Journal of Statistical Research *ISSN 0256 - 422 X*
2018, Vol. 52, No. 1, pp. 1-17

GUIDANCE FOR PRACTITIONERS ON THE CHOICES OF SOFTWARE IMPLEMENTATION FOR FRAILTY MODELS: SIMULATIONS AND AN APPLICATION IN DETERMINING THE BIRTH INTERVAL DYNAMICS

MOHAMMAD EHSANUL KARIM*
School of Population and Public Health, University of British Columbia, 2206 East Mall Vancouver, BC V6T 1Z3; and Centre for Health Evaluation and Outcome Sciences (CHÉOS) St. Paul's Hospital, 588-1081 Burrard St, Vancouver, BC V6Z 1Y6, Canada.
Email: ehsan.karim@ubc.ca

JAHIDUR RAHMAN KHAN
Centre for Research and Action in Public Health, Health Research Institute Faculty of Health, University of Canberra, Australia.
Email: jkhan@isrt.ac.bd

SUMMARY

In clustered survival analysis applications, researchers frequently fit frailty models using parametric and nonparametric approaches to obtain the estimates for the parameters associated with the survival model covariates and heterogeneity (frailty). Availability of the off-the-shelve implementations and freely available `R` software packages makes it convenient for the practitioners to fit these complicated models easily. Even though there has been a couple of studies assessing the stability of the older packages (e.g., `survival`, `coxme`) under a variety of scenarios, some of the newer implementations (e.g., `frailtySurv`, `JM` and `parfm`) have not gone through similar rigorous assessment. It is worth evaluating these new software implementations, and comparing them with the older packages. In the current work, via simulations, we will examine the estimates from all of these popularly used software implementations under a variety of scenarios when the corresponding assumptions related to the baseline hazard and frailty distributions are misspecified. Additionally, true heterogeneity parameter, censoring patterns and number of clusters were varied in the simulations to assess respective impacts on the estimates. From these simulations, we observed that when there is a large number of clusters and mild censoring, Cox PH frailty models fitted using a newer semiparametric estimation technique (from the `frailtySurv` package) produced regression and heterogeneity parameter estimates that were associated with unusually large bias and variability. On the other hand, when the true heterogeneity parameter is substantially large, the Cox PH frailty models fitted using the `coxme` package were often producing highly variable estimates of the heterogeneity parameter. The simulation findings then guided our choice of appropriate frailty model in the context of determining the birth interval dynamics in Bangladesh.

Keywords and phrases: Survival; Clustering; Sensitivity analysis; Simulation; R.

AMS Classification: 62N86

* Corresponding author
© Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka 1000, Bangladesh.

1 Introduction

In a general survival analysis scenario, researchers fit time-to-event data accounting for possible censoring. Traditional survival analysis models, such as Weibull regression model and Cox's proportional hazards (Cox PH) model (Cox, 1972; Zhang, 2016; Fleming and Harrington, 1991; Andersen, 1993) require statistical independence between lifetimes under consideration. Weibull regression is a parametric regression model that requires baseline hazard function specification, whereas, Cox PH is a semi-parametric model that allows the hazard function to be unspecified, and models covariates through a regression model. The assumption of independence is violated when the collected data consist of clusters. Ignoring correlated nature of the data in Weibull and Cox PH regression models may lead to incorrect standard errors and consequently misleading inferences (Henderson and Oman, 1999; Therneau and Grambsch, 2000; Duchateau and Janssen, 2007). To correct for such cluster-dependency, a cluster-specific random effect is introduced in the survival models to account for the cluster-specific susceptibility to the new events (Aalen, 1988). The random effect, popularly known as the 'frailty' term in the model, accounts for the unexplained heterogeneity arising from the clustered event times. If these frailties are known, the survival times are conditionally independent. A regression parameter (associated with a covariate) in the frailty model is considered a fixed effect, and interpreted as the effect of changing a covariate on a subject's average response conditional on the cluster (Vaupel et al., 1979; Zeger et al., 1988; Albert, 1999; Kelly, 2004).

Over the years, many different approaches of fitting the frailty models are proposed in the literature. They are generally extensions of Cox PH or Weibull regression models, based on how researchers want to incorporate the baseline hazards in the analyses (Clayton, 1978; Sahu et al., 1997). They may also differ with respect to the distribution of frailties as well as estimation methods. Gamma and log-normal distributions are frequently assumed as frailty distributions, based on empirical evidence or mathematical convenience (Therneau and Grambsch, 2000; Abbring and Van Den Berg, 2007; Hougaard, 2012). Once the frailties are generated from either of these distributions, fitting frailty models requires complicated likelihood integration over these frailties. Using sophisticated statistical and computational techniques, statisticians have developed a number of ways to get estimates of the heterogeneity parameter (e.g., θ) as well as the parameters associated with the covariates (e.g., βs) from these frailty models (Klein, 1992; Nielsen et al., 1992; Guo and Rodriguez, 1992; Therneau et al., 2003; McGilchrist and Aisbett, 1991; Munda et al., 2012; Duchateau and Janssen, 2007; Wienke, 2010; Rizopoulos et al., 2008, 2009; Rizopoulos, 2010; Gorfine et al., 2006; Zucker et al., 2008). However, the performances of various approaches vary to some degree with respect to various scenarios considered. In the previous literature, various estimation procedures for frailty models were compared (Cortinas Abrahantes and Burzykowski, 2005), effects of sample size (Karim, 2008) and cluster size (Abdulkarimova, 2013) were assessed for various frailty models. Fortunately, empirical researchers do not need to worry about solving such complicated computational issues due to the availability of numerous freely available software packages.

