Estimating parametric survival distributions for stroke outcomes using statistical and graphical diagnostics

Vijayan S[1], **Kavitha S[2]***

[1]*Research Scholar, Department of Statistics, Periyar University, Salem-11, Tamil Nadu, India.*
Vijaystatistician1210@gmail.com

[2*]*Assistant Professor, Department of Statistics, Periyar University, Salem -11, Tamil Nadu, India.*
pustatkavitha@gmail.com

***Abstract*: This paper proposes the performance of Weibull, Gamma, Log-logistic, and Lognormal distributions for modeling the survival times of stroke patients using real-world data. Model performance was evaluated using log-likelihood, AIC, BIC, and graphical diagnostics, including PP plots, QQ plots, and fitted density and cumulative distribution curves. The comparative analysis identifies the distribution that best captures the observed survival pattern, providing a robust foundation for risk assessment and survival prediction.**

***Index Terms:* Parameter model, log-likelihood, Akaike Information Criterion, Bayesian Information Criterion, Survival analysis, real-life data.**

I. INTRODUCTION

The importance of distribution selection was emphasized in these studies, which aimed to compare the survival estimates for 2019 and 2020 [3]. These studies analyzed colorectal cancer data using a combination of Cox semi-parametric models and Weibull parametric models. This illustrates the use of parametric approaches when the risk structures are well-defined. It is also used in predicting limited mean survival time and for health technology assessment [4]. Survival analysis establishes itself as a crucial statistical framework in biomedical research, clinical trials, and reliability engineering, providing methodological tools for modeling event outcomes over time and investigating factors affecting survival patterns. Choosing an appropriate parametric distribution is central to obtaining accurate estimates of survival probabilities and hazard functions. Early contributions utilized classical distributions such as Weibull, exponential, and log-normal, demonstrating their relevance in medical applications and contrasting them with non-parametric techniques, including the Kaplan-Meier estimator, and semi-parametric approaches such as the Cox regression model [1,2]. Advances in 2022 further expanded the flexibility of the parametric survival model [5], providing a comprehensive review of traditional and extended distributions, including the Weibull, Gamma, log-normal, log-logistic, and generalized gamma families, while subsequent studies introduced modified Weibull extensions and multi-time scale parameter models to improve the fit for diverse and censored datasets [6,7]. This presents modern computational techniques and recent advancements (2023–2024), particularly in Bayesian parametric modeling and its applications in high-dimensional data. Paolucci et al. [8] utilize Bayesian Weibull models for medical survival prediction, while investigations on breast cancer datasets from Gunma University Hospital [9] and spatially dependent survival data [10] demonstrate the potential of extended hazard frameworks. Distribution fitting studies, such as those involving Weibull, log-logistic, and Gompertz distributions, compare classical and flexible families on real-world datasets [11–14]. By 2025, research will have expanded through the introduction of advanced treatment models and flexible parametric distributions. The accelerated Weibull, generalized gamma, and hybrid log-logistic-Weibull models are used to understand complex risk structures, account for long-term survivors, and handle heterogeneous patient populations [15–17]. The current study explores generalized gamma, accelerated Weibull, and log-logistic distributions for modern medical datasets, as well as advanced Bayesian and mixture

parameter models suitable for high-dimensional survival data [18–20]. The current study uses four possible parametric distributions, Weibull, Gamma, Log-logistic, and Lognormal, to predict the survival of stroke patients in an emerging context. Model fit is assessed through statistical criteria such as AIC, BIC, and log-likelihood, as well as graphical methods including PP and QQ plots and comparisons of PDF, CDF, and ECDF. The objective is to identify the most appropriate parametric model for accurate survival prediction and risk assessment in stroke patients.

II. LITERATURE OF REVIEW

Applications of parametric survival analysis in medical datasets focus on improving model flexibility for baseline data. The use of flexible and generalized gamma models [1] shows that extended distributions can better describe breast cancer survival times compared to traditional models. Based on this, the generalized log-logistic proportional hazards model [2] contributed further flexibility by accommodating different hazard patterns, making it suitable for complex medical data.

Comparative assessments among Cox regression and parametric alternatives consistently point to the strengths of fully specified distributions. Studies on colorectal cancer demonstrate that Weibull-based models perform better than semi-parametric methods when smooth risk estimation is required [3].

Later simulation-based work investigated how parametric models perform under extrapolation, especially for estimating the median survival time defined in long-term predictions [4].

