An Application of Gamma Generalized Linear Model for Estimation of Survival Function of Diabetic Nephropathy Patients

Gurprit Grover, Alka Sabharwal* and Juhi Mittal

Department of Statistics, University of Delhi, Delhi, India

Abstract: Diabetic nephropathy (DN) is a generic term referring to deleterious effect on renal structure and/or function caused by diabetes mellitus. World Health Organization estimates that diabetes affects more than 170 million people worldwide and this number may rise to 370 million by 2030. The rate of rise in Serum Creatinine (SrCr) is a well-accepted marker for the progression of Diabetic Nephropathy (DN). In this paper, survival functions of type 2 diabetic patients with renal complication are estimated. Firstly, most appropriate distribution for duration of diabetes is selected through minimum Akaike Information Criterion value, Gamma distribution is found to be an appropriate model. Secondly, the parameters estimates of the selected distribution are obtained by fitting a Generalized Linear Model (GLM), with duration of diabetes as the response variable and predictors as SrCr and number of successes (number of times SrCr values exceed its normal range (1.4 mg/dl)). These covariates are linked with the response variable using two different link functions namely log and reciprocal links. Using the estimates of parameters obtained from generalized linear regression analysis, survival functions for different durations under both the links are estimated. Further we compared the estimated survival functions under both the links with Kaplan Meier (KM) estimates graphically. Findings suggested that the Kaplan Meier estimate and Gamma distribution under both links provided a close estimate of survival functions. Median survival time is 16.3 years and 16.8 years obtained from KM method and Gamma GLM respectively.

Keywords: Akaike Information Criterion, Gamma distribution, generalized linear models, Kaplan Meier method, log link, reciprocal link, serum creatinine, survival distributions.

1. INTRODUCTION

Type 2 diabetes mellitus (DM) is characterized by abnormal glucose-mediated insulin secretion and impaired insulin action in the liver and peripheral tissue. It is recognized as a global health problem nowadays, and it has been projected that the number of type 2 diabetic patients will rise from an estimated 135 million in 1995 to 300 million in 2025 [1]. Moreover, the Asia Pacific region is considered to be on the average of an emerging diabetes epidemic [2]. The chronic hyperglycemia of diabetes is associated with long-term damage, dysfunction, and failure of various organs, especially eyes, kidneys, nerves, heart, and blood vessels.

Diabetic nephropathy (DN) is one of the major chronic complication of diabetes which is characterized by increased albumin urinary excretion and loss of renal function [3]. It is also, the commonest cause of end-stage renal disease in the U.K., accounting for 20% of all patients requiring renal replacement therapy, in patients with either type 1 or type 2 diabetes [4]. DN develops in 20-40% of patients within 10 to 15 years after the onset of diabetes [5]. With the development of kidney complications, Glomerular Filtration Rate (GFR) starts to fall and SrCr level starts to increase. Various studies have shown the importance of measurement of Estimated Glomerular Filtration Rate (eGFR) and SrCr level for predicting the development of DN. The rate of rise in SrCr, a well accepted marker for the progression of DN,(creatinine value 1.4 mg/dl to 3.0 mg/dl) is the indicator for impaired renal function [6] and the normal level of creatinine is 0.8 mg/dl to 1.4 mg/dl [7].

Survival analysis or time-to-event data analysis is predominately used in medical science, where the interest is in observing the duration of time until an event of interest (death, onset of disease etc.) occurs. One of the ways of analyzing these occurrences over a period of time is by observing the survivability patterns, also known as survival curves. A survival curve is a summary display of the pattern of survival rates over time, where survival rate is a statistical index that summarizes the probable frequency of specific outcomes for a group of patients at a particular point in time. These survival curves are very useful in comparing two or more survivability patterns. Several approaches have been proposed in literature for estimating the survival functions. Kaplan-Meier, Nelson Aalen and life table methods are some nonparametric techniques used for estimating survival functions and hazard rates of survival data. Another approach is a parametric method, under which the survival times are assumed to follow a specific mathematical distribution. Several survival distributions have been proposed and the identification of a suitable one is a crucial step. Akaike Information Criterion (AIC) is most widely known and used model selection tool among

*Address correspondence to this author at the Kirori Mal College, University of Delhi, Delhi-110007, India; Tel: +919311589805;
E-mail: sabharwal_alka@hotmail.com

E-ISSN: 1929-6029/13

statisticians. AIC was introduced by Hirotugu Akaike in his seminal 1973 paper “Information Theory and an Extension of the Maximum Likelihood Principle" [8]. He proposed a framework wherein both model estimation and selection could be simultaneously accomplished. The computation of AIC is remarkably simple and the model with lowest AIC value is selected. Once the distribution of survival times is chosen it may be used to estimate the survival function and hence can be used to plot survival curves [9].

Linear regression analysis is customarily used when we want to estimate the response variable on the basis of several predictors, with an assumption that response variables are normally distributed. Another approach is Generalized Linear Models (GLM), which are an extension of the linear modeling process. They extend the ideas of regression analysis to a wider class of problems involving the relationship between a response and one or more explanatory variables. Also, they can be used for a model which follows probability distribution other than the Normal distribution such as Poisson, Binomial, Gamma and others. The link function is used to model responses when a dependent variable is assumed to be nonlinearly related to the predictors. In conclusion, the GLM can be used to predict responses both for dependent variables with continuous and non-continuous distribution and for dependent variables which are non-linearly related to the predictors.

Hakulinen and Tenkanen [11] estimated the relative survival rates of lung cancer patients by assuming a binomial distribution and applying generalized linear model approach with log-log link. Karem [12] applied general and generalized linear models for determining which combination of effects allows for the optimal prediction of survival for lung cancer patients. They showed that a full effects generalized linear model outperforms the general linear model. They also illustrated that a selection of Poisson distribution and log link function leads to excellent assessment. Yuan, Hong and Shyr [13] also studied the survival patterns of lung cancer patients by applying Cox proportional hazard models. Akram, Ullah and Taj [14] investigated the survival pattern of cancer patients using the non-parametric and parametric modeling strategies. They applied Kaplan-Meier method and Weibull model based on Anderson-Darling test to the real life time data of cancer patients.