Availability of the existing routines and a number of new software packages (e.g., freely available `R` packages) enables researchers to easily fit frailty models in survival modelling applications with clustered lifetime data. However, these software implementations of frailty models differ with respect to types of regression models used (e.g., parametric and semiparametric), the assumed base-

line hazard and frailty distributions for the estimation of the respective parameters as well as the optimization technique. In real-life applications, where true baseline hazard and frailty distributions are generally unknown, the choice of a parametric or a non-parametric baseline hazard function, as well as the assumed frailty distributions in the frailty model, is rather arbitrary (Khan et al., 2016; Mahmood et al., 2013). The ramifications of the potential violation of this assumption on the parameters associated with the survival model covariates (β's) and frailties (θ) are hard to assess empirically. Two previous studies have compared various statistical properties of a couple of software implementations (e.g., `R` packages `survival` and `coxme`) that offer to fit various frailty models (Kelly, 2004; Hirsch and Wienke, 2012).

In this work, using various simulation scenarios, with respect to the mis-specification of baseline hazard distributions and frailty distributions, we assessed how sensitive the estimates (β's and θ) are from the newer software implementations of these frailty models. For that, we have compared `frailtySurv`, `JM` and `parfm` as well as the previous packages. Using the same scenarios, we will also assess the impact of the varying the number of clusters, percentage of censoring, varying heterogeneity parameter and compare the findings with the existing literature. We will then use these simulation findings to guide our choice of appropriate frailty model in an empirical context of determining the birth interval dynamics in Bangladesh (Khan et al., 2016).

2 Methods

Mathematically, to impose a frailty parameter in a Cox model, we need to add a new random effect (ν), i.e., an unobserved random variable that is implicitly common for all the observations in the same cluster and it's log transformation is assumed to be randomly distributed with mean 0 and variance θ. This term acts multiplicatively in the hazard function along with the baseline hazards and the covariate function (see Web-Appendix 1) (Therneau and Grambsch, 1998). These frailties are assumed to originate from a frailty distribution $f_\nu(\nu)$ belonging to some assumed parametric family of distributions, say, gamma or log-normal. The role of the frailty is thus to revise the hazard function for each cluster, so that clusters with higher frailties have a proportionally higher risk of failure. However, the estimation procedure is substantially different than that of the Cox or Weibull model fitting, and requires maximizing more challenging likelihoods (Nielsen et al., 1992; McGilchrist and Aisbett, 1991; Ripatti and Palmgren, 2000; Rizopoulos et al., 2008, 2009; Rizopoulos, 2010) (see Web-Appendix 2 for a general description of the likelihood under frailty).

Table 1 summarizes the frailty model fitting approaches for estimation of parameters associated with the covariates and heterogeneity parameter. Web-Appendix 3 includes the software implementation details of each of these methods. In the application section, for comparison, we will use a conventional Cox PH approach as well.

Table 1: Approaches of the frailty model fitting and software implementations considered in the current study.

Approach	Baseline hazard	Estimation approach	Frailty distribution	Software (R) package
`Cox.EM.G`	Unspecified	$\sim$ EM[1]	Gamma	`survival`
`Cox.REML.N`	Unspecified	$\sim$ REML[2]	Log-normal	`survival`
`Cox.ML.N`	Unspecified	Maximizing likelihood[3]	Log-normal	`coxme`
`Cox.Sp.G`	Unspecified	Semiparametric[4]	Gamma	`frailtySurv`
`Cox.Sp.N`	Unspecified	Same as above[4]	Log-normal	`frailtySurv`
`Weib.JM.G`	Weibull	Maximizing likelihood (joint model)[5]	Gamma	`JM`
`Weib.ML.G`	Weibull	Maximizing marginal likelihood[6]	Gamma	`parfm`
`Weib.ML.N`	Weibull	Same as above[6]	Log-normal	`parfm`

[1] Estimated via penalized methods to approximate EM approach (Klein, 1992; Nielsen et al., 1992; Guo and Rodriguez, 1992; Therneau et al., 2003).

[2] Estimated via penalized methods to approximate REML approach (McGilchrist, 1993).

[3] Estimated via maximizing the likelihood (Cortinas Abrahantes and Burzykowski, 2005; Therneau et al., 2003; Ripatti and Palmgren, 2000)

[4] Estimated via Gorfine et al.'s semiparametric estimation technique (Gorfine et al., 2006; Zucker et al., 2008).

[5] Estimated via maximizing the maximum likelihood (closed form) (Rizopoulos et al., 2008, 2009; Rizopoulos, 2010).

[6] Estimated via maximizing the marginal likelihood using the Laplace transform of the frailty distribution (Munda et al., 2012; Duchateau and Janssen, 2007; Wienke, 2010).

3 Simulation Settings

We generated clustered survival data from a specified shared frailty model, with the following expression of the hazard function:

$$S(t_{ij}|\boldsymbol{\beta}_j, \boldsymbol{\omega}_j) \quad = \quad \Lambda_0(t_{ij}) \exp(\boldsymbol{X}'_{ij}\boldsymbol{\beta}_j + \boldsymbol{Z}'_{ij}\boldsymbol{\omega}_j),$$

where Λ_0 is the cumulative baseline hazard, ω_j, a transformation of ν_j, is the frailty value of cluster j, $\boldsymbol{\beta}'_j = (\beta_1, \beta_2)$ is the regression coefficient vector, and $\boldsymbol{X}_{ij}$ is the covariate vector consisting of two covariates X_1 and X_2, for subject i in cluster j.

Table 2 lists 18 different simulation scenarios under consideration varying censoring percentage, assumed frailty and baseline hazard distributions. We fixed the number of subjects in each cluster to $K = 25$ and the number of clusters to $N = 15$. In these settings, we set heterogeneity parameter, $\theta = 2$ and set regression coefficients $\boldsymbol{\beta}'_j = (\beta_1, \beta_2) = (\log(2), \log(3))$. We considered two censoring proportions: 0.15 and 0.85. Frailty values were sampled from gamma, log-normal or inverse Gaussian distribution respectively (see Web-Table 4.2). Both covariates were generated from normal distributions, and censoring was generated under log-normal distribution. In these simulations, the baseline hazard was specified by the cumulative baseline hazard from Weibull or log-logistic distribution respectively (see Web-Table 4.1 and Web-Figure 4.1 in Web-Appendix 4). For conducting sensitivity analyses, we have additionally considered 108 other simulation scenarios varying the

number of clusters from moderate to large (e.g., $N = 30$ and 60), and varying heterogeneity parameters from very small to very large (e.g., $\theta = 0.1$, 0.5, 1 and 3). We performed a Monte Carlo study with 500 iterations in each scenario.

