CAUCHY –Jurnal Matematika Murni dan Aplikasi
Volume 7(1) (2021), Pages 13-21
p-ISSN: 2086-0382; e-ISSN: 2477-3344

A Combination of Generalized Linear Mixed Model and LASSO Methods for Estimating Number of Patients Covid 19 in the Intensive Care Units

Alona Dwinata[1,2], Khairil Anwar Notodiputro[2]*, Bagus Sartono[2]

[1]Mathematics Education Study Program, Raja Ali Haji Maritime University,
Tanjungpinang
[2]Department of Statistics, IPB University, Bogor

Email: alonadwinata@umrah.ac.id, khairil@apps.ipb.ac.id*, bagusco@apps.ipb.ac.id
**Corresponding Author*

ABSTRACT

Generalized linear mixed models (GLMM) combined with the L_1 penalty (Least Absolute Shrinkage and Selection Operator/LASSO) is called LASSO GLMM. LASSO GLMM reduces overfitting and selects predictor variables in modeling. The aim of this study is to evaluate the performance model for predicting Covid-19 patients with certain congenital disease that require ICU based on the results of blood tests laboratory and patient's vital signs. This study used binary response variables, 1 if the patient was admitted to the ICU and 0 if the patient was not admitted to the ICU. The fixed effect predictor variables are the results of blood tests laboratory and patient's vital signs. The random effect predictor variable is patient's congenital disease. The result showed that the average of accuracy and AUC from LASSO GLMM is more than the average of accuracy and AUC from LASSO GLM by using 5% level of significance. Respiratory rate and Lactate show a significance effect to predict the ICU needs of Covid-19 patients. The random effects patient's congenital disease has significance effect at 5% level of significance. It means that the ICU needs for Covid-19 patients varies among patient's congenital disease. We can conclude that GLMM LASSO with the random effect of patient's congenital diseases has better modeling performance to predict the ICU needs of Covid-19 patients based on the results of blood tests laboratory and patient's vital signs. The results of this modeling can quickly detect Covid-19 patients who need the ICU and can help medical staff use ICU resources optimally.

Keywords: Covid 19; GLMM; glmmLasso; LASSO

INTRODUCTION

Generalized linear model (GLM) is an approach that can be used to model the effect of predictor variables on response variables derived from exponential family distribution. For observations in certain groups there is usually a correlation between observations then the GLM study is expanded to include random effects on linear predictors. When the GLM model added a random effect, the model called Generalized Linear Mixed Models (GLMM) [1]. GLMM modeling has a problem with the number of predictor variables used in relation to complexity in modeling. The more predictor variables used in modeling, the estimation is very unstable [2]. The existence of predictor variables that are not related

Submitted: Pebruary 14, 2021 Reviewed: April 20, 2021 Accepted: October 12, 2021
DOI: http://dx.doi.org/10.18860/ca.v7i1.11575

to the response variables in the model will cause overfitting problems. To improve the accuracy of the model prediction, a penalty is added in modeling [3].

The addition of penalty function in modeling was carried out by Tibshirani (1996) using the L1 penalty, namely $\lambda \sum_{j=1}^{p} |\beta_j|$ which is called Least Absolute Shrinkage and Selection Operator (LASSO). Lambda (λ) in the L1 penalty function is a shrinkage parameter (λ) that determines the amount of shrinkage regression coefficient. LASSO reduces overfitting and selects predictor variables in modeling [4]. Modeling with a combination of GLM and GLMM with LASSO techniques in this study are called LASSO GLM and LASSO GLMM. Researchers have discussed various problems on LASSO GLM, such as Arnold and Tibshirani (2016) [5], Hossain et al. (2015) [6], Zhang and Zou (2014) [7], Simon et al. (2013) [8], Friedman et al. (2010) [9]. The LASSO GLM optimizes the objective function by using coordinate descent optimization. This algorithm is available in the R programming language, namely glmnet package [9].

