International Journal of Mathematical, Engineering and Management Sciences
Vol. 4, No. 1, 219–231, 2019
https://dx.doi.org/10.33889/IJMEMS.2019.4.1-019

The Generalized Gamma Shared Frailty Model under Different Baseline Distributions

Sukhmani Sidhu[1*], Kanchan Jain[2], Suresh K. Sharma[3]
Department of Statistics
Panjab University, Chandigarh, 160014, India
E-mails: [1]sukhmani.15@gmail.com, [2]jaink14@gmail.com, [3]ssharma643@yahoo.co.in
[*]*Corresponding author*

(Received November 15, 2017; Accepted June 19, 2018)

Abstract
In the analysis of clustered survival data, shared frailty models are often used when observations in the same group share common unknown risk factors or frailty. There is dependence in the event times belonging to the same group, while event times from different groups are conditionally independent given their covariates. In such models, the known effect on survival time is described using the baseline distribution and regression coefficients while the unknown effect is described through a frailty distribution. In this paper, the Gompertz, log-logistic, and generalized exponential distributions are studied as baseline distributions, under a shared frailty effect described by the generalized gamma distribution. Their hazard functions have been compared and their applicability under different settings and performance with generalized gamma frailty has been explored. These models are fitted to three real life datasets using Bayesian estimation methods and compared using the Bayesian Information Criteria (AIC, BIC, and DIC) and the Bayes Factor.

Keywords- Gompertz hazard, Log-logistic hazard, Generalized exponential hazard, Bayesian information criteria, Bayes factor.

1. Introduction

In survival studies, there are often observations that need to be grouped together on the basis of the study centre, hospital, city, etc. In all such cases, individuals belonging to the same group or cluster are exposed to the same environmental factors. Such factors are unknown and are depicted in the models as a common unknown risk or shared frailty of a group. This causes a dependence in survival times of a group while observations of different groups are conditionally independent. In shared frailty models, the effect on survival times due to known covariates or treatment effects is described using a baseline distribution and regression coefficients while effect due to unknown risk factors or randomness in the data is described using a frailty distribution.

The baseline distribution describes the overall common risk to all individuals in the study. The distribution of event times in survival studies is generally discussed in terms of the hazard function. The hazard function is the instantaneous probability of failure over time and hence the form of the baseline hazard function, $h_0(t)$ is of considerable importance. It may be increasing, decreasing or constant over time and thus, different distributions are applicable in different studies.

The frailty distribution in the shared frailty model is helpful in describing a dependence structure where the observations within the clusters are correlated as they share a common frailty, while clusters among themselves are conditionally independent given the covariates. In grouped data, it is expected that there will be high heterogeneity among clusters. However, as the factors affecting

International Journal of Mathematical, Engineering and Management Sciences
Vol. 4, No. 1, 219–231, 2019
https://dx.doi.org/10.33889/IJMEMS.2019.4.1-019

frailty are unknown, it would be more appropriate to describe the frailty effect using a flexible distribution whose probability density function (PDF) can take various shapes. The generalized gamma distribution (GGD) includes many distributions as its particular cases viz. Weibull, Gamma, exponential, and log-normal distributions (Khodabin and Ahmadabadi, 2010) and hence it is a convenient choice to model frailty (Sidhu et al., 2018).

Earlier work done on the generalized gamma shared frailty model (Balakrishnan and Peng, 2006; Chen et al., 2013) shows how this distribution can estimate the heterogeneity among clusters more efficiently than other distributions. However, there are computational problems due to the integral in the unconditional likelihood not being in a closed form, making it an unfavourable choice to model frailty. Bayesian estimation of the parameters of the model (Sidhu et al., 2018) can make estimation easier and faster as prior information can be incorporated in the model or the problems of multiple modes or non-convergence can be diagnosed quicker.

In this article, the applicability of the model under three different baseline distributions viz. Gompertz, log-logistic, and generalized exponential, have been explored. The distributions are widely applicable in survival studies for different types of datasets and have been modelled with a gamma frailty, inverse Gaussian frailty, and log-normal frailty (Manton et al., 1986; Klein et al., 1999; Wienke et al., 2003, 2005; Locatelli et al., 2004; Hens et al., 2009; Hanagal and Dabade, 2013; Hanagal and Sharma, 2013) among other distributions, but have not yet been used in conjunction with the generalized gamma frailty effect.

In Section 2, the generalized gamma shared frailty model with non-informative right censoring is described. This is followed by a discussion on some basic properties of the baseline hazard functions of Gompertz, log-logistic, and generalized exponential distributions in Section 3. In Section 4, an outline of the estimation procedure is given and these models are compared using real life datasets in Section 5.

2. Shared Frailty Model with Censoring

Consider G groups of survival times with n_i observations (i = 1, 2… G) per cluster. For the j^{th} individual in the i^{th} cluster, the survival time, censoring indicator, and the vector of s known covariates are denoted by t_{ij}, δ_{ij}, and $\boldsymbol{X_{ij}}$ respectively.

