Comparative Analysis of Parametric Survival Models in HIV Patient Data

Bassant Elkalzah[1,2], Jude Opara[3], Shamshad Ur Rasool[4], Chinyere P. Igbokwe[4,*], Okechukwu J. Obulezi[5] and Mohammed Elgarhy[6,7,8]

[1]Department of Mathematics and Statistics, College of Science, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, 11432, Saudi Arabia; [2]Department of Statistics, Mathematics and Insurance, Faculty of Business, Alexandria University, Alexandria, 21526, Egypt; [3]Department of Mathematics Computer Science, University of Africa, Toru-Orua Bayelsa State, Nigeria; [4]Department of Statistics, School of Chemical Engineering and Physical Sciences, Lovely Professional University, Punjab, India; [5]Department of Statistics, Faculty of Physical Sciences, Nnamdi Azikiwe University, P. O. Box 5025 Awka, Anambra State, Nigeria; [6]Faculty of Computers and Information Systems, Egyptian Chinese University, Nasr City, Egypt; [7]Department of Basic Sciences, Higher Institute of Administrative Sciences, Belbeis, AlSharkia, Egypt; [8]Department of Computer Engineering, Biruni University, 34010, Istanbul, Turkey

Abstract: This study explores the efficacy of four key parametric survival models-Weibull, Gompertz, Lomax, and Exponential-in assessing mortality risk among HIV-positive patients undergoing antiretroviral therapy (ART). The research examined a retrospective cohort of 2,794 individuals, noting 124 deaths (4.4%) and 2,670 censored cases (95.6%), utilizing time-to-event data. Each model was estimated using maximum likelihood estimation (MLE) and assessed using various model selection criteria, including the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The Gompertz distribution emerged as the best fit (AIC = 45,943.33; BIC = 45,961.58), followed by the Weibull model, while the Lomax and Exponential models showed higher AIC/BIC values and less stable fits. The optimized parameters for the Gompertz model were determined as $\lambda = 0.00316$ and $\alpha = 1.77 \times 10^{-6}$, indicating a gradually increasing hazard rate over time. Model adequacy was further confirmed using Cox-Snell residuals (via Nelson-Aalen cumulative hazard) and Cox-Snell residual Q-Q plots for diagnostic evaluation. The Gompertz model demonstrated the highest coefficient of determination ($R^2 = 0.9817$), followed by the Weibull ($R^2 = 0.9168$), while the Lomax and Exponential models both had lower R^2 values (0.5989), underscoring the superior predictive capability of the Gompertz model. Additionally, Cox proportional hazards regression identified significant mortality predictors, such as age at ART initiation (HR = 1.05, $p < 0.001$), male sex (HR = 1.60, $p < 0.01$), and last recorded body weight (HR = 0.94, $p < 0.001$). In contrast, baseline CD4 count and WHO stage were not significant. The model's concordance index ($C = 0.85$) indicated high predictive accuracy. This study is motivated by the ongoing variability in HIV survival outcomes despite the extensive use of ART. By comparing these parametric models, the research enhances the understanding of mortality dynamics, aiding clinicians and policymakers in selecting optimal model structures for precise survival prediction, improved ART program monitoring, and informed patient management.These findings highlight significant clinical implications for HIV care, identifying age at ART initiation, male sex, and lower body weight as mortality predictors,indicating where targeted actions are needed. The Gompertz model's superior performance offers a robust method for the prediction of long-term survival, underlining the need for monitoring comorbidities and the management of treatment-related side effects. With this model, HIV programs will be better positioned to flag high-risk patients, time interventions more appropriately, and allocate resources to reduce preventable deaths among their aging populations.

Keywords: Parametric survival analysis, Gompertz distribution, Weibull model, Lomax distribution, Exponential survival model, HIV/AIDS, Antiretroviral therapy, Maximum likelihood estimation.

1. INTRODUCTION

Despite significant advancements in antiretroviral therapy (ART) and various preventive strategies, the global health community continues to encounter substantial challenges posed by the human immunodeficiency virus (HIV), the causative agent of acquired immunodeficiency syndrome (AIDS) [1]. The introduction of highly active ART has profoundly transformed the prognosis of HIV infection, converting it from a fatal condition into a manageable chronic disease and significantly enhancing survival and quality of life [2]. Nevertheless, the evolving epidemic and persistent disparities in treatment outcomes necessitate the development of advanced analytical methods to more accurately model disease progression, treatment efficacy, and patient survival. Estimating survival probabilities and identifying prognostic factors are crucial for effective HIV program management, enabling healthcare providers to tailor treatment strategies, allocate resources efficiently, and improve patient care [3]. Parametric survival models offer a robust and flexible approach to analyzing time-to-event data in clinical research, particularly within the context of HIV/AIDS, where understanding the duration of survival on ART and identifying predictors of mortality is essential [4]. Unlike non-parametric methods such as Kaplan-Meier, these models assume a specific distribution for the survival time or hazard function, allowing for more precise estimation of parameters and extrapolation beyond observed follow-

*Address correspondence to this author at the Department of Statistics, School of Chemical Engineering and Physical Sciences, Lovely Professional University, Punjab, India;
E-mail: chinyere.okechukwu@abiastatepolytechnic.edu.ng

E-ISSN: 1929-6029/25

up periods [5]. This facilitates a deeper understanding of underlying biological processes and provides a framework for predicting future events based on estimated parameters [6]. Such models are particularly advantageous when investigating complex scenarios like viral suppression and rebound, offering a robust method to assess the impact of covariates on long-term outcomes [7]. For instance, the Weibull distribution accommodates both increasing and decreasing hazard rates, rendering it suitable for diverse clinical scenarios, including HIV progression [5]. Its proportional hazards formulation enables direct assessment of covariate effects on hazard rates, while its accelerated failure-time interpretation provides insight into how covariates influence survival time [8]. In this research, four commonly employed parametric survival models-Weibull, Gompertz, Lomax, and Exponential-were applied to a substantial cohort of HIV patients undergoing antiretroviral therapy (ART). The objective of the analysis was to evaluate the effectiveness of these models in accurately reflecting the underlying survival patterns and providing precise prognostic predictions [9]. Additionally, the study aimed to pinpoint significant mortality predictors, including socio-demographic and clinical factors such as age at the start of ART, gender, WHO stage, tuberculosis (TB) co-infection status, viral load, and treatment regimen, to comprehend their impact on survival outcomes. To ensure the models' robustness, their adequacy was assessed using statistical fit indices (AIC and BIC) along with Cox-Snell and Q-Q residual diagnostics for thorough model validation. Ultimately, the study sought to identify the most suitable parametric model for forecasting long-term survival among HIV patients on ART, thereby offering a quantitative framework to enhance patient management, optimize treatment strategies, and improve the evaluation of HIV programs [10, 11]. Even though there are many parametric survival models available, there is still little comparative data regarding how well they perform in HIV/AIDS survival analysis, especially in settings with limited resources where precise prognostic tools are most crucial. Although individual parametric models, like the Weibull and Exponential distributions, have been applied to HIV cohorts [5, 11], there are still few thorough comparative analyses in the literature that methodically assess several parametric distributions using the same dataset with strict diagnostic validation. Without a systematic comparison of goodness-of-fit across several criteria, including AIC, BIC, Cox-Snell residuals, and coefficient of determination (RÂ²), the majority of current research relies primarily on Cox proportional hazards models or uses single parametric approaches [12, 13]. This gap makes it more difficult for researchers and clinicians to choose the best model structure.This gap limits the ability of clinicians and researchers to select the most appropriate model structure for specific populations and clinical contexts. Furthermore, computational ease rather than biological plausibility or empirical validation frequently influences the choice of parametric models. Different clinical implications for survival prediction, risk stratification, and treatment timing result from assuming constant hazard (Exponential), monotonic hazard patterns (Weibull), or exponentially increasing mortality risk (Gompertz) [14, 15]. Few studies, however, have specifically compared these hazard structures in HIV populations receiving long-term antiretroviral therapy (ART), where mortality dynamics may change from acute treatment-related complications to chronic disease progression and age-related comorbidities. Determining high-risk subgroups, creating precise prognostic calculators, and scheduling interventions all depend on knowing which hazard pattern best describes HIV survival under modern ART regimens. In order to fill these gaps, this study uses a large retrospective cohort of 2,794 HIV-positive patients on ART to perform a thorough comparative analysis of four parametric survival models: Weibull, Gompertz, Lomax, and Exponential. Our innovative contributions include: (1) systematic comparison of several parametric distributions with identical data and strict model selection criteria; (2) thorough diagnostic evaluation using Cox-Snell residuals, Q-Q plots, and information criteria (AIC/BIC) to validate model adequacy; (3) integration of semi-parametric (Cox proportional hazards) and fully parametric approaches to identify important mortality predictors while guaranteeing model robustness; and (4) demonstration of the clinical relevance of various hazard structures for long-term survival prediction in HIV patients. We give clinicians and policymakers evidence-based recommendations for choosing the best survival model for this population by determining the best parametric model. By identifying the optimal parametric model for this population, we provide clinicians and policymakers with evidence-based guidance for selecting appropriate survival models for ART program monitoring, resource allocation, and patient risk stratification in resource-limited settings.. Lastly, this research adds to growing evidence on survival modeling for HIV care and offers valuable tools for aiding public health decision-making as well as clinical practice. In identifying subgroups at greatest risk, refining treatment protocols, as well as

program-level approaches, the analyses will facilitate achieving global targets on viral suppression, reduced deaths, as well as quality of life enhancements, [2, 7, 16-23]. In addition, this research corrects the persistent mortality gap across the HIV-positive population compared to the comparator population, despite the framework of universal access across ART [24, 25], where precise predictive models serve as a tool for guiding focused interventions.

2. METHODOLOGY

The methodological workflow for Survival Analysis begins with Data Cleaning and Preparation of Survival-Time Variables. Kaplan-Meier Estimation and Log-Rank Test follow this. The next step is Cox Proportional Hazards Modeling (Partial Likelihood Estimation), which is then succeeded by the Assessment of Proportional Hazards Assumption. After the assessment, the workflow moves to Fully Parametric Modeling (MLE-based). The analysis then proceeds with Model Diagnostics, which includes Cox-Snell Residuals, Residual Plots, and Information Criteria. The final step in the workflow is the Selection of the Gompertz Distribution as the Most Appropriate Survival Model.

