

# A Bayesian Approach for Estimating Onset Time of Nephropathy for Type 2 Diabetic Patients Under Various Health Conditions

Gurprit Grover<sup>1</sup>, Alka Sabharwal<sup>1</sup> & Juhi Mittal<sup>1</sup>

<sup>1</sup> Department of Statistics, University of Delhi, Delhi, India

Correspondence: Alka Sabharwal, Kirori Mal College, University of Delhi, Delhi, India. Tel: 91-931-158-9805.

E-mail: sabharwal\_alka@hotmail.com

Received: March 10, 2013 Accepted: April 9, 2013 Online Published: April 25, 2013

doi:10.5539/ijsp.v2n2p89

URL: <http://dx.doi.org/10.5539/ijsp.v2n2p89>

## Abstract

Nephropathy is a life-threatening complication of diabetes mellitus and is the leading cause of end-stage renal disease. The rate of rise in Serum Creatinine (SrCr) is a well-accepted marker for the progression of Diabetic Nephropathy (DN). The objective of this paper is to estimate the DN onset times and for this a retrospective data from 132 type 2 diabetic patients were collected as per American Diabetes Association (ADA) standards. This data is divided into four groups: Less Advance DN (LADN) group ( $1.4\text{mg/dl} \leq \text{SrCr} \leq 1.9\text{ mg/dl}$ ), Advance DN (ADN) group ( $\text{SrCr} > 1.9\text{ mg/dl}$ ), Non-Informative group ( $\text{SrCr} < 1.4\text{ mg/dl}$  & duration of diabetes  $\leq 15$  years) and Controlled group ( $\text{SrCr} < 1.4\text{ mg/dl}$  & duration of diabetes  $> 15$  years). For estimating the time of onset of DN for each group we have applied Bayesian approach and Bayes estimate is obtained under squared error loss function for the unknown scale parameter. We found that patients with DN as complication will reach less advance and advance stages of DN in approximately 15.012 years and 16.890 years respectively. And it was also found that patients with duration of diabetes less than 15 years (non-informative group) will be free from DN for at least 7.781 years. And, patients with duration of diabetes greater than 15 years (controlled group) will not develop any complication up to 20.109 years. This estimation can be used to predict the future onset times of new type 2 diabetic patients.

**Keywords:** Akaike Information Criterion, Bayesian analysis, conjugate prior, diabetic nephropathy, Jeffreys non-informative prior, serum creatinine, squared error loss function, survival distribution

## 1. Introduction

Diabetes Mellitus (DM) is one of the most common non-communicable chronic disease and its current pandemic threatens to be a rapidly expanding burden in the future for both the developed and developing countries. By the year 2025 the number of persons with diabetes is expected to increase by 41% (51 to 72 million) in developed countries and by 170% (84 to 228 million) in developing countries (Engelgau, Narayan, Saaddine, & Vinicor, 2003). Type 2 diabetes accounts for more than 90% of all diabetes cases and is the main driver of diabetes epidemic (Mohan & Pradeepa, 2009).

The chronic complications of DM affect many organ systems and are responsible for majority of morbidity and mortality associated with the disease (Harrison, 2008). Diabetic Nephropathy (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 (Bojestig, Arnqvist, Hermansson, Karlberg, & Ludvigsson, 1994). With development of kidney complications, 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 to 3.0 mg/dl) is the indicator for impaired renal function (Adler et al., 2003) and the normal level of creatinine is 0.8 to 1.4 mg/dl (Dabla, 2010). DN develops in 20–30% of patients with type 2 diabetes, and 3–8% of type 2 diabetic's progress to ESRD (ADA, 2000).

In many investigative fields, including engineering and medical research, researchers are interested in estimating the time until an event of interest occurs (also known as survival time). In a field involving live subjects like medical, one is interested in estimating the time until death of the subject from the beginning of observation time, onset of a disease etc. Predictive models are used in a variety of medical domains for estimating these survival

times. These models can be viewed as a tool for dealing with uncertainty. A subjective Bayesian analysis is one of the approaches for estimating the survival time. The Bayesian analysis is the only statistical theory that combines modeling inherent uncertainty and statistical uncertainty (Al-Kutubi, 2011). It is a statistical procedure that endeavors to estimate parameters of an underlying distribution based on the observed distribution. The main difference between Bayesian analysis and some classical statistical methods is that we use not only the sample information but also some information about the parameter  $\theta$  in Bayesian analysis. Essential in the Bayesian approach is to view the parameter  $\theta$  as a value of some random variable  $X$  with a known distribution (rather than viewing  $\theta$  as an unknown constant). This completely specified (discrete or continuous) density over the parameter space  $\Theta$  is called the prior density, which reflects past experience about the parameter  $\theta$ . Usually, given the states of a random variable  $X$ , a conditional probability is attached to this variable, say  $f(x|\theta)$ , and a prior density of the parameter  $\theta$ , say  $g(\theta)$ , is specified based on previous knowledge. Also, in Bayesian analysis, the prior density and the sample information are combined via Bayes theorem to obtain the posterior density of  $\theta$  given the sample information  $X$ ,  $g(\theta|x)$ .

Bayesian analysis requires the assertion of a prior distribution for the unknown parameters. The prior distribution can be viewed as representing the current state of knowledge, or current description of uncertainty, about the model parameters prior to data being observed. Priors are either non-informative or informative. A non-informative prior is a function which is used when little or no prior information is available about the unknown parameter  $\theta$ . The commonly used non-informative prior is the Jeffreys prior, which is proportional to the square root of the determinant of the Fisher information matrix (Jeffreys, 1961). Informative prior is used when there is substantial knowledge about a phenomenon under investigation. For monotonic hazard rate, a number of distributions have been proposed and perhaps the most widely used among these are Weibull and gamma distributions. Both of these distributions have increasing/decreasing hazard rate depending on their shape parameters and are found to be useful in modeling and analyzing life time data in the fields of medicine, biology, engineering sciences and others.