Some researchers have discussed variable selection procedures in GLMM using the L1 penalty, including Thomson and Hossain (2018) [10], Groll and Tutz (2014) [2], Schelldorfer et al. (2011) [11], Ibrahim et al. (2010) [12]. The LASSO GLMM produces stable estimations because penalty L1 can select the important predictor variables used in GLMM [2]. The GLMMs using the L1 penalty are useful whenever there is a grouping structure among high dimensional observations [11]. Previous studies also have found an algorithm for estimating the maximum likelihood in the GLMM model with the addition of the L1 penalty function. The penalized loglikelihood function maximize using gradient ascent algorithm, this algorithm is called GLMMLasso [13]. The GLMMLasso algorithm in the R programming language is included in the glmmLasso package [14].

In this study, researchers apply LASSO GLM and LASSO GLMM to predict the ICU needs for Covid-19 patients. The surge in Covid-19 cases is putting enormous pressure on the health care system. Intensive Care Units (ICU) is one of the health facilities needed by patients with Covid-19 confirmation. The study examines the prediction of ICU for Covid-19 patients. The ICU needs for Covid-19 patients were analyzed using the results of blood tests laboratory, vital signs and the patient's congenital disease. The predictor variables for blood test laboratory results and patient's vital signs were fixed effect, whereas predictor variables for patient's congenital disease were assumed to be fixed effect for LASSO GLM and random effect for LASSO GLMM. Previous researchers have discussed the performance of LASSO GLM and LASSO GLMM modeling on rainfall data, the results showed that modeling with LASSO GLMM has better performance than LASSO GLM [15]. To predict the ICU needs for Covid-19 patients based on laboratory results of blood tests, patient's vital signs and congenital disease, researchers conducted modeling with LASSO GLM and LASSO GLMM. The aim of this study is to evaluate the model's performance in predicting Covid-19 patients with certain congenital disease groups that require ICU based on the results of blood tests laboratory and patient's vital signs.

METHODS

Data

The study used data from patients confirmed by Covid-19 at the Sírio-Libanês Hospital, São Paulo, Brasilia. Data were collected after 12 hours of confirmed Covid-19 patients undergoing treatment in the hospital. Total data were 98 patients, with 52 ICU patients and 46 non-ICU patients.

The study used binary response variables, 1 if the patient was admitted to the ICU and 0 if the patient was not admitted to the ICU. The fixed effect predictor variables for modeling totally used 32 variables, 26 variables from the results of blood tests laboratory and 6 variables patient's vital signs. The fixed effect predictor variables used in modeling can be seen in Table 1. Researchers assumed patient's congenital disease as fixed effect predictor variables in modeling using LASSO GLM and a random effect predictor variable in modeling using LASSO GLMM.

Table 1. Research Variables

Variable	Variable Name	Type	Information
Y	The Covid-19 patient's status	Binary	1 = ICU patient, 0 = Non-ICU patient
X1	ALBUMIN	Numeric	fixed effect
X2	BE_VENOUS	Numeric	fixed effect
X3	BIC_VENOUS	Numeric	fixed effect
X4	BILLIRUBIN	Numeric	fixed effect
X5	CALCIUM	Numeric	fixed effect
X6	CREATININ	Numeric	fixed effect
X7	FFA	Numeric	fixed effect
X8	GGT	Numeric	fixed effect
X9	GLUCOSE	Numeric	fixed effect
X10	HEMATOCRITE	Numeric	fixed effect
X11	HEMOGLOBIN	Numeric	fixed effect
X12	LACTATE	Numeric	fixed effect
X13	LEUKOCYTES	Numeric	fixed effect
X14	LINFOCITOS	Numeric	fixed effect
X15	NEUTROPHILES	Numeric	fixed effect
X16	P02_VENOUS	Numeric	fixed effect
X17	PC02_VENOUS	Numeric	fixed effect
X18	PCR	Numeric	fixed effect
X19	PH_VENOUS	Numeric	fixed effect
X20	PLATELETS	Numeric	fixed effect
X21	POTASSIUM	Numeric	fixed effect
X22	SAT02_VENOUS	Numeric	fixed effect
X23	SODIUM	Numeric	fixed effect
X24	TTPA	Numeric	fixed effect
X25	UREA	Numeric	fixed effect
X26	DIMER	Numeric	fixed effect
X27	BLOODPRESSURE_DIASTOLIC	Numeric	fixed effect
X28	BLOODPRESSURE_SISTOLIC	Numeric	fixed effect
X29	HEART_RATE	Numeric	fixed effect
X30	RESPIRATORY_RATE	Numeric	fixed effect
X31	TEMPERATURE	Numeric	fixed effect
X32	OXYGEN_SATURATION	Numeric	fixed effect