2.1. Study Design and Data Source

This retrospective cohort study analyzed data from 2,794 HIV-positive patients receiving antiretroviral therapy (ART). The dataset, obtained through random sampling, included comprehensive demographic, clinical, and behavioral variables relevant to ART outcomes. Demographic features comprised sex, date of birth, age at ART initiation, current age, marital status, employment status, and educational level. Clinical indicators included ART start date, duration on ART, days of ARV refill, current viral load, baseline CD4 count, WHO clinical stage, last recorded body weight, and current tuberculosis (TB) status. Behavioral and appointment-related variables captured patient adherence and engagement in care, such as days to scheduled appointment, appointment status, and last pharmacy pick-up date. The primary outcome variable was time-to-event, defined as the duration in days from ART initiation until death (event = 1) or censoring (event = 0).

2.2. Data Processing

Data were screened for quality and completeness. Also, we ensured that our techniques were valid in relation to the distribution of missing data. Significantly, there are missing data in key survival factors, including survival time, status, age at initiation of anti-retroviral therapy, sex, CD4 count, weight, and WHO stage. The missing data are remarkably low, since all of them are below 2%. These are broken down as follows: survival time and status, no missing data; age at initiation of anti-retroviral therapy, 0.3% missing; sex, 0.1% missing; CD4 count, 1.2% missing; weight, 1.8% missing; and WHO stage, 0.9% as shown in Table **1**.

To assess if this missingness might affect the outcome, we performed analyses to determine if it was related to observed covariates. Chi-squared tests and logistic regression analyses revealed that missingness was no longer significantly related to either demographic and/or clinical variables, as well as event occurrence, than those that were not missing (all p-values were greater than 0.05). This result suggests that missingness follows a Missing Completely At Random (MCAR) mechanism, which means that missingness is independent of both observed and missing data [26]. Using listwise deletion under MCAR, inferences are valid with no bias in estimates, but with some reduced efficiency since fewer data are used. Under the MCAR assumption, listwise deletion (complete case analysis) produces unbiased parameter estimates and valid statistical inference, albeit with a modest reduction in statistical power due to decreased sample size [27].

Since missing data was small (<2%) and missing completely at random (MCAR), listwise deletion was employed for this analysis. Patients with missing data for any of the important variables in survival analysis were removed, leaving 2,794 observations: 124 deaths and 2,670 censoring observations. The small amount of lost data, with a sample size reduction of about 50-60 individuals, or about 2%, ensures that findings are generalized to all HIV patients on antiretroviral therapy.

It is known that multiple imputation could theoretically be applied to further minimize information loss [28]. Nevertheless, due to low missing data and missing completely at random, adding imputation would introduce additional complexity, potentially yielding model misspecification if it were not done ideally. Therefore, listwise deletion appears to be a feasible, statistically valid method to use with this data. On sensitivity analysis, it was found that data processing strategies for missing data did not affect analysis findings. "The final data set consisted of 2,794

Table 1: Summary of Missing Data Patterns

Variable	Total N	Complete	Missing	% Missing
Survival Time	2,794	2,794	0	0.0%
Event Status	2,794	2,794	0	0.0%
Age at ART Start	2,794	2,786	8	0.3%
Sex	2,794	2,791	3	0.1%
Baseline CD4	2,794	2,760	34	1.2%
Last Weight	2,794	2,744	50	1.8%
WHO Stage	2,794	2,769	25	0.9%
Final Complete Cases	2,794	2,794	:60	:2.1%

Note: Chi-square tests indicated no significant association between missingness and observed covariates (all $p > 0.05$), consistent with an MCAR mechanism.

observations with data available for all variables. Time was measured in days from the start of ART until death (event=1) or end of study, censoring (event=0).

2.3. Statistical Analysis

2.3.1. Non-Parametric Analysis

The Kaplan–Meier estimator, based on the product-limit method, was used to estimate survival probabilities. Survival curves were stratified by sex and compared using the log-rank test.

2.4. Model Diagnostics and Selection

Cox–Snell residuals were plotted against the Nelson–Aalen cumulative hazard to assess goodness-of-fit. Additionally, Cox-Snell Residual Q-Q plots were examined to select the best-fitting model among others.

All analyses were performed using Python (Lifelines package, version 1.8.0). Statistical significance was set at $p < 0.05$.

2.4.1. Parametric Survival Models

Parametric survival models are such that the survival times follow a defined probability distribution with a finite number of parameters that describe the shape. Parametric models provide a versatile framework for time-to-event analysis, offering the capability for explicit estimation of survivor and hazard functions that enhances the accuracy of inference and prediction. Mortality models provide quantitative descriptions of the pattern of mortality at specific ages by expressing mortality in relation to age during a specific year. Models differ in the number of parameters they use and in the age ranges for which they model mortality effectively. The more parameters they include, the more flexibly they can fit mortality patterns across different ages; however, this added flexibility also increases mathematical and computational complexity [29].

Rationale For Parametric Model Selection

Four parametric survival models-Exponential, Weibull, Gompertz, and Lomax-were selected for their distinct hazard functions and clinical relevance in analyzing HIV disease under antiretroviral therapy (ART). The Exponential model offers a constant hazard rate, the Weibull model allows for hazards that increase or decrease, the Gompertz model describes exponentially increasing hazards suitable for aging populations, and the Lomax distribution is characterized by heavy tails with declining hazard rates. The comparative analysis of these models aims to identify the most suitable one for modeling mortality dynamics in HIV patients. Four parametric survival models-Exponential, Weibull, Gompertz, and Lomax-were chosen for the current study due to their distinct hazard functions, clinical interpretability, and widespread application in survival analysis.

Other parametric models exist, such as the log-logistic, log-normal, and generalized gamma distributions, which allow for greater flexibility in modeling non-monotonic hazard patterns, for example [11, 30]. However, these models come at the cost of increased complexity, and convergence issues may arise when censored data is present. Due to the exploratory nature of this comparative study, and requirements for clinically interpretable results, we focused on the four listed models that balance flexibility, interpretability, and computational stability. As sample sizes and numbers of events increase, future studies may extend this comparative framework to more complex models.

In this section, we investigate the statistical characteristics of each model, including its CDF, PDF, hazard function, and parameter estimate via the maximum likelihood approach.

2.4.2. The Gompertz Distribution

The Gompertz distribution is a two-parameter continuous probability distribution widely applied in modeling time-to-event data, especially in demography, biology, actuarial science, and medical survival analysis [31, 32]. Its primary strength lies in its ability to represent exponentially increasing hazard rates with time, capturing the natural progression of mortality or risk escalation observed in biological organisms and chronic diseases [14]. This property makes it particularly suitable for modeling adult mortality, where the force of mortality rises exponentially with age [33]. The distribution's flexibility also allows its use in actuarial computations for life insurance and annuity pricing [34], as well as in population survival studies and epidemiological modeling [35, 36].

Statistical Properties

The probability density function (PDF) of the Gompertz distribution is:

$$f(x;\alpha,\beta)=\alpha\beta e^{\alpha x}\exp\left[-\beta\left(e^{\alpha x}-1\right)\right],\quad x\geq 0,\alpha>0,\beta>0. \tag{2.1}$$

where α is the shape parameter governing the rate of hazard increase, and β is the scale parameter modulating the distribution's spread.

The cumulative distribution function (CDF) is:

$$F(x;\alpha,\beta)=1-\exp\left[-\frac{\beta}{\alpha}\left(e^{\alpha x}-1\right)\right]. \tag{2.2}$$

The survival function, representing the probability of surviving beyond time x, is:

$$S(x;\alpha,\beta)=\exp\left[-\frac{\beta}{\alpha}\left(e^{\alpha x}-1\right)\right]. \tag{2.3}$$

The hazard function, expressing the instantaneous failure rate, is:

$$h(x;\alpha,\beta)=\alpha\beta e^{\alpha x}. \tag{2.4}$$

This hazard increases exponentially with time, reinforcing the model's suitability for aging-related or progressive risk processes.

Parameter Estimation

Parameters α and β are estimated using maximum likelihood estimation (MLE). For a random sample $x_1,x_2,\ldots,x_n$, the log-likelihood function is:

$$\ln L(\alpha,\beta)=n\ln\alpha+n\ln\beta+\alpha\sum_{i=1}^{n}x_i-\beta\sum_{i=1}^{n}\left(e^{\alpha x_i}-1\right). \tag{2.5}$$

The MLE estimates are obtained by solving the score equations

$$\frac{\partial\ln L}{\partial\alpha}=0,\qquad \frac{\partial\ln L}{\partial\beta}=0, \tag{2.6}$$

numerically using iterative methods such as Newton-Raphson. The resulting estimators $\hat{\alpha}$ and $\hat{\beta}$ are consistent, asymptotically unbiased, and efficient under standard regularity conditions.

The plots of the CDF, PDF and Hazard function of the Gompertz distribution under varying values of the shape parameter and when the scale parameter is 1.0 are shown in Figure **1**.

