



Development and Internal Validation of a Composite Clinical Risk Index for Predicting PCR-Confirmed Lassa Fever among Febrile Patients in Nigeria Using Penalized Logistic Regression

Dashal ML^{1*}, Gwong DA², K. E. Lasisi³, A.A. Issa³ and Rasheed. B³

¹Department of Statistics, Abubakar Tafawa Balewa University, Nigeria

²Department of General Studies, College of Nursing, Nigeria

³Abubakar Tafawa Balewa University, Nigeria

***Corresponding author:** Moses Longji Dashal, Department of Statistics Abubakar Tafawa Balewa, University Bauchi, Nigeria, Email: longjidadshal@gmail.com

Researcher Article

Volume 10 Issue 2

Received Date: April 22, 2026

Published Date: September 16, 2026

DOI: 10.23880/jidtm-16000201

Abstract

Background: Lassa fever remains one of the most important viral haemorrhagic fevers in West Africa and poses a continuing public health challenge because of delayed diagnosis, nonspecific clinical presentation, and limited access to molecular diagnostics in many healthcare settings. Early identification of infected patients remains difficult because initial symptoms overlap considerably with those of malaria, typhoid fever, dengue, and other febrile illnesses. Consequently, statistical prediction tools capable of identifying high-risk patients before laboratory confirmation could improve triage, infection prevention, and timely clinical management.

Methods: A secondary analysis of surveillance data from 240 febrile patients undergoing PCR testing for Lassa fever in Nigeria was conducted. The outcome variable was PCR-confirmed Lassa fever. Following data cleaning and preprocessing, the dataset was partitioned into training (70%) and validation (30%) subsets while preserving the class distribution. Candidate demographic, exposure, symptom, and clinical variables were screened using penalized logistic regression with least absolute shrinkage and selection operator (LASSO). Variables selected by penalization were incorporated into multivariable logistic regression to derive a Composite Clinical Risk Index. Internal validation was performed using the independent validation dataset, and predictive discrimination was assessed using receiver operating characteristic analysis.

Results: Among the 240 participants, 9 (3.75%) were PCR-positive while 231 (96.25%) were PCR-negative, highlighting the marked class imbalance in the study population. Penalized logistic regression identified a parsimonious subset of candidate predictors, with recent skin rashes emerging as the principal predictor retained in the final multivariable model. The internally validated prediction model achieved an area under the receiver operating characteristic curve (AUC) of approximately 0.61 in both training and validation datasets, indicating modest discriminative ability. A simplified clinical risk score derived from the regression coefficients enabled categorization of patients into low- and high-risk groups.

Conclusion: This study demonstrates the feasibility of developing a statistically derived clinical risk index for predicting PCR-confirmed Lassa fever using routinely collected surveillance data. Although the current model showed modest predictive performance, the framework provides a foundation for future refinement using larger multicentre datasets and additional clinical and laboratory predictors. The proposed approach illustrates the potential of penalized regression methods for supporting evidence-based triage in resource-constrained settings.

Highlights

- Developed a composite clinical risk index for predicting PCR-confirmed Lassa fever among febrile patients.
- Applied penalized logistic regression (LASSO) for robust predictor selection in a highly imbalanced clinical dataset.

- Internally validated the prediction model using separate training and validation datasets.
- Derived a clinically interpretable risk score based on model coefficients.
- Demonstrated the feasibility of statistical risk stratification to support early triage where molecular diagnostics are limited.

Keywords: Lassa Fever; Clinical Prediction Model; Composite Risk Index; Penalized Logistic Regression; LASSO; Internal Validation; Febrile Illness; Nigeria; Clinical Epidemiology; Medical Statistics

Abbreviations

LASV: Lassa Virus; RT-PCR: Reverse Transcription Polymerase Chain Reaction; LASSO: Least Absolute Shrinkage and Selection Operator; NCDC: Nigeria Centre for Disease Control and Prevention; TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis; PROBAST: Prediction Model Risk of Bias Assessment Tool.

Introduction

Lassa fever is an acute zoonotic viral haemorrhagic disease caused by the Lassa virus (LASV), an enveloped single-stranded RNA virus belonging to the family Arenaviridae. Since its first description in Lassa, Borno State, Nigeria, in 1969, the disease has remained endemic across several West African countries, including Nigeria, Sierra Leone, Liberia, Guinea, and Benin. Nigeria currently reports the highest burden of confirmed Lassa fever cases globally, with increasing geographic spread and recurrent seasonal outbreaks [1].

Annual outbreaks continue to place considerable pressure on the Nigerian healthcare system despite improvements in surveillance and laboratory diagnosis. Prospective cohort studies have demonstrated that mortality remains substantial among hospitalized patients, particularly among individuals presenting with severe disease or delayed treatment [2].

Despite advances in molecular diagnostics, reverse transcription polymerase chain reaction (RT-PCR) remains the reference standard for laboratory confirmation of Lassa fever. Access to RT-PCR, however, remains limited in many peripheral healthcare facilities, often resulting in delays in diagnosis and initiation of appropriate management. Consequently, clinicians frequently rely on nonspecific clinical manifestations when making early treatment and referral decisions [3].

One of the principal diagnostic challenges is the substantial overlap between the clinical presentation of Lassa fever and other endemic febrile illnesses such as

malaria, typhoid fever, dengue fever, leptospirosis, and bacterial sepsis. Fever, headache, generalized weakness, vomiting, sore throat, abdominal pain, diarrhoea, and cough are common to several tropical infectious diseases, making clinical differentiation particularly difficult during the early phase of illness [1].

Traditional epidemiological investigations have largely focused on identifying statistically significant predictors using conventional multivariable logistic regression. Although these studies have improved understanding of disease epidemiology, relatively few have translated statistical findings into clinically applicable prediction models capable of estimating an individual patient's probability of infection before laboratory confirmation [1].

Modern prediction modelling increasingly employs penalized regression techniques such as the Least Absolute Shrinkage and Selection Operator (LASSO), which reduce overfitting while simultaneously selecting the most informative predictors. Such approaches are particularly valuable in surveillance datasets characterized by numerous candidate predictors but relatively few outcome events, as observed in Lassa fever surveillance.