Miller (1980) obtained the posterior distribution for scale and shape parameters of the two-parameter gamma distribution using non-informative and conjugate priors. Soland (1969) examined the Weibull process with unknown scale and shape parameters as a model for Bayesian decision making. He used a family of prior distributions that places continuous distributions on the scale parameter and discrete distribution on the shape parameter. Soland has applied the Bayesian treatment of Weibull process for two purposes, (i) to make probability statements about the mean life and reliability of a long-life component both before and after life testing and (ii) estimation of the probability distribution of the number of replacement items needed by a group of users during a specified future time interval. Sun (1997) investigated the non-informative priors for two-parameter Weibull distribution with single and both parameters unknown by including Jeffreys and reference priors. Zhang and Meeker (2005) described the Bayesian methods using conjugate prior distributions for type II censored data from a Weibull distribution with given shape parameter. They have also investigated the exponential distribution as a special case of Weibull distribution. Al-Kutubi (2011) estimated the scale parameter and survival function for Weibull distribution under squared error loss function using the standard Bayes estimation method which involves in taking the Jeffreys prior as the marginal distribution of the unknown scale parameter. The same has been obtained by Al Omari and Ibrahim (2011) for the right censored data. They have also obtained the estimate of survival function by using an extension of Jeffreys prior. Haq and Dey (2011) considered the problem of Bayesian estimation of the parameters of gamma distribution under squared error loss function by assuming different independent informative priors as well as joint priors for both shape and scale parameters. They have also compared these estimates for different choices of hyperparameters using a simulation study. Al Omari, Ibrahim, Adam, and Arasan (2012) estimated the survival function and hazard rate of the Weibull distribution for the right censored data by using Bayesian estimator with Jeffreys and extension of Jeffreys prior.

The objective of this paper is to estimate the survival time which is here defined as the time from diagnosis of diabetes to the onset of DN using Bayesian estimation approach. For this retrospective data from 132 type 2 diabetic patients were collected as per American Diabetes Association (ADA) standards from the data base of Dr. Lal's Path Lab through house to house survey and the study was terminated on November 2007. The following variables were recorded for each patient: (i) value of serum creatinine (SrCr), (ii) fasting blood glucose (FBG), (iii) systolic blood pressure (SBP), (iv) diastolic blood pressure (DBP), (v) low density lipoprotein (LDL), (vi) duration of diabetes and (vii) age at which diabetes was diagnosed. Grover, Gadpayle and Sabharwal (2012) have used multiple linear regression (MLR) and logistic regression models by taking the above variables as predictors and have shown that DN can be estimated on the basis of SrCr value only. So to achieve our objective the data is divided into two groups: DN group and Non Diabetic Nephropathy (NDN) group on the basis of SrCr value. The

patients in DN group have SrCr level greater than 1.4 mg/dl and the patients in NDN group have SrCr level less than 1.4 mg/dl. The DN group is further divided into two groups Less Advance Diabetic Nephropathy (LADN) group ( $1.4 \text{ mg/dl} \leq \text{SrCr} \leq 1.9 \text{ mg/dl}$ ) and Advance Diabetic Nephropathy (ADN) group ( $\text{SrCr} > 1.9 \text{ mg/dl}$ ) (Grover, Gadpayle, & Sabharwal, 2010; Lewis, 1993). The NDN group has also been further divided into two groups: Non-Informative and Controlled on the basis of duration of diabetes. Now for estimating the time of onset of DN for each group we have applied Bayesian approach. For applying this method firstly testing of distribution of duration of diabetes is done for each group. Secondly, appropriate priors are selected and posterior distributions are obtained for all the four groups. Thirdly, the Bayes estimate under squared error loss function is obtained for the unknown scale parameter for each group. Finally, we compare the onset DN times of type 2 diabetic patients of real data with simulated study under several conditions. A wide variety of Bayesian applications have been considered earlier by taking appropriate distributions and priors under various loss functions which adequately determines the unknown parameters. In this paper, the research interest is focused on the generation of models for estimating the survival time of type 2 diabetic patients who had different renal health. This estimation can be used to predict the future onset times of new type 2 diabetic patients and thus helps in developing treatment comparison designs. Besides introduction this paper includes three more sections. In section 2 developments of the models are introduced. Section 3 applies the models to the data set of type -2 diabetic patients of corresponding groups and some concluding remarks are made in section 4.

**2. Methodology**

The study consists of  $n$  patients who are divided into two groups namely DN group and NDN group with  $n_1$  and  $n_2$  patients respectively. The  $n_1$  patients under DN group are further divided into two groups on the basis of mean SrCr values of sizes  $m_1$  and  $n_1 - m_1$ . The patients with the mean SrCr value lying between 1.4 mg/dl to 1.9 mg/dl are classified as LADN group and the patients with the mean SrCr value greater than 1.9 mg/dl are classified as ADN group. Also the  $n_2$  patients under NDN group are divided into two groups of sizes  $m_2$  and  $n_2 - m_2$  on the basis of duration of diabetes. The patients with less than 15 years of diabetic history belong to non-informative group and the patients with more than 15 years of diabetic history belong to controlled group.

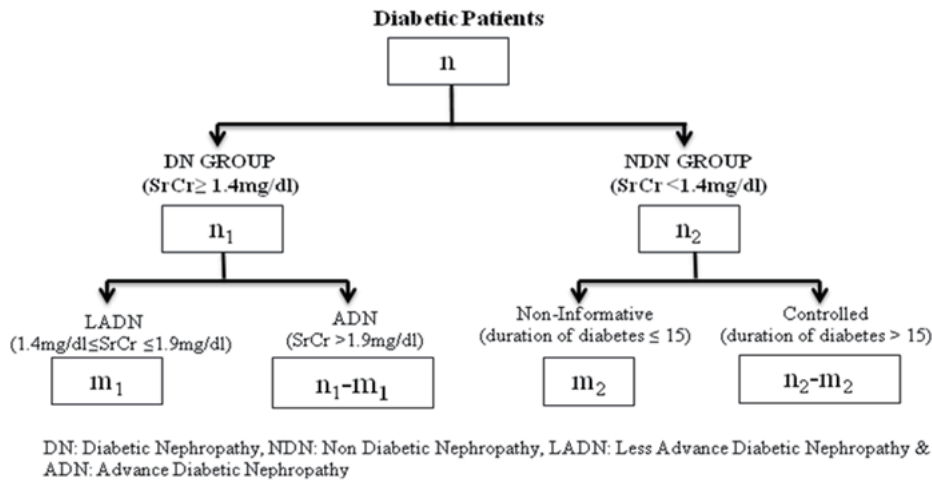


Figure 1. Classification of  $n$  type 2 diabetic patients under various groups

**2.1 Development of Models**

**2.1.1 Model for Less Advance Diabetic Nephropathy Group**

Let the random variables  $T_1, T_2, \dots, T_{m_1}$  denote the duration of diabetes for  $m_1$  individuals who are under LADN group. These random variables follow Weibull distribution with given shape parameter  $\gamma$  and unknown scale parameter  $\lambda$ . The probability density function of Weibull  $(\lambda, \gamma)$  is given by:

$$f(T_i|\lambda, \gamma) = \frac{\gamma}{\lambda} \left(\frac{T_i}{\lambda}\right)^{\gamma-1} e^{-(T_i/\lambda)^\gamma}; \lambda > 0, \gamma > 0 \ \& \ T_i > 0; \ i = 1, 2, \dots, m_1 \tag{1}$$

