Hacettepe Journal of
Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 53 (4) (2024), 1158 – 1177
DOI : 10.15672/hujms.1320996

Research Article

Exploring shared frailty models for cluster-specific risk estimation: A study on diabetes patients with a history of acute coronary syndrome

Kazeem Adedayo Adeleke*[1,2,3], Harshal Deshmukh[1,2], Alan Rigby[1],
Thozhukat Sathyapalan[1,2], Joseph John[4]

[1] *Hull York Medical School, University of Hull, Hull, United Kingdom*
[2] *Allam Diabetes Centre, Hull University Teaching Hospital NHS Trust, Hull, United Kingdom*
[3] *Department of Mathematics, Obafemi Awolowo University, Ile-Ife, Nigeria*
[4] *Department of Cardiology, Castle Hill Hospital, Kingston upon Hull, United Kingdom*

Abstract

This study proposes the use of semiparametric log-normal shared frailty models to analyze time-to-event data for individuals with similar features referred to as clusters. Shared frailty models are useful for modeling and estimating common risk in the lifetimes of individuals in these clusters. While various methods have been proposed for estimating shared frailty models, few studies have explored the use of the pseudo-full-likelihood method. In this study, the pseudo-full-likelihood and hierarchical likelihood approaches were used to construct and estimate parameter estimates and check for asymptotic properties via simulations. Log-normal semiparametric frailty model was used to obtain cluster-specific frailty based on the semiparametric log-normal shared frailty distribution. The results of both methods were compared, and prediction intervals for a random effect were obtained. To further investigate the existence of shared frailty in diabetes patients and a history of acute coronary syndrome (STEMI and NSTEMI), data from UK Biobank was used. The results suggest the presence of frailty within the clusters and indicate cluster time dependence in the study population. Overall, this study highlights the potential benefits of using the pseudo-full-likelihood method in shared frailty modeling and provides insights into the impact of observed variabilities on hazards within clusters.

Mathematics Subject Classification (2020). 62N01, 62N02, 62Nxx

Keywords. Log-normal frailty model, pseudo-full-likelihood, shared frailty model, acute coronary syndrome, diabetes mellitus

*Corresponding Author.

Email addresses: adedayobright@gmail.com (K.A. Adeleke), harshaldeshmukh@gmail.com (H. Deshmukh), asr1960@hotmail.com (A. Rigby), thozhukat.sathyapalan@nhs.net (T. Sathyapalan), joseph.john@nhs.net (J. John)

Received: 03.07.2023; Accepted: 15.06.2024

1. Introduction

The underline assumption or shape of the baseline hazard function dictates if a model is semiparametric or parametric in a Cox regression model. Cox regression model is considered to be semiparametric if no assumption(s) is/are made about the nature of the baseline hazard function. The celebrated Cox regression model has provided tremendously successful tools for exploring the association of covariates with failure time and survival distributions. It has also been used for studying the effect of a primary covariate while adjusting for other variables. Of course, In semiparametric frailty model, hazard function is left unspecified, resulting in different estimation strategies when compared with the parametric frailty models. In epidemiology or clinical data, one of the main features is clustering or dependence on some unobserved covariate. This could be due to geographical location, common genes, and so on [38]. The shared frailty model requires careful consideration of various estimation techniques. This approach, initially proposed as a means of modeling cluster-specific unobserved effects in prior studies [6, 7], has since gained widespread use in the field [37]. However, the accurate estimation of frailty parameters remains a critical challenge in the implementation of these models. The study by some authors [6, 23, 28, 31] investigated the Expectation Maximization (EM) algorithm approach for parameter estimation of the semiparametric Gamma Shared frailty model. The problem with the EM algorithm is that variance estimates of the estimated parameters are not readily available. Hanagal and Sharma [20] conducted research on bivariate survival times using parametric shared frailty models, where the frailty term was specified as an inverse Gaussian distribution and the baseline distribution was known to be log-logistic. The use of the profile likelihood method in semiparametric models was established to show that profile likelihood and ordinary likelihood are the same so long as the nuisance parameter has been profiled [27]. The development of a new method to handle any parametric frailty distribution with finite time was conducted and applied to analyze correlated survival times and also investigated large sample properties [14]. In a study to estimate the parameters of the Gamma semiparametric frailty model using the pseudo-full-likelihood (PFL) method [41], the results showed consistency in variance estimation. Several methods of estimation and diagnostics for model adequacy and inference in frailty models have also been discussed in the literature [1–4, 12, 13, 15, 25, 34, 39, 40]. A result from simulating and fitting semiparametric shared frailty models through R-package showed that parameter estimators are asymptotically normally distributed [26]. This paper is motivated by the work of [16], who constructed a prediction interval in the log-normal semiparametric frailty model using the hierarchical likelihood (HL) method and obtained a standard error of the random effects. Our aim is to construct and estimate the parameter estimates and check for the asymptotic properties of the log-normal semiparametric frailty model, using PFL approach and HL methods. An illustration was carried out with survival times of related individuals such as twins or acute coronary syndrome (ACS) (STEMI and NSTEMI). Acute myocardial infarction is myocardial necrosis resulting from obstruction of a coronary artery. Symptoms include chest discomfort with or without dyspnea, nausea, and diaphoresis. Diagnosis is by ECG and the presence or absence of serologic markers. For St-segment-elevation myocardial infraction, emergency reperfusion is via fibrinolytic drugs, percutaneous intervention, or coronary artery bypass graft surgery. For the nonST-segment-elevation myocardial infarction, reperfusion is via percutaneous intervention or coronary artery bypass graft surgery.

2. Methods

Frailty refers to an unobserved random effect that accounts for the inherent variability in the risk of an event occurring among individuals or groups with similar characteristics.

This concept is analogous to the idea of random effects in mixed models, where frailty represents unmeasured factors that affect an individual's susceptibility to the event of interest [37]. Frailty models are random effect models for time-to-event data, where the random effect has a multiplicative effect on the baseline hazard function [22]. It is an extension of the most popular Cox proportional hazard model. Ignoring the existence of frailty term in the analysis of survival time data, when heterogeneity is present will leads to underestimation of parameters with higher standard errors [29]. By incorporating frailty into survival models, we can better account for heterogeneity and dependency within the data. To establish notation, assume that given n q-dimensional vector of covariates X, the underline conditional hazard rate for Cox regression model [9] is defined as:

$$\lambda(t \mid x) = \frac{1}{\Delta}\Pr\{t \leq T < t + \Delta t \mid T \geq t, X = x\}. \tag{2.1}$$

$$\lambda(t \mid x) = \lambda_0(t)\psi(x) \tag{2.2}$$

where,

$$\psi(x) = \exp(X\beta)$$

Consider n clusters with cluster i containing n_i observations $i = 1, 2, \dots, n_i$. Let T_{ij}, C_{ij}, and X_{ij} denote respectively the event time, censoring time for individual j in cluster i and observed p-vector of covariates X_{ij}, $1 \leq i \leq n$, $1 \leq j \leq n_i$ and an associated unobserved frailty Y_i, $(1 \leq i \leq n)$. The indicator variable $\delta_{ij} = I(T_{ij} \leq C_{ij})$. Let Y_i be the frailty which induce independence among cluster members. If Y_i could be measured and included in the model, then $\theta \to 0$ and we obtain Cox marginal PH model. Suppose individuals in the same clusters share the same value called frailty, then the conditional hazard function at time T_{ij} for the j^{th} subject in the i^{th} cluster is given as:

$$\lambda_{ij}(t_{ij}, x_{ij}, y) = \lambda_0(t_{ij})\exp(x_{ij}^{\top}\beta + y_i) \tag{2.3}$$

Let $u_i = \exp(y_i)$ then,

$$\lambda_{ij}(t_{ij}, x_{ij}, y) = \lambda_0(t_{ij})u_i\exp(x_{ij}^{\top}\beta) \tag{2.4}$$

(2.4) is a shared frailty model as it represents the model for j^{th} subject in the i^{th} cluster that share the same frailty factor. In this paper, we choose to use log-normal semiparametric frailty model as well as PFL method of estimation due to its efficiency, computational simplicity as well as its approximation to the likelihood function.