In the present study retrospective data from type 2 diabetic patients was collected as per American Diabetes Association (ADA) standards from the data base of Dr. Lal’s path lab through house to house survey. The dataset consists of a total of 132 patients records, including both DN (60 uncensored observations) and Non Diabetic Nephropathy (72 censored observations) patients. The main aim of this research is to study the survival curves of DN patients, by estimating and interpreting survivor functions from data of type 2 diabetic patients. To achieve this, we first obtained an appropriate distribution for the duration of diabetes by fitting six different survival distributions. On the basis of AIC criterion, Gamma distribution is found to be the most appropriate model for given data. Since survival functions i.e. probability of occurrence of an event of interest which is DN, can be estimated only for uncensored observations, we considered only DN patients and estimated the survival functions using GLM. For performing the generalized linear regression analysis the duration of diabetes is taken as the response variable and predictors as SrCr and number of successes (number of times SrCr values exceed its normal range (1.4 mg/dl)). These covariates are linked with the response variable using two different link functions namely log and reciprocal link, as the AIC values are approximately same. The parameters of Gamma distribution are then estimated by using results of GLM under both the links. And estimated parameters of above distribution are used for estimating the survival functions. We also compared these survival functions with the traditional Kaplan Meier estimates.

Although some work has been done in literature regarding the estimation of survival functions but to the best of our knowledge this is the first investigation about survivability pattern of nephropathy patients arising out of type 2 diabetes only using generalized linear models. These patterns helps in developing treatment comparison design and could be important pointers for the medical fraternity to guide the patients about likely outcome i.e. end-stage renal disease. The remainder of the paper is organized as follows. In section 2 developments of models is discussed. Section 3 applies the model to the dataset of type 2 diabetic patients and some concluding remarks are made under section 4.

2. METHODOLOGY

The study consists of n type 2 diabetic patients who are divided into two groups namely DN and Non Diabetic Nephropathy (NDN) with r and $n-r$ patients respectively. The patients who do not develop DN during the course of study are censored observations.

2.1. Model Selection for the Duration of Diabetes

When fitting a fully parametric model, the survival times are assumed to follow a statistical distribution. Several different distributions have been proposed, and the identification of a suitable one is a crucial step. Here we apply six different survival distributions namely two-parameter exponential, Gamma, Weibull, lognormal, inverse Gaussian and Rayleigh distributions. A special feature of all these distributions is that they all belong to exponential family of distribution [15]. Weibull distribution is an important model used in medicine since it is a flexible distribution that allows a monotonous increasing and decreasing hazard rate. Gamma distribution, for its compliance with the patient data, is a suitable distribution to use in survival analysis. Exponential distribution which is a specific version of Gamma distribution is also used for survival analysis; it has been used by Lee and Go [16] to model the cancer survival data. The positively skewed distributions where the average values are low, variances are high and the values are not negative, generally accord with lognormal distribution. Generally, positively skewed data plays an important role in modeling medical data. Both Weibull and lognormal distributions are used to analyze positively skewed data. The inverse Gaussian distribution also provides much flexibility in modeling, when early occurrences of failures are dominant in a life time distribution and its failure rate is expected to be non-monotonic. It is almost an increasing failure rate distribution when it is slightly skewed and hence is also applicable to describe lifetime distribution which is not dominated by early failures. Gamma and lognormal distributions are preferable when hazard rises to a peak before decreasing. Rayleigh distribution which is a special case of Weibull distribution (when shape parameter is 2), is widely used to model events that occur in different fields such as medicine, social and natural science. Because of the similar and overlapping features of these distributions, an appropriate model selection is a vital step of any survival study [10]. The probability density function, survival function and the respective likelihood function of aforementioned distributions are defined as follows:

Two Parameter Exponential Distribution

$$\mathrm{f}(\mathrm{t}\,|\,\theta,\sigma)=\frac{1}{\sigma}\exp\left(\frac{-\left(\mathrm{t}-\theta\right)}{\sigma}\right);\ \sigma>0\ \&\ \mathrm{t}\geq\theta \tag{2.1}$$

$$\mathrm{S}(\mathrm{t})=\exp\left(\frac{-\left(\mathrm{t}-\theta\right)}{\sigma}\right) \tag{2.2}$$

When the censoring times T_i and onset time t_i are different for each patients, the likelihood function for the n survival times is defines as [9],

$$\mathrm{L}=\prod_{\mathrm{i}=1}^{\mathrm{n}}\Pr\left(\mathrm{t}_{\mathrm{i}},\delta_{\mathrm{i}}\right)=\prod_{\mathrm{i}=1}^{\mathrm{n}}\left[\mathrm{f}\left(\mathrm{t}_{\mathrm{i}}\right)\right]^{\delta_{\mathrm{i}}}\left[\mathrm{S}\left(\mathrm{T}_{\mathrm{i}}\right)\right]^{1-\delta_{\mathrm{i}}} \tag{2.3}$$

where δ indicates whether duration time t is uncensored $\left(\delta=1\right)$ or censored $\left(\delta=0\right)$. Then using equations 2.1 and 2.2 the likelihood function for the two parameter exponential distribution can be written as,

$$\mathrm{L}=\prod_{\mathrm{i}=1}^{\mathrm{n}}\left[\frac{1}{\sigma}\exp\left(\frac{-\left(\mathrm{t}_{\mathrm{i}}-\theta\right)}{\sigma}\right)\right]^{\delta_{\mathrm{i}}}\left[\exp\left(\frac{-\left(\mathrm{T}_{\mathrm{i}}-\theta\right)}{\sigma}\right)\right]^{1-\delta_{\mathrm{i}}} \tag{2.4}$$

$$\mathrm{L}=\frac{1}{\sigma^{\mathrm{r}}}\exp\left(-\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\frac{\delta_{\mathrm{i}}\left(\mathrm{t}_{\mathrm{i}}-\theta\right)}{\sigma}+\frac{\left(1-\delta_{\mathrm{i}}\right)\left(\mathrm{T}_{\mathrm{i}}-\theta\right)}{\sigma}\right)\right)$$

where r is the number of uncensored observations.

