IDENTIFYING KEY PREDICTORS OF GESTATIONAL DIABETES MELLITUS USING PENALIZED LOGISTIC REGRESSION

[1]Olayiwola T. A, [2]Olayiwola O. D and [2]Olayiwola O. M
[1]Department of Biochemistry, FUNAAB
[2]Department of Statistics, FUNAAB
Tijesuniolayiwola06@gmail.com

Abstract

Gestational diabetes mellitus (GDM) remains a major public health concern in Nigeria, with rising prevalence and substantial maternal and neonatal complications. Several studies have identified potential risk factors; however, most were limited to descriptive statistics and bivariate analyses, offering little insight into variable selection in settings with many correlated predictors. This study applies penalized logistic regression to identify key determinants of GDM symptoms using a real-world antenatal dataset from Nwose et al. (2023). After extensive data cleaning, which resulted in 17 complete cases and 17 predictors. Both ridge and lasso logistic regression models were fitted to address multicollinearity and prevent overfitting. Model comparison using Akaike Information Criterion (AIC) indicated that the lasso model (AIC = 12.16) outperformed the ridge model (AIC = 34). Lasso penalization further enabled variable selection, identifying gestational week, family history of type 2 diabetes mellitus, and polycystic ovary syndrome as the most influential predictors of GDM symptoms. The results highlighted the importance of familial metabolic risk and reproductive health factors in GDM screening.

Keywords: Gestational Diabetes Mellitus (GDM), Penalized Logistic Regression, LASSO Regression, Predictor Selection, Maternal Health Analytics.

1. Introduction

Gestational diabetes mellitus (GDM) is defined as glucose intolerance first detected during pregnancy. Its incidence is increasing globally, with recent studies reporting a prevalence of approximately 11% in Nigeria (Azeez et al., 2021). Risk factors commonly associated with GDM include advanced maternal age, high body mass index (BMI), sedentary lifestyle, and family history of diabetes. Early identification of risk factors is essential for timely intervention to prevent adverse pregnancy outcomes.

Onyenekwe et al. (2019) looked at the prevalence of GDM and associated risk factors in a population of pregnant women in Enugu, South East Nigeria and reported high prevalence of GDM in the sample. Risk factors for GDM identified were aged Gestational age, gravidity, parity, miscarriages and live births. John et al. (2019) looked retrospectively at the prevalence of GDM and its risk factors in pregnant women attending antenatal clinic in Rivers state around 2017. They reported that the prevalence of GDM was 10.5%. Positive history of GDM in previous pregnancy was the only independent risk factor. GDM mothers had a significantly higher risk of developing pre-eclampsia. Neonates of GDM mothers were at increased risk of fetal macrosomia and neonatal admissions. They concluded that the prevalence of GDM was high and that those with GDM were at increased risk of developing fetal and maternal complications. A history of GDM in previous pregnancy was an essential risk factor for subsequent GDM.

On the knowledge of GDM among pregnant Women, Ladan and Ibrahim (2023) noted that there was only 41% of awareness of GDM among pregnant women attending antenatal care in North West, Nigeria, concluding a poor knowledge of GDM in that region. They reported that the major sources of information for the awareness were friends/neighbors and health workers. The knowledge of common risk factors for GDM/DM were mainly family history of type 2 DM and obesity. Azeez et al. (2021) did a meta-analysis of the prevalence and determinants of gestational diabetes mellitus in Nigeria from studied done been done between 2000 and 2020. They reported that the pooled prevalence of GDM in Nigeria was 11.0%. The most frequently reported determinants of GDM were previous macrosomic babies, maternal obesity, family history of diabetes, previous miscarriage, and advanced maternal age.

Omo-Aghoja et al. (2023) did a prospective cohort analytical observational study of blood glucose levels amongst two cohorts of women who attended antenatal care at the obstetric unit of Delta State University Teaching Hospital, Oghara. They reported the prevalence of GDM was 31.3% and 9.4%, respectively, for cases and controls. Adeoye et al. (2025) studied incidence, risk factors and pregnancy outcomes of GDM in Ibadan, Southwest Nigeria. They reported that the cumulative incidence of GDM was 20.7%. The mean time for the diagnosis of GDM was 25.4±1.42 weeks of gestation. Identified significant risk factors were maternal age ≥35 years, maternal obesity and a previous history of congenital anomaly. Women with GDM had a higher risk for elective CS. Odoh et al. (2025) presented the prevalence and predictors of GDM in Enugu, Nigeria using the International Federation of Gynecology and Obstetrics (FIGO) and reported prevalence of 5.5%, with obesity and family history of diabetes mellitus as predictors of GDM.