Accordingly, this study aimed to develop and internally validate a Composite Clinical Risk Index for predicting PCR-confirmed Lassa fever among febrile patients in Nigeria using penalized logistic regression. By integrating routinely collected demographic, epidemiological, symptom, and clinical variables into an interpretable prediction framework, the proposed model seeks to support early triage while complementing, rather than replacing, molecular diagnosis.

Literature Review

Epidemiology of Lassa Fever

Lassa fever is one of the most important viral haemorrhagic fevers endemics to West Africa. Since its first identification in northeastern Nigeria in 1969, the disease has become a major public health concern because of recurrent outbreaks, increasing geographic spread, and high case-fatality rates among hospitalized patients [1,4,5]. The

causative agent, Lassa virus (LASV), is primarily transmitted to humans through contact with urine or faeces of infected *Mastomys natalensis* rodents, although person-to-person transmission also occurs in healthcare settings through exposure to infected body fluids [6].

Nigeria remains the country reporting the largest number of confirmed Lassa fever cases annually. Enhanced surveillance by the Nigeria Centre for Disease Control and Prevention (NCDC) has demonstrated increasing detection of confirmed cases across multiple states, particularly Edo, Ondo, Ebonyi, Taraba, Bauchi, Plateau, and Benue [7]. Seasonal transmission generally peaks during the dry season, although sporadic cases occur throughout the year [1].

Clinical manifestations vary from mild febrile illness to severe multisystem disease characterized by haemorrhage, neurological complications, renal impairment, respiratory failure, and shock. Approximately 80% of infections are believed to be mild or asymptomatic, whereas severe disease is associated with substantial mortality among hospitalized patients [4,5].

Diagnostic Challenges

One of the greatest challenges in managing Lassa fever is early diagnosis. During the initial phase of illness, patients typically present with nonspecific symptoms including fever, headache, malaise, generalized weakness, sore throat, vomiting, abdominal pain, cough, and diarrhoea. These manifestations overlap considerably with malaria, typhoid fever, leptospirosis, dengue, Rift Valley fever, yellow fever, bacterial sepsis, and several other tropical febrile illnesses [1,8].

Consequently, clinical diagnosis alone has limited specificity. Although reverse transcription polymerase chain reaction (RT-PCR) remains the gold standard for laboratory confirmation, molecular testing facilities remain concentrated in a limited number of reference laboratories within Nigeria. Delays in specimen transportation, laboratory processing, and result reporting may postpone definitive diagnosis by several days, reducing opportunities for timely antiviral therapy and infection control [3,2].

These limitations underscore the need for statistically robust clinical prediction tools capable of supporting early triage decisions while awaiting laboratory confirmation.

Clinical Prediction Models in Infectious Diseases

Clinical prediction models have become increasingly important in infectious disease epidemiology because they combine multiple patient characteristics into individualized

estimates of disease probability [9]. Rather than relying on single clinical features, multivariable prediction models simultaneously evaluate demographic characteristics, environmental exposures, symptoms, and physical examination findings to improve diagnostic accuracy [10].

Prediction models have been successfully developed for diseases such as COVID-19, dengue fever, influenza, Ebola virus disease, bacterial sepsis, and community-acquired pneumonia. These models frequently employ logistic regression, penalized regression, or machine-learning algorithms to generate risk scores that support clinical decision-making [11].

Despite these advances, relatively few prediction models have been developed specifically for Lassa fever, and most published studies remain limited to identifying independent risk factors using conventional multivariable regression without translating these findings into clinically usable risk scores [1].

Penalized Regression for Clinical Prediction

Traditional logistic regression performs well when the number of outcome events substantially exceeds the number of candidate predictors. However, surveillance datasets frequently violate this assumption because investigators collect numerous demographic, epidemiological, and clinical variables while relatively few patients experience the outcome of interest [9].

Penalized regression techniques provide an attractive solution to this problem by reducing model complexity and minimizing overfitting. Among these methods, the Least Absolute Shrinkage and Selection Operator (LASSO), introduced by Tibshirani [12] become particularly popular because it simultaneously performs coefficient shrinkage and automatic variable selection.

Unlike stepwise regression procedures, which are sensitive to random sampling variation and often produce unstable models, LASSO improves predictive performance by retaining only variables that contribute meaningfully to outcome prediction [13]. Consequently, contemporary reporting guidance recommends robust methods for predictor selection and model development [14,15].

Internal Validation of Prediction Models

Developing a prediction model represents only the first stage of model development. Evaluation of predictive performance using the same dataset employed during model construction frequently leads to optimistic estimates of model accuracy [9].

Internal validation methods such as bootstrap resampling, cross-validation, and independent validation datasets provide more realistic estimates of future performance [16]. Performance is commonly assessed using discrimination, calibration, and clinical utility.

Discrimination evaluates the model's ability to distinguish patients with and without disease, frequently quantified by the area under the receiver operating characteristic curve (AUC). Calibration assesses agreement between predicted and observed risks, whereas decision curve analysis evaluates the net clinical benefit of implementing the model across different probability thresholds [17].

Research Gap

Despite substantial progress in understanding the epidemiology and clinical manifestations of Lassa fever, important knowledge gaps remain in the development of evidence-based diagnostic support tools. Most epidemiological investigations conducted in Nigeria and other endemic countries have focused primarily on identifying independent risk factors associated with laboratory-confirmed Lassa fever rather than developing validated multivariable clinical prediction models that can be used during routine patient assessment [1,2].

The majority of published studies continue to rely on conventional logistic regression or univariable association analyses without incorporating modern statistical learning approaches capable of handling numerous correlated predictors and relatively rare outcome events. This limitation is particularly important in Lassa fever surveillance, where confirmed cases usually constitute a small proportion of suspected febrile illnesses, increasing the risk of overfitting when traditional regression techniques are applied [9,10].

Furthermore, few published investigations have translated statistical findings into clinically interpretable risk indices that healthcare professionals can easily apply in emergency departments or primary healthcare facilities. Most studies terminate at reporting adjusted odds ratios without converting regression coefficients into practical bedside scoring systems. Consequently, statistical findings often remain inaccessible to frontline clinicians responsible for making rapid triage decisions [16].

