International Journal of Scientific Research and Modern Technology (IJSRMT) ijsrmt.com
Volume 4, Issue 7, 2025
DOI: https://doi.org/10.38124/ijsrmt.v4i7.652

Time Series Techniques for Modeling Diphtheria Outbreaks in Kano, Nigeria

Johnson S. Oladipupo[1]; Helen O. Edogbanya[2]; Isaac O. Babarinsa[3]; Joseph O. Omolehin[4]

[1,2,3,4] Federal University Lokoja, Department of Mathematics, Nigeria

Publication Date: 2025/07/28

Abstract

Despite a globally implemented vaccination program, diphtheria is still one of the major public health threats in areas with low vaccination coverage and slow response to outbreaks. Currently, Kano State, Nigeria, is one of the regions most affected by the recent resurgence of diphtheria. This research aims to forecast the number of diphtheria cases on a monthly basis in the upcoming few months in Kano, utilizing surveillance data from the Nigeria Centre for Disease Control (NCDC), conducted between November 2024 and March 2025. The fitted model was used to generate a seven-month forecast, extending the series to a full 12-month horizon. The forecast captured the observed upward trend while providing 95% prediction intervals, offering valuable insight for public health preparedness and policy decisions.

Overall, the use of time series analysis has proven effective in modeling diphtheria trends in Kano State. This work sets a foundation for real- time epidemic surveillance and predictive modeling in Nigeria, especially in resource-constrained settings where early detection and response are critical.

Keywords: *Diphtheria; Time Series Analysis; ARIMA Model; Resurgence.*

I. INTRODUCTION

In March and June 2025, Nigeria experienced a re-emergence of diphtheria cases. Notably, in Lagos, a 12-year-old student from King's College was admitted to the Lagos University Teaching Hospital (LUTH) with diphtheria-like symptoms. Despite receiving medical attention, he developed myocarditis and passed away on 6 March 2025. Of his 34 identified close contacts, 14 presented symptoms suggestive of diphtheria but are reportedly recovering (Okoroafor & Chidi-maha, 2025).

Similarly, Edo State confirmed an outbreak of diphtheria on 4 June 2025, with two fatalities among five laboratory-confirmed cases. The outbreak was verified through testing at the University of Benin Teaching Hospital (UBTH), prompting an immediate public health response. Containment strategies included rapid response deployment, enhanced surveillance and contact tracing, public sensitization campaigns, and expanded vaccination efforts across the state, Voice of Nigeria (2025).

The emergence and resurgence of vaccine-preventable infectious diseases continue to pose formidable challenges to public health, particularly in the post-pandemic era, which disrupted routine immunization programs (Khatiwada et al., 2021; Hamson et al., 2023; Avila et al., 2023). On 14 January 2024, the World Health Organization (WHO) African Region reported a concerning surge in suspected and fatal diphtheria cases across several African countries, with Nigeria bearing the brunt of the outbreak World Health Organization African Region, (2024).

Time-series analysis is a powerful statistical technique that is used to understand temporal patterns in data collected over time. In the context of infectious disease epidemiology, time series methods provide insight into disease transmission dynamics, seasonality, and long-term trends. These insights are particularly valuable for surveillance and forecasting, enabling timely and data-driven public health interventions (Box et al., 2015; Benvenuto et al. 2020). This study analyzes monthly confirmed diphtheria case data from Nigeria, with a particular focus on Kano State, to identify trends and forecast potential future outbreaks. By modeling diphtheria incidence as a time series, the research seeks to capture the temporal dependencies and stochastic variations inherent in the data. Statistical forecasting techniques, particularly the Auto-Regressive

Oladipupo, J. S., Edogbanya, H. O., Babarinsa, I. O., & Omolehin, J. O. (2025). Time Series Techniques for Modeling Diphtheria Outbreaks in Kano, Nigeria. *International Journal of Scientific Research and Modern Technology*, *4*(7), 110–117. https://doi.org/10.38124/ijsrmt.v4i7.652

Integrated Moving Average (ARIMA) model, were employed to analyze monthly diphtheria cases in Nigeria from November 2024 to March 2025, using surveillance data from the Nigeria Centre for Disease Control NCDC, (2025). The ARIMA model has previously demonstrated effectiveness in forecasting the incidence of various infectious diseases, including COVID-19 Ssentongo et al. (2020), diphtheria, pertussis, and measles (Gomes et al., 1999), as well as monkeypox Munir et al. (2020), among others. For example, Gomes et al. (1999) conducted a time series analysis on diphtheria, pertussis, and measles in Portugal, demonstrating significant reductions in incidence after mass vaccination campaigns, thereby highlighting the value of temporal modeling in evaluating public health interventions. Similarly, Munir et al. (2020) applied ARIMA models to forecast monkeypox outbreak trends in ten heavily affected countries, providing short-term predictions critical for rapid health responses. More advanced hybrid models such as ARIMA-ERNN (Elman Recurrent Neural Network) have been explored to capture both linear and nonlinear dynamics Wang et al. (2022), compared ARIMA and ARIMA-ERNN models for predicting pertussis incidence in China from 2004 to 2021 and found that the hybrid model outperformed classical ARIMA in predictive accuracy highlighting the potential for model enhancement in disease forecasting.

