Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v10n1p82

Estimating unknown heterogeneity in head and neck cancer survival: a parametric shared frailty approach

By Sen et al.

Published: 26 April 2017

Electronic Journal of Applied Statistical Analysis
Vol. 10, Issue 01, April 2017, 82-92
DOI: 10.1285/i20705948v10n1p82

Estimating unknown heterogeneity in head and neck cancer survival: a parametric shared frailty approach

Subhradev Sen*[a], Meenu Tom[b], Geetha M.[c], and Satheesan B.[d]

[a]*School of Business, Alliance University, Bangalore, India*
[b]*Department of Biostatistics, St. Thomas College, Palai, Kerala, India*
[c]*Department of Radiation Oncology, Malabar Cancer Centre, Thalassery, Kerala, India*
[d]*Department of Surgical Oncology, Malabar Cancer Centre, Thalassery, Kerala, India*

Published: 26 April 2017

The term frailty was introduced by Vaupel et al., 1979, to indicate that different individuals are at risks even though on the surface they may appear to be quite similar with respect to the measurable attributes such as age, sex, habits etc. The term frailty can be utilized to represent an unobservable random effect shared by subjects with similar risks in the analysis of time to event data and/or mortality rates. In this article, we make use of the parametric shared frailty models to a real life data for identifying the distributional form of baseline hazard function. The gamma shared frailty, with disease stages as clusters, with log-logistic baseline hazard model came out to be the best choice for modeling survival data of Head and Neck cancer patients treated with radiotherapy. The suitability of the best-chosen model is justified considering two significant covariates, namely, age of the patients and habit of their alcohol consumption. We obtain the estimates of frailty (or unknown heterogeneity) for five stages of disease taken as clusters for Gamma- log-logistic shared frailty model.

keywords: Shared frailty, gamma distribution, hazard models, survival, head and neck cancer

*Corresponding author:subhradev.stat@gmail.com

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

1 Introduction

The Cox model (see Cox, 1972) is a well-recognized statistical technique for analyzing survival data. The Cox model is based on a modeling approach to the analysis of survival data. The purpose of the model is to simultaneously explore the effects of several variables on survival. When it is used to analyze the survival of patients in a clinical trial, the model allows us to isolate the effects of treatment from the effects of other variables. Importantly, correct inference based on those proportional hazards models needs independent and identically distributed samples.

Sometimes, subjects may be exposed to different risk levels, even after controlling for known risk factors; this is because some relevant covariates are often unavailable to the researcher or even unknown. Also, the study population may be divided into clusters so that subjects from the same cluster behave more cohesively than those in other clusters. In the medical field, *frailty* is a term that is used more frequently. It originates from gerontology where it is used to indicate that frail people have an increased risk for morbidity and mortality. In statistical literature, frailty is a random component designed to account for variability due to unobserved individual-level factors that is otherwise unaccounted for by the other predictors in the model, see for more insight Vaupel et al., 1979; Lancaster, 1979.

The Frailty model is the extension of the Cox proportional hazards models, which is suitable for the individual or clustered survival data. Frailty models are the survival models analog to regression models that account for heterogeneity and random effect. Univariate frailty models take into account that the population is not homogeneous. Heterogeneity may be explained by covariates, but when important covariates have not been observed, this leads to unobserved heterogeneity. The idea is to suppose that different patients possess different frailties and patients more *frail* (or *prone*) tend to have an intended event earlier than those who are less frail. A shared frailty model is a random effect model where the frailties are common among groups of individuals or clusters and are randomly distributed across groups.

In this article, we consider the parametric approach of shared frailty models where the variable frailty is assumed to have different probability distributions (such as gamma and inverse Gaussian) under the three different parametric baseline hazard functions viz. exponential, Weibull and log-logistic (see more details on shared frailty models Hougaard, 2012; Therneau and Grambsch, 2000; Duchateau and Janssen, 2007 and references therein).

