JURNAL EMACS
e-ISSN: 2686-2573
[Engineering, MAthematics and Computer Science] Vol.5 No.3 September 2023: 181-187 DOI: 10.21512/emacsjournal.v5i3.10540

Bayesian Accelerated Failure Time Model for Risk Pregnancy Detection

Dennis Alexander[1], Sarini Abdullah[2], Adam Fahsyah Nurzaman[3]

[1] Mathematics Department, School of Computer Science,
[3] Information Systems Department, School of Information Systems,
Bina Nusantara University,
Jakarta, Indonesia 11480
[2] Faculty of Mathematics and Natural Sciences, Department of Mathematics,
Universitas Indonesia,
Depok, Indonesia 16424
dennis.alexander@binus.edu; sarini.abdullah@sci.ui.ac.id;
adam.nurzaman@binus.edu

*Correspondence: dennis.alexander@binus.edu

Abstract – *Preeclampsia (PE) also known as a hypertension during third trisemester of pregnancy. PE, is one of the most feared complications of pregnancy because it can potentially become serious complications in the future, including mother and fetus's death. The goal of this study is other than to have a bettter undestanding about risk factor in pregnancy by modelling the relationship between several factors and the time until deliveries under the PE condition. Data on 924 patients at obstetric and gynecology department in Jakarta were used in the analysis. Accelerated Failure Time (AFT) model was proposed to indentify some risk factors that influenced the condition. Model parameters were estimated using Bayesian method. Due to imbalance data, undersampling method will be used as a pre-procesing stage. Ratio between PE and non-PE data will be 60:40. Flat prior and posterior sample will be used using MCMC simulation with 12,000 iterations (including 2,000 iterations as a burnin stage) to get a convergen result. The iteration was repeated for 100 times so that the chosen data from undersampling was not error and biased. A consistent result for credible interval of the mean result was considered as the factors that affect PE condition consistently. From this study, there are two factors that have consistent Credible Interval result, Body Mass Index (BMI) and Mean Arterial Pressure (MAP).*

Keywords: *Censoring; Convergent; MCMC; BMI; MAP*

I. INTRODUCTION

Pregnancy is a phase in a woman's life where her body undergoes significant changes to support the growth and development of the baby she is carrying. The process begins when a fertilized egg is attached to the wall of the uterus by a sperm cell. For nine months, the organs in a woman's body adapt to provide a safe and optimal place for the fetus to grow and develop. However, there is a condition that may occur in the last trimester of pregnancy called Preeclampsia (PE). PE is a condition that can occur in pregnant women when the gestational age reaches 20 weeks or more. According to the International Society for the Study og Hypertension in Pregnancy (ISSHP), one of the common symptoms in PE is that the patient has high blood pressure or hypertension with categories ≥ 140 mmHg systolic and ≥ 90 mmHg diastolic(Brown et al. 2018). The National Institute for Health and Care Excellence (NICE) argues that PE can be caused by several factors, such as having high blood pressure in previous pregnancies, diabetes, pregnancy over the age of 40, having a body mass index (BMI) ≥ 35 kg/m, or a family history of PE (UK, 2019). Until now, there is no definite cure for PE, but giving aspirin to pregnant women is a best practice because it is considered quite effective in reducing the risk of PE if identified before 16 weeks of pregnancy (Rolnik et al, 2022).

The total number of PE cases in Indonesia is relatively small, which is around 5-10% of the total cases of pregnant women (Wirakusuma et al, 2019). However, one-third of PE cases lead to premature births that are potentially dangerous for the baby, such as increased chances of cerebral palsy, respiratory disorders, hypertension, insulin immunity, or even obesity. Moreover, mothers with PE have two to five

times the potential for hypertension, cardiovascular, and cerebrovascular conditions (Rolnik et al, 2022).