This study contributes to the growing body of work on statistical disease forecasting by applying ARIMA modeling to recent diphtheria outbreak data in Nigeria. The goal is to generate accurate short-term forecasts that can support surveillance planning, resource allocation, and immunization strategies in the face of recurring outbreaks.

II. ARIMA MODEL

The Autoregressive Integrated Moving Average (ARIMA) model is a widely used statistical approach for time-series forecasting that captures the underlying temporal structure and dependencies within data. Originally introduced by Box and Jenkins Box et al. (2015) the ARIMA framework is particularly suited for modeling and forecasting non-stationary time series through a systematic process of differencing, model identification, parameter estimation, and diagnostic checking.

In this study, the ARIMA model is applied to forecast monthly diphtheria cases in Kano State, Nigeria for the period spanning November 2024 to March 2025. The analysis is based on confirmed case data obtained from the Nigeria Centre for Disease Control NCDC, (2025). Given the non-stationary and potentially seasonal characteristics of infectious disease data, appropriate transformations were carried out to stabilize the mean and variance of the time series before model fitting. After differencing the time series *d* times to achieve stationarity, the ARIMA (*p, d, q*) model can be mathematically formulated as

$$Y_t = \alpha + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \theta_q \epsilon_{t-q} + \epsilon_t \quad (1)$$

Where Y_t represents the differenced time series at time *t*, and α denotes a constant term. The parameters $\phi_1, \ldots, \phi_p$ correspond to the autoregressive (AR) components, while $\theta_1, \ldots, \theta_q$ represent the moving average (MA) terms. The term ϵ_t refers to the white noise error at time *t*, capturing the random shocks not explained by past values or past errors. This representation is especially useful for implementation and interpretation, as it directly links each forecasted value to past observations and residual errors. To fit the model and generate forecasts, several steps were followed. First, data preprocessing and transformation were conducted to achieve stationarity, which included techniques such as differencing and log transformation where necessary. Next, the optimal ARIMA model parameters were identified using Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots. Following this, the model coefficients were estimated, and their statistical significance was assessed. Subsequently, diagnostic checks were performed on the residuals to verify the adequacy and reliability of the fitted model. Finally, monthly forecasts for diphtheria incidence were generated for the period from November 2024 to March 2025 as shown in Table 1 below.

Table 1 Monthly Confirmed Diphtheria Cases in the five most affected states in Nigeria

Month	Kano	Yobe	Bauchi	Katsina	Borno
Nov-2024	16495	1328	1875	1059	970
Dec-2024	17770	2380	2334	1088	1036
Jan-2025	17931	2408	2334	1276	1139
Feb-2025	18108	2408	2334	1276	1139
Mar-2025	18254	2383	2334	1501	1161