Our objective is to find suitable parametric model for baseline hazard function under different distributional assumptions in frailty for modeling head and neck cancer survival data when clustered as per disease stages. The aim is to find the estimates for unknown heterogeneity (frailty) for different clusters (owing the fact that under same disease stage subjects experience similar frailness) under parametric shared frailty modeling.

Rest of the article is organized as follows. The basic concept of parametric shared frailty model is described in section 2. The gamma shared frailty and inverse Gaussian shared frailty models are described in subsections 2.1 and 2.2, respectively, while in subsection 2.3, discussion is made on baseline hazard assumptions. Section 3 deals with

the real data application of the parametric shared frailty models with known covariates while subsections 3.1 and 3.2 deals with the data description and results of the data analysis, respectively. Finally, section 4 concludes.

2 The parametric shared frailty models

In this section, we discuss about the basic concept and theoretical structure of parametric shared frailty models. We consider here the parametric baseline hazard assumption under known covariates.

Suppose there are n clusters and that cluster i has n_i observations and associates with the unobserved frailty $Z_i(1 \leq i \leq n)$. The vector $\mathbf{X}_{ij}(1 \leq i \leq n, 1 \leq j \leq n_i)$ contains the covariate information for event time T_{ij} of the j^{th} observation in the i^{th} cluster. The survival times in cluster $i(1 \leq i \leq n)$, conditional on the frailty term Z, are assumed to be independent and their hazard functions are of the form

$$\mu(t|\boldsymbol{X}_{ij}, Z_i) = Z_i \mu_0(t) \exp(\boldsymbol{\beta'} \boldsymbol{X}_{ij}) \tag{1}$$

where $\mu_0(t)$ denotes the baseline hazard functions and $\boldsymbol{\beta}$ denotes vector of the fixed effect parameter to be estimated. The frailties $Z_i(i = 1, 2, \ldots, n)$ are assumed to be independently and identically distributed random variables with density function $f(z)$. Various frailty distributions have been proposed in the literature (see for more comprehensive overview in this field Duchateau and Janssen, 2007; Van den Berg, 2001). The main assumption of a shared frailty model is that all individuals in cluster i share the same value of frailty $Z_i(i = 1, 2, \ldots, n)$. The lifetimes are assumed conditionally independent with respect to the shared (common) frailty. Within the clusters, this shared frailty is the cause of dependence between lifetimes. We can derive the joint conditional multivariate survival function for the individuals in the i^{th} cluster. Conditional on frailty Z_i which is shared by all individuals in cluster i, we have

$$\begin{aligned} S(t_{i1}, t_{i2}, \ldots, t_{in_i}|\mathbf{X}_i, Z_i) &= S(t_{i1}|\mathbf{X}_{i1}, Z_i) S(t_{i2}\mathbf{X}_{i2}, Z_i) \ldots S(t_{in_i}\mathbf{X}_{in_i}, Z_i) \\ &= \exp\left\{ -Z_i \sum_{j=1}^{n_i} H_0(t_{ij}) e^{\boldsymbol{\beta'} \boldsymbol{X}_{ij}} \right\} \end{aligned} \tag{2}$$

Where $H_0(t) = \int_0^t \mu_0(s)ds$ denotes the cumulative baseline hazard function and $\mathbf{X}_i = (\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{in_i})$ is the covariate matrix of the individuals in the i^{th} cluster.