The survival times considered in this paper are right censored, that is, t_{ij} is the true survival time (T_{ij}) when the event is observed and is the censoring time (C_{ij}) when the event is not observed due to any reason such as loss in follow-up or failure to complete the study. In other words,

$$\delta_{ij} = \begin{cases} 1, & \text{if } t_{ij} = T_{ij}, \text{that is, event is observed} \\ 0, & \text{if } t_{ij} = C_{ij}, \text{that is, event is censored.} \end{cases}$$

Associated with each cluster is the shared frailty effect z_i, which acts multiplicatively on the hazard function, increasing or decreasing the hazard for the j^{th} individual in the i^{th} cluster. The conditional hazard function for j = 1, 2… n_i, i = 1, 2… G, is given by

$$h\left(t_{ij}\middle|z_i, \boldsymbol{X_{ij}}\right) = z_i\, h_0\left(t_{ij}\right) e^{\boldsymbol{X_{ij}'\beta}}, \tag{1}$$

International Journal of Mathematical, Engineering and Management Sciences
Vol. 4, No. 1, 219–231, 2019
https://dx.doi.org/10.33889/IJMEMS.2019.4.1-019

where $\boldsymbol{\beta}$ is a vector of *s* regression coefficients corresponding to the known covariates in the study.

The corresponding conditional survival function is written as

$$S(t_{ij}|z_i, \boldsymbol{X_{ij}}) = \boldsymbol{e}^{-z_i H_0(t_{ij}) e^{X'_{ij}\boldsymbol{\beta}}}, \quad (2)$$

where $H_0(t_{ij})$ is the cumulative hazard function.

Under the assumption that the event times and censoring times are independent and that the censoring in the study is non-informative, the conditional likelihood for the frailty model is given as

$$L = \prod_{i=1}^{G} \prod_{j=1}^{n_i} \left[h(t_{ij}|z_i, \boldsymbol{X_{ij}})\right]^{\delta_{ij}} S(t_{ij}|z_i, \boldsymbol{X_{ij}}).$$

Integrating *L* over the entire range of the frailty variable *Z*, the unconditional likelihood function $\boldsymbol{L}$, is obtained as

$$\boldsymbol{L} = \int \int \ldots \int \prod_{i=1}^{G} \prod_{j=1}^{n_i} \left[h(t_{ij}|z_i, \boldsymbol{X_{ij}})\right]^{\delta_{ij}} S(t_{ij}|z_i, \boldsymbol{X_{ij}}) f(z_i) dz_i \quad (3)$$

where $f(.)$ denotes the PDF of the generalized gamma frailty distribution (*GGD (b, d, k)*) given by

$$f(z: b, d, k) = \frac{d}{\Gamma(k) b} \left(\frac{z}{b}\right)^{dk-1} e^{\left(\frac{z}{b}\right)^d}, \quad z, b, d, k > 0. \quad (4)$$

Note that *d* and *k* are the shape parameters and *b* is the scale parameter of the GGD. To make the parameters of the model identifiable, we set the mean of the frailty parameter equal to 1. Hence,

$$E(z) = \frac{b\,\Gamma\left(k + \frac{1}{d}\right)}{\Gamma(k)} \Rightarrow b = \frac{\Gamma(k)}{\Gamma\left(k + \frac{1}{d}\right)}.$$

This gives

$$V(z) = \frac{\Gamma(k)\Gamma\left(k + \frac{2}{d}\right)}{\Gamma\left(k + \frac{1}{d}\right)^2} - 1.$$

3. Baseline Distributions

3.1 Gompertz Distribution

The Gompertz distribution has been widely used in survival studies, especially in demographical or biological areas. In actuarial studies, it has been used to describe an exponentially increasing mortality with age. Its PDF, hazard function, and cumulative hazard function are given by

International Journal of Mathematical, Engineering and Management Sciences
Vol. 4, No. 1, 219–231, 2019
https://dx.doi.org/10.33889/IJMEMS.2019.4.1-019

$$f(t) = \omega e^{\alpha t} e^{-\frac{\omega}{\alpha}[e^{\alpha t}-1]}, \qquad \omega > 0, t > 0,$$
$$h(t) = \omega e^{\alpha t},$$
$$H(t) = \frac{\omega}{\alpha}[e^{\alpha t} - 1].$$

The hazard function for this distribution increases from ω at $t = 0$ to ∞ as $t \to \infty$ for $\alpha > 0$ and is applicable in studies where the hazard is increasing at an exponential rate with respect to time. For $\alpha < 0$, the distribution has a decreasing hazard which converges to $-\omega/\alpha$ as time increases. This implies that a certain proportion of the population never experiences the event. While this may be useful in certain cases like cure rate models, but in this paper, $\alpha > 0$ has been considered.