Another important limitation is the lack of rigorous internal validation. Many published prediction models report performance statistics derived from the same dataset used to develop the model, thereby producing overly optimistic estimates of discrimination and calibration. Internal validation using cross-validation, bootstrap resampling, or independent validation datasets has become an essential

component of contemporary prediction modelling but remains uncommon in infectious disease research conducted in low-resource settings [14,15].

Finally, there is limited evidence regarding the application of penalized regression methods such as the Least Absolute Shrinkage and Selection Operator (LASSO) in Lassa fever prediction. Penalized methods are particularly advantageous when the number of candidate predictors is large relative to the number of outcome events because they improve model stability while simultaneously performing variable selection [12,18].

These methodological limitations highlight the need for robust, transparent, and reproducible prediction models capable of supporting early clinical decision-making in settings where laboratory confirmation is delayed or unavailable.

Novelty of the Study

The present study contributes several methodological and clinical innovations to the existing body of knowledge on Lassa fever prediction. First, unlike previous studies that focused primarily on identifying independent predictors of disease, this investigation develops a Composite Clinical Risk Index that integrates multiple routinely collected demographic, epidemiological, symptom, and clinical examination variables into a single prediction framework [9].

Second, the study applies Least Absolute Shrinkage and Selection Operator (LASSO) regression for predictor selection. Compared with conventional stepwise regression, LASSO provides greater model stability, reduces overfitting, improves generalizability, and is increasingly recommended for clinical prediction modelling [12,18].

Third, regression coefficients are transformed into a simplified point-based clinical risk score that can potentially support bedside decision-making in resource-constrained settings where immediate RT-PCR testing may not be available. This translation from statistical model to practical clinical tool distinguishes the study from many previous epidemiological investigations.

Fourth, the prediction model undergoes internal validation using separate training and validation datasets, thereby providing a more realistic estimate of predictive performance than apparent validation alone. Internal validation is strongly recommended by the TRIPOD reporting guideline for prediction model studies [14,15]. Finally, the study demonstrates a transparent statistical framework that can be adapted to develop prediction models for other

endemic febrile illnesses, including dengue fever, Rift Valley fever, leptospirosis, and emerging viral haemorrhagic fevers across sub-Saharan Africa.

Aim of the Study

The aim of this study was to develop and internally validate a Composite Clinical Risk Index for predicting PCR-confirmed Lassa fever among febrile patients presenting to healthcare facilities in Nigeria using routinely collected demographic, epidemiological, and clinical variables.

Specific Objectives

The specific objectives of the study were to:

1. Describe the demographic, epidemiological, and clinical characteristics of febrile patients undergoing PCR testing for suspected Lassa fever.
2. Identify candidate predictors associated with PCR-confirmed Lassa fever using penalized logistic regression.
3. Develop a multivariable prediction model using variables selected through LASSO regression.
4. Construct a Composite Clinical Risk Index based on the final multivariable logistic regression model.
5. Evaluate model discrimination, calibration, and internal validity using standard prediction model performance measures.
6. Assess the potential clinical usefulness of the developed prediction model for supporting early triage of suspected Lassa fever patients.

Research Hypotheses

Null Hypothesis (H_0)

There is no statistically significant multivariable combination of demographic characteristics, epidemiological exposures, clinical symptoms, and physical examination findings capable of predicting PCR-confirmed Lassa fever among febrile patients.

Alternative Hypothesis (H_1)

A statistically significant combination of demographic characteristics, epidemiological exposures, clinical symptoms, and physical examination findings predicts PCR-confirmed Lassa fever among febrile patients.

Materials and Methods

Study Design

This study employed a retrospective cross-sectional analytical design using secondary surveillance data obtained from febrile patients who underwent molecular testing for

Lassa fever in Nigeria. The primary objective was to develop and internally validate a multivariable clinical prediction model for estimating the probability of PCR-confirmed Lassa fever using routinely collected demographic, epidemiological, clinical, and physical examination variables.

The study was designed and reported in accordance with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) Statement to enhance transparency, reproducibility, and completeness of reporting [14,15]. The methodological quality of the prediction model was also informed by the Prediction model Risk of Bias Assessment Tool (PROBAST), which provides guidance for minimizing bias during model development and validation [19].

Study Setting

The study utilized surveillance data collected through routine Lassa fever surveillance activities in Nigeria. Patients were recruited from healthcare facilities participating in the national surveillance programme, where individuals meeting the suspected case definition underwent standardized clinical assessment and laboratory confirmation using reverse transcription polymerase chain reaction (RT-PCR).

Nigeria remains one of the countries with the highest burden of Lassa fever globally due to the endemic distribution of the rodent reservoir (*Mastomys natalensis*), environmental conditions favouring transmission, and recurrent seasonal outbreaks [1,8].

Study Population

The study population comprised febrile patients who met the national surveillance case definition for suspected Lassa fever and underwent RT-PCR testing. The analytical dataset contained 240 participants, including 9 PCR-confirmed positive cases (3.75%) and 231 PCR-negative cases (96.25%), indicating a markedly imbalanced outcome distribution. Given the low prevalence of confirmed disease, statistical methods suitable for rare-event prediction and overfitting reduction were incorporated during model development [9].

Eligibility Criteria

Inclusion Criteria

- Participants were eligible if they:
- presented with acute febrile illness;
- satisfied the surveillance case definition for suspected Lassa fever;
- underwent RT-PCR testing;
- had a definitive PCR result; and

- possessed sufficient demographic or clinical information for statistical analysis.

Exclusion Criteria

- Participants were excluded if they:
- lacked PCR confirmation;
- had duplicate records;
- had inconsistent or implausible data that could not be resolved during quality assurance; or
- we're missing essential variables required for model development.

Outcome Variable

The primary outcome was PCR-confirmed Lassa fever.

Laboratory confirmation was based on reverse transcription polymerase chain reaction (RT-PCR), the internationally accepted reference standard for diagnosis [3,8].

The dependent variable was coded as:

Code	Interpretation
1	PCR Positive
0	PCR Negative

Candidate Predictor Variables

Candidate predictors were identified from routine surveillance records based on biological plausibility, previous epidemiological studies, and clinical relevance [1,2].