Actually, the condition of PE can be predicted by using a risk scoring system based on the medical conditions of pregnant women. However, risk scoring has shortcomings including less effective performance in predicting PE and cannot measure the specific risk of individual patients (Syaharutsa & Purwosunu, 2018). In addition to the risk scoring method, there is a logistic regression method that can be used to measure the specific risk of each patient individually but has the disadvantage of lacking flexibility in choosing different gestational ages to categorize the severity level of PE (Badriyah et al, 2018).

In addition to these two methods, there is a method that can predict the time until an event occurs or what is often referred to as survival analysis (Wright et al, 2020). This study uses data on PE in pregnant women until delivery, where the condition during labor can experience PE (complete data) or without PE condition (censored data) (Xue et al, 2020). For conditions like this, where only one event succeeds and other events are declared failures, it is called survival analysis or often also referred to as failure time analysis. One of the most common models in survival analysis is Accelerated Failure Time (AFT) where this model aims to predict a failure value of an event by taking into account censored and uncensored events (Alvares et al. 2021).

The Bayesian method is applied in estimating model parameters. This is done by considering the flexibility of the Bayesian estimation results. The Bayesian method is not only based on data alone, but also considers the opinions of experts in related fields, in this case through the use of a prior distribution as an initial guess. Thus, it is expected that the combination of the prior with information from the data can obtain more complete (updated) information, namely through the posterior distribution (Hu et al, 2021).

The data of pregnant women used in this study were obtained from an obstetrics and gynecology department of one of the hospitals in Jakarta, namely 924 patients with censored data (giving birth without PE condition) as many as 860 patients and uncensored data (giving birth with PE condition) as many as 64 patients. Due to the censored data ≥ 90%, before forming the model, it is necessary to pre-process the data. The processing is done by taking some random data from the censored group (non-PE group) or what is often referred to as undersampling so that the ratio between uncensored and censored data is 60% and 40%. The process will be repeated up to 100 times to anticipate the possibility of errors in data selection.

Based on the explanation listed above, this study is focused on some factors that may influence PE condition consistently based on secondary data from one hospital in Jakarta. There are some limitations for this study, such as the methodology used is survival analysis, especially AFT with Bayesian approach with the ratio between data PE and non-PE is 60:40.

.

II. METHODS

2.1 Survival Analysis

Survival analysis is one of the important fields in statistics that is often used in the world of health, biology, epidemiology, engineering, and demography. In general, survival analysis can be defined as a set of statistical procedures in data analysis that has a response variable in the form of the length of time an event of interest occurs. So, survival analysis can also be referred to as time-to-event analysis.

If the observed event is the event of contracting a disease, death, or other negative events, then the survival time can also be referred to as failure time. There are three important requirements in determining survival time. The initial time of the start of observation must be clearly defined, the time measurement scale has been determined either in units of days, weeks, months, or years, and the last requirement is that the end time of observation is when the observed event occurs or completed. Based on Cox & Oakes' explanation, of course there is data obtained that can be classified based on these three conditions or what is referred to as censoring (Abdullah, 2023).

2.2 Censoring

In general, censoring is divided into three major groups, namely right, left, and interval censoring. Right censoring is a condition where the end of observation has occurred before the event occurs so that the exact time of the event is not recorded. Left censoring is one type of censoring that rarely occurs because the event has occurred before the start of the observation so that usually researchers immediately exclude the data from the observation. While this last type, interval censoring, is a type of censoring where the initial status has been recorded at the occurs in the middle of the observation period but the patient or individual who is the object of research does not check up the condition or examination that should be done periodically so as to result in empty or unknown conditions and when the patient does the last check-up there is a change in the status of the patient (Savell et al, 2015).

2.3 Survival Function

There are some methods to summarize and describe survival time. Two common qunatitative methods are survival and hazard function. Survival time of an observed individual, t, can be described as observed value from variable T and has non-negative value. Variable T also has a different value so that it has a probability distribution. For example, random variable T has a probability distribution with density function (pdf), $F_T(t)$, and cummulative distribution function (cdf) can be shown as $F_T(t) = \Pr(T \leq t) = \int_0^t f_t(w)\, dw$ which describe probability survival time less than or equal to a value of t. Whereas survival function, $S_T(t)$, describes a probability of survival time after a value of t. Simply can be said that $S_T(t) = Pr\ (T > t)$. If T is a continuos random variable, the interaction between survival and distribution function can be written as: $S_T(t)=Pr(T > t)=1-Pr(T \leq t) = 1-F_T(t)$.