Research methods

Modeling was carried out to predict the ICU needs for Covid-19 patients based on the results of blood tests laboratory, vital signs and congenital diseases. There are many predictor variables used in modeling. We select the variables to determine the important predictor variables, then a simpler model is obtained by adding the L1 penalty function to the model. The algorithms of this research were as follows:

1. LASSO GLM modeling predicted the ICU needs for Covid-19 patients
   a. Determine the optimum lambda value
   b. LASSO GLM modeling used the R package glmnet
   c. Analyze the parameters from the modeling results
   d. Determine the model accuracy.
2. LASSO GLMM modeling predicted the ICU needs for Covid-19 patients
   a. Determine the optimum lambda value
   b. LASSO GLMM modeling used the R package glmmLasso
   c. Analyze the parameters from the modeling results
      The random effects in modeling used hypothesis $H_0: \sigma^2 = 0$. This hypothesis was tested by using likelihood ratio, $G^2 = 2(\text{loglik}_{LASSOGLMM} - \text{loglik}_{LASSOGLM})$. If $G^2 > \chi^2_{(db=1,\alpha=0.05)}$ then H_0 is rejected.
   d. Determine the model accuracy.

To evaluate the performance of LASSO GLM and LASSO GLMM, researchers have chosen the best model to predict the hospitalization needs of a patient with Covid-19. The best model was selected based on accuracy and AUC. The steps for selecting the best model were as follows:

a. Partition data with a composition of 80% modeling data and 20% validation data. Data partitioning was performed 30 times
b. Modeling the LASSO GLM and LASSO GLMM used modeling data for each replication
c. Assessing model performance based on AUC and accuracy values using validation data for each replication
d. Statistically perform a performance difference of LASSO GLM and LASSO GLMM used paired sample t-test.

RESULTS AND DISCUSSION

1. LASSO GLM Modeling

LASSO GLM selects variables based on λ. The λ optimum is obtained when the binomial deviance value is minimum. Cross validation plot to optimize LASSO GLM shrinkage parameters is shown in Figure 1.

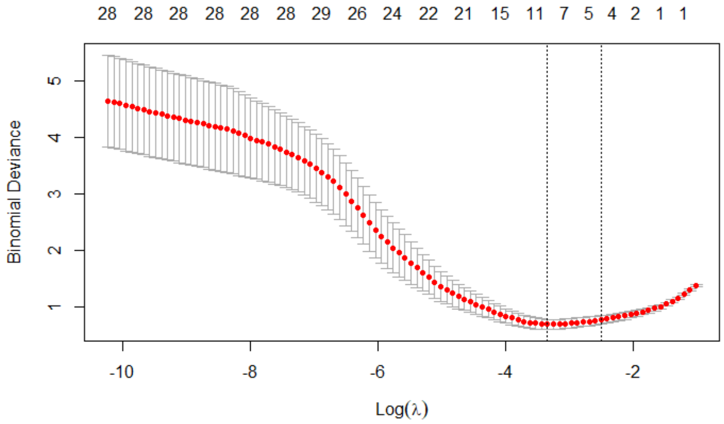

Figure 1. Cross validation plot to optimize GLM LASSO shrinkage parameters

Based on Figure 1, the optimum λ was 0.024. The predictor variables included in the modeling are fixed effect predictor variables. There are 26 features laboratory blood test results and 6 patient's vital signs, and a patient's congenital disease as dummy variable.