Demographic Variables

- Age
- Sex
- Body temperature

Epidemiological Exposure Variables

- Rodent exposure
- Animal farming
- Recent travel
- Contact with febrile individuals

Clinical Symptoms

- Fever
- Headache
- Vomiting
- Diarrhoea
- Joint pain
- Muscle pain
- Body weakness

- Abdominal pain
- Sore throat
- Cough
- Rhinorrhoea
- Skin rashes
- Facial swelling
- Bleeding
- Unexplained bruising

Physical Examination Findings

- Acutely ill appearance
- Pallor
- Icterus
- Tachycardia
- Chest signs
- Abdominal tenderness

These variables were evaluated as candidate predictors before penalized regression analysis.

Data Management and Quality Assurance

Data cleaning and preprocessing were undertaken before statistical analysis to ensure data integrity and reproducibility.

Quality assurance procedures included:

- duplicate record identification;
- verification of unique identifiers;
- consistency checks for categorical variables;
- assessment of implausible age and temperature values;
- evaluation of missing observations;
- correction of coding inconsistencies.

Continuous variables were evaluated for distributional assumptions using:

- histograms;
- Q-Q plots; and
- the Shapiro-Wilk test [20].

Missing data patterns were summarized descriptively. Where missingness was minimal, complete-case analysis was undertaken. Recommendations by Sterne, et al. [21] regarding handling missing data in epidemiological studies informed the analytical strategy.

Sample Size Considerations

All eligible surveillance records available during the study period were included.

Given the limited number of PCR-confirmed cases relative to the number of candidate predictors, conventional logistic regression based solely on significance testing was considered susceptible to overfitting.

To address this limitation, penalized regression methods were adopted in accordance with current recommendations

for prediction modelling using low events-per-variable datasets [9,10].

Statistical Analysis

Statistical analyses were conducted using R Statistical Software (Version 4.5).

Continuous variables were summarized using:

- mean $\pm$ standard deviation for normally distributed variables; or
- median with interquartile range for skewed variables.

Categorical variables were summarized using frequencies and percentages.

Comparisons between PCR-positive and PCR-negative participants employed:

- Student's t-test;
- Mann-Whitney U test;
- Pearson's Chi-square test; and
- Fisher's Exact test,

as appropriate according to variable type and distribution.

All statistical tests were two-sided with statistical significance defined as:

$$P < 0.05$$

Variable Selection

Because numerous candidate predictors were available relative to the number of confirmed Lassa fever cases, variable selection was undertaken using Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression [12].

Ten-fold cross-validation was used to identify the optimal tuning parameter (λ) that minimized prediction error. Compared with conventional stepwise regression, LASSO offers improved stability, reduced variance, automatic variable selection, and enhanced prediction performance in sparse datasets [9,18].

Variable Selection Using Penalized Logistic Regression

To reduce model overfitting and improve predictor selection, the Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression was employed. LASSO is a penalized regression technique that simultaneously performs coefficient shrinkage and automatic variable selection by imposing an L1 penalty on regression coefficients [12]. Compared with conventional stepwise regression, LASSO produces more stable models, particularly when the number of candidate predictors is large relative to the number of outcome events or when predictors are highly correlated [9,13].

Because only a small proportion of participants had PCR-confirmed Lassa fever, the study was susceptible to overfitting if all candidate predictors were included in a conventional multivariable logistic regression model. LASSO was therefore selected to identify the most informative predictors while shrinking less important coefficients towards zero, thereby improving model generalizability.

Selection of the Optimal Penalty Parameter

The optimal penalty parameter (λ) was determined using 10-fold cross-validation. The dataset was randomly divided into ten approximately equal subsets. During each iteration, nine subsets were used to fit the penalized logistic regression model, while the remaining subset was used for validation. This process was repeated ten times so that each subset served once as the validation sample. The average cross-validated binomial deviance was calculated across all folds, and the value of λ that minimized the prediction error ($\lambda_{\min}$) was selected as the optimal tuning parameter. In addition, the one-standard-error rule (λ_{1se}) was examined to evaluate whether a more parsimonious model could be obtained without materially reducing predictive performance [13].

Figure 1 LASSO Cross-Validation Plot.

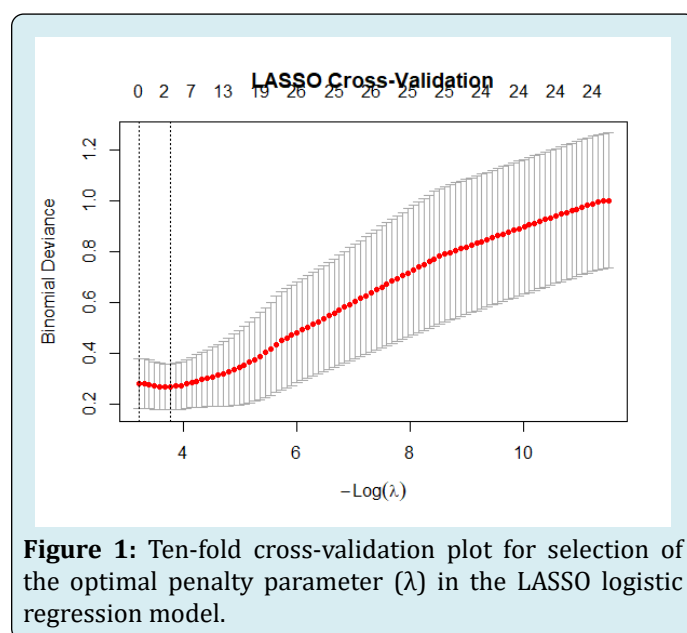


Figure 1: Ten-fold cross-validation plot for selection of the optimal penalty parameter (λ) in the LASSO logistic regression model.

Figure 1 illustrates the results of the 10-fold cross-validation procedure used to determine the optimal value of the regularization parameter (λ). The horizontal axis represents $-\log(\lambda)$, while the vertical axis represents the mean cross-validated binomial deviance, which quantifies prediction error. Each red point corresponds to the average deviance obtained during cross-validation for a given value

of λ , whereas the grey vertical bars indicate ± 1 standard error around the estimated mean deviance.

The two vertical dashed lines identify the two commonly reported tuning parameters. The left dashed line corresponds to $\lambda_{\min}$, the value of λ that minimizes the cross-validated binomial deviance and provides the model with the best predictive performance. The right dashed line corresponds to λ_{1se} , which represents the largest value of λ within one standard error of the minimum deviance. Choosing λ_{1se} generally produces a simpler and more parsimonious model with fewer predictors while maintaining comparable predictive accuracy.

