Bayesian Competing Risk Model for Medical Data

Nadya Devana[1], Sarini Abdullah[1*]

[1] *Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Indonesia, Depok, Indonesia*
[*] *Corresponding author. Email: sarini@sci.ui.ac.id*

ABSTRACT
Problems considering group assignment is often found in health area, where the groups represent whether a patient will be recovered or not, at high risk of relapse or not, and many other. While the occurrence of these events could be modelled using classification methods, more insights on the time of occurrence is cannot be provided. Thus, a more comprehensive method is required, which could be answered by the survival model. Competing risk model is one of the statistical methods that could be used for modelling the occurrence of several competing events, in which we can produce not just the probability of the events occurrence, but more specifically the probability of that events at a given time period. In this study, we propose the use of Bayesian competing risk model to predict whether a patient will give birth under the complication of Pre-Eclampsia (PE) condition, for a given gestational time. Data on patients in their first trimester of pregnancy from one of the hospitals in Jakarta were used in the analysis. Non-informative prior distributions were set for the parameters, data were assumed to follow a Weibull distribution, and upon obtaining the posterior distribution, Markov Chain Monte Carlo (MCMC) was implemented for posterior sampling. The result showed fast convergence, as only 30,000 iterations are required to achieve it, and several important predictors were identified.

Keywords: *Bayesian approach, flat prior, gestational time, MCMC, survival analysis.*

1. INTRODUCTION

In medical research, there are many problems in which the analysis requires classification methods. To exemplify, predicting whether people with certain characteristics is at a high or low risk of diabetes [1], or heart disease [2], identification of factors that could explain high or low likelihood of having cancer [3] [4] [5], and understanding factors contributing to development of pneumonia [6]. With the promising progress on computational technology and machine learning methods, implementing these models might aid clinicians in order to determine suitable treatment, or intervention of treatment, to direct the patients in a more favorable condition.

There are several classification methods that are often used in medical problems, including decision tree (DT) [1] [7] which is user-friendly, in the sense that it does not require a complex mathematical formulation for the interpretation. Thus, DT is preferred to other classification methods, and commonly applied in medical research. However, DT is prone to overfit. As a remedy, Random Forest (RF) [8] offers a more robust approach that by assembling the DTs, resulting in high accuracy and less likely to be overfit. Nevertheless, RF cannot provide the assignment rule, and the amount of contribution of each predictor has on the classification cannot be measured as in regression. Whereas in medical problems, the main concern is not always about the accuracy. Instead, it is important to measure the magnitude of effect of each variable in determining the class assignment. For this concern, logistic regression method [9] [10] can be the solution. It can measure the effect of each variable on the classification, also can be used to calculate specific probability of each observation against a group if certain conditions are given. Even its accuracy can surpass the RF's accuracy [11].

While being able to produce the probability of being a member of a certain group (or class), logistic regression cannot the answer the need for the probability that an object will be in a certain class until a certain time period, that is, the incorporating the time, known as time to event, where the event is said to happen as the object is assigned in a specified class. Survival analysis is a suitable method for this purpose [12]. Survival analysis can combine the effects of several factors that might explain the duration of time until something happens.

To accommodate the effect of other factors in explaining time until the event occurs, it is necessary to consider the regression method. There are several regression methods in survival analysis, one of the most used is Cox Proportional Hazard (Cox PH) regression [13]. Cox PH is often used because of its flexibility, where we do not need to determine the time distribution, which is not always easy in real application data.

However, the Cox PH model cannot provide metrics, such as survival function that represents the probability of an event to occur after a certain time interval. This condition is unfavorable for some medical research, in which the aim is not just being able to measure the effect of the contributing factors, but more importantly is to make predictions. Parametric approach such as the Accelerated Failure Time (AFT) model may serve this purpose. Yet, the time to events is only considered as uncensored data (observed events) or censored data (unobserved events). While in some medical cases, it is possible that there is more than one events to consider and the events are competing in the sense that when one event occurs, the other will cease to happen. Implementing AFT model in this condition might be misleading.