Survival function can be described as a density function ($f_T(t)$), $S_T(t)=Pr(T > t) = \int_t^\infty f_T(w)\, dw$. If those

functions are substituted between density and cummulative distribution, survival function, then survival can be written as: $f_T(t) = (dF_T(t))/dt = d(1-S_T(t))/dt = -(dS_T(t))/dt = -S'_T(t)$. If T a random discrete variable, then it needs a different technique because discrete variable has a rounding factor.

For example, specific value of variable T (t_j, j = *1,2,3,...*) with probability mass fnction (pmf), $p(t_j)$ = Pr(T-t_j), for j = *1,2,...* which $t_1 < t_2 < \cdots$. Survival function for discrete variable of T can be shown as $S_T(t) = \Pr(T > t) = \sum_{tj>t} p(t_j)$ (Turkson et al, 2021).

Characteristic of survival function for random variable T are:

- $0 \leq S_T(t) \leq 1, \forall t$
- A non increasing function.
- A right continuous function
- $\lim_{t\to 0} S_T(t) = 1$ dan $\lim_{t\to\infty} S_T(t) = 0$ (Abdullah, 2023)

2.4 Accelerated Failure Time

If there are two populations A and B with different survival function (($S_A(t)$ and ($S_B(t)$) and both functions have connected each other through accelerated failure rate λ:

$S_A(t) = S_B(t/\lambda)$. Rate can be defined as a factor that can accelerate or decelerate survival function or also known as acceleration factor. As an ilustration, a dog has a accelerated factor of 7 of their age compared to human (λ = 1/7). Based on the function above, accelerated failure time (AFT) can be shown as $S_A(t) = S_B(t/\lambda(t))$ for every t. If there are some observed covariates, with x = ($x1,x2,x3,...,xp$)' are the vector of those covariates, then mathematically AFT model will be $S(t|x) = S_0(t/e^{\eta(x)})$ with $e^{\eta x} = e(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)$ as an accelaration factor which give an information about how big the impact of covariate to survival time, β = ($\beta_1, \beta_2, ..., \beta_p$) as a regression parameter, and S_0 as a survival function baseline.

From the interaction of between survival and hazard function, hazard function can be shown as:

$$S(t|\boldsymbol{x}) = S_0\left(\frac{t}{e^{\eta(\boldsymbol{x})}}\right)$$

$$-\ln S(t|\boldsymbol{x}) = -\ln S_0\left(\frac{t}{e^{\eta(\boldsymbol{x})}}\right)$$

$$\frac{d}{dt}(-\ln S(t|\boldsymbol{x})) = \frac{d}{dt}\left[-\ln S_0\left(\frac{t}{e^{\eta(\boldsymbol{x})}}\right)\right]$$

$$h(t|\boldsymbol{x}) = e^{-\eta(\boldsymbol{x})}\, h_0\left[\frac{t}{e^{\eta(\boldsymbol{x})}}\right].$$

The interacction among covariates with survival time can be described as a linear model using logarithm function, $ln(T) = \mu + \beta x_i + \sigma W_i$ with μ is an intercept, σ is a parameter scale, and W is an accepted error which assumed to have a distribution (Alvares et al, 2021). There are some common distributions for modelling survival analysis, Exponential, Weibull, Log-Logistik, and Log-Normal Distribution.