Then, the likelihood function is given as (Klein & Moeschberger, 2003),

$$L(\lambda, \gamma | T) = \prod_{i=1}^{m_1} [f(T_i | \lambda, \gamma)] \quad (2)$$

$$L(\lambda, \gamma | T) = \frac{\gamma^{m_1}}{\lambda^{(m_1)\gamma}} \prod_{i=1}^{m_1} T_i^{\gamma-1} e^{-\sum_{i=1}^{m_1} (T_i/\lambda)^\gamma} \quad (3)$$

To perform the Bayesian analysis, a gamma prior  $\text{Gamma}(\alpha, \beta)$  is assumed for the scale parameter  $\lambda$  with probability density function,

$$g(\lambda) = \frac{\beta^\alpha}{\Gamma} \lambda^{\alpha-1} e^{-\beta\lambda}; \quad \alpha > 0, \beta > 0, \lambda > 0 \quad (4)$$

The posterior distribution of  $\lambda$  using Bayes theorem is given as follows,

$$g_{LADN}(\lambda | T, \gamma) = \frac{g(\lambda) \times L(\lambda, \gamma | T)}{\int_{\Theta} g(\lambda) \times L(\lambda, \gamma | T) d\lambda} \quad (5)$$

$$g_{LADN}(\lambda | T, \gamma) = \frac{\left(\frac{\beta^\alpha}{\Gamma} \lambda^{\alpha-1} e^{-\beta\lambda}\right) \times \left(\frac{\gamma^{m_1}}{\lambda^{m_1\gamma}} \prod_{i=1}^{m_1} T_i^{\gamma-1} e^{-\sum_{i=1}^{m_1} (T_i/\lambda)^\gamma}\right)}{\int_0^\infty \left(\frac{\beta^\alpha}{\Gamma} \lambda^{\alpha-1} e^{-\beta\lambda}\right) \times \left(\frac{\gamma^{m_1}}{\lambda^{m_1\gamma}} \prod_{i=1}^{m_1} T_i^{\gamma-1} e^{-\sum_{i=1}^{m_1} (T_i/\lambda)^\gamma}\right) d\lambda} = \frac{\lambda^{\alpha-m_1\gamma-1} e^{-(\sum_{i=1}^{m_1} (T_i/\lambda)^\gamma + \beta\lambda)}}{\int_0^\infty \lambda^{\alpha-m_1\gamma-1} e^{-(\sum_{i=1}^{m_1} (T_i/\lambda)^\gamma + \beta\lambda)} d\lambda} \quad (6)$$

The Bayes estimate of  $\lambda$  is obtained under the squared error loss function which is the posterior mean estimated as follows,

$$\hat{\lambda} = E^{\lambda | T}[\lambda] = \int_0^\infty \lambda g_{LADN}(\lambda | T, \gamma) d\lambda \quad (7)$$

This estimate is used to obtain the mean onset time of DN for the  $m_1$  patients under this group.

### 2.1.2 Model for Advance Diabetic Nephropathy Group

Let the random variables  $T_{m_1+1}, T_{m_1+2}, \dots, T_{n_1}$  denote the duration of diabetes for  $n_1 - m_1$  individuals who are under ADN group. These random variables follow gamma distribution with given shape parameter  $\gamma$  and unknown scale parameter  $\lambda$  whose probability density function for the  $i^{th}$  random variable is given by,

$$f(T_i | \gamma, \lambda) = \frac{\lambda^\gamma}{\Gamma} T_i^{\gamma-1} e^{-\lambda T_i}; \quad \lambda > 0, \gamma > 0, i = m_1 + 1, m_1 + 2, \dots, n_1 \quad (8)$$

And the corresponding likelihood function can be obtained using Equation (2) as,

$$L(\lambda, \gamma | T) = \frac{\lambda^{(n_1-m_1)\gamma}}{(\Gamma)^{(n_1-m_1)}} \prod_{i=m_1+1}^{n_1} T_i^{\gamma-1} e^{-\lambda \sum_{i=m_1+1}^{n_1} T_i} \quad (9)$$

Previous studies suggest that conjugate prior is a flexible prior for gamma distribution. A prior is said to be a conjugate prior for a family of distributions if the prior and posterior distributions are from the same family, which means that the form of the posterior has the same distributional form as the prior distribution. Thus, assuming a conjugate prior for the scale parameter  $\lambda$ , whose probability density function is defined in Equation (4). And the posterior distribution is obtained by using Equation (5) as,

$$\begin{aligned} g_{ADN}(\lambda | T, \gamma) &= \frac{\frac{\lambda^{(n_1-m_1)\gamma}}{(\Gamma)^{(n_1-m_1)}} \prod_{i=m_1+1}^{n_1} T_i^{\gamma-1} e^{-\lambda \sum_{i=m_1+1}^{n_1} T_i} \cdot \frac{\beta^\alpha}{\Gamma} \lambda^{\alpha-1} e^{-\beta\lambda}}{\int_0^\infty \frac{\lambda^{(n_1-m_1)\gamma}}{(\Gamma)^{(n_1-m_1)}} \prod_{i=m_1+1}^{n_1} T_i^{\gamma-1} e^{-\lambda \sum_{i=m_1+1}^{n_1} T_i} \cdot \frac{\beta^\alpha}{\Gamma} \lambda^{\alpha-1} e^{-\beta\lambda} d\lambda} \\ &= \frac{(\beta + \sum_{i=m_1+1}^{n_1} T_i)^{\alpha+(n_1-m_1)\gamma} \lambda^{\alpha+(n_1-m_1)\gamma-1} e^{-\lambda(\beta + \sum_{i=m_1+1}^{n_1} T_i)}}{\Gamma(\alpha + (n_1 - m_1)\gamma)} \end{aligned} \quad (10)$$

which is also a Gamma distribution with parameters  $(\alpha + (n_1 - m_1)\gamma, \beta + \sum_{i=m_1+1}^{n_1} T_i)$ .

The Bayes estimate of  $\lambda$  can be obtained in the similar as done for LADN group,

$$\hat{\lambda} = \frac{\alpha + (n_1 - m_1)\gamma}{\beta + \sum_{i=m_1+1}^{n_1} T_i} \quad (11)$$

This estimate is further used to obtain the mean onset time of DN for  $(n_1 - m_1)$  patients under this group.