Now, averaging expression in (2) with respect to Z_i, we get the marginal survival function as

$$\begin{aligned} S(t_{i1}, t_{i2}, \ldots, t_{in_i}|\mathbf{X}_i) &= E\left[S(t_{i1}, t_{i2}, \ldots, t_{in}|\mathbf{X}_i, Z_i\right] \\ &= E\left[\exp\left\{ -Z_i \sum_{j=1}^{n_i} H_0(t_{ij}) e^{\boldsymbol{\beta'} \boldsymbol{X}_{ij}} \right\} \right] \end{aligned}$$

Which can be further written as

$$S(t_{i1}, t_{i2}, \ldots, t_{in_i}|\mathbf{X}_i) = \mathcal{L}\left(\sum_{j=1}^{n_i} H_0(t_{ij}) e^{\boldsymbol{\beta}' \boldsymbol{X}_{ij}}\right) \tag{3}$$

where $\mathcal{L}(.)$ is the Laplace transformation of the frailty variable.
Thus, the multivariate survival function is expressed as the Laplace transform of the frailty distribution, evaluated at the cumulative baseline hazard. Assuming independence between clusters, the joint survival function for all event-time data is now the product of the survival functions of all the clusters and is given by

$$S(t_{11}, t_{12} \ldots, t_{nn_n}|\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) = \prod_{i=1}^{n} \mathcal{L}\left(\sum_{j=1}^{n_i} H_0(t_{ij}) e^{\boldsymbol{\beta}' \boldsymbol{X}_{ij}}\right)$$

The unconditional (univariate) survival functions can be expressed by means of the Laplace transform

$$\begin{aligned} S(t_{ij}|\mathbf{X}_{ij}) &= E\left[S(t_{ij}|\mathbf{X}_{ij}, Z_i\right] \\ &= E\left[\exp\left\{-Z_i H_0(t_{ij}) e^{\boldsymbol{\beta}' \boldsymbol{X}_{ij}}\right\}\right] \\ &= \mathcal{L}\left(H_0(t_{ij}) e^{\boldsymbol{\beta}' \boldsymbol{X}_{ij}}\right) \end{aligned} \tag{4}$$

2.1 Gamma shared frailty models

The standard assumption about frailty in shared frailty models is that it follows a gamma distribution. Gamma distribution is preferred due to its flexible mathematical properties and the simple form of Laplace transformation. Specific form of dependence between event times in clusters is implied by each frailty distribution, e.g., the gamma distribution models late dependence in shared frailty models (see Duchateau and Janssen, 2007).
Assuming for frailty a gamma distribution with expectation unity and variance σ^2, we get by using (3), multivariate survival function for the i^{th} cluster as

$$S(t_{i1}, t_{i2}, \ldots, t_{in_i}|\mathbf{X}_i) = \mathcal{L}\left(\sum_{j=1}^{n_i} H_0(t_{ij}) e^{\boldsymbol{\beta}' \boldsymbol{X}_{ij}}\right) = \left(1 + \sigma^2 \sum_{j=1}^{n_i} H_0(t_{ij}) e^{\boldsymbol{\beta}' \boldsymbol{X}_{ij}}\right)^{-\frac{1}{\sigma^2}} \tag{5}$$

Different parametric functions can be assumed for baseline hazard function, but theory of parameter estimation is developed in a very general situation. The regression parameters $\boldsymbol{\beta}$ and the variance of frailty σ^2 are the parameters usually estimated in parametric gamma frailty models. If we assume the $\boldsymbol{\theta}$ is the vector of unknown parameter involved in baseline hazard function μ_0, then to derive unconditional likelihood function we consider first the conditional likelihood in the case of n clusters of size $n_i (i = 1, 2, \ldots, n)$

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) = \prod_{i=1}^{n} \int_0^{\infty} \prod_{j=1}^{n_i} \left[z_i \mu_0(t_{ij}; \boldsymbol{\theta}) e^{\boldsymbol{\beta}' \mathbf{x}_{ij}}\right]^{\delta_{ij}} \exp\left\{-z_i H_0(t_{ij}; \boldsymbol{\theta}) e^{\boldsymbol{\beta}' \mathbf{x}_{ij}}\right\} f(z_i; \sigma^2) dz_i \tag{6}$$