Gamma Distribution

$$f\left(t\,|\,\lambda,\gamma\right)=\frac{\lambda^{\gamma}}{\left[\gamma\right.}t^{\gamma-1}e^{-\lambda t};\ \lambda>0,\gamma>0\ \&\ \mathrm{t}>0 \tag{2.5}$$

$$\mathrm{S}(\mathrm{t})=\left[1-\mathrm{I}\left(\lambda\mathrm{t},\gamma\right)\right] \tag{2.6}$$

where $\mathrm{I}\left(\lambda\mathrm{t},\gamma\right)$ is the incomplete Gamma function defined as,

$$\mathrm{I}\left(\lambda\mathrm{t},\gamma\right)=\frac{1}{\left[\gamma\right.}\int_{0}^{\lambda\mathrm{t}}\mathrm{u}^{\gamma-1}\mathrm{e}^{-\mathrm{u}}\mathrm{du} \tag{2.7}$$

$$\mathrm{L}=\frac{\lambda^{\gamma\mathrm{r}}}{\left(\left[\gamma\right.\right)^{\mathrm{r}}}\prod_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathrm{t}_{\mathrm{i}}^{\gamma-1}\right)^{\delta_{\mathrm{i}}}\exp\left(-\lambda\sum_{\mathrm{i}=1}^{\mathrm{n}}\delta_{\mathrm{i}}\mathrm{t}_{\mathrm{i}}\right)\prod_{\mathrm{i}=1}^{\mathrm{n}}\left[1-\mathrm{I}\left(\lambda\mathrm{T}_{\mathrm{i}},\gamma\right)\right]^{\left(1-\delta_{\mathrm{i}}\right)} \tag{2.8}$$

Weibull Distribution

$$f\left(t\,|\,\lambda,\gamma\right)=\frac{\gamma}{\lambda}\left(\frac{t}{\lambda}\right)^{\gamma-1}e^{-\left(t/\lambda\right)^{\gamma}};\ \lambda>0,\gamma>0\ \&\ \mathrm{t}>0$$

$$\mathrm{S}(\mathrm{t})=\exp\left(-\,\mathrm{t}/\lambda\right)^{\gamma}$$

$$\mathrm{L}=\frac{\gamma^{\mathrm{r}}}{\lambda^{\gamma\mathrm{r}}}\prod_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathrm{t}_{\mathrm{i}}^{\gamma-1}\right)^{\delta_{\mathrm{i}}}\exp\left(-\left[\sum_{\mathrm{i}=1}^{\mathrm{n}}\delta_{\mathrm{i}}\left(\mathrm{t}_{\mathrm{i}}/\lambda\right)^{\gamma}+\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(1-\delta_{\mathrm{i}}\right)\left(\mathrm{T}_{\mathrm{i}}/\lambda\right)^{\gamma}\right]\right)$$

Lognormal Distribution

$$\mathrm{f}\left(\mathrm{t}\,|\,\mu,\sigma\right)=\frac{1}{\mathrm{t}\sigma\sqrt{2\pi}}\exp\left[\frac{-1}{2\sigma^{2}}\left(\log\mathrm{t}-\mu\right)^{2}\right];$$

$$-\infty<\mu<\infty,\sigma>0,\mathrm{t}\geq0$$

$$S(t)=\frac{1}{\sigma\sqrt{2\pi}}\int_{t}^{\infty}\frac{1}{x}\exp\left[\frac{-1}{2\sigma^2}(\log x-\mu)^2\right]=$$
$$1-\Phi\left[\log\left(\frac{t\exp(-\mu)}{\sigma}\right)\right]$$

where $\Phi(.)$ is cumulative distribution function of standard normal variable.

$$L=\frac{1}{\left(\sigma\sqrt{2\pi}\right)^{r}\prod_{i=1}^{n}(t_i)^{\delta_i}}\exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n}\delta_i(\log t_i-\mu)^2\right]$$
$$\prod_{i=1}^{n}\left[1-\Phi\left[\log\left(\frac{T_i\exp(-\mu)}{\sigma}\right)\right]\right]^{(1-\delta_i)}$$

Inverse Gaussian Distribution

$$f(t\mid\mu,\lambda)=\sqrt{\frac{\lambda}{2\pi t^3}}\exp\left[\frac{-\lambda(t-\mu)^2}{2\mu^2 t}\right];\ t>0,\mu>0\ \&\ \lambda>0$$

$$S(t)=1-\sqrt{\frac{\lambda}{2\pi}}\int_{0}^{t}\left(\frac{1}{x}\right)^{3/2}\exp\left[\frac{-\lambda x}{2\mu^2}\left(1-\frac{\mu}{x}\right)^2\right]dx$$

$$L=\left(\frac{\lambda}{2\pi}\right)^{r/2}\frac{1}{\prod_{i=1}^{n}\left(\sqrt{t_i^3}\right)^{\delta_i}}\exp\left[\frac{-\lambda\sum_{i=1}^{n}\delta_i(t_i-\mu)^2}{2\mu^2 t_i}\right]\prod_{i=1}^{n}\left[S(T_i)\right]^{(1-\delta_i)}$$

Rayleigh Distribution

$$f(t\mid\sigma)=\frac{t}{\sigma^2}\exp\left[\frac{-t^2}{2\sigma^2}\right];\ t\geq 0\ \&\ \sigma>0$$

$$S(t)=\exp\left[\frac{-t^2}{2\sigma^2}\right]$$

$$L=\prod_{i=1}^{n}t_i^{\delta_i}\frac{1}{\sigma^{2r}}\exp\left[-\left(\sum_{i=1}^{n}\frac{\delta_i t_i^2}{2\sigma^2}+\sum_{i=1}^{n}\frac{(1-\delta_i)T_i^2}{2\sigma^2}\right)\right]$$

Maximum likelihood estimates (MLE) of the parameters are obtained by taking log of likelihood function and solving the likelihood equations simultaneously (refer Appendix).

Akaike Information Criterion for Model Selection

Model selection or the distribution assumption made about the response variable is a topic of special relevance in medical studies and can have a critical impact on the conclusions drawn. Regardless of which type of model is fitted and how the variables are selected to be in the model, it is important to evaluate how well the model represents the data, since the use of an inappropriate statistical model may give rise to misleading conclusions. A survival model is adequate if it represents the survival patterns in the data to an acceptable degree. This aspect of a model is known as goodness of fit. Akaike Information Criterion provides an attractive basis for model selection and is defined as,

$$AIC=-2\log(L)+2(p+c) \tag{2.9}$$

The term $-2\log(L)$ is well known among statistician as the "deviance", p and c is the number of parameters and covariates respectively in the model. The model with the smallest AIC value is preferred i.e., if the AIC is smaller for the first model than the second then the former is preferred [17-19].