2.1. Log-normal frailty distribution

The use of log-normal as a frailty distribution emanated from the properties of generalized models with the standard assumption that random effect U_i follows a zero mean and variance σ^2 [11]. The function is given as;

$$f(u) = \frac{1}{u\sqrt{2\pi\sigma^2}}\exp\left(\frac{\log(u^2)}{2\sigma^2}\right) \tag{2.5}$$

with mean $E(u) = \exp(\sigma^2/2)$ and $var(u) = \exp(\sigma^2)(\exp(\sigma^2)^{-1})$. Log-normal frailty distribution has no explicit evaluation of Laplace transformation but allows a relatively simple extension into the multivariate case. Figure 1 represents the plot of log-normal distribution at various values of standard deviations (σ). Literature revealed that there are some commonly used baseline hazard distribution viz: Exponential, Weibull, Gompertz, Gamma, e.t.c., as for this work, where model is purely semiparametric, therefore, the baseline distribution is unspecified.

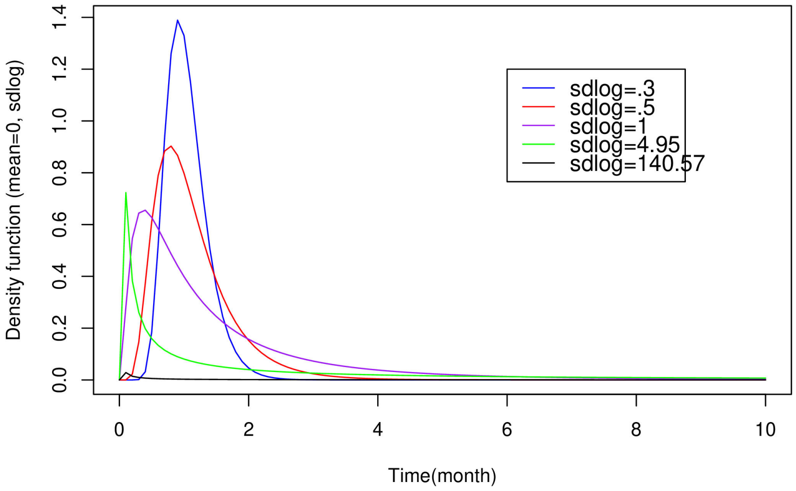

Figure 1. Log-normal distribution at various values of standard deviation.

2.2. Pseudo-full-likelihood estimation method

Here, a log-normal semiparametric frailty model was proposed using PFL method of estimation for gamma frailty model [14, 26, 38]. We apply the model to a diabetes data from UK Biobank. Given that the frailty u_i is independent of X_{ij} and has the density function, $f(u_i;\theta)$ where θ is unknown parameter. Then, the unconditional full likelihood function of the data is given as

$$L(\beta,\theta,\Lambda_0) = \prod_{i=1}^{n} \int \prod_{j=1}^{m_i} \{\lambda_{ij}(T_{ij}|X_{ij},u)\}^{\delta_{ij}} S_{ij}(T_{ij}|X_{ij},u) f(u)du \tag{2.6}$$

$$L(\beta,\theta,\Lambda_0) = \prod_{i=1}^{n}\prod_{j=1}^{m_i} \{\lambda_0(T_{ij})\exp(X'\beta)\}^{\delta_{ij}} \prod_{i=1}^{n} \int u_i^{N_{i\cdot}(t)} \exp\{-uH_{i\cdot}(t)\} f(u)du \tag{2.7}$$

2.3. Pseudo-full-likelihood estimation method

Here, a log-normal semiparametric frailty model was proposed using PFL method of estimation for gamma frailty model [14, 26, 38]. We apply the model to a diabetes data from UK Biobank. Given that the frailty u_i is independent of X_{ij} and has the density function, $f(u_i;\theta)$ where θ is unknown parameter. Then, the unconditional full likelihood function of the data is given as

$$L(\beta,\theta,\Lambda_0) = \prod_{i=1}^{n} \int \prod_{j=1}^{m_i} \{\lambda_{ij}(T_{ij}|X_{ij},u)\}^{\delta_{ij}} S_{ij}(T_{ij}|X_{ij},u) f(u)du \tag{2.8}$$

$$L(\beta,\theta,\Lambda_0) = \prod_{i=1}^{n}\prod_{j=1}^{m_i} \{\lambda_0(T_{ij})\exp(X'\beta)\}^{\delta_{ij}} \prod_{i=1}^{n} \int u_i^{N_{i\cdot}(t)} \exp\{-uH_{i\cdot}(t)\} f(u)du \tag{2.9}$$

2.4. Hierarchical likelihood approach

The HL uses the Laplace approximation when the numerical integration is intractable, giving a statistically efficient estimation in frailty models [4,18,19,24]. Estimation method using the HL for the inference of frailty models was proposed to solve the semiparametric models in frailty models together with the corresponding estimation procedure. This method helps semiparametric models in resolving the problem especially once it involves clustered survival data. The approach follows that of [5,16–18] for model with competing risk data. Below is the H likelihood for the log-normal frailty model given in (2.4);

$$l = l(\beta, \lambda_0, \theta) = \sum_{ij} l_{1ij} + \sum_{i} l_{2i} \tag{2.10}$$

where

$$\sum_{ij} l_{1ij} = \sum_{ij} \delta_{ij} \{\log \lambda_0(y_{ij}) + \kappa_{ij}\} - \sum_{ij} \Lambda_0(y_{ij}) \exp(\kappa_{ij})$$

and

$$l_{2i} = l_{2i}(\theta; v_i) = \frac{1}{2}\log(2\pi\theta) - \frac{1}{2}v_i^2$$

is the log of density function for v_i with variance θ, $\lambda_0 = (\lambda_{01}, \lambda_{02}, \ldots, \lambda_{0\tau})^T$, $\kappa_{ij} = x_{ij}^T\beta + v_i$.

From equation 2.1 we estimate the parameters (β, v) with $v = (v_1, \ldots, v_q)^T$. In [18,19], we noticed that the dimension of λ_0 increases with sample size n, hence, the proposition of Profile HL l^* with λ_0 eliminated.

$$l^* = l|_{\lambda_0 = \hat{\lambda}_0} = \sum_{ij} \delta_{ij}\kappa_{ij} - \sum_{k} d_k \log\left\{\sum_{(ij \in R(k))} \exp(\kappa_{ij})\right\} + \sum_{i} l_{2i}$$

Here in log-normal frailty model, l^* becomes the kernel of the penalized partial likelihood as used in [33]. From (2.4), we estimate the parameters (β, θ), and random effect v.

2.5. Prediction intervals for random effects

Following series of research works in literature [18,19] the asymptotic covariance matrix of $\hat{\beta}$ and $\hat{v} - v$ is the inverse, i.e. the Hessian matrix H without nuisance parameter given by

$$H(\hat{\beta}, v) = -\begin{pmatrix} \frac{\partial^2 l^*}{\partial\beta^2} & \frac{\partial^2 l^*}{\partial\beta\partial v} \\ \frac{\partial^2 l^*}{\partial v\partial\beta} & \frac{\partial^2 l^*}{\partial v^2} \end{pmatrix}$$

$$H(\hat{\beta}, v) = -\begin{pmatrix} X^T W^* X & X^T W^* Z \\ Z^T W^* X & Z^T W^* Z + R \end{pmatrix}$$

with X_{nxp} matrix whose i^{th} row vector is X_{ij}^T, Z_{nxq} group indicator matrix, W_{nxn} symmetric matrix given in [19] and $R = \text{diag}\left[\frac{\partial^2 l_{2i}}{\partial v^2}\right]_{q\times q}$ diagonal matrix with

$$\text{Var}(\hat{\beta}) = (X^T V^{-1} X) \quad \text{with} \quad V = W^{n-1} + ZR^{-1}Z^T \tag{2.11}$$

From the bottom left-hand corner of H^{-1}, from above the variance of $\hat{v} - v$ is:

$$\text{Var}(\hat{v} - v) = \left[(Z^T W^* Z + R) - (Z^T W^* X)(X^T W^* X)^{-1}(X^T W^* Z)\right] \tag{2.12}$$

Also, from the bottom right-hand corner of H^{-1},the 95% prediction interval for random effect is given by

$$\hat{v}_i \pm 1.96 \times \text{SE}(\hat{v}_i) \tag{2.13}$$

where, $SE(\hat{v}_i) = \sqrt{\text{Var}(\hat{v} - v)}$ is the estimated standard error obtained from the H^{-1} matrix [10,16,36].