with $f(z_i;\sigma^2) = \frac{z_i^{\frac{1}{\sigma^2}-1}\exp(-z_i/\sigma^2)}{\sigma^{2/\sigma^2}\Gamma(1/\sigma^2)}$ denoting the probability density function (PDF) of gamma distribution with mean unity and variance σ^2 and δ_{ij} as event indicator.
The expression in (6) can be written as

$$L(\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2) = \prod_{i=1}^{n} \frac{\left[z_i\mu_0(t_{ij};\boldsymbol{\theta})e^{\boldsymbol{\beta'}\mathbf{x}_{ij}}\right]^{\delta_{ij}}}{y_i^{\frac{1}{\sigma^2}+d_i}\sigma^{2/\sigma^2}\Gamma(1/\sigma^2)} \int_0^{\infty} (y_i z_i)^{1/\sigma^2+d_i-1}\exp(y_i z_i) y_i dz_i \tag{7}$$

where $y_i = \frac{1}{\sigma^2} + \sum_{j=1}^{n_i} H_0(t_{ij};\boldsymbol{\theta})e^{\boldsymbol{\beta'}\mathbf{x}_{ij}}$ and $d_i = \sum_{j=1}^{n_i}\delta_{ij}$, number of observed events in cluster i. Finally, we have the unconditional log-likelihood function of the shared gamma frailty model (see for details Duchateau and Janssen, 2007) as

$$\begin{aligned} L(\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2) = {} & \sum_{i=1}^{n}\left[d_i\ln\sigma^2 + \ln\Gamma(1/\sigma^2+d_i) - \ln\Gamma(1/\sigma^2)\right] - \sum_{i=1}^{n}(1/\sigma^2+d_i) \\ & \ln\left(1+\sigma^2\sum_{j=1}^{n_i}H_0(t_{ij};\boldsymbol{\theta})\exp\boldsymbol{\beta'}\mathbf{x_{ij}}\right) + \sum_{i=1}^{n}\sum_{j=1}^{n_i}\delta_{ij}\{\boldsymbol{\beta'}\mathbf{x_{ij}} + \ln\mu_0(t_{ij};\boldsymbol{\theta})\} \end{aligned} \tag{8}$$

In parametric shared gamma frailty model with known covariates, the unobserved frailty $Z_i (i = 1, 2, \ldots, n)$ in each cluster can be estimated using counting process (see Nielsen et al., 1992) as follows

$$\hat{Z}_i = \frac{1/\hat{\sigma^2} + \sum_{j=1}^{n_i}\delta_{ij}}{1/\hat{\sigma^2} + \sum_{j=1}^{n_i}H_0(t_{ij};\hat{\boldsymbol{\theta}})\exp\left(\hat{\boldsymbol{\beta'}}\mathbf{x}_{ij}\right)} \tag{9}$$

Here, $\hat{\sigma^2}$, $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$ are the estimates of σ^2, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, respectively.

2.2 Inverse Gaussian shared frailty models

As mentioned earlier, gamma distribution is used as a frailty distribution for several reasons, however, it has drawbacks (see Kheiri et al., 2007) for example it may weaken the effect of covariates. Alternative to the gamma distribution Hougaard, 1984 introduced inverse Gaussian as a frailty distribution.
The inverse Gaussian distribution has many similarities to standard Gaussian distribution (see Chhikara, 1988). Furthermore, it provides much flexibility in modeling, when early occurrences of failures are dominant in a lifetime distribution and its failure rate is expected to be non-monotonic.
In such situations, the inverse Gaussian distribution might provide a suitable choice for the lifetime model. Moreover, for the inverse Gaussian distribution the surviving population becomes more homogeneous with respect to time, whereas for gamma distribution the relative heterogeneity is constant. The inverse Gaussian distribution has unimodal density and is the member of exponential family. While its shape resembles the other

skewed density functions, such as lognormal and gamma (see Hanagal and Dabade, 2013 for more detailed discussion). These properties of inverse Gaussian distribution motivate us to use inverse Gaussian as frailty distribution in case of shared heterogeneity.
The similar theoretical line up is followed as given in equations, (3) and (4) above and the frailty distribution for the cluster $i(i = 1, 2, \ldots, n)$ is replaced with the inverse Gaussian distribution.