2.2. Generalized Linear Model

After the selection of an appropriate model a GLM is fitted to duration of diabetes of DN patients to estimate the survival function. The GLM was developed to fit regression models for univariate response data that follows a very general distribution called the exponential family. The Binomial, Poisson, normal, Gamma, inverse Gaussian distributions are some members of this family. All the members of exponential family of distribution can be defined by the following probability density function,

$$f(t;\theta,\phi)=\exp\left(\frac{t\theta-b(\theta)}{a(\phi)}+c(t,\phi)\right) \tag{2.10}$$

where $a(.)$, $b(.)$ and $c(.)$ are specific functions. The parameter θ is the natural location parameter, and ϕ is often called a dispersion parameter. The function $a(\phi)$ is generally of the form $a(\phi)=\phi.\omega$ where ω is known constant. If $t_i;\ i=1,2,...,n$ represents the response values, then the GLM is given by,

$$g(\mu_i)=g\left[E(t_i)\right]=\underset{\sim}{X}_i'\underset{\sim}{\beta}\ ;i=1,2,...,n \tag{2.11}$$

where $\underset{\sim}{X}_i$ is a vector of regression variables or covariates for the i^{th} observation and $\underset{\sim}{\beta}$ is the vector of regression coefficients. Every GLM has three components: a response variable distribution, a linear predictor $\eta=X'\underset{\sim}{\beta}$ that involves the regression

variables and a link function $\eta_i = g(\mu_i)$ that connects the linear predictor to the natural mean of the response variable. The most commonly used link functions are reciprocal, log, logistic, identity and power link [20].

2.3. Gamma Generalized Linear Model for Estimating the Survival Function

In this section, Gamma GLM is applied to estimate the survival function by a Gamma response variable under the influence of the predictors SrCr and the number of successes (which is defined as the number of times SrCr is greater than 1.4mg/dl). The probability density function of the response variable is same as defined in equation 2.5. Various link functions are commonly used, depending on the assumed distribution of dependent variable. Log and reciprocal links are often used for Gamma distribution. These links can be defined as,

Link	$\eta_i = g(\mu_i)$
Log	$\eta_i = \log(\mu_i)$
Reciprocal	$\eta_i = 1/\mu_i$

The choice of correct link function is important for inference and predictions. To access the quality of the link function AIC values are compared and the model with the least AIC values is considered to be the appropriate model [21]. Once the parameters are estimated the survival function can be estimated using equations equation 2.6 defined in section 2.1.

3. APPLICATION

A retrospective study of 132 type 2 diabetic patients, who were diagnosed as diabetic as per ADA standards from the data base of Dr. Lal's path lab (a reputed NABL certified path lab), is conducted. Up-to-date pathological reports of the patients were collected through house to house survey. The data regarding the duration of diabetes and other factors like age at which diabetes was diagnosed; Fasting Blood Glucose (FBG), Diastolic Blood Pressure (DBP), Systolic Blood Pressure (SBP), Low Density Lipoprotein (LDL) and values of SrCr are recorded for each patient. Under this section the model defined before is applied on the last available record of every variable corresponding to each patient to assess the latest renal health of type 2 diabetic patient. The rate of rise in the value of SrCr is an important marker for prediction of DN. Thus using the values of SrCr the data has been classified into two categories namely DN (SrCr ≥ 1.4mg/dl) and NDN (SrCr < 1.4mg/dl) groups and it was found at the end of study that out of 132 patients there are only 60 (45.45 %) DN cases and 72 (54.55%) are NDN cases. Using results of Grover, Gadpayle and Sabharwal [22] we have also considered an additional factor namely number of successes s, which measures the number of time SrCr value was recorded greater than 1.4 mg/dl out of total number of times test is recommended for a particular patient, as a factor for predicting the progression of DN in type 2 diabetic patients.

3.1. Model Selection for Duration of Diabetes

AIC, a statistic that trades off model's likelihood against its complexity is used to compare the viability of different parametric models. We found that Gamma model based on 132 patients (including censored observations) has higher likelihood than the other models and lower AIC value of 89.3945, indicating that this distribution is most accurate for the duration of diabetes. The AIC values with the MLE of parameters are presented in Table **1**.

Table 1: Akaike Information Criterion (AIC) of Six Different Distributions

Model	parameters	MLE	AIC values
Two Parameter Exponential	(θ,σ)	(5.6463, 17.30589)	259.1541
Gamma	(γ,λ)	(4.5330, 0.3470)	89.3945
Weibull	(γ,λ)	(2.3420, 14.796)	208.9497
Lognormal	(μ,σ)	(1.0664, 0.485)	191.689
Inverse Gaussian	(μ,λ)	(13.0631, 59.2175)	442.7841
Rayleigh	(σ)	(15.5780)	116.4871

3.2. Gamma Generalized Linear Model with Log Link Applied on Duration of Diabetes to Estimate the Survival Functions

Survival functions can be estimated only for the uncensored observations (i.e. the patients who develop DN during the period of study), we have re-estimated the parameters of Gamma distribution by fitting GLM to the data of duration of diabetes of DN patients only. Firstly, Gamma GLM model with log link is applied to estimate the duration of diabetes (response variable) for each patient, depending on two independent predictors SrCr and number of successes. The GLM model used to estimate the survival function is defined in the following equation,

$$\text{Log}\left(\text{Dur}\right)_i = \beta_0 + \beta_1(\text{SrCr})_i + \beta_2(\text{No.Succ})_i \qquad (3.1)$$

Where,

Dur: duration of diabetes

SrCr: serum creatinine

No.Succ: number of time SrCr exceeds its normal range.