3. Simulation study

3.1. Simulation study application

To study the numerical implementation and performance of the proposed model through PFL method, we conduct a simulation study based on semiparametric log-normal frailty distribution and obtain prediction intervals for the random effects [26]. We begin by generating data for semiparametric frailty model (2.4) with varied clusters of sizes k = (2,3,4, and 5) and for sample size N = (50, 100, 200, 300). We made use of Type IV Pareto distribution baseline hazard, $h = dca(dt)^{c-1}[1+(dt)^c]^{-1}$, where a, c and d are location, shape and scale parameters. The regression parameters $\beta_1 = 5, \beta_2 = \log(3)$ and $variance = \theta = \gamma^2 = 2$. We sample two time-independent covariates $X_{1j} \sim \mathcal{N}(0,1)$ and $X_{2j} \sim \text{U}(1,3)$ from normal and uniform distributions respectively. A right censoring rate is fixed at approximately 35%. Our simulation is based on 1000 replications and the $mean(\hat{\beta})$, standard deviation $SE(\hat{\beta})$, the mean square error $MSE(\hat{\beta})$ are also obtained. Likewise, for random frailty parameter, θ or σ^2, the mean(θ), SE(θ), mean square error, MSE(θ) and prediction intervals are obtained. For model fitting simulation and computations, we used frailtysurv in R programming.

3.2. Results of the simulations

An important issue is the choice between different estimation methods, yet on the same frailty model. The choice between PFL and HL estimation methods. In our simulation study, both methods converge and show no problem. It only takes PFL little time than it takes HL to converge. Our results show the effect of varying clusters and sample sizes on both methods. Also knowing that we are using log-normal shared frailty model, the observed difference shows that one method is more efficient, less bias and provides minimum MSE when compared to the other. In the simulation, we maintain the same levels of heterogeneity and censoring, and obtained the mean, absolute bias, standard deviation $SE(\hat{\beta})$, and the mean square error $MSE(\hat{\beta})$, likewise, for random frailty parameter, θ the mean, absolute bias, $SE(\theta)$, mean square error $MSE(\theta)$. All approaches (PFL and HL) produced on average similar estimates of the parameter β and θ. Tables 1,2, and 3 reported the mean, absolute bias, and the mean square error $MSE(\hat{\beta})$. Obviously, Table 1 shows the empirical means of the parameter estimates, which are bias and as well as shows some variability from the actual value. The degree of the variability becomes more stable as sample size increases, see Figures 2 and 3 respectively. More differences are seen in the estimates produced by PFL and HL methods, Figure 3. The parameters become more stable as sample size increases within clusters sizes. Table 2 and 3 highlight the effect of clusters (k) and sample sizes (n) on absolute bias and the MSE of the parameters respectively. It is shown that for a comparable setting, the absolute bias increases greatly, as cluster and sample size increases see Figure 6. In the presence of moderate censored data (35%), the MSE Figure 4 and 5 decreases greatly and continues further. In addition, juxtaposing the two methods under the same conditions as shown in Figure 4 which displays the effect of cluster sizes, the MSE values decrease as cluster size increases irrespective of sample size. This implies that Hl model provides the best estimates in terms of MSE than the PFL estimates of the log-normal shared frailty model. In addition, all used approaches are semi-parametric estimation methods that consider the baseline hazard as unknown. More importantly, the overview average simulation (convergence) time increases as sample and cluster sizes increase, Table 4. Generally, it takes less time to converge when using PFL method to the HL estimation method. Thereby showing that the PFL method is less time-consuming.

4. Application

4.1. Study population

The UK Biobank recruited approximately 502,000 men and women aged 37 - 73 years from the general population during the period of 2006 - 2010 [8]. Participants attended one of 22 assessment centers across England, Wales, and Scotland [30,35]. At the assessment centers, participants completed an electronically signed consent form, a touchscreen questionnaire and physical measurements, as previously described by [21,30,32,35]. Our study looks into the cohort of Participant with Diabetes Melitus and have complicated acute coronary syndrome ACS (STEMI and NSTEMI). The cohort study comprises in total 855 participants from UK Biobank who were diabetes with complications. From these, 366 participants had reported cases of MI and each were accessed for both STEMI and NSTEMI. Time (in months) from diagnosis of patients till death or end of study were obtained for each participant at both levels (STEMI and NSTEMI). To understand the concept of frailty among participant with ACS, we used Baseline characteristics factors which includes Age (years), and Sex, Glycated hemoglobin (HbA1c X mmol/mol) values, Date of first diabetes diagnosis, Date of MI, Date of death, Date of STEMI, Date of NSTEMI and censoring status taking value one (status = 1) if the participant is dead or observe STEMI and or NSTEMI and zero (status = 0) otherwise.

4.2. Result

Table 5 displays the estimated coefficients, standard errors, and p-values for various covariates obtained by applying the shared frailty model presented in equation (2.4) using both the PFL and HL methods. The covariates included in the table are HbA1c (a measure of blood sugar control), Age, Sex (with Male as the reference group), ACSSTEMI (a type of heart attack), and theta (frailty parameter in the model). The covariate HbA1c has a significant positive effect on the hazard of death, with estimated coefficients of 0.0259 and 0.0050 in the PFL and HL methods, respectively. This suggests that higher HbA1c levels are associated with an increased risk of death, although the effect is stronger in the PFL method. Age is not significantly associated with the hazard of death, with estimated coefficients of -0.0278 and 0.1218 in the PFL and HL methods, respectively, and p-values greater than 0.05. Being male is significantly associated with an increased hazard of death, with estimated coefficients of 0.3995 and 0.7142 in the PFL and HL methods, respectively, and p-values less than 0.001. The covariate ACSSTEMI is also significantly associated with an increased hazard of death, with estimated coefficients of 0.0493 and 0.0014 in the PFL and HL methods, respectively, and p-values less than 0.05. Finally, for the frailty parameter theta, in both methods (PFL and HL) are significantly associated with the hazard of death, that is (positive dependence between event times in the clusters of participants) with estimated coefficients of 6.7410 and 14.6100 in the PFL and HL methods, respectively, and p-values less than 0.001. These results suggest that there is significant variability among study participants that cannot be explained by the observed covariates.

Comparing the two Cumulative Baseline Hazard plots from PL and HL, both display the cumulative hazard over a 60-month period and beyond, signifying the total risk of an event occurring up to each point in time. The first plot (PFL-Figure 8a) features a stepwise increase but highlights a more pronounced escalation in risk as time progresses, particularly towards the later stages of the 60-month period. While second plot (HL-Figure 8b) shows a consistent increase in risk, with a stepwise pattern reflecting the occurrence of events and steady increments initially, followed by varying increments, indicating periods of fluctuating risk intensity. Both plots use rug plots at the bottom to mark individual event times. While the first plot (PFL) focuses on the gradual growth of hazard that becomes more significant over time, suggesting notable periods of higher risk later in the timeline, second plot (HL) emphasizes a consistent risk with periods of varying intensity. Table 6 provides prediction intervals for the estimates of the covariates using the PFL and HL estimation methods. The prediction intervals give a range of values within which we can expect the true value of the estimate to fall with a certain level of confidence. For the covariate HbA1c, the prediction interval for the PFL method is (0.0147, 0.0372), while for the HL method, it is (-0.0065, 0.0165). This indicates that the estimates of the effect of HbA1c on the hazard of death are relatively precise in the PFL method, while there is more uncertainty in the HL method. For the covariate Age, the prediction interval for the PFL method is (-0.0550, -0.0006), while for the HL method, it is (0.0937, 0.1501). This suggests that the estimates of the effect of Age on the hazard of death are more uncertain in the PFL method than in the HL method. For the covariate Sex-Male, the prediction interval for the PFL method is (0.1941, 0.6048), while for the HL method, it is (0.3886, 1.0398). This indicates that the estimates of the effect of being male on the hazard of death are relatively precise in both methods, with a larger effect size in the HL method. For the covariate ACSSTEMI, the prediction interval for the PFL method is (-0.0029, 0.0957), while for the HL method, it is (-0.1623, 0.1651). This suggests that the estimates of the effect of ACSSTEMI on the hazard of death are more uncertain in the HL method than in the PFL method. For the frailty parameter, theta, the prediction interval for the PFL method is (5.8436, 7.6390), while for the HL method, it is (10.5241, 18.6958). This indicates that the estimates of the frailty parameter are relatively precise in the PFL method, while there is more uncertainty in the HL method. The PFL and HL intervals for Age differ significantly, with the PFL interval ranging from -0.0550 to -0.0006 and the HL interval ranging from 0.0937 to 0.1501. For Sex-Male, the PFL interval ranges from 0.1941 to 0.6048, while the HL interval is much wider, ranging from 0.3886 to 1.0398. This suggests that there is greater uncertainty in the estimate of the effect of sex on the outcome variable when using the HL method. The PFL interval for the variable ACSSTEMI ranges from -0.0957 to -0.0029, indicating a high level of uncertainty, while the HL interval ranges from -0.1623 to 0.1651. Finally, for the frailty parameter (theta), the PFL interval ranges from 5.8436 to 7.6390, while the HL interval is wider and ranges from 10.5241 to 18.6958. This suggests that the PFL method provides a more precise estimate of the true value of θ.

