ASM Sc. J., 13, 2020
https://doi.org/10.32802/asmscj.2020.sm26(1.1)

A Comparison Between Bayesian and Frequentist Approach in the Analysis of Risk Factors for Female Cardiovascular Disease Patients in Malaysia

Nurliyana Juhan[1,2], Yong Zulina Zubairi[3*], Zarina Mohd Khalid[1], Ahmad Syadi Mahmood Zuhdi[4]

[1]*Department of Mathematical Sciences, Faculty of Science Universiti Teknologi Malaysia, 81310 Johor Bahru, Malaysia*
[2] *Preparatory Centre for Science and Technology, Universiti Malaysia Sabah, 88400 Sabah, Malaysia*
[3]*Centre for Foundation Studies in Science, University of Malaya, 50603 Kuala Lumpur, Malaysia*
[4]*Cardiology Unit, University Malaya Medical Centre, 50603 Kuala Lumpur, Malaysia*

Cardiovascular disease (CVD) is the number one killer among women in Malaysia and globally, with over two million deaths each year. In this study, two modelling approaches namely Bayesian approach and frequentist approach were considered to identify associated risk factors in CVD among female patients presenting with ST Elevation Myocardial Infarction (STEMI) and to obtain feasible model to fit the data. Comparisons were made to find the best model. A total of 874 STEMI female patients from 18 participating hospitals across Malaysia in the National Cardiovascular Disease Database-Acute Coronary Syndrome (NCVD-ACS) registry year 2006-2013 were analysed. Univariate and multivariate analysis were performed for both Bayesian and frequentist approaches. Six variables namely smoking, dyslipidaemia, myocardial infarction (MI), renal disease, Killip class and age group were found to be significant at the multivariate analysis. The standard errors obtained from the Bayesian approach were much smaller than the frequentist approach. Also, the model fit using Bayesian approach was much better than the frequentist as the deviance value produced by the Bayesian approach was smaller. The Bayesian analysis provides a better alternative to the frequentist approach in the analysis of the risk factors associated with mortality among female CVD patients.

Keywords: cardiovascular; female; risk factor; bayesian; frequentist

I. INTRODUCTION

Cardiovascular disease (CVD) is the number one killer among women in Malaysia and globally, with over two million deaths each year (Department of Statistics Malaysia Official Portal, 2016; World Heart Federation, 2017). Although known as a men's disease, in some cases its effects can be worse in women (Garcia *et. al.,* 2016). Some of the symptoms in women might be different to that in men. Women are more probable to have sudden numbness or weakness of the face, arm or leg, especially on one side of the body, shortness of breath, nausea, vomiting and jaw pain (Gonsalves *et. al.,* 2017). Also, women often have misconceptions about these CVD symptoms with breast cancer signs (Miller, 2016). Thus, they are often under-diagnosed and under-treated when compared to men (Garcia *et. al.*, 2016). Early detection of risk factors can help in reducing and control CVD event among women.

*Corresponding author's e-mail: yzulina@um.edu.my

In this study, two modelling approaches namely Bayesian and frequentist were considered to identify associated risk factors in CVD among female patients presenting with ST Elevation Myocardial Infarction (STEMI) and to obtain feasible model to describe the data. Comparison of the results, interpretations and limitations of both approaches were made to enrich the discussion as well as to obtain robust conclusions. Both Bayesian and frequentist approaches offer good statistical models, and each is important in the development of the other approach. Bayes risk of an estimator is the average risk over the prior distribution which leads to a problem statement known as credible interval (Koslovsky *et. al.*, 2018). In a frequentist framework, data is modelled probabilistically and inferential statement about expressed in the form of confidence interval (Torman and Camey, 2015).

Frequentist approach depend on the work of classical statisticians such as Fisher, Pearson and Neyman which referred to the traditional framework used to fit statistical models and conduct hypothesis testing. The term frequentist represented the probability which is defined as the frequency of observing an event if an experiment was repeated over infinite number of times (Christ and Desjardins, 2018). As for the Bayesian approach, statistical conclusions about the unknown parameters, are made in terms of a subjective belief, view of probability conditional on the observed data and existing knowledge (Gelman *et. al.*, 2008). These subjective beliefs are properly incorporated into statistical distributions about the unknown parameters known as prior distributions. Despite the difference in the philosophies, applying both approaches in solving medical related problem is useful in practice.