All of the studies above but Adeoye et al. (2025) were majorly descriptive statistics and contingency table analysis of data on GDM collected. Adeoye et al. (2025) evaluated the association between GDM and pregnancy outcomes also using bivariate log-binomial regression models. This study analyzes a real-world Nigerian antenatal dataset of Nwose et al. (2023), GDM register with risk factors for screening selection criteria pilot dataset. In order to evaluate risk factors related to the presence of GDM symptoms, and its prediction, this study employed penalized logistic regression. This approach overcomes the problem of logistic regression of overfitting if there are many variables and multicollinearity. Penalized logistic regression modifies the loss function by adding a penalty term that depends on the magnitude of the regression coefficients. Also, the penalty discourages large coefficients, preventing extreme values that can lead to overfitting. In addition, some penalties, particularly Lasso, shrink the coefficients of less important variables exactly to zero, effectively removing them from the model and creating a more parsimonious, interpretable solution (Yan et al., 2022). The choice of penalized logistic regression was necessitated by the real life GDM dataset situations, including large number of independent variables, multicollinearity among independent variables and small sample size.

2. Methods

The dataset used for this study from Nwose et al. (2023), is publicly available. GDM, if unmanaged can complicate pregnancy outcomes. Selective screening of GDM is a common policy hence, the need for complete medical records of patients. The extent and pattern that incomplete documentation of patients' records can prevent recall of antenatal patients requires elucidation. Initial data were collected in 2018, which continued in 2019 at Eku Baptist Government Hospital

(EBGH). Demographic data were complete in all patients, but incomplete documentation was observed with as much as 98%. 301/391 lacked complete data about 95% of the cases, this was solely due to missing height measurements. In 2020, records of 123 case files were reviewed for effectiveness of phone contacts to do telehealth, and with simultaneous GDM risk assessment. 98/123 have phone details on medical records, of which 41/98 cases followed up were reached hence constituted the pilot dataset (Nwose et al., 2023). Dataset included maternal demographic variables (age, BMI, parity), lifestyle risk factors, and clinical indicators. Data cleaning involved removing empty rows, standardizing categorical fields, and recoding the outcome variable (GDM symptoms) as a binary indicator, and removing missing values cases. This further reduces the sample size.

Penalized Logistic Regression

Penalized logistic regression (PLR) extends the traditional logistic regression model by introducing a penalty term to the loss function. Like the effect of the penalty terms on ordinary least square (OLS) regression, this adjustment helps to prevent overfitting, manage high-dimensional data, and improve model generalization where binary outcomes are required.

Whereas OLS regression returns outcomes for a dependent variable that can be of any arbitrary magnitude, the basic logistic regression model predicts a binary outcome by modelling the log-odds of the dependent variable as a linear combination of the independent variables. The probability pi that an event occurs is given by

$$p_i = \frac{1}{1 + e^{-X_i \beta}}$$

where X_i represents the predictor variables, and β represents the coefficients. The model is fitted by maximizing the likelihood function

$$\mathbb{L}(\beta) = \prod_{i=1}^{N} p_i^{y_i} (1 - p_i)^{1 - y_i}$$

where y_i are the binary outcomes, i indexes each observation and N is the total number of observations. Hence, the logistic regression model calculates the probability of the event $y_i = 1$ based on the predictors X_i.

Logistic Regression with LASSO and Ridge

In Ridge logistic regression, the l_2 penalty term is imposed, and the objective is to minimise the penalized cost function:

$$-\left[\sum_{i=1}^{N} y_i \log(p_i) + (1 - y_i) \log(1 - p_i)\right] + \lambda \sum_{j=1}^{p} \beta_j^2$$

In LASSO logistic regression, the absolute values of β are penalised, and the objective is to minimise the cost function:

$$-\left[\sum_{i=1}^{N} y_i \log(p_i) + (1 - y_i) \log(1 - p_i)\right] + \lambda \sum_{j=1}^{p} \left|\beta_j\right|$$

where λ is a tuning parameter which control the strength of the penalty applied to the model coefficients. Selecting the optimal value of these parameters is crucial for balancing the bias-variance trade-off and achieving the best predictive performance. Generally speaking, as the tuning parameter increases, the penalty imposed increases and hence both ridge and lasso estimate coefficients decrease (Yan et al., 2022).

