Bayesian Model Averaging for Selection of a Risk Prediction Model for Death within Thirty Days of Discharge: The SILVER-AMI Study

Terrence E. Murphy[1,*], Sui W. Tsang[1], Linda S. Leo-Summers[1], Mary Geda[1], Dae H. Kim[2], Esther Oh[3], Heather G. Allore[1], John Dodson[4], Alexandra M. Hajduk[1], Thomas M. Gill[1] and Sarwat I. Chaudhry[1]

[1]*Yale University School of Medicine, New Haven, CT, USA*

[2]*Harvard University School of Medicine, Boston, MA, USA*

[3]*Johns Hopkins University School of Medicine, Baltimore, MD, USA*

[4]*New York University School of Medicine, New York, NY, USA*

Abstract: We describe a selection process for a multivariable risk prediction model of death within 30 days of hospital discharge in the SILVER-AMI study. This large, multi-site observational study included observational data from 2000 persons 75 years and older hospitalized for acute myocardial infarction (AMI) from 94 community and academic hospitals across the United States and featured a large number of candidate variables from demographic, cardiac, and geriatric domains, whose missing values were multiply imputed prior to model selection. Our objective was to demonstrate that Bayesian Model Averaging (BMA) represents a viable model selection approach in this context. BMA was compared to three other backward-selection approaches: Akaike information criterion, Bayesian information criterion, and traditional p-value. Traditional backward-selection was used to choose 20 candidate variables from the initial, larger pool of five imputations. Models were subsequently chosen from those candidates using the four approaches on each of 10 imputations. With average posterior effect probability ≥ 50% as the selection criterion, BMA chose the most parsimonious model with four variables, with average C statistic of 78%, good calibration, optimism of 1.3%, and heuristic shrinkage of 0.93. These findings illustrate the utility and flexibility of using BMA for selecting a multivariable risk prediction model from many candidates over multiply imputed datasets.

Keywords: Risk prediction, AMI, Bayesian model averaging, AIC, BIC, backward-selection.

1. INTRODUCTION

Although several prior studies have described the considerations relevant in developing prognostic models [1-3], with that of Harrell among the most frequently cited [4], two issues have yet to be adequately addressed. The first is how to select a model for a new outcome or population when a large number of candidate variables are worth considering. The second is how to integrate multiple imputation into the selection process itself. In this paper we illustrate how Bayesian Model Averaging (BMA), introduced by Raftery [5] and described by Hoeting *et al*., [6] offers utility and flexibility in addressing these two issues.

Most clinical researchers are familiar with traditional backward-selection based on p-values (BW) and many have heard of the Akaike information criterion (AIC) [7] and the Bayesian information criterion (BIC) [8]. Although introduced 20 years ago, few clinical researchers have heard of BMA. The foremost conceptual difference is the intrinsic assumption in BMA that no single model is entirely accurate because of the uncertainty fundamental to statistical modeling.

Because the AIC, BIC and BW approaches each choose a single optimal model and subsequently use it to predict future values, we collectively refer to them as the discrete approaches. In contrast, BMA evaluates all possible combinations of the candidate variables to identify the best fitting model and a subset of others whose performance is close to the best model. BMA explicitly acknowledges the uncertainty in model selection by calculating a separate association from each of the best fitting models.

The second defining characteristic of BMA is its calculation of the posterior effect probability (PEP) for each of the variables among the best fitting models. In contrast with the rather abstract meaning of a traditional p-value [9, 10], the PEP is the probability that the given predictor's association with the outcome is not equal to zero. Although larger absolute values reflect stronger evidence that a predictor is important, the relative values of PEP among any group of candidate variable can also be used to choose the most potent predictors in a given modeling situation, even when they do not approach values near the intuitive threshold of 50% or greater.

[*] Address correspondence to this author at the Yale University School of Medicine, 300 George St, Suite 775, New Haven, CT 06511, USA; Tel: (203) 737-2295; Fax: (203) 785-4823; E-mail: terrence.murphy@yale.edu, sarwat.chaudhry@yale.edu

E-ISSN: 1929-6029/19

In this report we demonstrate the utility of BMA in selecting a risk prediction model for death within 30 days of discharge among older persons hospitalized for acute myocardial infarction in the Comprehen SIVe Evaluation of Risk Factors in Older Patients with Acute Myocardial Infarction study (SILVER-AMI, NIH/NHLBI R01HL115295) [11]. We illustrate that BMA chooses a model similar to the discrete approaches, while providing a more objective criterion (posterior effect probability) for deciding when to retain variables differentially selected across imputed datasets.