Although, the Gompertz distribution provides a close fit to adult mortality in developed countries, this distribution cannot be used where the risk of failure is fairly constant as $t \to \infty$. Such phenomenon is experienced often in clinical trials relating to the treatment of a disease or in reliability studies of machines.

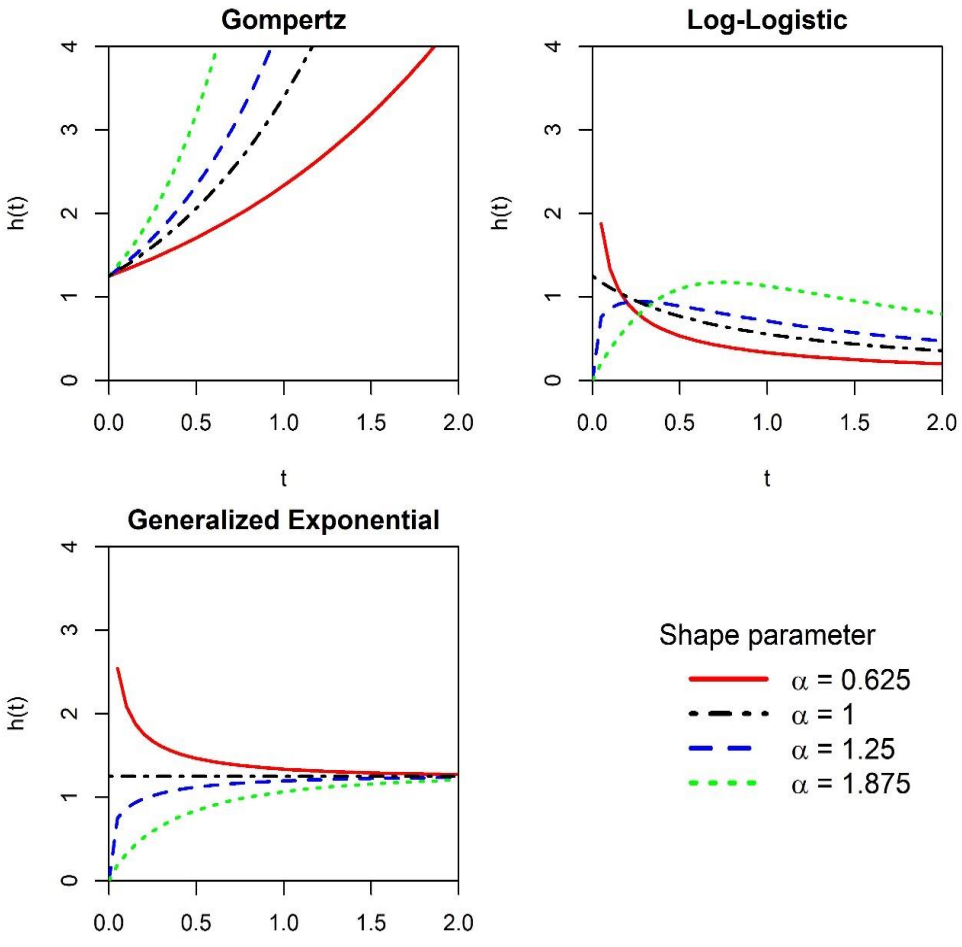

Figure 1. Hazard function plots for Gompertz, log-logistic, and generalized exponential distribution for $\omega = 1.25$ and varying shape parameter α

International Journal of Mathematical, Engineering and Management Sciences
Vol. 4, No. 1, 219–231, 2019
https://dx.doi.org/10.33889/IJMEMS.2019.4.1-019

3.2 Log-Logistic Distribution

The two parameter log-logistic (LL) distribution has a fairly flexible hazard function as compared to the Gompertz distribution as shown by Figure 1. The PDF, hazard function, and cumulative hazard function are given by

$$f(t) = \frac{\alpha\omega(\omega t)^{\alpha-1}}{[1 + (\omega t)^{\alpha}]^2}, \qquad \alpha, \omega, t > 0,$$
$$h(t) = \frac{\alpha\omega(\omega t)^{\alpha-1}}{1 + (\omega t)^{\alpha}},$$
$$H(t) = \log[1 + (\omega t)^{\alpha}].$$

Its hazard function is either decreasing or hump shaped as the shape parameter α varies. It fits datasets that have a hazard rate that is either initially high or increases rapidly at the beginning of the study to a finite maximum and later gradually reduces as $t \to \infty$.

Even though it is one of the closest alternatives to the Weibull distribution, this distribution cannot be used in cases where the dataset has an increasing hazard throughout the study period.

