Scientific Journal of
Biomedical Engineering & Biomedical Science

Research Article

Parametric and Semi-Parametric Survival Models with Application to Diabetes Data - 🔓

Musa Ganaka Kubi[1]*, Lasisi KE[1] and Bello Abdulkadr Rasheed[2]

[1]*Department of Mathematical Sciences, Abubakar Tafawa Balewa University, Bauchi, Bauchi State Nigeria*

[2]*Department of Mathematics and Statistics, Federal Polytechnic, Bauchi, Bauchi State, Nigeria*

***Address for Correspondence:** Musa Ganaka Kubi, Department of Mathematical Sciences, Abubakar Tafawa Balewa University, Bauchi, Bauchi State, Nigeria, Tel: +234-806-790-3755;
E-mail: ganakamusak@gmail.com

Submitted: 26 October 2022; **Approved:** 29 November 2022; **Published:** 30 November 2022

Cite this article: Kubi MG, Lasisi KE, Rasheed BA. Parametric and Semi-Parametric Survival Models with Application to Diabetes Data. Sci J Biomed Eng Biomed Sci. 2022 Nov 30;3(1): 001-010.

Abstract

In survival analysis, correlated survival data with potential censoring are frequently encountered. Accelerated Failure Time (AFT) and Cox Proportional Hazards Model (CPHM) models that traditionally assume independent responses may not be suitable when there are dependencies among observed survival periods. In this study, the performance of parametric and semi-parametric survival models with and without random effect were compared. Data on diabetes mellitus was collected from three selected Hospitals in Nasarawa State for the period of five years (2016-2020). The results from the analysis revealed that the Weibull AFT model with inverse Gaussian Random effect distribution had the least AIC and BIC values indicating that it outperformed the other models considered in the study. Based on empirical results, it was found that area of residence, level of education, alcohol consumption status and body mass index were the risk factors of Diabetes mellitus. The study recommends that health expert should use the Weibull AFT model with inverse Gaussian Random effect distribution for predicting the risk factors of diabetes mellitus especially when the data are correlated.

Keywords: AFT; CPHM; Diabetes; Hazard; Survival

INTRODUCTION

Survival analysis is a set of statistical techniques for data analysis where the dependent variable is the time until the occurrence of an event of interest. This event of interest can be death, occurrence of a disease, failure of a device or recovery from a disease. Survival analysis is hampered by censoring, or when complete information on a subject's event is not accessible, therefore general methods of statistical inference are not valid [1]. Numerous disciplines, including social sciences, engineering, and public health, use survival analysis. For example, in the medical sciences, the term "time to event" might refer to the amount of time until a cancer study's tumors reappear. Survival analysis is the term that is most frequently used and recognized, despite the fact that various disciplines may emphasize slightly different methodologies and procedures [2].

There are basically three main methods used in the analysis of survival data. This includes the non-parametric, semi-parametric and parametric survival models. In the past, non-parametric methods such as the log-rank test and the Kaplan-Meier (K-M) which requires no distributional assumption were used to estimate the survival function, compare the survival experiences of various groups and also estimate the survival probability in clinical set up [3]. The baseline hazard function incorporated in the Cox Proportional Hazards (CPH) model does not need to follow any probability distribution, making it a semi-parametric survival model that is used for the study of time to event data [4,5]. The parametric survival models used in analyzing survival data give similar results but each has its own unique procedure usually under specific assumptions or no assumptions [3].

Globally, the prevalence of chronic diseases like Diabetes Mellitus (DM) is rising, and these conditions are linked to reduced quality of life and increased financial burden. It is therefore crucial to develop preventive interventions for these conditions [6]. Along with infectious diseases and dietary issues, the burden of chronic diseases is increasing in developing countries. Although they account for a sizable amount of disease burden in African nations, chronic diseases are not adequately prevented or controlled [7]. There are few risk factors that are connected to nutrition and lifestyle choices that are common to the three primary chronic diseases: cancer, diabetes, and cardiovascular disease. These include conditions that are on the rise in many African nations, such as high blood pressure, high cholesterol, tobacco usage, excessive alcohol use, insufficient consumption of fruits and vegetables, obesity, and being overweight or sedentary [8].