Wald Chi-square statistics shows that for estimating the duration of disease, the predictors SrCr and number of successes are significant as p-values are less than 0.0100. We also failed to reject the hypothesis that Gamma GLM model fits the data well, as the p-value is greater than 0.0500. The detailed results of Gamma GLM are presented in Table **2**. The fitted Gamma GLM model with log link is as follows,

$$\text{Log}\left(\text{Dur}\right)_i = 3.1710 - 0.3520(\text{SrCr})_i + 0.0356(\text{No.Succ})_i \qquad (3.2)$$

The GLM model also gives the estimate of dispersion parameter which is the reciprocal of the shape parameter. Hence, an estimate of shape parameter γ of the Gamma distribution is 17.2414 and using the mean of fitted values of response variable the scale parameter λ is 1.0111. Using these parameter estimates survival functions for different time periods are estimated by applying equation 2.6 and are presented in Table **4**. A graphical comparison of survival function with Kaplan Meier method is presented in Figure **1**. Figure **1** supports the claim that a Gamma distribution is an effective description to model the data of diabetic patients with renal complication. It also shows that the median survival times obtained from Kaplan Meier method and Gamma GLM is 16.3 years and 16.8 years respectively.

3.3. Gamma Generalized Linear Model with Reciprocal Link Applied on Duration of Diabetes to Estimate the Survival Functions

The assessment of different link functions in GLM is performed on the basis of AIC values of the fitted model. Since, AIC values of Gamma GLM models under reciprocal links is 324.1400 which is very close to the one obtained under log link, we also estimated survival functions using reciprocal links. Taking the same variables and performing similar steps as done

Table 2: Gamma Generalized Linear Regression Model for Estimating the Duration of Diabetes Based on Serum Creatinine (SrCr) and Number of Successes (No.Succ) with Log Link

LINK	**Log**					
Model	$\text{Log}\left(\text{Dur}\right)_i = \beta_0 + \beta_1(\text{SrCr})_i + \beta_2(\text{No.Succ})_i$					
			95% Confidence Interval		**Wald Chi-sq for hypothesis parameter =0**	
Variable	**Parameter estimates**	**Standard Errors**	**Lower**	**Upper**		**P-value**
Intercept	3.1710	0.1931	2.7930	3.550	269.7000	0.0000
SrCr	-0.3520	0.1350	-0.6170	-0.0870	6.7590	0.0090
No.Succ	0.0356	0.0100	0.0160	0.0570	12.4910	0.0000
Dispersion	0.0580	0.0110	0.0400	0.0840		
AIC Value	323.1220					
Fitted Model	$\text{Log}\left(\text{Dur}\right)_i = 3.1710 - 0.3520(\text{SrCr})_i + 0.0356(\text{No.Succ})_i$					

Degrees of freedom for each predictor is 1.

Model with no intercept can be defined as $\text{Log}\left(\text{Dur}\right)_i = \beta_1(\text{SrCr})_i + \beta_2(\text{No.Succ})_i$

for log link, the Gamma GLM model under this link is defined as,

$$\frac{1}{(\text{Dur})_i} = \beta_0 + \beta_1(\text{SrCr})_i + \beta_2(\text{No.Succ})_i \qquad (3.3)$$

Where,

Dur: duration of diabetes

SrCr: serum creatinine

No.Succ: number of time SrCr exceeds its normal range.

Wald Chi-square statistics shows that for estimating the duration of disease, the predictors SrCr and number of successes are significant as p-values are less than 0.0500. We also failed to reject the hypothesis that Gamma GLM model fits the data well, as the p-value is greater than 0.0500. The detailed results of analysis are presented in Table **3**. The fitted Gamma GLM under reciprocal link is defined as,

Table 3: Gamma Generalized Linear Regression Model for Estimating the Duration of Diabetes Based on Serum Creatinine (SrCr) and Number of Successes (No.Succ) with Reciprocal Link

LINK	Reciprocal					
Model	$\frac{1}{(\text{Dur})_i} = \beta_0 + \beta_1(\text{SrCr})_i + \beta_2(\text{No.Succ})_i$					
Variable	Parameter estimates	Standard Errors	95% Confidence Interval		Wald Chi-sq for hypothesis parameter =0	P-value
			Lower	Upper		
Intercept	0.0410	0.0119	0.0170	0.0640	11.7010	0.0010
SrCr	0.0190	0.0082	0.0030	0.0350	5.4360	0.0200
No.Succ	-0.0020	0.0006	-0.0030	-0.0010	11.2500	0.0010
Dispersion	0.0590	0.0111	0.0410	0.0860		
AIC Value	324.1400					
Fitted Model	$\frac{1}{(\text{Dur})_i} = 0.0410 + 0.0190(\text{SrCr})_i - 0.0020(\text{No.Succ})_i$					

Degrees of freedom for each predictor is 1.

Model with no intercept can be defined as $\frac{1}{(\text{Dur})_i} = \beta_1(\text{SrCr})_i + \beta_2(\text{No.Succ})_i$

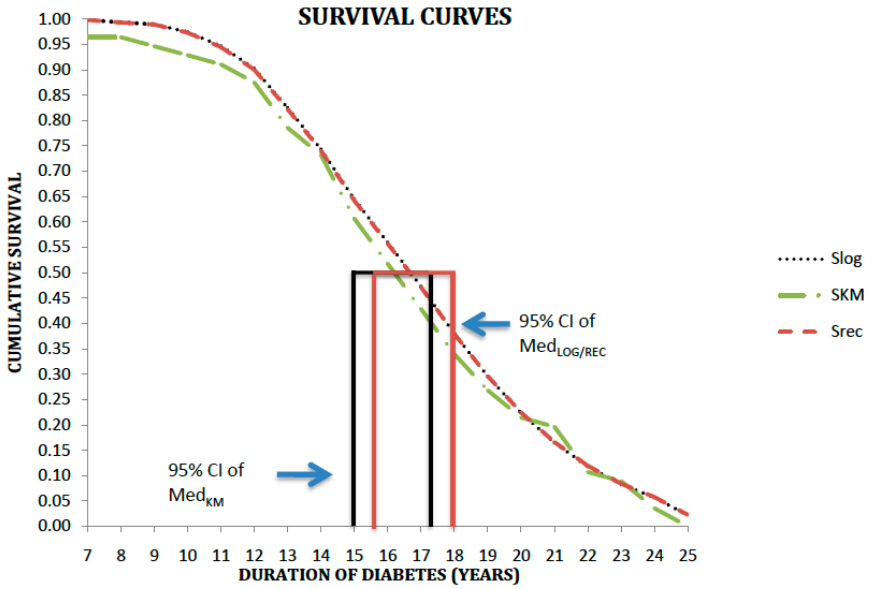

Figure 1: Survival Curves plotted using Gamma Generalized Linear Model (reciprocal and log links) and Kaplan Meier Method.