3.3 Generalized Exponential Distribution

The generalized exponential (GE) distribution has been considered as the third baseline distribution in this article due to its simplicity in analysing data with an increasing, decreasing or flat hazard rate. The corresponding PDF, hazard function, and cumulative hazard function are given by

$$f(t) = \alpha\omega e^{-\omega t}[1 - e^{-\omega t}]^{\alpha-1}, \qquad \alpha, \omega, t > 0,$$
$$h(t) = \frac{\alpha\omega e^{-\omega t}[1 - e^{-\omega t}]^{\alpha-1}}{1 - [1 - e^{-\omega t}]^{\alpha}},$$
$$H(t) = -\log[1 - (1 - e^{-\omega t})^{\alpha}].$$

The hazard function for GE distribution varies when the shape parameter α changes. At $\alpha = 1$, the hazard is constant and equal to ω, while it increases from 0 to ω for $\alpha > 1$ and decreases from ∞ to ω for $\alpha < 1$. The main advantage of using this distribution to model the baseline hazard is that it incorporates a case of a constant hazard with respect to time, which may be true for the common effect in some survival studies.

4. Estimation Procedure

4.1 Likelihood Function

After replacing the conditional hazard and conditional survival functions (expressions (1) and (2) respectively) in the likelihood function given by (3), and using Gauss Laguerre quadrature rule, $\boldsymbol{L}$ reduces to

$$\boldsymbol{L} = \prod_{i=1}^{G}\left\{\left(\prod_{j=1}^{n_i}[h_0(t_{ij})\eta_{ij}]^{\delta_{ij}}\right) I_i\right\}. \tag{5}$$

The integral I_i is approximated as

International Journal of Mathematical, Engineering and Management Sciences
Vol. 4, No. 1, 219–231, 2019
https://dx.doi.org/10.33889/IJMEMS.2019.4.1-019

$$I_i \approx \frac{1}{A_i^{D_i+1}} \sum_n w_n u_n^{D_i} f\left(\frac{u_n}{A_i}\right) \tag{6}$$

where, $\eta_{ij} = e^{\boldsymbol{X}_{ij}'\boldsymbol{\beta}}$, $D_i = \sum_{j=1}^{n_i} \delta_{ij}$ and $A_i = \sum_{j=1}^{n_i} H_0(t_{ij})\eta_{ij}$.

The $h_0(t_{ij})$ and $H_0(t_{ij})$ are defined in sub-sections 3.1, 3.2, and 3.3 for the three models, that is, Gompertz, log-logistic, and generalized exponential baseline, under a generalized gamma frailty distribution. In the sequel, these models shall be referred to as Models I, II, and III respectively.

4.2 MCMC Algorithm for Bayesian Estimates

In the Bayesian method of estimation, the parameters of the model are considered to be random variables. The dataset in the form of the likelihood function $\boldsymbol{L}(x|\theta)$, is clubbed with the prior information (prior density) $p(\theta)$, available about the parameter. This gives the posterior density $\pi(\theta|x)$, where

$$\pi(\theta|x) \propto \boldsymbol{L}(x|\theta)p(\theta).$$

Since we generally have limited information about the parameters of the model, prior distributions are taken to be flat by using distributions with high variances. The Normal distribution, *N* (0, 1000) is used for regression coefficients while Gamma (0.001, 0.001) is used for all other non-negative parameters.

Assuming that all parameters of the model are independently distributed and using the likelihood function given by (5) in conjunction with prior densities, we obtain the conditional posterior densities of each parameter as

$$\pi(\alpha|\omega,\boldsymbol{\beta},d,k) \propto \prod_{i=1}^{G}\left\{\left(\prod_{j=1}^{n_i}\left[h_0(t_{ij})\right]^{\delta_{ij}}\right) I_i\right\} p(\alpha),$$

$$\pi(\omega|\alpha,\boldsymbol{\beta},d,k) \propto \prod_{i=1}^{G}\left\{\left(\prod_{j=1}^{n_i}\left[h_0(t_{ij})\right]^{\delta_{ij}}\right) I_i\right\} p(\omega),$$

$$\pi(d|\alpha,\omega,\boldsymbol{\beta},k) \propto \prod_{i=1}^{G} I_i p(d),$$

$$\pi(k|\alpha,\omega,\boldsymbol{\beta},d) \propto \prod_{i=1}^{G} I_i p(k),$$

$$\pi(\beta_r|\alpha,\omega,\boldsymbol{\beta_r},k) \propto \prod_{i=1}^{G}\left\{\left(\prod_{j=1}^{n_i}\left[\eta_{ij}\right]^{\delta_{ij}}\right) I_i\right\} p(\beta_r).$$

Here $\boldsymbol{\beta_r} = (\beta_1, \beta_2, .., \beta_{r-1}, \beta_{r+1}, .. \beta_s)$ and $p(.)$ are the prior densities of the model parameters.

International Journal of Mathematical, Engineering and Management Sciences
Vol. 4, No. 1, 219–231, 2019
https://dx.doi.org/10.33889/IJMEMS.2019.4.1-019

In order to obtain the estimates of the parameters, a sample is drawn using the Metropolis-Hastings Algorithm (Metropolis and Ulam, 1949; Metropolis et al., 1953; Hastings, 1970) from each posterior density and conclusions are drawn.