Table 2: List of 18 main simulating scenarios under consideration. In the following settings, we have considered parameters associated with covariates $\boldsymbol{\beta}_j' = (\beta_1, \beta_2) = (\log(2), \log(3))$ and 25 subjects in each clusters. These two covariates were generated from normal (0,1) distributions, and censoring was generated from log-normal distribution. For sensitivity analyses, we have considered additional 108 simulation settings by (I.) varying number of clusters to moderate (30) and large (60), and by (II.) varying heterogeneity parameters (e.g., 0.1, 0.5, 1 and 3).

Scenario	No. of clusters	% censoring	True frailty distribution	True baseline hazard distribution	Heterogeneity parameter θ
1	15	15	Gamma	Weibull	2
2	15	15	Gamma	Log-logistic	2
3	15	15	Log-normal	Weibull	2
4	15	15	Log-normal	Log-logistic	2
5	15	15	Inverse-Gaussian	Weibull	2
6	15	15	Inverse-Gaussian	Log-logistic	2
7	15	45	Gamma	Weibull	2
8	15	45	Gamma	Log-logistic	2
9	15	45	Log-normal	Weibull	2
10	15	45	Log-normal	Log-logistic	2
11	15	45	Inverse-Gaussian	Weibull	2
12	15	45	Inverse-Gaussian	Log-logistic	2
13	15	85	Gamma	Weibull	2
14	15	85	Gamma	Log-logistic	2
15	15	85	Log-normal	Weibull	2
16	15	85	Log-normal	Log-logistic	2
17	15	85	Inverse-Gaussian	Weibull	2
18	15	85	Inverse-Gaussian	Log-logistic	2

4 Results

4.1 Robustness of θ Estimates

The accuracy of the heterogeneity parameter estimates ($\hat{\theta}$) depends on the assumption of the frailty distribution. Gamma and log-normal distributions are assumed for frailty in most popularly used frailty models and associated implementations. For example, when gamma distribution is assumed during the data generation process, and parameter estimation model specifies the same distribution for frailty, the estimated θ parameters are associated with least bias (e.g., see Web-Figure 5.1 and for the log-normal case, see Web-Figure 5.2: dotted lines represent the true parameter value). Therefore, it is not surprising that when the inverse-Gaussian distribution is assumed for the data generation and parameter estimation model uses either gamma or Gaussian distribution for frailty, the estimated

θ parameters will be associated with substantial bias (see Figure 1). In this figure, the θ estimates from Gorfine et al.'s semiparametric approach (Gorfine et al., 2006; Zucker et al., 2008) had an unusually smaller interquartile range (IQR) under gamma frailty assumption when the true frailty was generated from Inverse Gaussian. We also see almost the same result when the censoring rate is moderate (45%, see Web-Figure 13.36).

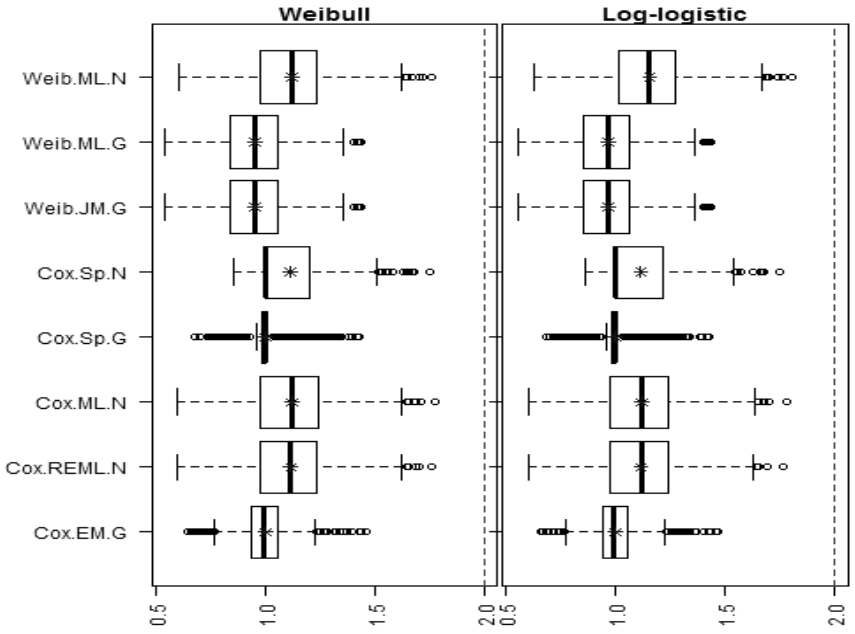

Figure 1: θ estimates when frailty generated from Inverse Gaussian ($N = 60$, censoring 15%).

4.2 Robustness of β Estimates

4.2.1 Robustness of β_1 and β_2 Estimates when Censoring is Mild

Estimation of θ is not generally of primary interest. One is usually more interested in estimating regression coefficients in the survival model. When survival times were generated from Weibull distribution and frailties were generated from gamma or log-normal distribution, in terms of estimating β_1 (associated with the first covariate in the model X_1), all the frailty estimation models are fairly robust to frailty parameter specification in the frailty model (e.g., see left-hand side of Web-Figures 5.1 - 5.2). Interestingly, when survival times were generated from log-logistic distribution (shown in the right-hand side of Web-Figures 6.1 - 6.2), parametric frailty models were slightly overestimating β_1. This phenomenon was also consistent for parameter β_2 estimates (associated with the second

covariate in the model X_2; see Web-Figures 7.1 - 7.2). Again, the most surprising result comes from the β_1 and β_2 estimates from Gorfine et al.'s semiparametric approach (Gorfine et al., 2006; Zucker et al., 2008) when frailty generated from Inverse Gaussian (see Figure 2 and Web-figure 7.3). The mean of the parameter estimates are further from the truth as well as the empirical variance is noticeably larger than that of other estimates.