### 2.1.3 Model for Non-Informative Group

Let the random variables  $T_1, T_2, \dots, T_{m_2}$  denote the duration of diabetes of  $m_2$  diabetic patients for whom there is no information about the onset of DN. These random variables follow gamma distribution with given shape parameter  $\gamma$  and unknown scale parameter  $\lambda$ . The likelihood function with appropriate sample size can be obtained using Equation (9) in section 2.1.2. Since, there is a lack of information regarding the development of DN we assume a non-informative Jeffreys prior for the scale parameter  $\lambda$ . The probability density function of  $\lambda$  under this prior is given by,

$$g(\lambda) \propto \sqrt{I(\lambda)} = k\sqrt{I(\lambda)} \quad (12)$$

Where,  $k$  is the constant of proportionality and  $I(\lambda)$  is the information matrix numerically represented as,

$$I(\lambda) = -m_2 \cdot E\left(\frac{\partial^2 \ln f(T, \lambda, \gamma)}{\partial \lambda^2}\right) \quad (13)$$

Then using Equation (7) from appendix the posterior distribution is obtained as,

$$\begin{aligned} g_{Non-Inf}(\lambda|T, \gamma) &= \frac{k/\lambda \sqrt{m_2\gamma} \times \lambda^{m_2\gamma} (\bar{y})^{-m_2} \prod_{i=1}^{m_2} T_i^{\gamma-1} e^{-\lambda \sum_{i=1}^{m_2} T_i}}{\int_0^\infty k/\lambda \sqrt{m_2\gamma} \times \lambda^{m_2\gamma} (\bar{y})^{-m_2} \prod_{i=1}^{m_2} T_i^{\gamma-1} e^{-\lambda \sum_{i=1}^{m_2} T_i} d\lambda} \\ &= \frac{\lambda^{m_2\gamma-1} e^{-\lambda \sum_{i=1}^{m_2} T_i} (\sum_{i=1}^{m_2} T_i)^{m_2\gamma}}{m_2\gamma} \end{aligned} \quad (14)$$

which is also a gamma distribution with parameters  $(m_2\gamma, \sum_{i=1}^{m_2} T_i)$ .

And Bayes estimator of  $\lambda$  under squared error loss function is given as,

$$\hat{\lambda} = \frac{m_2\gamma}{\sum_{i=1}^{m_2} T_i} \quad (15)$$

Using this estimated value of  $\lambda$  we have obtained the mean duration of time during which the patients under this group will not develop DN complication.

### 2.1.4 Model for Controlled Group

The patients for whom the development of diabetic nephropathy is not observed are said to be controlled. So for this let, the random variables  $T_{m_2+1}, T_{m_2+2}, \dots, T_{n_2}$  denote the duration of diabetes for  $n_2 - m_2$  observations. These random variables follow gamma distribution with given shape parameter  $\gamma$  and unknown scale parameter  $\lambda$ . An exponential prior is assumed for scale parameter  $\lambda$  as the patients under this group have approximately constant hazard of developing diabetic nephropathy. This distribution is regarded as a special case of Weibull distribution when shape parameter equal to one and its density function can be represented as,

$$g(\lambda) = \alpha e^{-(\lambda\alpha)}; \alpha > 0, \lambda > 0 \quad (16)$$

where  $\alpha$  is the scale parameter. Then the posterior distribution is given by the following equation,

$$g_{Controlled}(\lambda|T, \gamma) = \frac{\lambda^{(n_2-m_2)\gamma} e^{-\lambda(\sum_{i=m_2+1}^{n_2} T_i + \alpha)} (\sum_{i=m_2+1}^{n_2} T_i + \alpha)^{(n_2-m_2)\gamma+1}}{(n_2 - m_2)\gamma + 1} \quad (17)$$

which is also a gamma distribution with parameters  $((n_2 - m_2)\gamma + 1, \sum_{i=m_2+1}^{n_2} T_i + \alpha)$ .

The Bayes estimator under squared error loss function can be obtained by using Equation (7) and is given as follows,

$$\hat{\lambda} = \frac{(n_2 - m_2)\gamma + 1}{\sum_{i=m_2+1}^{n_2} T_i + \alpha} \quad (18)$$

The estimated value of  $\lambda$  is used to obtain the mean duration of time during which the patients under controlled group will be free from any complications.

### 2.2 Model Checking

Most statistical methods assume an underlying distribution (model) in the derivation of their results. However when we assume that the data follow a specific distribution, we are making an assumption (Khan & Hakkak, 2012). If such a model does not hold then the conclusion from such analysis may be invalid. Hence model verification is necessary to check whether we have achieved the goal of choosing the right model. To check the model appropriateness we have considered two methods. Firstly, by plotting graphs of the survival function of the fitted distribution. Secondly, by applying Akaike information criterion (AIC) which is one of the most commonly used information criteria for model selection and is defined by the following equation,

$$AIC = -2 \log(L) + 2p \quad (19)$$

Where  $L$  refers to the likelihood under the fitted model and  $p$  is the number of parameters in the model. Because larger likelihoods are preferable and we are working on the negative log scale, smaller values of this AIC points to better model. That is if the AIC is smaller for the first model than the second then the former is preferred (Lindsey & Jones, 1998; Acquah, 2010).

### 3. Application

The methods discussed in section 2 are applied to the data obtained through house to house survey of type 2 diabetic patients who were referred for pathological tests to Dr. Lal path lab, Delhi, India. Retrospective study has been conducted on the collected data. The patients with co-morbidity like retinopathy, cardiovascular diseases; neuropathy has been excluded as our study focus only on diabetic nephropathy. Thus, a total of 132 patients were selected who were diagnosed as diabetic as per American Diabetes Association (ADA) standards. The ADA and the National Institutes of Health recommend eGFR calculated from serum creatinine at least once a year in all people with diabetes for detection of kidney dysfunction (Dabla, 2010). On the basis of SrCr the event of interest nephropathy is observed for each patient and it is found at the end of study that out of 132 patients there are only 60 (45.45%) DN cases and remaining 72 (54.55%) are NDN cases. Table 1 represents descriptive statistics of 132 type 2 diabetic patients giving minimum, maximum, range and mean  $\pm$  S.D of the variables, age at diagnosis, duration of diabetes, fasting blood glucose(FBG), systolic blood pressure(SBP), diastolic blood pressure(DBP), low density lipoprotein(LDL) and SrCr for two groups: DN and NDN group. The DN cases are divided into two groups: LADN group ( $1.4 \text{ mg/dl} \leq \text{SrCr} \leq 1.9 \text{ mg/dl}$ ) and ADN group ( $\text{SrCr} > 1.9 \text{ mg/dl}$ ). Also, the NDN cases are divided into two groups on the basis of duration of diabetes: non-informative group (duration of diabetes  $\leq 15$  years) and controlled group (duration of diabetes  $> 15$  years). Table 2 represents number of patients and observed mean onset time of disease for each of the four groups.