The organization of this study is as such; it begins with a brief information on CVD, frequentist and Bayesian approach in Section I, followed by materials and methods in Section II. Next, is the results of proposed models and followed by a discussion of the findings of the analysis in Section III. Finally, summary is given in Section IV.

II. MATERIALS AND METHODS

A. Source of Data

A total of 874 ST-Elevation Myocardial Infarction (STEMI) female patients from 18 participating hospitals across Malaysia in the National Cardiovascular Disease Database-Acute Coronary Syndrome (NCVD-ACS) registry for the years 2006-2013 were analysed. Data was collected from the time the female patient with STEMI was admitted to the hospital till 30 days post discharge. Variables were categorized into demographic, risk factors, comorbidities, clinical presentation and treatment. As for the clinical presentation, the Killip classification predicts the chances of survival within 30 days in patients with an acute heart attack, in which Killip class IV having a higher chance of dying (Killip and Kimball, 1967). This NCVD registry study was approved by the Medical Review & Ethics Committee (MREC), Ministry of Health (MOH) Malaysia in 2007 (Approval Code: NMRR-07-20-250). MREC waived informed consent for NCVD.

B. Statistical Methods

For the frequentist approach, traditional framework is used to fit statistical model. As the dependent variable is of the binary form with "1" represented death and "0" alive or otherwise, logistic regression was used in the development of the prognostic models. Generally, the logistic function can be written as:

$$F(x)=\frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \tag{1}$$

where F(x) is the probability of the dependent variable equalling a case, β_0 is the intercept from the linear regression equation and β_1 is the regression coefficient. While the inverse logistic function, logit can be defined as:

$$\text{logit}(x)=\ln\left(\frac{F(x)}{1-F(x)}\right)=\beta_0+\beta_1 x \tag{2}$$

Univariate and multivariate analysis were performed for both frequentist and Bayesian approaches. Purposeful variable selection method based on Hosmer & Lemeshow (2000) was applied. Maximum likelihood method is applied as a standard method in parameter estimation in a logistic regression model. The frequentist maximum likelihood method assumes that existing data are random subsets from a population of interest, where the unknown parameters are assumed to be fixed with an associated random (Discacciati, Orsini and Greenland, 2015). The maximum likelihood method

produces estimation of the coefficients and their corresponding standard errors while the frequency of an event interpretation of probability will produce the definitions of p-values and confidence intervals.

For the Bayesian model considered in this study, Markov chain Monte Carlo (MCMC) refers to the algorithms that combine the logic of simulation known as Monte Carlo methods with a mathematical random process called Markov chains was applied in the estimation and inference of model parameters (Jackman, 2009). Different from traditional frequentist approach, Bayesian approach treats the unknown parameters in a model as random variables and the posterior distribution of a parameter is used to measure uncertainty (Ntzoufras, 2009).

The likelihood function corresponds to the probability of the data given all the plausible values for the parameters is assigned as Bernoulli distribution with the parameter μ, defined as a logistic. Non-informative prior was assigned for the regression coefficients, β due to lack of information on the regression parameters. The prior distribution, $p(\theta)$ is multiplied by a likelihood function, $p(y|\theta)$ and then divided by the distribution of the data, $p(y)$which resulting in the posterior distribution, $p(\theta|y)$. All Bayesian inferences are made from this posterior distribution. The derivation of posterior distribution is through Bayes' theorem as in Equation 3.

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)} \quad (3)$$

The denominator is usually omitted as it does not have any parameters and it is a constant. Thus, Bayes' theoremis re-expressed as:

$$p(\theta|y) \propto p(\theta)p(y|\theta) \quad (4)$$

The probability of the parameters given the data are able to be obtained through Equation 3 and 4. The posterior distribution is then estimated using MCMC method.

Three multiple parallel chains with different initial points were applied in the simulation work. At the univariate level, models were developed by running the multiple chains for 10,000 iterations each with the initial 1000 burn-in. While at the multivariate level, simulation runs were fixed at 100,000 iterations, with the initial 10,000 burn-in samples were discarded from the analysis. Comparisons between frequentist and Bayesian approach were made to find the best model. All the analyses were performed using SPSS statistical software (version 24, IBM SPSS Statistics, USA) and Just Another Gibbs Sampling (JAGS) software in R interface.