2. METHODS

2.1. Design of SILVER-AMI

All explanatory variables and outcomes used in this report come from the SILVER-AMI study, a prospective longitudinal study of adults 75 years and older hospitalized for AMI. The design of SILVER-AMI and its variables have been previously described [11]. The outcome in this analysis is a binary indicator of death within 30 days of hospital discharge. Because our objective is to demonstrate a model selection process in detail, this study is restricted to selection of the final model and its internal validation rather than external validation in a separate cohort.

2.2. Selection of Candidate Variables and Multiple Imputation of Missing Values

More than 100 explanatory variables plausibly associated with AMI in older persons were initially selected from demographic, cardiac, and geriatric domains, based on prior risk prediction models and clinical judgement. After examining their distributions, the following rules excluded variables of dubious predictive value: missing > 20% and prevalence either < %5 or > 95%. In-hospital interventions with strong indication bias not addressable in the risk model were also excluded, leaving a pool of 83 candidates for the prognostic model. Missing data was assumed missing-at-random and imputed 10 times. We will show that the choice of 10 imputations provides a convenient context for comparing selection among all the approaches. Figure **1** illustrates the stages of the selection process.

2.3. Reducing the Pool of Potential Predictors From 80+ to 20 Candidate Variables

In order to compare selection among the discrete model approaches and BMA, we needed to first reduce our initial pool of 83 variables to a group of 20. This is because of the high computational burden of processing a large number of candidate variables (>

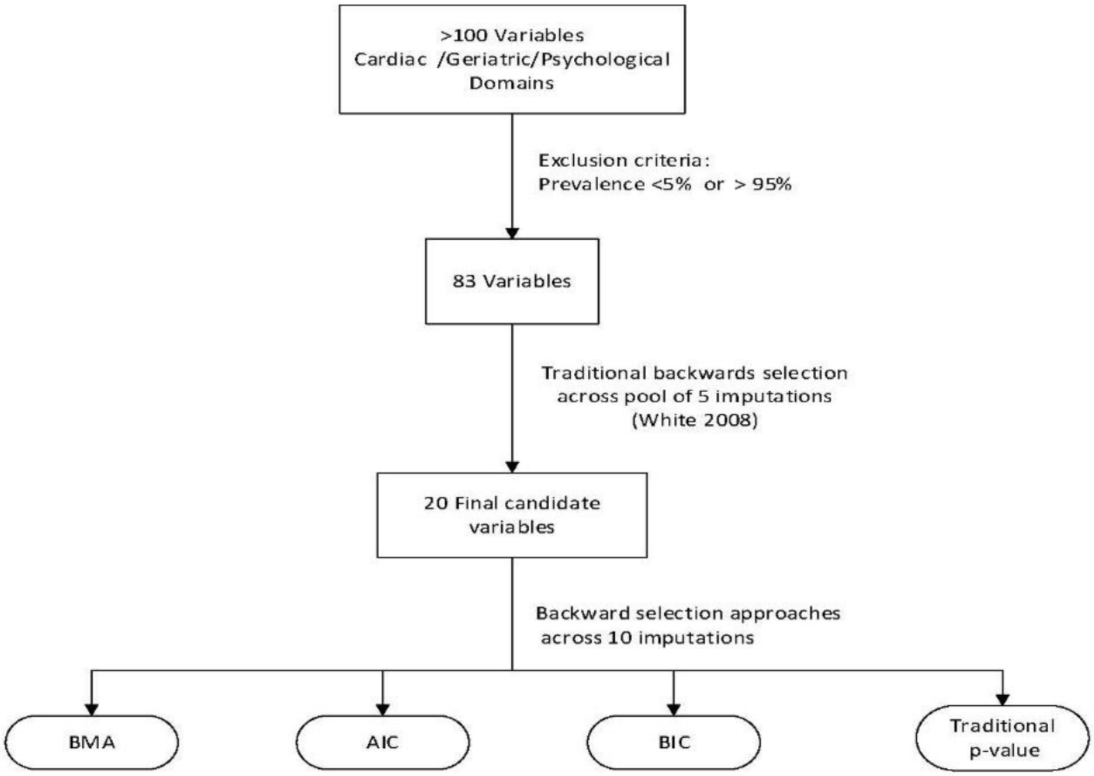
>100 Variables
Cardiac /Geriatric/Psychological Domains
Exclusion criteria:
Prevalence <5% or > 95%
83 Variables
Traditional backwards selection across pool of 5 imputations (White 2008)
20 Final candidate variables
Backward selection approaches across 10 imputations
BMA
AIC
BIC
Traditional p-value

Figure 1: Stages of Selection for the Final Multivariable Risk Prediction Model.