2.5 Log-Normal Distribution

Log-Normal distribution has two main variables, mean and variance. Survival function with log-normal distribution can be shown as $S_T(t) = \int_t^\infty f(w)\, dw = 1 - \int_0^t f(w)\, dw = 1 - \Phi\left(\frac{\ln t-\mu}{\sigma}\right)$; $t \geq 0$. Where $\Phi\left(\frac{\ln t-\mu}{\sigma}\right)$ is a cumulative distibution function from normal standard of $\left(\frac{\ln t-\mu}{\sigma}\right)$. So that the function can be reshown as:

$$S_T(t) = 1 - \Phi\left(\frac{\ln t-\mu}{\sigma}\right)$$

$$1 - S_T(t) = \Phi\left(\frac{\ln t-\mu}{\sigma}\right)$$

$$\Phi^{-1}[1 - S_T(t)] = \frac{\ln t-\mu}{\sigma}$$ (Badriyah et al, 2018)

2.6 Bayesian Inference

In the world of statistics there are two commonly used approaches to probability interpretation, Bayesian and Frequentist. The main difference between the two approaches lies in the nature of the probabilities. The classical (frequencyist) method, which is the most widely used statistical method, treats parameters as fixed unknown constants and describes probabilities as relative frequencies of events. Slightly different is Bayesian which treats parameters as random variables.

The term "Bayesian" comes from an English statistician and Prebiterian priest in the eighteenth century, Thomas Bayes. Bayes wondered in solving a question "what is the probability of an event?". He tried to use conditional probability to create an algorithm that would be used in calculating limits on unknown random parameters (Psioda & Ibrahim, 2019)

2.7 Likelihood

Probability function for bayesian inference is called as likelihood. This function is implemented for estimating parameter in AFT model. Likelihood in survival analysis needs an asumption which censored, and survival time must be independent. Without this asumption, then likelihood function cannot be implemented to censored data in the analysis. If the time variable T expresses the time of occurrence of the observed event, then the likelihood function of the survival model can be divided into four different parts according to the information from the data used.

In uncensored observations, the survival time is observed and information can be obtained about the probability that the event occurs at that time. So, the density function can be used to form a likelihood function.

In right-censored observations, where the event occurs after the observation time, it is found that the value of the variable T is smaller than a certain time t. Since $T > t$, then the survival function is used. Since $T > t$, the survival function $S_t(t)$ will be used to construct the likelihood function.

In left-censored observations, where the individual has experienced the event before the observation time, the value of $T < t$ is obtained. Therefore, the likelihood function can be constructed using the cumulative distribution function, $F_t(t)$=1-$S_t(t)$.

In observations with interval censoring, information is obtained that the event occurs at a certain time interval. Thus the survival time is in [L,R]. The information of the individual can be expressed by $Pr(L < T < R) = S(L) - S(R)$ (Ma et al, 2021).

2.8 Probability Distribution

There are 2 kinds of probability distribution, prior and posterior distribution. Prior distribution is a probability of the uncertain qunaity before some evidences are included to the calculation parameter Previous experiment and expert

judgement can be used as a prior distribution in a vector form of the observed data. Non informatif prior or flat prior is a condition where the distribution of likelihood is relatively flat which means it has minimum impact for the posterior distribution. Posterior distribution usually called as the result of the observed data. In bayesian theorem, posterior is a conditional probability on quantitative data which combine between likelihood and prior distribution. Estimated mean posterior can be concluded as an estimation of coefficient parameter. A proper posterior is hard to be analyzed because it has complex integral calculation and multidimension. To solve it, the posterior distibution will be analyzed using Markov Chain Monte Carlo (MCMC) (Alvares et al. 2021).

2.9 MCMC Method

Markov Chain Monte Carlo (MCMC) method can be used as part of Bayesian inference as it focuses on sampling the posterior distribution, which is usually the difficult part when using analytical checks. Markov Chain itself is a series of events whose distribution depends only on the outcome of the previous event. The advantage of using Markov Chain is that it has coverage of the targeted variable distribution as long as it implements the correct algorithm. Having a different focus to Markov Chain, Monte Carlo focuses on sampling from the posterior distribution. So basically the MCMC method is a simulation of sampling from the posterior distribution and calculating the posterior sum of interest of the targeted distribution while each sample depends on the previous calculation (Ma et al, 2021).