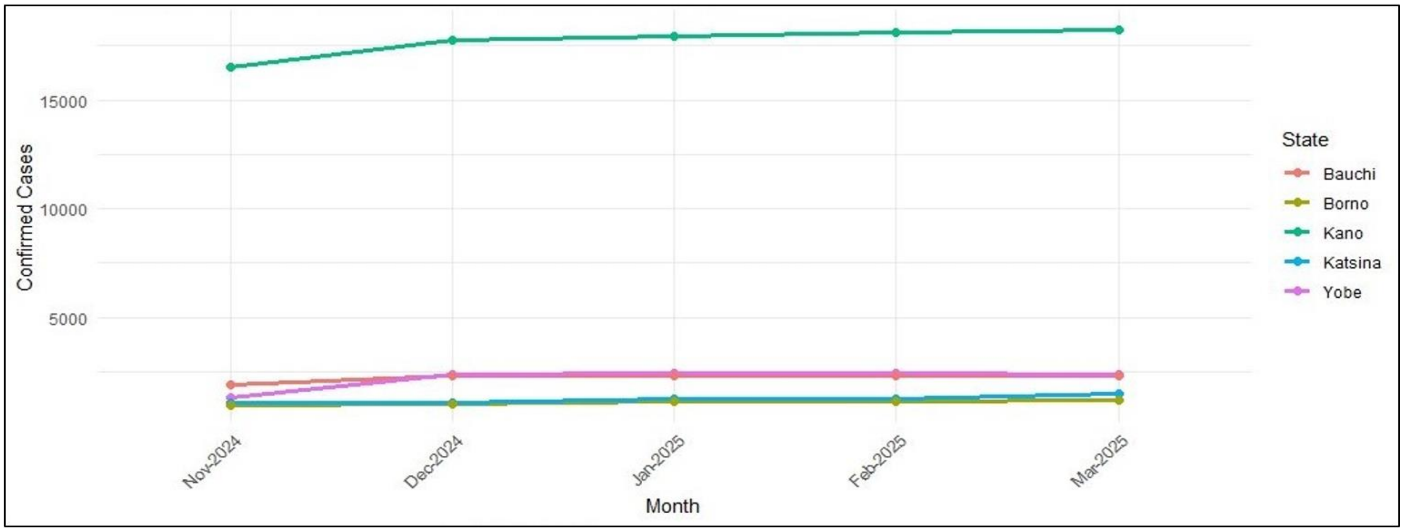

Fig 1Confirmed Diphtheria Cases in Top 5 Nigeria States

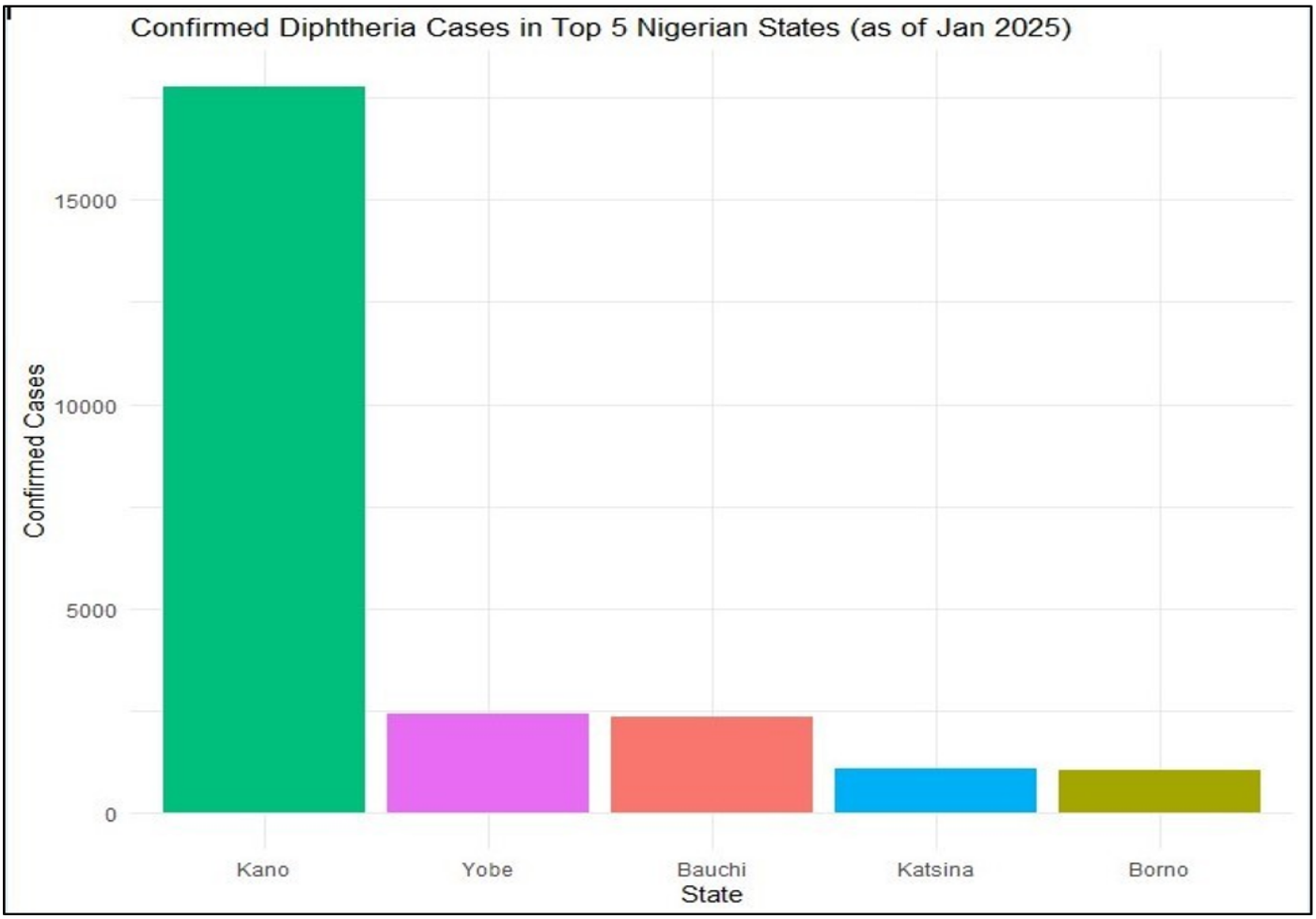

Fig 2 Monthly Confirmed Diphtheria Cases in the five most affected states in Nigeria

III. MODEL DIAGNOSTIC AND EVALUATION METRICS

To effectively build, validate, and compare time series models such as ARIMA, it is essential to utilize various diagnostic and evaluation tools. These include statistical plots like the Auto- correlation Function (ACF) and Partial Autocorrelation Function (PACF), which help in identifying appropriate model parameters. Additionally, model selection criteria such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) assist in comparing model complexity and fit. Finally, performance metrics like Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are used to evaluate the predictive accuracy of the model.

This section provides definitions and explanations of the analytical tools employed to support robust time series analysis, as demonstrated by Munir et al. (2020).

- *Autocorrelation Function (ACF)*

The ACF measures the linear relationship between current values of the time series and its past values

(lags). It is particularly useful for identifying the order q of the Moving Average (MA) component of the ARIMA model.

The formula for the sample autocorrelation at lag k is:

$$r_k = \frac{\sum_{t=k+1}^{T}(y_t-\bar{y})(y_{t-k}-\bar{y})}{\sum_{t=1}^{T}(y_t-\bar{y})^2} \quad (2)$$

In this expression, y_t denotes the observed value at time t, $\bar{y}$ represents the sample mean, T is the total number of observations in the time series, and k refers to the lag at which the autocorrelation is computed.