Abbreviations: BMA = Bayesian Model Averaging, AIC = Akaike Information Criterion, BIC = Bayesian Information Criterion.

20) in three of the approaches. Due to its use of a p-value criterion allowing sequential elimination of candidate variables, only BW readily accommodates a large number of candidate variables. Our SAS macros for selection with AIC and BIC did not employ a p-value criterion for successive elimination of variables. Consequently for more than 20 variables, the simultaneous consideration of multiple models involved high memory usage. In like fashion, the R package BMA [12], with its memory intensive task of evaluating all possible variable combinations, is limited to no more than 30 variables. To balance computational stability against adequate exercising of each selection approach, we compared the four approaches over these same 20 candidates. This was implemented with the suggestion of White *et al*. (2011) [13] by employing backward selection on a pool of five imputed datasets (rather than 10) to choose the 20 variables with the strongest multivariable associations. Because of the large size of this pooled dataset, the p-value for retention in the model was lowered to identify the group of 20 that collectively did the best in describing the variability of the outcome. Those 20 subsequently served as the candidates for the final multivariable chosen by BMA and the reference approaches

2.4. Comparing BMA with The Discrete Approaches

To perform a rigorous comparison of the backward-selection approaches, we employed the full complement of ten imputations. We applied backward-selection using multivariable logistic regression on each imputation with AIC [7], BIC [8], BW (p-value ≤ 0.05) and BMA [5]. While BW is widely recognized, AIC and BIC are information criteria that add the maximum likelihood of the model to a penalty that rises with number of variables. The AIC adds a modest penalty for each variable while BIC adds a larger penalty, causing it to choose more parsimonious models, and both are designed such that lower values represent better model fit. Common to the AIC, BIC and BW approaches is that each selects a single, optimal model assumed to be the true model of the outcome.

In counterpoint, BMA examines all the possible combinations of candidate variables and after selecting the best one, retains a subset of others within a range known as Occam's window, described in detail in Madigan and Raftery (1994) [14]. For variables in any of the best fitting models, BMA subsequently calculates a posterior effect probability (PEP), i.e., the probability the variable is associated with the outcome, which is compared against a minimal threshold. To illustrate a threshold of 50%, any candidate with PEP ≥ 50% is included in the final model. The choice of threshold allows flexibility for the myriad scenarios arising from varying sample sizes, candidate variables and multiply imputed datasets.

One issue that has not previously been addressed in prior statistical methods papers is how to select a risk prediction model over multiple imputations. Because missing values result in different subsets of the data at any given stage, they can bias model selection. For this reason, selection needs to take place over multiply imputed datasets. The question then becomes how to decide on a final set of variables when different "best" models are chosen from different imputations.

The application of model selection to 10 imputations provides a convenient way to resolve this issue. For BMA, the PEP values can simply be averaged across the ten imputations and compared against a threshold value such as 50%. For the discrete approaches, we suggest the following. When any variable has been chosen in at least five of the imputations, we interpret it as having a rough probability ≥ 50% of inclusion in the final model. The 50% threshold balances elimination of variables unlikely to predict the outcome against being overly selective. We will compare the final models chosen by each approach using this threshold of 50% probability of association (BMA) or inclusion (discrete approaches). We will demonstrate that this threshold facilitates a comparison of approach-based variable selection across the imputed datasets.

2.5. Comparing Final Models from the Four Different Approaches

The models chosen by the four approaches were compared for discrimination, calibration, optimism, and shrinkage. Discrimination is the ability of a model to correctly predict, on average, whether a given person will experience the outcome of interest, measured as the average C statistic across the imputations [15]. Calibration is a model's ability to consistently calculate the probability of the outcome across its entire range, with acceptable performance commonly interpreted as a p-value > 0.05 for the Hosmer-Lemeshow statistic in each imputation [16]. Optimism estimates the degree to which a model's predictive performance will be reduced when applied to an external dataset. We evaluate the optimism of the model's C statistic following Harrell [4], by fitting the final model to 100 bootstrapped samples of the development data and subsequently predicting performance in the original data.