Due to the real life GDM dataset situations of large number of independent variables, multicollinearity among independent variables and small sample size, the interest is on lasso penalized logistic regression, since the ridge regression still retains almost all of the independent variables, while reducing their effects. The feature selection approach of lasso regression makes it desirable in this particular data case. However, comparison of both regression for the dataset will be made using Akaike information criterion (AIC).

3. Results

The following section presents the results of the application of penalized logistic regression to the GDM dataset, consisting of GDM symptom response variable and 17 independent variables, with correlation between some of the variables, inducing multicollinearity in the dataset. The independent variables included age of women, gravida, Gestational Age at the time of Registration (weeks), Height (cm), Weight (kg), BMI, SBP, DBP, sedentary lifestyle, Family History of Diabetes, History of T2DM/GDM, Miscarriage, fetal/neonatal death, Polycystic ovary and Macrosomia.

Ridge penalized logistic regression model

First ridge regression is fitted to the dataset and the shrinkage of the coefficients is visualized as lambda increases. This process is represented in Figure 1. As seen from figure 1, the whole seventeen independent variables are in the likelihood and their coefficients shrank to zero around the value of – 4 of – log of lambda.

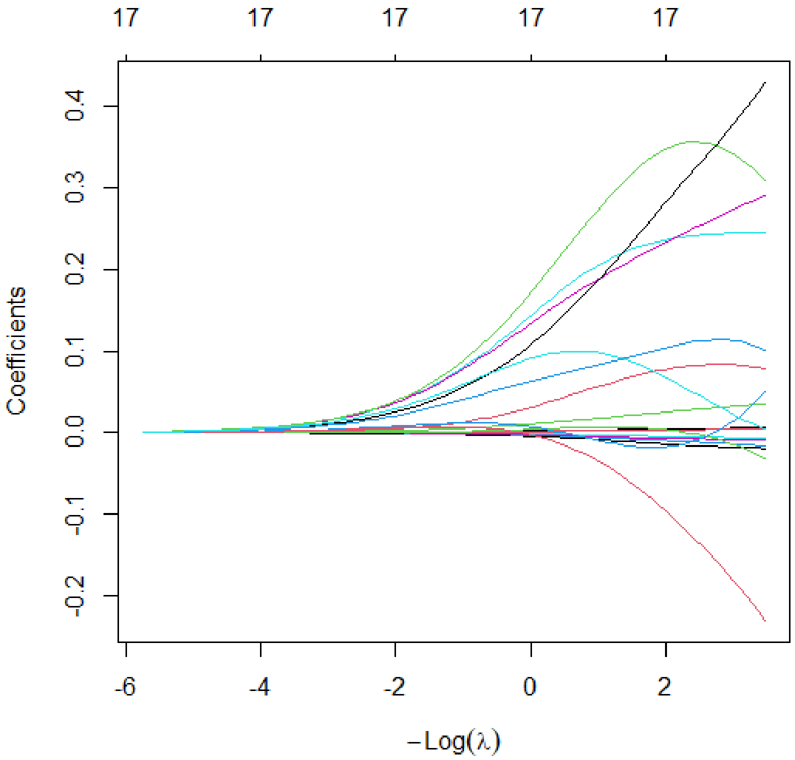

Figure 1. Plot of coefficients shrinkage as lambda increases with Ridge

Looking at the cross-validation plot in figure 2, which displays the cross-validation error according to the – log of lambda. The dashed vertical line indicates that the – log of the optimal value of lambda is approximately 1, which is the value that minimizes the prediction error. This lambda value will give the best penalty parameter and the most accurate model. The exact value of lambda was obtained as 2.707955.