2.3 Baseline hazard function assumption

As indicated at the introduction, under the parametric approach, the baseline hazard is defined through a parametric function involved with a vector of its parameters and is estimated together with the regression coefficients, and the frailty parameter(s). Here we consider the exponential, Weibull and log-logistic as parametric distributions for the baseline hazard for each case of frailty assumption, viz. gamma and inverse Gaussian. Table 1 indicates the forms of the baseline hazard functions for the exponential, Weibull and log-logistic distributions.

Exponential hazard function. The constant hazard function, λ, is a consequence of

Table 1: Baseline hazard function assumption

Distribution	$\mu_0(t)$	$H_0(t) = \int_0^t \mu_0(s)ds$	Parameter space
Exponential	λ	λt	$\lambda > 0$
Weibull	$\lambda p t^{p-1}$	λt^p	$\lambda > 0, p > 0$
Log-logistic	$\frac{e^\alpha k t^{k-1}}{1+e^\alpha t^k}$	$\ln(1 + e^\alpha t^k)$	$\alpha \in \mathbb{R}, k > 0$

the memoryless property of the exponential distribution: the distribution of the subjects remaining survival time given that he or she has survived till time t does not depend on t. In other words, the probability of death in a time interval $[t, t + \delta t]$ does not depend on the starting point, t.

Weibull hazard function. For the Weibull distribution, the hazard function depends on t. We can see that, depending on whether p is greater than or less than 1, the hazard can increase or decrease with increasing t. This is often more realistic than the assumption of a constant hazard function (as in the exponential case). Since the exponential distribution is a special case of the Weibull with $p = 1$, one way of analyzing the hazard rate is to fit the (more general) Weibull model and then test whether $p = 1$.

Log-logistic hazard function. The log-logistic distribution has a fairly flexible functional form, it is one of the parametric survival time models in which the hazard function may be decreasing, increasing, as well as hump-shaped, that is it initially increases and then decreases, depending upon the values of α and k.
With these above considerations, we approach to find out the best possible combination for frailty and baseline hazard for the below-mentioned data set and thereby find the

estimates for frailty for each selected cluster (disease stage in our case). We consider Akaike information criteria (AIC), see Akaike, 1974, for model selection. Smaller the value of AIC better is the model. We apply R software to perform the statistical analysis. The detailed result with discussion is presented in the next section.

3 Application in real life data

As indicated at the introduction, in this section, we consider parametric shared frailty approach for modeling the data on cancer survival. The best model is reached through a search for best possible distribution for the baseline hazard under gamma frailty and inverse Gaussian frailty. We assume three popular lifetime distributions, viz. exponential, Weibull and log-logistic, in modeling baseline hazard for both the cases.

3.1 Head and Neck cancer survival data

The survival (time to death) data on 244 subjects with squamous cell carcinoma in four sites viz. Larynx, Oropharynx, Hypopharynx, and Nasopharynx treated at the Radiation Oncology department of Malabar Cancer Centre, Thalassery, India, from January 2010 to December 2013 are collected retrospectively from case record files. Latest survival status (death or alive) are captured with the last follow up time at the end of July 2015. Five different stages (stage-I, stage-II, stage-III, stage-IVA, stage-IVB) of disease are found while capturing the data. In this data collection, other factors such as age, sex, smoking and alcohol habits of the patients are considered as possible known cofactors, other than unknown heterogeneity shared stage wise, affecting the survival time.