Finally, shrinkage describes the extent to which the associations between predictors and outcome estimated during model development may be inflated relative to their performance in external data. We employ the heuristic estimator of van Heuwelingen and le Cessie to estimate the amount of shrinkage where higher values (i.e., closer to 1) reflect better performance [17]. Note that the estimates of optimism and shrinkage constitute internal validation, which, while drawing only from development data, provide insight regarding model performance in external datasets.

3. RESULTS

3.1. Method-specific Selection of Candidate Variables across Multiply Imputed Datasets

Shown in Table **1** are the number of times each variable was chosen in the discrete approaches over the 10 imputations. The last column presents the average PEP value calculated by BMA. Taking BW and its overall count of chosen variables as a reference (76 out of possible 200), AIC chose twice as many (159) whereas BIC chose nearly half as many (40). With one notable exception, BMA mimics the selection pattern of BIC. For example, the Short Form 12 measure of general health was consistently chosen by BIC, which concurs with the high PEP value (83%) calculated by BMA. BIC and BMA also agree in their selection of age, dyspnea, social support relaxation, and the telephone interview of cognitive status (TICS) score of mental function. The notable difference is seen in the variable denoting length of stay in hospital. Never chosen by BIC, the latter rates an average PEP of 83% from BMA.

3.2. Final Model Selected Across Multiply Imputed Datasets by Each Approach

The primary objective of this report is to demonstrate that BMA is a viable method for model selection. While BMA can provide its own sophisticated form of estimation, the final multivariable associations for comparing the approach-specific models were

Table 1: Frequencies of Candidate Variable Selection by Approach

Twenty Candidate Variables chosen by Backward Selection across Pool of Five Imputations	Percent of Times Chosen over 10 Imputations for Optimal Model by each Discrete Method			Average Posterior Effect Probability (Percent)
	AIC	BIC	BW	BMA
Acute kidney disease	100	0	100	4.7
Age in years at admission	100	100	100	99.2
End of month need help with finances	0	0	0	0.0
ESAS - Depressed	100	0	0	0.0
ESAS - Drowsy	100	0	0	0.0
ESAS - Dyspnea	100	50	90	47.2
First systolic blood pressure	100	0	30	4.2
Grip strength frailty	0	0	0	0.0
In-hospital bleeding event	70	0	0	0.0
In-hospital heart failure	100	0	0	0.0
In-hospital hyperglycemia	0	0	0	0.0
Length of stay	70	0	0	82.7
Month prior walking	100	0	0	13.0
PHQ evidence of depression	100	0	80	8.1
Prior history of CABG	50	0	0	0.0
Short form 12 general health	100	100	100	82.7
Social support: no one to relax with	100	50	90	38.0
TICS total score	100	100	100	81.6
Unintended WL	100	0	70	33.0
Visually impaired	100	0	0	0.0
Total number of variables chosen over 10 imputations (out of Possible 200)	159	40	76	N/A

Abbreviations: AIC = Akaike information criterion; BIC = Bayesian information criterion; BW = backward selection with p-value of 0.05; BMA = Bayesian model averaging; ESAS = Edmonton Symptom Assessment Scale; PHQ = patient health questionnaire; CABG = Coronary artery bypass-graft; TICS = modified telephone interview for cognitive status; WL= weight loss; N/A = not applicable.

estimated using logistic regression. Each approach's final model was fit to each of the 10 imputations and coefficients calculated using Rubin's rules in the SAS procedure MIanalyze [18].

Table **2** compares the final models chosen by approach across the 10 imputations by model performance. The AIC approach selected 17 variables while BW chose eight. These two approaches produce the highest accuracy, as measured by their respective C statistics of 88% and 84%. They also exhibit the worst performance in optimism and shrinkage, suggesting they will not hold up well in external datasets. In choosing five variables, BIC was slightly more liberal than BMA's choice of four. In both approaches age, short form 12 general health, and TICS were chosen, i.e., continuous measures of longevity, general physical health, and mental function, each carrying immediate face validity for mortality among older persons. BIC also chose dyspnea which was narrowly rejected by BMA (average PEP of 47%). Whereas BIC chose lack of social support for relaxing, BMA chose length of hospitalization, a surrogate for severity of the index AMI. The models chosen by BIC and BMA have respective C statistics of 79% and 78%, and the best values for optimism (1.7% and 1.3%) and

Table 2: Comparing Final Models Chosen by Four Different Backward-Selection Approaches from 20 Candidate Variables across 10 Multiple Imputations