LASSO GLM modeling used the R package glmnet. The plot of the LASSO GLM coefficient for each log λ value can be seen in Figure 2. The regression coefficient with non-zero values results from the LASSO GLM modeling is shown in Table 2.

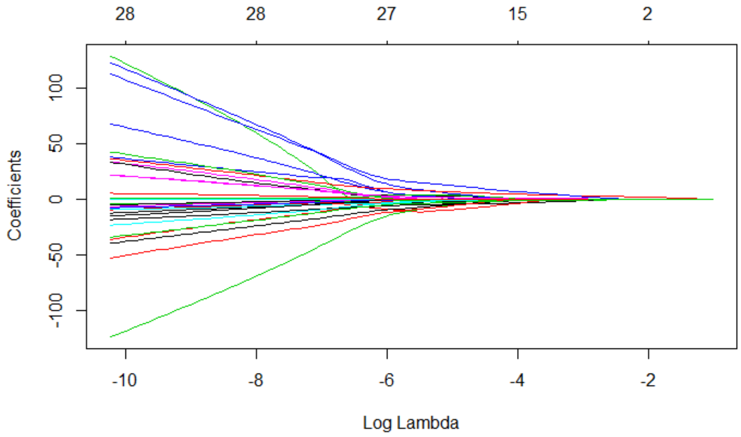

Figure 2. Plot of LASSO GLM coefficients for each shrinkage parameter value

Table 2. LASSO GLM coefficient

Variables	Coefficient
LACTATE	-0.45
P02_VENOUS	-1.82
SODIUM	-0.03
BLOODPRESSURE_SISTOLIC	0.83
RESPIRATORY_RATE	3.78
OXYGEN_SATURATION	4.26
HTN	0.23
Disease Group 1	-0.52

2. LASSO GLMM Modeling

The same as LASSO GLM, LASSO GLMM also required optimum λ in modeling. Figure 3 shows the binomial deviance value for each value of λ. The optimum λ is 19.6 that obtained when the smallest deviance.

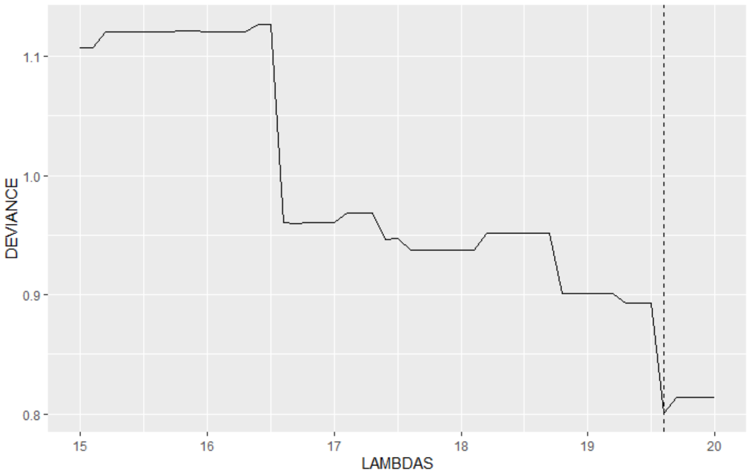

Figure 3. Cross validation plot for optimizing LASSO GLMM shrinkage parameters

There are 26 features laboratory blood test results, 6 patient's vital signs and a patient's congenital disease as random effect in LASSO GLMM. The modeling use R package glmmLasso. The plot of the LASSO GLMM coefficient spread for each λ can be seen in Figure 4. The regression coefficients go to zero along to the increasing λ. The regression coefficient of the LASSO GLMM modeling with λ = 19.6 result 4 non-zero predictor variables that is shown in Table 3.