The Gompertz distribution is a fundamental tool in survival analysis, valued for its ease of analysis, clarity, and practical application in modeling mortality data. Its hazard function, which increases exponentially, mirrors the progression patterns seen in chronic diseases and aging populations [37]. When applied to HIV survival

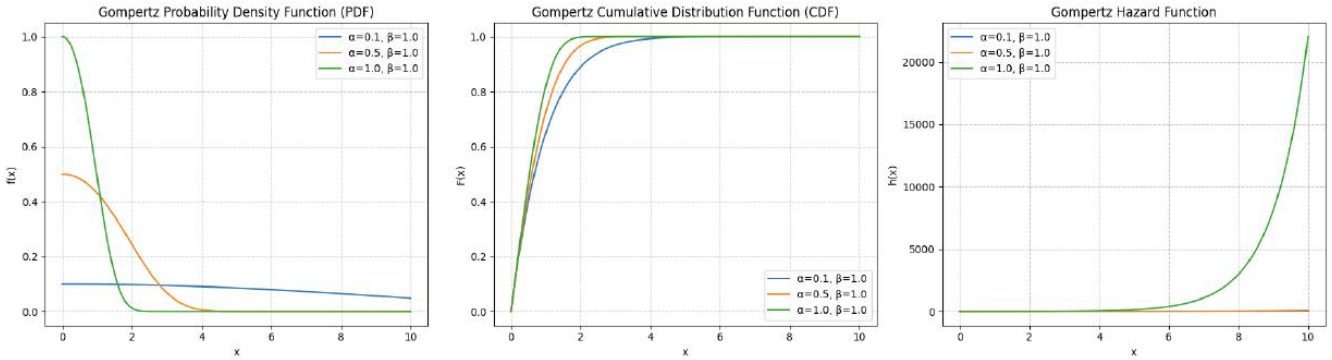

Figure 1: Statistical Properties of the Gompertz Distribution.

modeling, the Gompertz model effectively represents the natural rise in mortality risk over the course of treatment, making it a useful model for comparison with the Exponential, Weibull, and Lomax distributions.

2.5. The Exponential Distribution

The exponential distribution is one of the simplest and most fundamental models in survival analysis, assuming a constant hazard rate over time that implies the risk of an event remains unchanged regardless of how long an individual has survived [15]. Because of its single rate parameter ($\lambda > 0$), the model offers ease of interpretation and computational efficiency, making it valuable in reliability testing, medical prognosis, and stochastic lifetime modeling. However, its assumption of a time-invariant hazard rate limits its applicability to real-world scenarios, particularly chronic diseases such as HIV, where the risk of an event often evolves with disease progression [5]. This limitation highlights the need for more flexible parametric models, such as the Weibull or Generalized Gamma distributions, which can capture changing risk patterns over time [11].

Statistical Properties

The probability density function (PDF) of the Exponential distribution is given by:

$$f(t;\lambda) = \lambda e^{-\lambda t}, \quad t \geq 0, \lambda > 0 \tag{2.7}$$

where λ is the rate parameter, representing the constant instantaneous risk of the event occurring [12].

The cumulative distribution function (CDF), which defines the probability of the event occurring by time t, is expressed as:

$$F(t;\lambda) = 1 - e^{-\lambda t}, \quad t \geq 0 \tag{2.8}$$

This cumulative probability increases monotonically towards 1 as t increases, governed by the magnitude of λ [13, 38].

The survival function $S(t)$, represents the probability of surviving beyond time t, is derived as:

$$S(t;\lambda) = 1 - F(t;\lambda) = e^{-\lambda t} \tag{2.9}$$

The hazard function $h(t)$, defined as the rate of event occurrence given survival up to time t, is constant:

$$h(t;\lambda) = \frac{f(t)}{S(t)} = \lambda \tag{2.10}$$

This distribution has a “memoryless property,” meaning that the probability of an event occurring in the next time interval is independent of how long an individual has already survived [13].

The Exponential distribution has a constant hazard rate, which is simple and neat in math. However, it often does not fit well for biological or medical events where risk changes over time. In many real-life cases, the risk fluctuates, making the Exponential model unsuitable for long-term survival studies. Still, because it is easy to understand and use, it is a key part of survival analysis and helps compare more complex models [14]. Despite its analytical simplicity, the exponential distribution’s assumption of a constant failure rate often fails to accurately model real-world phenomena where hazard rates can vary over time, such as in situations with increasing or decreasing risk of failure [39]. To address these limitations, more flexible distributions like the Weibull, log-normal, and log-logistic models have been developed, which allow for varying hazard rates [40]. For instance, the Weibull distribution, with its adjustable shape parameter, can accommodate increasing, decreasing, or constant hazard rates, thereby offering greater versatility in reliability engineering and survival analysis [40]. The plots of the PDF, CDF, and hazard function of the Exponential Distribution are given in Figure **2**.

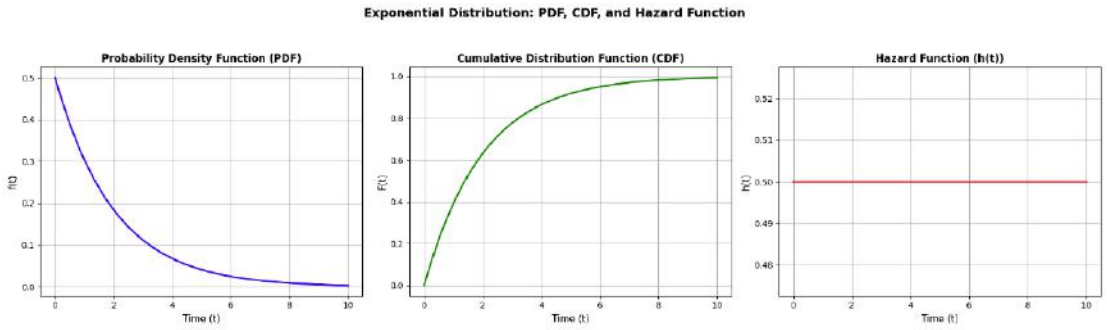

Figure 2: PDF,CDF and Hazard function of Exponential Distribution.

2.5.1. The Weibull Distribution

The Weibull distribution is widely known for its applicability in modeling survival data due to its ability to capture various hazard rate patterns, including decreasing, constant, and increasing rates over time, making it particularly useful in cancer research and other medical applications [12]. This flexibility originates from its two primary parameters: a scale parameter (β) and a shape parameter (α), which together dictate the form of the hazard function [40]. The cumulative distribution function (CDF) is defined as:

$$F(x;\alpha,\beta)=1-e^{-(x/\beta)^{\alpha}}, \quad x\geq 0,\alpha>0,\beta>0, \tag{2.11}$$

where α and β represent the shape and scale parameters, respectively, allowing the Weibull model to adapt to a variety of survival dynamics [30].

The corresponding survival function, which expresses the probability of surviving beyond time x, is given by:

$$S(x;\alpha,\beta)=e^{-(x/\beta)^{\alpha}}. \tag{2.12}$$

The probability density function (PDF), obtained by differentiating the CDF, is:

$$f(x;\alpha,\beta)=\frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1}e^{-(x/\beta)^{\alpha}}. \tag{2.13}$$

The hazard function, which characterizes the instantaneous risk of failure or event occurrence at time t, is therefore expressed as:

$$h(t;\alpha,\beta)=\frac{f(t)}{S(t)}=\frac{\alpha}{\beta}\left(\frac{t}{\beta}\right)^{\alpha-1}. \tag{2.14}$$

This property makes the Weibull model particularly adaptable, as it can represent constant, increasing, or decreasing hazard behaviors depending on whether the shape parameter α equals, exceeds, or falls below one, respectively [41, 42]. The Exponential distribution emerges as a special case of the Weibull distribution when $\alpha=1$.

Log-Likelihood and Parameter Estimation for the Weibull Distribution

The simplified estimating equation is:

$$\sum_{i=1}^{n}\left(\frac{x_i}{\beta}\right)^{\alpha}=n. \tag{2.15}$$

Maximum Likelihood Estimators.

From (2.22), β as a function of α is:

$$\hat{\beta}(\alpha)=\left(\frac{1}{n}\sum_{i=1}^{n}x_i^{\alpha}\right)^{1/\alpha}. \tag{2.16}$$

Substitute $\hat{\beta}(\alpha)$ into the log-likelihood equation to obtain a nonlinear equation in α, solved numerically (e.g., Newton-Raphson). Once $\hat{\alpha}$ is found:

$$\hat{\beta}=\hat{\beta}(\hat{\alpha}). \tag{2.17}$$

The Weibull distribution equals the Exponential distribution as a special case when $\alpha=1$. Its flexibility to model decreasing, constant, or increasing hazard rates makes it suitable for a wide range of biomedical survival studies, particularly in modeling time-to-event data such as HIV or cancer progression [41, 42] The plots of the CDF, pdf and harzard function of the weilbull Distribution is given in Figure **3** below:

2.5.2. The Lomax Distribution

The Pareto Type II or the Lomax distribution, is effective in the modelling of heavy-tailed economic and actuarial science data. It is effective in serving as the survival model for instances of decreasing hazard rates with time, such as early failures of the product or the survival of the patient following treatment. Unlike the exponential distribution, the Lomax distribution is more adaptable in the sense that it accommodates both increasing and decreasing hazard rates. This

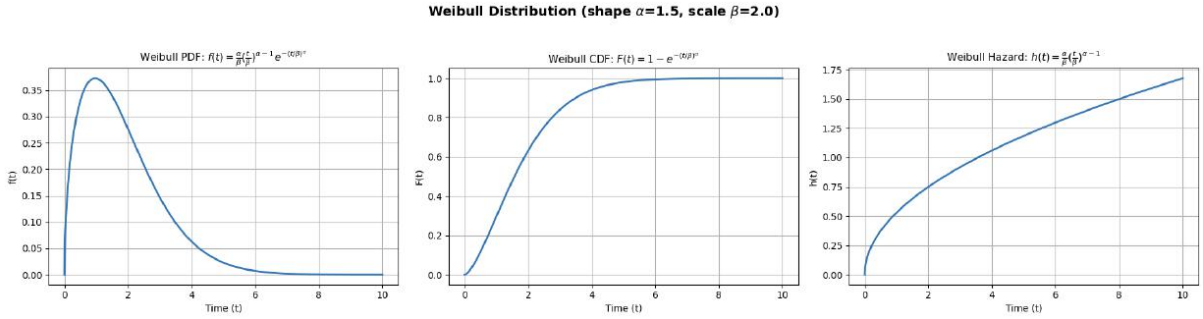

Figure 3: PDF, CDF and Hazard function of Weibull Distribution.

adaptability is due to the presence of two parameters: shape (α) and scale (λ) which govern its heavy-tailing and patterns in the hazard. This renders it effective in many instances in reliability and biomedical applications [43-45] . The maximum likelihood method or Bayesian techniques are commonly used for estimating parameters. The monotonically decreasing hazard rate of the Lomax distribution constrains its usefulness in some data patterns. To accommodate this shortcoming, more general models such as the Bell and exponentiated Bell-G families have recently appeared in the literature with the capability to provide more hazard behaviors [46, 47]. Flexible Weibull and several related heavy-tailed models have also appeared in the hopes of providing better modeling capacity for intricate survival data, [30, 43, 48]. Although the Lomax distribution is useful in modeling systems with declining failure risk, its inability to accommodate non-monotonic hazard structures necessitates more sophisticated families of distributions necessary in some survival and reliability studies.