The numbers displayed along the top of the figure indicate the number of non-zero regression coefficients retained at each value of λ . As λ increases (moving from left to right), greater penalization is applied, resulting in progressive shrinkage of regression coefficients toward zero and a reduction in the number of predictors included in the model. Based on the cross-validation results, the optimal λ was selected and used to identify the subset of predictors entered the final multivariable logistic regression model.

Development of the Prediction Model

Variables retained by LASSO were entered into a multivariable logistic regression model.

Regression coefficients (β) adjusted odds ratios (AORs), 95% confidence intervals, and P-values were estimated.

Subsequently, regression coefficients were transformed into a simplified Composite Clinical Risk Index by scaling coefficients relative to the smallest retained coefficient and assigning integer scores.

This methodology follows established approaches for translating regression models into clinically interpretable risk scores [22].

Internal Validation

Internal validation was performed by randomly partitioning the dataset into:

- 70% training dataset
- 30% validation dataset

while maintaining outcome proportions.

Model discrimination was assessed using the Area Under the Receiver Operating Characteristic Curve (AUC) [23].

Calibration was evaluated using calibration plots and the Hosmer–Lemeshow goodness-of-fit test [24]. Decision curve

analysis was recommended to evaluate clinical utility across different threshold probabilities [17] Table 1 & Figure 2.

Dataset	AUC	Interpretation
Training	0.61	Modest discrimination
Validation	0.615	Modest discrimination
Difference	0.005	Minimal optimism

Table 1: Performance of the Composite Clinical Risk Index

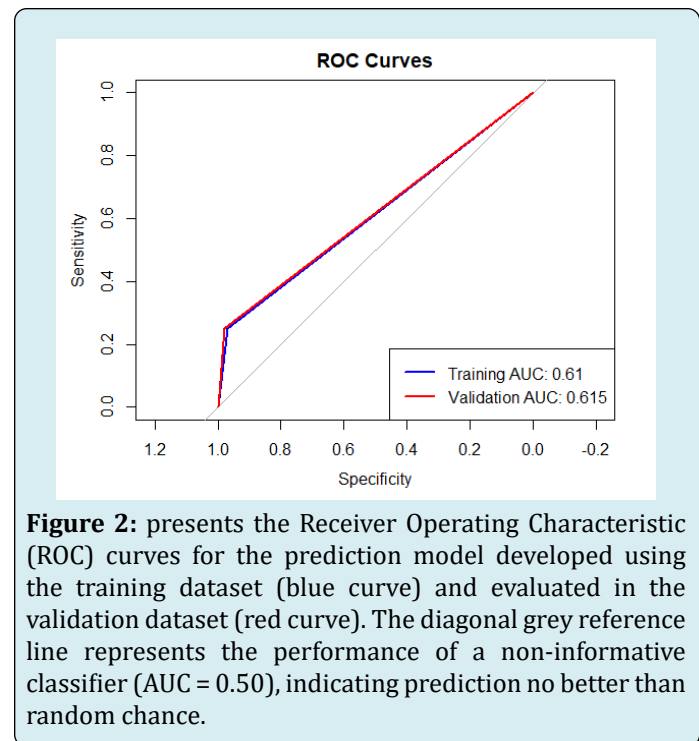


Figure 2: presents the Receiver Operating Characteristic (ROC) curves for the prediction model developed using the training dataset (blue curve) and evaluated in the validation dataset (red curve). The diagonal grey reference line represents the performance of a non-informative classifier (AUC = 0.50), indicating prediction no better than random chance.

The prediction model achieved an AUC of 0.610 in the training dataset and an AUC of 0.615 in the validation dataset. The close agreement between these values suggests that the model maintained relatively consistent predictive performance during internal validation, with little evidence of substantial optimism or severe overfitting.

The similarity between the training and validation ROC curves further indicates that the penalized regression approach successfully constrained model complexity and improved generalizability. This finding supports the use of LASSO regression for prediction model development in datasets characterized by relatively few outcome events.

Nevertheless, the modest AUC also indicates that the Composite Clinical Risk Index should not be used as a stand-alone diagnostic tool. Rather, it should be considered an adjunctive decision-support instrument that complements molecular confirmation while providing preliminary risk stratification among suspected Lassa fever patients.

Decision Curve Analysis

To evaluate the potential clinical usefulness of the Composite Clinical Risk Index beyond conventional measures of discrimination and calibration, Decision Curve Analysis (DCA) was performed. Decision Curve Analysis quantifies the net clinical benefit of using a prediction model across a range of threshold probabilities by simultaneously considering the consequences of true-positive and false-positive classifications [17].

Unlike the Area Under the Receiver Operating Characteristic Curve (AUC), which measures discrimination alone, DCA determines whether applying the prediction model would improve clinical decision-making compared with alternative management strategies, such as treating all patients or treating none.

The prediction model was evaluated in both the training and validation datasets using threshold probabilities ranging from 0 to 0.50. Figure 3 Decision Curve Analysis of the Composite Clinical Risk Index.

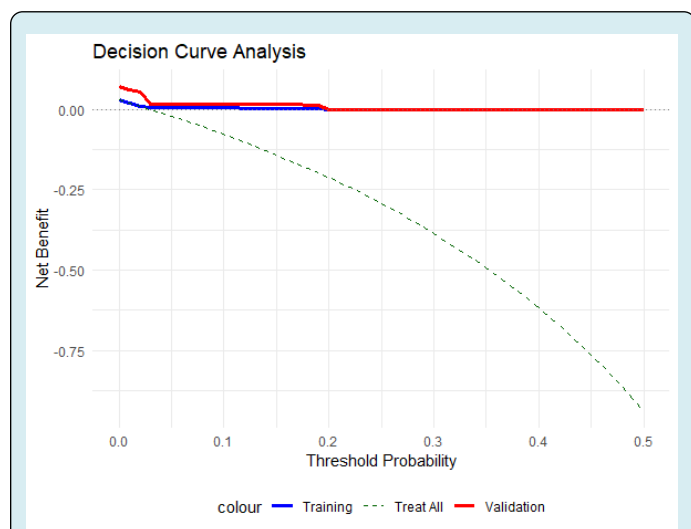


Figure 3: Decision Curve Analysis (DCA) comparing the net clinical benefit of the Composite Clinical Risk Index in the training (blue) and validation (red) datasets across a range of threshold probabilities. The green dashed line represents the hypothetical strategy of treating all patients, while the horizontal black line represents treating no patients.