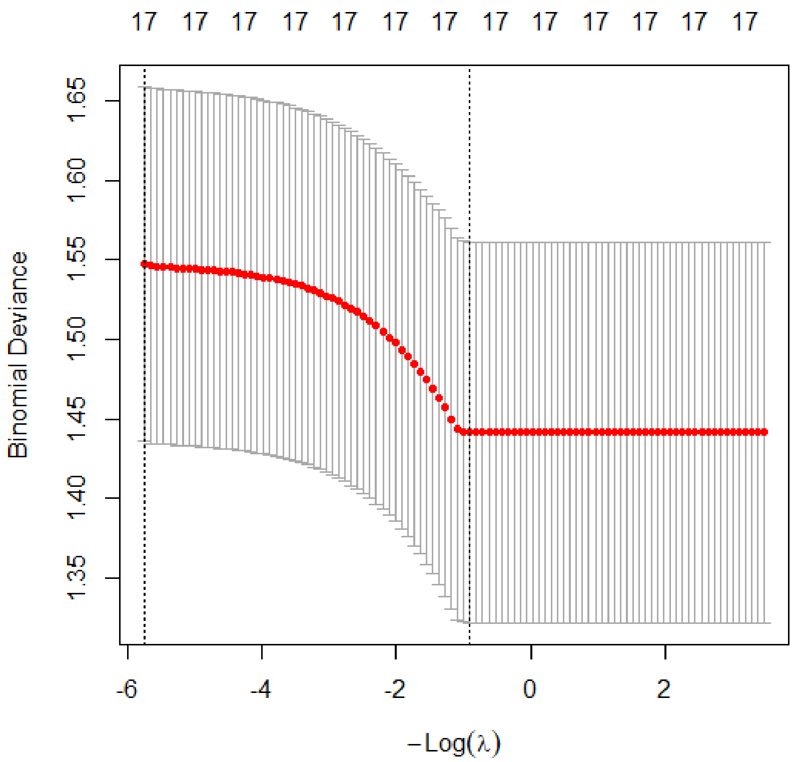

Figure 2. Cross-validation plot to find the best penalty parameter with Ridge

The final ridge model with the best lambda value of 2.707955 is fitted to the GDM dataset by refitting the final model without penalization, using only the selected predictors from ridge regression, which of course was all independent variables. The coefficients estimates are given in Table 1 and the reported AIC value for the model was 34.

Table1: Estimate from ridge penalized logistic regression

Coefficients:	Estimate	Std. Error	z value	p-value
(Intercept)	-3443.00	47550000.00	0	1
age	1.69	48350.00	0	1
gravida	1.13	110600.00	0	1
gestation week	2.45	36730.00	0	1
height	21.57	296200.00	0	1
weight	-29.04	354500.00	0	1
bmi	68.99	861800.00	0	1
sbp	-0.72	18440.00	0	1
dbp	1.26	29670.00	0	1
PH	-0.61	225500.00	0	1
sedentary lifestyle	-18.12	880200.00	0	1
family history of DB	10.98	568700.00	0	1
history of T2DB/GDM	-2.58	1517000.00	0	1

Miscarriage	110.50	953800.00	0	1
fetal death	-74.31	811200.00	0	1
polycystic ovary	-15.68	434700.00	0	1
fetal growth	68.78	1553000.00	0	1

Lasso penalized logistic regression model

Shifting attention to the lasso regression of interest, it can be observed that lower number of independent variables had their shrinkage close to zero unlike the ridge regression and from figure 3, this happened around 1.5 of – log of lambda.