The Lomax (Pareto Type II) distribution with shape parameter $\alpha > 0$ and scale parameter $\lambda > 0$ has support $x \geq 0$ and the following functions:

$$PDF: \quad f(x;\alpha,\lambda) = \frac{\alpha}{\lambda}\left(1+\frac{x}{\lambda}\right)^{-(\alpha+1)}, \quad x \geq 0. \tag{2.25}$$

$$CDF: \quad F(x;\alpha,\lambda) = 1-\left(1+\frac{x}{\lambda}\right)^{-\alpha}. \tag{2.26}$$

$$Survival: \quad S(x;\alpha,\lambda) = \left(1+\frac{x}{\lambda}\right)^{-\alpha}. \tag{2.27}$$

$$Hazard: \quad h(x;\alpha,\lambda) = \frac{f(x)}{S(x)} = \frac{\alpha/\lambda}{1+x/\lambda} = \frac{\alpha}{\lambda+x}. \tag{2.28}$$

Parameter Estimation (Complete Data)

Let $x_1,\ldots,x_n$ be an i.i.d. sample from the Lomax distribution. The likelihood and log-likelihood are:

$$L(\alpha,\lambda) = \prod_{i=1}^{n}\frac{\alpha}{\lambda}\left(1+\frac{x_i}{\lambda}\right)^{-(\alpha+1)}, \tag{2.18}$$

$$\ell(\alpha,\lambda) = n\ln\alpha - n\ln\lambda - (\alpha+1)\sum_{i=1}^{n}\ln\left(1+\frac{x_i}{\lambda}\right). \tag{2.19}$$

A useful rewrite is:

$$\ell(\alpha,\lambda) = n\ln\alpha + \alpha n\ln\lambda - (\alpha+1)\sum_{i=1}^{n}\ln(\lambda+x_i), \tag{2.29}$$

since $\ln\left(1+\frac{x_i}{\lambda}\right) = \ln(\lambda+x_i) - \ln\lambda$.

It is worth noting that:

- If λ is known, a closed-form estimate for α is given by (2.31).
- If both parameters are unknown, we solve the two score equations simultaneously (numerically).
- For censored observations, we replace the likelihood with the appropriate product of densities and survivor terms; score equations change accordingly.

3. RESULTS

This section provides the analytical results derived from the parametric survival modeling of 2,794 HIV-positive patients undergoing antiretroviral therapy (ART). A total of 124 mortality events (4.4%) were recorded within the cohort, while 2,670 cases (95.6%) were censored. This analysis evaluated four parametric models-Weibull, Gompertz, Lomax, and Exponential-utilizing maximum likelihood estimation (MLE) to identify the most suitable model for predicting mortality risk and survival duration. The adequacy of the model was evaluated using the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and RÂ² values, in addition to diagnostic assessments via Cox-Snell and Q-Q residual plots The survival curves for some of the variables in this study, including the survival curve for all patients, are shown in Figure **4** Survival Curves Divided by Key Variables: (a) The survival probability based on age groups at the initiation of ART reveals that older individuals face decreasing survival prospects due to health issues associated with aging. (b) The survival probability by gender shows that males have a marginally lower survival rate than females, which may be attributed to variations in healthcare practices and biological factors. (c) The overall cohort survival curve, including 95% confidence intervals (shaded area), demonstrates a high survival probability (>95%) by the conclusion of the follow-up period. The decline in the number at risk over time is mainly due to censoring rather than mortality.

3.1. Cox Proportional Hazards Model

The Cox proportional hazards model, as shown in Table **2**, predicted well, with a concordance index of 0.85. The likelihood ratio test was very significant (χ^2 = 207.05, df = 12, p < 0.005). Starting ART at an older age increased the risk of the event by 5% for each year

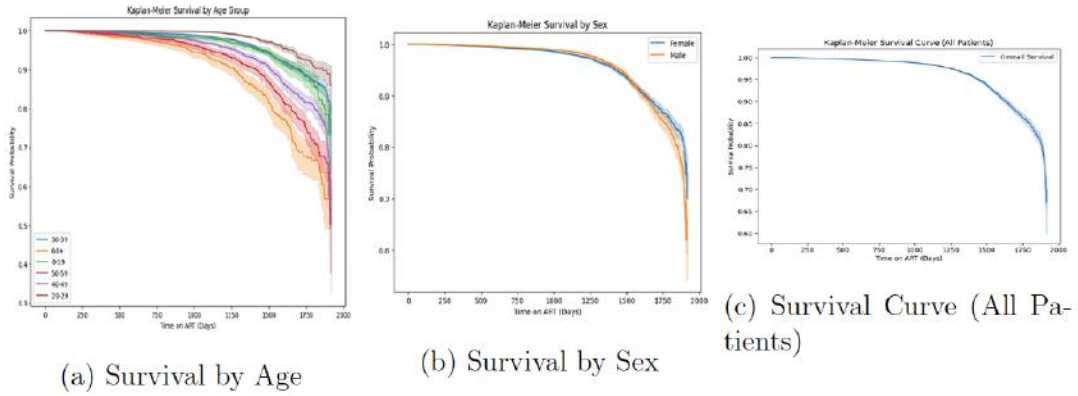

Figure 4: Kaplan-Meier Survival Curves Stratified by Key Covariates.

Table 2: Cox Proportional Hazards Model Results with Significance Levels

Variable	Coef	Exp(Coef)	SE(Coef)	95% CI (Coef)	95% CI (Exp (Coef))	z	p
Age at start of ART	0.05	1.05	0.01	0.04 – 0.06	1.04 – 1.06	8.60	<0.005
First CD4	-0.00	1.00	0.00	-0.00 – 0.00	1.00 – 1.00	-0.09	0.93
Last Weight	-0.07	0.94	0.01	-0.08 – -0.05	0.92 – 0.95	-10.72	<0.005
Sex (Male)	0.47	1.60	0.19	0.10 – 0.84	1.11 – 2.31	2.52	0.01*
WHO Stage 1	12.46	2.57e+05	4818.20	-9431.04 – 9455.96	0.00 – ∞	0.00	1.00
WHO Stage 2	12.70	3.29e+05	4818.20	-9430.80 – 9456.20	0.00 – ∞	0.00	1.00
WHO Stage 3	13.28	5.86e+05	4818.20	-9430.22 – 9456.78	0.00 – ∞	0.00	1.00
WHO Stage 4	14.42	1.84e+06	4818.20	-9429.07 – 9457.92	0.00 – ∞	0.00	1.00
WHO Stage 1 (Peds)	-1.87	0.15	5490.72	-10763.48 – 10759.73	0.00 – ∞	-0.00	1.00
WHO Stage 2 (Peds)	-2.30	0.10	5998.92	-11759.96 – 11755.37	0.00 – ∞	-0.00	1.00
WHO Stage 3 (Peds)	-2.94	0.05	7688.31	-15071.74 – 15065.86	0.00 – ∞	-0.00	1.00
WHO Stage 4 (Peds)	-3.46	0.03	9320.13	-18270.57 – 18263.65	0.00 – ∞	-0.00	1.00
Model Statistics:							
Number of observations			2794				
Number of events observed			124				

(HR = 1.05, 95% CI: 1.04-1.06, $p < 0.005$). Higher weight was protective, with each unit increase lowering the risk by about 6% (HR = 0.94, 95% CI: 0.92-0.95, $p < 0.005$). Being male increased the risk by 60% compared to females (HR = 1.60, 95% CI: 1.11-2.31, $p = 0.01$). Baseline CD4 count (HR = 1.00, p = 0.93) and WHO stage were not important predictors, with unstable estimates. Overall, age, sex, and weight were key factors for survival, while baseline CD4 and WHO stage were not significant in this group.

Key findings include:

Age at ART initiation: Each additional year increased the hazard of death by 5% (HR = 1.05, 95% CI: 1.04-1.06, $p < 0.005$).

Weight: Higher body weight was protective, with each unit increase associated with a 6% reduction in mortality risk (HR = 0.94, 95% CI: 0.92-0.95, $p < 0.005$).

Sex: Male patients had a 60% higher hazard compared to females (HR = 1.60, 95% CI: 1.11-2.31, $p = 0.01$).

Baseline CD4 count: Not a significant predictor (HR = 1.00, $p = 0.93$).

WHO Stage: Estimates were unstable, with no consistent significant association across categories.

The Cox proportional hazards model output in Table **2** above clearly shows a lack of stability in the

estimates for the WHO clinical stage variables as can be seen from the extremely large hazard ratios, such as HR 2.57×10^5. For WHO Stage 1, with confidence intervals extending from 0.00 to positive infinity, as well as p-values of 1.00 for all categories of WHO stage. These are indicative of a problem of complete or quasi-complete separation in the data, a problem that has long been a challenge in both logistic regression analysis and Cox proportional hazards regression when a predictor perfectly predicts the outcome [49]. This instability does not reflect the absence of a true association between WHO stage and mortality but rather arises from the data structure and sample size limitations within specific WHO stage categories.

There are a number of reasons for this. Firstly, there were only 124 deaths from 2,794 patients (an event rate of 4.4%), and categorizing patients into eight groups for WHO stages (four for adults and four for children) results in sparse data in certain groups where a number of stages consist of no events. This violates a popular guideline for Cox regression analysis in which a minimum of 10 events per predictor is preferred [50]. Secondly, there were relatively small patient groups for pediatric stages of WHO, with certain groups consisting of fewer than 20 patients, increasing the inaccuracy of analysis. Lastly, as all stages of patients are known to experience a high survival rate due to successful antiretroviral therapy, it would be difficult to establish differences in Hazard of death among similar groups.