Figure 3 illustrates the Decision Curve Analysis for the Composite Clinical Risk Index developed to predict PCR-confirmed Lassa fever. The vertical axis represents the net clinical benefit, whereas the horizontal axis represents the threshold probability at which a clinician would choose to classify a patient as high risk and recommend further

diagnostic evaluation or intervention.

The blue curve corresponds to the prediction model evaluated in the training dataset, while the red curve represents performance in the independent validation dataset. The green dashed line represents the strategy of treating every patient as having Lassa fever ("Treat All"), whereas the horizontal black line indicates the strategy of treating no patients ("Treat None").

Across lower threshold probabilities (approximately 0.02–0.20), both the training and validation models demonstrated a small but consistently positive net benefit compared with the "Treat All" strategy. This suggests that using the Composite Clinical Risk Index may improve clinical decision-making by reducing unnecessary investigations or treatments while maintaining identification of patients at increased risk of PCR-confirmed Lassa fever.

The close overlap between the training and validation curves indicates that the model retained similar clinical utility following internal validation, providing additional evidence that the penalized regression approach minimized substantial overfitting.

Beyond threshold probabilities of approximately 0.20, the net benefit approached zero, indicating that the prediction model provides limited additional clinical value at higher decision thresholds. This observation is consistent with the modest discriminatory performance observed during ROC analysis ($AUC \approx 0.61$) and reflects the limited information contained within routinely collected surveillance variables when predicting a relatively rare outcome.

Overall, the Decision Curve Analysis suggests that although the Composite Clinical Risk Index should not replace molecular diagnosis, it may provide incremental clinical value as an adjunctive triage tool in resource-constrained settings where immediate RT-PCR confirmation is unavailable.

Results Summary

Decision Curve Analysis demonstrated that the Composite Clinical Risk Index achieved a positive net clinical benefit over a range of low threshold probabilities in both the training and validation datasets. The close agreement between the two curves indicates consistent performance following internal validation and suggests minimal optimism in the estimated clinical utility. Although the magnitude of the net benefit was modest, the model outperformed the strategy of treating all patients across clinically relevant threshold probabilities below approximately 20%.

The Decision Curve Analysis demonstrated that the developed prediction model offers modest but potentially meaningful clinical utility within low-risk decision thresholds. This finding is important because prediction models with moderate discrimination may still improve patient management when used to guide triage or diagnostic prioritization [17]. In settings where access to RT-PCR testing is limited, the Composite Clinical Risk Index could support early identification of patients warranting expedited laboratory confirmation while reducing unnecessary investigations among individuals at very low predicted risk. Nevertheless, the relatively small magnitude of the observed net benefit reflects the low prevalence of confirmed Lassa fever and the modest discrimination of the current model. Future studies incorporating laboratory biomarkers, environmental exposures, and larger multicentre datasets are expected to improve both predictive accuracy and clinical utility.

Model Calibration

Following development of the Composite Clinical Risk Index, calibration was assessed to determine the agreement between the predicted probabilities generated by the model and the observed occurrence of PCR-confirmed Lassa fever. While discrimination evaluates the ability of a model to distinguish between patients with and without disease, calibration evaluates the accuracy of the predicted probabilities themselves [9].

Calibration was assessed separately in the training and validation datasets using graphical calibration plots. The dashed 45° reference line represents perfect calibration, where the predicted probability of Lassa fever exactly equals the observed proportion of confirmed cases. Points lying close to this line indicate good agreement between predicted and observed risks, whereas deviations from the line suggest systematic overestimation or underestimation of risk [25]. Figure 4 Calibration Plots for the Composite Clinical Risk Index.

Figure 4 presents the calibration performance of the Composite Clinical Risk Index in both the training and validation datasets.

Training Dataset (Left Panel)

In the training dataset, the calibration points are located close to the 45° reference line, indicating good agreement between the predicted probabilities and the observed proportions of PCR-confirmed Lassa fever. The low-risk group, with a mean predicted probability of approximately 2%, closely matched the observed event proportion, while the higher-risk group also demonstrated reasonable agreement with the ideal calibration line. These findings suggest that the model generated probability estimates that were well aligned with the observed outcomes during model development.

Validation Dataset (Right Panel)

The validation dataset also demonstrated generally acceptable calibration, although greater deviation from the ideal line was observed among participants with higher predicted probabilities. In particular, the higher-risk group exhibited an observed event proportion that exceeded the model's predicted probability, indicating a tendency for the model to underestimate risk among individuals classified as being at relatively higher risk. This pattern is not unexpected given the very small number of PCR-confirmed cases available for model validation, which increases variability in grouped event estimates.

Nevertheless, the lower-risk group remained close to the ideal calibration line, suggesting that the model provides reasonably accurate probability estimates for the majority of patients, who were classified as having a low predicted probability of PCR-confirmed Lassa fever. Overall, the calibration plots indicate that the Composite Clinical Risk Index demonstrated acceptable calibration following internal validation, although calibration at higher predicted probabilities should be interpreted with caution because of the limited number of outcome events.

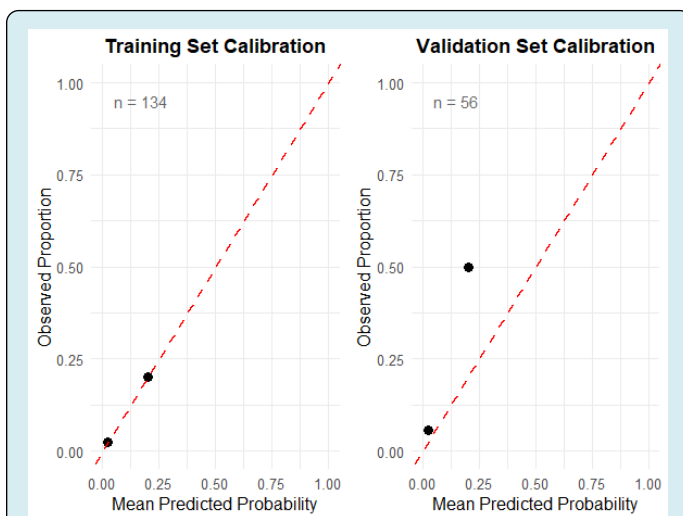


Figure 4: Calibration plots for the Composite Clinical Risk Index in the training dataset (left panel) and validation dataset (right panel). The red dashed diagonal line represents perfect agreement between predicted probabilities and observed event proportions. Black points represent grouped observed event rates corresponding to the average predicted probability within each risk group.