Empirically, Kassa TH [9] used Bayesian Accelerated Failure Time model and classical Accelerated Failure Time to identify the determinants risk factors of diabetes mellitus patients in Addis Ababa, Ethiopia and found that Body Mass Index, age category, types of diabetic disease, alcohol consumption, diabetic complication, cholesterol level, blood pressure, family history of diabetic, fasting blood sugar, density lipoprotein, comorbidity, triglyceride level and smoking habit are significantly related to the survival time until death of diabetic patients. Lomo SI, et al. [10] used K-Meier curves and CPH regression to study the factors that affect the survival time of patients with type two DM in Indonesia and found that age, gender, diagnosis complication, comorbidity, intermittent blood glucose levels and treatment profile are the factors effecting survival of patients with type II DM. Naim S, et al. [11] studied the survival time of DM patients with hemodialysis in Indonesia using K-M curves and CPH regression and found that age and gender has significant effect on the survival DM patients. Uloko AE, et al. [12] study the prevalence and risk factors for diabetes mellitus in Nigeria via systematic review and meta-analysis and found that family history of DM, urban dwelling, unhealthy dietary habits, cigarette smoking, older age, physical inactivity and obesity were the risk factors for the pooled prevalence of DM. Hordofa SB, et al. [13] utilized multivariable CPH regression model to model the survival time of DM patients who were under follow-up at Nekemte referral Hospital, Ethiopia and found that body mass index, tobacco use, alcohol use, diabetic type diagnosed, blood pressure, and family history of diabetes mellitus were significantly related to survival of diabetic patients. Adedotun AF, et al. [14] study the survival time distribution for DM patients at the National Hospital Abuja using CPH model and found that the distributions of survival time of patients differ based on the four age categories. The study revealed further that there is no sex related multiplicative impact. Belay A, et al. [15] utilized K-M and CPH model to study the time to recovery of DM patients in Minlik Referral Hospital, Ethiopia and found that sex, Spdrt and regimen contribute significantly to survival time to recovery of patients. Zhao Z, et al. [16] examine the survival of type 2 DM and the risk factors for mortality in one suburb cohort of Beijing, China using K-M analysis, and CPH model and found that older age, higher systolic blood pressure, lower body mass index and lower estimated glomerular filtration rate are significant risk factors of mortality from type 2 diabetes. Badmus NI, et al. [17] examines and upgrades a two-parameter double exponential distribution to a four-parameter beta double exponential distribution model by compounding the baseline distribution and beta link function to fit and analyzed death-cases data resulted from COVID-19 in Africa and Non-African Countries. The proposed Beta Double Exponential Model (BDEM) proved flexible and robust to fits than other distributions given the fact that the data is skewed. Similarly, Adeniran AT, et al. [18] invested an alternative and less rigorous method of deriving Gaussian probability distribution. The study found an approach independent of Lemas and Theorems and free from rigorous mathematical analysis.

Most existing literatures on studies related to survival time of diabetes patients utilized K-M and Cox regression survival models and only few were conducted in Nigeria. In addition, in survival analysis, correlated survival data with potential censoring are frequently encountered. This comprises research with multiple focus groups in which participants are grouped according to clinical or other environmental characteristics that affect anticipated survival time. For data that occurs in such a setting, the traditional cox proportional hazards and AFT models that assume observations as independent are inadequate. For Instance, Vincent and Ismaila performed parametric survival analysis of Tuberculosis data collected from Federal Medical Centre Bida. Three AFT parametric survival (Exponential, Weibull and Log logistic) were fitted. It was found that the Weibull model performed better. However, the study focused on studying the effect of the fixed covariate on the survival time without considering the cluster-specific (Hospital) random effect on the survival time. Furthermore, Cox proportional Hazard and parametric proportional hazard model with random effect has been proposed to account for such dependencies. However, less attention has been giving to Accelerated failure time model by incorporating a random effect parameter that will account for the dependencies of the correlated data. In this study, we fit an AFT model with random effect that will account for such dependencies using a diabetes mellitus data collected from three hospitals and the performance of the model was compared with the conventional AFT and cox proportional hazard model.

MATERIALS AND METHODS

Data on diabetes mellitus was collected from the three selected Hospitals for the period of five years (2016-2020).

Techniques of data analysis

Cox proportional hazards models with random effect: A semi-parametric model (Cox-proportional Hazards model) with random effect can be formulated as:

$$\lambda_i\left(t\right) = \lambda_0\left(t\right) exp(\beta_1 z_{1i}, \beta_2 z_{2i}, \ldots, \beta_p z_{pi} + \propto_j) \qquad (1)$$

Where λ_0 is the baseline hazard function, β_i is a vector of fixed effects corresponding to the covariates vectors z_i *and* $\propto_j$ is the per-subject random effect denotes the random effect associated with the j^th cluster. The random effect can be thought of as an intercept that modifies the linear predictors. This approach retains the full flexibility of Cox regression while accommodating associations among individual response times.

AFT parametric survival models with random effect parameter: In this study, we introduce a random effect component to the AFT model that accounts for lack of independencies by introducing a random effect component as:

$$\log T_i = \mu + \propto' X_i + \sigma\epsilon_i + b \qquad (2)$$

Where $\propto'$ is a vector of unknown regression coefficient, σ is a scale parameter, μ is the intercept parameter, the ϵ_i is the independently and identically distributed random errors, and the *b* is the cluster-specific random effects which are assumed to be independent, identically distributed random variables with density function *p*(*b*). Here we have assumed that the random effect *b* follows gamma and inverse Gaussian distribution with mean zero and variance θ, as defined as in the density function below:

$$f\left(Z\right) = \frac{Z^{\left(\frac{1}{\theta}\right)^{-1}}}{\theta^{\frac{1}{\theta}} \tilde{A}(\frac{1}{\theta})} \exp\left(\frac{-X_i}{\theta}\right), \theta > 0 \qquad (3)$$

Where $\tilde{A}(.)$ is the gamma function, it corresponds to the gamma distribution $Gam(\mu,\theta)$ with μ fixed to 1 for identifiability and its variance is θ the associate Laplace transform is:

$$L\left(\mu\right) = \left(1 + \frac{\mu}{\theta}\right)^{-\theta}, \theta > 0 \qquad (4)$$