5. Discussion

When we fixed the cluster and vary the sample sizes, to see the impact of sample size on the parameters, estimates obtained by HL using the log-normal shared frailty model are very high and increased as sample size increased compared to the cases of PFL which remains almost the same as sample size increases. Figure 3 clearly shows the effect of sample size on the parameter estimates and MSE (Figure 4) at fixed cluster size. It is obvious that the MSE (Figure 5) of the estimates obtained asymptotically reduces as sample size increases with HL showing the best method. On the other hand, one can remark that as sample size increases and at the highest given cluster (k=5) the $MSE(\hat{\beta}_1)$ of the PFL and HL are asymptotically the same but for the MSE of variance of frailty, $MSE(\theta)$ of PFL provides the minimum MSE compared to HL method, meaning that PFL provided the best estimates in terms of variance of frailty. It can also be remarked that in the presence of heterogeneity, and under normal censoring (35%) settings, the estimates of shared frailty models using the PFL model exhibit less MSE than the HL method. The average simulation time increases as sample and cluster sizes increase Table

3. Generally, it takes less time to run when using PFL methods than the HL estimation method. Thereby showing that the PFL method is less time-consuming. In Table 6, the prediction intervals for the HL method are wider than those for the PFL method for all covariates, except for HbA1c. This suggests that the HL method is associated with greater uncertainty in the estimates compared to the PFL method. However, it's important to note that the precision of the estimates should also be considered in addition to the width of the prediction intervals when evaluating the quality of the statistical analysis.

6. Summary

Despite its wide applications, which researchers have not fully appreciated its usefulness, this estimation method provides contemporary alternative method of estimating parameters of shared frailty model. Our approach of PFL estimation method which is almost the same as ordinary likelihood method has proved to be a useful method of computation in shared frailty model if some conditions are met. Using a UK Biobank data, we demonstrated herein, with strong evidence, a reasonably survival of participant from diabetes diagnose till either acute coronary syndrome (ACS), death or loss to follow up. There exists a high dependent within the clusters of participants with STEMI and NSTEMI and independent across the participant given an estimated value of $\theta > 0$. HbA1c are contributing significantly to the mortality rate from diagnose to STEMI or NSTEMI. This translated to higher risk among those taking Glucose. There is also a strong correlation between observed variable and unobserved variable as shown by the estimated correlation coefficients. Stability of our estimate is strongly supported by the prediction interval obtained in Table 6 as well as the trace plot (Figure 7) using PFL estimation method for semiparametric log-normal shared frailty model.

Acknowledgements

We would like to thank the editor and all the anonymous reviewers and specifically Prof. Alan Rigby for providing professional comments, and advice on the article. This study makes use of data obtained from UKBiobank.

Author contributions. All the co-authors have contributed equally in all aspects of the preparation of this submission.

Conflict of interest statement. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding. There is no support nor funding for the work

Data availability. Data used for this research can be found in UKBiobank as described in the article.

References

[1] K.A. Adeleke and G. Grover, *Parametric frailty models for clustered survival data: Application to recurrent asthma attack in infants*, J. Stat. Appl. Probab. **6**, 89-99, 2019.

[2] S.A. Adham and A.A. AlAhmadi, *Gamma and inverse Gaussian frailty models: A comparative study*, Journal of Mathematics and Statistics Invention, 2321-4767, 2016.

[3] T.A. Balan and H. Putter, *frailtyEM: An R package for estimating semiparametric shared frailty models*, J. Stat. Softw. **90**, 1-29, 2019.

[4] P. Barker and R. Henderson, *Small sample bias in the gamma frailty model for univariate survival*, Lifetime Data Anal. **11**, 265-84, 2005.

[5] N.J. Christian, I.D. Ha and J.H. Jeong, *Hierarchical likelihood inference on clustered competing risks data*, Stat. Med. **35** (2), 251-67, 2016.

[6] D. Clayton and J. Cuzick, *Multivariate generalizations of the proportional hazards model*, J. Roy. Statist. Soc. Ser. A **148**, 82-108, 1985.

[7] D.G. Clayton, *A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence*, Biometrika **65**, 141-51, 1978.

[8] R. Collins, *What makes UK Biobank special?*, Lancet **379** (9822), 1173-1174, 2012.

[9] DR. Cox, *Regression models and lifetables*, J. R. Stat. Soc. Ser. B. Stat. Methodol. **34**, 187-202, 1972.

[10] I.L. Do Ha, *On estimation of random effect in Poisson HGLMs*, J. Kor. Data Anal. Soc. **19** (1), 375-83, 2008.

[11] L. Duchateau and P. Janssen, *The Frailty Model*, Springer Verlag, New York, 2008.

[12] D.G. Enki, A. Noufaily and C.P. Farrington, *A time-varying shared frailty model with application to infectious diseases*, Ann. Appl. Stat. **8** (1), 430-447 2014.

[13] W.S. Gachau, Frailty models with applications in medical research: observed and simulated data, PhD Thesis, University of Nairobi, Nairobi, 2014.

[14] M. Gorfine, D.M. Zucker, and L. Hsu, *Prospective survival analysis with a general semiparametric shared frailty model: A pseudo full likelihood approach*, Biometrika **93** (3), 735-41, 2006.

[15] U.S. Govindarajulu, H. Lin, K.L. Lunetta and R.B. D'Agostino Sr, *Frailty models: applications to biomedical and genetic studies*, Stat. Med. **30** (22), 2754-64, 2011.

[16] I.D. Ha and G.H. Cho, *On prediction of random effects in log-normal frailty models*, J. Kor. Data Anal. Soc. **20**, 203-209, 2009.

[17] I.D. Ha, J.H. Jeong and Y. Lee, *Statistical Modelling of Survival Data with Random Effects: h-likelihood Approach*, Springer, 2017.

[18] I.D. Ha, Y. Lee and J.K. Song, *Hierarchical likelihood approach for frailty models*, Biometrika **88** (1), 233, 2001.

[19] I.D. Ha and Y. Lee, *Estimating frailty models via Poisson hierarchical generalized linear models*, J. Comput. Graph. Statist. **12** (3), 663-81, 2003.

[20] D.D. Hanagal and R. Sharma, *Analysis of diabetic retinopathy data using shared inverse Gaussian frailty model*, Model Assist. Stat. Appl. **8**, 103-19, 2013.

[21] F.K. Ho, S.R. Gray, P. Welsh, F. Petermann-Rocha, H. Foster, H. Waddell, J. Anderson, D. Lyall, N. Sattar, J.M. Gill and J.C. Mathers, *Associations of fat and carbohydrate intake with cardiovascular disease and mortality: prospective cohort study of UK Biobank participants*, BMJ **368**, 2020.

[22] N. Keyfitz and G. Littman, *Mortality in a heterogeneous population*, Popul. Stud. **33**, 333-342, 1979.

[23] J.P. Klein, *Semiparametric estimation of random effects using the Cox model based on the EM algorithm*, Biometrics **1**, 795-806, 1992.

[24] Y. Lee, J.A. Nelder and Y. Pawitan, *Generalized Linear Models with Random Effects: Unified Analysis via H-Likelihood*, CRC Press, 2018.

[25] S. Mahmood, B. Zainab, and A.M. Latif, *Frailty modeling for clustered survival data: an application to birth interval in Bangladesh*, J. Appl. Stat. **40** (12), 2670-80, 2013.

[26] J.V. Monaco, M. Gorfine and L. Hsu, *General semiparametric shared frailty model: estimation and simulation with frailtySurv*, J. Stat. Softw. **86**, 2018.

[27] S.A. Murphy, A. W. Van der Vaart, *On profile likelihood*, J. Am. Stat. Assoc. **95** (450), 449-65, 2000.

[28] G.G. Nielsen, R.D. Gill, P.K. Andersen and T.I. Sørensen, *A counting process approach to maximum likelihood estimation in frailty models*, Scand. J. Stat. **1**, 25-43, 1992.

[29] A.W. Oyekunle, K.A. Adeleke and A.A. Olosunde, *Gamma and Inverse Gaussian Frailty Models with Time-varying co-variates Based on Some Parametric Baseline Hazards* , Afr. Stat. **15** (1), 2199-2224, 2020.

[30] L.J. Palmer, *UK Biobank: bank on it*, Lancet **369** (9578), 1980-1982, 2007.

[31] E. Parner, *Asymptotic theory for the correlated gamma-frailty model*, Ann. Stat. **26**, 183-214, 1998.