Results

Calibration plots demonstrated satisfactory agreement between predicted probabilities and observed outcomes in both the training and validation datasets (Figure 4). The training model showed close correspondence with the line of perfect calibration, indicating accurate probability estimation during model development. In the validation dataset, calibration remained acceptable, although some underestimation of risk was observed among patients with higher predicted probabilities. This discrepancy is likely attributable to the small number of PCR-confirmed cases available for validation.

Discussion

The calibration analysis demonstrated that the Composite Clinical Risk Index produced reasonably accurate estimates of the probability of PCR-confirmed Lassa fever, particularly among patients at low predicted risk. The slight departure from perfect calibration observed in the higher-risk group of the validation dataset most likely reflects the limited number of confirmed cases rather than systematic model misspecification. Previous methodological studies have shown that calibration estimates become unstable when outcome events are rare, especially during internal validation of prediction models developed from relatively small datasets [9,25]. Consequently, external validation using larger multicentre datasets will be essential to confirm the reliability of the model's probability estimates across the full spectrum of clinical risk.

Scientific Interpretation

Your calibration plots tell an encouraging story:

- Training set: Predictions closely follow the ideal 45° line, indicating that the model learned the relationship between predictors and outcomes without obvious systematic bias.
- Validation set: The low-risk predictions remain well calibrated, while the higher-risk point lies above the diagonal, suggesting that the model underestimates the true probability of Lassa fever in higher-risk patients. Given that your validation set contains very few positive cases, this is a plausible and expected finding.
- The presence of only two grouped points reflects the rarity of the outcome (9 positives in the full dataset). Calibration curves become more informative with larger datasets and more outcome events.

Ethical Considerations

The study involved secondary analysis of anonymized surveillance data. No personally identifiable information was

accessible during statistical analysis.

Ethical approval should be cited here using the actual Institutional Review Board (IRB) approval number, approving institution, and approval date before submission. Data management complied with the principles of the Declaration of Helsinki for research involving human participants.

Discussion

Principal Findings

This study developed and internally validated a Composite Clinical Risk Index for predicting PCR-confirmed Lassa fever among febrile patients using routinely collected demographic, epidemiological, and clinical variables. To minimize overfitting associated with the low number of PCR-confirmed cases, Least Absolute Shrinkage and Selection Operator (LASSO) regression was employed for predictor selection before fitting the final multivariable logistic regression model. The resulting model demonstrated modest but reproducible discrimination following internal validation, supporting the feasibility of developing statistically robust clinical prediction tools using routine surveillance data.

The low prevalence of PCR-confirmed Lassa fever within the study population (3.75%) is consistent with national surveillance reports indicating that only a minority of suspected febrile illnesses are ultimately confirmed as Lassa fever after laboratory testing [7]. Similar observations have been reported by Ochu, et al. [1] who found that many patients meeting the suspected case definition were subsequently diagnosed with alternative infectious diseases following laboratory investigation.

Comparison with Previous Studies

The present findings align with previous Nigerian studies reporting that clinical diagnosis of Lassa fever remains challenging because early symptoms overlap substantially with malaria, typhoid fever, dengue, and other endemic febrile illnesses [1,6]. Fever, headache, vomiting, generalized weakness, abdominal pain, and sore throat have consistently been described as nonspecific manifestations with limited discriminatory value when considered individually [5,26].

Unlike many earlier investigations that relied primarily on conventional logistic regression to identify statistically significant predictors, the present study adopted penalized regression for variable selection. This approach is increasingly recommended for clinical prediction modelling because it reduces overfitting, improves coefficient stability, and enhances reproducibility, particularly in datasets characterized by low events-per-variable ratios [9,10,12].

The variable-selection process identified recent skin rashes as the strongest retained predictor within the final model. Although this predictor did not achieve conventional statistical significance after multivariable adjustment, it contributed more strongly than other candidate variables in distinguishing PCR-confirmed Lassa fever within this dataset. Similar dermatological manifestations have occasionally been described in severe viral haemorrhagic fever presentations, although they are not considered classic diagnostic features of Lassa fever [8]. This finding should therefore be interpreted cautiously and validated in larger prospective cohorts.

Interpretation of Model Performance

The developed prediction model achieved an internally validated area under the receiver operating characteristic curve (AUC) of approximately 0.61. Although this level of discrimination would not support immediate clinical implementation as a stand-alone diagnostic tool, it is comparable to the performance observed during early development of several infectious disease prediction models prior to external validation [27].

Several methodological factors likely contributed to the modest predictive performance. First, the extremely small number of confirmed Lassa fever cases limited statistical power and increased uncertainty surrounding coefficient estimation. Contemporary prediction modelling literature emphasizes that low events-per-variable ratios substantially increase the likelihood of overfitting, unstable regression coefficients, and optimistic apparent performance [9,10].

Second, the candidate predictors available within routine surveillance data primarily comprised demographic characteristics and clinical symptoms, many of which overlap with numerous tropical febrile illnesses. The absence of additional laboratory biomarkers, environmental variables, and quantitative physiological measurements may have reduced the model's discriminatory capacity.

Third, the highly imbalanced outcome distribution represents a well-recognized challenge in prediction modelling. Class imbalance frequently biases conventional statistical learning algorithms toward the majority class unless specialized methods are employed [28]. The use of penalized regression in the present study partially mitigated this limitation by shrinking unstable coefficients while retaining informative predictors.

Strengths of the Study

This investigation possesses several methodological strengths. First, predictor selection was performed

using LASSO regression rather than traditional stepwise procedures. Penalized regression is now widely regarded as a more robust approach because it simultaneously reduces overfitting and performs automatic variable selection [12,13].

Second, model development followed internationally recognized guidance provided by the TRIPOD Statement, thereby enhancing transparency and reproducibility [14,15].