Detailed derivation of the expression in sub-section 4.1 and the algorithm for estimation in sub-section 4.2 have been given in Sidhu et al. (2018).

4.3 Model Comparison

In the presence of numerous choices to model datasets, it becomes important to be able to compare different models to choose the one that provides the best fit. For this purpose, the Akaike Information Criteria (Akaike, 1974), Bayesian Information Criteria (Schwarz, 1978), Deviance Information Criteria (Spiegelhalter et al., 2002), and Bayes Factor are used.

For *p* number of parameters in the model and *n* number of observations in the dataset, the Akaike Information Criteria (*AIC*), Bayesian Information Criteria (*BIC*), and Deviance Information Criteria (*DIC*) are defined as

$$AIC = -2\log L\left(x|\widehat{\boldsymbol{\theta}}\right) + 2p,$$
$$BIC = -2\log L\left(x|\widehat{\boldsymbol{\theta}}\right) + \log(n)\, p,$$
$$DIC = -2\log L\left(x|\widehat{\boldsymbol{\theta}}\right) + 2p_D.$$

where

$$p_D = E[-2logL(x|\boldsymbol{\theta})] - \left[-2logL\left(x|\widehat{\boldsymbol{\theta}}\right)\right].$$

The models with lower values of *AIC*, *BIC*, and *DIC* are preferred. Generally, these three methods give concurrent results. However, *AIC* penalizes the number of parameters less strongly than *BIC* and may not be preferable.

DIC incorporates the effective number of parameters p_D and is frequently used when the estimates are obtained using the output from a Markov chain. However, *DIC* suffers from theoretical drawbacks and may not be a suitable choice (Spiegelhalter et al., 2014).

The Bayes factor (*BF*) as a model choice criterion is defined as

$$BF_{10} = \frac{P(x|M_1)}{P(x|M_0)}$$

where

$$P(x|M_i) = \int P(x|\boldsymbol{\theta}, M_i)\pi(\boldsymbol{\theta}|\, M_i)\, d\boldsymbol{\theta}\,.$$

Although, $(2\ log\ BF_{10})$ is approximately equal to the difference in the *BIC* values for the models, we use the method given by Kass and Raftery (1995), to compute $P(x|M)$ from the MCMC sample obtained for each of the parameters in the model.

International Journal of Mathematical, Engineering and Management Sciences
Vol. 4, No. 1, 219–231, 2019
https://dx.doi.org/10.33889/IJMEMS.2019.4.1-019

$$P(x|M) \approx \left(\frac{\sum_{i=1}^{N} L(x|\boldsymbol{\theta}^{(i)})^{-1}}{N}\right)^{-1}$$

where $\boldsymbol{\theta}^{(i)}$ is posterior sample of size *N* obtained using the Metropolis-Hastings algorithm from posterior densities in sub-section 4.2, for the vector of parameters $\theta = (\alpha, \omega, \boldsymbol{\beta}, d, k)$ for the model *M*.

A value of more than 10 for $(2 \log BF_{10})$ indicates an extremely strong positive evidence to favour M_1 over M_0 while a value between 0 and 2 is insufficient evidence to favour either model (Kass and Raftery, 1995). A value between 2 - 6 or 6 - 10 indicates a mild or a moderately strong evidence respectively, to prefer the numerator model.

5. Applications

To demonstrate the use of Models I, II, and III, the Metropolis-Hastings algorithm is applied to the Catheter Infection (CI) dataset by McGilchrist and Aisbett (1991), chronic granulotomous disease (CGD) dataset by Fleming and Harrington (2011), and the tumorigenesis study on rats (RATS) by Mantel et al. (1977).

CI dataset contains recurrent infection times from the use of catheters for 38 patients using a portable dialysis machine. The two infection times per patient are grouped together in a cluster. Other information available is the censoring status of each infection time, patient's age, gender (0 - male, 1 - female), and disease (Glomerulo Nephritis (GN), Acute Nephritis (AN), and Polycystic Kidney Disease (PKD)).

CGD dataset contains time to first serious infection from the use of gamma interferon in chronic granulotomous disease. There are 128 observations recorded at 13 centres along with their treatment status (placebo or rIFN-g), sex (0 - female, 1 - male), age, pattern of inheritance (0 - X-linked, 1 - Autosomal Recessive), use of corticosteroids (0 - Used, 1 - Did not use), and use of prophylactic antibiotics (0 - Used, 1 - Did not use). We group the dataset on the basis of the institute of treatment, although this data can be grouped together on the basis of multiple other factors like institution category, recurrent times, or pattern of inheritance.

RATS dataset consists of times to appearance of the tumor in three rats (one treatment and two control) that belong to the same litter. The original dataset had grouping on the basis of gender as well, but the male population was heavily censored. Hence, we consider only the 50 female litters for analysis in this article. The data available are treatment status (0 - control, 1 - Treatment) and censoring indicator.