3.2 Results and discussions

In this sub-section, we represent the results of parametric shared frailty model when applied to the above-mentioned data. First, after general discussion about the data, we present the results for gamma frailty model and reach the best possible combination of frailty and baseline hazard under known covariates and then we continue the same exercise changing the frailty assumption as inverse Gaussian distribution. Finally, the best model is found out comparing gamma frailty and inverse Gaussian frailty with judiciary chosen baseline hazard for each case and estimates for frailty is obtained for the best model. It is to be noted that, under different frailty assumption the best suited baseline hazard function may be different for the same data set but in our case it was not so, as we can see below.

In the data, we have in total 244 squamous cell carcinoma cases treated with radiotherapy comprising 223(91%) males and only 21(9%) female cases. We collected data for four sites in Head and Neck, viz. site-1 (Larynx) 34% cases, site-2 (Oropharynx) 23% cases, site-3 (Hypopharynx) 34% cases and site-4 (Nasopharynx) 9% cases. The average age of the patients is 59 years with standard deviation of 10 years, while the median age is found to be 60 years. The data shows 200(82%) cases are smokers and 151(62%) cases posses alcoholic habit. In terms of disease stage, we have 9% with stage-I, 16%

stage-II, 30% are stage-III, 38% are stage-IVA and 7% in stage-IVB. The 68(28%) cases experienced the event (death) and rests are censored. For analysis, the unit for the survival times is taken in days.

Now, we start with gamma frailty assumption with three different base line hazard functions as mentioned in the previous section. Initially, we considered three covariates, viz. age, alcohol habit and smoking habit of the patients, but it is observed that smoking habits do not contribute significantly in the models (see table 2 for details) and hence we proceed with only two significantly contributing covariates, age and alcohol habit (P-value less than or equal to 0.05 is taken as significant level).The result obtained is given in the table 3. From table 3, we see that log-logistic distribution is the logical

Table 2: Gamma shared frailty models, baseline hazards, covariates, AIC value

Frailty distribution	Baseline Hazard	Covariates	Estimate	SE	P-Value	AIC
Gamma	Exponential	Age	0.031	0.014	0.023	1181.884
		Alcohol	0.542	0.287	0.058	
		Smoking	-0.071	0.378	0.851	
	Weibull	Age	0.032	0.014	0.019	1181.645
		Alcohol	0.573	0.287	0.046	
		Smoking	-0.073	0.378	0.846	
	Log-logistic	Age	0.030	0.014	0.034	1181.323
		Alcohol	0.568	0.287	0.048	
		Smoking	-0.079	0.377	0.833	

choice (as AIC value is the least) for the baseline hazard function under gamma shared frailty with clustering done with respect to disease stage. The same exercise is performed changing the frailty distribution as inverse Gaussian distribution with known covariates age and alcohol habit. The result obtained is given in table 4.

From table 4, we see that again log-logistic distribution is the logical choice (as AIC value is the least in this case also) for the baseline hazard function under inverse Gaussian shared frailty with clustering done with respect to disease stage. Hence, for a data like ours we suggest to take log-logistic distribution as a favorable choice for baseline hazard in estimating unknown heterogeneity in survival analysis where natural cluster is given priority for similar risk experience.

With the above findings, we, now, proceed further to identify the best model considering baseline hazard as log-logistic failure rate function. We take, as previous, model selection criterion, AIC, and find the best model as depicted in table 5.

It is clear from table 5 that gamma shared frailty model with log-logistic baseline

Table 3: Gamma shared frailty models, baseline hazards, significant covariates, AIC value

Frailty distribution	Baseline Hazard	Covariates	Estimate	SE	P-Value	AIC
Gamma	Exponential	Age	0.030	0.013	0.023	1179.918
		Alcohol	0.522	0.265	0.049	
	Weibull	Age	0.032	0.013	0.019	1179.682
		Alcohol	0.552	0.266	0.038	
	Log-logistic	Age	0.030	0.014	0.035	1179.366
		Alcohol	0.545	0.266	0.040	

Table 4: Inverse Gaussian shared frailty models, baseline hazards, significant covariates, AIC value