[32] F. Petermann-Rocha, S. Parra-Soto, S. Gray, J. Anderson, P. Welsh, J. Gill, N. Sattar, F.K. Ho, C. Celis-Morales, J.P. Pell, *Vegetarians, fish, poultry, and meat-eaters: Who has higher risk of cardiovascular disease incidence and mortality? A prospective study from UK Biobank*, Eur. Heart J. **42** (12), 1136-1143, 2021.

[33] S. Ripatti and J. Palmgren, *Estimation of multivariate frailty models using penalized partial likelihood*, Biometrics **56** (4), 1016-22, 2000.

[34] Y.R. Su and J.L. Wang, *Semiparametric efficient estimation for shared-frailty models with doubly-censored clustered data*, Ann. Stat. **44** (3), 1298, 2016.

[35] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray and B. Liu, *UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age*, PLoS Med. **12** (3), e1001779, 2015.

[36] F. Vaida and R. Xu, *Proportional hazards model with random effects*, Stat Med **19** (24), 3309-3324, 2000.

[37] J.W. Vaupel, K.G. Manton and E. Stallard, *The impact of heterogeneity in individual frailty on the dynamics of mortality*, Demography **16**, 439-54, 1979.

[38] A. Wienke, *Frailty Models in Survival Analysis*, Boca Raton: Chapman and Hall/CRC, 2010.

[39] X. Xue, *Multivariate survival data under bivariate frailty: an estimating equation approach*, Biometrics **1**, 1631-1637, 1998.

[40] N. Zare and F. Moradi, *Parametric frailty and shared frailty models applied to waiting time to first pregnancy*, International Conference on Applied Mathematics and Pharmaceutical Sciences, 598-600, 2012.

[41] D.M. Zucker, M. Gorfine and L. Hsu, *Pseudo-full likelihood estimation for prospective survival analysis with a general semiparametric shared frailty model: Asymptotic theory*, J. Statist. Plann. Inference **138** (7), 1998-2016, 2008.

Table 1. Effects of cluster size (k) on parameters estimates for HL and PFL methods.

	mean	n=300		n=200		n=100		n=50	
		HL	**PFL**	**HL**	**PFL**	**HL**	**PFL**	**HL**	**PFL**
k=2	$\hat{\beta}_1$	-4.219	-4.912	9.166	-4.574	-4.054	-4.719	2.864	-4.412
	$\hat{\beta}_2$	-0.948	-1.103	2.043	-1.060	-0.909	-1.058	1.110	-0.916
	θ	3.519	4.097	1.408	3.797	3.104	3.614	1.856	3.001
k=3	$\hat{\beta}_1$	-4.091	-4.762	9.080	-4.750	-4.038	-4.701	-3.943	-4.591
	$\hat{\beta}_2$	-0.935	-1.088	2.034	-1.089	-0.922	-1.073	-0.863	-1.005
	θ	3.480	4.051	1.441	4.006	3.352	3.902	3.036	3.534
k=4	$\hat{\beta}_1$	-4.567	-4.762	9.386	-4.762	-4.381	-4.568	-3.913	-4.555
	$\hat{\beta}_2$	-1.043	-1.088	2.115	-1.088	-1.013	-1.056	-0.865	-1.007
	θ	3.885	4.051	1.641	4.051	3.633	3.789	3.136	3.651
k=5	$\hat{\beta}_1$	7.967	-3.829	9.274	-4.457	-4.256	-4.438	-4.261	-4.444
	$\hat{\beta}_2$	1.803	-0.897	2.100	-1.044	-1.001	-1.044	-0.978	-1.020
	θ	1.318	3.166	1.535	3.686	3.489	3.638	3.332	3.474

Table 2. Absolute bias of the parameter estimates for different values of n and k.

	Absolute Bias	n=300		n=200		n=100		n=50	
		HL	**PFL**	**HL**	**PFL**	**HL**	**PFL**	**HL**	**PFL**
k=2	$\hat{\beta}_1$	9.912	9.219	9.850	9.166	9.719	9.054	9.412	2.136
	$\hat{\beta}_2$	2.202	2.046	2.198	2.043	2.156	2.007	2.014	0.012
	θ	2.097	1.519	1.967	1.408	1.614	1.104	1.001	0.144
k=3	$\hat{\beta}_1$	9.762	9.091	9.750	9.080	9.701	9.038	9.591	8.943
	$\hat{\beta}_2$	2.187	2.033	2.187	2.034	2.172	2.021	2.104	1.962
	θ	2.051	1.480	2.006	1.441	1.902	1.352	1.534	1.036
k=4	$\hat{\beta}_1$	9.762	9.567	9.574	9.386	9.568	9.381	9.555	8.913
	$\hat{\beta}_2$	2.187	2.142	2.159	2.115	2.154	2.111	2.106	1.964
	θ	2.051	1.885	1.797	1.641	1.789	1.633	1.651	1.136
k=5	$\hat{\beta}_1$	2.967	8.829	9.457	9.274	9.438	9.256	9.444	9.261
	$\hat{\beta}_2$	0.705	1.995	2.142	2.100	2.143	2.100	2.118	2.076
	θ	0.682	1.166	1.686	1.535	1.638	1.489	1.474	1.332

Table 3. MSE of parameter estimates in terms of sample and cluster sizes.

		PFL				HL			
Sample size		**50**	**100**	**200**	**300**	**50**	**100**	**200**	**300**
k=2	MSE($\hat{\beta}_1$)	1.961	1.408	0.966	0.791	1.684	1.350	0.927	0.759
	MSE($\hat{\beta}_2$)	1.011	0.720	0.502	0.407	0.869	0.690	0.482	0.390
	MSE(θ)	3.465	2.634	1.858	1.533	2.976	2.526	1.782	1.470
k=3	MSE($\hat{\beta}_1$)	1.119	0.767	0.535	0.435	0.961	0.736	0.513	0.417
	MSE($\hat{\beta}_2$)	0.681	0.470	0.328	0.266	0.585	0.451	0.314	0.255
	MSE(θ)	2.096	1.510	1.052	0.854	1.800	1.448	1.009	0.819
k=4	MSE($\hat{\beta}_1$)	0.934	0.572	0.390	0.435	0.803	0.549	0.374	0.417
	MSE($\hat{\beta}_2$)	0.581	0.372	0.255	0.266	0.499	0.357	0.245	0.255
	MSE(θ)	1.938	1.143	0.782	0.854	1.665	1.096	0.750	0.819
k=5	MSE($\hat{\beta}_1$)	0.698	0.464	0.320	0.009	0.669	0.445	0.306	0.008
	MSE($\hat{\beta}_2$)	0.452	0.313	0.216	0.401	0.433	0.300	0.207	0.385
	MSE(θ)	1.478	0.962	0.665	0.019	1.417	0.922	0.637	0.118

Table 4. Simulation runtime for the methods.

	300	**200**	**100**	**50**
PFL				
k2	(227.54ś52.12)	(96.42ś20.76)	(24.62ś5.67)	(6.10ś1.64)
k3	(332.03ś63.4)	(146.93ś30.09)	(36.20ś8.66)	(10.28ś3.28)
k4	(332.74ś63.40)	(192.57ś41.38)	(50.15ś10.34)	(14.59ś3.88)
k5	(333.43ś54.98)	(290.00ś139.65)	(58.17ś11.54)	(17.41ś4.77)
HL (HL)				
k2	(421.46ś96.68)	(178.46ś37.18)	(44.61ś9.29)	(11.34ś3.05)
k3	(617.18ś18.77)	(271.41ś28.77)	(66.92ś14.87)	(19.11ś6.09)
k4	(618.93ś39.21)	(356.93ś16.21)	(92.95ś18.59)	(27.12ś7.22)
k5	(619.05ś46.39)	(539.12ś20.71)	(107.82ś20.44)	(31.60ś7.44)

Table 5. Estimates of PFL and HL estimation methods.

Covariate	**PFL**			**HL**		
	Estimate	Std. Error	p-value	Estimate	Std. Error	p-value
HbA1c	0.0259	0.0117	0.0269	0.0050	0.0058	< 0.001
Age	-0.0278	0.0283	0.3260	0.1218	0.0144	< 0.001
Sex-Male	0.3995	0.2139	0.0618	0.7142	0.1661	< 0.001
ACSSTEMI	0.0493	0.0483	0.0080	0.0014	0.0835	< 0.001
θ	6.7410	0.9351	< 0.001	14.6100	2.0846	< 0.001

Table 6. Prediction interval of the estimates.