Therefore, in this paper, we propose the competing risk model for analyzing medical data with competing events. Data on patients in Obstetrics department from one of the hospitals in Jakarta, were used in the analysis. The competing events are pre-eclampsia (PE) and non-PE. Preeclampsia is an obstetric syndrome with new-onset hypertension accompanied by protein in the urine after 20 weeks of gestation. A pregnant woman has two possibilities, between giving birth before or after the development of preeclampsia. If the analysis was carried out using the AFT approach, cases of women who gave birth not in a preeclampsia condition (censored data) were considered to have preeclampsia if the observation time was extended. Meanwhile, a woman who is pregnant has a maximum time limit for her pregnancy. It is impossible for the gestation period to be extended beyond the maximum time. Therefore, this concept is not appropriate in cases of preeclampsia.

To be able to estimate the survival function we must be able to estimate the regression coefficients or regression parameters in the model. Frequentist approach, such as maximum likelihood estimator (MLE) is commonly employed. However, the outcome from MLE is only based on information from data. Bayesian method provides a more comprehensive approach and having a more flexible and robust results. Bayesian approach uses additional information which is a prior distribution that represents expert judgment. If only based on the data, it will be at risk of misleading information when data validity is questioned, or when the data is noisy. Combining expert judgment through prior distribution with new data will provide posterior estimates that include a more complete explanation, so that the model will be further improved.

Considering the advantages provided by Bayesian method, in this paper we will propose a competing risk model that implements the Bayesian method for parameter estimation. The proposed method will be implemented to analyze preeclampsia data. The aim is to be able to estimate the survival function of whether the patient will give birth in a preeclampsia or non-preeclampsia condition given the gestational age.

2. MATERIAL AND METHODS

2.1. Dataset

The dataset was retrieved from Obstetrics and Gynecology department at one of the hospitals in Jakarta, Indonesia. This dataset consists of 946 observations and 27 features. However, this study only focuses on data where the pregnancy outcome is live birth, resulting in 933 observations.

On the preeclampsia feature, there are 65 records for patient who develop preeclampsia. These features will be utilized to predict preeclampsia. The features of the dataset are as follows:

Table 1. Preeclampsia Dataset

No	Feature	Description	Data Type
1	First Pregnant	{'No'=0, 'Yes'=1}	Categorical
2	Class Age >40	{'Patient's age <40 years' = 0, 'Patient's age >40 years' = 1}	Categorical
3	Class BMI >35	{'Patient's BMI <35 kg/m^2' = 0, 'Patient's BMI >35 kg/m^2' = 1}	Categorical
4	Conception	{'Spontaneous' =0, 'IVF'=1}	Categorical
5	Previous Gestational HT	{'No'=0, 'Yes'=1}	Categorical
6	Previous PE	{'No'=0, 'Yes'=1}	Categorical
7	Renal Disease	{'No'=0, 'Yes'=1}	Categorical
8	SLE	{'No'=0, 'Yes'=1}	Categorical
9	APS	{'No'=0, 'Yes'=1}	Categorical
10	Diabetes Mellitus Type 1	{'No'=0, 'Yes'=1}	Categorical
11	Diabetes Mellitus Type 2	{'No'=0, 'Yes'=1}	Categorical
12	Chronic HT	{'No'=0, 'Yes'=1}	Categorical
13	Any Family	{'No'=0,	Categorical

	History of PE	'Yes'=1}	
14	Previous Gestational Diabetes	{'No'=0, 'Yes'=1}	Categorical
15	Cardiac Disease	{'No'=0, 'Yes'=1}	Categorical
16	Smoking	{'No'=0, 'Yes'=1}	Categorical
17	Use of Aspirin	{'No'=0, 'Yes'=1}	Categorical
18	Use of Anti HT Drug	{'No'=0, 'Yes'=1}	Categorical
19	Use of Insulin	{'No'=0, 'Yes'=1}	Categorical
20	CRL	According to the fetus's Crown-Rump Length (mm)	Numerical
21	Pregnancy Outcome	{'Live Birth' = 1, 'Intrapartum' = 2, 'Neonatal' = 3, 'Stillbirth' = 4}	Categorical
22	Weeks Gestational Delivery	Gestational Age of Delivery in Weeks	Numerical
23	Preeclampsia	{'No'=0, 'Yes'=1}	Categorical
24	MAP	Mean arterial pressure	Numerical
25	UtA-PI	Mean uterine artery pulsatility index	Numerical
26	Opthalmica	Opthalmica	Numerical
27	PLGF Concentration	Serum placental growth factor	Numerical