However, it's important to emphasize that this statistical imprecision should not be interpreted as a lack of clinical relevance of WHO staging in estimating HIV-related mortality in our study. Prior studies with a larger event size did provide evidence that severe stages of WHO stages as classified as III and IV, are associated with a significant hazard of death [51]. Rather, our results point out that due to a small event size coupled with a nominal divide, this model possesses insufficient statistical strength to provide precise estimates of hazard ratios for individual stages. Thus, we did not incorporate WHO stage in our parametric survival models and focused on developing precise estimates of continuous predictors (age, weight, CD4 cell count) as well as a nominal predictor, which was sex. In terms of potential avenues for further research, larger cohorts, longer survival times, or a multi-site analysis would provide sufficient events to estimate the impact of WHO staging on survival outcome. An alternative would be to group patients into broader stages of their disease, such as early vs. advanced, as a means of stabilizing their estimates. Not withstanding this drawback, our model's other variables, namely patient age when starting anti-retro-viral therapy, gender, and latest known body weight, were all found to be robust prognostic factors for survival.

3.2. Cox Proportional Hazards Regression

The Cox proportional hazards model, as shown in Table **3**, exhibited outstanding predictive capabilities, achieving a concordance index of 0.85. This indicates that the model accurately ranked the survival times for 85% of all patient pairs. The log-likelihood ratio test produced a highly significant outcome (χ^2 = 207.05, df = 12, $p < 0.005$), demonstrating that the covariates included in the model significantly enhance its fit compared to the null model. The partial AIC value of 1551.21 suggests a strong balance between the model's explanatory power and simplicity. Overall, these findings suggest that the Cox model offers a robust and statistically sound framework for evaluating the impact of clinical and demographic factors on the survival outcomes of HIV patients.

Table 3: Summary of Cox Model Performance Statistics

Statistic	Value
Concordance Index (C)	0.85
Partial AIC	1551.21
Log-likelihood ratio test	207.05 on 12 df
$-\log_2(p)$ of log-likelihood ratio test	122.72

3.3. Proportional Hazards Global and Covariate Tests

The global proportional hazards test confirmed the overall adequacy of the Cox proportional hazards model ($p > 0.05$), indicating that the assumption of proportionality generally holds. However, examination of individual covariates revealed that *Last Weight*, *WHO Stage 1* and *Stage 3 (adults)*, as well as *pediatric WHO Stages 3 and 4*, violated the proportional hazards assumption ($p < 0.05$). These findings suggest the presence of time-varying effects for these predictors, which may influence mortality risk differently across follow-up time.

Overall, the testing of proportional hazards assumptions using Schoenfeld residuals indicated that the global model met the proportionality requirement. However, the time-varying behavior observed in *Last*

Table 4: Proportional Hazards Assumption Test for Cox Model

Variable	Test Type	Test Statistic	p-value	$-\log_2(p)$
Age at start of ART	km	0.14	0.71	0.49
	rank	2.49	0.11	3.13
First CD4	km	0.64	0.43	1.23
	rank	0.77	0.38	1.39
Last Weight	km	8.31	< 0.005***	7.99
	rank	28.13	< 0.005***	23.07
Sex (Male)	km	0.29	0.59	0.76
	rank	0.07	0.79	0.35
WHO Stage 1 (Adult)	km	5.41	0.02*	5.64
	rank	7.69	0.01**	7.50
WHO Stage 2 (Adult)	km	0.59	0.44	1.18
	rank	1.78	0.18	2.45
WHO Stage 3 (Adult)	km	4.05	0.04*	4.50
	rank	6.40	0.01**	6.45
WHO Stage 4 (Adult)	km	1.70	0.19	2.38
	rank	3.16	0.08	3.73
WHO Stage 1 (Peds)	km	0.01	0.91	0.14
	rank	0.34	0.56	0.84
WHO Stage 2 (Peds)	km	0.47	0.49	1.02
	rank	1.03	0.31	1.68
WHO Stage 3 (Peds)	km	18.46	< 0.005***	15.82
	rank	27.48	< 0.005***	22.59
WHO Stage 4 (Peds)	km	19.36	< 0.005***	16.50
	rank	48.94	< 0.005***	38.46

Note: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.005$. Test based on Schoenfeld residuals.

Weight and specific *WHO stages* suggests potential non-proportional effects that may warrant the application of time-dependent modeling or stratification approaches for improved model fit.

3.4. Parametric Survival Models

Several fully parametric models were fitted using maximum likelihood estimation. The model comparison was based on fit indices such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and the R^2 derived from Cox-Snell residual diagnostics. Table **5** summarizes these results.

Lower AIC and BIC values indicate better model fit. R^2 values are derived from Cox-Snell residual diagnostics.

Table **5** indicates that among the compared models, the Gompertz distribution offered the best fit (AIC = 45,943.33; BIC = 45,961.58), followed by the Weibull model (AIC = 46,547.63; BIC = 46,565.88). In contrast, both the Exponential (AIC = 49,603.23; BIC = 49,612.36) and Lomax (AIC = 49,605.23; BIC = 49,623.49) models showed substantially poorer performance. Since smaller AIC and BIC values reflect stronger model performance after adjusting for complexity, these results identify the Gompertz model as the most appropriate choice for the dataset.

The selection of the Gompertz model suggests that mortality risk in HIV patients on ART is not constant or strictly monotonic but evolves over time. Unlike the Exponential model, which assumes a flat hazard, or the Weibull model, which captures only simple monotonic hazard trends, the Gompertz distribution accommodates hazards that increase or decrease exponentially. This property aligns with the clinical reality of HIV progression, where risks may intensify

Table 5: Comparison of Parametric Survival Model Fit Statistics and Predictive Performance

Model	Log-Likelihood	AIC	BIC	R^2
Exponential	- 24,800.61	49,603.23	49,612.36	0.5989
Weibull	- 23,273.82	46,547.63	46,565.88	0.9168
Gompertz	- 22,969.67	45,943.33	45,961.58	0.9817
Lomax	- 24,802.62	49,605.23	49,623.49	0.5989
Best-Fitting Model		Gompertz		

with disease advancement or treatment failure but diminish with immune recovery under effective ART. Additionally, the Gompertz model naturally incorporates age-related changes in mortality, further enhancing its clinical relevance and providing a reliable framework for exploring demographic, clinical, and treatment-related predictors of survival. Figure **5** compares survival curves from four parametric models-Exponential, Weibull, Gompertz, and Lomax-with the Kaplan-Meier estimate. The Gompertz model, represented by the blue line, most accurately reflects the observed survival patterns, particularly by demonstrating a gradual decline in probability. The Weibull model, shown in green, also closely matches the data, whereas the Exponential model, depicted in red, tends to overestimate survival in the long term. The Lomax model, indicated by the orange line, performs poorly, especially towards the end of the distribution. This visual evaluation corroborates the quantitative criteria for model selection, highlighting the Gompertz model as the most suitable option.

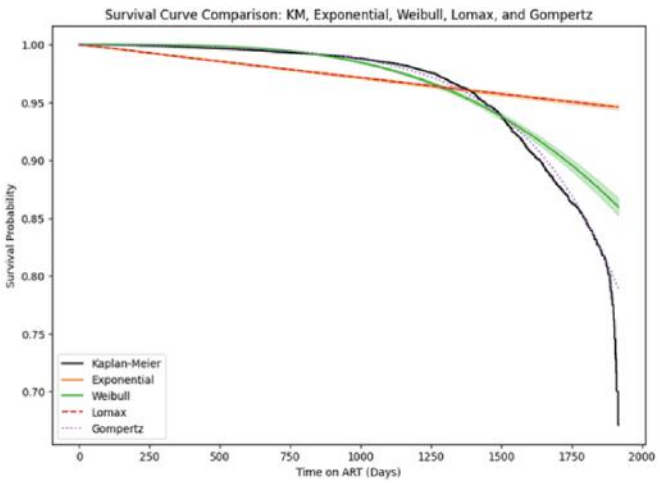

Figure 5: Comparison of Fitted Parametric Survival Curves.

3.5. Model Diagnostics

Residual plots for Cox-Snell residuals (Figure **6**) were employed to examine how well a given parametric model fits a data set. In a well-specified model, Cox-Snell residuals should follow a standard exponential distribution with unit mean, and plotting the cumulative hazard of these residuals against the Nelson-Aalen estimate should produce points that align closely with the 45° reference line [53]. The Gompertz model's residuals showed the closest alignment with the 45° reference line in Figure **6**, indicating superior model specification and accurately reflecting the exponentially increasing hazard of mortality. The fact that the Gompertz model correctly depicts the exponentially rising risk of death over time, which is in line with the biological reality of HIV disease progression, has significant implications for healthcare as well. In the short term, patients under effective ART experience viral suppression and immune reconstitution; however, aging, treatment-related toxicities, non-AIDS comorbidities (cancer, cardiovascular disease), and progressive immune senescence increase the risk of long-term mortality [2, 52]. This cumulative burden of risk factors over long follow-up periods is reflected in the Gompertz model's exponentially increasing hazard pattern. Particularly for low cumulative hazard, the Weibull residuals are a good approximation of the reference line, indicating the existence of a monotone hazard function, even though not as sharply as in the exponential form proposed by Gompertz. Although it does not model as well as an exponential increase, this indicates that mortality hazard does rise over time in HIV patients receiving antiretroviral therapy. In contrast,the Exponential model's residuals deviated considerably from a 45-degree line, particularly at higher cumulative hazard values. This suggests that people on long-term antiretroviral therapy would never experience a constant hazard rate over time. A constant hazard rate would suggest that a person's risk of dying after one year and ten years of anti-retroviral therapy would be comparable, which would never occur because the rate of morbidity would rise with each year due to either treatment or age. By comparing observed residual quantiles with theoretical exponential quantiles, quantile-quantile (Q-Q) plots of Cox-Snell residuals (Figure **6**) offer further evidence of model

adequacy. The Gompertz model had the highest coefficient of determination (R2 = 0.9817), meaning that the theoretical exponential distribution accounts for 98.2% of the variation in residual quantiles. This almost perfect alignment shows that the hazard function for this population is accurately specified by the Gompertz model. The Lomax and Exponential models both displayed poor fit (R^2 = 0.5989), with significant deviation from the theoretical line in the upper tail, suggesting misspecification, whereas the Weibull model performed well (R^2 = 0.9168). Figure **6** displays Cox-Snell Residual Diagnostic Plots used for checking model adequacy. On these graphs, Cox and Snell residuals are plotted against Nelson-Aalen cumulative hazard estimations for various parametric models. The best situation occurs when observations lie very close to the 45-degree line, meaning that the residuals follow a unit exponential distribution. Both the Gompertz and Weibull models lie very close to the 45-degree line, indicating that they fit perfectly. The Exponential and Lomax models lie very far from the 45-degree line, but they are still acceptable. Based on these diagnostic graphs, it can be concluded that the best model among all these models is still Gompertz.