	Final Models Chosen by Backward-Selection Approaches on 20 Candidate Variables over 10 Imputations (N= 2000 per imputation)			
Performance Metrics of Selected Model	**AIC (17 variables)**	**BIC (5 variables)**	**BW (8 variables)**	**BMA (4 variables)**
Variables Selected Across 10 Imputations [a]	Acute kidney disease Age in years ESAS - depressed ESAS - drowsy ESAS - dyspnea First diastolic BP In- hospital bleeding In- hospital HF Length of stay Month prior walking PHQ - depression Prior history CABG SF12genHealth Social Support NR TICS score Unintended WL Visually impaired	Age in years ESAS - dyspnea SF12genHealth Social support NR TICS score	Acute kidney disease Age in years ESAS - dyspnea PHQ - depression SF12genHealth Social Support NR TICS score Unintended WL	Age in years Length of stay SF12genHealth TICS score
Calibration (minimum p-value of H-L statistic)	0.08	0.19	0.46	0.10
Discrimination (C statistic higher is better)	88%	79%	84%	78%
Heuristic Shrinkage[b] (higher is better)	0.85	0.92	0.90	0.93
Optimism (lower is better)	3.5%	1.7%	2.0%	1.3%

[a] chosen from ≥ 5 imputations (AIC, BIC, and BW) or with average posterior probability ≥ 50% (BMA).
[b] from Van Heuwelingen JC and le Cessie S (LR– number of terms) / (LR).
abbreviations: AIC = Akaike information criterion; BIC = Bayesian information criterion; BMA = Bayesian model averaging; BP = blood pressure; BW = backward selection with p-value of 0.05; ESAS = Edmonton Symptom Assessment Scale; HF = heart failure; HL = Hosmer-Lemeshow goodness of fit ; LR = likelihood ratio chi-square; NR = no relaxation; PHQ = patient health questionnaire; SF12genHealth = general health on Short Form 12; TICS = modified telephone interview for cognitive status; WL= weight loss.

shrinkage (0.92 and 0.93). Based on these criteria, for this case BMA has performed as well as the discrete approaches.

4. DISCUSSION

We have demonstrated the selection process for a prognostic model of mortality within 30 days of discharge among older persons hospitalized for AMI. While largely following the practices of Harrell [4], we provide details on selecting a model from a large initial pool of candidate variables over multiply imputed datasets. We propose that the decision of which variables to retain over imputed datasets, after proper consideration for power constraints, be based on a threshold for either an average PEP value (BMA) or probability of inclusion (AIC, BIC, BW) that strikes a balance between excessive parsimony and retention of weak predictors. Two of the approaches, (AIC and BW) yielded models with an unjustifiably large number of variables. At the cost of reduced discrimination, the other two approaches (BIC and BMA) yielded leaner models with better internal validation, with BMA narrowly edging out BIC.

The strengths of this study include the source data, which draws exclusively from community dwelling adults age 75 or greater from 94 hospitals across the U.S. This data includes variables from diverse domains, including demographic, clinical, psychosocial and functional [19]. A second strength is that short-term mortality is a highly objective outcome which permits a high level of discrimination. The primary limitation for model selection with BMA is that, due to the computational burden of fitting all possible sub-models, its eponymous R package is limited to processing ≤ 30 candidate variables. This is a clear disadvantage relative to traditional BW selection, which can seamlessly handle a much larger number of candidate variables. A limitation of the outcome of 30-day mortality is its low rate of incidence (2.8%). Having started with a large number of potential predictors, we are far from complying with the optimal of ≥20 events-per-variable for selection of prognostic models [20]. A second limitation is that our internal validation used the final models from each approach rather than replicating the selection of the final model from the entire pool of candidates, suggesting optimism may be higher than that estimated here [21]. Despite these limitations, the primary objective was to demonstrate that BMA is a viable selection approach across multiply imputed datasets. Assuming the stated limitations affect the discrete model selection approaches in equivalent measure, BMA performed as well as the discrete approaches, and its calculation of PEP facilitates variable selection across multiple imputations.

5. CONCLUSION

In conclusion, we have demonstrated that relative to the discrete model approaches, BMA chose a leaner model that demonstrates strong face validity and performed well in internal validation. This study demonstrates that BMA is a viable technique for building a prognostic model from a large group of potential predictors that include missing data, and that relative values of the average of the posterior effect probabilities across multiple imputations can be used to facilitate final selection of variables.