2.2. Bayesian Competing Risk Model

A study whose analysis involves time-to-event data is called a survival analysis. The purpose of survival analysis is to measure the time between the time of diagnosis and the time of the event occurred. When more than one event (failure) is considered, one of the statistical problems that can be characterized is competing risk. Competing risk occurs when there is at least more than one cause of failure, but only one type of failure can actually occur. The purpose of competing risk data is to assess the relationship between related predictors and events, or the corresponding survival probabilities of events where competing risk of other means to fail [12].

Summary information over time survival experiences in survival analysis is to look at the survival curves. The most widely used empirical method for estimating survival curves is the Kaplan Meier (KM) approach. However, this curve is described only for situations when there is only one event to analyze. In the case of competing risks, the KM survival curve may not be as informative as when there is only one risk, because it is based on an independent assumption about competing risks that cannot be verified. One alternative that can be used in the case of competing risk is the Cumulative Incidence Function (CIF) which involves marginal probability. CIF is derived from cause-specific hazard function [14].

Let T_c be a random variable representing the duration of observation (time-to-event) of the cause c, for $c = 1, 2, \dots, C$, where c considers as censored observations. For competing risk data, time observed to the earliest cause $T = min\ (T_1, T_2, \dots, T_c)$ and indicator $\delta = c$, if $T = T_c$.

The cause-specific hazard function for cause c is defined by:

$$h_c(t) = \left\{\frac{P(t \le T < t + \Delta t, \delta = c | T \ge t)}{\Delta t}\right\} \quad (1)$$

The cumulative incidence function for cause c is defined as:

$$F_c(t) = P(T_c \le t) = \int_0^t \ h_c(u)\, S(u)\, du, \quad (2)$$

where $S(t)$ is the overall survival function, it is defined as:

$$S(t) = P(T_c > t) = exp\left\{-\sum_{c=1}^{C} \ \int_0^t \ h_c(u)\, du\right\} \quad (3)$$

Let the censoring indicator for the i-th individual's is defined as $\Delta_{ic} = 1$ if $\delta_i = c\varepsilon(1, 2, \dots, C)$ and $\Delta_{ic} = 0$ otherwise. The covariates are defined as X_{ic} with the proportional hazard assumption as:

$$h_c(X_{ic}, \beta_c) = h_{0c}(t_i) exp\ (\beta_{ic} X_{ic}) \quad (4)$$

with $h_{0c}(t_i) = \lambda_c\ \alpha_c\ t^{\alpha_c - 1}$ for event c specified as a Weibull baseline hazard function, where λ_c is the shape and α_c is the scale parameter. $\beta_c = (\beta_{1c}, \beta_{2c}, \dots, \beta_{ic})^T$ are the regression coefficients for each covariate, respectively, for $c = 1,2$.

Further, the cumulative incidence function $F_c(t)$ is extended as:

$$F_c(t) = \int_0^t h_c(X_{ic}, \beta_c) \exp\{-\sum_{c=1}^{C} \int_0^u h_l(X_{il}, \beta_l)\, dv\} du \quad (5)$$

Assuming independence among observations for each cause c, the likelihood function is then given by

$$L_c = \prod h_c(X_{ic}, \beta_c)$$

Non informative priors were set for the model parameter. $\beta_{jc} \sim N(0, \sigma_{jc}^2)$, with σ_{jc}^2 is set to a large value (relative to the scale of the corresponding covariate). For the shape parameter, a gamma distribution is set as the prior, that is $\lambda_c \sim \Gamma(a, b)$ where $a > 0, b > 0$. Multiplying the prior and likelihood produce the posterior, in which the parameters will be sampled.