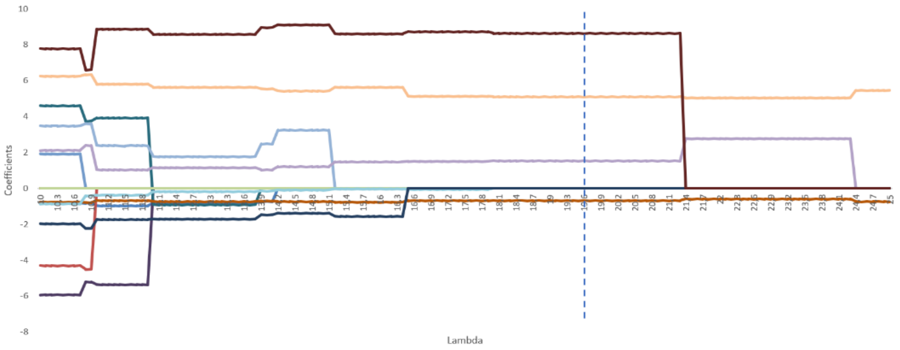

Figure 4. Plot of LASSO GLMM coefficients for each shrinkage parameter

Table 3. LASSO-Penalized Logistic Mixed Effects Regression Model (GLMM-Lasso)

Fixed Effects	Coefficient	Standard Error	Z	P(>|Z|)
(Intercept)	-6.88	0.54	-12.636	0.000
LACTATE	-0.65	0.38	-1.70	0.08
BLOODPRESSURE_SYSTOLIC	1.55	1.88	0.82	0.41
RESPIRATORY_RATE	5.11	1.25	4.09	0.04
OXYGEN_SATURATION	8.64	5.36	1.61	0.11

The patient's congenital disease as random effect had standard deviation 0.8262 with $G^2 = 4.12$ dan $\chi^2_{(db=1,\alpha=0.05)}$ =3.84. Then, H_0 is rejected. It means that the random effects for patient's congenital disease was significant at 5% level of significance.

3. Selection of the best model

Data were divided randomly with a composition of 80% modeling data and 20% validation data. Furthermore, there are 79 patients as modeling data and 19 patients as validation data. Data partitioning was carried out in 30 replications. The optimum λ is obtained based on the modeling data taken for each replication.

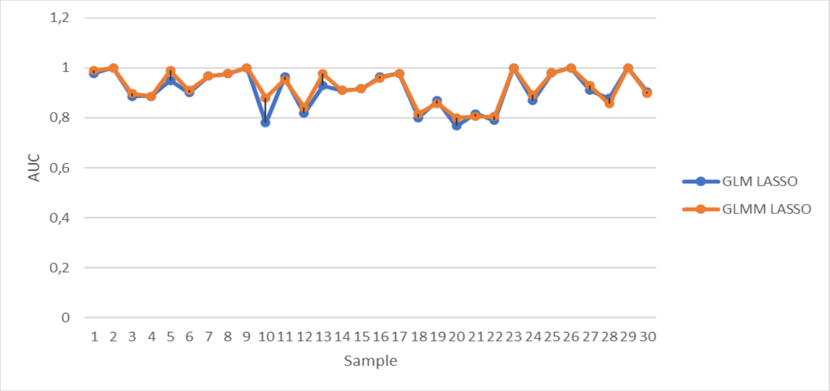

Figure 5. AUC of LASSO GLM and LASSO GLMM for 30 replications

Furthermore, the LASSO GLM and LASSO GLMM modeling was carried out for each replication. Assessment of modeling performance use the accuracy and AUC in the validation data. Comparison of the accuracy and AUC of 30 replications for each model is shown in Figure 5 and Figure 6.