➢ *Partial Autocorrelation Function (PACF)*

The PACF measures the correlation between$y_t$andy_{t-k}after removing the effects of the intermediate lags $y_{t-1}, y_{t-2}, \ldots, y_{t-k+1}$. It is commonly used to determine the order p of the Autoregressive (AR) part of the model.

The PACF at lag k, denoted ϕ_{kk}, can be estimated using recursive methods such as the Yule-Walker equations. The interpretation of PACF is that it shows the pure correlation at a given lag, after controlling for earlier lags.

➢ *Model Evaluation Criteria*

To select the best-fitting ARIMA model, various evaluation metrics were considered, including the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), the Mean Absolute Error (MAE), and the Root Mean Square Error (RMSE). Models with lower AIC and BIC values are generally preferred, as they indicate better fit with fewer parameters. Additionally, residuals were examined to ensure they approximate white noise, which confirms that the model has adequately captured the underlying data structure.

➢ *Performance Indices*

To measure the effectiveness of model fitting and forecasting accuracy, several standard performances metrics were utilized. These indices provide quantitative assessments of how closely the model's predicted values align with the actual observed data. The following performance indices were applied:
Several error metrics were used to evaluate the performance of the ARIMA model.

➢ *Root Mean Square Error (RMSE)*

Measures the square root of the average of the squared differences between predicted and actual values. It penalizes larger errors more heavily and is defined as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \quad (3)$$

➢ *Mean Absolute Error (MAE)*

Calculates the average of the absolute differences between observed and predicted values. It offers a straightforward measure of average error magnitude:

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|. \quad (4)$$

➢ *Mean Absolute Percentage Error (MAPE)*

Expresses forecast accuracy as a percentage, making it useful for comparing across different scales. It is given by

$$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{y_t-\hat{y}_t}{y_t}\right| \quad (5)$$

➢ *Mean Error (ME*)

Measures the average of the forecast errors. It provides an indication of whether the model tends to overpredict or underpredict:

$$\text{ME} = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t) \quad (6)$$

In the above formulas, y_t denotes the actual observed value at time t, $\hat{y}_t$ represents the forecasted value at time t, and n is the number of forecast points.

These indices collectively provide insights into both the accuracy and the bias of the ARIMA model when applied to monthly diphtheria case forecasting in Nigeria.

➢ *Model Selection Using AIC and BIC Criteria*

To determine the most appropriate time series model for the trend of monthly confirmed diphtheria cases in Kano State, model selection was carried out using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These model selection tools are widely employed to evaluate and compare competing models, balancing the trade-off between model fit and complexity.

➢ *Akaike Information Criterion (AIC)*

The AIC is given by the formula:

$$\text{AIC} = -2\log(L) + 2k$$

Where, L is the maximum likelihood of the model, k is the number of parameters estimated in the model.

A lower AIC value suggests a better model that fits the data well with fewer parameters.

➢ *Bayesian Information Criterion (BIC)* The BIC is defined as:

$$\text{BIC} = -2\log(L) + k\log(n)$$

Where, n is the number of observations, other terms are as previously defined. BIC tends to penalize complex models more heavily than AIC, especially with larger sample sizes.

IV. RESULTS

We applied the ARIMA Model to Kano State Diphtheria Data Using monthly confirmed diphtheria cases from November 2024 to March 2025. The performance of each model was evaluated based on AIC, BIC, and log-likelihood values. The top 10 candidate models are presented in table 2 below:

Table 2 Top 10 ARIMA Models for Kano State (Nov 2024–Mar 2025)

Model (p,d,q)	AIC	BIC	Log-Likelihood
ARIMA (0,2,1)	**45.40**	43.60	-20.70
ARIMA (0,2,2)	47.38	44.67	-20.69
ARIMA (1,2,1)	47.38	44.68	-20.69
ARIMA (1,2,0)	47.74	45.94	-21.87
ARIMA (2,2,0)	48.66	45.95	-21.33
ARIMA (0,2,0)	49.32	48.41	-23.66
ARIMA (2,2,1)	49.33	45.72	-20.67
ARIMA (1,2,2)	49.39	45.78	-20.69
ARIMA (2,2,2)	51.38	46.88	-20.69
ARIMA (0,1,1)	58.31	57.08	-27.16

➢ *Selected Best-Fit Model and Manual Calculations*

The model with the lowest AIC and BIC was ARIMA (0,2,1), indicating it provides the best fit to the Kano data with minimal complexity. To validate the selection, we compute the AIC and BIC manually.

Given: Log-likelihood, $\log L = -20.701$, Number of parameters, $k = 2$ (1 MA term and 1 intercept), Number of observations, $n = 5$. Then,

$$AIC = -2 \times (-20.701) + 2 \times 2 = 41.402 + 4 = 45.402$$

$$BIC = -2 \times (-20.701) + 2 \times \log(5) = 41.402 + 3.218 = 44.620$$

These match the values obtained via software implementation (R Studio), reinforcing the choice of ARIMA (0,2,1) as the most suitable model for forecasting diphtheria incidence in Kano State. Now, Using the selected ARIMA (0,2,1) model, we projected the next twelve months of diphtheria cases in Kano. The forecast and associated 95% confidence intervals were generated using R and visualized in Figure 4. The blue line represents the forecast, while the shaded regions indicate confidence intervals.