This analysis was performed using R version 4.0.3 and the Markov chain Monte Carlo (MCMC) method is used to exemplify the competing risk modeling implemented in BUGS syntax.

3. ANALYSIS AND RESULT

Figure 1 contains the descriptive statistics of the patients in the study sample for categorical data and Figure 2 for numerical data.

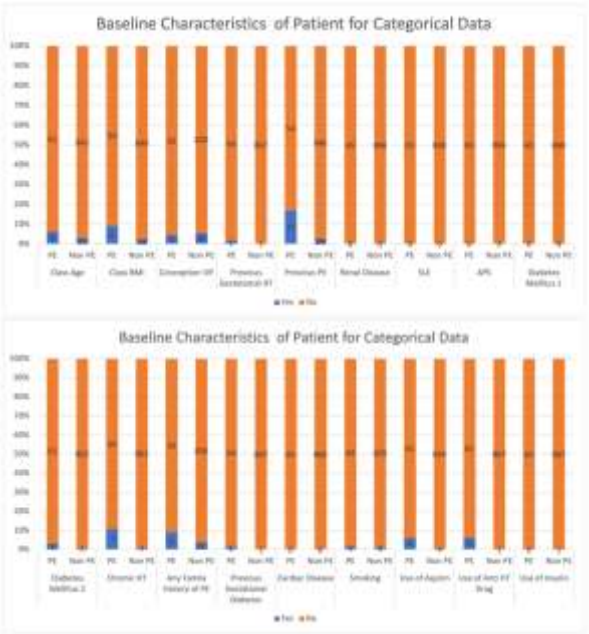

Figure 1 Statistical Descriptive for Categorical Data

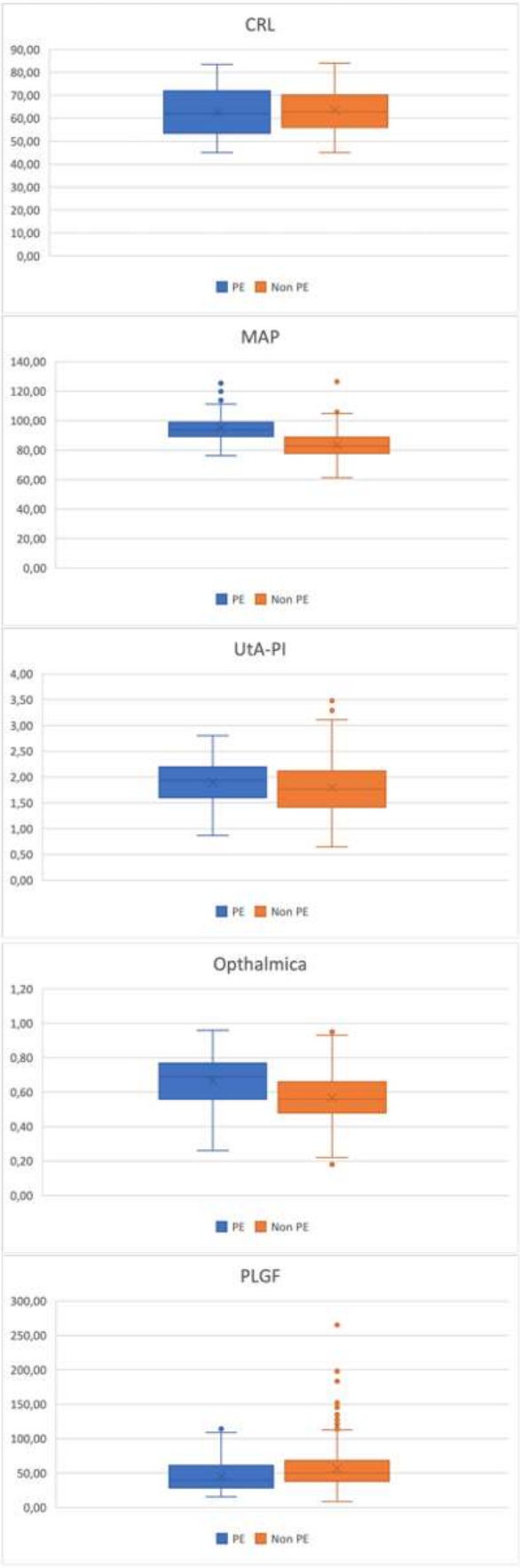

Figure 2 Statistical Descriptive for Numerical Data

As given in Figure 1, in this dataset there are several features that have the same value in all observations, such as Renal disease, SLE, APS, Diabetes Mellitus I, Cardiac Disease, and Use of Insulin. So it will impact the model and also we can't take any insights from these features. Therefore, we will drop these features from the dataset to get a better model.