2.10 Chain

Every time MCMC runs, it is important to check the performance of the chains. Although several literatures have different arguments regarding the selection of the number of chains for MCMC, they all agree to use more than one chain for computation. The idea of using more than one chain is that there is at least one chain that can converge to the targeted distribution and is expected to explore all the features associated with the distribution (Hu et al, 2021).

2.11 Burn-in

Markov chains are highly dependent on their initial values even if the iteration has started over a long period of time. Therefore, if the chosen initial value is sufficiently different from the posterior mode, then the dependency will make the chain converge more slowly than usual. It may take several iterations before the chain enters the high probability region where it represents the targeted distribution. To make the chain converge faster, the first n iterations will be removed as a burn-in stage to make the chain have good initial values (Ma et al, 2021).

2.12 Thinning

The slow decay of autocorrelation usually makes the chain mixing poor. To avoid this condition, inference should be recorded at every iteration i-th of the chain where i-th is set to a high value so that it will become independent. This strategy is called thinning. In addition, thinning will make the results obtained less than usual because the model only records the i-th and multiples of its iterations (Hu et al, 2021).

2.13 MCMC Convergence

After all the iterations are completed, the next thing to analyze is how to determine the plausibility point to believe that the selected samples accurately approximate the stationary distribution of the Markov chain. The convergence of a chain can be used to check the efficiency with which the chain can approach the stationary distribution. A well-known technique for checking the convergence of chains using a formal statistical test was established by Gelman-Rubin in 1992. The test is based on a comparison between the variables used in the chain and the variance between the chains (Psioda & Ibrahim 2019).

2.14 Credible Interval

In Bayesian Inference, credible interval is also called as a confidence onterval. The goal of this interval is not only for sumarizing the result, but also to picture the parameter's inconsistency. Credible interval for paramter is defined as $[c_p, c_{1-p}]$ where c_p and c_{1-p} are estimated as a quantile p^{th} and $(1 - p)^{th}$ from posterior distribution.

The most common credible interval is 2.5% dan 97.5% or 95% credible interval. This interval is symmetry because it has removed both sides of the distribution. 95% credible interval can be used to check the significance of estimated model parameter. So, it can be concluded that there is 95% probability of accepted parameter under this interval (Hu et al, 2021).

III. RESULTS AND DISCUSSION

3.1 Dataset

Dataset for this research was using medical records from obstretric and gynecology department in one hospital in Jakarta, Indonesia. Data consist of 924 patients with 860 of them delivered without PE condition and 64 of them delivered with PE condition. There are 13 variables which might affect to PE condition, the details are shown in Table I.

Table I. Data Variable Definition

No	Variable	Data Type	Description
1	FirstPregnant	Categoric	First pregnancy
2	Conception_IVF	Categoric	Pregnancy with In Vitro Fertilization
3	PreviousPE	Categoric	PE condtion at the last pregnancy
4	ChronicHT	Categoric	Hypertension
5	AnyFamilyHistoryofPE	Categoric	PE condition that affects family
6	Age	Numeric	Patient's age
7	BMI	Numeric	Body mass index
8	CRL_mm	Numeric	Crown Rump Length
9	MAP	Numeric	Mean Arterial Pressure
10	PLGFConcentration	Numeric	Placental growth factor concentration from blood
11	MeanUtAPI	Numeric	Mean value of uterine artery pulsatility
12	Opthalmica	Numeric	Blood vessel in eyes
13	PE	Categoric	Giving birth with PE condition

In general, the condtion of the data is imbalance, where PE cases are much less than non-PE. This imbalance condition needs to be fixed before building the model because it might worsen the capability of the model in predicting parameters.