Table 1. Descriptive statistics of 132 type 2 diabetic patients giving minimum, maximum, range and mean  $\pm$  standard deviation of age at diagnosis, duration of diabetes, fasting blood glucose (FBG), diastolic blood pressure (DBP), systolic blood pressure (SBP), low density lipoprotein (LDL) and serum creatinine (SrCr) for two groups i.e. DN and NDN

Variable	Statistic	DN group	NDN group
Age at diagnosis (years)	Minimum	29	35
	Maximum	56	58
	Range	27	23
	mean $\pm$ S.D	45.003 $\pm$ 5.28	44.011 $\pm$ 4.36
Duration of disease (years)	Minimum	6.1	5.6
	Maximum	29	27
	Range	22.9	21.4
	mean $\pm$ S.D	14.0931 $\pm$ 5.0528	10.2784 $\pm$ 5.7
FBG (mg/dl)	Minimum	120	62
	Maximum	242	186
	Range	122	124
	mean $\pm$ S.D	142.035 $\pm$ 14.39	133.8027 $\pm$ 17.48
DBP (mm Hg)	Minimum	76	68
	Maximum	112	95
	Range	36	27
	mean $\pm$ S.D	91.9695 $\pm$ 9.423	82.3919 $\pm$ 6.0789
SBP (mm Hg)	Minimum	110	110
	Maximum	160	160
	Range	50	50
	mean $\pm$ S.D	142.8214 $\pm$ 13.8815	125.1214 $\pm$ 12.4007
LDL (mg/dl)	Minimum	68	62
	Maximum	132	186
	Range	64	124
	mean $\pm$ S.D	107.4417 $\pm$ 14.2667	91.7973 $\pm$ 18.75007
SrCr (mg/dl)	Minimum	1.2	0.71
	Maximum	2.21	1.39
	Range	1.01	0.92
	mean $\pm$ S.D	1.6686 $\pm$ 0.28233	0.9982 $\pm$ 0.15084

Table 2. Number of subjects and observed mean onset time of disease of LADN, ADN, Non-informative & Controlled Group

Group	Number of type 2 diabetic patients	Observed mean onset time
LADN	33	15.125
ADN	27	16.720
NON-INFORMATIVE	54	7.807
CONTROLLED	18	20.256

LADN: Less Advance Diabetic Nephropathy, ADN: Advance Diabetic Nephropathy.

### 3.1 Estimating the Mean Onset Time of DN for Patients Under the LADN Group

The motivation of our model is to estimate the survival time (from the diabetes diagnosis till the onset of DN) under different renal health status. There are 33 patients under LADN group with SrCr level between 1.4 mg/dl to 1.9 mg/dl. The model assumption of duration of diabetes for patients under this group is tested by calculating AIC values for different distribution models and the model with least AIC value is considered to be the best fit. For this group Weibull distribution has the minimum AIC value which came out to be 274.556 as given in Table 3. Thus, it is evident that Weibull distribution with  $\gamma = 2.621$ ,  $\lambda = 17.112$  as maximum likelihood estimates (MLE) of shape and scale parameters respectively fits the data well. The fitting is also tested by plotting PP plots (Figure 2(a)) and observed that Weibull distribution is a good fit over the entire range of survival times. For Bayesian analysis the distribution of  $\lambda$  is of primary interest so we introduce a probability model for the scale parameter  $\lambda$ .

Previous studies suggest that a diabetic patient takes approximately 15 years to develop DN (Dabla, 2010) which is also confirmed by Grover, Gadpayle and Sabharwal (2010) on same data, so a gamma prior with 21.720 and 1.450 as the shape and scale parameters respectively is an appropriate prior for this group (as mean of  $G(21.720, 1.450)$  is 14.970). And the Bayes estimate is obtained under squared error loss function as  $\hat{\lambda} = 16.840$ . Then using this Bayes estimate of  $\lambda$  we have generated 5 samples each of size 1000 from Weibull distribution with 2.621 and 16.840 as shape and scale parameter respectively, to estimate the mean onset time of DN. The samples mean are computed and are presented in Table 4(a). Hence, it can be concluded that a diabetic patient with DN as the complication will reach its less advance stage in approximately 15.012 years from the diagnosis of diabetes which matches with the observed mean value (Table 2).