A total of 30,000 iterations are considered. The posterior summaries of mean, standard deviation (SD), and 95% confidence intervals of regression parameter are given in Table II. It is defined as β_{i1} for non-preeclampsia cases and β_{i2} for preeclampsia cases.

Table II showed which variables give positive or negative risk for PE and non-PE cases. Furthermore, 95% credible interval for betas that between negative and positive values are having possibilities that the regression coefficient may be zero. Therefore, there is insufficient evidence that predictors can explain the incidence of preeclampsia.

Trace plots and density plots for each variable are given in Figure 3, Figure 4 and Figure 5. It shows that most of the trace plots is convergent which mean that the prior distribution is well-calibrated. The examples of the cumulative incidence curve for two different characters of pregnant women are given in Figure 6 and Figure 7.

Table 2 Posterior Estimates of The Parameter Observed

Parameter	Mean	SD	(2.5%,97.5%)
First Pregnant (h_1)			
β_{11}	-0.231	0.075	(-0.376, -0.085)
β_{12}	0.667	0.324	(0.048, 1.316)
Class Age (h_2)			
β_{21}	0.117	0.215	(-0.322, 0.521)
β_{22}	0.796	0.680	(-0.649, 1.998)
Class BMI (h_3)			
β_{31}	0.036	0.245	(-0.460, 0.493)
β_{32}	-0.068	0.511	(-1.126, 0.874)
Conception IVF (h_4)			
β_{41}	0.603	0.184	(0.233, 0.947)
β_{42}	-0.126	0.669	(-1.571, 1.039)
Previous PE (h_5)			
β_{51}	0.025	0.243	(-0.480, 0.481)
β_{52}	1.057	0.425	(0.214, 1.880)
Diabetes Mellitus Type 2 (h_6)			
β_{61}	0.940	0.503	(-0.135, 1.831)
β_{62}	0.811	0.990	(-1.366, 2.527)
Chronic HT (h_7)			
β_{71}	0.178	0.414	(-0.695, 0.928)
β_{72}	-0.436	0.545	(-1.544, 0.593)
Any Family History of PE (h_8)			
β_{81}	0.135	0.189	(-0.250, 0.489)
β_{82}	0.346	0.464	(-0.642, 1.203)
Smoking (h_9)			
β_{91}	0.039	0.355	(-0.712, 0.676)
β_{92}	1.101	1.309	(-1.990, 3.112)
Use of Aspirin (h_{10})			
β_{101}	-1.178	0.556	(-2.376, -0.204)
β_{102}	1.645	0.639	(0.281, 2.796)
Use of Anti HT Drug (h_{11})			
β_{111}	-1.602	1.300	(-4.740, 0.341)
β_{112}	1.004	0.680	(-0.431, 2.246)
CRL (h_{12})			
β_{121}	-0.001	0.005	(-0.011, 0.009)
β_{122}	0.022	0.015	(-0.007, 0.052)
MAP (h_{13})			
β_{131}	-0.014	0.006	(-0.029, -0.002)
β_{132}	0.103	0.015	(0.073, 0.132)
UtA-PI (h_{14})			
β_{141}	-0.234	0.078	(-0.394, -0.086)
β_{142}	0.477	0.288	(-0.097, 1.038)
PLGF Concentration (h_{15})			
β_{151}	-0.001	0.002	(-0.004, 0.002)

β_{152}	-0.018	0.008	(-0.033, -0.003)
Opthalmica (h_{16})			
β_{161}	-0.093	0.281	(-0.667, 0.452)
β_{162}	1.523	1.000	(-0.463, 3.450)