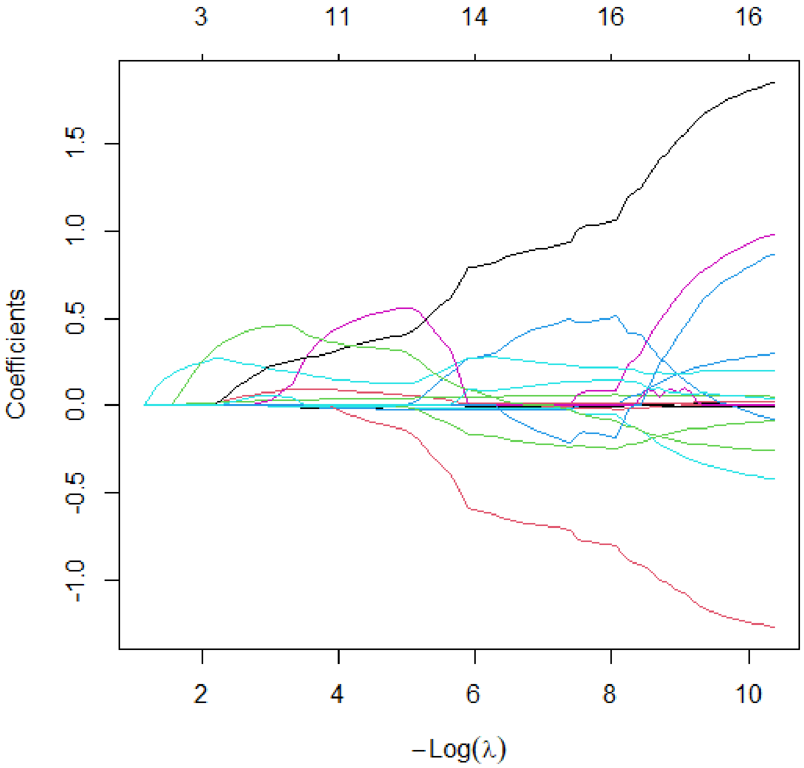

Figure 3. Plot of coefficients shrinkage as lambda increases with Lasso

A view of the cross-validation plot in figure 4, which displays the cross-validation error according to the – log of lambda. The dashed vertical line indicates that the – log of the optimal value of lambda is approximately 1.5, which is the value that minimizes the prediction error. This lambda value will give the best penalty parameter and the most accurate model. The exact value of lambda was obtained as 0.1347756.

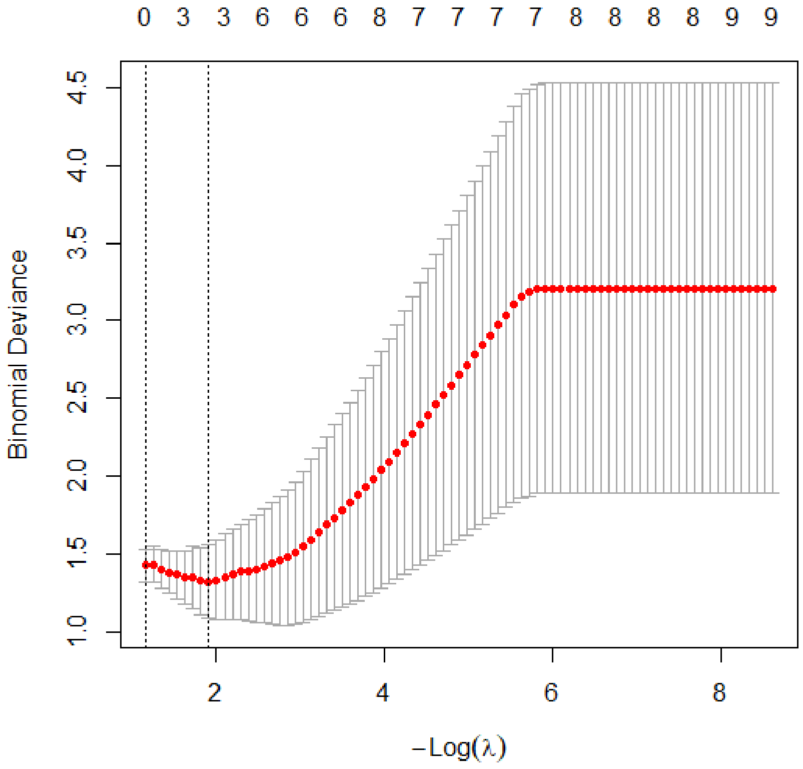

Figure 4. Cross-validation plot to find the best penalty parameter with Lasso

Fitting final model without penalization and using only the selected independent variables from lasso penalized logistic regression and with the best lambda value of 0.1347756, results in the estimates shown in table 2. The AIC value of 12.162 was reported for the final model. The important predictors for GDM from the pool of correlated independent variables were seen from the model to be gestation week, family history of T2DB and polycystic ovary. However, the dataset does not give enough statistical evidence from the lasso penalized logistic regression model that these independent variables do in fact influence the incidence of GDM, since their p-values were greater than 5% level of significance. This may be attributed to the small sample size of 17, which comprised complete cases, after data cleaning had been carried out.

Table2: Estimate from lasso penalized logistic regression

Coefficients	Estimate	Std. Error	z value	p-value
(Intercept)	-22.292	16.4596	-1.354	0.176
gestation week	0.899	0.6868	1.309	0.191
family history of DB	5.1137	5.6247	0.909	0.363
polycystic ovary	26.4179	6340.65	0.004	0.997

4. Discussion

Ridge and lasso penalized logistic regression were applied to the GDM dataset of Nwose et al. (2023). Comparing the model adequacy using AIC from both models, showed that ridge had a value of 34, while lasso regression had a value of 12.162, effectively preferred for the dataset over ridge model. The most important predictors of GDM from lasso regression were gestation week, family history of T2DB and polycystic ovary, among the seventeen predictors in the dataset. The dataset did not provide statistical evidence of effects of these predictors on GDM and this may be attributed to the small sample size of 17, comprising complete cases. From the study is was seen that the issues of overfitting and multicollinearity with large predictors can be overcome by using lasso penalized logistic regression. The model suggested that familial metabolic predisposition remains a critical risk factor in this population. This supports implementing early GDM screening among women with such histories. Future studies resulting in datasets, should include larger samples and clinical OGTT-confirmed GDM diagnoses.