Covariate	**PFL**		**HL**	
	Lower	Upper	Lower	Upper
HbA1c	0.0147	0.0372	-0.0065	0.0165
Age	-0.0550	-0.0006	0.0937	0.1501
Sex-Male	0.1941	0.6048	0.3886	1.0398
ACSSTEMI	-0.0029	0.0957	-0.1623	0.1651
θ	5.8436	7.6390	10.5241	18.6958

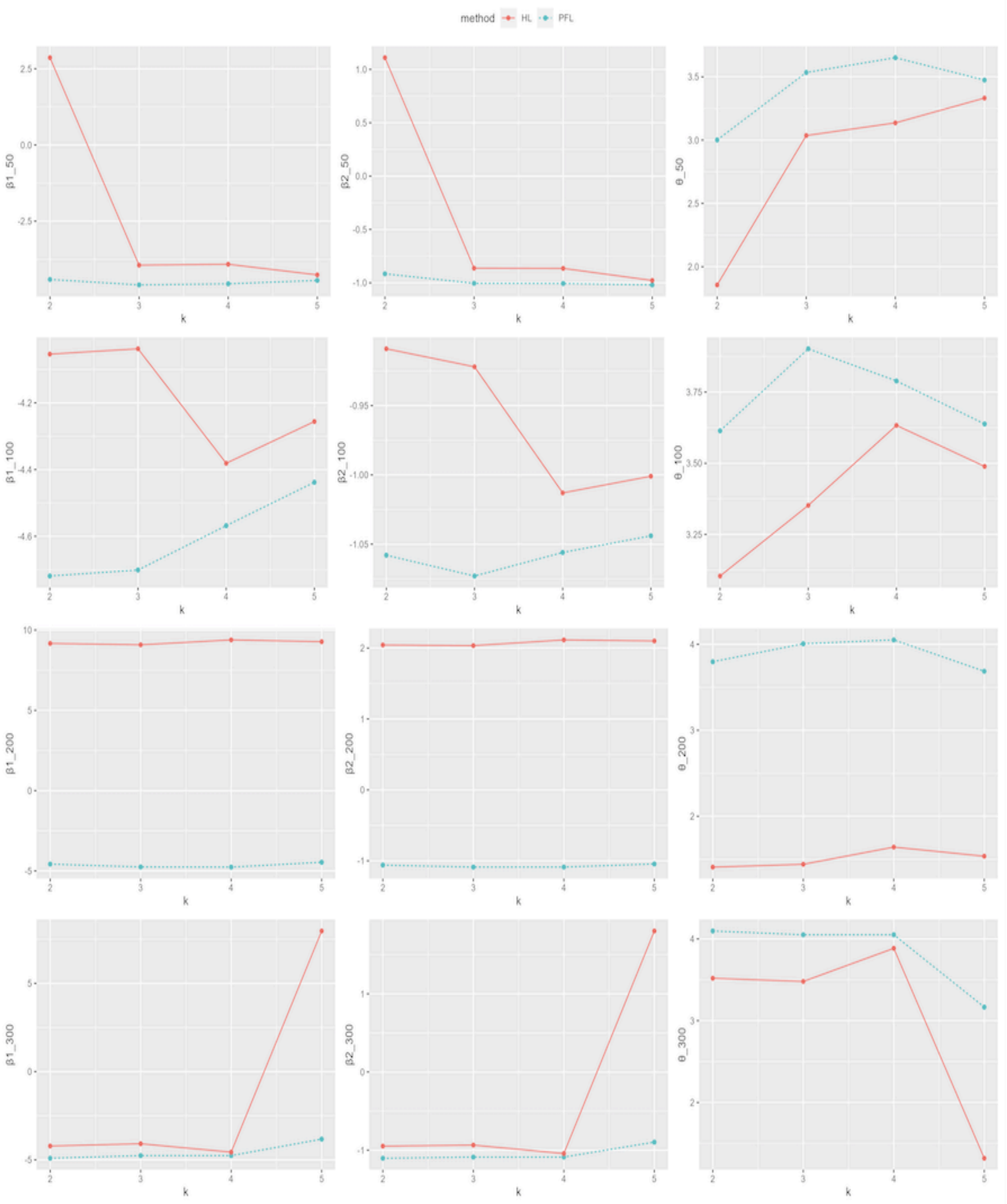

Figure 2. Effect of cluster size on parameter estimates.

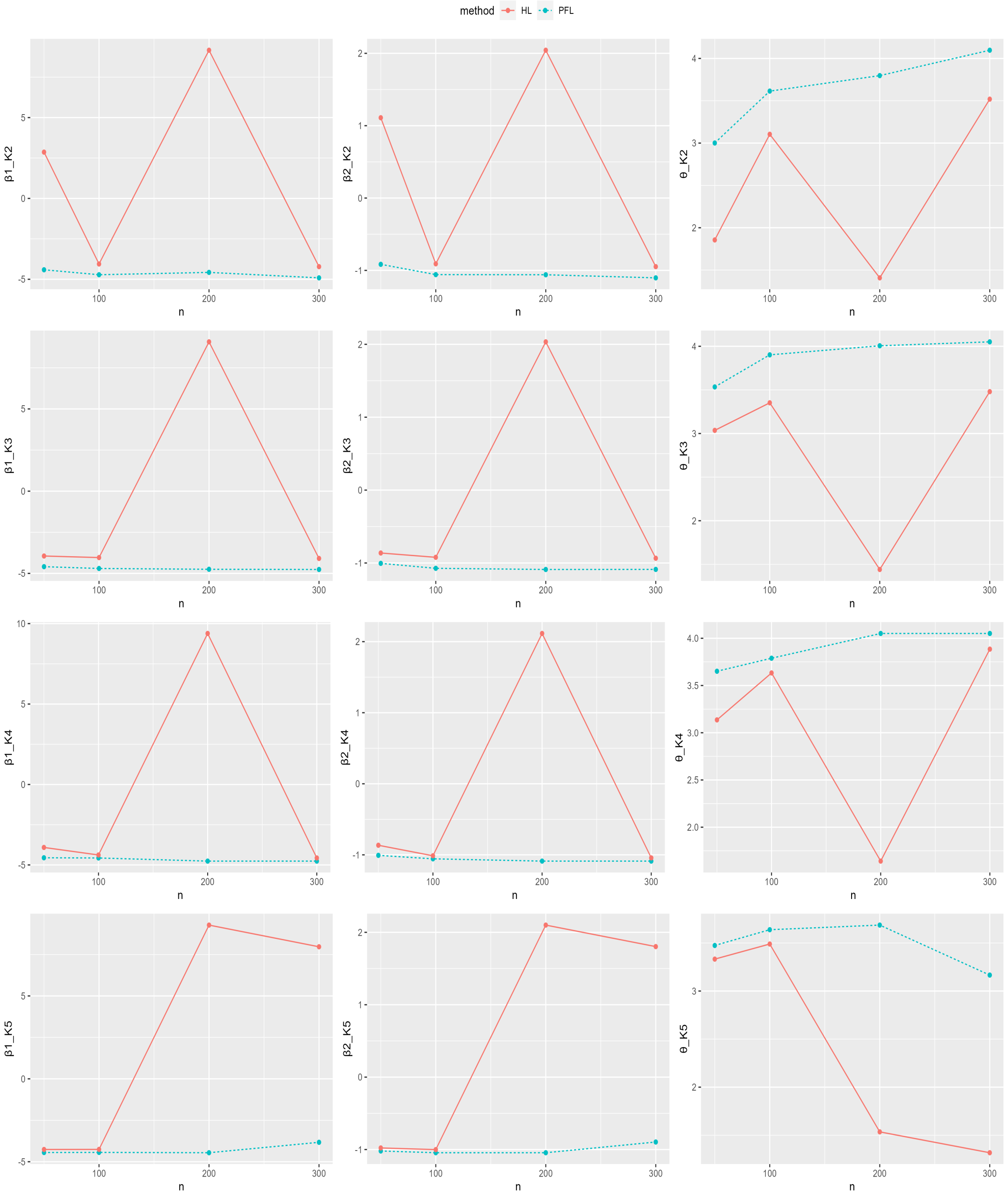

Figure 3. Effect of sample size on parameter estimates.