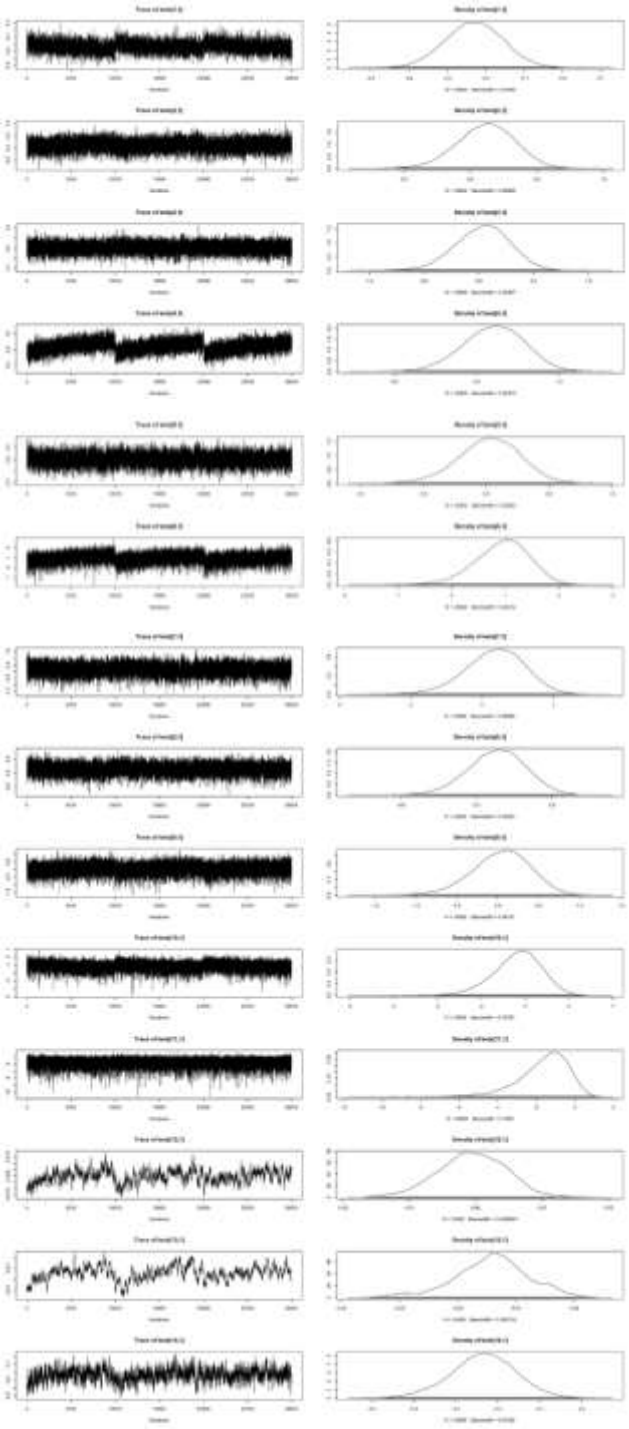

Figure 3 Trace Plots and Desity Plots for Each Variable (1)