Table 1. Predicted hazard function plots for the datasets modelled on a Weibull baseline under a shared gamma frailty

Dataset	CI	CGD	RATS
α	1.16	0.96	3.93
σ_{FR}^2	0.2820	0.1238	0.4888

To illustrate how different these datasets are, we obtain preliminary estimates of the parameters using "frailtypack" package in R (Rondeau et al., 2012). The built-in function "frailtyPenal" is

International Journal of Mathematical, Engineering and Management Sciences
Vol. 4, No. 1, 219–231, 2019
https://dx.doi.org/10.33889/IJMEMS.2019.4.1-019

used where Weibull distribution is taken to model the baseline hazard and the shared frailty is assumed to follow gamma distribution. Table 1 shows the estimated value of α, the shape parameter of the Weibull distribution and σ_{FR}^2, the variance of the frailty term. For the Weibull distribution, $\alpha > 1$ indicates an increasing hazard while $\alpha < 1$ points to a decreasing hazard with respect to time. As per Table 1, the estimated frailty variance signifies a low to moderate degree of heterogeneity among the clusters. Figure 2. shows the predicted hazard function for the three datasets.

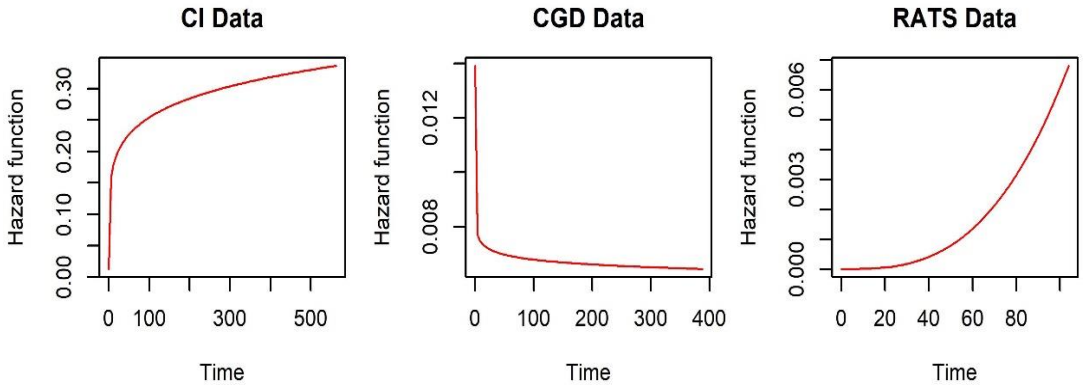

Figure 2. Predicted hazard function plots for the datasets modelled on a Weibull baseline under a shared gamma frailty

As per Figure 2, the hazard functions for CI data and RATS data are both increasing with respect to time, hence GE and Gompertz distributions can be used to model the baseline effect. CGD data indicates a decreasing hazard with time and ideally log-logistic baseline should give the best fit. In order to find the most appropriate baseline for these three datasets, they are fitted to Models I, II, and III using the estimation procedure mentioned in Section 4.

Table 2. Parameter estimates (credible intervals) of CI dataset under the generalized gamma shared frailty model

Baseline	Model I Gompertz	Mode II LL	**Model III** **GE**
α	0.0032	1.598	1.595
β_1 (Age)	0.0093 (-0.0178, 0.0442)	0.0146 (-0.0057,0.0397)	0.0073 (-0.0198,0.0353)
β_2 (Gender)	-1.787 (-2.814, -0.8454)	-1.2660 (-2.009, -0.7155)	-1.889 (-2.782,-1.006)
β_3 (GN)	0.2977 (-0.873,1.335)	0.1708 (-0.554, 0.9953)	0.2298 (-0.8304,1.2363)
β_4 (AN)	0.6979 (-0.4713,1.8834)	0.4095 (-0.4480,1.3316)	0.6062 (-0.4675, 1.6151)
β_5 (PKD)	-0.5774 (-2.714,1.1854)	-0.7581 (-1.9990,0.6649)	-0.7449 (-2.340,0.8435)
σ_{FR}^2	0.5465 (0.0560,1.8355)	0.1325 (0.0116,0.5396)	0.4065 (0.0531,1.1790)
AIC	700.92	695.66	**681.51**
BIC	721.89	716.64	**702.49**
DIC	**655.50**	668.84	669.26

International Journal of Mathematical, Engineering and Management Sciences
Vol. 4, No. 1, 219–231, 2019
https://dx.doi.org/10.33889/IJMEMS.2019.4.1-019

Table 2 lists the estimated parameters and the credible intervals of the covariates for CI dataset. The value of $\hat{\alpha}$, shows that the population hazard increases with time. Gender is a significant factor affecting hazard as its credible interval does not contain zero. Negative value for $\widehat{\beta_2}$ indicates a lower hazard for the female population.