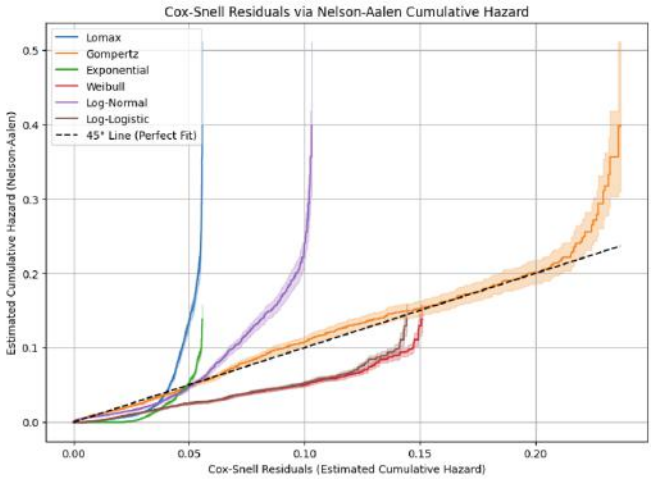

Figure 6: Cox-Snell Residual Diagnostic Plots.

Clinical Implications of Diagnoses

Clinical risk assessment and patient care are directly impacted by the Gompertz model's superior performance. First, it recommends that time-dependent risk scores that rise exponentially with ART duration and patient age be included in mortality risk assessment instruments for HIV patients. Long-term mortality risk will consistently be underestimated by static risk calculators that assume constant hazard, as suggested by the exponential model. Second, especially for patients who have been on ART for long periods of time, the exponentially increasing risk highlights the significance of proactive management of age-related comorbidities, such as cardiovascular screening, cancer surveillance, and bone health monitoring [2, 14]. Third, follow-up intensity and resource allocation should be dynamically modified based on treatment duration, with more frequent monitoring for long-term survivors who are at higher risk of dying irrespective of virological suppression. Among the models we fit, the Gompertz distribution emerged as the best fit. It had the smallest AIC value of 45,943.33 and BIC value of 45,961.58, and its residuals tracked closest to the 45-degree line compared with those of other models. The Weibull model is a suitable alternative for situations requiring computational simplicity. Conversely, the Exponential model underestimates long-term mortality risk and is not recommended for HIV survival predictions. Unreliable and inconsistent results were obtained using Lomax. This model does not assume a declining risk with very heavy tails; instead, it fits well with a progressive pattern of survival and mortality among patients with HIV/AIDS. Figure **7** presents Quantile-Quantile (Q-Q) plots comparing observed Cox-Snell residual quantiles to theoretical exponential quantiles for model diagnostics. The Gompertz distribution clearly shows the best fit with R^2 = 0.9817 and its residuals lying very close to perfect fit. The Weibull distribution fits well with R^2 = 0.9168 but exhibits some discrepancies at higher quantiles. However, the Exponential and Lomax distributions display very poor fit with R^2 = 0.5989 and indicate some non-linear behavior at higher quantiles, thus confirming their inefficiency. These findings, alongside Cox-Snell residual diagnostics from Figure **6**, strongly advocate for the Gompertz distribution as the preferred parametric model for the HIV cohort.

4. DISCUSSION

This study applied parametric survival models to a large retrospective cohort of 2,794 HIV-positive patients on ART, with 124 deaths observed. Our findings demonstrate that demographic and clinical factors, particularly age, sex, and baseline weight, significantly influence mortality risk. Identifying demographic factors such as age, gender, and clinical markers such as WHO stage and recent weight, as major predictors of mortality risk, provides useful information for the customization of patient care and enhancing ART adherence programs. The outstanding goodness-of-fit statistics also highlight its potential utility in making predictions on patient outcomes and

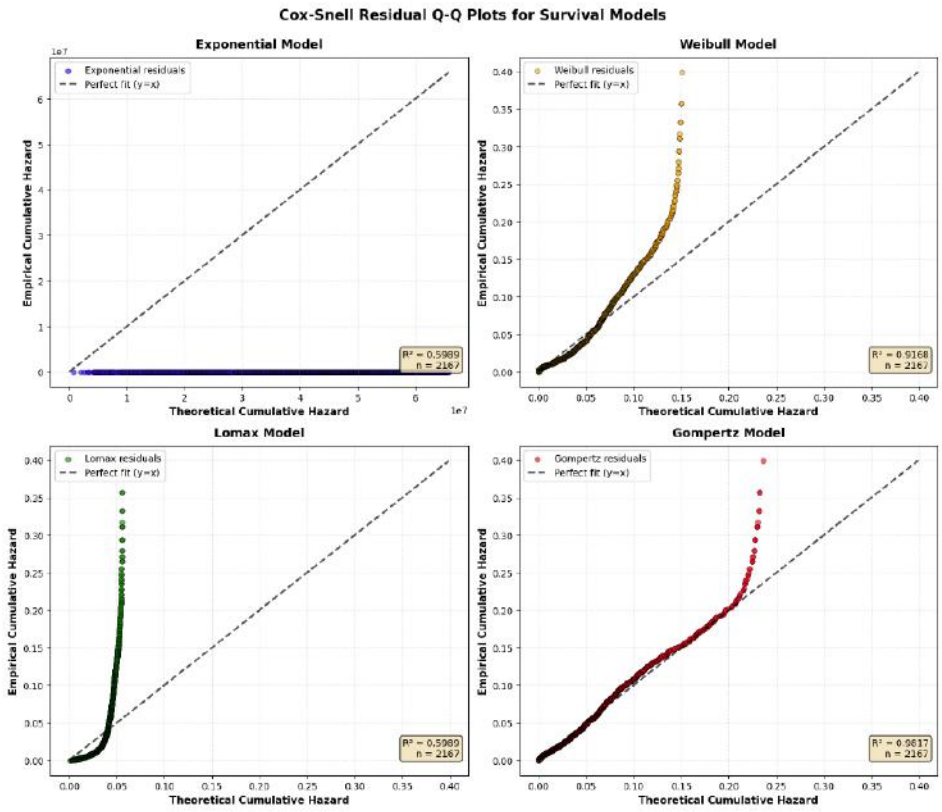

Figure 7: Quantile-Quantile (Q-Q) Plots of Cox-Snell Residuals.

informing clinical practices. In addition, the results of this analysis, especially the comparative efficiency of various parametric models, can help develop more precise prognostic tools for HIV patients who have varied treatment protocols [15]. These tools can eventually help improve personalized treatment regimens so that medical professionals can properly respond to each patient's individual needs. By incorporating these findings into everyday practice, clinicians can promote improved health outcomes and maximize the use of healthcare system resources.

The Kaplan-Meier estimator revealed differences in survival by sex, though log-rank testing indicated only modest evidence of statistical significance. These results align with prior studies reporting sex-based disparities in ART outcomes, with male patients often exhibiting poorer survival [53, 54]. The Cox proportional hazards model confirmed that older age and lower body weight are associated with higher mortality, consistent with evidence that advanced age and poor nutritional status worsen ART outcomes [7]. Male sex was also identified as an independent risk factor, increasing hazard by 60%, which has been attributed to differences in healthcare-seeking behavior and adherence patterns. Baseline CD4 count did not emerge as a strong predictor, which contrasts with earlier studies [51], possibly due to improved ART initiation policies that reduce reliance on CD4 thresholds.

The observed instability in WHO clinical stage estimates reflects a limitation of the data structure rather than a lack of clinical relevance. The wide confidence intervals and non-significant p-values results from a sparse distribution of event data across stage categories and an overall low event rate of 4.4%. This phenomenon is called quasi-complete separation, leads to inflated parameter estimates with undefined standard errors [55, 56], particularly in pediatric WHO stages where sample sizes are often under 20 patients. Among adults, there is a preponderance of event data within advanced stages III and IV, but against the rule for stable Cox regression modeling requiring at least 10-15 event observations per predictor [55]. It is important to note that the statistical instability of WHO staging does not diminish its useful role as a predictor for survival rates among HIV patients.It has been revealed that patients with an advanced stage of WHO have an odds ratio of death, and it's significantly higher among ART-naive patients and immunosuppressed patients who initiate antiretroviral therapy [49, 56] However, this study's cohort largely comprised stable ART patients with high survival, resulting in insufficient

data to accurately calculate stage-specific hazards. This reflects a challenge in HIV survival research, where improving ART effectiveness and declining mortality rates may hinder the quantification of traditional risk factors like WHO staging due to limited statistical power. To overcome this limitation, future research should employ several strategies. Firstly, combining data from various locations or countries can enhance the number of events, allowing for more accurate estimation of WHO stage effects [13]. Secondly, merging WHO stages into larger groups (such as early stages I-II versus advanced stages III-IV) might enhance the stability of estimates while maintaining clinical relevance. Thirdly, using WHO staging as a time-varying covariate instead of a baseline predictor could more effectively capture its evolving relationship with mortality as patients move through different disease stages during treatment [7]. Fourthly, alternative analytical methods like Bayesian hierarchical models with informative priors derived from historical data can stabilize estimates when event rates are low [56]. Despite the instability in WHO stage estimates, our model successfully identified strong predictors-age, sex, and body weight-that offer clinically actionable risk stratification for HIV patients on ART.