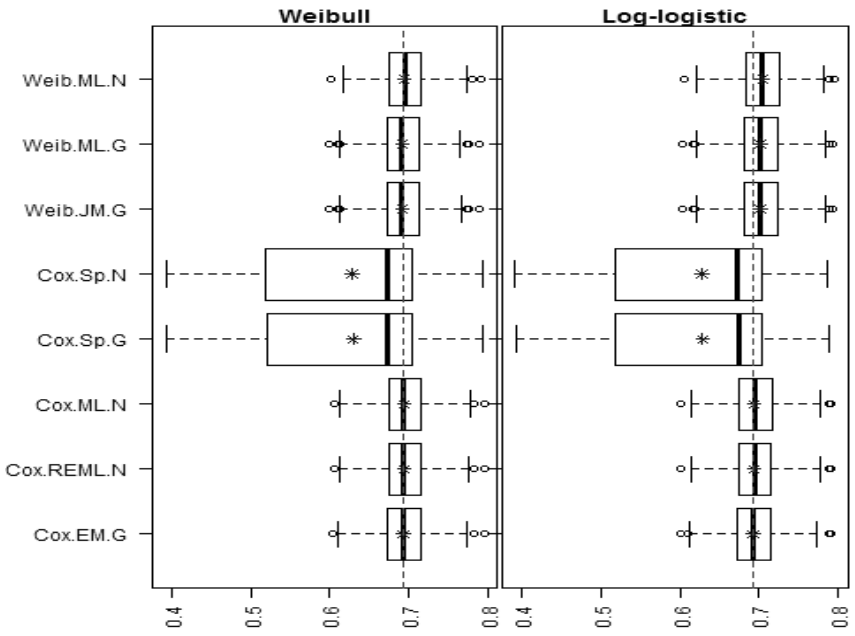

Figure 2: β_1 estimates when frailty generated from Inverse Gaussian($N = 60$, censoring 15%).

4.2.2 Robustness of β_1 and β_2 Estimates when Censoring is Severe

Interestingly, when we incorporated more censoring in the data generation process (85% censoring instead of 15% that was applied earlier), parameters estimates for β_1 and β_2 are less biased (at the cost of higher variability) irrespective of survival or frailty distribution assumed (see Web-Figures 8.1-8.3 and Web-Figures 9.1-9.3) as was seen in (Nielsen et al., 1992; Petersen et al., 2006; Hsu et al., 2007; Hirsch and Wienke, 2012). With severe censoring, θ estimates were also associated with less bias but often more variability irrespective of the frailty distribution chosen (see Web-Figures 10.1-10.3 compared to Web-Figures 5.1, 5.2 and Figure 1). For moderate censoring, the variabilities are in between (see Web-Figures 13.28-13.36). Again, estimates of β_1 and β_2 from Gorfine et al.'s semiparametric approach (Gorfine et al., 2006; Zucker et al., 2008) were associated with a large number of outliers when frailty generated from log-normal under moderate censoring

(see Web-figures 13.31 and 13.32). Moreover, when frailty generated from an inverse Gaussian distribution under moderate censoring, these estimates were very different than those from the rest of the approaches (see Web-figures 13.34 and 13.35). For mild censoring, the median and IQR values of these estimates were similar to those from the rest of the approaches, but were still associated with a large number of outliers (see Web-figures 13.46 and 13.47).

4.3 Effects of Changing Number of Clusters

When we used a smaller number of clusters ($N = 15$) with 25 subjects per cluster, the parameter estimates, we noticed higher empirical variance as expected (compared to $N = 60$, see Web-Figures 13.10-13.27, even under different censoring patterns). However, the general patterns remained the same (See Web-Figure 11.1-11.5). We also assessed these patters under different censoring patters (moderate censoring from Web-Figures 13.10-13.18; severe censoring from Web-Figures 13.19-13.27), and the patters were consistent (and with higher variability as censoring rate increased). One noticeable feature was that aberrant behavior of Gorfine et al.'s semiparametric methods (Gorfine et al., 2006; Zucker et al., 2008) was not visible in scenarios with small number of clusters (See Web-Figure 11.7-11.8). When we increased the number of clusters to 90, we see patterns similar to the case of $N = 60$ clusters (see Figure 3 and Web-Figures 12.1 and 12.2). The θ parameter is estimated with higher precision (judged by empirical IQRs) with larger number of clusters (see Web-Figures 12.2 and 13.36) when frailty was generated from an inverse Gaussian distribution.

In general, a larger number of clusters and a smaller censoring percentages contribute to smaller standard errors for the β estimates (see Web-Figures 13.1 - 13.9).

4.4 Effects of Changing the Heterogeneity Parameter θ

We changed the true θ values from 0.1 to 3, and the general patterns were mostly the same (see Web-Figures 13.67-13.174). When we set the true heterogeneity parameter small (e.g., $\theta = 0.1$), we do not see much difference in θ estimates from different approaches. However, as we gradually increase this parameter (to $\theta = 0.5$), we see that the estimates from the Cox frailty maximizing likelihood estimation approach under normality assumption are associated with more outliers than those from the other approaches (e.g., Web-Figure 13.96). The estimates from REML approach are close, but associated with slightly less outliers. When this parameter increases to $\theta = 3$, this phenomenon is more visible (e.g., one extreme example is Web-Figure 13.153). When the number of clusters was increased to 60 and censoring was mild (15%), all parameter estimates ($\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\theta}$) from Gorfine et al.'s semiparametric methods (Gorfine et al., 2006; Zucker et al., 2008) were associated with substantial bias (see Web-Figures 13.37-13.39).