3.2 Undersampling

In computing systems, imbalanced data is the unequal distribution of data between groups. A larger amount of positive data (majority) compared to the amount of negative data (minority) is an imbalanced data condition. Data imbalance leads to misclassification, where the classifier favors the majority data. Minority data will be considered as noise and outliers and can reduce the performance of the classifier.

Data imbalance cases can generally be handled with oversampling and undersampling methods. Both methods are solutions to data imbalance based on processing or handling part of the data. The oversampling method is used to handle data imbalance by creating synthetic data in the minority data. The undersampling method reduces the amount of majority data until the data has the same distribution of the amount of data. The undersampling method is often used by considering the speed and time efficiency factor (Hairani et al, 2020).

This method was chosen because the Bayesian Accelerated Failure Time Model will consider the PE group as the main event and the non-PE group as the censored event, so it would be better if the main event is more than the censored event. For the model to have a good ability to identify factors that affect PE conditions consistently, the proportion of PE and non-PE data is made into 60% and 40%. The modeling process will be repeated 100 times to reduce bias or error when undersampling occurs.

3.3 Building Model

The model building process uses the Bayesian Accelerated Failure Time method, which is one of the methods of survival analysis that serves to calculate the coefficients of the predictor variables in the data. There are several distributions that are quite often used in this method, some of which are Weibull, Log-Normal, and Log-Logistic distributions. Finding the right distribution for the AFT model is the first step that must be done so that the model formed can be a good model. Therefore, in this study, the three distributions will be compared using the R Studio application.

Based on the Akaike Information Criterion (AIC) value obtained, the Log-Normal distribution is the most appropriate distribution to use in this study with the lowest AIC value of 833.8 followed by the Log-Logistic Distribution with a value of 837.9 and the Weibull Distribution with a value of 846.8.

After finding the right distribution, declaration, and fitting for all variables into matrix form is required. This is done so that the model can clearly distinguish categorical and numerical variables so that it is easier when the calculation process is carried out. The data used is divided into two types of data, namely right-censored data and uncensored data. Uncensored data is data on patients who gave birth with PE conditions. Meanwhile, right-censored data is data on patients who gave birth without PE. The division of data based on tensors and uncensored will influence the formation of groups that will be the main focus when forming the model.

AFT model with Log-Normal distribution will be built using Bayesian simulation approach. The total simulation conducted amounted to 12,000 iterations with the first 2,000 iterations considered as burn-in, thinning value = 10, and chain value = 3. The iteration process is a process where the coefficients obtained in the previous calculation process will be reused in the next iteration and the results obtained in the first 2,000 iterations will be discarded because they are considered inconsistent or referred to as burn-in. The thinning value = 10 indicates that the recording process is only done once every 10 iterations. Since there are 10,000 iterations performed (excluding the burn-in iteration), there are 1,000 values of each coefficient, therefore the results obtained in this modeling process are in the form of an average of all stored results. The value of chain = 3 indicates that there will be 3 repetitions in model building. If the resulting value for each repetition is within a certain value limit, the color of each chain will be mixed and there will not be 1 or 2 colors that dominate (Figure 1).

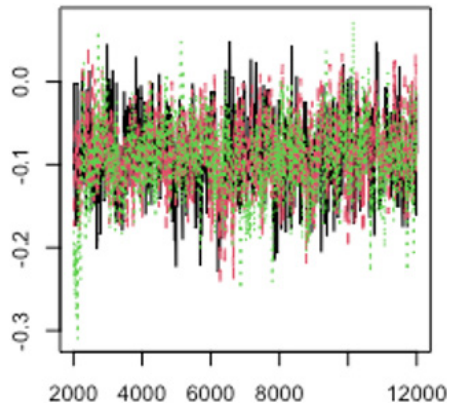

Figure 1. Trace Plot with 3 Chains

After completing one cycle of Bayesian AFT modeling, the whole process will be repeated 99 more times and then the average value is sought. The following Table II is the result of the coefficient obtained when modeling.