Table 3. Parameter estimates (credible intervals) of CGD dataset under the generalized gamma shared frailty model

Baseline	Model I Gompertz	**Mode II** **LL**	Model III GE
α	0.001	1.0140	0.9080
β_1 (Treatment)	-1.116 (-1.679,-0.6316)	-1.0040 (-1.5760,-0.5075)	-1.0510 (-1.6197,-0.5180)
β_2 (Sex)	1.0880 (0.399,1.9349)	1.2650 (0.4722,2.1237)	1.0250 (-0.0122,2.0490)
β_3 (Age)	-0.0355 (-0.0665,-0.0072)	-0.0290 (-0.0539,-0.0013)	-0.0337 (-0.0671,-0.0027)
β_4 (Inherit)	0.8008 (0.2111,1.3690)	0.8023 (0.2679,1.3519)	0.6940 (0.0165,1.3086)
β_5 (Steroids)	1.594 (0.5615,2.7660)	1.526 (0.2675,2.7350)	1.489 (-0.1157,2.6469)
β_6 (Propylac)	-0.5649 (-1.3170,0.1767)	-0.4338 (-1.0990,0.2220)	-0.5850 (-1.3800,0.7190)
σ_{FR}^2	0.2525 (0.0129,1.1271)	0.2117 (0.0065,0.7212)	0.2097 (0.0100,0.7190)
AIC	171.76	**170.29**	207.20
BIC	204.89	**203.42**	240.34
DIC	121.64	115.19	**99.89**

Table 3, presents the results from CGD dataset fitted to the three models. As per the results, the factors treatment, age, and inherit significantly affect the hazard rate. While the hazard is lower for individuals under treatment and older in age, there is an increased hazard rate for individuals with Autosomal Recessive Inheritance. Additionally, Models I and II also indicate a significantly higher risk for the female patients and those not using corticosteroids. As there is a difference in the results of these models, it becomes important to look at the results of the model comparison. Both *AIC* and *BIC* indicate that Model II is the best model for the dataset, hence the two factors sex and steroids should also be considered as significant factors affecting survival times.

Table 4. Parameter estimates (credible intervals) of RATS dataset under the generalized gamma shared frailty model

Baseline	**Model I** **Gompertz**	Mode II LL	**Model III** **GE**
α	0.0529	4.018	6.536
β_1 (Treatment)	1.0220 (0.3741,1.7337)	1.0550 (0.3240,1.6295)	1.0950 (0.3691,1.5710)
σ_{FR}^2	0.5940 (0.0774,2.1016)	0.25320 (0.02250,1.2438)	0.4618 (0.0200,0.7866)
AIC	498.57	487.511	**487.16**
BIC	513.63	502.56	**501.02**
DIC	487.50	488.20	**482.63**

International Journal of Mathematical, Engineering and Management Sciences
Vol. 4, No. 1, 219–231, 2019
https://dx.doi.org/10.33889/IJMEMS.2019.4.1-019

As per the results reported in Table 4 for RATS dataset, the treatment given significantly increases the hazard rate thus implying an increased risk of tumors with the treatment.

The best model as per *AIC* and *BIC* values for CI and RATS datasets (Tables 2 and 4) is the GE baseline while LL baseline is best for CGD dataset (Table 3). *DIC* values however, do not give us the same conclusions. Also, none of the *AIC*, *BIC*, or *DIC* values indicate whether one model is significantly better than the other. For this reason, Table 5 listing the values of (2*logBF*) for a pair of models is considered. The table lists only positive values as the reciprocal model is the negative of the value listed, that is, $(2logBF_{01}) = -(2logBF_{10})$.

Table 5. $2logBF_{10}$ values for comparing M_1/M_0

Dataset	Numerator / Denominator	Gompertz	LL	GE
CI Data	Gompertz	-	*	3.24
	LL	6.62	-	**9.87**
	GE	*	*	-
CGD Data	Gompertz	-	68.58	*
	LL	*	-	*
	GE	65.33	**133.91**	-
RATS Data	Gompertz	-	*	0.06
	LL	3.46	-	**3.52**
	GE	*	*	-

* Indicates a negative value for $(2logBF_{10})$
\- Combinations where numerator and denominator model is same, model comparison invalid

The values in Table 5 indicate that, for CI data, the GE and Gompertz baseline provides a significantly better fit than the LL baseline as the corresponding values are greater than 6 (6.62 and 9.87). However, GE baseline is only slightly better than the Gompertz baseline as the corresponding value (3.24) falls between 2 to 6.

For CGD dataset, all positive values are greater than 10 indicating a marked difference in the fit provided by the three baselines with LL being the best model with value of 68.58 when compared with Gompertz and 133.91 when compared with GE. GE provides the worst fit among the three with Gompertz being favourable among GE and Gompertz with a value of 65.33.