III. RESULTS AND DISCUSSIONS

Descriptive statistics was performed, and the results are shown in Table 1. Majority of the STEMI female patients were from ethnic Malay with more than 50% followed by Chinese (20%) and Indian (19.9%). More than 50% of female patients fell into the less than 65-year-old age group. The most prevalence risk factor for STEMI female patients was hypertension (74.5%), followed by diabetes mellitus (55.6%) and dyslipidaemia (35.8%) respectively. Presence of comorbidity variables are generally low among female patients, yet the most relevant comorbidity was myocardial infarction (MI) followed by renal disease and cerebrovascular disease. Most of the STEMI patients were in Killip class I or II on presentation. For the treatment, cardiac catheterisation was the most undergone procedure followed by the percutaneous coronary intervention (PCI).

Table 1. Patients' characteristics

Characteristic			**N (%)**
Demographic	Ethnicity	Malay	489 (55.9)
		Chinese	175 (20.0)
		Indian	174 (19.9)
		Others	36 (4.1)
	Age group	<65	495 (56.6)
		≥65	379 (43.4)
Risk factor	Diabetes Mellitus	No	388 (44.4)
		Yes	486 (55.6)
	Hypertension	No	223 (25.5)
		Yes	651 (74.5)
	Smoking status	Never	786 (89.9)
		Active/former	88 (10.1)
	Dyslipidaemia	No	561 (64.2)
		Yes	313 (35.8)
	Family history of CVD	No	792 (90.6)
		Yes	82 (9.4)
Comorbiditie s	MI History	No	784 (89.7)
		Yes	90 (10.3)
	Chronic lung disease	No	859 (98.3)
		Yes	15 (1.7)

	Cerebrovascular disease	No	842 (96.3)
		Yes	32 (3.7)
	Peripheral vascular disease	No	870 (99.5)
		Yes	4 (0.5)
	Renal disease	No	832 (95.2)
		Yes	42 (4.8)
Clinical presentation	Killip Class	Class I	521 (59.6)
		Class II	219 (25.1)
		Class III	44 (5.0)
		Class IV	90 (10.3)
Treatment	PCI	No	629 (72.0)
		Yes	245 (28.0)
	Cardiac catheterisation	No	573 (65.6)
		Yes	301 (34.4)

Although not shown, of the fifteen variables, seven are found to be significant at the univariate analysis. The seven significant variables were then included into the multivariate analysis to identify the prognostic factors. For the frequentist model, using purposeful selection method, six variables of the seven were observed to be significantly associated with mortality of female CVD patients namely dyslipidaemia, myocardial infarction, smoking, renal disease, Killip class and age group of the patients. Similar results were obtained for the final multivariate Bayesian model.

Comparison between frequentist and Bayesian multivariate models based on the odds ratio, standard errors and credible intervals/confidence intervals are shown in Table 2. Generally, there were no significant differences between the Bayesian and frequentist estimates for most of the variables in the model. However, the odds ratio in the Bayesian is larger than that of the frequentist. It is worthwhile to note that, the odds ratio that fall within the confidence intervals also fall in the credible intervals. The credible intervals appear to be wider than the confidence limits. The addition of informative priors might however be helpful in narrowing the gap of credible interval and provide precise choice between the null and alternative hypothesis (Fosu, Jackson and Twum, 2016). Others suggested that by increasing the sample size, the numerical difference between both types of interval will decrease where in most common cases, the frequentist confidence intervals and Bayesian credible intervals lead to very comparable conclusions (Albers *et al.*, 2018).

Also, the results shown that the standard errors associated with the Bayesian model are much smaller than the frequentist model. This means that the Bayesian model can be more relied upon than the frequentist model and even with non-informative priors, the Bayesian model may well predict better. This is in line with a study by Guure (2015) where improvement in results was observed through the use of a non-informative prior. Besides, the model fit using Bayesian approach was much better than the frequentist as the deviance value produced by the Bayesian approach from Table 2 was smaller. This is supported by few studies which also in comparing the two methods found that the Bayesian model performed better (Pascale, Nicoli and Spagnolini, 2014; Wong and Ismail, 2016).