Table II. Coefficient Model Parameter

Variabel	Coef	Coef Lower	Coef Upper
FirstPregnant	0.00098	-0.05421	-0.00094
Age	-0.01054	-0.03598	0.00464
BMI	0.00755	0.00285	0.04947
Conception_IVF	-0.00292	-0.01690	0.02494
PreviousPE	-0.07233	-0.04581	0.04653
ChronicHT	-0.01438	-0.03291	0.01166
AnyFamilyHistoryofPE	-0.04164	-0.02918	0.01397
CRL_mm	-0.01419	-0.08699	0.08314
MAP	-0.02673	-0.13391	-0.01196
MeanUtAPI	-0.01549	-0.08841	0.06109
PLGFConcentration	0.02584	-0.10723	0.02427
Opthalmica	0.00391	-0.03447	0.00601

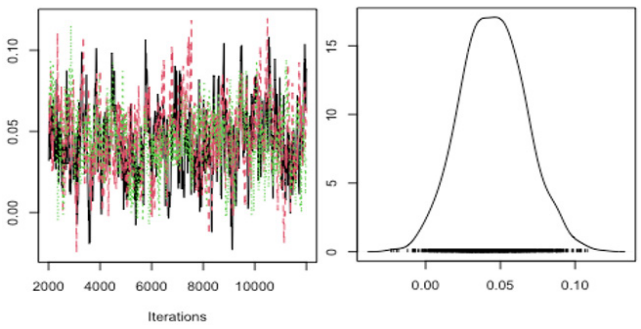

Figure 2. Trace and Density Plot PLGFConcentration

A convergent result is shown by Trace and Density Plot above. The trace plot above consists of three colors, if the three colors are well mixed and there are no one or two colors that dominate, it can be said that each iteration carried out in the formation of the model takes place well and produces coefficient values that are in a certain range of values consistently. While the density plot shows that each variable has one mode which indicates that the results obtained are normally distributed.

In Table 2. there are columns "Coef Lower" and "Coef Upper" which are the values of the credible intervals of 2.5% and 97.5%. Through the Coef Lower and Coef Upper values, it can be seen that there are three factors, namely BMI and MAP, which have CoefUpper and CoefLower values that are in the positive-positive or negative-negative range (while other variables are between negative and positive numbers). The consistency in the credible interval value is an indication that the three factors can be categorized as factors that affect PE conditions consistently.

IV. CONCLUSION

Preeclampsia (PE) is a high blood pressure condition that can occur in pregnant women with gestational age above 20 weeks. Until now, the clinical factors that cause PE and the drugs that can cure it are not yet known. Although the number of PE conditions only ranges from 5-10%, PE conditions are considered dangerous because they can have long-term effects for mothers and their babies, such as hypothermia, hypoglycemia, and polycythemia. The condition of PE in pregnant women can be predicted using survival analysis model and one of the commonly used methods is Accelerated Failure Time (AFT). In addition to using survival analysis, this study also uses the Bayesian method for calculating model parameters.

The data used in this study is medical record data from the obstetrics and gynecology department of one of the hospitals in Jakarta. The data contains 924 patients with 860 patients giving birth normally or without PE condition and 64 patients giving birth with PE condition. The low data of patients with PE condition causes an unbalanced condition and needs data preprocessing. The data preprocessing process chosen is the undersampling method where data without PE condition (censored data) will be randomly selected with a smaller quantity so that the ratio between PE and non-PE data is 60:40.

The model is built using the Bayesian AFT method with a Log-Normal distribution. The total number of simulations performed is 12,000 iterations with the first 2,000 iterations considered as the burn-in stage, the thining value is 10, and the chain is 3. Model building will be carried out 100 times and the results obtained will be averaged to reduce errors in the undersampling process performed.

The coefficient values generated from this Bayesian AFT model have a Trace Plot with 3 colors that are evenly mixed and there is only 1 mode in the Density Plot which indicates that all coefficients are in the same range of values or converging.