For RATS dataset however, GE and Gompertz baselines provide a slightly better fit than LL baseline with corresponding values of 3.52 and 3.46. However, both GE and Gompertz provide a comparable fit (value falling between 0 and 2) with GE/Gompertz = 0.06.

When the variance of the shared frailty term (σ_{FR}^2) under gamma distribution (Table 1) is compared with the best models suggested under the generalized gamma distribution (Models in bold font in Tables 2-4), it appears that the variance tends to be underestimated under gamma frailty. The models under different frailty distributions can also be compared using similar methodology to lend credence to an improved variance estimation under the generalized gamma

distribution when using Bayesian methods, as has been performed earlier for the estimation under classical inference (Balakrishnan and Peng, 2006; Chen et al., 2013). Heterogeneity among the clusters indicates a source of variation in the hazard that has not been accounted for in the study. Since this can impact the results of the study, it is important for researchers to be able to diagnose the source of variation in order to have more meaningful results.

6. Conclusions

In this paper, the use of the Gompertz, log-logistic and generalized exponential distribution to model the baseline hazard under a generalized gamma shared frailty effect has been studied. Bayesian estimation is possible for all models and has been illustrated with three real life applications. Model comparison is performed using the Akaike Information Criteria, Bayesian Information Criteria, Deviance Information Criteria, and Bayes Factor and the best model is suggested for each dataset. The variance of the frailty term is also found to be underestimated when the shared frailty effect follows the gamma distribution.

Conflict of Interest

All authors have contributed equally in this work. The authors declare that there is no conflict of interest for this publication.

Acknowledgement

The first author would like to thank the University Grants Commission, Govt. of India, for providing financial support. The authors also acknowledge the aid given by Department of Science and Technology, Govt. of India, under PURSE grant.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723.

Balakrishnan, N., & Peng, Y. (2006). Generalized gamma frailty model. *Statistics in Medicine*, *25*(16), 2797-2816.

Chen, P., Zhang, J., & Zhang, R. (2013). Estimation of the accelerated failure time frailty model under generalized gamma frailty. *Computational Statistics & Data Analysis*, *62*, 171-180.

Fleming, T. R., & Harrington, D. P. (2011). *Counting processes and survival analysis* (Vol. 169). John Wiley & Sons.

Hanagal, D. D., & Dabade, A. D. (2013). A comparative study of shared frailty models for kidney infection data with generalized exponential baseline distribution. *Journal of Data Science*, *11*(1), 109-142.

Hanagal, D. D., & Sharma, R. (2013). Modeling heterogeneity for bivariate survival data by shared gamma frailty regression model. *Model Assisted Statistics and Applications*, *8*(2), 85-102.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97-109.

Hens, N., Wienke, A., Aerts, M., & Molenberghs, G. (2009). The correlated and shared gamma frailty model for bivariate current status data: An illustration for cross-sectional serological data. *Statistics in Medicine*, *28*(22), 2785-2800.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773-795.

Khodabin, M., & Ahmadabadi, A. (2010). Some properties of generalized gamma distribution. *Mathematical Sciences*, *4*(1), 9-28

Klein, J. P., Pelz, C., & Zhang, M. J. (1999). Modeling random effects for censored data by a multivariate normal regression model. *Biometrics*, *55*(2), 497-506.

Locatelli, I., Lichtenstein, P., & Yashin, A. I. (2004). The heritability of breast cancer: a Bayesian correlated frailty model applied to Swedish twins data. *Twin Research and Human Genetics*, *7*(2), 182-191.

Mantel, N., Bohidar, N. R., & Ciminera, J. L. (1977). Mantel-Haenszel analyses of litter-matched time-to-response data, with modifications for recovery of interlitter information. *Cancer Research*, *37*(11), 3863-3868.

Manton, K. G., Stallard, E., & Vaupel, J. W. (1986). Alternative models for the heterogeneity of mortality risks among the aged. *Journal of the American Statistical Association*, *81*(395), 635-644.

McGilchrist, C. A., & Aisbett, C. W. (1991). Regression with frailty in survival analysis. *Biometrics*, 461-466.

Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, *44*(247), 335-341.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087-1092.

Rondeau, V., Mazroui, Y., & Gonzalez, J. R. (2012). Frailtypack: an R package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *Journal of Statistical Software*, *47*(4), 1-28.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461-464.

Sidhu, S., Jain, K., & Sharma, S. K. (2018). Bayesian estimation of generalized gamma shared frailty model. *Computational Statistics*, *33*(1), 277-297.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), *64*(4), 583-639.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), *76*(3), 485-493.

Wienke, A., Arbeev, K. G., Locatelli, I., & Yashin, A. I. (2003). A simulation study of different correlated frailty models and estimation strategies. *Technical Report, MIPDR Working Paper WP*.

Wienke, A., Arbeev, K. G., Locatelli, I., & Yashin, A. I. (2005). A comparison of different bivariate correlated frailty models and estimation strategies. *Mathematical Biosciences*, *198*(1), 1-13.