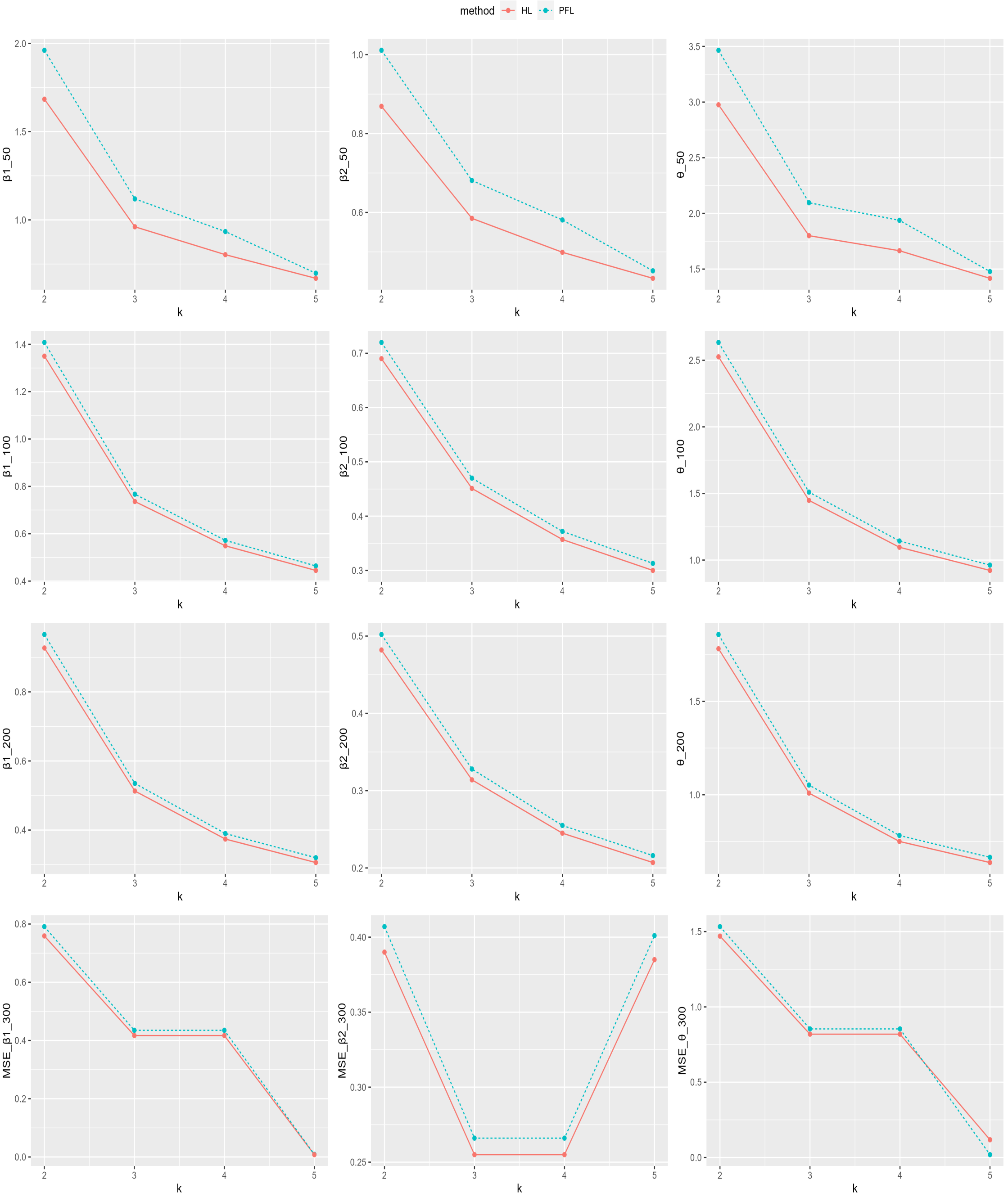

Figure 4. MSE and cluster size.

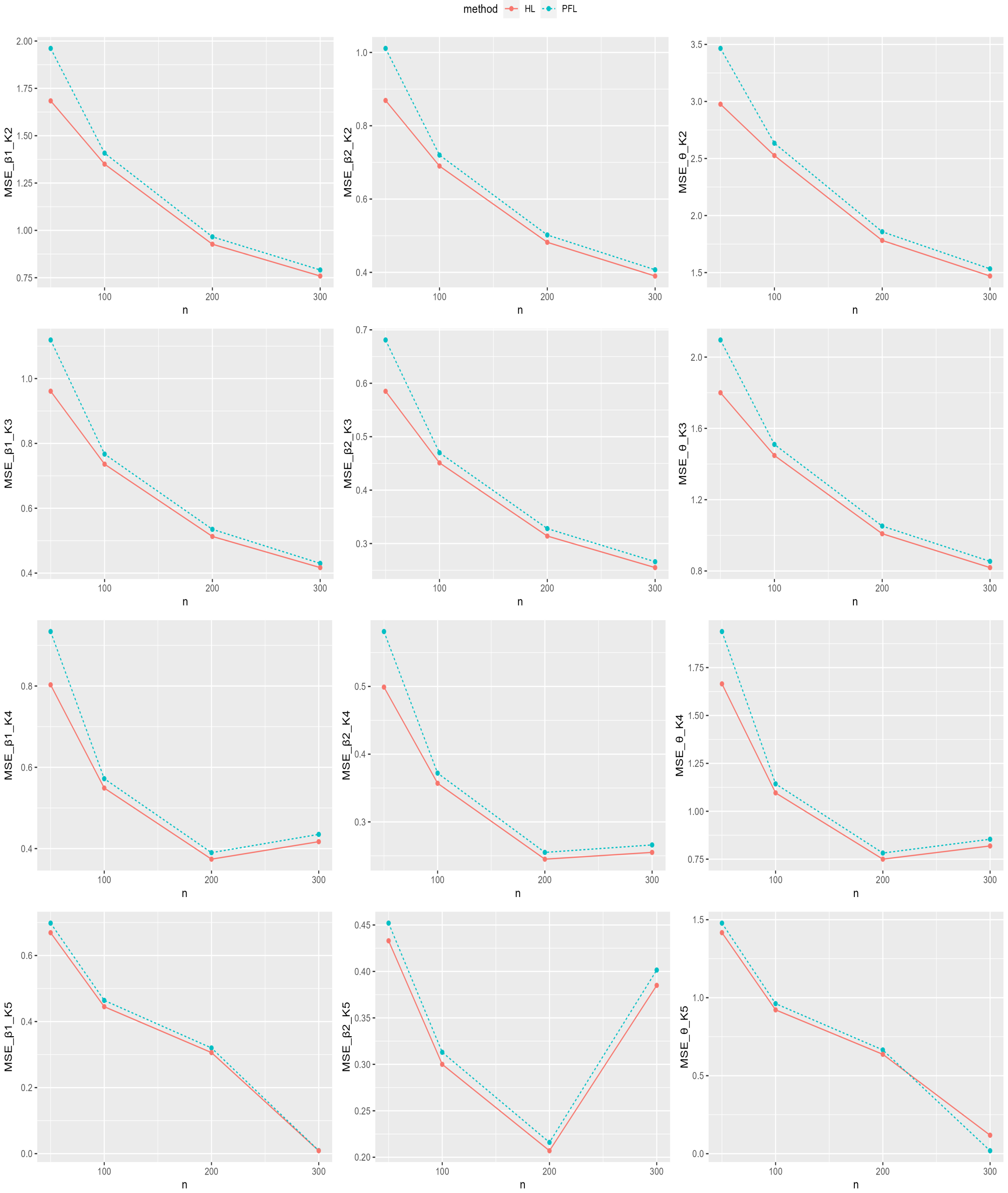

Figure 5. MSE and sample sizes.