The comprehensive methodological review of survival and reliability parameter distributions further clarified their historical development and practical benefits in research [5]. Additional contributions were found in advancing Bayesian approaches, such as the modified Weibull extension distribution, for modeling censored clinical data, emphasizing improved inference under uncertainty [6]. Flexible parametric models also gained traction for analyzing data with multiple time scales, demonstrating their value in multidimensional clinical time series [7].

Recent advancements have introduced Bayesian parametric frameworks for medical prediction tasks, and this plays a significant role in providing robust methods for incorporating prior information [8]. Its real-world applications in breast cancer confirm the utility of parametric survival methods in identifying patient-specific risk patterns [9]. Spatial dependent survival models extend these concepts by addressing geographical variation in health outcomes using excess risk approaches [10].

III. METHODS

This structured method allows for a rigorous comparison of applicant parametric distributions and ensures robust survival prediction for stroke patients. This retrospective study analyzed survival data from stroke patients. The primary outcome was time-to-event (T), measured in months from diagnosis to stroke occurrence or censoring. The censoring indicator (δ) was coded as: Probability Density Function (PDF) represents the instantaneous risk of the event at time t, Weibull, gamma, lognormal, loglogistic distribution

$$f(t;\lambda,k) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} \exp\left[-\left(\frac{t}{\lambda}\right)^{k}\right], t > 0 \quad (1)$$

$$f(t;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} t^{\alpha-1} \exp\left(-\frac{t}{\beta}\right), t > 0 \quad (2)$$

$$f(t;\mu,\sigma) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln t-\mu)^2}{2\sigma^2}\right], t > 0 \quad (3)$$

$$f(t;\kappa,\lambda) = \frac{\kappa}{\lambda}\left(\frac{t}{\lambda}\right)^{\kappa-1} \left[1+\left(\frac{t}{\lambda}\right)^{\kappa}\right]^{-2}, t > 0 \quad (4)$$

The Cumulative Distribution Function (CDF) gives the probability that the event occurs on or before time t, Weibull, gamma, lognormal, loglogistic distribution

$$F(t;\theta) = P(T \le t) = \int_0^t f(u;\theta)\, du \quad (5)$$

$$F(t) = 1 - \exp\left[-(t/\lambda)^k\right] \quad (6)$$

$$F(t) = \gamma(\alpha, t/\beta)/\Gamma(\alpha) \quad (7)$$

$$F(t) = \Phi((\ln\ t - \mu)/\sigma) \quad (8)$$

$$F(t) = (t/\lambda)^{\kappa}/[1 + (t/\lambda)^{\kappa}] \quad (9)$$

Survival function gives the probability of surviving beyond time t

$$S(t;\theta) = 1 - F(t;\theta) \quad (10)$$

Likelihood Function for Censored Data For $n$patients with observed times $t_i$and event indicators δ_i

$$L(\theta) = \prod_{i=1}^{n}[f(\ t_i;\theta)^{\delta_i} S(t_i;\theta)^{1-\delta_i}] \quad (11)$$

The log-likelihood is used for maximum likelihood estimation of parameters, estimates $\hat{\theta}$are obtained by maximizing $\ell(\theta)$.

$$\ell(\theta) = \sum_{i=1}^{n}[\ \delta_i \ln\ f(t_i;\theta) + (1-\delta_i)\ln\ S(t_i;\theta)] \quad (12)$$

Empirical and Theoretical Quantiles for Diagnostics is CDF (ECDF)

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(T_i \le t) \quad (13)$$

PP-plot: Compares $\hat{F}_n(t_i)$with $F(t_i;\hat{\theta})$

QQ-plot: Compares observed quantiles $\boldsymbol{t_{(i)}}$with theoretical quantiles

$$Q_i^{\text{theoretical}} = F^{-1}(\tfrac{i-0.5}{n};\hat{\theta}), Q_i^{\text{observed}} = t_{(i)} \quad (14)$$

Model Compare distributions, the following criteria were computed Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC)

$$\text{AIC} = 2k - 2\ln\ L(\hat{\theta}) \quad (15)$$

$$\text{BIC} = k\ln\ (n) - 2\ln\ L(\hat{\theta}) \quad (16)$$

$k$is the number of estimated parameters and $n$is the sample size. Lower AIC and BIC indicate better model fit.

III. ANALYSIS

This research used secondary data collected from the Kaggle website, specifically a 10-year dataset of medical records from stroke patients. R software is used to carry out the entire analysis, using the following packages: fitdistrplus for parametric distribution fitting and MLE estimation. ggplot2 for visualization of PDFs, CDFs, and ECDFs. Table 1 shows that values in the Comparative analysis of survival models demonstrate that the Weibull distribution provides the most accurate representation of survival times in this dataset.