Heterogeneity exists in the model if and only $\theta > 0$. So, the large values of θ reflect a greater degree of heterogeneity among groups and a stronger association within groups. The conditional survival and hazard function of the gamma frailty distribution is given by Gutierrez RG [19]:

$$S_\theta = \left[1 - \theta \ln\left(S\left(t\right)\right)\right]^{\frac{-1}{\theta}} \text{ and}$$

$$h_\theta\left(t\right) = h\left(t\right)\left[1 - \theta \ln(S\left(t\right))\right]^{-1} \qquad (5)$$

where S(t) and h(t) are the survival and the hazard functions of the baseline distributions. According to Hougaard, kendall's Tau measures the association between any two event times from the same cluster in the multivariate case in a Gamma distribution. The associations within group members are measured by Kendall's, which is given by:

$$\tau = \frac{\theta}{\theta + 2} \epsilon(0,1) \qquad (6)$$

The probability density function of an inverse Gaussian shared distributed random variable with parameter θ > 0 is given by:

$$f_Z(X_i) = \left(\frac{1}{2\pi\theta}\right)^{\frac{1}{2}} X_i^{-3/2} \exp\left(\frac{-\left(X_i - 1\right)^2}{2\theta Z_i}\right), \theta > 0, X > 0 \qquad (7)$$

For easy indemnification, we assume X has expected value equal to one and variance θ. The Laplace transformation of the inverse Gaussian distribution is:

$$L\left(s\right) = \exp\left[\frac{1 - \left(1 + 2\theta s\right)^{\frac{1}{2}}}{\theta}\right], \theta > 0, s > 0 \qquad (8)$$

For the inverse Gaussian frailty distribution, the conditional survival and hazard function is given by Gutierrez RG [19] in (9) and (10), respectively:

$$S_\theta\left(t\right) = \exp\{\frac{1}{\theta}(1 - \left[1 - 2\theta \ln\left\{S\left(t\right)\right\}\right]^{\frac{-1}{2}}\}, \theta > 0 \qquad (9)$$

and

$$h_\theta(t) = \exp\{\frac{1}{\theta}(\left[1-2\theta\ln\{S(t)\}\right]^{\frac{-1}{2}}\},\theta > 0 \tag{10}$$

where S(t) and h(t) are the survival and the hazard functions of the baseline distributions. With multivariate data, an Inverse Gaussian distributed frailty yields a Kendall's Tau given by:

$$\tau = \frac{1}{2} - \frac{1}{\theta} + 2\frac{\exp\left(\frac{2}{\theta}\right)}{\theta^2}\int_{\frac{2}{\theta}}^{\infty}\frac{\exp(-\mu)}{\mu}d\mu, \tau(0,\frac{1}{2}) \tag{11}$$

On the log survival time scale, the random effect can be thought of as an unobserved covariate that describes certain decreases or increases in event timings for distinct clusters. Within a cluster, all observations have a same unobserved random effect. The log of the survival time has a location-scale distribution in several survival time distributions, such as the Log-normal, Weibull, and Log-logistic distributions. Conditional on the random effects, the survivor function in (2) can be written in the form

$$S_{ij}(t/b_i) = s_0(\frac{loT - \mu - \propto_i Z_i - b}{\sigma} | b) \tag{12}$$

One assumption of the parametric model is that the survival time is assumed to follow a distribution with density function f(t). The AFT survival models considered in this study are: Exponential, Weibull, Log-Normal and Log-logistic survival distributions.

The Weibull AFT model

Survival time t is a positive random variable with Weibull probability density function can be expressed as:

$$f(t;\mu,\alpha) = \frac{\alpha}{\mu}\left(\frac{t}{\mu}\right)^{\alpha-1}\exp[\left(-\frac{t}{\mu}\right)^{\alpha}] \tag{13}$$

where, μ > 0 and α > 0 and the baseline hazard function of the distribution becomes:

$$h(t;\mu,\alpha) = \frac{\alpha}{\mu}\left(\frac{t}{\mu}\right)^{\alpha-1} \tag{14}$$

This yield the following survivorship functions: $S(t) = \exp[\left(-\frac{t}{\mu}\right)^{\alpha}]$ and the cumulative hazards function becomes $\mathrm{H}(\mathrm{t}) = \left(\frac{t}{\mu}\right)^{\alpha}$

Depending on the value of α, the hazard function can increase or decrease with increasing survival time. Hence, the Weibull model can yield an accelerated failure time model. Independent observations (t_i, δ_i), i ,...,n with survival time t_i and censoring indicator δi which has value of one if ith observation is not censored and zero when the ith observation is censored and let α be the unknown parameter. The likelihood function is:

$$L(\alpha) = \prod_{\boldsymbol{i}=1}^{\boldsymbol{n}}\{\boldsymbol{f_i}(\boldsymbol{t_i})\}^{\boldsymbol{ó_i}}\{\boldsymbol{S_i}(\boldsymbol{t_i})\}^{1-\boldsymbol{ó_i}}\}$$

$$= \prod_{\boldsymbol{i}=1}^{\boldsymbol{n}}\{\boldsymbol{h}(\boldsymbol{t_i})\}^{\boldsymbol{ó_i}}\ \boldsymbol{S_i}(\boldsymbol{t_i})\}$$

$$= \prod_{i=1}^{n}\{\left(\frac{\alpha}{\mu}\left(\frac{t}{\mu}\right)^{\alpha-1}\right)^{\boldsymbol{ó_i}}\exp[\left(-\frac{t}{\mu}\right)^{\alpha}\} \tag{15}$$

Re-parameterizing the Weibull distribution using λ= μ^-α then $h_0(t) =$ λα $\mathrm{t}^{\alpha-1}$ will be the baseline hazard function. Now incorporate covariates X in the hazard function, the Weibull regression models become:

$$h(t;\mu,\propto) = \lambda\alpha t^{\alpha-1}\exp(X \propto) \tag{16}$$

The exponential AFT model: The time data is skewed to the right with exponential distribution, the time of survival for a set of covariates ***X***, which is called, accelerated failure time is expressed as:

$$T = \exp(\mu + \propto' X + \sigma\varepsilon) \tag{17}$$

Where σ is the error component.