$$\frac{1}{(\text{Dur})_i} = 0.0410 + 0.0190(\text{SrCr})_i - 0.0020(\text{No.Succ})_i \qquad (3.4)$$

Estimate of shape and scale parameter under this model are obtained as 16.9491 and 0.9945 respectively. Using these estimates and equation (2.6) survival functions are estimated and are shown in Table **4**. Figure **1** graphically displays the estimated survival distribution of Kaplan Meier compared with Gamma distribution. Figure supports the claim that Gamma distribution is an effective description to model the data of concern of our study. It also shows that the median survival times obtained from Gamma GLM under reciprocal link is 16.8 years. Figure **2** displays a brief algorithm of the above procedure.

4. DISCUSSION

Type 2 diabetes accounts for more than 90% of all diabetic cases and is the main driver of diabetes epidemic [23]. DN is one of the serious, progressive, complications associated with diabetes and is the leading cause of End Stage Renal Disease (ESRD) in the United States [24]. Diabetic nephropathy is the leading cause of kidney disease in patients starting renal replacement therapy and affects approximately 40% of type 1 and type 2 diabetic patients [25]. The complexity of renal complication, challenges the research community to apply technological advances to develop prevention treatment programs which will lighten the burden of diabetes. The present study focused on estimating the survival functions of type 2 diabetic patients who develop renal complication.

The primary objective of our study was to determine an appropriate distribution for the duration of diabetes. AIC is generally used for the identification of an optimum model in a class of competing models. This criterion has been used here over the conventional Chi-square goodness of fit test as it overcomes its major disadvantage, i.e. while using chi-square test more than one distribution may fit well to the data, hence making the final model selection difficult. Gamma distribution was found to be an appropriate model for the duration of diabetes of type 2 diabetic patients and is widely used and plays an important role in reliability field and survival analysis.

In our study, patients with same survival time have different renal health status. The renal health status of

Table 4: Survival Functions by using Gamma Generalized Linear Model under Log and Reciprocal Links and Kaplan Meier Method

Duration of disease	$S_{Log}(t)$	$S_{Rec}(t)$	$S_{KM}(t)$
$7 \le t_i < 8$	0.9983	0.9981	0.9643
$9 \le t_i < 10$	0.9897	0.9890	0.9464
$10 \le t_i < 11$	0.9745	0.9733	0.9286
$11 \le t_i < 12$	0.9467	0.9448	0.9107
$12 \le t_i < 13$	0.9025	0.9000	0.8750
$13 \le t_i < 14$	0.8253	0.8223	0.7857
$14 \le t_i < 15$	0.7436	0.7407	0.7321
$15 \le t_i < 16$	0.6450	0.6425	0.6071
$16 \le t_i < 17$	0.5599	0.5580	0.5179
$17 \le t_i < 18$	0.4730	0.4718	0.4286
$18 \le t_i < 19$	0.3795	0.3791	0.3393
$19 \le t_i < 20$	0.2956	0.2959	0.2679
$20 \le t_i < 21$	0.2238	0.2246	0.2143
$21 \le t_i < 22$	0.1648	0.1660	0.1964
$22 \le t_i < 23$	0.1183	0.1196	0.1071
$23 \le t_i < 24$	0.0829	0.0841	0.0893
$24 \le t_i < 25$	0.0567	0.0578	0.0357
$t_i \ge 25$	0.0221	0.0228	0.0000

*No uncensored case in interval $8 \le t_i < 9$.