With very low AIC and BIC values and a shape parameter that indicates increasing risk over time, the Weibull model closely matches the observed survival behavior. In contrast, the gamma and lognormal models show moderate fits, while the log-logistic model performs poorly. Overall, the Weibull distribution emerges as the most appropriate choice for modeling survival outcomes. Suggests the Weibull model is more effective a capturing the underlying distribution in the real-life dataset.

Figure 1 provides graphical diagnostics to assess how well the Weibull, Gamma, Log-logistic, and Lognormal distributions represent observed survival times. Of these, the Weibull model shows the closest agreement with the empirical data. Its density curve fits the histogram well, and both the QQ and PP plots have points very close to the reference line, indicating a strong fit in terms of magnitudes and probabilities. The Gamma and Lognormal models perform moderately well, as they follow the empirical distribution but show clear deviations, especially in the upper tail.

In contrast, the Log-logistic model deviates greatly, especially in the left tail of the PDF and in both the QQ and CDF comparisons, indicating that its hazard structure is not consistent with the data. The Gamma and log-normal models are acceptable and provide a less accurate fit, and the log-logistic model performs poorly. In comparison, the Weibull model shows superior performance.

Table 1. Comparison of Parametric Survival Models for Stroke Patients

Model	Shape	Scale	LL	AIC	BIC
Weibull	1.864	7.235	-2154.61	4315.21	4388.88
Gamma	2.455	0.381	-2183.74	4375.48	4540.34
Log-Logistic	2.353	5.638	-2260.47	4526.94	4516.20
Lognormal	1.645	0.738	-2249.40	4502.80	4328.62

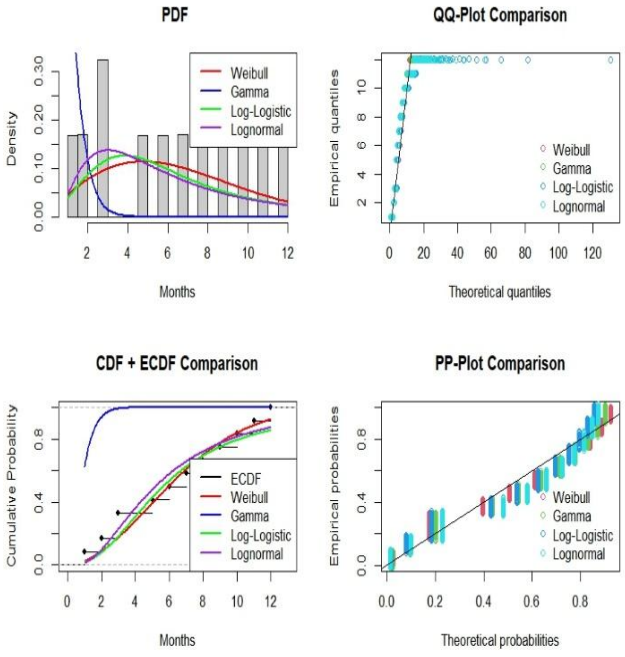

Figure 1. *Parameter estimates for the fitted models using stroke patients' survival time (in months) and their average glucose levels*

V. CONCLUSION

In the survival analysis of stroke patients, among the possible distributions used – Weibull, gamma, log-logistic and lognormal – the Weibull distribution performed best, exhibiting the lowest AIC and BIC values and the highest log-likelihood. The Weibull model can be used to estimate that the risk of stroke events increases over time. Future studies could incorporate relevant variables such as age, gender, hypertension, BMI, and glucose levels of stroke patients into the analysis. Simultaneously, exploring time-dependent covariates would help in modeling risk factors that change over time.

REFERENCES

[1] Yavari, P., Abadi, A., Amanpour, F., & Bajdik, C. (2012). Applying conventional and saturated generalized gamma distributions in parametric survival analysis of breast cancer. Asian Pacific Journal of Cancer Prevention, 13(5), 1829-1831. https://doi.org/10.7314/apjcp.2012.13.5.1829

[2] Khan, S. A., & Khosa, S. K. (2016). Generalized log-logistic proportional hazard model with applications in survival analysis. Journal of Statistical Distributions and Applications, 3(1), 16. https://doi.org/10.1186/s40488-016-0054-z

[3] Umar, U., & Muhammad, M. H. Comparative Analysis of the Cox Semi-parametric and Weibull Parametric Models on Colorectal Cancer Data. https://doi.org/10.11648/j.ijdsa.20200601.15

[4] Gallacher, D., Kimani, P., & Stallard, N. (2021). Extrapolating parametric survival models in health technology assessment: a simulation study. Medical Decision Making, 41(1), 37-50. https://doi.org/10.1177/0272989X20973201