The survivorship function may be obtained by expressing in terms of time as:

(t, X, ∝) = (-t $e^{-\propto' X}$) and the hazard function of the exponential regression model is

h(t,X, ∝)= $e^{-\propto' X}$.

The log-logistic AFT model: Multiple covariate log-logistic accelerated failure time may be expressed as:

$$logT = \exp(\mu + \propto' X + \sigma\varepsilon) \tag{18}$$

Where σ is the scale parameter and ε is the residual (unexplained) variation in the transformed survival time. The survivorship function for the model in (18) is $s(t, X, \propto, \sigma) = ([1+\exp(z))^{-1}$

Where z is the standardized log-time outcome variable, that is;

$$z = \frac{(y - \mu - \propto_i X)}{\sigma}$$

and

$$Y = \mathrm{In}(t) \tag{19}$$

The odds of a survival time of at least t are,

$$OR = \frac{s(t,x,\propto,\sigma)}{1-s(t,x,\propto,\sigma)} = \exp(-z), \tag{20}$$

assumes that the covariate is dichotomous and coded 0 or 1. The odds-ratio at time *t* from the ratio the odds of a survival time evaluated at x= 0 and x= 1 is:

$$OR\left(x=1,x=0\right)=\frac{\frac{\exp[-\left(y-\mu-\propto_1 x_1\right)]}{\sigma}}{1+\frac{\exp[-\left(y-\mu-\propto_1 x_0\right)]}{\sigma}}$$

$$=exp\left(\frac{\propto_1}{\sigma}\right) \qquad (21)$$

This is independent of time.

The lognormal AFT model: The log-normal model assumes that ε ~N (0, 1). Let h(t) be the hazard function of T for the model (11) when β=0 i.e. $\beta0 = \beta_1 = ... = \beta_p = 0$. Then, h(t) has the following functional form:

$$h\left(t\right)=\frac{\phi\left(\frac{\log(t)}{\sigma}\right)}{[1-\phi\left(\frac{\log(t)}{\sigma}\right)\sigma t} \qquad (22)$$

where $\phi\left(t\right)=\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-t^2}{2}\right)$ is the probability density function,

and $\phi\left(t\right)=\int_{-\infty}^{t}\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-u^2}{2}\right)du$ is the cumulative distribution function of the standard normal distribution. The survival function (t/X) at any covariate x can be expressed as:

$$s\left(t\,/\,X\right)=\phi[\mu+\propto_1^{*}x_1+\ldots+\propto_p^{*}x_p-\propto\log\left(t\right)] \qquad (23)$$

Where $\propto=\frac{1}{\sigma},\propto_j^{*}=\frac{\propto j}{\sigma}$

for $j=0,1,\ldots,p$

This is the final survival model with intercept depending with t.

Method of estimation

According to Gutierrez RG [19], given the covariates information under assumptions of non-informative right -censoring and of independence between the censoring time and the survival time random variables, the marginal log-likelihood of the observed data is given by:

$$l_{marg}\left(\varphi,\propto,\theta;z,X\right)=\prod_{i=1}^{s}[(\prod_{j=1}^{n_i}(h_0\left(y_{ij}\right)exp\left(X_{ij}^{T}\propto\right)^{\delta_{ij}})$$

$$X\int_0^{\infty}z_i^{d_i}exp(-z_i\sum_{j=1}^{n_i}h_0\left(y_{ij}\right)exp\left(X_{ij}^{T}\propto\right)f(z_i)dz_{i]}$$

$$=\prod_{i=1}^{s}[(\prod_{j=1}^{n_i}(h_0(y_{ij})exp\left(X_{ij}^{T}\propto\right)^{\delta ij})\times$$

$$\left(-1\right)^{di}L^{di}(\sum_{j=1}^{n_i}h_0\left(y_{ij}\right)exp(X_{ij}^{T}\propto)] \qquad (24)$$

Taking the logarithm, the marginal likelihood is:

$$l_{marg}\left(\varphi,\propto,\theta;z,X\right)=\sum_{i=1}^{s}\{[\sum_{j=1}^{n_i}\delta_{ij}(\log(h_0\left(y_{ij}\right))$$

$$+X_{ij}^{T}\propto)]+\log[\left(-1\right)^{di}L^{d}([\sum_{j=1}^{n_i}h_0\left(y_{ij}\right)\exp(X_{ij}^{T}\propto)])]\} \qquad (25)$$

Where $d_i=\sum_{j=1}^{n_i}\delta_{ij}$ is the number of event in the $_i{}^{th}$ cluster, and $L^{(q)}(.)$ is the q^{th} derivative of the Laplace transform of the random effect distribution defined as:

$$L^{(q)}\left(s\right)=E\left[\exp\left(-Zs\right)\right]=\int_0^{\infty}\exp\left(-Z_{is}\right)f\left(Z_i\right)dz_i,s\geq 0$$

where φ represents a vector of parameters of the baseline hazard function, $\propto$ the vector of regression coefficients and θ the variance of the random effect. Estimates of φ, $\propto$, θ are obtained by maximizing the marginal log-likelihood of the above; this can be done if one is able to compute higher order derivatives $L^{(q)}(.)$ of the Laplace transform up to $q=\max\{d_1,\ldots,d_s\}$.