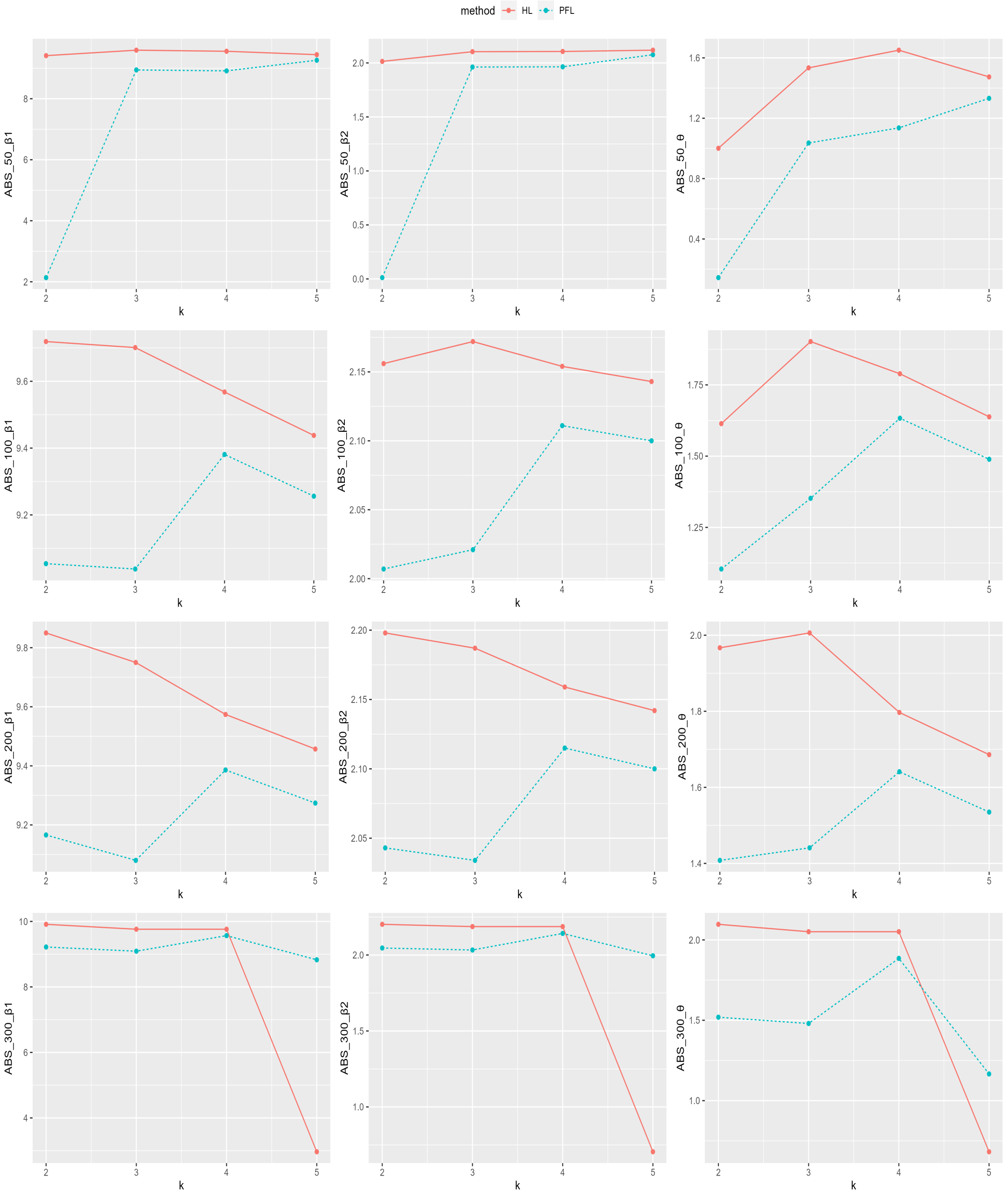

Figure 6. Absolute bias of estimates based on sample sizes and clusters.

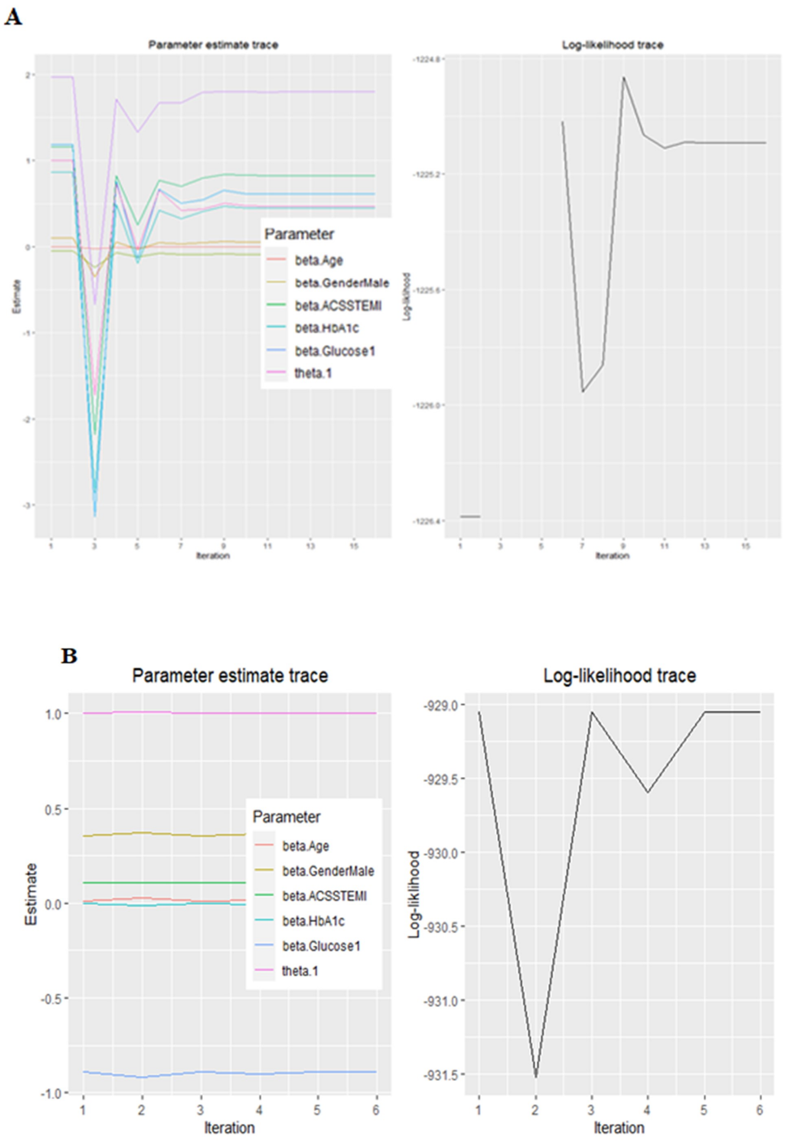

Figure 7. Parameter estimate trace and parameter log-likelihood trace using PFL (A) and HL (B).

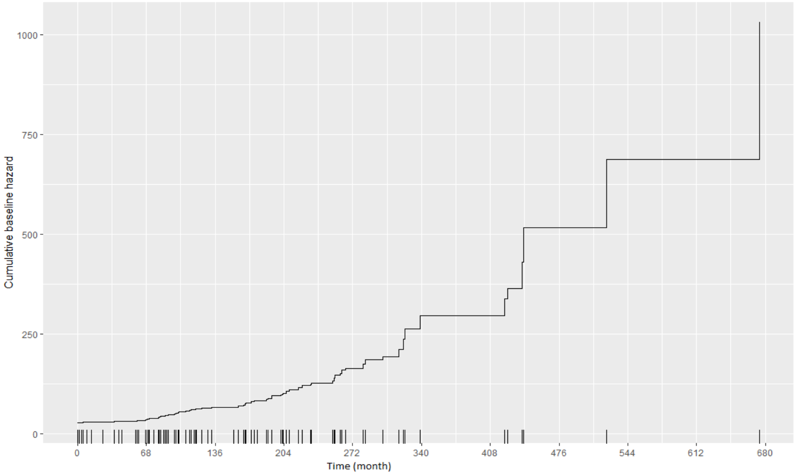

Figure 8a

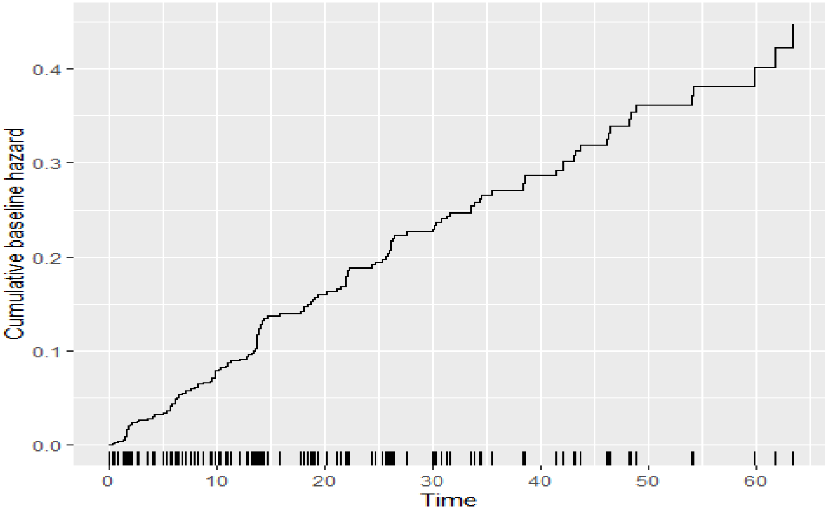

Figure 8b:

Figure 8. Cummulative Baseline Hazards plot for PFL (a) and for HL (b).