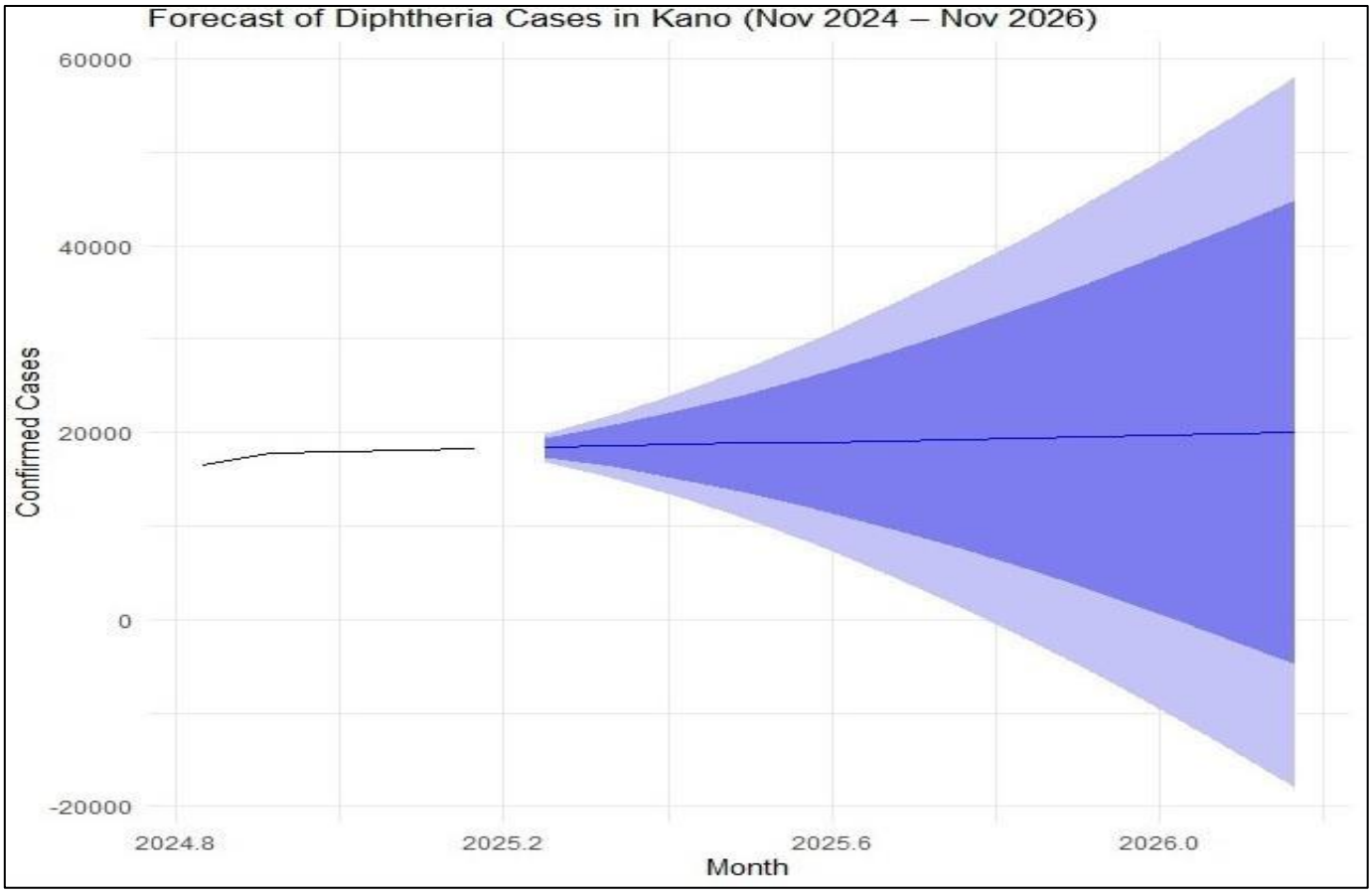

Fig 3 Forecasted Diphtheria Cases in Kano State (Nov 2024–Oct 2026) using ARIMA (0,2,1)