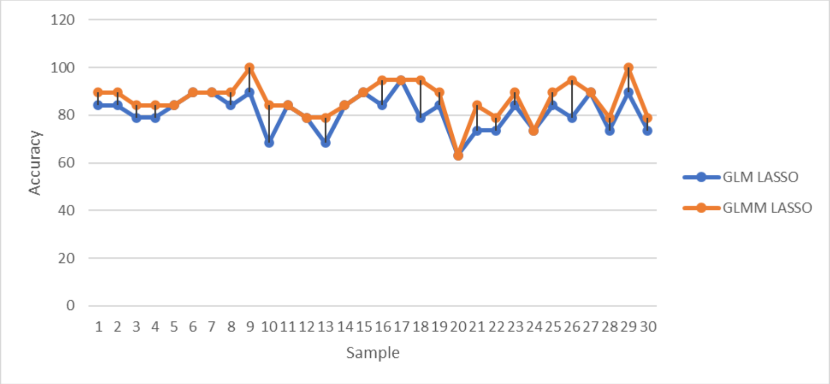

Figure 6. Accuracy of LASSO GLM and LASSO GLMM for 30 replications

The performance differences of LASSO GLM and LASSO GLMM can be statistically stated by paired sample t-test of AUC and accuracy. The results of the paired sample t-test for these two models can be seen in Table 4. The hypothesis about accuracy or AUC of the two models is as follows:

- H_0: Average accuracy of LASSO GLM is less than or equal to average accuracy of LASSO GLMM
- H_0: Average AUC of LASSO GLM is less than or equal to average AUC of LASSO GLMM

Table 4. The paired sample t-test of accuracy and AUC

Criteria	t-stat	p-value
accuracy	5.5746	0.0000
AUC	2.2058	0.0178

The t-test results in Table 4 showed the p-value for accuracy and AUC less than 0.05. It means the average of accuracy and AUC from LASSO GLMM is more than the average of accuracy and AUC from LASSO GLM by using 5% level of significance.

Discussion

The ability to identify patients who need the ICU is needed. The solution to this problem can be done by identifying the most important variables that affect the ICU needs for Covid-19 patients. The paired sample t-test of accuracy and AUC in Table 4 showed that modeling with LASSO GLMM has better performance than LASSO GLM. Figure 4 shows the effect of the predictor variables for each lambda value. By using lambda 19.6, this model produced four non-zero fixed effect predictor variables which are the focus of attention to predict the ICU needs of Covid-19 patients, namely Lactate, Blood pressure systolic, Respiratory rate and Oxygen saturation. Among these four predictors, only respiratory rate had a significant effect at the 5% level of significance and Lactate had a significant effect at the 10% level of significance. Meanwhile, Blood Pressure Systolic and Oxygen Saturation had no significant effect.

The odds ratio of respiratory rate was 165.67. It meant that the odds of Covid-19 patient required the ICU was 165.67 higher given an increase of a unit respiratory rate (respirations per minute/rpm) than before the increase. Covid-19 damages the respiratory system. Respiratory rate is one measure used to identify respiratory tract infections immediately before and during the first days of symptoms. The normal respiratory rate for adults at rest is 12 to 20 rpm [16]. The findings of a study suggest that the stability of nightly respiratory rate measurements in healthy individuals at night rest is a useful metric for tracking changes in health [16].

The odds ratio of Lactate was 0.52. It meant that the odds of a Covid-19 patient required the ICU was 0.52 lower given an increase of a unit Lactate (mmol/L) than before the increase. Arterial lactatemia higher than central vein (a reversed Delta a-cv lactate) indicates a disturbance in the mitochondrial metabolism of lung cells caused by severe inflammation [17]. An increase in one unit of venous blood lactate reduces reversed delta a-cv lactate.

LASSO GLMM produced an AUC of 0.96. This means that GLMM LASSO has good predictive performance in predicting the ICU needs of Covid-19 patients. The random effects patient's congenital disease was significant at 5% level of significance. It means that the ICU needs for Covid-19 patients varies among patient's congenital disease. We can conclude that GLMM LASSO with the random effect of patient’s congenital diseases has better modeling performance to predict the ICU needs of Covid-19 patients.

CONCLUSIONS

In this study, modeling with LASSO GLMM has better performance to predict the ICU needs of Covid-19 patients than LASSO GLM. LASSO GLMM has good predictive performance in predicting the ICU needs of Covid-19 patients with an AUC 0.96. Respiratory rate has a significant effect at 5% level of significance and Lactate has a significant effect at 10% level of significance in LASSO GLMM. Respiratory rate shows the largest significance effect to predict the ICU needs of Covid-19 patients. Random effects of patient congenital disease had a significant effect on covid-19 patients requiring ICU at 5% level of significance. It means that the ICU needs for Covid-19 patients varies among patient's congenital disease. We can conclude that GLMM LASSO with the random effect of patient’s congenital diseases has better modeling performance to predict the ICU needs of Covid-19 patients based on the results of blood tests laboratory and patient‘s vital signs.