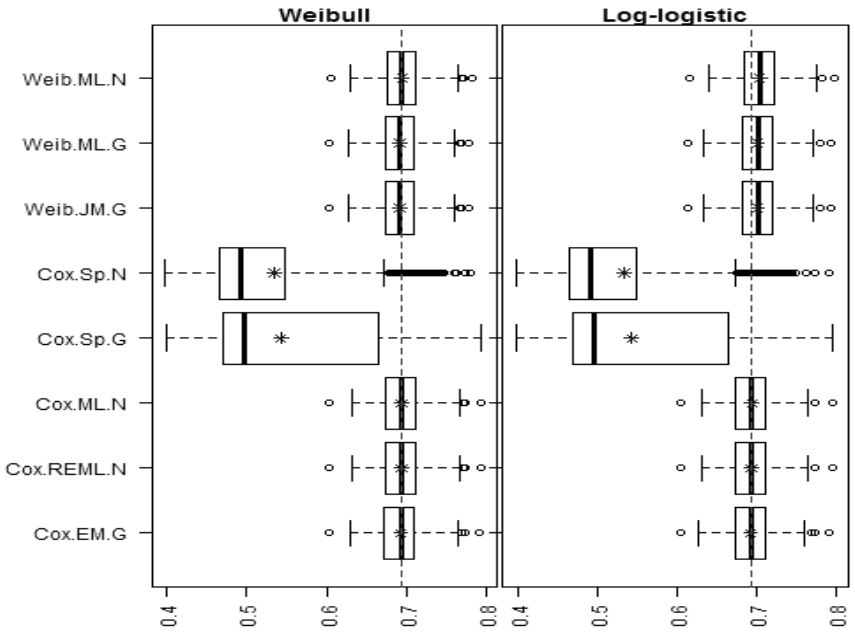

Figure 3: β_1 estimates when frailty generated from Inverse Gaussian and used very large number of clusters($N = 90$, censoring 15%).

5 Empirical Data Analysis: Determining the Birth Interval Dynamics in Bangladesh

A longer spacing between consecutive births ('birth-interval') decreases the number of children per women. This practice has beneficial effects on population size, maternal and child health status of a country. Bangladesh is a country with excessive population, and understanding the current practice of birth interval as well as its determinants is imperative for designing evidence-based strategies. Nationally representative Bangladesh demographic and health survey (BDHS) has been conducted in Bangladesh since 1993 to provide the demographic and health characteristics of populations with certain time interval. The 2014 BDHS used a two-stage stratified cluster sampling design, with average 120 households per cluster. A total of $17,989$ households were selected for survey where 96 percent were successfully interviewed from the sampled households of 2014 BDHS (NIPORT, Mitra and Associates, and ICF, 2016).

In this study, birth intervals are obtained from the Bangladesh Demographic and Health Surveys (BDHS) of the year 2014 (NIPORT, Mitra and Associates, and ICF, 2016). Women who have at least one live birth in preceding five years of the respective survey year, and have more than two children

were selected from the BDHS 2014 data and all birth intervals corresponding to the selected women were used in the analysis. The duration between last birth and interview date is considered as open birth interval (censored time), where the duration between two successive live births is closed birth interval (event time). That means birth intervals are considered as time-to-event data. Among the important covariates for modeling birth intervals that were reported in earlier studies (Khan et al., 2016; Mahmood et al., 2013; Chakraborty et al., 1996), those available in BDHS data were selected for this study. Web-Table 14.1 describes summary statistics of the covariates under consideration.

Among the few studies on birth interval in Bangladesh, considered birth intervals as independent time-to-event data and did not consider heterogeneity (Khan and Raeside, 1998; Islam et al., 2010). Birth-intervals within a geographical cluster (community) and mother can be considered as correlated because mothers from the same community could share a certain type of unobserved environmental factor, quality care of health facility, cultural practices etc., where children from the same mother could share certain unobserved characteristics like genetic factors, biological factors, knowledge etc. Therefore, modeling birth intervals without considering within mother or within-cluster level correlation may lead to incorrect inference when exploring the determinants of the birth spacing. A few recent studies have analyzed birth intervals of women from Bangladesh after adjusting the heterogeneity (frailty) due to cluster and mother (Khan et al., 2016; Mahmood et al., 2013). The inclusion of frailty terms was deemed helpful in exploring the between mother and between cluster variation on the length of birth interval. While fitting frailty models in this application, we were unable to obtain estimates from the following approaches `Weib.ML.N`, `Cox.Sp.N` and `Cox.Sp.G`. It was mostly due to the fact that they caused software to crash. We suspect that handling large dataset may be an issue for these software implementations. Moreover, performances of the later two approaches in some of our simulation settings were not encouraging as well.

5.1 Mother Level Frailty Model

Results show that effects of the maternal age at birth are significant in all models except `Weib.JM.G`, which indicates the likelihood of the subsequent birth decreases as the mother's age at birth increases (see Table 3). But the effect size varies across the models, where fixed effect Cox PH (non-frailty) model and the parametric gamma frailty model show the largest and smallest effects, respectively. Survival status of the index child has a statistically significant effect on the likelihood of having the next child in all the considered models. Comparing all the models it has been found that the mothers with previous children alive are less likely to have next birth compared to the mothers who lost their previous child, where effect size is comparatively small in the fixed effect Cox PH model. Family composition is an important predictor for the length of birth intervals (Setty-Venugopal and Upadhyay, 2002). All models reveal that mothers with one child are more likely to have the next birth compared to the mothers with a balanced number of children (one girl and one boy). As expected, mothers with two girls are less likely to wait longer for the next birth. Comparing the models with and without frailty, the smallest effect size is noticed in the fixed effects Cox PH model, where semiparametric and parametric frailty models show an increase in effect size for family composition. Birth interval tends to be significantly shorter for mothers from rural region compared to that of from urban region, which is true for all the models considered. In all of models under consideration, the

distribution of birth intervals changes according to the administrative division. Results reveal earlier that women from Chittagong and Sylhet divisions are more likely to have the next birth compared to women from Barisal division, whereas women from the Khulna, Rajshahi and Rangpur divisions show relatively lower likelihood of the next birth. Interestingly, there is no significant difference found in the distribution of birth interval among the maternal education categories in fixed effects Cox PH and semiparametric frailty model. However, consistent with the literature (Setty-Venugopal and Upadhyay, 2002; Tulasidhar, 1993), the duration of birth interval in the current analysis is shorter among the mothers with primary and no formal educated in parametric frailty models. Estimates of the frailty variance indicate that the lengths of birth intervals varies with mother.