Frailty distribution	Baseline Hazard	Covariates	Estimate	SE	P-Value	AIC
Inverse Gaussian	Exponential	Age	0.030	0.013	0.026	1180.862
		Alcohol	0.515	0.265	0.052	
	Weibull	Age	0.031	0.013	0.022	1180.737
		Alcohol	0.543	0.266	0.041	
	Log-logistic	Age	0.029	0.014	0.040	1180.404
		Alcohol	0.536	0.266	0.044	

Table 5: AIC values for Gamma and Inverse Gaussian frailty models with Log-logistic baseline hazard

Baseline Hazard	Frailty distribution	AIC
Log-logistic	Gamma	1179.366
	Inverse Gaussian	1180.404

hazard function comes out to be the best model in this situation as AIC value is lesser as compared to that in case of inverse Gaussian frailty assumption.
Now we get estimates of frailty for each cluster (disease stage) by applying the formula given in (9) and with log-logistic baseline hazard function consideration. The estimates obtained are given in table 6.

Table 6: Frailty estimates with disease stage as clusters

Disease Stage(cluster)	Frailty estimates
Stage-III	0.298
Stage-IVA	0.804
Stage-I	1.193
Stage-IVB	1.292
Stage-II	1.356

4 Conclusion

In this article, parametric shared frailty models are applied to a real life data for identifying the distributional form of baseline hazard function. When we applied parametric shared frailty approach considering stages of cancer as clusters, we identified that log-logistic distribution's hazard rate is a logical choice for baseline hazard function for the model taking gamma and inverse Gaussian distribution as individual choice for frailty distribution. Finally, the gamma shared frailty with log-logistic baseline hazard model came out to be the best choice for modeling survival data of head and neck cancer patients, with specifically mentioned sites, treated with radiotherapy.
The suitability of the best-chosen model is justified considering two significant covariates, namely age of the patients and habit of their alcohol consumption. We obtain the estimates of frailty (or unknown heterogeneity) for five stages of disease taken as clusters for Gamma- log-logistic shared frailty model. It is found that patients with stages IVB, II, I have estimates of frailty variable more than unity and they are more frail as compared to the patients with stages III and IVA, respectively.

Acknowledgment

The authors would like to thank the anonymous referees for their valuable suggestion and comments leading to an improvement in the earlier draft of the paper. The authors would also like to place their sincere thanks to the entire staffs and faculty members of Cancer Registry & Epidemiology department and Medical Records division of Malabar Cancer Centre, Thalassery, India, for their kind support in providing case files for capturing the

data electronically. The second author would like to thank Prof. K. K. Jose, Prof. Joy Jacob and Mr. Akhil Vijayan of St. Thomas College, Palai, India, for their kind support and encouragement during the study.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Chhikara, R. (1988). *The Inverse Gaussian Distribution: Theory: Methodology, and Applications*, volume 95. CRC Press.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, 34:187–220.

Duchateau, L. and Janssen, P. (2007). *The frailty model*. Springer Science & Business Media.

Hanagal, D. D. and 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.

Hougaard, P. (1984). Life table methods for heterogeneous populations: distributions describing the heterogeneity. *Biometrika*, 71(1):75–83.

Hougaard, P. (2012). *Analysis of multivariate survival data*. Springer Science & Business Media.

Kheiri, S., Kimber, A., and Meshkani, M. R. (2007). Bayesian analysis of an inverse gaussian correlated frailty model. *Computational statistics & data analysis*, 51(11):5317–5326.

Lancaster, T. (1979). Econometric methods for the duration of unemployment. *Econometrica: Journal of the Econometric Society*, pages 939–956.

Nielsen, G. G., Gill, R. D., Andersen, P. K., and Sørensen, T. I. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian journal of Statistics*, pages 25–43.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling survival data: extending the Cox model*. Springer Science & Business Media.

Van den Berg, G. J. (2001). Duration models: specification, identification and multiple durations. *Handbook of econometrics*, 5:3381–3460.

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454.