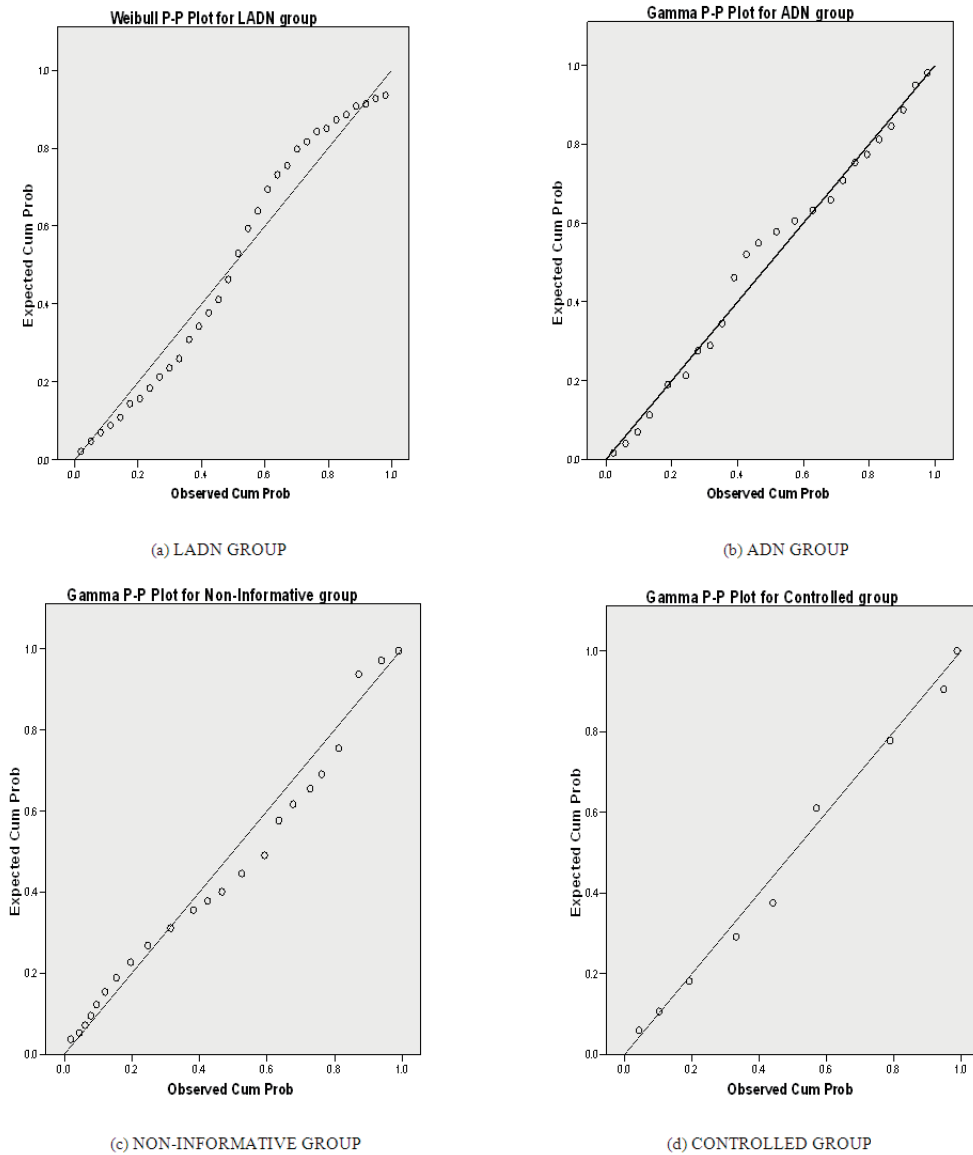


Figure 2. Probability plots (P-P) using MLEs estimate for LADN, ADN, non-informative and controlled groups

### 3.2 Estimating the Mean Onset Time of DN for Patients Under the ADN Group

Out of 132 patients there are 27 patients with SrCr greater than 1.9 mg/dl. Gamma model is considered to be the best model as it has the minimum AIC value as 105.511. Thus, it is evident that gamma distribution with  $\gamma = 15.304$ ,  $\lambda = 0.869$  as shape and scale parameters respectively fits the data well. The fitting is also tested by plotting P-P plots (Figure 2(b)). Applying similar approach as done for LADN group we have assumed a conjugate



gamma prior for  $\lambda$  with 17.304 and 0.990 as the shape and scale parameters respectively for Bayesian analysis. And the Bayes estimate  $\lambda$  is obtained under squared error loss function using Equation (11) as 0.904. Then using this Bayes estimate of  $\lambda$  we have generated 5 samples each of size 1000 from gamma distribution with 15.304 and 0.904 as shape and scale parameter respectively and mean onset times of DN is computed for each sample and presented in Table 4(a). It can be concluded that a diabetic patient with DN as a complication will reach its advance stage in 16.890 years (approximately) from the time of diagnosis of diabetes.

### 3.3 Estimating the Mean Duration of Diabetes for Patients Under the Non-Informative Group

For non-informative group we have collected data regarding the duration of diabetes for 54 type 2 diabetic patients. The level of SrCr for the patients under this group is less than 1.4 mg/dl (i.e. no DN complication) and the duration of diabetes is less than 15 years. Since, the duration of disease is less than 15 years we have only partial information about renal complication of patients under this group. For this group gamma distribution has the least AIC value. Thus, it is evident that the gamma distribution with  $\gamma = 15.122$ ,  $\lambda = 1.937$  as shape and scale parameters respectively fits the data well. The fitting is also tested by plotting P-P plots (Figure 2(c)) and it was observed that the gamma distribution is a good fit over the entire range of duration of diabetes. Since, there is no prior knowledge about the scale parameter  $\lambda$  we have assumed a non-informative Jeffreys prior. The Bayes estimate is obtained under squared error loss function as  $\hat{\lambda} = 1.939$ . Then using this estimate Gamma (15.122, 1.939) is generated for 5 samples each of size 1000 and their respective means are shown in Table 4(a). Thus, it can be concluded that the diabetic patients under this group are free from DN complication for at least 7.781 years (approximately).

### 3.4 Estimating the Mean Duration of Diabetes for Patients Under the Controlled Group

The controlled group is defined as a group consisting of patients with SrCr level less than 1.4 mg/dl and duration of diabetes greater than 15 years. Out of 132 observations there are only 18 under this group and these observations represents the duration of diabetes. Because of the limited number of observations under this group firstly a Bootstrap sample of length 32 is generated. And then model is selected on the basis of these 50 observed (18) and generated (32) observations. Gamma model is considered to be the best model as it has the minimum AIC value as 162.186. The fitting is also tested by plotting P-P plots (Figure 2(d)) and it was observed that the gamma distribution is a good fit over the entire range of duration of diabetes. Therefore, we conclude that gamma distribution with  $\gamma = 78.211$ ,  $\lambda = 3.886$  as shape and scale parameters respectively fits the data well. The scale parameter  $\lambda$  is assumed to follow exponential distribution with hyperparameter  $\alpha = 0.050$ . The Bayes estimate is obtained under squared error loss function as  $\hat{\lambda} = 3.886$ . Then using this estimate Gamma (78.211, 3.886) is generated for 5 samples each of size 1000 and their respective means are shown in Table 4(a). Hence, it can be concluded that the patients under this group will not develop any complications for 20.109 years (approximately) if all the factors like FBG, LDL etc. are controlled. The calculations and analysis are performed using SPSS for Windows, Version 15 and MATLAB, Version 6.5 statistical packages.

Table 3. AIC values, parameter estimates using MLE and Bayesian approach for LADN, ADN, non-informative and controlled group

Group	Model	AIC Value	Selected Model	MLE		Bayesian	
				Estimate of parameters( $\gamma, \lambda$ )	Standard Error	Estimate of parameter $\lambda$	Standard Error
LADN	Weibull	274.556	Weibull	(2.621, 17.112)	(0.361, 0.963)	16.840	0.941
	Gamma	322.077					
	Lognormal	300.036					
ADN	Weibull	139.233	Gamma	(15.304, 0.869)	(28.222, 1.500)	0.904	0.044
	Gamma	105.511					
	Lognormal	399.448					
Non-Informative	Weibull	1824.253	Gamma	(15.122, 1.937)	(18.753, 0.053)	1.939	0.065
	Gamma	203.723					
	Lognormal	439.386					
Controlled	Weibull	1667.262	Gamma	(78.211, 3.886)	(239.603, 2.112)	3.886	1.971
	Gamma	162.186					
	Lognormal	401.121					

Table 4(a). Mean of 5 generated samples each of size 1000, estimate of parameter is based on real data for computing mean duration of disease under LADN, ADN, non-informative and controlled group

Sample	LADN	ADN	Non-informative	Controlled
	Weibull(2.621, 16.840)	Gamma(15.304, 0.904)	Gamma(15.122, 1.939)	Gamma(78.211, 3.886)
I	15.112	17.008	7.77	20.099
II	14.796	16.727	7.711	20.09
III	15.207	16.929	7.91	20.045
IV	15.017	16.871	7.798	20.133
V	14.927	16.913	7.716	20.178
Mean	15.012	16.890	7.781	20.109