Apart from that, it was also identified that there are 3 factors that have consistent credible interval values, namely BMI and MAP. BMI stands for body mass index which measures the ratio between height and weight. MAP stands for mean arterial pressure which measures the average arterial blood pressure in mmHg. Based on the results of this study, it can be concluded that these two factors consistently influence the condition of PE.

Regardless of the results obtained, this study is limited to the ratio of PE and non-PE data, which is 60:40, so if the ratio used changes, the results obtained may be different. This research can also be continued by looking at the effect of censoring on the data used in modeling whether it will affect the model's ability to predict the timing of PE conditions. In addition, other survival analysis methods, such as competing risk analysis, can be used in building similar models to compare the results of the parameters of the factors that affect PE conditions and predict the time of occurrence of PE conditions.

REFERENCES

Abdullah, S. (2023). Analisis Survival: Konsep dan Aplikasi dengan R. Bumi Aksara.

Alvares, D., Lázaro, E., Gómez-Rubio, V., & Armero, C. (2021). Bayesian survival analysis with BUGS. Statistics in Medicine, 40(12), 2975–3020.

Badriyah, T., Tahrir, M., & Syarif, I. (2018). Predicting the risk of preeclampsia with history of hypertension using logistic regression and naive bayes. 2018 International Conference on Applied Science and Technology (ICAST), 399–403.

Brown, M. A., Magee, L. A., Kenny, L. C., Karumanchi, S. A., McCarthy, F. P., Saito, S., Hall, D. R., Warren, C. E., Adoyi, G., & Ishaku, S.

Hairani, H., Saputro, K. E., & Fadli, S. (2020). K-means-SMOTE untuk menangani ketidakseimbangan kelas dalam klasifikasi penyakit diabetes dengan C4.5, SVM, dan naive Bayes. Jurnal Teknologi Dan Sistem Komputer, 8(2), 89–93.

Hu, G., Xue, Y., & Huffer, F. (2021). A comparison of Bayesian accelerated failure time models with spatially varying coefficients. Sankhya B, 83(Suppl 2), 541–557

Ma, Z., Xue, Y., & Hu, G. (2021). Geographically weighted regression analysis for spatial economics data: A Bayesian recourse. International Regional Science Review, 44(5), 582–604.

Psioda, M. A., & Ibrahim, J. G. (2019). Bayesian clinical trial design using historical data that inform the treatment effect. Biostatistics, 20(3), 400–415.

Rolnik, D. L., Nicolaides, K. H., & Poon, L. C.(2022). Prevention of preeclampsia with aspirin. American Journal of Obstetrics and Gynecology, 226(2), S1108–S1119.

Savell, C. T., Borsotto, M., Woodson, S., Dahl, E., Needham, W., Ellor, J., & Korzun, J. (2015). Expert Structures and Coating Analysis Tool (ESCAT).

Syaharutsa, D. M., & Purwosunu, Y. (2018). A Scoring System for Preeclampsia Screening Based on Maternal and Biophysical Factors: Result from a 3 Month Cohort Study in Jakarta, Indonesia. Advanced Science Letters, 24(9), 6361–6365.

Turkson, A. J., Ayiah-Mensah, F., & Nimoh, V. (2021). Handling censoring and censored data in survival analysis: a standalone systematic literature review. International Journal of Mathematics and Mathematical Sciences, 2021, 1–16.

(UK), N. G. A. (2019). Hypertension in Pregnancy: Diagnosis and Management (NG133).

Wirakusuma, G., Surya, I. G. P., & Sanjaya, I. N. H. (2019). Rendahnya kadar placental growth factor (PlGF) serum merupakan faktor risiko terjadinya preeklamsia. Medicina, 50(1).

Wright, D., Wright, A., & Nicolaides, K. H. (2020). The competing risk approach for prediction of preeclampsia. American Journal of Obstetrics and Gynecology, 223(1), 12–23.

Xue, Y., Schifano, E. D., & Hu, G. (2020). Geographically weighted Cox regression for prostate cancer survival data in Louisiana. Geographical Analysis, 52(4), 570–587.