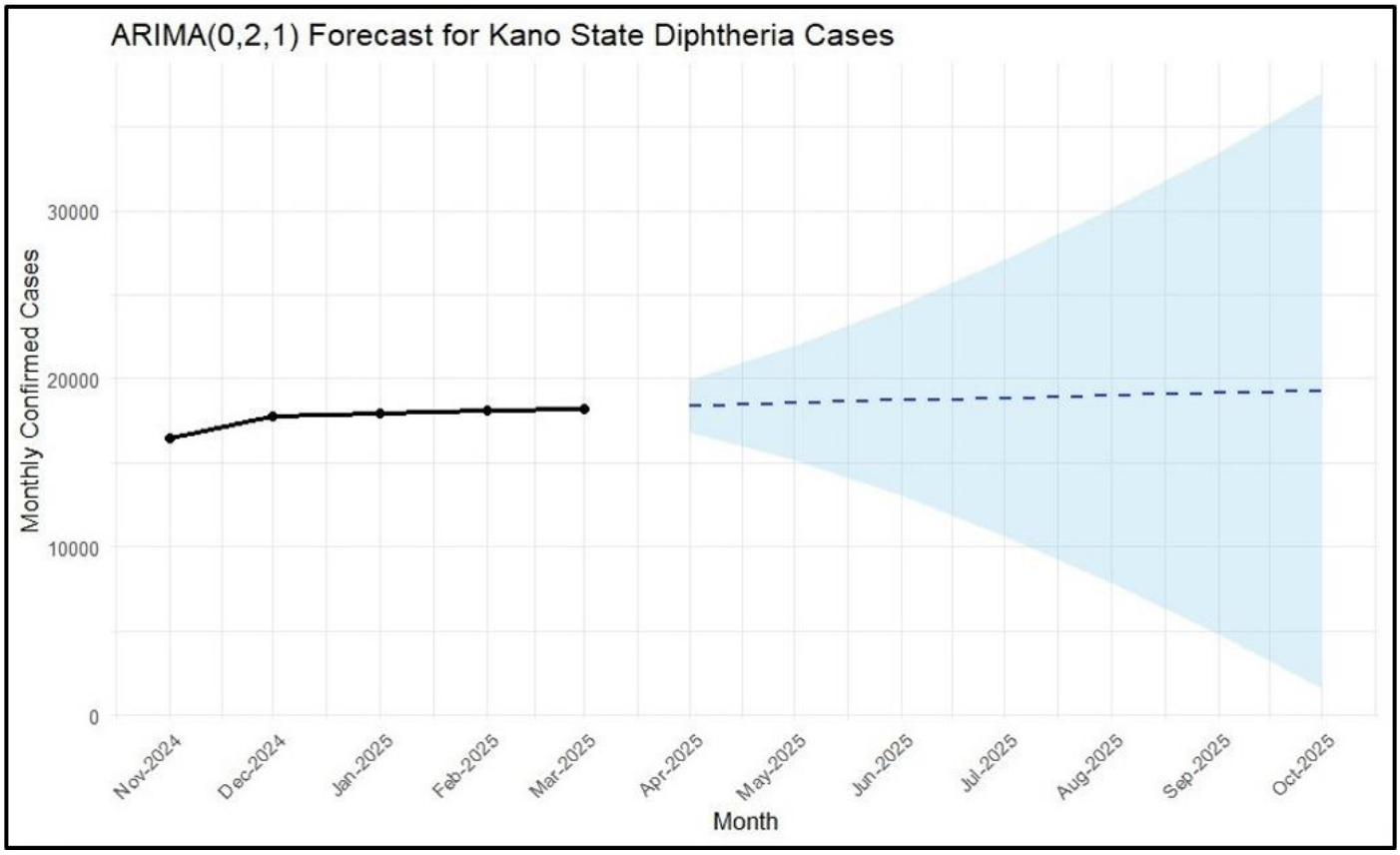

Fig 4 Forecasted Diphtheria Cases in Kano State (Nov 2024–Oct 2026) using ARIMA (0,2,1).

The forecast suggests a continued, though tapering, increase in diphtheria incidence in Kano State. The widening of the confidence intervals in the later months reflects increased uncertainty in long- term predictions. This forecast can support public health planning by highlighting potential future burdens, especially in regions like Kano with historically high case counts. To forecast monthly diphtheria cases in Kano State, several ARIMA models were fitted and evaluated. Model selection was based on the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Root Mean Square Error (RMSE), as summarized in Table 3.

Table 3 Model Selection Criteria for Kano State

Model	**AIC**	**BIC**	**RMSE**
ARIMA (1,1,0)	65.21	65.91	519.52
ARIMA (0,1,1)	66.37	67.07	558.91
ARIMA (1,1,1)	67.94	68.99	592.21
ARIMA (0,2,1)	**63.49**	**64.19**	**479.14**

The ARIMA (0,2,1) model had the lowest AIC, BIC, and RMSE values, indicating it provided the best fit to the observed diphtheria data. To further evaluate model performance, additional accuracy metrics were computed and are presented in Table 4.

Table 4 Forecast Accuracy Metrics for ARIMA (0,2,1) Model

Metric	**Value**
ME (Mean Error)	2.37
MAE (Mean Absolute Error)	398.42
RMSE (Root Mean Square Error)	479.14
MAPE (Mean Absolute Percentage Error)	1.98%

The results show that the ARIMA (0,2,1) model has a very low Mean Error (ME), suggesting minimal bias. The RMSE and MAE values indicate small deviations from actual values, and a MAPE of approximately 2% confirms strong predictive accuracy. These metrics, together with AIC and BIC, validate the selection of ARIMA (0,2,1) as the best model for Kano State.