Table 4(b). Mean of 5 generated samples each of size 1000, estimate of parameter is based on simulated data for computing mean duration of disease under LADN, ADN, non-informative and controlled group

Sample	LADN	ADN	Non-informative	Controlled
	Weibull(2.621, 17.017)	Gamma(15.304, 0.887)	Gamma(15.122, 2.002)	Gamma(78.211, 3.706)
I	15.103	17.334	7.481	20.989
II	15.126	17.047	7.529	21.081
III	15.513	17.254	7.519	20.561
IV	15.343	17.194	7.683	21.129
V	15.102	17.237	7.516	20.287
Mean	15.237	17.213	7.546	20.809

#### 4. Discussion

The aim of this paper is to estimate the onset time of DN/duration of diabetes of type 2 diabetic patients under various renal health conditions. The different renal health status of patients is defined on the basis of SrCr as it is considered to be an important marker for predicting DN. The major use of estimating the survival time of diabetic patients is that the current as well as the future life time of new diabetic patients can be predicted.

To achieve the aim of our study we have divided the data of 132 type 2 diabetic patients into DN and NDN groups. The 60 patients belonging to DN group are further classified into two groups namely LADN and ADN groups of sizes 33 and 27 respectively. The classification is done on the basis of SrCr level, the first group has SrCr level between 1.4 mg/dl to 1.9 mg/dl and the second group has SrCr level greater than 1.9 mg/dl. This classification is in accordance with the previous studies (Grover, Gadpayle, & Sabharwal, 2010; Lewis 1999). Also, the NDN group is further classified into non-informative and controlled group of sizes 54 and 18 respectively. The classification of these two groups is done on the basis of SrCr level and duration of diabetes. The SrCr level for both the groups is less than 1.4 mg/dl and the duration of diabetes for the patients under the non-informative group is less than 15 years and for the controlled group is greater than 15 years.

For estimating the survival time/duration of disease for each group we have applied Bayesian analysis. Bayesian framework is particularly useful in circumstances where a conventional approach to statistical analysis may be difficult or misleading. It is a preferable estimating method when the data with small sample size is available. For applying the Bayesian approach to first group i.e. LADN group the duration of diabetes for 33 subjects are found to follow Weibull distribution. As, Weibull distribution is a statistical model which has a broad range of applications in life testing and reliability theory with the major advantage of providing reasonably accurate failure analysis and failure forecasts with infinitesimally small samples. A gamma prior is assumed for scale parameter of fitted Weibull distribution and Bayes estimate is obtained under squared error loss function. The mean onset time of DN for patients under LADN group is found to be 15.012 years (approx.) indicating the minimum time taken by the diabetic patient to reach less advance stage of DN from time of diagnosis of diabetes. Similar approach is applied to ADN group where durations of diabetes of 27 patients are found to follow gamma distribution. Gamma distribution has also been used as a universal statistical law for various complex stochastic events in various fields such as physical, chemical and biological systems, among others. A flexible choice of a prior here is a conjugate prior as gamma distribution does not have a closed form of posterior distribution and we may need to use numerical methods to obtain posterior mean etc. The mean onset of DN of patients under this group is found to be 16.890 years (approx.) indicating minimum time taken by the diabetic patient to reach advance stage of DN from time of diagnosis of diabetes. For the non-informative group the duration of diabetes for 54 subjects are found to follow gamma distribution. Since there does not exist any prior knowledge about the parameter a non-informative

Jeffreys prior is assumed for the scale parameter. Jeffreys priors work well for single parameter models, but not for models with multidimensional parameters. The mean duration of diabetes for subjects under this group is found to be 7.781 years (approx.) indicating the minimum time period during which the renal complication will not occur from time of diagnosis of diabetes. Lastly, this approach is applied to controlled group for whom the subjects are again found to follow gamma distribution. An exponential prior is assumed for scale parameter and mean duration of diabetes is found to be 20.109 years (approx.). This indicates that patients under this group are free from any complications for at least 20.109 years (approx.) provided all the covariates are maintained under normal range. The above onset times corresponding to each group are computed on the basis of parameter estimates obtained from real data (shown in Table 4(a)). The mean onset times for each group are further computed on the basis of parameter estimates obtained from simulated data as shown in Table 4(b), these results are consistent with the results of Table 4(a) and the observed mean onset times shown in Table 2.

As we can see in the methodology section, the Bayesian approach is computationally more complex and time-consuming compared to the maximum likelihood strategy. Here, one may ask this question ‘why we should estimate the model parameters using the Bayesian method?’ When data with small sample size is available, the maximum likelihood method may lead to larger standard errors (Kazemnejad, Zayeri, Hamzah, Gharaaghaji, & Salehi, 2010). In these situations, the Bayesian approach (which may result in smaller standard errors) could be the preferable estimating method. In our study, Bayesian approach resulted in smaller standard errors than the ML method except for the non-informative group (Table 3). This is probably due to the fact that performance of MLE is quite close to that of the Bayes estimators with respect to the non-informative prior. Thus, if we have no prior information on the unknown parameters, then we can use either MLE or Bayes estimates (Kundu, 2008).

The findings of this paper are consistent with previous studies which states that kidney damage rarely occurs in first 10 years of diabetes, and 15 to 25 years will usually pass before kidney failure occurs. For people who live with diabetes for more than 25 years without any signs of kidney failure, the risk of ever developing it decreases (Dabla, 2010). Application of these models can also be used for estimating survival times of other complication of type-2 diabetes such as retinopathy, CVD and others. However, our study population has been confined to a certain region and therefore may not be a representative of the entire diabetic population of the country.