Frequentist approach is more straightforward and computationally relatively simple as there is no numerical integration is needed (Fosu *et. al.,* 2016; Yap *et. al.,* 2015). In the frequentist model, the maximum likelihood estimation (MLE) method used has attractive properties such as consistency, asymptotic normality and asymptotic unbiasedness when the sample size is sufficiently large. As the sample size increases, the estimated coefficients asymptotically approach the population values (Cole *et al.,* 2014). However, due to this asymptotic behaviour, the maximum likelihood method produced poor and unreliable results in terms of p-values and estimation of parameters for small sample size (King and Ryan, 2002; Morey et. al., 2016). Other study suggested using maximum likelihood estimates for sample sizes above 500 and not suitable for studies with less than 100 samples (Long 1997). As in this study, the sample size of 874 is sufficiently enough as the frequentist approach produced almost similar estimation of the parameters as the Bayesian approach.

Unlike frequentist approach, Bayesian approach involves integral solutions to marginal distributions in the computation of the posterior distribution. However, integrations are difficult through analytical methods for high dimensional problems making exact inference on the posterior distribution and analytical solutions are impossible. Thus, in this study, simulation-based methods such as the MCMC algorithms (Metropolis et al., 1953; Hastings, 1970) is used for estimation of posterior distributions. In this study, 100,000 iterations with 10,000 burn-in was sufficed to achieve convergence as other studies suggested that at least 1000 and up to a million iterations were used for estimation respectively (Besag et. al., 1991; Torman and Camey, 2015; Wong and Ismail, 2016).

In model estimation and inference there are several advantages in using a Bayesian MCMC approach.

Table 2. Bayesian and frequentist estimations in final multivariate model for female patients

Variable	Bayesian			Frequentist (MLE)		
	β	**SE**	**OR (95% Credible Interval)**	***β***	**SE**	**OR (95% Confidence Interval)**
Dyslipidaemia	-0.636	0.048	0.529 (0.310, 0.866)	-0.644	0.250	0.525 (0.322, 0.858)
Myocardial infarction	0.639	0.062	1.895 (0.980, 3.490)	0.628	0.319	1.874 (1.295, 2.014)
Smoking	-0.984	0.089	0.374 (0.141, 0.979)	-0.915	0.452	0.401 (0.493, 0.963)
Renal disease	0.827	0.077	2.286 (1.026, 3.938)	0.828	0.392	2.289 (1.598, 3.478)
Killip Class II	0.861	0.054	2.366 (1.362, 3.088)	0.858	0.278	2.359 (1.553, 2.717)
Killip Class III	1.120	0.083	3.065 (1.274, 7.308)	1.139	0.423	3.124 (1.362, 7.169)
Killip Class IV	2.813	0.057	16.660 (11.971, 20.995)	2.773	0.297	16.000 (12.685, 20.981)
Age ≥ 65	1.108	0.045	3.028 (1.819, 4.731)	1.088	0.232	2.967 (1.881, 4.680)
Deviance	544.000			562.000		

First, the Bayesian approach is considered flexible as it able to involve prior information about the underlying parameters with information from past or existing experience. Also, the Bayesian approach provided a degree of uncertainty in the model, hence produced more realistic predictions and protected against overfitting of models more than frequentist approaches (Guure et. al., 2015). Additionally, the Bayesian approach allows updating knowledge, add observations and calculate probabilities for complex functions instead of testing a null hypothesis over and over again (van de Schoot et. al., 2014; Torman and Camey, 2015).

IV. SUMMARY

The risk factors identified were smoking, dyslipidaemia, MI, renal disease, Killip class and age group as they were found to be significant at the multivariate analysis for both frequentist and Bayesian approach. The Bayesian approach provided better estimates and appeared to be promising as an alternative to the traditional frequentist in the analysis of the risk factors associated with mortality for female CVD patients. Also, it can be a useful tool in guiding clinicians in decision making, interpretation of diagnosis and proper treatment to patients.

V. ACKNOWLEDGEMENTS

The authors wish to thank all medical staffs and non-medical staffs who participated in the Malaysian NCVD-ACS registry.

VI. REFERENCES

Albers, CJ et al. 2018, 'Credible Confidence: A Pragmatic View on the Frequentist vs Bayesian Debate', Collabra: Psychology, The Regents of the University of California, 4(1).

Christ, TJ & Desjardins, CD 2018, 'Curriculum-Based Measurement of Reading: An Evaluation of Frequentist and Bayesian Methods to Model Progress Monitoring Data', Journal of Psychoeducational Assessment, 36(1), pp. 55–73.

Cole, SR, Chu, H & Greenland, S 2014, 'Maximum Likelihood, Profile Likelihood, and Penalized Likelihood: A Primer', American Journal of Epidemiology. Oxford University Press, 179(2), pp. 252–260.