Model diagnostics revealed partial violations of the proportional hazards assumption, particularly for WHO staging and body weight, underscoring the importance of testing this assumption in HIV survival studies. This justified our use of fully parametric survival models, which provide greater flexibility in capturing time-varying hazards. Among the parametric models, the Gompertz distribution offered the best fit, as evidenced by the lowest AIC and BIC values, and Cox-Snell residuals closely aligned with the 45° line. The Weibull model followed closely, while the Exponential and Lomax distributions failed to adequately capture long-term mortality risk. The superiority of the Gompertz model is consistent with its capacity to model monotonic hazard increases over time, a pattern well aligned with HIV progression and treatment-related dynamics [52, 57]. Importantly, the Gompertz model also provided biologically interpretable results, capturing the increased hazard associated with aging and prolonged ART exposure. Integrating longitudinal markers such as body weight and WHO staging into such flexible parametric frameworks may improve patient-level risk prediction and enable earlier intervention strategies [7, 53].

The strengths of this study include the use of multiple survival modeling approaches, robust diagnostics, and model comparison based on information criteria. Furthermore, the large sample size and random sampling enhance generalizability. However, limitations include potential unmeasured confounding (e.g., ART adherence, socioeconomic factors) and the exclusion of patients with incomplete records. The Lomax model failed to converge, highlighting challenges in estimating heavy-tailed distributions with censored HIV data. Our findings emphasize the need for sex-specific and age-sensitive interventions in ART programs. Weight monitoring should be prioritized as an early warning indicator of adverse outcomes. Methodologically, this study highlights the value of combining semi-parametric and parametric models, with the Gompertz distribution emerging as particularly suitable for HIV survival analysis. Future research should explore dynamic survival models incorporating time-dependent covariates, including longitudinal viral load and CD4 trajectories.

Overall, both non-parametric and regression-based approaches highlight the clinical importance of age, sex, and weight as predictors of mortality among HIV patients on ART. The Cox regression confirmed their significance with robust hazard estimates, while parametric modeling demonstrated that the Gompertz distribution best captured survival dynamics. These findings suggest that the hazard of mortality in this population increases with time and is strongly shaped by demographic and baseline health factors.

5. CONCLUSION

The Gompertz model is specifically designed to address an exponentially rising hazard rate, a pattern frequently seen in chronic illnesses where the risk of death increases with age or the length of the disease [10]. This feature makes it particularly apt for modeling mortality in HIV-positive individuals, where the ongoing progression of the disease and prolonged antiretroviral treatment result in a variable hazard function. In contrast, models like the exponential assume a steady hazard, while the Weibull and Lomax distributions provide greater flexibility in depicting monotonic and heavy-tailed hazard patterns, respectively.

However, choosing a parametric survival model requires careful consideration of the data's structure, as incorrect assumptions about the hazard shape can lead to skewed estimates and unreliable predictions [11]. Therefore, validating the selected model using fit indices and ensuring it aligns with clinical knowledge is

essential for credible survival analysis. Incorporating expert opinion, as shown in some studies, can further refine survival projections and enhance model reliability, especially when differentiating between parametric models with similar statistical fits [58].

Although the Gompertz model was most effective in capturing the survival dynamics of this group, examining other flexible distributions-such as the Type I heavy-tailed Weibull or the alpha power transformed inverse Lindley distribution-could improve predictive accuracy by accommodating complex hazard structures and long-tailed survival patterns often found in clinical datasets [59]. Overall, survival outcomes for HIV-positive patients on ART were affected by demographic and clinical factors, with age, sex, and body weight being significant predictors. While the Cox model provided reliable hazard estimates, the parametric analysis showed that the Gompertz distribution best represents the survival trajectory in this population. These findings underscore the importance of flexible parametric models in enhancing the precision of survival predictions and guiding targeted interventions in HIV management.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest.

DATA AND CODE AVAILABILITY STATEMENT

The data and python codes used in this study are publicly available on https://github.com/obulezi12345-svg/AI-Powered-Mortality-Prediction-for-HIV-AIDS-Patients-on-ART-in-Nigeriahttps://github.com/obulezi12345-svg/AI-Powered-Mortality-Prediction-for-HIV-AIDS-Patients-on-ART-in-Nigeria

REFERENCES

[1] Md Faiyazuddin, *et al*. The impact of artificial intelligence on healthcare: a comprehensive review of advancements in diagnostics, treatment, and operational efficiency. In: Health Science Reports 2025; 8(1): e70312. https://doi.org/10.1002/hsr2.70312

[2] Trickey A, Zhang L, Sabin CA, Sterne JAC. Life expectancy of people with HIV on long-term antiretroviral therapy in Europe and North America: a cohort study. In: The Lancet Healthy Longevity 2022; 3: S2. https://doi.org/10.1016/S2666-7568(22)00063-0

[3] Althoff KN, *et al*. The shifting age distribution of people with HIV using antiretroviral therapy in the United States. In: Aids 2022; 36(3): 459-471. https://doi.org/10.1097/QAD.0000000000003128

[4] Teklehaymanot AN, Lemma TB, Gudina EK, Getnet M, Amdisa D, Dadi LS. Predictors of Mortality among Adult People Living with HIV and Its Implications for Appointment Spacing Model Approach Care in Southwest Ethiopia. In: 2020. https://doi.org/10.21203/rs.3.rs-35093/v1

[5] Kumssa TH, Mulu A, Mihret A, Asfaw ZG. Competing risks multi-state model for time-to-event data analysis of HIV/AIDS: a retrospective cohort national datasets, Ethiopia. In: BMC Infectious Diseases 2024; 24(1): 1412. https://doi.org/10.1186/s12879-024-10280-9

[6] Korenromp EL, Williams BG, Schmid GP, Dye C. Clinical prognostic value of RNA viral load and CD4 cell counts during untreated HIV-1 infection—a quantitative review. In: PloS One 2009; 4(6): e5950. https://doi.org/10.1371/journal.pone.0005950

[7] Dessie ZG, Zewotir T, Mwambi H, North D. Modelling of viral load dynamics and CD4 cell count progression in an antiretroviral naive cohort: using a joint linear mixed and multistate Markov model. In: BMC infectious diseases 2020; 20(1): 246. https://doi.org/10.1186/s12879-020-04972-1

[8] Haushona N, Esterhuizen TM, Thabane L, Machekano R. An empirical comparison of time-to-event models to analyse a composite outcome in the presence of death as a competing risk. In: Contemporary Clinical Trials Communications 2020; 19: 100639. https://doi.org/10.1016/j.conctc.2020.100639

[9] Payne CF, Houle B, Chinogurei C, Herl CR, Kabudula CW, Kobayashi LC, Salomon JA, Manne-Goehler J. Differences in healthy longevity by HIV status and viral load among older South African adults: an observational cohort modelling study. In: The Lancet HIV 2022; 9(10): e709-e716. https://doi.org/10.1016/S2352-3018(22)00198-9

[10] Glaubius R, *et al*. Disease progression and mortality with untreated HIV infection: Evidence synthesis of HIV seroconverter cohorts, antiretroviral treatment clinical cohorts and population-based survey data. In: Journal of the International AIDS Society 2021; 24: e25784. https://doi.org/10.1002/jia2.25784

[11] Mustefa YA, Chen D-G. Accelerated failure-time model with weighted least-squares estimation: application on survival of HIV positives. In: Archives of Public Health 2021; 79(1): 88. https://doi.org/10.1186/s13690-021-00617-0

[12] Omer ME, Mustafa M, Ali N, Rahman NHA. Non-Mixture Cure Model Estimation in Bladder Cancer Patients: A Novel Approach with Exponentiated Weibull Exponential Distribution. In: Asian Pacific Journal of Cancer Prevention: APJCP 2023; 24(12): 4167. https://doi.org/10.31557/APJCP.2023.24.12.4167

[13] Hussain S, Rashid MS, Ul Hassan M, Ahmed R. The generalized exponential extended exponentiated family of distributions: Theory, properties, and applications. In: Mathematics 2022; 10(19): 3419. https://doi.org/10.3390/math10193419

[14] Skalski JR, Whitlock SL. Vitality models found useful in modeling tag-failure times in acoustic-tag survival studies. In: Animal Biotelemetry 2020; 8(1): 26. https://doi.org/10.1186/s40317-020-00213-z

[15] Dzinza R, Ngwira A. Comparing parametric and Cox regression models using HIV/AIDS survival data from a retrospective study in Ntcheu district in Malawi. In: Journal of Public Health Research 2022; 11(3): 22799036221125328. https://doi.org/10.1177/22799036221125328

[16] Delva W, Eaton JW, Meng F, Fraser C, White RG, Vickerman P, Boily M-C, Hallett TB. HIV treatment as prevention: optimising the impact of expanded HIV treatment programmes. In: PLoS Medicine 2012; 9(7): e1001258. https://doi.org/10.1371/journal.pmed.1001258

[17] Onyekwere CK, Nwankwo CK, Abonongo J, Asogwa EC, Shafiq A. Economic growth dynamics: a machine learning-augmented nonlinear autoregressive distributed lag model of

asymmetric effect. In: Innovation in Computer and Data Sciences 2025; 1(1): 9-31. https://doi.org/10.64389/icds.2025.01125

[18] Onyekwere CK, Nwankwo CK, Obulezi OJ, Ezeilo CI. The featurevalue paradox: Unsupervised discovery of strategic archetypes in the smartphone market using machine learning. In: Journal of Artificial Intelligence in Engineering Practice 2025; 2(2): 65-72. https://doi.org/10.21608/jaiep.2025.420689.1024

[19] Onyekwere CK, Nwankwo CK, Apameh DG. A Hybrid Machine Learning Framework for Multi-Objective Performance Optimization and Anomaly Detection in Maritime Operations. In: Innovation in Computer and Data Sciences 2026; 2(1). https://doi.org/10.64389/icds.2026.02131

[20] Asogwa EC, Nwankwo MP, Oguadimma EE, Okechukwu CP, Suleiman AA. Hybrid LSTM-CNN deep learning framework for stock price prediction with google stock and reddit sentiment data. In: Innovation in Computer and Data Sciences 2025; 1(1): 32-50. https://doi.org/10.64389/icds.2025.01126