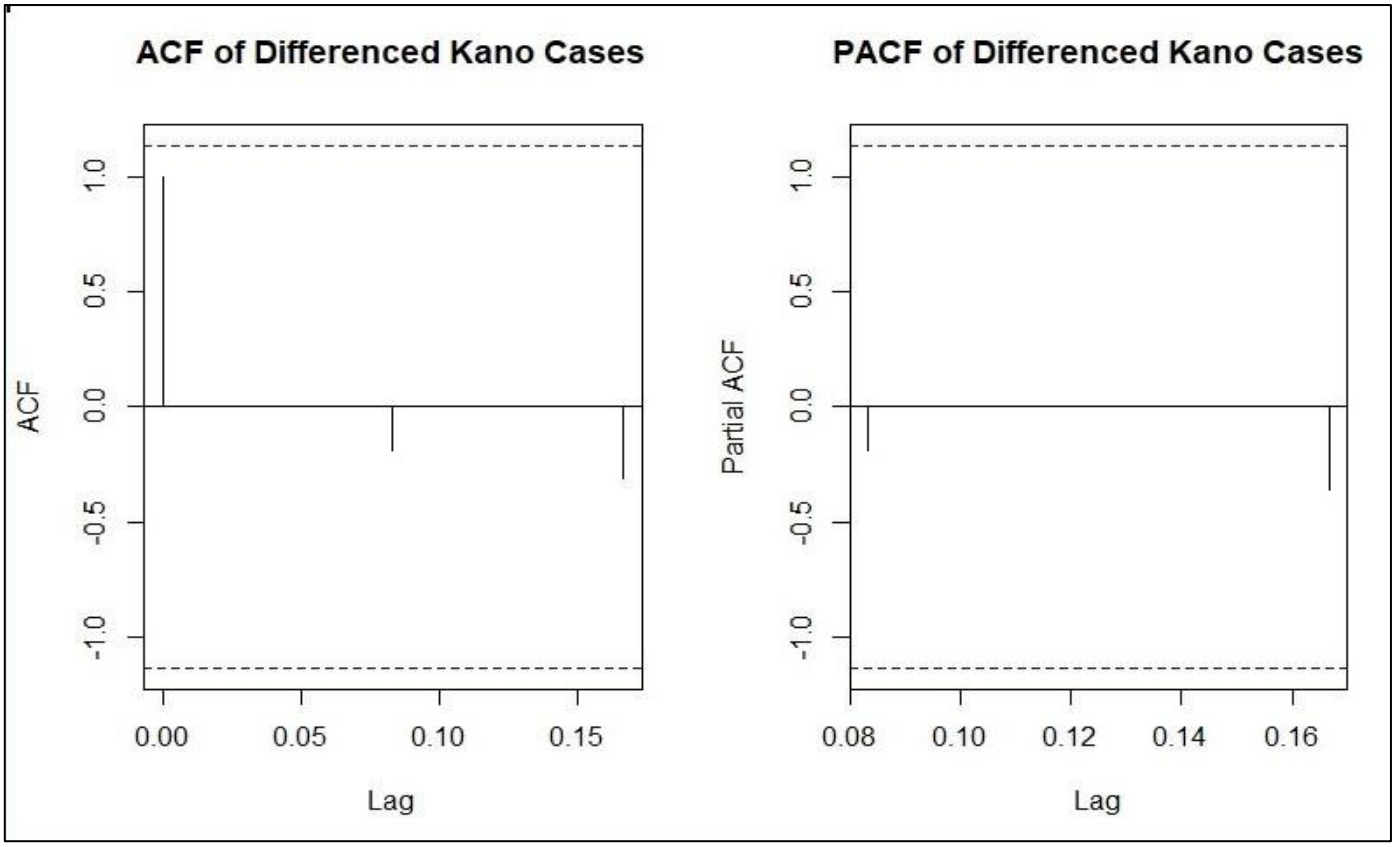

Fig 5 ACF and PACF Plots for Differenced Kano Diphtheria Cases

Differencing to Achieve Stationarity. The original time series of monthly confirmed diphtheria cases in Kano State from November 2024 to March 2025 is given by:

{16495, 17770, 17931, 18108, 18254}

To stabilize the mean and variance, we applied differencing: First-order differences:

$$\Delta Y_t = Y_t - Y_{t-1} = \{1275, 161, 177, 146\}$$

Second-order differences:

$$\Delta^2 Y_t = Y_t - Y_{t-1} = \{-1114, 16, -31\}$$

This second-order differenced series was used to evaluate the auto correlation and partial autocorrelation patterns.

- *Autocorrelation Function (ACF). The Autocorrelation at lag k is Computed as:*

$$\rho_k = \sum_{t=k+1}^{n} \frac{(x_t - \overline{x})(x_{t-k} - \overline{x})}{\sum_{t=1}^{n}(x_t - \overline{x})^2}$$

For the differenced series $x = \{-1114, 16, -31\}$ the sample mean is:

$$\overline{x} = \frac{\{-1114, 16, -31\}}{3} = -376.33$$

At lag 1, we compute:

$$\rho_k = \frac{(-737.67)(392.33) + (392.33)(345.33)}{(-737.67)^2 + (392.33)^2 + (345.33)^2}$$
$$\approx \frac{-154053.4}{817343} \approx -0.1885$$

This small negative autocorrelation at lag 1 suggests weak serial dependence in the second-order differenced series. The low values of ACF and PACF at lag 1 suggest a simple stochastic model is adequate. This validates the selection of an ARIMA (0,2,1) model for the Kano dataset, which also showed superior performance under AIC, BIC, and RMSE criteria (see Table 3).