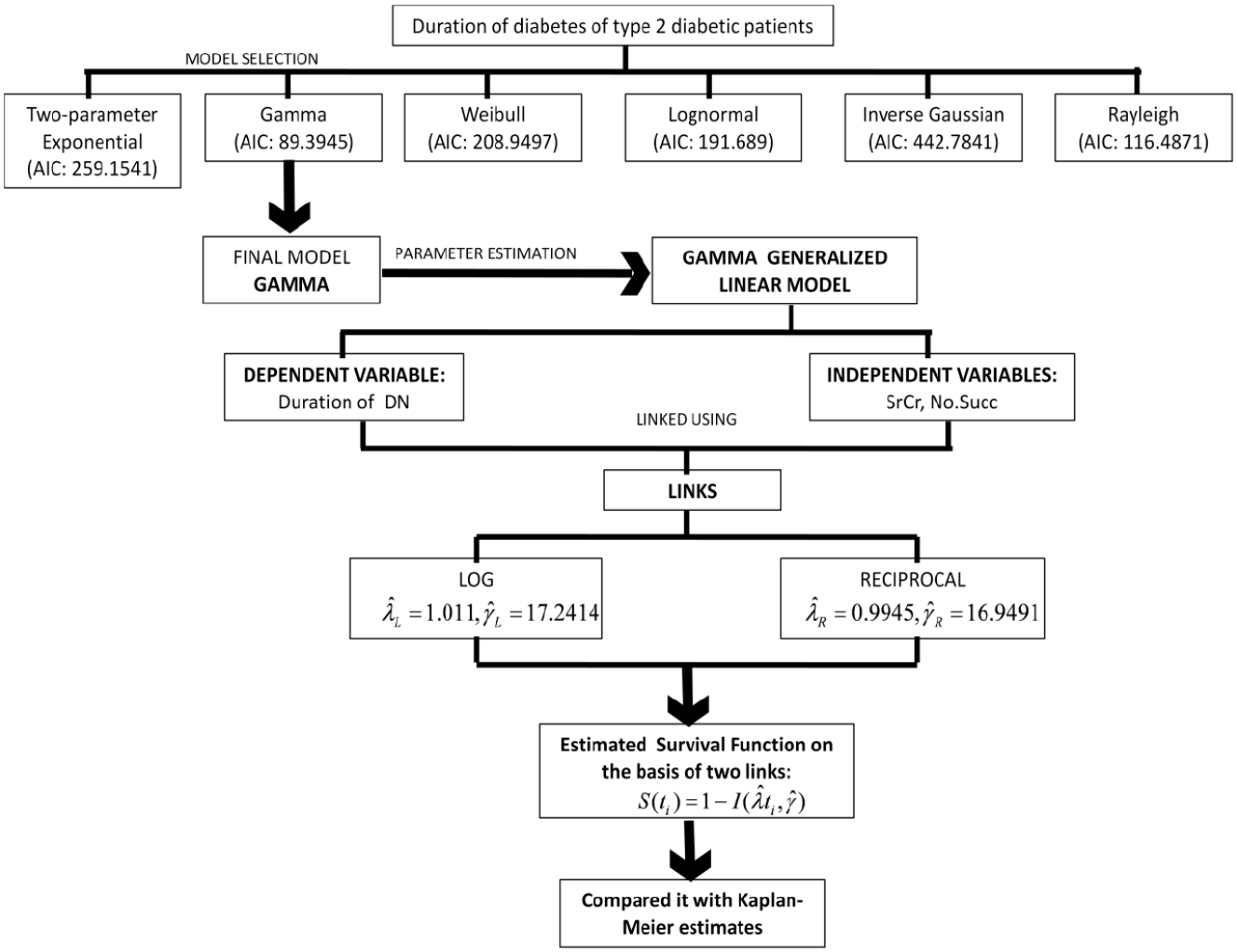

Figure 2: Algorithm for estimating the survival functions of type 2 diabetic patients with renal dysfunction.

a patient is determined on the basis of SrCr, as it is considered to be an important marker for the prediction of DN [7]. Grover, Gadpayle and Sabharwal [22] have shown that renal heath of a patient can also be predicted on the basis of the number of times the SrCr value exceeds its normal range (1.4 mg/dl). Therefore, SrCr and the number of times SrCr exceeds its normal range are the two most significant factors for the prediction of DN. Gamma GLM model is applied for estimating the mean duration of diabetes on the basis of these two covariates. GLM models are an extension of linear modeling process that enables us to analyze the dataset which follow probability distributions other than normal distribution, such as Poisson, Gamma, Binomial etc. A link function is used in GLM to link the random component of the model, the probability distribution of the response variable, to the systematic component of the model (predictors). Log and reciprocal link are commonly used for the Gamma distribution and the choice of link function is made on the basis of the minimum AIC value. It was found that the mean duration of diabetes of patients who develop DN under log and reciprocal link is 17.0522 and 16.9687 respectively. After the estimation of shape and scale parameters by applying these two links, the expression of survival function of Gamma distribution was used to compute the respective survival functions. A simple test for the model adequacy is to compare the overall (Kaplan- Meier) survival curve to the model-based predicted survival. Ideally, for any group of patients the two should be close, if not identical [19]. As Kaplan Meier curves provide a visual depiction of the raw data, the failure times (the “steps” down) and the censoring times (the vertical bars), they also provide a mathematical estimate of the underlying probability distribution. The survival functions obtained are then compared graphically with the traditional Kaplan Meier estimate. The estimates of the survival functions obtained under the two links and, by applying Kaplan Meier method are found to be approximately same.

The findings of this paper are consistent with the previous studies, which states that renal complication starts and approaches to advance stage in a diabetic patient in approximately 15 to 17 years [26]. Application of Gamma distribution under the same approach can be applied to estimate the onset time of other diabetic complication with appropriate covariates. Also, other distributions can be explored with GLM models to estimate the survival function of different diabetic complications. However, our study is confined to a certain region and therefore may not be a representative of the entire diabetic population of the country.

APPENDIX: MAXIMUM LIKELIHOOD ESTIMATE

1. Weibull Distribution

$$f\left(t\,|\,\lambda,\gamma\right)=\frac{\gamma}{\lambda}\left(\frac{t}{\lambda}\right)^{\gamma-1}e^{-\left(t/\lambda\right)^{\gamma}};\ \lambda>0,\gamma>0\ \ \&\ \ t>0$$

The log likelihood of Weibull distribution is,

$$\log L=r\log\gamma-r\gamma\log\lambda+(\gamma-1)\sum_{i=1}^{n}\delta_i\log t_i-\sum_{i=1}^{n}\delta_i\left(t_i/\lambda\right)^{\gamma}-\sum_{i=1}^{n}\left(1-\delta_i\right)\left(\frac{T_i}{\lambda}\right)^{\gamma}$$

The likelihood equations for λ and γ are,

$$\sum_{i=1}^{n}\delta_i t_i^{\gamma}+\sum_{i=1}^{n}\left(1-\delta_i\right)T_i-r\lambda^{\gamma}=0 \qquad \text{(A.1)}$$

$$\frac{r}{\gamma}-r\log\lambda+\sum_{i=1}^{n}\delta_i\log t_i-\sum_{i=1}^{n}\delta_i\left(t_i/\lambda\right)^{\gamma}\log\left(\frac{t_i}{\lambda}\right)-\sum_{i=1}^{n}\left(1-\delta_i\right)\left(T_i/\lambda\right)^{\gamma}\log\left(\frac{T_i}{\lambda}\right)=0 \qquad \text{(A.2)}$$

The MLE of λ and γ are obtained by solving the above two equations simultaneously, using Newton-Raphson method. Newton – Raphson method is one of the best iterative methods in numerical analysis because it's very fast and the error of this iterative method is quadratic approximation. An iterative procedure is a technique of successive approximations, and each approximation is called iteration. If the successive approximations approaches the solutions vary closely then the iterations converge. The Newton – Raphson method requires an initial value of each unknown parameters. The initial estimates of parameters are taken from graphical method and the iterative procedure is continued till any two successive iterations are approximately same up to 4 places of decimal.

2. Two-Parameter Exponential Distribution

$$\theta=t_{(1)}\ \&\ \sigma=\frac{\sum_{i=1}^{n}\delta_i\left(t_i-\theta\right)+\sum_{i=1}^{n}\left(1-\delta_i\right)\left(T_i-\theta\right)}{r};$$

where $t_{(1)}$ = minimum observed onset time (A.3)

3. Rayleigh Distribution

$$\sigma=\sqrt{\frac{\sum_{i=1}^{n}\delta_i t_i^2+\sum_{i=1}^{n}\left(1-\delta_i\right)T_i^2}{2r}} \qquad \text{(A.4)}$$

The MLE estimates of parameters for the other distribution i.e, Gamma, lognormal and inverse Gaussian are obtained using MATLAB software.