[21] Ugbor G, Jamal F, Khan S, Shawki AW. Generative AI for drug discovery: Accelerating molecular design with deep learning using Nigerian local content. In: Innovation in Computer and Data Sciences 2025; 1(1): 66-77. https://doi.org/10.64389/icds.2025.01128

[22] Nnaekwe K, Ani E, Obieke V, Okechukwu C, Usman A, Othman M. Forecasting seasonal rainfall with time series, machine learning and deep learning. In: Innovation in Computer and Data Sciences 2025; 1(1): 51-65. https://doi.org/10.64389/icds.2025.01127

[23] Asogwa EC, Okechukwu OM, Charles EI, Tochukwu B. A Medical Intelligent Process Model Using Ontology Based Technique. In: International Journal of Simulation: Systems, Science & Technology 2024; 25(1). https://doi.org/10.5013/IJSSST.a.25.01.03

[24] Madut DB, Park LP, Yao J, Reddy EA, Njau B, Ostermann J, Whetten K, Thielman NM. Predictors of mortality in treatment experienced HIV-infected patients in northern Tanzania. In: Plos One 2020; 15(10): e0240293. https://doi.org/10.1371/journal.pone.0240293

[25] de la Mora L, Mallolas J, Ambrosioni J. Epidemiology, treatment and prognosis of HIV infection in 2024: A practical review. In: Medicina ClÃLınica (English Edition) 2024; 162(11): 535-541. https://doi.org/10.1016/j.medcle.2023.12.010

[26] Little RJA, Rubin DB. Statistical analysis with missing data. John Wiley & Sons 2019. https://doi.org/10.1002/9781119482260

[27] Pedersen AB, Mikkelsen EM, Cronin-Fenton D, Kristensen NR, Pham TM, Pedersen L, Petersen I. Missing data and multiple imputation in clinical epidemiological research. In: Clinical Epidemiology 2017; 157-166. https://doi.org/10.2147/CLEP.S129785

[28] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. In: BMJ 2009; 338. https://doi.org/10.1136/bmj.b2393

[29] Cohen JE, Bohk-Ewald C, Rau R. Gompertz, Makeham, and Siler models explain Taylor's law in human mortality data. In: Demographic Research 2018; 38: 773-841. https://doi.org/10.4054/DemRes.2018.38.29

[30] Ramaki Z, Alizadeh M, Tahmasebi S, Afshari M, Contreras-Reyes JE, Yousof HM. TheWeighted FlexibleWeibull Model: Properties, Applications, and Analysis for Extreme Events. In: Mathematical and Computational Applications 2025; 30(2): 42. https://doi.org/10.3390/mca30020042

[31] Jafari AA, Tahmasebi S. Gompertz-power series distributions. In: Communications in Statistics-Theory and Methods 2016; 45(13): 3761-3781. https://doi.org/10.1080/03610926.2014.911904

[32] Witten M, Satzer W. Gompertz survival model parameters: Estimation and sensitivity. In: Applied Mathematics Letters 1992; 5(1): 7-12. https://doi.org/10.1016/0893-9659(92)90125-S

[33] Wilson DL. A comparison of methods for estimating mortality parameters from survival data. In: Mechanisms of Ageing and Development 1993; 66(3): 269-281. https://doi.org/10.1016/0047-6374(93)90014-I

[34] Albrecher HN, Bladt M, Bladt M. Multivariate fractional phase-type distributions. In: Fractional Calculus and Applied Analysis 2020; 23(5): 1431-1451. https://doi.org/10.1515/fca-2020-0071

[35] Hussam E, Almetwally EM. Statistical inference on the perk distribution under progressive-stress Type-II Adaptive Progressive Hybrid Censoring. In: 2024. https://doi.org/10.21203/rs.3.rs-5361936/v1

[36] Di Palo C. On a closed-form expression and its approximation to Gompertz life disparity. In: Demographic Research 2023; 49: 1-12. https://doi.org/10.4054/DemRes.2023.49.1

[37] Li H, Tan KS, Tuljapurkar S, Zhu W. Gompertz law revisited: Forecasting mortality with a multi-factor exponential model. In: Insurance: Mathematics and Economics 2021; 99: 268-281. https://doi.org/10.1016/j.insmatheco.2021.03.018

[38] Lin GD, Dou X, Kuriki S. The bivariate lack-of-memory distributions. In: Sankhya A 2019; 81(2): 273-297. https://doi.org/10.1007/s13171-017-0119-1

[39] Rajitha CS, Akhilnath A. Generalization of the Lindley distribution with application to COVID-19 data: Rajitha CS and A. Akhilnath. In: International Journal of Data Science and Analytics 2025; 20(2): 291-311. https://doi.org/10.1007/s41060-022-00369-2

[40] Osman AA, Esse AA, Muse AH. Analyzing factors affecting age at first birth among married women in Somalia: a Bayesian shared frailty modelling approach using SDHS 2020. In: BMC Women's Health 2025; 25(1): 346. https://doi.org/10.1186/s12905-025-03900-2

[41] Kearns B, Stevens J, Ren S, Brennan A. How uncertain is the survival extrapolation? A study of the impact of different parametric survival models on extrapolated uncertainty about hazard functions, lifetime mean survival and cost effectiveness. In: Pharmacoeconomics 2020; 38(2): 193-204. https://doi.org/10.1007/s40273-019-00853-x

[42] Alnssyan B, Ahmad Z, Malela-Majika J-C, Seong J-T, Shafik W. On the identifiability and statistical features of a new distributional approach with reliability applications. In: AIP Advances 2023; 13(12). https://doi.org/10.1063/5.0178555

[43] Tolba AH, Muse AH, Fayomi A, Baaqeel HM, Almetwally EM. The Gull Alpha Power Lomax distributions: Properties, simulation, and applications to modeling COVID-19 mortality rates. In: Plos One 2023; 18(9): e0283308. https://doi.org/10.1371/journal.pone.0283308

[44] Al-Essa LA, Abdel-Hamid AH, Alballa T, Hashem AF. Reliability analysis of the triple modular redundancy system under step-partially accelerated life tests using Lomax distribution. In: Scientific Reports 2023; 13(1): 14719. https://doi.org/10.1038/s41598-023-41363-3

[45] Ferreira PH, Ramos E, Ramos PL, Gonzales JFB, Tomazella VLD, Ehlers RS, Silva EB, Louzada F. Objective Bayesian analysis for the Lomax distribution. In: Statistics & Probability Letters 2020; 159: 108677. https://doi.org/10.1016/j.spl.2019.108677

[46] Benkhelifa L. Modi linear failure rate distribution with application to survival time data. In: arXiv preprint 2025; arXiv:2509.20831.

[47] Imran M, Alsadat N, Tahir MH, Jamal F, Elgarhy M, Ahmad H, Johannssen A. The development of an extended Weibull model with applications to medicine, industry and actuarial sciences. In: Scientific Reports 2024; 14(1): 12338. https://doi.org/10.1038/s41598-024-61308-8

[48] Nkomo W, Oluyede B, Chipepa F. Type I heavy-tailed family of generalized Burr III distributions: properties, actuarial measures, regression and applications. In: Statistics in Transition New Series 2025; 26(1): 93-115. https://doi.org/10.59139/stattrans-2025-006

[49] Mansournia MA, Geroldinger A, Greenland S, Heinze G. Separation in logistic regression: causes, consequences, and control. In: American Journal of Epidemiology 2018; 187(4): 864-870. https://doi.org/10.1093/aje/kwx299

[50] Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. In: Statistics in Medicine 2006; 25(1): 127-141. https://doi.org/10.1002/sim.2331

[51] Sabin CA. Do people with HIV infection have a normal life expectancy in the era of combination antiretroviral therapy? In: BMC Medicine 2013; 11(1): 251. https://doi.org/10.1186/1741-7015-11-251

[52] Crowther MJ, Look MP, Riley RD. Multilevel mixed effects parametric survival models using adaptive Gauss-Hermite quadrature with application to recurrent events and individual participant data meta-analysis. In: Statistics in Medicine 2014; 33(22): 3844-3858. https://doi.org/10.1002/sim.6191

[53] Shoko C, Chikobvu D. A superiority of viral load over CD4 cell count when predicting mortality in HIV patients on therapy. In: BMC Infectious Diseases 2019; 19(1): 169. https://doi.org/10.1186/s12879-019-3781-1

[54] Engsig FN, *et al*. Long-term mortality in HIV-positive individuals virally suppressed for. 3 years with incomplete CD4 recovery. In: Clinical Infectious Diseases 2014; 58(9): 1312-1321.

[55] Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. In: American Journal of Epidemiology 2007; 165(6): 710-718. https://doi.org/10.1093/aje/kwk052

[56] Greenland S, Mansournia MA. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. In: Statistics in Medicine 2015; 34(23): 3133-3143. https://doi.org/10.1002/sim.6537

[57] Brennan AT, Maskew M, Sanne I, Fox MP. The interplay between CD 4 cell count, viral load suppression and duration of antiretroviral therapy on mortality in a resource-limited setting. In: Tropical Medicine & International Health 2013; 18(5): 619-631. https://doi.org/10.1111/tmi.12079

[58] Harvey NC, McCloskey E, Kanis JA, Compston J, Cooper C. Cost-effective but clinically inappropriate: new NICE intervention thresholds in osteoporosis (Technology Appraisal 464). In: Osteoporosis International 2018; 29(7): 1511-1513. https://doi.org/10.1007/s00198-018-4505-x

[59] Tekelehaimanot AN, Belachew T, Gudina EK, Getnet M, Amdisa D, Dadi LS. Predictors of mortality among adult people living with HIV and its implications for appointment spacing model approach care. In: Ethiopian Journal of Health Sciences 2021; 31(5). https://doi.org/10.4314/ejhs.v31i5.3

Received on 12-11-2025 Accepted on 17-12-2025 Published on 30-12-2025

https://doi.org/10.6000/1929-6029.2025.14.82