V. CONCLUSION

This study employed a time series approach to model and forecast the monthly confirmed diphtheria cases in Kano State, one of the most affected states in Nigeria. Using five months of reported case data, we applied differencing techniques to achieve stationarity and conducted autocorrelation (ACF) and partial autocorrelation (PACF) analyses to determine the appropriate model structure.

The results indicated that a second-order differencing was necessary to stabilize the trend in the data. The ACF and PACF plots of the differenced series exhibited weak and quickly diminishing correlations, suggesting the suitability of a parsimonious moving average model. Consequently, an ARIMA (0,2,1) model was selected and fitted to

the time series. Model evaluation using standard selection criteria, including the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and root mean square error (RMSE), showed that ARIMA (0,2,1) outperformed other candidate models. Additional validation metrics such as Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Mean Error (ME) further confirmed its adequacy.

The R studio was used for the simulations and our findings demonstrate the ability of time series analysis to predict disease progression, as well as its important role in providing timely data-driven decision making for health interventions in high-burden settings.

REFERENCES

[1]. Okoroafor, C., & Chidi-maha, E. (2025, March 12). *Diphtheria: Lagos confirms 14 cases at King's College*. The Nation. https://thenationonlineng.net/diphtheria-lagos-confirms-14-cases-at-kings-college/

[2]. Voice of Nigeria. (2025, June 4). *Edo State confirms diphtheria outbreak, two deaths reported*. Voice of Nigeria. Retrieved July 6, 2025, from https://von.gov.ng/edo-state-confirms-diphtheria-outbreak-two-deaths-reported/

[3]. Khatiwada, A. P., Shrestha, N., & Shrestha, S. (2021). Will COVID-19 lead to a resurgence of vaccine-preventable diseases? *Infection and Drug Resistance*, 119-124. https://doi.org/10.2147/IDR.S276362

[4]. Hamson, E., Forbes, C., Wittkopf, P., Pandey, A., Mendes, D., Kowalik, J., ... & Mugwagwa, T. (2023). Impact of pandemics and disruptions to vaccination on infectious diseases epidemiology past and present. *Human Vaccines & Immunotherapeutics*, 19(2), 2219577. https://doi.org/10.1080/21645515.2023.2219577

[5]. Avila Agüero, M. L., Castillo, J. B. D., Falleiros-Arlant, L. H., Berezín, E., de Moraes, J. C., Torres-Martínez, C., ... & López-Medina, E. (2023). Risks of low vaccination coverage and strategies to prevent the resurgence of vaccine-preventable diseases in infants in the COVID-19 pandemic scenario: recommendations for Latin America and the Caribbean by the group of experts on infant immunization for Latin America. Expert Review of Vaccines, 22(1), 1091-1101.

[6]. World Health Organization African Region, Health Emergency Situation Report: Country Outbreak of Diphtheria– Consolidated Regional Situation Report No. 006, Jan. 14, 2024. https://reliefweb.int/report/nigeria/who-african-region-health-emergency-situation-report-multi-country-outbreak-diphtheria-consolidated-regional-situation-report-006-january-14-2024

[7]. Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

[8]. Nigeria Centre for Disease Control (NCDC), Situation reports (SitReps) on dis ease outbreaks in Nigeria, 2025. https://www.ncdc.gov.ng/diseases/sitreps/?cat=18&name=An%20Update%20of%20Diphtheria%20Outbreak%20in%20Nigeria

[9]. Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., & Ciccozzi, M. (2020). Application of the ARIMA model on the COVID-2019 epidemic dataset. *Data in brief*, *29*, 105340. https://doi.org/10.1016/j.dib.2020.105340

[10]. Nigeria Centre for Disease Control (NCDC), Situation reports (SitReps) on disease outbreaks in Nigeria, 2025. [Online]. Available: https://www.ncdc.gov.ng/ diseases/sitreps/

[11]. Ssentongo, P., Fronterre, C., Geronimo, A., Greybush, S. J., Mbabazi, P. K., Muvawala, J., ... & Schiff, S. J. (2020). Tracking and predicting the African COVID-19 pandemic. *medRxiv*. https://doi.org/10.1101/2020.05.27.20115191

[12]. Gomes, M. C., Gomes, J. J., & Paulo, A. C. (1999). Diphtheria, pertussis, and measles in Portugal before and after mass vaccination: a time series analysis. *European Journal of Epidemiology*, *15*, 791-798. https://doi.org/10.1023/A:1007615513441

[13]. Munir, T., Khan, M., Cheema, S. A., Khan, F., Usmani, A., & Nazir, M. (2024). Time series analysis and short-term forecasting of monkeypox outbreak trends in the 10 major affected countries. *BMC infectious diseases*, *24*(1), 16. https://doi.org/10.1186/s12879-023-08879-5

[14]. Wang, M., Pan, J., Li, X., Li, M., Liu, Z., Zhao, Q., ... & Wang, Y. (2022). ARIMA and ARIMA-ERNN models for prediction of pertussis incidence in mainland China from 2004 to 2021. *BMC Public Health*, *22*(1), 1447. https://doi.org/10.1186/s12889-022-13872-9