Comparing all models, the parametric gamma frailty models estimate a higher unobserved heterogeneity component (roughly twice) than semi-parametric gamma frailty models. Model with log-normal frailty assumption is also giving smaller heterogeneity than the parametric gamma frailty model. Associated computational times are reported in Web-Appendix 14.2.

5.2 Estimates from the Empirical Data Analysis: Community Level Frailty Model

Results show that effects of all variables are identical with mother level frailty model, although the effect sizes are decreased in community level frailty model (see Web-Table 14.2 in Section 14). In this dataset, considering the community levels, we had small number of clusters, but cluster sizes were large, which was the opposite for the mother level. Consequently, comparing to all models with mother level frailty models, the size of the unobserved heterogeneity is much smaller in community level than the mother level. Comparing all community level frailty models, the parametric gamma frailty models estimate a higher unobserved heterogeneity component than semi-parametric gamma frailty models. Model with log-normal frailty is also giving smaller heterogeneity than the parametric gamma frailty model.

6 Discussion

Frailty models are used for analyzing correlated survival-time data (Duchateau and Janssen, 2007). Availability of software routines has made it easier for the researchers to fit this rather complicated model in practical data analysis applications. In the literature, we have found two studies that have compared a number of statistical properties of two software implementations (`survival` and `coxme`) (Kelly, 2004; Hirsch and Wienke, 2012). Since then, additional software packages have emerged that offers fitting of these frailty models. In this current work, we have assessed the robustness the estimates from these newer packages (`frailtySurv`, `JM` and `parfm`), and compared with the previous packages. Using extensive simulation scenarios, we showed that, Cox frailty models are more robust than the parametric Weibull models when the baseline hazard distribution is misspecified.

Our simulation study shows that estimates of the heterogeneity parameter are highly sensitive to the misspeficication of the frailty distribution, which may lead to noticeable bias. However, the

Table 3: Estimates of the parameters and corresponding standard errors of the different survival models with and without mother level frailty for birth interval of Bangladesh (Bangladesh demographic and health survey 2014).

Variable/Method	Cox PH	`Cox.EM.G`	`Cox.REML.N`	`Cox.ML.N`	`Weib.JM.G`	`Weib.ML.G`
	$\hat{\beta}$(SE)	$\hat{\beta}$(SE)	$\hat{\beta}$(SE)	$\hat{\beta}$(SE)	$\hat{\beta}$(SE)	$\hat{\beta}$(SE)
Mother age at birth	-0.194^a (0.037)	-0.148^a (0.042)	-0.148^a (0.042)	-0.152^a (0.042)	-0.060 (0.046)	-0.059^b (0.026)
Survival status						
Alive	-0.732^a (0.047)	-0.841^a (0.052)	-0.831^a (0.052)	-0.826^a (0.052)	-0.890^a (0.056)	-0.900^a (0.031)
Maternal education						
Secondary	0.020 (0.099)	0.039 (0.115)	0.036 (0.115)	0.035 (0.114)	0.088 (0.130)	0.067 (0.072)
Primary	0.144 (0.099)	0.174 (0.115)	0.169 (0.115)	0.168 (0.114)	0.238 (0.130)^c	0.239^a (0.072)
No education	0.124 (0.099)	0.172 (0.115)	0.157 (0.115)	0.156 (0.114)	0.224 (0.130)^c	0.233^a (0.072)
Family composition						
1 boy	0.389^a (0.049)	0.458^a (0.053)	0.457^a (0.053)	0.453^a (0.053)	0.576^a (0.057)	0.577^a (0.031)
1 girl	0.404^a (0.048)	0.468^a (0.052)	0.465^a (0.052)	0.462^a (0.052)	0.581^a (0.055)	0.584^a (0.031)
2 boys	-0.032 (0.057)	-0.032 (0.062)	-0.031 (0.062)	-0.031 (0.062)	-0.040 (0.068)	-0.045 (0.038)
2 girls	0.162^a (0.053)	0.181^a (0.058)	0.178^a (0.058)	0.178^a (0.058)	0.209^a (0.063)	0.217^a (0.035)
$>$2 children	-0.148^a (0.045)	-0.273^a (0.049)	-0.282^a (0.049)	-0.273^a (0.049)	-0.367^a (0.056)	-0.373^a (0.031)
Division						
Chittagong	0.167^a (0.048)	0.178^a (0.057)	0.178^a (0.057)	0.177^a (0.057)	0.203^a (0.065)	0.206^a (0.033)
Dhaka	0.018 (0.051)	0.009 (0.060)	0.014 (0.060)	0.014 (0.06)	0.004 (0.069)	0.004 (0.037)
Khulna	-0.270^a (0.065)	-0.301^a (0.077)	-0.297^a (0.077)	-0.296^a (0.076)	-0.350^a (0.088)	-0.349^a (0.049)
Rajshahi	-0.243^a (0.060)	-0.272^a (0.070)	-0.269^a (0.071)	-0.267^a (0.070)	-0.312^a (0.081)	-0.313^a (0.044)
Rangpur	-0.215^a (0.057)	-0.247^a (0.067)	-0.238^a (0.068)	-0.237^a (0.067)	-0.283^a (0.077)	-0.283^a (0.042)
Sylhet	0.474^a (0.047)	0.519^a (0.056)	0.519^a (0.056)	0.517^a (0.056)	0.582^a (0.065)	0.581^a (0.034)
Place of residence						
Rural	0.210^a (0.031)	0.223^a (0.037)	0.222^a (0.037)	0.221^a (0.037)	0.244^a (0.043)	0.244^a (0.024)
Variance of random effect	-	0.117	0.129	0.119	0.242	0.241