## References

- Acquah, H. de-Graft. (2010). Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship. *Journal of Development and Agricultural Economics*, 2(1), 1-6.
- Adler, A. I., Stevens, R. J., Manley, S. E., Bilous, R. W., Cull, C. A., & Holman, R. R. (2003). Development and progression of nephropathy in type 2 diabetes: Observation and modeling from the United Kingdom Prospective Diabetes Study. *Kidney International*, 63, 225-232.
- Al Omari, M. A., & Ibrahim, N. A. (2011). <http://dx.doi.org/10.1046/j.1523-1755.2003.00712.x>
- Alkutubi, H. S. (2011). On best estimation for parameter Weibull distribution. *World Academy of Science, Engineering and Technology*, 75, 281-283.
- Al Omari, M. A., & Ibrahim, N. A. (2011). Bayesian survival estimator for Weibull distribution with censored data. *Journal of Applied Sciences*, 11(2), 393-396. <http://dx.doi.org/10.3923/jas.2011.393.396>
- Al Omari, M. A., Ibrahim, N. A., Adam, M. B., & Arasan, J. (2012). Bayesian survival and hazard estimate for Weibull censored time distribution. *Journal of Applied Sciences*, 12(12), 1313-1317. <http://dx.doi.org/10.3923/jas.2012.1313.1317>
- American Diabetes Association. (2000). Diabetic Nephropathy (Position Statement). *Diabetes Care*, 23(7), 569-572.
- Bojestig, M., Arnqvist, H. J., Hermansson, G., Karlberg, B. E., & Ludvigsson, J. (1994). Declining incidence of nephropathy in insulin-dependent diabetes mellitus. *New England Journal of Medicine*, 330, 15-18. <http://dx.doi.org/10.1056/NEJM199401063300103>
- Crockcroft, D. W., & Gault, M. H. (1976). Prediction of creatinine clearance from serum creatinine. *Nephron*, 16, 31-41. <http://dx.doi.org/10.1159/000180580>
- Dabla, P. K. (2010). Renal function in diabetic nephropathy. *World J. Diabetes*, 1(2), 48-56. <http://dx.doi.org/10.4239/wjd.v1.i2.48>

- Engelgau, M. M., Narayan, K. M. V., Saaddine, J. B., & Vinicor, F. (2003). Addressing the burden of diabetes in the 21st century: Better care and primary prevention. *Journal of American Society of Nephrology*, 14, S88-S91. <http://dx.doi.org/10.1097/01.ASN.0000070143.71933.B0>
- Grover, G., Gadpayle, A. K., & Sabharwal, A. (2010). Identifying patients with diabetic nephropathy based on serum creatinine under zero truncated model. *Electronic Journal of Applied Statistical Analysis*, 3(1), 28-43. <http://dx.doi.org/10.1285/i20705948v3n1p28>
- Grover, G., Gadpayle, A. K., & Sabharwal, A. (2012). Identifying patients with diabetic nephropathy based on serum creatinine in the presence of covariates in type-2 diabetes: A retrospective study. *Biomed Res India*, 23(4), 1-11.
- Haq, A., & Dey, S. (2011). Bayesian estimation of Erlang distribution under different prior distributions. *Journal of Reliability and Statistical Studies*, 4(1), 1-30.
- Harrison. (2008). *Principles of Internal Medicine* (17th ed.). USA: McGraw-Hill.
- Jeffreys, H. (1961). *Theory of Probability*. Oxford: Oxford University Press.
- Kazemnejad, A., Zayeri, F., Hamzah, N. A., Gharaaghaji, R., & Salehi, M. (2010). A Bayesian analysis of bivariate ordered categorical responses using a latent variable regression model: application to diabetic retinopathy data. *Scientific Research and Essays*, 5(11), 1264-1273.
- Khan, M. A., & Hakkak, A. A. (2012). Bayesian estimation of the parameter of generalized exponential distribution using Markov chain Monte Carlo method in open BUGS for informative sets of prior. *Journal of Arts, Science and Commerce*, 3(2), 96-106.
- Klein, J. P., & Moeschberger, M. L. (2003). *Survival Analysis Techniques for Censored and Truncated data*. New York, USA: SpringerVerlag.
- Kundu, D. (2008). Bayesian inference and life testing plan for the Weibull distribution in presence of progressive censoring. *Technometrics*, 50(2), 144-154. <http://dx.doi.org/10.1198/004017008000000217>
- Lindsey, J. K., & Jones, B. (1998). Choosing among generalized linear models applied to medical data. *Statistics in Medicine*, 17, 59-68. [http://dx.doi.org/10.1002/\(SICI\)1097-0258\(19980115\)17:1<59::AID-SIM733>3.0.CO;2-7](http://dx.doi.org/10.1002/(SICI)1097-0258(19980115)17:1<59::AID-SIM733>3.0.CO;2-7)
- Miller, R. B. (1980). Bayesian analysis of the two-parameter gamma distribution. *Technometrics*, 22(1), 65-69. <http://dx.doi.org/10.2307/1268384>
- Mohan, V., & Pradeepa, R. (2009). Epidemiology of diabetes in different regions of India. *Health Administrator*, 22(1, 2), 1- 18.
- Soland, R. M. (1969). Bayesian analysis of the Weibull process with unknown scale and shape parameters. *IEEE Transactions on Reliability Analysis*, 18, 181-184. <http://dx.doi.org/10.1109/TR.1969.5216348>
- Sun, D. (1997). A note on noninformative priors for Weibull distributions. *Journal of Statistical Planning and Inference*, 61(2), 319-338. [http://dx.doi.org/10.1016/S0378-3758\(96\)00155-3](http://dx.doi.org/10.1016/S0378-3758(96)00155-3)
- Zhang, Y., & Meeker, W. Q. (2005). Bayesian life test planning for the Weibull distribution with given shape parameter. *Metrika*, 61, 237-249. <http://dx.doi.org/10.1007/s001840400334>

### Appendix: Derivation of Jeffreys Prior

The probability density function of  $\lambda$  under non-informative Jeffreys prior can be written as,

$$g(\lambda) = k \sqrt{-m_2 \cdot E\left(\frac{\partial^2 \ln f(T, \lambda, \gamma)}{\partial \lambda^2}\right)} \quad (1)$$

Where,  $k$  is the constant of proportionality and

$$\ln f(T, \lambda, \gamma) = \gamma \log \lambda - \lambda T + (\gamma - 1) \log T - \log \bar{\gamma} \quad (2)$$

The following equations are obtained by partially differentiating Equation (2) w.r.t  $\lambda$

$$\frac{\partial \ln f(T, \lambda, \gamma)}{\partial \lambda} = \frac{\gamma}{\lambda} - T \quad (3)$$

$$\frac{\partial^2 \ln f(T, \lambda, \gamma)}{\partial \lambda^2} = -\frac{\gamma}{\lambda^2} \quad (4)$$

And,

$$E\left(\frac{\partial^2 \ln f(T, \lambda, \gamma)}{\partial \lambda^2}\right) = E\left(-\frac{\gamma}{\lambda^2}\right) \quad (5)$$

Then  $I(\lambda)$  is obtained as,

$$I(\lambda) = m_2 \frac{\gamma}{\lambda^2} \int_0^{\infty} f(T, \lambda, \gamma) dT = m_2 \frac{\gamma}{\lambda^2} \quad (6)$$

So, the Jeffreys prior information is given by,

$$g(\lambda) = k \sqrt{m_2 \frac{\gamma}{\lambda^2}} \quad (7)$$