FUNDING

SILVER-AMI is supported by the National Institutes of Health (NIH) National Heart, Lung, and Blood Institute (NHLBI) grant R01HL115295 and the Yale Coordinating Center is located at the Yale Program on Aging/Claude D.Pepper Older Americans Independence Center (P30 AG021342). Dr. Dodson is supported by the NIH National Institute of Aging (NIA) grant R03AG045067, a T. Franklin Williams Scholarship Award (funding provided by: Atlantic Philanthropies, Inc, the John A. Hartford Foundation, the Alliance for Academic Internal Medicine-Association of Specialty Professors, and the American College of Cardiology), and is the recipient of a Clinical Research Loan Repayment award from the NHLBI; Dr. Gill is the recipient of an Academic Leadership Award (K07AG043587) from the National Institute on Aging; Dr. Hajduk is supported by a NIA Training Grant in Geriatric Clinical Epidemiology (T32AG1934); Dr. Kim is supported by a Paul B. Beeson Clinical Scientist Development Award in Aging (K08AG051187) from the National Institute on Aging; Dr. Oh is supported by K23AG043504-01 (NIH/NIA). Dr. Allore is supported by R01 AG047891 (NIH/NIA).

REFERENCES

[1] Burnham K, Anderson D. Multimodel inference: understanding AIC and BIC in Model Selection. Sociological Methods & Research 2004; 33(2): 261-304. https://doi.org/10.1177/0049124104268644

[2] Hurvich CM, Tsai C-L. The impact of model selection on inference in linear regression. The American Statistician 1990; 44(3): 214-7

[3] Posada D, Buckley TR. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst Biol 2004; 53(5): 793-808. https://doi.org/10.1080/10635150490522304

[4] Harrell FE Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996; 15(4): 361-87. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

[5] Raftery AE, Richardson S. Model selection for generalized linear models via GLIB, with application to epidemiology. In *Bayesian Biostatistics* (Berry DA, Stangl DK, Eds.), New York: Marcel Dekker 1996; pp. 321-354.

[6] Hoeting JA, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. Statistical Science 1999; 14(4): 382-417.

[7] Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control 1974; AC-19(6): 716-23. https://doi.org/10.1109/TAC.1974.1100705

[8] Schwarz G. Estimating the dimension of a model. Annals of Statistics 1978; 6(2): 461-4. https://doi.org/10.1214/aos/1176344136

[9] Goodman S. A dirty dozen: twelve p-value misconceptions. Semin Hematol 2008; 45(3): 135-40. https://doi.org/10.1053/j.seminhematol.2008.04.003

[10] Wasserstein RL, NLL. The ASA's statement on p-values: context, process, and purpose. Am Stat 2016; 70: 129-33. https://doi.org/10.1080/00031305.2016.1154108

[11] Dodson JA, Geda M, Krumholz HM, Lorenze N, Murphy TE, Allore HG, *et al*. Design and rationale of the comprehensive evaluation of risk factors in older patients with AMI (SILVER-AMI) study. BMC Health Serv Res 2014; 14: 506.PMCPMC4239317.

[12] Raftery A, Hoeting JA, Volinsky C, Painter I, Yeung KY. R package 'BMA' 2015.

[13] White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med 2011; 30(4): 377-99. https://doi.org/10.1002/sim.4067

[14] Madigan D, Raftery A. Model selection and accounting for model uncertainty in graphical models using Occam's window. Journal of the American Statistical Association 1994; 89: 1535-1546. https://doi.org/10.1080/01621459.1994.10476894

[15] Hanley JA, BJ M. The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. Radiology 1982; 143(1): 29-36. https://doi.org/10.1148/radiology.143.1.7063747

[16] Hosmer DW, S L. Applied Logistic Regression. New York: Wiley 2013.

[17] van Houwelingen JC, le Cessie S. Predictive value of statistical models. Stat Med 1990; 8: 1303-25. https://doi.org/10.1002/sim.4780091109

[18] Rubin DB. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons Inc., New York 1987. http://dx.doi.org/10.1002/9780470316696

[19] Newell MC, Henry JT, Henry TD, Duval S, Browning JA, Christiansen EC, *et al*. Impact of age on treatment and outcomes in ST-elevation myocardial infarction. Am Heart J 2011; 161(4): 664-72. https://doi.org/10.1016/j.ahj.2010.12.018

[20] Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996; 49(12): 1373-9. https://doi.org/10.1016/S0895-4356(96)00236-3

[21] Steyerberg E. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating (pages 87-89): Springer 2008; p. 500.

Received on 12-02-2019 Accepted on 15-03-2019 Published on 05-04-2019

https://doi.org/10.6000/1929-6029.2019.08.01