p-value: $a < 0.01$, $b < 0.05$, $c < 0.10$

effects of frailty distribution misspeficication on the regression coefficient estimations are generally minimal. This is especially true for Cox PH frailties. These findings are coherent with the previous literature (Hsu et al., 2007). We also found that the heterogeneity parameter estimates are generally insensitive to the choice of baseline hazard function. However, not all frailty estimation technique implementations based on Cox PH are the same. In terms of the software implementations of these frailty models, most of these packages resulted in consistent estimates under most simulation scenarios we have considered. Interestingly, we had two new observations from the results of our extensive simulations. The first one is that the heterogeneity parameter (θ) estimates from the `coxme` package were associated with unusual variability when the true heterogeneity parameter was large. Secondly, and most notably, when Cox PH frailty models are fitted using the Gorfine et al.'s semiparametric estimation technique (Gorfine et al., 2006; Zucker et al., 2008) (e.g., `Cox.Sp.G` and `Cox.Sp.N` from package `frailtySurv`), the regression parameter estimates are sometimes associated with unusually large variability when dealing with a large number of clusters and mild censoring. If the practitioners are dealing with such scenarios (i.e., when the estimate of θ is unusually high or low, or when there is a large number of clusters and mild censoring present), we would advise them to consider refitting the models using other approaches as sensitivity analyses, and look for inconsistency, if any.

For the empirical data analysis, all the approaches considered show that different socioeconomic and demographic variables are useful in explaining women's birth spacing distribution in Bangladesh. The size of the effects is relatively high in parametric models compared to the semiparametric models. This is consistent with our findings when the survival data were not generated from Weibull distribution and the estimates of regression parameters were overestimated. Based on our simulation experience, we prefer Cox PH frailty estimates for our data analysis. A sizable mother to mother variation is also noticed among the all frailty models, which indicates the importance of frailty effect in birth interval modeling. However, the sizes of the estimates of heterogeneity parameter were generally smaller for Cox frailty approaches compared to those from Weibull frailty approaches. In addition, a more uniform sized community to community variation is also noticed, although the size is much smaller than mother to mother variation. Comparing the estimates of regression coefficients, we notice that the effects of covariates are smaller when frailty effects are ignored. In general, frailty models seem to account for unobserved heterogeneity; regression estimates and their standard errors increase when frailty is introduced into the model. These findings are consistent with the previous literature (Khan et al., 2016; Mahmood et al., 2013).

In summary, most of the popular frailty implementations we have considered in this work were able to produce reasonable estimates (e.g., packages `JM` (fastest in our application), `parfm` (slowest), `survival` and `coxme`). In general, Cox frailties were found to be more robust compared to parametric approaches and the corresponding software implementations (e.g., `survival` and `coxme`) were found to be more stable than the others. In our application, we found the frailty model fitting function in `coxme` package to be much faster than those in `survival` (see Web-Appendix 14.2). However, overall, in the same application, the frailty model fitting function in `JM` package was the fastest, while that in `parfm` was much slower in fitting the model. Most notably, in this research, we have identified that the implementation of a relatively newer method, such as Gorfine et

al.'s semiparametric estimation technique (Gorfine et al., 2006; Zucker et al., 2008) (e.g., in package `frailtySurv`) produced some unexpected results in some specific settings. Future research needs to identify whether these issues could be mitigated by more careful implementation of the approach in the software package.

This work has some limitations and potential to extend in future research endeavours. In this work, we did not explore hypothesis testing either of the regression coefficients or the heterogeneity parameter of the frailty models. There are other software implementations of frailty approaches proposed in the literature. The versions we considered are freely available and the most popularly used by practitioners, but the list is not exhaustive (Rondeau et al., 2012; Do Ha et al., 2012). We also did not consider commercial software implementations or approaches that consider variable selection via shrinkage, AIC or BIC approaches to choose parsimonious models (Abdulkarimova, 2013; Munda et al., 2012; Androulakis et al., 2012). In our simulation, we have considered only two continuous covariates, and both were normally distributed. In practice, data analysis may involve a wide variety of variables (e.g., of categorical and continuous nature). Future simulations could explore more complex covariate settings.

Our simulation study findings highlights the consequences of violating the assumptions required by different types of frailty model implementations. Based on our simulation findings and implications of these results in the subsequent data analysis context, practitioners can make informed decision about which type of frailty models and associated software implementations would be likely useful in different situations. Researchers should consider comparing various existing and new frailty model fits on their dataset to identify if any of them produce unusual results. We considered the default settings offered by off-the-shelf software packages that are freely available. Hopefully this will facilitate appropriate use of these frailty approaches by practitioners.

Acknowledgements

We thank A. H. M. Mahbub Latif (Professor, Graduate School of Public Health, St. Luke's International University, Tokyo, Japan) for helpful comments, his input on computing issues and overall guidance in designing the study and framing the writing.

References

Aalen, O. O. (1988), "Heterogeneity in survival analysis," *Statistics in Medicine*, 7, 1121–1137.

Abbring, J. H. and Van Den Berg, G. J. (2007), "The unobserved heterogeneity distribution in duration analysis," *Biometrika*, 97, 87–99.

Abdulkarimova, U. (2013), "Frailty Models For Modelling Heterogeneity," Ph.D. thesis, McMaster University.

Albert, P. S. (1999), "Longitudinal data analysis (repeated measures) in clinical trials," *Statistics in Medicine*, 18, 1707–1732.

Andersen, P. (1993), *Statistical models based on counting processes*, Springer.

Androulakis, E., Koukouvinos, C., and Vonta, F. (2012), "Estimation and variable selection via frailty models with penalized likelihood," *Statistics in Medicine*, 31, 2223–2239.

Chakraborty, N., Sharmin, S., and Islam, M. A. (1996), "Differential pattern of birth intervals in Bangladesh." *Asia-Pacific Population Journal*, 11, 73–86.

Clayton, D. G. (1978), "A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence," *Biometrika*, 141–151.

Cortinas Abrahantes, J. and Burzykowski, T. (2005), "A version of the EM algorithm for proportional hazard model with random effects," .