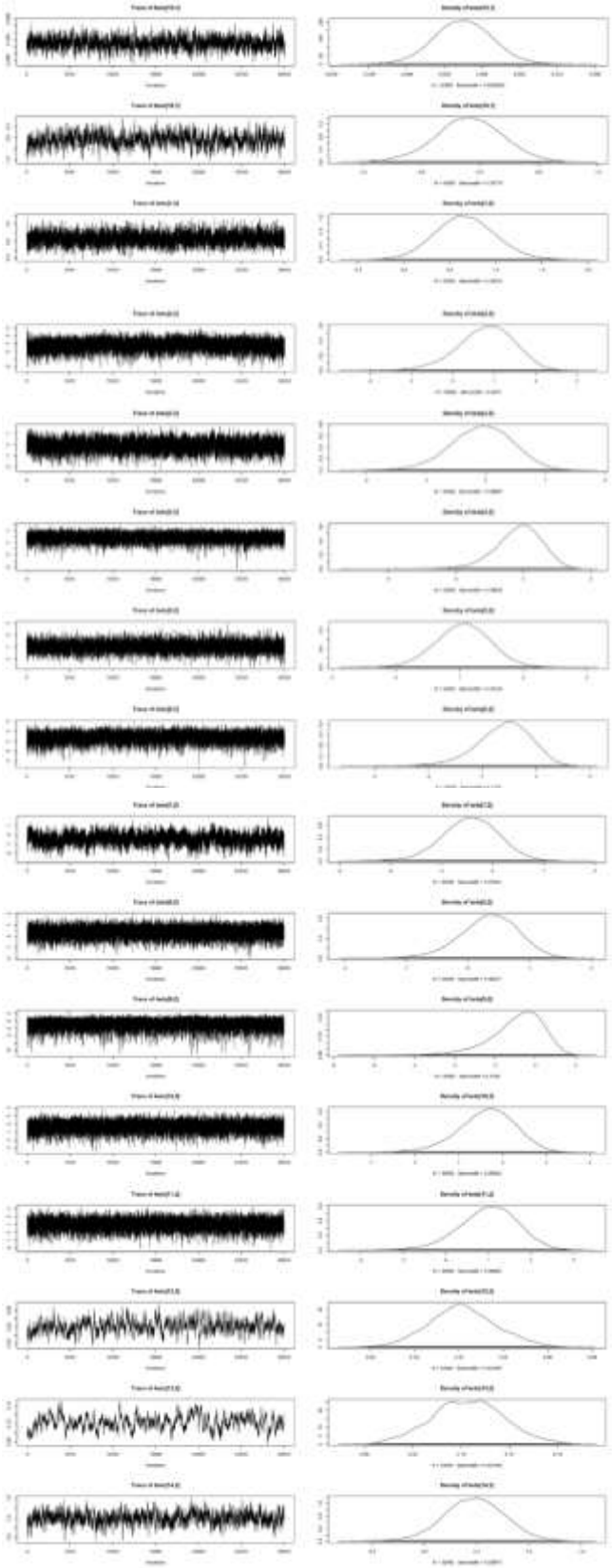

Figure 4 Trace Plots and Desity Plots for Each Variable (2)

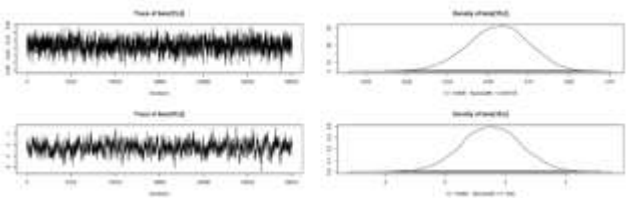

Figure 5 Trace Plots and Desity Plots for Each Variable (3)

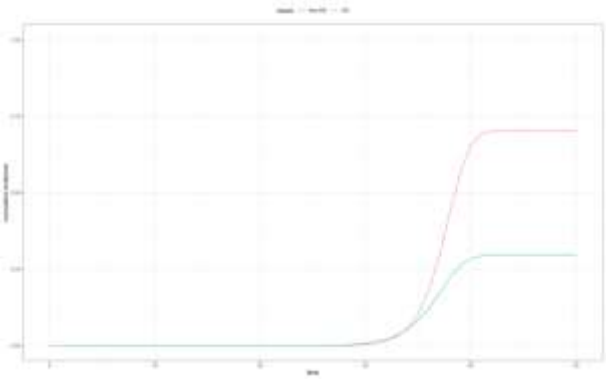

Figure 6 Cumulative Incidence of Preeclampsia for a specific cause (1)