[5] Taketomi, N., Yamamoto, K., Chesneau, C., & Emura, T. (2022). Parametric distributions for survival and reliability analyses, a review and historical sketch. Mathematics, 10(20), 3907. https://doi.org/10.3390/math10203907

[6] Feroze, N., Tahir, U., & Noor-ul-Amin, M., et.al., (2022). Applicability of modified weibull extension distribution in modeling censored medical datasets: a Bayesian perspective. Scientific Reports, 12(1), 17157. https://doi.org/10.1038/s41598-022-21326-w

[7] Batyrbekova, N., Bower, H., Dickman, P. W., Ravn Landtblom, A., Hultcrantz, M., Szulkin, R., ... & Andersson, T. M. (2022). Modelling multiple time-scales with flexible parametric survival models. BMC medical research methodology, 22(1), 290. https://doi.org/10.1186/s12874-022-01773-9

[8] Paolucci, I., Lin, Y. M., Albuquerque Marques Silva, J., Brock, K. K., & Odisio, B. C. (2023). Bayesian parametric models for survival prediction in medical applications. BMC Medical Research Methodology, 23(1), 250. https://doi.org/10.1186/s12874-023-02059-4

[9] Tasfa Marine, B., & Mengistie, D. T. (2023). Application of parametric survival analysis to women patients with breast cancer at Jimma University Medical Center. BMC cancer, 23(1), 1223. https://doi.org/10.1186/s12885-023-11685-6

[10] A. V. R. Amaral, F. J. Rubio, M. Quaresma, et al., (2023). Extended excess hazard models for spatially dependent survival data. arXiv e-prints, arXiv-2302. https://ui.adsabs.harvard.edu/link_gateway/2023arXiv230209392R/doi:10.48550/arXiv.2302.09392

[11] Bello FM, Musa FN, Hassan IA. (2023) Distribution of computerized datasets to fit Weibull, Log Logistic and Gompertz survival models. International Journal of Scientific Advances2023;4(3). https://doi.org/10.51542/ijscia.v

[12] Li, X., Marcus, D., Russell, J., Aboagye, E. O., et.al (2024). Weibull parametric model for survival analysis in women with endometrial cancer using clinical and T2-weighted MRI radiomic features. BMC Medical Research Methodology, 24(1), 107. https://doi.org/10.1186/s12874-024-02234-1

[13] Hanada, K., & Kojima, M. (2024). Bayesian Parametric Methods for Deriving Distribution of Restricted Mean Survival Time. arXiv preprint arXiv:2406.06071.https://doi.org/10.48550/arXiv.2406.06071

[14] Musa, F. N, Usman, A. & Amoto, A. (2024). Comparing the performance of Weibull, Log Logistic and Gompertz survival models on oncological data," Int. J. Sci. Adv., vol. 5, no. 3, 2024. https://doi.org/10.51542/ijscia.v5i3.7

[15] Borges, P., & Rodrigues, A. (2025). Estimating Quantiles and Cure Rate in Survival Data: A Parametric Regression Framework Using the Exponentiated Weibull Distribution. Journal of Statistical Theory and Applications, 1-32. https://doi.org/10.1007/s44199-025-00121-2

[16] Sadiq, I. A., Kajuru, J. Y., & Doguwa, S. I., et.al. (2025). Survival analysis in advanced lung cancer using Weibull survival regression model: estimation, interpretation, and clinical application. Journal of Statistical Sciences and Computational Intelligence,1(2),106-123. https://doi.org/10.64497/jssci.30

[17] Abd Elgawad, M. A., Usman, A.,& Doguwa, S. I. (2025). A hybrid Log-Logistic-Weibull Regression Model for survival analysis in leukemia patients and radiation data. Journal of Radiation Research and Applied Sciences, 18(3), 101836. https://doi.org/10.1016/j.jrras.2025.101836

[18] Alshawarbeh, E., et al. (2025). Gamma exponentiated generalized family of distributions: Properties and applications. Scientific Reports, 15, Article 23470. https://doi.org/10.1038/s41598-025-23470-5

[19] Kaindal, S., & Venkataramana, B. (2025). A comparative analysis of parametric survival models and machine learning methods in breast cancer prognosis. Scientific Reports, 15, Article 15696. https://doi.org/10.1038/s41598-025-15696-0
.

[20] Chu, J., Wang, Y., Sun, N., et al. (2025). A parametric survival model with Bayesian structural equation based on multi-omics integration. BMC Bioinformatics. https://doi.org/10.1186/s12859-025-06338-3