Model diagnostic

For the parametric regression problem, analogs of the semi parametric, residual plots can be made with a redefinition of the various residuals to incorporate the parametric form of the baseline hazard rates (Klein and Moeschberger 2003). The first of such residual is the Cox–Snell residual that provides a check of the overall fit of the model. The Cox–Snell residual, r_j, is defined by:

$$r_j=\hat{H}(T_j\mid X_j)$$

Where H is the cumulative hazard function of the fitted model. If the model fits the data, then the r_j's should have a standard (λ=1) exponential distribution, so that a hazard plot of r_j versus the Nelson-Aalen estimator of the cumulative hazard of the r_j's should be a straight line with slope 1. The best model will have the plots of the cumulative hazard close to the line of the residuals.

RESULTS AND DISCUSSION

The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values for each model is presented in table 1. The cox proportional hazard model with random effect had the least AIC and BIC. This suggests its efficiency over the conventional cox proportional hazard model (without random effect). In addition, the performance of the AFT model with and without the random effect was also considered. The result in table 1 revealed that the Weibull AFT with Inverse Gaussian random effect model has the least AIC and BIC indicating that it outperformed the exponential, log-logistics and lognormal models with and without the random effect when predicting the survival time of diabetes patient. Hence, the researcher based the interpretation of the results for AFT models on Weibull AFT with Inverse Gaussian random effect model.

Table 2 presents the estimated parameters for CPHM with and without random effect. The model selection criterion indicated that the CPHM with random effect outperformed the conventional CPHM. Hence, the interpretation of the result was based on CPHM with random effect. The random effect in the cox proportional

Table 1: AIC and BIC for diabetes data.

	No RE		**GAMMA RE**		**Inverse Gaussian RE**	
Model	**AIC**	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**
CPH	2578.9	2691.8	**2571.5**	**2681.6**		
Exponential	1005.1	1131.6	1007.1	1138.1	1007.1	1138.1
Weibull	1111.7	1242.7	864.5	999.9	**864.3**	**999.8**
Lognormal	1054.1	1185.1	1056.1	1191.6	1056.1	1191.6
Log logistic	1055.7	1186.8	1057.7	1193.2	1057.7	1193.2

Source: Computed using STATA.

Table 2: Cox proportional Hazard models for diabetes data.

	CPH (No RE)				**CPH (With RE)**			
Variable/Category	**B**	**HR**	**S. E**	***p*-value**	**B**	**HR**	**S. E**	***p*-value**
Gender								
Male	**Ref**							
Female	-.1475	0.8629	0.1495	0.395	-.1253	0.8822	0.1533	0.471
Residential Area								
Urban	**Ref**							
Peri-Urban	.7682	2.1559	0.6416	0.010	.7864	2.1955	0.6534	0.008
Rural	.4965	1.6429	0.3771	0.031	.4901	1.6325	0.3759	**0.033**
Education								
Tertiary	**Ref**							
Secondary	.9285	2.5308	0.7165	0.001	.9448	2.5722	0.7278	0.001
Primary	.3403	1.4054	0.2843	0.092	.3445	1.4113	0.2864	0.050
Non-Formal	.4112	1.5087	0.4836	0.199	.4431	1.5576	0.5002	0.168
Occupation								
Unemployed	**Ref**							
Employed	-.2285	0.7957	0.2523	0.471	-.2285	0.7957	0.2521	0.471
Trading	-.1064	0.8991	0.3218	0.766	-.1300	0.8781	0.3138	0.716
Farming	-.2022	0.8169	0.2154	0.443	-.2064	0.8135	0.2140	0.433
Retire	-.4868	0.6146	0.2037	0.142	-.5073	0.6021	0.1998	0.126
Smoking Status								
Non-Smoker	**Ref**							
Smoker	.5327	1.7035	0.6650	0.172	.4565	1.5785	0.6248	0.249
Drinking Status								
Non-Drunker	**Ref**							
Drunker	.6467	1.9093	0.3235	0.000	.6666	1.9477	0.3321	0.000
Exercise								
Regular	**Ref**							
Sometimes	.1634	1.1775	0.4725	0.684	.1297	1.1385	0.4574	0.747
Not at all	.1733	1.1892	0.1866	0.269	.1431	1.1538	0.1836	0.368
Sal								
Salt Intake								
Not at all	**Ref**							
As in food	-.1557	0.8558	0.1886	0.480	-.1536	0.8576	0.1896	0.487
At table	.2209	1.2472	0.3196	0.389	.2258	1.2533	0.3207	0.377
Age								
Below 50	**Ref**							
50 – 59	-.0551	0.9464	0.1747	0.765	-.0638	0.9382	0.1730	0.729

60 – 69	.0527	1.0541	0.2377	0.815	.0508	1.0521	0.2376	0.822
Above 70	.1280	1.1366	0.3185	0.648	.1104	1.1167	0.3133	0.694
BMI								
Normal	**Ref**							
Underweight	.2079	1.2311	0.2445	0.295	.2226	1.2493	0.2483	0.263
Overweight	.1239	1.1319	0.1521	0.472	.1131	1.1197	0.1539	0.512
Obsessed	.9334	2.5431	0.1825	0.044	.9289	2.5317	0.1833	0.045
SBP								
Low	**Ref**							
Normal	-.0803	0.9228	0.1925	0.100	-.0946	0.9097	0.1903	0.650
Pre-hypertension	-.0056	0.9944	0.2925	0.523	-.0622	0.9397	0.1962	0.766
High	-.0125	0.9876	0.2195	0.953	.0103	1.0104	0.2237	0.963
DBP								
Low	**Ref**							
Normal	-1.7142	0.1801	0.1738	0.101	-.5090	0.6011	0.3151	0.100
Pre-hypertension	-.1135	0.8927	0.1586	0.523	-.1179	0.8888	0.1583	0.508
High	-.0484	0.9528	0.7771	0.953	-.0743	0.9284	0.7578	0.927
θ						0.0100	0.0247	
τ						0.0050		
					Likelihood ratio (θ): $\chi^2 =$ 0.26; prob. = 0.306			