5. Conclusion

While classical logistic regression are usually used for modelling incidence of GDM due to it binary outcome of presence or absence of GDM, the challenges of large independent variables or predictors of GDM, small sample size and correlation between the predictors makes the classical logistic regression unsuitable, and results in overfitting. Penalized logistic regression comes in handy for such datasets with its regularization process, giving room for shrinkage of coefficients but also setting some of them to exactly zero. That means it’s perfect for feature selection — retaining only important predictors. Lasso penalized logistic regression does exactly that and was applied to the GDM datasets used in this study that met the real life situation of large and correlated predictors and small sample size. The dataset set was from a study in south-eastern Nigeria and is available publicly. The lasso penalized regression reported gestation week, family history of T2DB and polycystic ovary to be the most important predictors of GDM. Hence, policy maker and health practitioners can focus scarce resources on education on T2DB and care of the polycystic ovary for pregnant women. Early knowledge of family history is important before pregnancy to enable pregnant women take care of themselves and receive appropriate treatment before, during and after pregnancy. Also, public health programs should incorporate targeted GDM screening and preventive counseling for at-risk women.

References

Adeoye I, Adedapo KS, Sonuga OO, Fagbamigbe, A. F. Adeleye, J. O. Olayemi, O. O., Omigbodun, A. O. and Bamgboye, A. E. (2025). Incidence, risk factors and pregnancy outcomes of gestational diabetes mellitus in Ibadan, Southwest Nigeria: a prospective cohort study. BMJ Open 2025;15:e095252. doi:10.1136/bmjopen-2024-095252

Azeez TA, Abo-Briggs T, Adeyanju AS. (2021). A systematic review and meta-analysis of the prevalence and determinants of gestational diabetes mellitus in Nigeria. Indian J Endocr Metab 2021; 25:182-90.

John D. H., Awoyesuku P. A., MacPepple D. A. and Kwosah N. J. (2019). Prevalence of Gestational Diabetes Mellitus and Maternal and Fetal Outcomes at the Rivers State University Teaching Hospital (RSUTH), Port Harcourt, Nigeria. JAMMR, 31(9): 1-16, 2019. DOI:10.9734/JAMMR/2019/v31i930319

Ladan AA, Ibrahim UA. (2023). Knowledge of Gestational Diabetes Mellitus among Pregnant Women Attending Antenatal Care in North West, Nigeria. NJGP. 2023;21(2):68 – 77

Nwose, E. U., Gbeinbo, F. D. and Bwititi, P. T. (2023). GDM register with risk factors for screening selection criteria pilot dataset. Mendeley Data, V1, doi:10.17632/r8s3j8hfdb.1

Odoh GU, Onwuka CI, Ugwu EO, Iloghalu EI, Nnagbo JE, Onwuka CI, Duru VC, Ifezuoke TD and Udealor PC. (2025). Prevalence and predictors of gestational diabetes mellitus in enugu, nigeria using the new international federation of gynecology and obstetrics (FIGO) diagnostic criteria. Niger J Clin Pract 2025; 28:648-53.

Omo-Aghoja LO, Onohwakpor EA, Adeyinka AT, Asaboro N. and Oyeye L. (2023). Prevalence of gestational diabetes mellitus, fetal and maternal outcomes of parturients with risk factors versus parturients without risk factors for gestational diabetes mellitus: A preliminary analysis of the comparative study of blood sugar levels at a tertiary hospital in southern Nigeria. African Journal of Tropical Medicine and Biomedical Research Vol. 6 No. 1 July 2023

Onyenekwe, M. B., Young, E. E., Nwatu, C. B., Okafor, C. I., Ugwueze, C. V. and Chukwu, S. N. (2019). Prevalence of Gestational Diabetes in South East Nigeria Using the Updated Diagnostic Guidelines. Int J Diabetes Metab 2019; 25:26–32. DOI: 10.1159/000500089

Yan, Y., Zhizhou Yang, Z., Semenkovich, T. R., Kozower B. D., Meyers, B.F., Nava, R. G., Kreisel, D. and Puri, V. (2022). Comparison of standard and penalized logistic regression in risk model development. JTCVS Open, vol. 9, Pp 303-316.ISSN 2666-2736. https://doi.org/10.1016/j.xjon.2022.01.016