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

Do Ha, I., Noh, M., and Lee, Y. (2012), "Frailtyhl: a package for fitting frailty models with h-likelihood," *R Journal*, 4, 28–36.

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

Fleming, T. and Harrington, D. (1991), *Counting processes and survival analysis*, Wiley New York.

Gorfine, M., Zucker, D. M., and Hsu, L. (2006), "Prospective survival analysis with a general semiparametric shared frailty model: A pseudo full likelihood approach," *Biometrika*, 735–741.

Guo, G. and Rodriguez, G. (1992), "Estimating a multivariate proportional hazards model for clustered data using the EM algorithm, with an application to child survival in Guatemala," *Journal of the American Statistical Association*, 87, 969–976.

Henderson, R. and Oman, P. (1999), "Effect of frailty on marginal regression estimates in survival analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61, 367–379.

Hirsch, K. and Wienke, A. (2012), "Software for semiparametric shared gamma and log-normal frailty models: an overview," *Computer Methods and Programs in Biomedicine*, 107, 582–597.

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

Hsu, L., Gorfine, M., and Malone, K. (2007), "On robustness of marginal regression coefficient estimates and hazard functions in multivariate survival analysis of family data when the frailty distribution is mis-specified," *Statistics in Medicine*, 26, 4657–4678.

Islam, S., Islam, M. A., and Padmadas, S. S. (2010), "High fertility regions in Bangladesh: a marriage cohort analysis," *Journal of Biosocial Science*, 42, 705–719.

Karim, M. E. (2008), "Comparative Study on the different approaches of Frailty Model: A simulation Study," Master's thesis, University of Dhaka, Dhaka.

Kelly, P. J. (2004), "A review of software packages for analyzing correlated survival data," *The American Statistician*, 58, 337–342.

Khan, H. and Raeside, R. (1998), "The determinants of first and subsequent births in urban and rural areas of Bangladesh." *Asia-Pacific Population Journal*, 13, 39–72.

Khan, J. R., Bari, W., and Latif, A. M. (2016), "Trend of determinants of birth interval dynamics in Bangladesh," *BMC Public Health*, 16, 934.

Klein, J. P. (1992), "Semiparametric estimation of random effects using the Cox model based on the EM algorithm," *Biometrics*, 795–806.

Mahmood, S., Zainab, B., and Latif, A. M. (2013), "Frailty modeling for clustered survival data: an application to birth interval in Bangladesh," *Journal of Applied Statistics*, 40, 2670–2680.

McGilchrist, C. (1993), "REML Estimation for Survival Models with Frailty," *Biometrics*, 49, 221–225.

McGilchrist, C. and Aisbett, C. (1991), "Regression with Frailty in Survival Analysis," *Biometrics*, 47, 461–466.

Munda, M., Rotolo, F., Legrand, C., et al. (2012), "parfm: Parametric frailty models in R," *Journal of Statistical Software*, 51, 1–20.

Nielsen, G. G., Gill, R. D., Andersen, P. K., and Sørensen, T. I. A. (1992), "A counting process approach to maximum likelihood estimation in frailty models." *Scandinavian Journal of Statistics*, 19, 25–44.

NIPORT, Mitra and Associates, and ICF (2016), "Bangladesh Demographic and Health Survey 2014," Tech. rep., National Institute of Population Research and Training (NIPORT), Mitra and Associates, and ICF International, Dhaka, Bangladesh, and Rockville, Maryland, USA: NIPORT, Mitra and Associates, and ICF International.

Petersen, L., Sørensen, T. I., Nielsen, G. G., and Andersen, P. K. (2006), "Inference methods for correlated left truncated lifetimes: parent and offspring relations in an adoption study," *Lifetime Data Analysis*, 12, 5–20.

Ripatti, S. and Palmgren, J. (2000), "Estimation of Multivariate Frailty Models Using Penalized Partial Likelihood," *Biometrics*, 56, 1016–1022.

Rizopoulos, D. (2010), "JM: An R package for the joint modelling of longitudinal and time-to-event data," *Journal of Statistical Software*, 35, 1–33.

Rizopoulos, D., Verbeke, G., and Lesaffre, E. (2009), "Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71, 637–654.

Rizopoulos, D., Verbeke, G., and Molenberghs, G. (2008), "Shared parameter models under random effects misspecification," *Biometrika*, 63–74.

Rondeau, V., Mazroui, Y., and 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, 1–28.

Sahu, S. K., Dey, D. K., Aslanidou, H., and Sinha, D. (1997), "A Weibull regression model with gamma frailties for multivariate survival data," *Lifetime Data Analysis*, 3, 123–137.

Setty-Venugopal, V. and Upadhyay, U. D. (2002), "Birth spacing: three to five saves lives." *Population Reports. Series L: Issues in World Health*, 1, 1–23.

Therneau, T. and Grambsch, P. (1998), "Penalized Cox models and frailty," *Working manuscript*.

— (2000), *Modeling Survival Data: Extending the Cox Model*, Springer.

Therneau, T. M., Grambsch, P. M., and Pankratz, V. S. (2003), "Penalized survival models and frailty," *Journal of Computational and Graphical Statistics*, 12, 156–175.

Tulasidhar, V. (1993), "Maternal education, female labour force participation and child mortality: evidence from the Indian census," *Health Transition Review*, 177–190.

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

Wienke, A. (2010), *Frailty models in survival analysis*, CRC Press.

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988), "Models for longitudinal data: a generalized estimating equation approach," *Biometrics*, 1049–1060.

Zhang, Z. (2016), "Parametric regression model for survival data: Weibull regression model as an example," *Annals of Translational Medicine*, 4.

Zucker, D. M., Gorfine, M., and Hsu, L. (2008), "Pseudo-full likelihood estimation for prospective survival analysis with a general semiparametric shared frailty model: Asymptotic theory," *Journal of Statistical Planning and Inference*, 138, 1998–2016.

Received: February 16, 2018

Accepted: July 31, 2018