Source: Computed using STATA.

hazard model is assumed to follow the Gamma distribution with mean 1 and variance equal to theta (θ). The heterogeneity in the population of the study which is used as a clusters as estimated by the selected model is θ = 0.0100 and the dependence within the clusters (hospitals) is measured by Kendall's tau is τ = 0.0050. A variance of zero (θ = 0) indicate that the random effect component does not contribute to the model. A likelihood ratio test for the hypothesis θ = 0 indicates a chi-square $\chi 2$ value of 0.26 resulting to an insignificant p-value of 0.306. The implication of this findings is that the random effect parameter had insignificant contribution to the model.

The categorical variables: area of residence (peri-urban and rural residence), education level (primary and secondary), drinking status and BMI (Obsessed) significantly contributed to the hazard of diabetes mellitus. The estimated values (B = 0.7804, HR = 2.1955, S.E. = 0.6534, p = 0.008 and B = 0.4901, HR = 1.6325, S.E. = 0.3759, p = 0.033) for peri-urban and rural residence indicates that peri-urban and rural residence had a higher risk of dying with diabetes mellitus by a factor 0.1955 (19.6%) and 0.6325 (63.25%) respectively times higher than urban residence when other covariates are held constant. Similarly, the estimated values (B = 0.3445, HR = 2.4113, S.E. = 0.2864, p = 0.050 and B = 0.9448, HR = 2.5576, S.E. = 0.7278, p = 0.001) indicated that diabetes mellitus patients with primary and secondary school education level had higher risk of dying from diabetes by a factor 0.4223 (41.1%) and 0.5576 (55.8%) times higher as compared to those with tertiary level of education respectively. Furthermore, the estimated values (B = 0.6666, HR = 1.9477, S.E. = 0.3321, p = 0.000) suggested that diabetes mellitus patients that are drunker had a higher risk of dying from diabetes by a factor 0.9477 (94.8%) times higher as compared to those that do not drink. The estimated values (B = 0.9289, HR = 2.5317, S.E. = 0.1833, p = 0.045) revealed that diabetes patient that are obese had a higher risk of dying with diabetes by a factor 0.1833 (18.3%) as compared with those that have normal weight. However, gender, occupation, smoking status, exercise, salt intake and age were insignificant contributing factors.

The results of data analysis based on Exponential, Weibull, Log-normal and Log-logistic AFT models with and without random effect using AFT survival analysis were presented in table 3. The model selection criterion presented in table 1 indicates that Weibull with Inverse Gaussian Random effect distribution which has the minimum AIC and BIC values of 864.3 and 999.8 appears to be appropriate model as compared with other models considered in this study. This implies that Weibull inverse Gaussian random effect AFT model is more efficient model to describe determinant factors of time-to-event of diabetes mellitus patient. From table 3, the random effect parameter in this model is assumed to follow an inverse Gaussian distribution with mean 1 and variance equal to theta (θ). The heterogeneity in the population of the hospitals which were used as a clusters as estimated by the selected model is θ = 0.0543 and the dependency within the clusters (hospitals) is measured by Kendall's tau is τ = 0.0264. A variance of zero (θ = 0) indicate that the random effect component does not contribute to the model. A likelihood ratio test for the hypothesis θ = 0 was presented at the bottom of table 3 and indicates a chi-square $\chi 2$ value of 4.08 which resulted to a highly significant p-value of 0.022. The implication of this findings is that, the random effect component had significant contribution to the model. The estimated Kendall's tau (τ = 0.0264) shows that there is weak dependency within the cluster for Weibull inverse Gaussian random effect model. The estimate of shape parameter in the Weibull inverse Gaussian random effect AFT model is u = 2.0409. This value shows the shape of hazard function is unimodal because the value is greater than unity i.e., it increases up to some time and then decreases. The estimated values, standard error, Time Ratio (TR), estimated parameters of baseline

Table 3: Results of Weibull AFT models.

