancer/acute-myeloid-leukaemia-aml/about-acute-myeloid-leukaemia

[3] Löwenberg, B., Downing, J.R. and Burnett, A. (1999) Acute Myeloid Leukemia. *The New England Journal of Medicine*, **341**, 1051-1062. https://doi.org/10.1056/NEJM199909303411407

[4] Feigl, P. and Zelen, M. (1965) Estimation of Exponential Survival Possibilities with Concomitant Information. *Biometrics*, **21**, 826-838. https://doi.org/10.2307/2528247

[5] Xu, Y. and Tsokos, C.P. (2012) Probabilistic Survival Analysis Methods Using Simulation and Cancer Data. *Problems of Nonlinear Analysis in Engineering Systems*, *English*/*Russian*, **18**, 47-59.

[6] Xu, Y., Keper, J. and Tsokos, C.P. (2011) Identify Attributable Variables and Interactions in Breast Cancer. *Journal of Applied Sciences*, **11**, 1033-1038. https://doi.org/10.3923/jas.2011.1033.1038

[7] O'brien, P.C. (1988) Comparing Two Samples: Extensions of the t, Rank-Sum, and Log-Rank Tests. *Journal of the American Statistical Association*, **83**, 52-61. https://doi.org/10.1080/01621459.1988.10478564

[8] Kleinbaum, D.G. and Klein, M. (2012) Kaplan-Meier Survival Curves and the Log-Rank Test. In: *Survival Analysis*. *Statistics for Biology and Health*, Springer, New York, 55-96. https://doi.org/10.1007/978-1-4419-6646-9_2

[9] Yamamoto, J.F. and Goodman, M.T. (2008) Patterns of Leukemia Incidence in the United States by Subtype and Demographic Characteristics, 1997-2002. *Cancer Causes Control*, **19**, 379-390. https://doi.org/10.1007/s10552-007-9097-2

[10] Massey Jr., F.J. (1951) The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, **46**, 68-78. https://doi.org/10.1080/01621459.1951.10500769

[11] Anderson, T.W. and Darling, D.A. (1954) A Test of Goodness-of-Fit. *Journal of the American Statistical Association*, **49**, 765-769. https://doi.org/10.1080/01621459.1954.10501232

[12] Haan, L. and Ferreira, A. (2006) Extreme Value Theory: An Introduction.

[13] Jenkinson, A.F. (1955) The Frequency Distribution of the Annual Maximum (or Minimum) Values of Meteorological Elements. *Quarterly Journal of the Royal Meteorological Society*, **81**, 158-171. https://doi.org/10.1002/qj.49708134804

[14] Fisher, R.A. and Tippett, L.H.C. (1928) Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, **24**, 180. https://doi.org/10.1017/S0305004100015681

[15] Hosking, J.R.M., Wallis, J.R. and Wood, E.F. (1985) Estimation of the Generalized Extreme-Value Distribution by the Method of Probability-Weighted Moments. *Technometrics*, **27**, 251-261. https://doi.org/10.1080/00401706.1985.10488049

[16] Park, H.W. and Sohn, H. (2006) Parameter Estimation of the Generalized Extreme Value Distribution for Structural Health Monitoring. *Probabilistic Engineering Mechanics*, **21**, 366-376. https://doi.org/10.1016/j.probengmech.2005.11.009

[17] Greenwood, J.A., Landwehr, J.M., Matalas, N.C. and Wallis, J.R. (1979) Probability Weighted Moments: Definition and Relation to Parameters of Several Distributions Express Able in Inverse Form. *Water Resources Research*, **15**, 1049-1054.

https://doi.org/10.1029/WR015i005p01049

[18] Kaplan, E.L. and Meier, P. (1958) Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, **53**, 457-481. https://doi.org/10.1080/01621459.1958.10501452

[19] Moore, D.F. (2016) Applied Survival Analysis Using R. Use R. Springer, Berlin. https://doi.org/10.1007/978-3-319-31245-3

[20] Stel, V.S., Dekker, F.W., Tripepi, G., Zoccali, C. and Jager, K.J. (2011) Survival Analysis. I: The Kaplan-Meier Method. *Nephron Clinical Practice*, **119**, c83-c88. https://doi.org/10.1159/000324758