Department of Statistics Malaysia Official Portal 2016, Statistics on Causes of Death, Malaysia, 2014.

Discacciati, A, Orsini, N & Greenland, S 2015, 'Approximate Bayesian logistic regression via penalized likelihood estimation with data augmentation', Stata Journal, 15(3), pp. 712–736.

Fosu, M, Jackson, O & Twum, S 2016, 'Bayesian and Frequentist Comparison: An Application to Low Birth Weight Babies in Ghana', British Journal of Applied Science & Technology, 16(2), pp. 1–15.

Garcia, M et al. 2016, 'Cardiovascular disease in women: Clinical perspectives', Circulation Research, 118(8), pp. 1273–1293.

Gelman, A, et al. 2008, 'A weakly informative default prior distribution for logistic and other regression models', Annals of Applied Statistics, 2(4), pp. 1360–1383.

Gonsalves, CA et al. 2017, 'Mass media narratives of women's cardiovascular disease: a qualitative meta-synthesis', Health Psychology Review, 11(2), pp. 164–178.

Guure, CB et al. 2015, 'Bayesian statistical inference of the loglogistic model with interval-censored lifetime data', Journal of Statistical Computation and Simulation. Taylor & Francis, 85(8), pp. 1567–1583.

Hastings, WK 1970, 'Monte Carlo sampling methods using Markov chains and their applications', Biometrika. Oxford University Press, 57(1), pp. 97–109.

Jackman, S. 2009, Bayesian analysis for the social sciences. Wiley.

Killip, T & Kimball, JT 1967, 'Treatment of myocardial infarction in a coronary care unit. A two year experience with 250 patients.', The American journal of cardiology, 20(4), pp. 457–64.

King, EN & Ryan, TP 2002, 'A Preliminary Investigation of Maximum Likelihood Logistic Regression versus Exact Logistic Regression', The American Statistician. Taylor & Francis, 56(3), pp. 163–170.

Koslovsky, MD et al. 2018, 'Using the EM algorithm for Bayesian variable selection in logistic regression models with related covariates', Journal of Statistical Computation and Simulation, 88(3), pp. 575–596.

Metropolis, N, et al. 1953, 'Equation of State Calculations by Fast Computing Machines', The Journal of Chemical Physics. American Institute of Physics, 21(6), pp. 1087–1092.

Miller, P 2016, 'Women and Cardiovascular Disease: What Can Health Care Providers Do to Reduce the Risks?', North Carolina Medical Journal, 77(6), pp. 406–409.

Morey, RD et al. 2016, 'The fallacy of placing confidence in confidence intervals', Psychonomic Bulletin & Review, 23(1), pp. 103–123.

Ntzoufras, I 2009, 'Bayesian Modeling Using WinBUGS', Book, pp. 506.

Van de Schoot, R, et al. 2014, 'A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research', Child Development. Wiley/Blackwell (10.1111), 85(3), pp. 842–860.

Scott Long, J 1997, 'Regression models for categorical and limited dependent variables.', Advanced Quantitative Techniques in the Social Sciences , 7.

Torman, VBL & Camey, SA 2015, 'Bayesian models as a unified approach to estimate relative risk (or prevalence ratio) in binary and polytomous outcomes', Emerging Themes in Epidemiology, 12(1).

Wong, RSY & Ismail, NA 2016, 'An application of Bayesian approach in modeling risk of death in an intensive care unit', PLoS ONE, 11(3), pp. 1–17.

World Heart Federation 2017, Women and CVD - Facts and tips - World Heart Federation - World Heart Federation. Available at: https://www.world-heart-federation.org/resources/women-cvd-facts-tips/.

Yap, C, Lin, X & Cheung, K 2015, 'Comparison of a frequentist and Bayesian response-adaptive randomisation approach in multi-stage phase II selection trials with multiple experimental arms', 16(Suppl 2), pp. 2015.

Pascale, A, Nicoli, M & Spagnolini, U 2014, 'Cooperative Bayesian Estimation of Vehicular Traffic in Large-Scale Networks', IEEE Transactions on Intelligent Transportation Systems, 15(5), pp. 2074–2088.

Hosmer, DW & Lemeshow, S 2000, Applied Logistic Regression, Wiley Series in Probability ans Statistics, pp.373.