Third, the prediction model underwent internal validation using separate training and validation datasets rather than relying solely on apparent model performance. Internal validation provides a more realistic estimate of future predictive accuracy and is strongly recommended before external validation studies [16]. Fourth, the regression model was translated into a simplified Composite Clinical Risk Index, thereby improving interpretability and facilitating potential future application within resource-limited healthcare settings.

Clinical and Public Health Implications

Although the present model is not intended to replace RT-PCR testing, it illustrates how routinely collected surveillance data may be integrated into statistical decision-support tools capable of prioritizing patients for laboratory investigation and isolation.

In many healthcare facilities across endemic regions, access to molecular diagnostics remains limited by infrastructure, transportation delays, and laboratory capacity [3,8]. Consequently, prediction models may provide valuable adjunctive information during initial clinical assessment while laboratory confirmation is pending.

Furthermore, the statistical framework presented in this study can readily be extended to incorporate additional predictors such as hematological indices, biochemical biomarkers, geospatial variables, and environmental exposures. Integration of such variables may substantially improve predictive discrimination in future multicentre studies.

Limitations

Several limitations should be acknowledged. First, the study analysed a relatively small number of PCR-confirmed Lassa fever cases, limiting statistical power and increasing uncertainty around predictor estimates. Future studies should include substantially larger multicentre cohorts to improve model precision. Second, the analysis was based on secondary surveillance data, restricting the availability of potentially important predictors that were not routinely

collected during surveillance activities.

Third, external validation was not performed. Consequently, the generalizability of the Composite Clinical Risk Index to other healthcare settings or populations remains unknown. Independent validation using geographically distinct cohorts represents the next essential stage of model evaluation [9].

Finally, although LASSO regression reduces overfitting, penalization cannot compensate entirely for limited sample size or low event frequency. Therefore, the proposed model should currently be considered a proof-of-concept prediction framework rather than a definitive clinical decision-support tool.

Future Research

Future investigations should prioritize prospective multicentre recruitment of suspected Lassa fever cases to increase the number of confirmed outcomes available for model development. Incorporation of laboratory biomarkers, ecological risk factors, and machine-learning approaches such as gradient boosting and ensemble methods may further improve predictive performance.

External validation across multiple endemic regions should be undertaken before clinical implementation. Decision curve analysis and impact studies are also recommended to determine whether the model improves clinical decision-making and patient outcomes.

References

- Ochu CL, Ntoimo L, Onoh I, Okonofua F, Meremikwu M, et al. (2023) Predictors of Lassa fever diagnosis in suspected cases reporting to health facilities in Nigeria. *Scientific Reports* 13: 6545.
- Duvignaud A, Jaspard M, Etafo IC, Gabillard D, Serra B, et al. (2021) Lassa fever outcomes and prognostic factors in Nigeria (LASCOPE): A prospective cohort study. *The Lancet Global Health* 9(4): E469-E478.
- NCDC (2019) National guidelines for Lassa fever case management. Nigeria Centre for Disease Control.
- Frame JD, Baldwin JM, Gocke DJ, Troup JM (1970) Lassa fever, a new virus disease of man from West Africa. *American Journal of Tropical Medicine and Hygiene* 19: 670-676.
- Cormick McJB, King IJ, Webb PA, Johnson KM, et al. (1987) Clinical observations and virological studies of Lassa fever.
- Richmond JK, Baglole DJ (2003) Lassa fever: Epidemiology, clinical features, and social consequences. *BMJ* 327(7426): 1271-1275.
- NCDC (2024) Lassa fever situation reports. Nigeria Centre for Disease Control
- WHO (2024) Lassa fever (Fact Sheet). World Health Organization.
- Steyerberg EW (2019) *Clinical Prediction Models*. 2nd (Edn.), Springer.
- Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE, et al. (2020) Minimum sample size for developing a multivariable prediction model. *BMJ* 368: m441.
- Wynants L, Calster BV, Collins GS, Riley RD, Heinze G, et al. (2020) Prediction models for diagnosis and prognosis of COVID-19 : systematic review and critical appraisal. *BMJ* 369: m1328.
- Tibshirani R (1996) Regression shrinkage and selection via the LASSO. *J R Stat Soc Series B* 58(1): 267-288.
- Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for GLMs via coordinate descent. *J Stat Softw* 33(1): 1-22.
- Moons KGM, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, et al. (2015) Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): Explanation and elaboration. *Ann Intern Med* 162(1): W1-W73.
- Collins GS, Reitsma JB, Altman DG, Moons KGM (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *Ann Intern Med* 163(1): 1-11.
- Harrell FE (2015) *Regression Modeling Strategies*. 2nd (Edn.), Springer.
- Vickers AJ, Elkin EB (2006) Decision curve analysis. *Medical Decision Making* 26(6): 565-574.
- Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Statistical Learning*. 2nd (Edn.), Springer.
- Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, et al. (2019) PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. *Annals of Internal Medicine* 170(1): W51-W58.
- Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). *Biometrika* 52(3-4): 199-215.

591-611.

21. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, et al. (2009) Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 338: b2393.
22. Sullivan LM, Massaro JM, D'Agostino RB (2004) Presentation of multivariate data for clinical use: The Framingham Study risk score functions. *Statistics in Medicine* 23(10): 1631-1660.
23. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143(1): 29-36.
24. Hosmer DW, Lemeshow S, Sturdivant RX (2013) *Applied Logistic Regression*. 3rd (Edn.). Wiley.
25. Calster VB, Lerneer McDJ, Smeden VM, Wynants L, Steyerberg EW (2019) Calibration: The Achilles heel of predictive analytics. *BMC Medicine* 17(230).
26. Okokhere P, Colubri A, Azubike C, Iruolagbe C, Osazuwa O, et al. (2018) Clinical and laboratory predictors of Lassa fever outcome in Nigeria. *Lancet Infect Dis* 18(6): 684-695.
27. He H, Garcia EA (2009) Learning from imbalanced data. *IEEE Trans Knowl Data Eng* 21(9): 1263-1284.
28. Okwonu FZ, Ahad NA, Hamid H, Sharipov OS (2023) Recursive prediction model: Lassa fever outbreak Nigeria. *Sains Malaysiana* 52(8): 2395-2406.