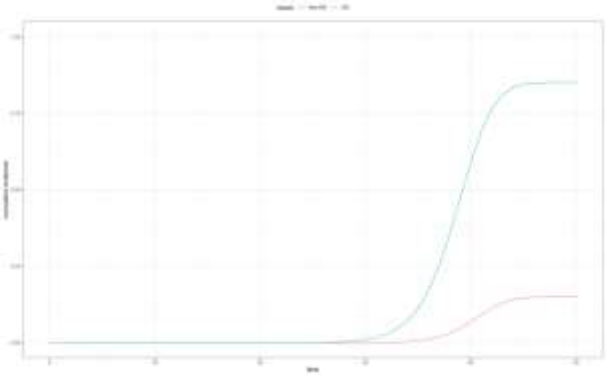

Figure 7 Cumulative Incidence of Preeclampsia for a specific cause (2)

4. CONCLUSION

In this study, competing risk with Bayesian approach has been proposed to predict the probability of someone having delivery with or without the development of preeclampsia based on maternal characteristics and Biomarkers. It shows that competing risk mode implemented with Bayesian approach is able to estimate the survival function on preeclampsia data. It can identify important factors explaining the delivery under preeclampsia condition and to produce personalized probability of delivery for a specific cause.

REFERENCES

[1] W. Chen, S. Chen, H. Zhang, T. Wu, A hybrid prediction model for type 2 diabetes using K-means and decision tree, in: Proceedings of the 8th International Conference on Software Engineering and Service Science (ICSESS), IEEE Press, Piscataway, NJ, 2017, pp. 386-390. DOI: 10.1109/icsess.2017.8342938

[2] S.S. Yadav, S. Jadhav, Deep convolutional neural network based medical image classification for disease diagnosis, in: Journal of Big Data, 2019, 6(113), pp. 1-18.

[3] I. Tougui, A. Jilbab, J. Mhamdi, Heart disease classification using data mining tools and machine learning techniques, in: Health Technol, 2020, 10, pp. 1137–1144.

[4] M. Amrane, S. Oukid, I. Gagaoua, T. Ensarİ, Breast cancer classification using machine learning, in: Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), IEEE Press, Piscataway, NJ, 2018, pp. 1-4.

[5] T.J. Brinker, A. Hekler, J.S. Utikal, N. Grabe, D. Schadendorf, J. Klode, et. al., Skin cancer classification using convolutional neural networks: systematic review, in: Journal of Medicine Internet Research, 2018, 20(10), pp. 1-8.

[6] G.S. Tandel, M. Biswas, O.G. Kakde, A. Tiwari, H.S. Suri, M. Turk, et. al., A review on a deep learning perspective in brain cancer classification, in: Cancers, 2019, 11(111), pp. 1-32.

[7] G. Ma, L. Zhang, G. Cui, Y. Cheng, Design of medical examination data mining system based on decision tree model, in: Journal of Physics: Conference Series, 2019, 1237(2), pp. 1-6.

[8] M.Z. Alam, M.S. Rahman, M.S. Rahman, A random forest based predictor for medical data classification using feature ranking, in: Informatics in Medicine Unlocked, 2019, pp. 1-12.

[9] M. Faisal, A. Scally, R. Howes, K. Beatson, D. Richardson, M.A. Mohammed, A comparison of logistic regression models with alternative machine learning methods to predict the risk of in-hospital mortality in emergency medical admissions via external validation, in: Health Informatics Journal, 2020, 26(1), pp. 34-44.

[10] T. Badriyah, M. Tahrir, I. Syarif, Predicting the risk of preeclampsia with history of hypertension using logistic regression and naive bayes, in: Proceedings of International Conference on Applied Science and Technology (ICAST), IEEE Press, Piscataway, NJ, 2018, pp. 399-403.

[11] K. Kirasich, T. Smith, B. Sadler Random forest vs logistic regression: binary classification for heterogeneous datasets, in: SMU Data Science Review, 2018, 1(3), pp. 1-24.

[12] D.G. Kleinbaum, M. Klein, Survival analysis - a self-learning text (Second Edition), Springer, USA, 2005.

[13] D.R. Cox, Regression models and life-tables, in: Journal of the Royal Statistical Society. Series B (Methodological), 1972, 34(2), pp. 187–220.

[14] D. Alvares, E. Lázaro, V. Gómez-Rubio, C. Armer, Bayesian survival analysis with BUGS, in: Statistics in Medicine, 2021, pp. 2975-3020.