REFERENCES

[1] King H, Aubert RE, Herman WH. Global burden of diabetes, 1995-2025: prevalence, numerical estimates, and projections. Diabetes Care 1998; 21: 1414-31. http://dx.doi.org/10.2337/diacare.21.9.1414

[2] Amos AF, McCarty DJ, Zimmet P. The rising global burden of diabetes and its complications: estimates and projections to the year 2010. Diabet Med 1997; 14(5): S1-S85. doi:10.1002/(SICI)1096-9136(199712)14:5+<S7::AID-DIA522>3.0.CO;2-R.

[3] Campbell RC, Ruggenenti P, Remuzzi G. Proteinuria in diabetic nephropathy: treatment and evolution. Current Diabetic Report 2007; 3(6): 497-504. http://dx.doi.org/10.1007/s11892-003-0014-0

[4] Craig KJ, Donovan K, Munnery M, Owens DR, Williams JD, Phillips AO. Identification and management of diabetic nephropathy in the diabetes clinic. Diabetes Care 2003; 26(6): 1806-11. http://dx.doi.org/10.2337/diacare.26.6.1806

[5] Mazze, Strock, Simonson, Bergenstal. Staged-diabetes management: A systematic approach. 2nd ed. John Wiley & Sons; 2005

[6] Adler AI, Stevens RJ, Manley SE, Bilous RW, Cull CA, Holman RR. Development and Progression of nephropathy in type 2 diabetes: Observation and modeling from the United Kingdom Prospective Diabetes Study. Kidney Int 2003; 63: 225-32. http://dx.doi.org/10.1046/j.1523-1755.2003.00712.x

[7] Dabla PK. Renal function in diabetic nephropathy. World J Diabetes 2010; 1(2): 48-56. http://dx.doi.org/10.4239/wjd.v1.i2.48

[8] Akaike H. Information theory and an extension of the maximum likelihood principle. Proceeding of 2nd International Symposium on information theory. Petrov BN, Caski F, Eds., Akademia Kiado, Budapest 1973; pp. 267-81.

[9] Klein JP, Moeschberger ML. Survival analysis techniques for censored and truncated data. Springer Verlag, New York, USA; 2003.

[10] Hayat EA, Suner A, Uyar B, Dursan O, Orman MN, Kitapcioglu G. Comparison of five survival models: breast cancer registry data from Ege university cancer research center. Turkiye Klinikeri J Med Sci 2010; 30(5): 1665-74. http://dx.doi.org/10.5336/medsci.2009-16200

[11] Hakulinen T, Tenkanen L. Regression analysis of relative survival rates. Appl Statist 1987; 36: 309-317. http://dx.doi.org/10.2307/2347789

[12] Karmen A. Application of the generalized linear model to the prediction of lung cancer survival. 2006; 1-18. http://analytics.ncsu.edu/sesug/2006/ST09_06.PDF

[13] Yuan X, Hong D, Shyr Y. Survival model and estimation for lung cancer patients 2007; 201-22. http://capone.mtsu.edu/dhong/YuanHongShyr07.pdf

[14] Akram M, Ullah MA, Taj R. Survival analysis of cancer patients using parametric and non-parametric approaches. Pakistan Vet J 2007; 27(4): 194-98. http://www.pvj.com.pk/pdf-files/27_4/194-198.pdf

[15] Exponential families. http://www.math.siu.edu/olive/ich3.pdf

[16] Lee ET, Go OT. Survival analysis in public health research. Annu Rev Public Health 1997; 18: 105-134. http://dx.doi.org/10.1146/annurev.publhealth.18.1.105

[17] Lindsey JK, Jones B. Choosing among generalized linear models applied to medical data. Statist Med 1998; 17: 59-68. http://dx.doi.org/10.1002/(SICI)1097-0258(19980115)17:1<59::AID-SIM733>3.0.CO;2-7

[18] Acquah H de-Graft. Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship. J Develop Agric Econom 2010; 2(1): 1-6. http://www.academicjournals.org/JDAE

[19] Bradburn MJ, Clark TG, Love SB, Altman DG. Survival analysis part III: Multivariate data analysis-choosing a model and assessing its adequacy and fit. Br J Cancer 2003; 89: 605-11. http://dx.doi.org/10.1038/sj.bjc.6601120

[20] Myers RH, Montgomery DC, Vining GG. Generalized Linear Models: With Application in Engineering and the SciencesJohn Wiley and Sons, New York 2002.

[21] Huettmann F, Linke J. Assessment of different link functions for modeling binary data to derive sound inferences and predictions. Springer-Verlag Berlin Heidelberg 2003; 43-48. http://individual.utoronto.ca/julialinke/ICCSA2003_HuettmannandLinke.pdf.

[22] Grover G, Gadpayle AK, Sabharwal A. Identifying patients with diabetic nephropathy based on serum creatinine under zero truncated model. Electron J Appl Statist Anal 2010; 3(1): 28-43. http://siba-ese.unile.it/index.php/ejasa/article/view/i20705948v3n1p28/3029.

[23] Mohan V, Pradeepa R. Epidemiology of diabetes in different regions of India. Health Administrator 2009; 22(1,2): 1- 18.

[24] American Diabetes Association: Position statement: Diabetic nephropathy. Diabetes Care 1999; 22 (Supp 1): S66-69.

[25] Gross JL, Azevedo MJD, Silveiro SP, Canani LH, Caramori ML, Zelmanovitz T. Diabetic nephropathy: Diagnosis, prevention, and treatment. Diabetes Care 2005; 28: 176-88. http://dx.doi.org/10.2337/diacare.28.1.164

[26] Grover G, Sabharwal A, Mittal J. A Bayesian approach for estimating onset time of nephropathy for type 2 diabetic patients under various health conditions. IJSP 2013; 2(2): 89-101. http://dx.doi.org/10.5539/ijsp.v2n2p89

Received on 25-06-2013 Accepted on 16-07-2013 Published on 31-07-2013

http://dx.doi.org/10.6000/1929-6029.2013.02.03.6