ACKNOWLEDGMENTS

The authors would like to thank to all persons who contributed to the improvement of this paper.

REFERENCES

[1] J. Jiang, *Linear and Generalized Linear Mixed Models and Their Applications*. 2007.

[2] A. Groll and G. Tutz, “Variable selection for generalized linear mixed models by L1-penalized estimation,” *Stat. Comput.*, 2014, doi: 10.1007/s11222-012-9359-z.

[3] T. Hastie, R. Tibshirani, and J. Friedman, *Springer Series in Statistics*, vol. 27, no. 2. New York, NY: Springer New York, 2008.

[4] R. Tibshirani, “Regression shrinkage and selection via the lasso: A retrospective,” *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 73, no. 3, pp. 273–282, 2011, doi: 10.1111/j.1467-9868.2011.00771.x.

[5] T. B. Arnold and R. J. Tibshirani, "Efficient Implementations of the Generalized Lasso Dual Path Algorithm," *J. Comput. Graph. Stat.*, vol. 25, no. 1, pp. 1–27, 2016, doi: 10.1080/10618600.2015.1008638.

[6] S. Hossain, S. E. Ahmed, and K. A. Doksum, "Shrinkage, pretest, and penalty estimators in generalized linear models," *Stat. Methodol.*, vol. 24, pp. 52–68, 2015, doi: 10.1016/j.stamet.2014.11.003.

[7] T. Zhang and H. Zou, "Sparse precision matrix estimation via lasso penalized D-trace loss," *Biometrika*, 2014, doi: 10.1093/biomet/ast059.

[8] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group lasso," *J. Comput. Graph. Stat.*, 2013, doi: 10.1080/10618600.2012.681250.

[9] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Stat. Softw.*, vol. 33, no. 1, pp. 1–22, 2010, doi: 10.18637/jss.v033.i01.

[10] T. Thomson and S. Hossain, "Efficient shrinkage for generalized linear mixed models under linear restrictions," *Sankhya Indian J. Stat.*, 2018.

[11] J. Schelldorfer, P. Bühlmann, and S. Van De Geer, "Estimation for High-Dimensional Linear Mixed-Effects Models Using ℓ1-Penalization," *Scand. J. Stat.*, vol. 38, no. 2, pp. 197–214, 2011, doi: 10.1111/j.1467-9469.2011.00740.x.

[12] J. G. Ibrahim, H. Zhu, R. I. Garcia, and R. Guo, "Fixed and Random Effects Selection in Mixed Effects Models," *Biometrics*, 2011, doi: 10.1111/j.1541-0420.2010.01463.x.

[13] J. Schelldorfer, L. Meier, and P. Bühlmann, "GLMMLasso: An algorithm for high-dimensional generalized linear mixed models using ℓ1-penalization," *J. Comput. Graph. Stat.*, vol. 23, no. 2, pp. 460–477, 2014, doi: 10.1080/10618600.2013.773239.

[14] A. Groll, "glmmLasso: Variable Selection for Generalized Linear Mixed Models by L1-Penalized Estimation," 2017.

[15] A. Muslim, M. Hayati, B. Sartono, and K. A. Notodiputro, "A Combined Modeling of Generalized Linear Mixed Model and LASSO Techniques for Analizing Monthly Rainfall Data," 2018, doi: 10.1088/1755-1315/187/1/012044.

[16] D. Miller *et al.*, "Analyzing changes in respiratory rate to predict the risk of COVID-19 infection.," vol. 2, pp. 1–10, 2020, doi: 10.1101/2020.06.18.20131417.

[17] G. Nardi *et al.*, "Lactate Arterial-Central Venous Gradient among COVID-19 Patients in ICU: A Potential Tool in the Clinical Practice," *Crit. Care Res. Pract.*, 2020, doi: 10.1155/2020/4743904.