	Weibull (No RE)				Weibull (Gamma)				Weibull (Inverse Gaussian)			
Variable/Category	**B**	**TR**	**S. E**	***p*-value**	**B**	**TR**	**S. E**	***p*-value**	**B**	**TR**	**S. E**	***p*-value**
Intercept	1.3731	3.9476	0.1616	0.000	2.3755	10.7564	0.2414	0.000	2.3753	10.7542	0.2414	0.000
Gender												
Male	**Ref**											
Female	0.0736	1.0764	0.0595	0.216	0.0203	1.0205	0.08854	0.819	0.0191	1.0193	0.0884	0.829
Residential Area												
Urban	**Ref**											
Peri-Urban	0.1611	1.1748	0.1011	0.111	-0.4491	.6382	0.1485	0.002	-0.4502	.6375	0.1484	0.002
Rural	0.0817	1.0851	0.0647	0.207	-0.2501	.7787	0.1150	0.030	-0.2498	.7790	0.1149	0.030
Education												
Tertiary	**Ref**											
Primary	0.0186	1.0188	0.06387	0.771	-0.1862	.8301	0.1022	0.069	-0.1863	.8300	0.1022	0.048
Secondary	0.0800	1.0833	0.1151	0.487	-0.5263	.5908	0.1406	0.000	-0.5270	.5904	0.1405	0.000
Non-formal	0.2442	1.2766	0.1164	0.036	-0.2519	.7773	0.1627	0.122	-0.2534	.7762	0.1625	0.119
Occupation												
Unemployed	**Ref**											
Employed	0.1251	1.1333	0.1124	0.266	0.1255	1.1337	0.1585	0.429	0.1257	1.1339	0.1583	0.427
Trading	0.1897	1.2089	0.1273	0.136	0.1367	1.1465	0.1795	0.446	0.1377	1.1476	0.1793	0.442
Farming	0.1068	1.1127	0.0968	0.270	0.9983	2.7137	0.1322	0.450	0.1002	1.1054	0.1320	0.448
Retire	0.0858	1.0896	0.1116	0.442	0.2296	1.2581	0.1670	0.169	0.2307	1.2595	0.1667	0.166
Smoking												
Non-Smoker	**Ref**											
Smoker	-0.1138	.8924	0.1404	0.417	-0.1716	.8423	0.2051	0.408	-0.1155	.8909	0.2050	0.415
Drinking Status												
Non-Drunker	**Ref**											
Drunker	0.0977	1.1026	0.0542	0.070	-0.3564	.7002	0.0854	0.000	-0.3573	.6996	0.0854	0.000
Exercise												
Regular	**Ref**											
Sometimes	-0.0680	.9343	0.1170	0.561	-0.1173	.8893	0.1995	0.557	-0.1155	.8909	0.1994	0.563
Not at all	-0.0345	.9661	0.0613	0.574	-0.0470	.9541	0.08202	0.567	-0.0452	.9558	0.0819	0.582
Salt Intake												
Not at all	**Ref**											
As in food	-0.0401	.9607	0.0715	0.575	0.0227	1.0230	0.1093	0.835	0.0229	1.0232	0.1994	0.834
At table	0.1441	1.1550	0.09031	0.111	-0.1439	.8660	0.1271	0.235	-0.1441	.8658	0.0820	0.256
Age												
Below 50	**Ref**											
50 – 59	-0.2135	.8078	0.0721	0.004	0.0358	1.0364	0.0912	0.694	0.0362	1.0369	0.0911	0.691
60 – 69	-0.1097	.8961	0.0873	0.209	-0.0617	.9402	0.1135	0.587	-0.0617	.9402	0.1134	0.586
Above 70	-0.1993	.8193	0.1041	0.055	-0.0662	.9359	0.1421	0.641	-0.0651	.9370	0.1420	0.647
BMI												
Norma weight	**Ref**											
Under weight	0.02884	1.0293	0.0788	0.714	-0.1634	.8493	0.0997	0.101	-0.1644	.8484	0.0996	0.049
Overweight	-0.1127	.8934	0.06084	0.064	-0.0413	.9595	0.0852	0.628	-0.0418	.9591	0.0852	0.628
Obsessed	-0.1305	.8777	0.1033	0.206	0.4501	1.5685	0.2317	0.052	-0.4496	0.6379	0.2314	0.032
SBP												
Low	**Ref**											

Normal	0.0021	1.0021	0.07262	0.977	0.0618	1.0637	0.1056	0.558	0.0605	1.0624	0.1055	0.566
Pre-hypertension	0.1370	1.1468	0.0800	0.087	-0.0434	.9575	0.1093	0.692	-0.0446	.9564	0.1092	0.684
High	0.0435	1.0445	0.0875	0.076	0.0423	1.0432	0.2011	0.165	-0.0425	.9584	0.2016	0.921
DBP												
Low	**Ref**											
Normal	-0.1630	.8496	0.1446	0.260	9.8822	19578.7486	5.0770	0.998	0.4674	1.5958	0.5261	0.998
Pre-hypertension	-0.0325	.9680	0.0626	0.603	0.07765	1.0807	0.08934	0.385	0.07791	1.0810	0.0893	0.383
High	-0.3894	.6775	0.3726	0.296	0.2033	1.2254	0.4014	0.618	0.2041	1.2264	0.4010	0.611
μ	2.0387		0.06916		2.0393				2.0409			
ρ	0.4905		0.01664		0.4904				0.4900			
θ					0.0524		0.0552		0.0543		0.0574	
τ					0.0255				0.0264			
					Likelihood ratio (θ): x^2 = 3.89; prob. =0.024				Likelihood ratio (θ): x^2 = 4.08; prob. = 0.022			

distributions and random effect parameter (θ) were presented in table 3. The Weibull with inverse Gaussian random effect distribution shows that area of residence, level of education, drinking status and BMI are statistically significant ($p < 0.05$) risk factors for diabetes patient. Whereas the gender, occupation, smoking status, exercise, salt intake, age, SBP and DBP were found to be statistically insignificant factor for diabetes patient. The diabetes patient that reside in peri-urban and rural area had lesser survival time by a factor 0.6375 and 0.7790 respectively than those that reside in urban area (TR < 1). The result of the analysis also showed that diabetes patient with primary and secondary level of education had lesser survival time by a factor 0.8300 and 0.5904 as compared with those that had tertiary level of education (TR < 1). Furthermore, it was discovered that diabetes patient with drinking status had lesser survival time by a factor 0.6996 (TR < 1) as compared to those that do not drink. In addition, diabetes patient with underweight and obese had lesser survival time by a factor 0.8484 and 0.6379 respectively (TR < 1) as compared to those with normal weight. This implies that BMI, area of residence, level of education and drinking status are the risk factors of diabetes. In addition, the result of the selected model revealed that BMI, area of residence, level of education and drinking status are the risk factors of diabetes.

Checking model adequacy of parametric baselines using Cox-Snell residuals plots

The Cox–Snell residuals are one way to investigate how well the model fits the data. The plots for fitted models of residuals for the selected models Weibull AFT with Inverse Gaussian Random effect for Diabetes data set via maximum likelihood estimation with cumulative hazard functions are given in figure 1. If the model fits the data, the plot of cumulative hazard function of residuals against Cox–Snell residuals should be approximately follow a straight line. The plot for both models makes straight lines through the origin suggesting that the selected models are appropriate for time-to-Event of Diabetes Data set respectively.

CONCLUSION

This study aimed at comparing the performance of parametric and semi-parametric survival models with application to clinical data sets. Specifically, the study compared the performance of conventional semi-parametric model with extended semi-parametric model (CPHM with random effect) and, conventional AFT models with the extended AFT models (with random effect). Finally, the performance of semi-parametric and parametric model with and without random effect were compared. Based on the results of the analysis as presented in the previous section, it was concluded that the Weibull AFT model with inverse Gaussian random effect distribution performed better than the other models considered in this study.

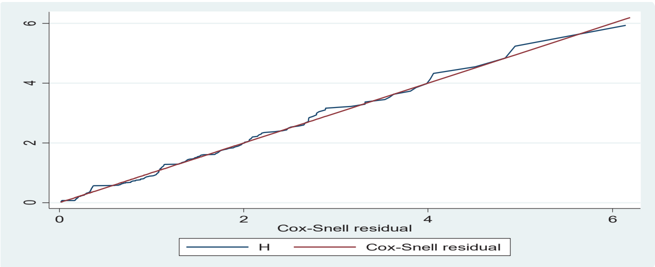

Figure 1:AFT Weibull model with inverse Gaussian RE distribution for diabetes data sets.

REFERENCES

1. Collett D. Modeling survival data in medical research. Boca Raton Florida: Chapman & Hall/CRC; 2003.
2. Xin J. Frailty models for the between center variation in survival following rectum cancer diagnosis. Masters Dissertation, Ghent University, Belgium. 2009.
3. Shankar PK, Screenivas V, Subrat KA. Comparison of cox proportional hazards model and lognormal accelerated failure time model: Application in time to event analysis of acute liver failure patients in India. Nepalese Journal of Statistics. 2019;3:21-40.
4. Cox DR. Regression models and life tables (with discussion). Journal of the Royal Statistical Society. 1972;187-220.
5. Cox DR. Partial likelihood. Biometrika. 1975;62:269-276.
6. Lahham HNM. Cardiovascular diseases and risk factors among diabetic patients. Nablus District, West Bank, Palestine. 2009.
7. WHO & AFRO. Cardiovascular diseases in African region: Current situation and perspectives report of the regional director. Fifty Fifth Sessions. Maputo, Mozambique. 2005.
8. Steyn K, Sliwa K, Hawken S, Commerford P, Damasceno A, Ounpuu S, Yusuf S. Risk factors associated with myocardial infarction in Africa. The inter heart Africa study. Circulation. 2005;112:3554-3561.

9. Kassa TH. Bayesian survival analysis of diabetes mellitus patients: A case study in tikur anbessa specialized hospital, Addis Ababa, Ethiopia. Journal of Reliability and Statistical Studies. 2018;37-56.

10. Lomo SI, Sugiyarto S, Darmawan E. A Survival analysis with cox regression interaction model of type II diabetes mellitus in Indonesian. Journal Kedokteran dan Kesehatan. 2021;15(1):81-90.

11. Naim S, Mahmudah M. Survival time of diabetes mellitus patients with hemodialysis: A study using survival analysis. Acta Medica Iranica. 2022;60(2):115-119.

12. Uloko AE, Musa BM, Ramalan MA, Gezawa ID, Puepet FH, Uloko AT, Sada KB. Prevalence and risk factors for diabetes mellitus in Nigeria: A systematic review and meta-analysis. Diabetes Therapy. 2018;9(3):1307-1316.

13. Hordofa SB, Debelo O. Statistical analysis of the survival of patients with diabetes mellitus: A case study at Nekemte Hospital, Wollega, Ethiopia. American Journal of Biometric and Biostatistics. 2020;4(1):6-12.

14. Adedotun AF, Odusanya OA, Okagbue HI, Ogundile OP. Analysis of reported cases of diabetes disease in Nigeria: A survival analysis approach. International Journal of Sustainable Development and Planning. 2022;17(2):643-647.

15. Belay A, Derebew B, Abebaw S. Survival analysis on time-to-recovery of diabetic patients at Minlik Referral Hospital, Ethiopia: Retrospective cohort study. Research square. 2021;1-25.

16. Zhao Z, Huo L, Wang L, Wang L, Fu Z, Li Y, Wu X. Survival of Chinese people with type 2 diabetes and diabetic kidney disease: A cohort of 12-year follow-up. BMC Public Health. 2019;19(1):1-8.

17. Badmus NI, Olanrewaju F, Adeniran AT. Modeling COVID-19 pandemic data with beta double exponential model. Asian journal of research in infectious diseases. 2020;5(6):66-79.

18. Adeniran AT, Faweya O, Ogunlade TO, Balogun KO. Derivation of Gaussian probability distribution: A new approach. Applied Mathematics. 2020;11(6):436.

19. Gutierrez RG. Parametric frailty and shared frailty survival models. Stata Journal. 2002;2(1):22-44.